DAGSTUHL
REPORTS

Volume 15, Issue 9, September 2025

Interactions in Constraint Optimization (Dagstuhl Seminar 25371)
Katalin Fazekas, Matti Järvisalo, Nina Narodytska, Peter J. Stuckey, and Christoph Jabs 1

Precision in Geometric Algorithms (Dagstuhl Seminar 25372)
Mikkel Abrahamsen, Sándor Kisfaludi-Bak, Linda Kleist, and Till Miltzow 21

Open Scholarly Information Systems: Status Quo, Challenges, Opportunities (Dagstuhl Seminar 25381)
Hannah Bast, Guillaume Cabanac, Paolo Manghi, Jian Wu, and Marcel R. Ackermann 38

Quantum Error Correction Meets ZX-Calculus (Dagstuhl Seminar 25382)
Miriam Backens, Aleks Kissinger, John van de Wetering, Michael Vasmer, and Sarah Meng Li 58

Retrieval-Augmented Generation – The Future of Search? (Dagstuhl Seminar 25391)
Matthias Hagen, Josiane Mothe, Smaranda Muresan, Martin Potthast, Min Zhang, Benno Stein, and Sebastian Heineking 71

Specification Engineering: Foundations for the Future of Software Development (Dagstuhl Seminar 25392)
Marsha Chechik, Eunsuk Kang, Shahar Maoz, Jan Oliver Ringert, and Allison Sullivan 160

Societal Impact of Computational Social Choice (Dagstuhl Seminar 25401)
Martin Lackner, Nicholas Mattei, Arianna Novaro, Clemens Puppe, and Ratip Emin Berker 183

ISSN 2192-5283

Published online and open access by
Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany. Online available at https://www.dagstuhl.de/dagpub/2192-5283

Publication date
May, 2026

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at `https://dnb.d-nb.de`.

Digital Object Identifier: 10.4230/DagRep.15.9.i

Aims and Scope
The periodical *Dagstuhl Reports* documents the program and the results of Dagstuhl Seminars and Dagstuhl Perspectives Workshops.
In principal, for each Dagstuhl Seminar or Dagstuhl Perspectives Workshop a report is published that contains the following:

- an executive summary of the seminar program and the fundamental results,
- an overview of the talks given during the seminar (summarized as talk abstracts), and
- summaries from working groups (if applicable).

This basic framework can be extended by suitable contributions that are related to the program of the seminar, e. g. summaries from panel discussions or open problem sessions.

Contact
Schloss Dagstuhl – Leibniz-Zentrum für Informatik
Dagstuhl Reports, Editorial Office
Oktavie-Allee, 66687 Wadern, Germany
reports@dagstuhl.de
https://www.dagstuhl.de/dagrep

Report from Dagstuhl Seminar 25371

Interactions in Constraint Optimization

Katalin Fazekas[*1], **Matti Järvisalo**[*2], **Nina Narodytska**[*3], **Peter J. Stuckey**[*4], **and Christoph Jabs**[†5]

1 **TU Wien, AT.** katalin.fazekas@tuwien.ac.at
2 **University of Helsinki, FI.** matti.jarvisalo@helsinki.fi
3 **VMware Research – Palo Alto, US.** n.narodytska@gmail.com
4 **Monash University – Caulfield, AU.** peter.stuckey@monash.edu
5 **University of Helsinki, FI.** christoph.jabs@helsinki.fi

Abstract

This report documents the Dagstuhl Seminar 25371 "Interactions in Constraint Optimization". Our Dagstuhl Seminar gathered 41 researchers from 15 countries, working on different constraint optimization paradigms. The report consists of an executive summary, and abstracts on tutorials, research talks, and panel discussions.

Seminar September 7–12, 2025 – https://www.dagstuhl.de/25371
2012 ACM Subject Classification Mathematics of computing → Combinatorial algorithms; Theory of computation → Discrete optimization
Keywords and phrases constraint programming, maximum satisfiability, mixed integer linear programming, optimization modulo theories, pseudo-boolean optimization
Digital Object Identifier 10.4230/DagRep.15.9.1

1 Executive Summary

Katalin Fazekas (TU Wien, AT)
Matti Järvisalo (University of Helsinki, FI)
Nina Narodytska (VMware Research – Palo Alto, US)
Peter J. Stuckey (Monash University – Caulfield, AU)

Optimization problems are one of the most common problems that real-world applications face: scheduling, rostering, hardware verification, vehicle routing, to name a few. Constraint optimization is the main technology to solve these problems in practice as it offers a principled way to explore the search space and find an optimal (or good enough) solution. However, as the complexity of real-world applications increases and these techniques are applied in safety critical applications, the need for more advanced features and techniques to be developed grows. In fact, the term constraint optimization can be seen as an umbrella term that covers a number of well-developed and practically useful technologies. Our main goal was to see how these technologies can leverage each other to increase practicality and adoption of optimization techniques.

This report documents the program and the outcomes of Dagstuhl Seminar 25371 "Interactions in Constraint Optimization". We gathered leading researchers from following research areas working on constraint optimization:

- Constraint Programming (CP),
- Mixed Integer Linear Programming (MIP),
- Boolean Satisfiability (SAT) / Maximum Satisfiability (MaxSAT),

[*] Editor / Organizer
[†] Editorial Assistant / Collector

- Satisfiability Modulo Theories (SMT) / Optimization Modulo Theories (OMT/MaxSMT), and
- Pseudo-Boolean optimization (PBO).

The main focus of the seminar was to foster interactions between these communities and consider how to build synergy between them in the coming decade. The seminar was structured to provide necessary background to the participants through tutorials presented by leading researchers in the corresponding subfield of optimization. These tutorials aimed to bring the audience up to speed on the state-of-the-art advances in each subarea. We had a number of contributing talks that reported on recent results and in-progress research developments. These fostered interesting discussions and outlined potential collaborations.

We held two panels to discuss future directions and collaborative opportunities. We believe we had fruitful discussions that triggered new ideas. The report presents a summary of talks, discussions and panel outcomes.

2 Table of Contents

Executive Summary
Katalin Fazekas, Matti Järvisalo, Nina Narodytska, and Peter J. Stuckey 1

Overview of Talks

Uncovering and Classifying Bugs in MaxSAT Solvers through Fuzzing and Delta Debugging
Armin Biere 5

Parallelising CP-SAT
Toby Davies, Frédéric Didier, Laurent Perron, and Peter J. Stuckey 5

MIP aspects of Google OR-tools CP-SAT solver
Frédéric Didier 5

Large Neighbourhood Search (LNS)
Pierre Flener 6

Towards Quantifying Fairness and Designing Interpretable Machine Learning: A Formal Methods Approach
Bishwamittra Ghosh 6

Tutorial: LP relaxations in MIP solving
Ambros Gleixner and Gioni Mexi 6

Constraint Optimisation as an accessible AI toolbox
Tias Guns 7

Tutorial: Maximum Satisfiability Solving
Alexey Ignatiev 7

IHS for PBO: Key Techniques in a State-of-the-Art Solver and Recent Developments
Hannes Ihalainen 8

Multi-Objective Optimization: A Pseudo-Boolean Perspective
Christoph Jabs 8

Core-Guided Linear Programming-based Maximum Satisfiability
George Katsirelos 9

Integrating Column Generation and Large Neighborhood Search
Lucas Kletzander 9

Solving the Identifying Code Set Problem with Grouped Independent Support
Anna Latour 10

Lower Bound in Branch-and-Bound (BnB) MaxSAT Solvers
Chu Min Li 10

Accelerating Column Generation via Template Pricing
Luke Marshall 11

My point of view on BDD/ZDD, SAT, and PB techniques for constraint optimization
Shin-ichi Minato 11

TT-Open-WBO-Inc: The SAT Engine of Modern Anytime MaxSAT
Alexander Nadel 12

SpotIT: Evaluating Text-to-SQL Evaluation with Formal Verification
Nina Narodytska 12

Certified Implicit Hitting Set Solving for Pseudo-Boolean Optimization
Jakob Nordström 13

Symbolic Conflict Analysis in Pseudo-Boolean Optimization
Albert Oliveras 13

Tutorial on Dantzig-Wolfe decomposition, column generation, and branch-price-and-cut
Elina Rönnberg 14

Quantum computing for discrete optimization: A glimpse into three technologies
Philine Schiewe 14

Weighted CP and related frameworks: soft arc consistency and bounds
Thomas Schiex 15

Introduction to Lazy Clause Generation
Peter J. Stuckey 15

Tutorial: Optimization in SMT
Nestan Tsiskaridze 16

Decision Diagrams for Discrete Optimization
Willem-Jan Van Hoeve 16

Tutorial: Optimization in CP
Hélène Verhaeghe 17

Working groups

How the optimization communities can work better together
Peter J. Stuckey and Luke Marshall 17

Panel discussions

Challenges and opportunities for the next 10 years
Ambros Gleixner, Alexey Ignatiev, Ciaran McCreesh, Thomas Schiex, and Christine Solnon 18

Participants 20

3 Overview of Talks

3.1 Uncovering and Classifying Bugs in MaxSAT Solvers through Fuzzing and Delta Debugging

Armin Biere (Universität Freiburg, DE)

Joint work of Tobias Paxian
Main reference Tobias Paxian, Armin Biere: "Uncovering and Classifying Bugs in MaxSAT Solvers through Fuzzing and Delta Debugging", in Proc. of the 14th International Workshop on Pragmatics of SAT co-located with the 26th International Conference on Theory and Applications of Satisfiability Testing (SAT 2023), Alghero, Italy, July 4, 2023, CEUR Workshop Proceedings, Vol. 3545, pp. 59–71, CEUR-WS.org, 2023.
URL https://ceur-ws.org/Vol-3545/paper5.pdf

We give a brief introduction into existing fuzzing and delta debugging techniques in automated reasoning with focus on SAT including model based testing of SAT solver APIs. Then we discuss extensions to MaxSAT and present results on cross checking MaxSAT solvers submitted to the last three MaxSAT evaluations through these techniques.

3.2 Parallelising CP-SAT

Toby Davies (Google – Pyrmont, AU), Frédéric Didier (Google – Paris, FR), Laurent Perron, and Peter J. Stuckey (Monash University – Caulfield, AU)

Main reference Toby O. Davies, Frédéric Didier, Laurent Perron, Peter J. Stuckey: "Parallelising Lazy Clause Generation with Trail Sharing", in Proc. of the Integration of Constraint Programming, Artificial Intelligence, and Operations Research, pp. 205–221, Springer Nature Switzerland, 2025.
URL https://doi.org/10.1007/978-3-031-95973-8_13

CP-SAT implements many complementary ideas and runs them in a parallel portfolio, but this is just part of its parallelism approach. Workers share information, about variable and objective bounds, and a subset of high-quality clauses. In addition to these traditional approaches "shared tree workers" also partition the search space, and perform "trail sharing" to complement traditional clause sharing.

3.3 MIP aspects of Google OR-tools CP-SAT solver

Frédéric Didier (Google – Paris, FR)

While CP-SAT focuses on pure integer-problems, a lot of its design is similar to the one of a "classic" MIP solver and they can be used to solve optimization problems in mostly the same way. After explaining the class of MIP problems that CP-SAT can deal with, I will dive into some of its implementation details. I will focus on the handling of linear constraints and how we use the LP relaxation of the problem, both being critical components like they are in MIP.

3.4 Large Neighbourhood Search (LNS)

Pierre Flener (Uppsala University, SE)

LNS is a hybrid solving technology for optimisation problems. A local-search algorithm invokes at every iteration a systematic-search solver (usually a CP solver) in order to explore a large neighbourhood that it constructs. I give a brief tutorial on LNS, show how to extend it to satisfaction problems, and outline its research frontier.

3.5 Towards Quantifying Fairness and Designing Interpretable Machine Learning: A Formal Methods Approach

Bishwamittra Ghosh (MPI-SWS – Kaiserslautern, DE)

Joint work of Bishwamittra Ghosh, Kuldeep Meel, Debabrota Basu, Dmitry Malioutov
Main reference Bishwamittra Ghosh, Debabrota Basu, Kuldeep S. Meel: "Justicia: A Stochastic SAT Approach to Formally Verify Fairness", in Proc. of the Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pp. 7554–7563, AAAI Press, 2021.
URL https://doi.org//10.1609/AAAI.V35I9.16925

Despite its widespread success, machine learning faces critical societal challenges, including unfair predictions and a lack of interpretability. My research, at the intersection of formal methods and machine learning, develops tools based on formal reasoning, satisfiability solving, and constrained optimization to formally quantify classifier fairness, identify sources of unfairness, and design inherently interpretable rule-based classifiers. In particular, I will discuss the application of stochastic satisfiability (SSAT) and maximum satisfiability (MaxSAT) in addressing fairness and interpretability, advancing the goal of trustworthy machine learning.

3.6 Tutorial: LP relaxations in MIP solving

Ambros Gleixner (HTW – Berlin, DE) and Gioni Mexi (Zuse-Institut Berlin, DE)

In this tutorial we survey the basics of how modern MIP solving uses, solves, and strengthens LP relaxations with pointers to recent trends and computational aspects. In particular, we touch upon valid inequalities and the geometry behind infeasibility analysis and Dantzig-Wolfe reformulation.

3.7 Constraint Optimisation as an accessible AI toolbox

Tias Guns (KU Leuven, BE)

URL https://wms.cs.kuleuven.be/chat-opt

Our research combines constraint optimisation with machine learning and explainability. For this, we need easy and efficient access to multiple solvers; as most of the work involves novel integrations. This is the case for industry projects, like our CP25 best application paper award on workforce scheduling, on our work evaluating LLMs for constraint modeling, and earlier neuro-symbolic projects like the Sudoku Assistant. To do this research, we had to build the CPMpy library. A Python library that translates to CP, SMT, ILP, PB and SAT solvers. To support all these translations, we converged on an elegant "waterfall" model, showing how much transformations are shared between the different transformations. We'll also highlight gaps and new possiblities: the need for standard APIs, for open source implementations of translations and components, and how this can enable cross-technology evaluations and dataset generation.

3.8 Tutorial: Maximum Satisfiability Solving

Alexey Ignatiev (Monash University – Clayton, AU)

This talk provides a tutorial on Maximum Satisfiability (MaxSAT) solving, a powerful declarative paradigm for tackling complex optimisation problems across a wide range of domains. The tutorial begins by defining the MaxSAT problem as an optimisation counterpart of Propositional Satisfiability. It then covers the fundamentals of representing practical problem constraints as propositional formulas in conjunctive normal form (CNF), using hard and soft clauses, illustrating how non-clausal constraints can be "clausified" using techniques like Tseitin transformation. The core of the talk focuses on state-of-the-art, core-guided MaxSAT algorithms. It explains the foundational principle of iteratively identifying and relaxing unsatisfiable cores. The presentation details the evolution of these methods, leading to the OLL algorithm, which introduces the key concept of relaxable cardinality constraints. Finally, it discusses practical considerations and performance enhancements as implemented in the open-source RC2 solver, including incremental SAT solving, incremental cardinality constraints, core exhaustion, Boolean lexicographic optimisation, and stratification.

3.9 IHS for PBO: Key Techniques in a State-of-the-Art Solver and Recent Developments

Hannes Ihalainen (University of Helsinki, FI)

Joint work of Hannes Ihalainen, Dieter Vandesande, André Schidler, Jeremias Berg, Bart Bogaerts, Matti Järvisalo
Main reference Hannes Ihalainen, Dieter Vandesande, André Schidler, Jeremias Berg, Bart Bogaerts, Matti Järvisalo: "Efficient and Reliable Hitting-Set Computations for the Implicit Hitting Set Approach", CoRR, Vol. abs/2508.07015, 2025.
URL https://doi.org//10.48550/ARXIV.2508.07015

Recently, the so-called implicit hitting set (IHS) approach has proven to be a succesful method for pseudo-Boolean optimization (PBO). To achieve practical competitiveness, IHS solvers incorporate many additional techniques in addition to the basic IHS algorithm. This talk provided an overview of the techniques implemented in a state-of-the-art PBO-IHS solver. In addition, the talk highlighted some recent advances in PBO-IHS: a symmetric core learning technique to tackle highly symmetric instances, and efforts to develop alternative methods – such as PB reasoning and stochastic local search – for hitting set computations, aimed at making solving trustworthy via proof logging and improving performance in practice.

3.10 Multi-Objective Optimization: A Pseudo-Boolean Perspective

Christoph Jabs (University of Helsinki, FI)

Joint work of Christoph Jabs, Jeremias Berg, Matti Järvisalo
Main reference Christoph Jabs, Jeremias Berg, Matti Järvisalo: "Engineering and Evaluating Multi-objective Pseudo-Boolean Optimizers", in Proc. of the Logics in Artificial Intelligence, pp. 115–134, Springer Nature Switzerland, 2026.
URL https://doi.org/10.1007/978-3-032-04587-4_8

Various real-world settings give rise to combinatorial optimization problems with multiple conflicting objectives, motivating the development of practical approaches to the challenging task of finding Pareto-optimal solutions to declarative models of multi-objective problems. In this talk we view multi-objective optimization through the lens of pseudo-Boolean constraints (MO-PBO) as an extension of propositional clauses and, at the same time, an important class of 0-1 linear constraints. We provide a first-of-kind cross-community evaluation of a selection of recently-proposed approaches applicable to MO-PBO, including first implementations of native MO-PBO algorithms we provide, as well as approaches based on integer linear programming techniques and a translation-based approach to MO-MaxSAT, providing insights into the current state-of-the-art approaches to MO-PBO, and a glimpse into how these paradigms compare more generally.

3.11 Core-Guided Linear Programming-based Maximum Satisfiability

George Katsirelos (INRAE – Palaiseau, FR)

Main reference George Katsirelos: "Core-Guided Linear Programming-Based Maximum Satisfiability", in Proc. of the 28th International Conference on Theory and Applications of Satisfiability Testing, SAT 2025, Glasgow, Scotland, August 12-15, 2025, LIPIcs, Vol. 341, pp. 17:1–17:17, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2025.
URL https://doi.org//10.4230/LIPICS.SAT.2025.17

The core-guided algorithm OLL is the basis of some of the most successful algorithms for MaxSAT in recent evaluations. It works by iteratively finding cores of the formula and transforming it so that it exhibits a higher lower bound. It has recently been shown to implicitly discover cores of the original formula, as well as a compact representation of its reasoning within a linear program. In this paper, we use and extend these results to design a practical MaxSAT solver. We show an explicit linear program which matches and usually exceeds the bound computed by OLL. We show that OLL can be restated as an algorithm that explicitly computes a feasible dual solution of this linear program. This restated algorithm naturally works with an arbitrary dual solution. It can in fact be used to improve any LP representation of the MaxSAT instance. This presents a large increase of the potential design space for such algorithms. We describe some potential improvements from this insight and show that an implementation outperforms the state of the art algorithms on the set of instances from the latest MaxSAT evaluation.

3.12 Integrating Column Generation and Large Neighborhood Search

Lucas Kletzander (TU Wien, AT)

Joint work of Lucas Kletzander, Tommaso Mannelli Mazzoli, Nysret Musliu, Pascal Van Hentenryck
Main reference Lucas Kletzander, Tommaso Mannelli Mazzoli, Nysret Musliu, Pascal Van Hentenryck: "Integrating Column Generation and Large Neighborhood Search for Bus Driver Scheduling with Complex Break Constraints", CoRR, Vol. abs/2505.02485, 2025.
URL https://doi.org//10.48550/ARXIV.2505.02485

This talk presents a framework that integrates column generation with large neighborhood search (LNS) to solve large bus driver scheduling problems with complex break constraints. Branch and price (B&P) is an established technique for such problems, using set partitioning as the master problem, and a resource constrained shortest path problem (RCSPP) as the pricing problem. However, the complex constraints require several optimizations to solve the high-dimensional sub-problem, in particular to split the sub-problem into disjoint sub-problems depending on characteristics like the break scheme, using exponential arc throttling to keep the sub-problem size small for as long as possible, and the use of k-d trees to deal with the current pareto frontier. However, scaling to very large instances is still not possible with B&P. Therefore LNS is used to select subsets of the current solution which are then optimized using column generation (CG). Instead of restarting CG for every subset, a column storage is introduced to reuse columns for future subsets, and a background thread works on the restricted master problem over the whole set of columns. LNS provides state-of-the-art results on instances of all sizes, and the tighter integration shows significant benefits over the naive combination of the technologies.

3.13 Solving the Identifying Code Set Problem with Grouped Independent Support

Anna Latour (TU Delft, NL)

Joint work of Anna L. D. Latour, Arunabha Sen, Kuldeep S. Meel
Main reference Anna L. D. Latour, Arunabha Sen, Kuldeep S. Meel: "Solving the Identifying Code Set Problem with Grouped Independent Support", in Proc. of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China, pp. 1971–1978, ijcai.org, 2023.
URL https://doi.org//10.24963/IJCAI.2023/219

An important problem in network science is finding an optimal placement of sensors in nodes in order to uniquely detect failures in the network. This problem can be modelled as an identifying code set (ICS) problem, introduced by Karpovsky et al. in 1998. The ICS problem aims to find a cover of a set S, such that the elements in the cover define a unique signature for each of the elements of S, and to minimise the cover's cardinality. In this work, we study a generalised identifying code set (GICS) problem, where a unique signature must be found for each subset of S that has a cardinality of at most k (instead of just each element of S). The concept of an independent support of a Boolean formula was introduced by Chakraborty et al. in 2014 to speed up propositional model counting, by identifying a subset of variables whose truth assignments uniquely define those of the other variables. In this work, we introduce an extended version of independent support, grouped independent support (GIS), and show how to reduce the GICS problem to the GIS problem. We then propose a new solving method for finding a GICS, based on finding a GIS. We show that the prior state-of-the-art approaches yield integer-linear programming (ILP) models whose sizes grow exponentially with the problem size and k, while our GIS encoding only grows polynomially with the problem size and k. While the ILP approach can solve the GICS problem on networks of at most 494 nodes, the GIS-based method can handle networks of up to 21 363 nodes; a $\tilde{4}0\times$ improvement. The GIS-based method shows up to a $520\times$ improvement on the ILP-based method in terms of median solving time. For the majority of the instances that can be encoded and solved by both methods, the cardinality of the solution returned by the GIS-based method is less than 10% larger than the cardinality of the solution found by the ILP method.

3.14 Lower Bound in Branch-and-Bound (BnB) MaxSAT Solvers

Chu Min Li (University of Amiens, FR)

Joint work of Shuolin Li, Chu-Min Li, Jordi Coll, Djamal Habet, Felip Manya
Main reference Shuolin Li, Chu-Min Li, Jordi Coll, Djamal Habet, Felip Manyà: "Improving the Lower Bound in Branch-and-Bound Algorithms for MaxSAT", in Proc. of the AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 – March 4, 2025, Philadelphia, PA, USA, pp. 11272–11281, AAAI Press, 2025.
URL https://doi.org//10.1609/AAAI.V39I11.33226

The MaxSAT problem is an optimization version of the satisfiability problem (SAT). A tight lower bound (LB) on the number of falsified soft clauses in a MaxSAT solution is crucial for the efficiency of Branch-and-Bound (BnB) MaxSAT solvers. To compute an LB, modern BnB solvers detect disjoint inconsistent subsets of soft clauses, called cores, using

unit propagation. A notable feature of these solvers is that soft clauses belonging to already detected cores cannot be reused to detect additional cores, limiting the number of cores that can be detected. In this paper, we propose an unlocking mechanism that allows the reuse of soft clauses in already detected cores while ensuring the soundness of LB. Experimental results show that this unlocking mechanism consistently improves the performance of a state-of-the-art BnB solver. In addition, it allowed us to win the first two places in the exact unweighted category of the MaxSAT Evaluation 2024.

3.15 Accelerating Column Generation via Template Pricing

Luke Marshall (Microsoft Research – Redmond, US)

Joint work of Luke Marshall, Santanu Dey, Prachi Shah

Column Generation (CG) is an iterative algorithm effective in solving large-scale integer programming problems. It often relies on the fast re-optimization of the primal simplex algorithm for suitable performance; however, CG is notorious for convergence issues (with degenerate problems).

Although there is much research on stabilization techniques to avoid these issues, I'll introduce a new approach "Template Pricing" that can converge orders of magnitude faster. I'll illustrate with a simple example and give insights why it works so well (with computational results). Specifically, the choice of "what" columns to add in each iteration can make a significant impact on re-optimization performance.

3.16 My point of view on BDD/ZDD, SAT, and PB techniques for constraint optimization

Shin-ichi Minato (Kyoto University, JP)

Joint work of Shin-ichi Minato, Jun Kawahara, Mutsunori Banbara, Takashi Horiyama, Ichigaku Takigawa, Yutaro Yamaguchi
Main reference Shin-ichi Minato, Jun Kawahara, Mutsunori Banbara, Takashi Horiyama, Ichigaku Takigawa, Yutaro Yamaguchi: "Fast enumeration of all cost-bounded solutions for combinatorial problems using ZDDs", Discrete Applied Mathematics, Vol. 360, pp. 467–486, 2025.
URL https://doi.org//10.1016/j.dam.2024.10.003

Recently I'm interested in integrating the techniques for enumeration, optimization, and satisfiability. A task of constraint optimization is strongly related to those techniques, to generate all cost-bounded solutions satisfying a given combinatorial constraint. In this talk, I will review a classical SAT-based constraint optimization method using BDDs, and then present the method of enumerating all cost-bounded solutions using ZDDs. I would like to discuss a future direction how to collaborate SAT/MaxSAT techniques, DD-based techniques and ILP/PB techniques.

3.17 TT-Open-WBO-Inc: The SAT Engine of Modern Anytime MaxSAT

Alexander Nadel (Technion – Haifa, IL & NVIDIA – Yokneam, IL)

Main reference Alexander Nadel: "TT-Open-WBO-Inc: An Efficient Anytime MaxSAT Solver", J. Satisf. Boolean Model. Comput., Vol. 15(1), pp. 1–7, 2024.
URL https://doi.org//10.3233/SAT-231504

MaxSAT extends the cornerstone NP-complete SAT problem from decision to the optimization of a linear objective. It has a wide range of applications in AI, CAD, planning, scheduling, and beyond. Many of these applications – especially in industry – require anytime solvers, which continuously produce improving solutions. This talk presents TT-Open-WBO-Inc, the de-facto SAT engine of modern anytime MaxSAT solvers. TT-Open-WBO-Inc was the SAT component in the winners of the last three MaxSAT Evaluations in all four anytime categories, differing only in their local search preprocessors. TT-Open-WBO-Inc's efficiency stems from effective problem approximations (via the BMO or Mrs. Beaver algorithms) combined with both classical local search and SAT-based local search (the Polosat algorithm). We will present the genealogy of TT-Open-WBO-Inc and devote most of the talk to its underlying algorithms.

3.18 SpotIT: Evaluating Text-to-SQL Evaluation with Formal Verification

Nina Narodytska (VMware Research – Palo Alto, US)

Joint work of Nina Narodytska, Rocky Klopfenstein, Yang He, Andrew Tremante, Yuepeng Wang, Haoze Wu

Text-to-SQL plays a central role in building natural language interfaces that enable non-expert users to query structured data sources. Given a natural language query N to a database, the goal is to generate a SQL query that retrieves the data requested by N. Due to its practical importance, a large number of Text-to-SQL frameworks have been developed over the last two years in both industry and academia. However, the evaluation of these frameworks' performance has received much less attention. Most evaluation relies on a static testing approach that compares a generated SQL query against a gold SQL query produced by a human annotator for the same natural language query N. The key issue is that such static testing performs an overoptimistic evaluation, as queries can be equivalent by chance, just on these test instances.

In this work, we propose an alternative approach in which we search for a database that reveals discrepancies in the results returned by these SQL queries. We implemented our evaluation framework, which is powered by efficient formal verification techniques, and conducted a performance analysis of ten SOTA Text-to-SQL frameworks on BIRD datasets. Our results reveal that their accuracy is significantly lower – by up to 15% – compared to the results reported by the static testing approach. We also perform a detailed analysis of the failure cases and provide iusefull insights about shortcomings of the benchmark datasets.

3.19 Certified Implicit Hitting Set Solving for Pseudo-Boolean Optimization

Jakob Nordström (University of Copenhagen, DK & Lund University, SE)

Joint work of Jakob Nordström, Benjamin Bogø, Xiamin Chen, Wietze Koops, Pinyan Lu, Marc Vinyals, Qingzhao Wu

Implicit hitting set (IHS) methods work well for SAT-based and pseudo-Boolean optimization, but have lacked proof logging due to the use of mixed integer programming (MIP) solvers for the hitting set subproblems. We replace MIP with a combination of local search and pseudo-Boolean solving, yielding the first fully certified IHS approach for pseudo-Boolean optimization with feasible proof generation and verification overhead. Our ongoing work focuses on improving the hitting-set phase and integrating it more tightly in the overall solving process, and on identifying MIP features crucial for efficient IHS solving.

3.20 Symbolic Conflict Analysis in Pseudo-Boolean Optimization

Albert Oliveras (UPC Barcelona Tech, ES)

Joint work of Robert Nieuwenhuis, Albert Oliveras, Enric Rodríguez-Carbonell, Rui Zhao
Main reference Robert Nieuwenhuis, Albert Oliveras, Enric Rodríguez-Carbonell, Rui Zhao: "Symbolic Conflict Analysis in Pseudo-Boolean Optimization", in Proc. of the 28th International Conference on Theory and Applications of Satisfiability Testing, SAT 2025, Glasgow, Scotland, August 12-15, 2025, LIPIcs, Vol. 341, pp. 23:1–23:18, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2025.
URL https://doi.org//10.4230/LIPICS.SAT.2025.23

In the the last two decades, a lot of effort has been devoted to the development of satisfiability-checking tools for a variety of SAT-related problems. However, most of these tools lack optimization capabilities. That is, instead of finding any solution, one is sometimes interested in a solution that is best according to some criterion.

Pseudo-Boolean solvers can be used to deal with optimization by successively solving a series of problems that contain an additional pseudo-Boolean constraint expressing that a better solution is required. A key point for the success of this simple approach is that lemmas that are learned for one problem can be reused for subsequent ones.

In this talk we go one step further and show how, by using a simple symbolic conflict analysis procedure, not only can lemmas be reused between problems but also strengthened, thus further pruning the search space traversal. In addition, we show how this technique automatically allows one to infer upper bounds in maximization problems, thus giving an estimation of how far the solver is from finding an optimal solution. Experimental results with our PB solver reveal that (i) this technique is indeed effective in practice, providing important speedups in problems where several solutions are found and (ii) on problems with very few solutions, where the impact of our technique is limited, its overhead is negligible.

3.21 Tutorial on Dantzig-Wolfe decomposition, column generation, and branch-price-and-cut

Elina Rönnberg (Linköping University, SE)

Joint work of Stephen J. Maher, Elina Rönnberg
Main reference Stephen J. Maher, Elina Rönnberg: "Integer programming column generation: accelerating branch-and-price using a novel pricing scheme for finding high-quality solutions in set covering, packing, and partitioning problems", Math. Program. Comput., Vol. 15(3), pp. 509–548, 2023.
URL https://doi.org//10.1007/S12532-023-00240-W

In optimisation, decomposition means making a reformulation of a problem into a set of simpler problems that are typically solved by an iterative scheme to produce a solution to the original problem. The purpose is to distribute the computational burden of the original problem onto the simpler ones in a way that pays off with respect to total solution time. Critical for obtaining this is to exploit problem structure in the decomposition and to design an efficient solution scheme for the simpler problems.

In MIP, having a strong formulation matters. Dantzig-Wolfe decomposition yields an extended formulation in a higher-dimensional space that is at least as strong as the original one, and sometimes it is very strong. This strength, however, comes at the cost of a very high-dimensional representation, so that the extended formulation cannot be expressed explicitly. In the context of column generation, which is used for solving linear programs, the extended formulation is called the master problem. Instead of representing the full problem, one uses a restricted master problem that includes only a small subset of the variables. By using a pricing problem, the master problem is iteratively extended to containing an LP optimal solution. To find an optimal integer solution, the standard approach is to use branch-price-and-cut, where column generation is integrated into branch-and-bound.

In this tutorial, the basic principles of Dantzig-Wolfe decomposition, column generation, and branch-price-and-cut are introduced. This is continued with a discussion on our ongoing work on optimality conditions and pricing for integrality, highlighting why I believe these to be interesting areas of future research.

3.22 Quantum computing for discrete optimization: A glimpse into three technologies

Philine Schiewe (Aalto University, FI)

Joint work of Alexey Bochkarev, Raoul Heese, Sven Jäger, Philine Schiewe, Anita Schöbel
Main reference Alexey Bochkarev, Raoul Heese, Sven Jäger, Philine Schiewe, Anita Schöbel: "Quantum computing for discrete optimization: A highlight of three technologies", European Journal of Operational Research, Vol. 329(3), pp. 747–766, 2026.
URL https://doi.org//10.1016/j.ejor.2025.07.063

Quantum optimization has emerged as a promising frontier of quantum computing, providing novel numerical approaches to mathematical optimization problems. In this presentation, we aim to provide an initial intuition for quantum-powered methods in the context of discrete optimization. To this end, we consider three quantum-powered optimization approaches that make use of different types of quantum hardware available on the market. To illustrate these approaches, we solve three classical optimization problems: the Traveling Salesperson

Problem, Weighted Maximum Cut, and Maximum Independent Set. We attempt to provide an intuition behind each approach, describe the corresponding high-level workflow, and highlight crucial practical considerations. In particular, we emphasize the importance of problem formulations and device-specific configurations, and their impact on the amount of resources required for computation (where we focus on the number of qubits). These points are illustrated with a series of experiments on three types of quantum computers: a neutral atom machine from QuEra, a quantum annealer from D-Wave, and gate-based devices from IBM.

3.23 Weighted CP and related frameworks: soft arc consistency and bounds

Thomas Schiex (INRA – Castanet-Tolosan, FR)

Joint work of Thomas Schiex, George Katsirelos, Simon de Givry, Pierre Montalbano, Martin Cooper, Tomas Werner
Main reference Martin C. Cooper, Simon de Givry, Martí Sánchez-Fibla, Thomas Schiex, Matthias Zytnicki, Tomás Werner: "Soft arc consistency revisited", Artif. Intell., Vol. 174(7-8), pp. 449–478, 2010.
URL https://doi.org//10.1016/J.ARTINT.2010.02.001

In Weighted Constraint Programming, constraints are replaced by infinite-valued cost functions that contribute to the definition of both feasibility and a criterion. In this talk, the relations of this "network of cost functions" model with MaxSAT, QP and (01)LP are explored and the so-called "local polytope" shown to provide lower bounds The extension of arc consistency to this setting strictly generalizes usual (G)AC and is shown to provide dual feasible solutions of this local polytope using efficient combinatorial algorithms. This added understanding paves the way to the definition of algorithms enforcing soft local consistencies on global cost functions, when they have suitable LP formulations. Experiments on the QAPLib, using the AllDifferent constraint, handled as a "Linear assignement Problem", show the interest of this approach.

3.24 Introduction to Lazy Clause Generation

Peter J. Stuckey (Monash University – Caulfield, AU)

Joint work of Peter J. Stuckey, Olga Ohrimenko, Michael Codish
Main reference Olga Ohrimenko, Peter J. Stuckey, Michael Codish: "Propagation via lazy clause generation", Constraints An Int. J., Vol. 14(3), pp. 357–391, 2009.
URL https://doi.org//10.1007/S10601-008-9064-X

Finite domain propagation solvers effectively represent the possible values of variables by a set of choices which can be naturally modelled as Boolean variables. In this introduction we describe how to mimic a finite domain propagation engine, by mapping propagators into clauses in a SAT solver. This immediately results in strong nogoods for finite domain propagation. But a naive static translation is impractical except in limited cases. We show how to convert propagators to lazy clause generators for a SAT solver. The resulting system introduces flexibility in modelling since variables are modelled dually in the propagation engine and the SAT solver. The approach has proven to be the state of the art approach

to CP solving. We show how we can improve the straightforward implementation to lazy create the Boolean representation of integer variables; how we can extend conflict analysis to generate stronger, more reusable explanations, and how the language of learning greatly effects the method.

3.25 Tutorial: Optimization in SMT

Nestan Tsiskaridze (Stanford University, US)

Satisfiability Modulo Theories (SMT) has become a central technology in formal methods, enabling expressive reasoning across domains such as verification, synthesis, security, planning, and beyond. In recent years, Optimization Modulo Theories (OMT) has emerged as a powerful extension of SMT that allows not only deciding satisfiability but also optimizing cost functions over models. This tutorial offers a comprehensive introduction to optimization in the SMT setting.

We begin with the foundations of OMT, introducing key techniques. Special emphasis is placed on the evolution of OMT – from its origins in early extensions of SMT solving, through theory- and objective-specific and symbolic approaches, to the recently proposed Generalized OMT (GOMT) framework that unifies all single and multi-objective optimization and arbitrary theory combinations.

The tutorial aims to equip participants with a conceptual and practical understanding of OMT: how it works, why it matters, and how it continues to expand the reach of SMT/OMT technology into new application areas.

3.26 Decision Diagrams for Discrete Optimization

Willem-Jan Van Hoeve (Carnegie Mellon University – Pittsburgh, US)

Main reference Willem-Jan van Hoeve: "An Introduction to Decision Diagrams for Optimization", in Tutorials in Operations Research: Smarter Decisions for a Better World pp. 117–145, 2024.
URL https://doi.org//10.1287/educ.2024.0276

Over the last decade, decision diagram-based optimization has emerged as a novel approach to solving discrete optimization problems. We provide an overview of this new methodology, focusing on three computational paradigms: (1) stand-alone decision diagram-based solvers, (2) integration into constraint programming, and (3) integration into integer linear programming. We discuss applications including graph theoretic problems, scheduling, and vehicle routing. In particular, combining decision diagrams with network flow theory – via a process called "column elimination" – has resolved previously unsolved benchmark instances for problems such as vehicle routing with time windows, pickup-and-delivery with time windows, and graph multi-coloring. These advancements highlight the potential of decision diagram-based optimization as a powerful tool for addressing complex optimization challenges across domains.

3.27 Tutorial: Optimization in CP

Hélène Verhaeghe (UC Louvain, BE)

Constraint Programming (CP) targets combinatorial (optimization) problems. CP solvers require the problem to be modeled using variables (and their domains) and constraints (and an objective). In CP, variables have usually boolean domains or finite integer domains, but can have other special domains (set, graphs, sequences,...). The constraints in CP are arithmetic, logical or global constraints. Global constraints represent a relation between a set of variables (which can vary). The CP solver utilizes dedicated algorithms, known as propagators, for each constraint. They are responsible for filtering invalid values from the domains, based on reasoning corresponding to the constraints. These propagators are coordinated by the fixpoint, which decides which needs to be called. The fixpoint is called during the construction of the search space to prune invalid sub-trees. The search tree is built following the search, a heuristic selected by the user. CP is modular and can adapt to various formulations of the problems.

4 Working groups

4.1 How the optimization communities can work better together

Peter J. Stuckey (Monash University – Caulfield, AU) and Luke Marshall (Microsoft Research – Redmond, US)

The group discussion focused on how communities working in CP, SAT, MaxSAT, MIP, SMT, decision diagrams, etc. can collaborate more effectively.

One concern was that the neighboring solver communities still operate in silos due to a lack of a shared incremental modelling interface. SAT and MIP have stable low-level cores (clauses, linear constraints) and widely used APIs, while CP has a diverse collection of global constraints with varying propagation strengths and no common, structured, in-memory layer to expose them. FlatZinc was acknowledged as useful for interchange, but it flattens away intent, is not incremental, and cannot serve as a modern programmatic API with callbacks, assumptions, or partial additions. There was not enough CP solver implementors in attendance to acertain if this has sufficient support, but it is something that should be discussed in the CP community.

Decomposition methods (Benders, column generation, implicit hitting set approaches) were highlighted as useful approaches to build hybrid systems. Decision diagrams were viewed as a potential way to communicate the structural relationships between variables (i.e., bounds, conflict graphs etc.) with the various solvers. However, there is currently no agreed format (text or binary) for sharing them. Similarly, while SAT and pseudo-Boolean solvers have mature proof logging, CP lacks lightweight, checkable certificates for global propagations, limiting reproducibility, post-mortem analysis, and learning opportunities. Also, many MIP solvers will likely never include proof logging due to lack of commercial interest. The idea of having a library of encoding tools to map decision diagrams to CP, SAT, MIP seems like a valuable tool to allow us to share complex information between paradigms.

The group advocated for a curated set of cross-community benchmark instances that can compare solver methods, i.e., illustrate exponential separations, pathological behaviors, or technology advantages. They also emphasized structured synthetic generators with tunable parameters to enable controlled empirical studies rather than relying on ad hoc or purely random sets. In addition, they identified that encoding quality should not be overlooked. CSPlib could be used as a starting point for this, since it already includes curated CP models, and each technology can probably be applied. Of course finding volunteers to take this on is a challenge

Parallel and portfolio solving emerged as another cross-cutting topic. Recent SAT parallelism has shown surprising near-linear gains, suggesting untapped potential. Open questions center on what to share (bounds, cuts, symmetry info, nogoods, structural summaries), while being mindful of communication costs. Decision diagram partitioning in branch and bound was mentioned as a promising natural workload splitter.

Finally, there was enthusiasm for solver introspection and learning: mining proof traces or search logs to attribute progress to strategies, guiding restart policies, feature extraction for algorithm selection, and predicting which paradigm or encoding will excel based on structural instance features.

5 Panel discussions

5.1 Challenges and opportunities for the next 10 years

Ambros Gleixner (HTW – Berlin, DE), Alexey Ignatiev (Monash University – Clayton, AU), Ciaran McCreesh (University of Glasgow, GB), Thomas Schiex (INRA – Castanet-Tolosan, FR), and Christine Solnon (INSA Lyon / Inria, FR)

The main goal of the panel is to discuss key technical directions and challenges for the coming decade in CP, SAT/MaxSAT, SMT and MIP, and identify synergy between these research areas and other areas that can promote fast development and adoption of the technology. The following schemes were discussed.

Hybrid methods and efficiency scheme. Optimization techniques span over a large number of paradigms, each of which has complementary strengths. For example, local search provides fast but possibly far from optimal solutions while complete search guarantees finding optimal solutions but is computationally demanding. However, with modern computational resources, hybrid search methods are increasingly viable paradigms that can take advantage of these resources, e.g. memory intensive methods like breadth-first search or decision diagrams. A recurring theme was the need to prioritize fast, high-quality solutions over provable optimality, particularly in real-world applications where a tradeoff between speed and suboptimal solutions is acceptable. Another important point is to take advantage of GPUs that are increasingly available. It will require change of the algorithms as existing algorithms are not well suited for GPUs. So, it is an interesting direction to pursue. Finally, the integration of CP's propagation mechanisms into decision diagrams and dynamic programming frameworks was proposed as a promising direction for solver interoperability and performance enhancement.

Modeling and adoption scheme. Modeling remains a central concern. The users might still find it challenging to express their problems even in the user friendly modeling languages like MiniZinc. Direct modeling to SAT or MIP paradigms is feasible for advanced users only. As problem complexity grows, e.g. multi objective optimization, interactive optimization, incremental solving, it is even more important for the community to provide the user a simple way to deal with challenging problems. One of the solutions offered was to use LLMs as an intermediate layer between the user and the model. The idea is that the user specifies the problem in natural language and using LLMs we translate it into a model. The process is iterative and requires user guidance and verification of correctness. This was identified as a promising approach. Finally, the panel discussed strategic directions for community investment, including shared benchmark libraries, accessible educational resources, and scalable solver infrastructure. We also need to invest in teaching optimization technology as it does help with adoption.

Verification and explainability scheme. The panel emphasized that verification and explanability are very important topics for the next decade. Proof logging is the central technique that allows verifying correctness of the solver output that has received significant attention in recent years. Indeed, it is an area under active development. Explanability is another key point as giving the user a solution (or no solution) answer might not be very satisfying. Explanability techniques need more development, starting from defining what exactly explanability is in a given context, the language of explanability and computing explanations efficiently. The panel advocated for modular solver architectures that facilitate cross-paradigm innovation and re-emphasized the importance of proof logging to ensure correctness and reproducibility.

Priority areas. If given substantial funding, the panel identified that priorities are would include parallelization, deployment in everyday decision-making contexts, and broader accessibility.

Participants

- Florent Avellaneda
UQAM – Montreal, CA
- Armin Biere
Universität Freiburg, DE
- Bart Bogaerts
KU Leuven, BE
- Toby Davies
Google – Pyrmont, AU
- Frédéric Didier
Google – Paris, FR
- Katalin Fazekas
TU Wien, AT
- Pierre Flener
Uppsala University, SE
- Bishwamittra Ghosh
MPI-SWS – Kaiserslautern, DE
- Ambros Gleixner
HTW – Berlin, DE
- Tias Guns
KU Leuven, BE
- Alexey Ignatiev
Monash University – Clayton, AU
- Hannes Ihalainen
University of Helsinki, FI
- Christoph Jabs
University of Helsinki, FI
- Matti Järvisalo
University of Helsinki, FI
- George Katsirelos
INRAE – Palaiseau, FR
- Zeynep Kiziltan
University of Bologna, IT
- Lucas Kletzander
TU Wien, AT
- Anna Latour
TU Delft, NL
- Chu Min Li
University of Amiens, FR
- Luke Marshall
Microsoft Research – Redmond, US
- Ciaran McCreesh
University of Glasgow, GB
- Gioni Mexi
Zuse-Institut Berlin, DE
- Shin-ichi Minato
Kyoto University, JP
- Alexander Nadel
Technion – Haifa, IL & NVIDIA – Yokneam, IL
- Nina Narodytska
VMware Research – Palo Alto, US
- Robert Nieuwenhuis
Barcelona, ES
- Jakob Nordström
University of Copenhagen, DK & Lund University, SE
- Andy Oertel
Lund University, SE
- Albert Oliveras
UPC Barcelona Tech, ES
- Anastasia Paparrizou
CNRS – Montpellier, FR
- Elina Rönnberg
Linköping University, SE
- Andre Schidler
Universität Freiburg, DE
- Philine Schiewe
Aalto University, FI
- Thomas Schiex
INRA – Castanet-Tolosan, FR
- Mohamed Siala
LAAS – Toulouse, FR
- Christine Solnon
INSA Lyon / Inria, FR
- Peter J. Stuckey
Monash University – Caulfield, AU
- Nestan Tsiskaridze
Stanford University, US
- Willem-Jan Van Hoeve
Carnegie Mellon University – Pittsburgh, US
- Hélène Verhaeghe
UC Louvain, BE
- Allen Z. Zhong
Monash University – Clayton, AU

Report from Dagstuhl Seminar 25372

Precision in Geometric Algorithms

Mikkel Abrahamsen*[1], Sándor Kisfaludi-Bak*[2], Linda Kleist*[3], and Till Miltzow*[4]

1 **University of Copenhagen, DK.** miab@di.ku.dk
2 **Aalto University, FI.** sandor.kisfaludi-bak@aalto.fi
3 **Universität Hamburg, DE.** linda.kleist@uni-hamburg.de
4 **Utrecht University, NL.** t.miltzow@googlemail.com

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 25372 "Precision in Geometric Algorithms". This seminar was an opportunity for a get together of researchers interested in geometric problems that require high precision of the coordinates to find a correct solution.

Seminar September 7–12, 2025 – https://www.dagstuhl.de/25372
2012 ACM Subject Classification Theory of computation → Design and analysis of algorithms
Keywords and phrases Computational Geometry, Real Complexity Theory
Digital Object Identifier 10.4230/DagRep.15.9.21

1 Executive Summary

Till Miltzow (Utrecht University, NL)

The Dagstuhl Seminar "Precision in Geometric Algorithms" (25372) brought together researchers working on computational geometry, real-complexity theory, and geometric computation models that require high-precision reasoning. The seminar aimed to understand how geometric problems behave when precision, real-number computation, and continuous models become central, and to explore the algorithmic, structural, and complexity-theoretic consequences.

The invited talks covered a broad spectrum: geometric graph theory in hyperbolic spaces; optimal convex-hull reconstruction from imprecise data; ER-complete recognition problems for geometric intersection graphs; oracle separations in the real polynomial hierarchy; new approximation schemes for geometric multimatching; the complexity of the boundary–boundary art gallery problem; and dynamic Steiner spanners in curved spaces. Together, these contributions showcased how precision constraints shape both the geometry and the complexity of algorithmic problems.

Several working groups produced substantial progress. One group extended Fréchet-distance techniques to more than two curves and proved meaningful lower bounds via reductions from 3OV. Another initiated the study of "Devil's Games," a class of infinite-move combinatorial games linked to the first-order theory of the reals. Others explored realization spaces of geometric graph representations, online packing of convex objects, sparse geometric emulators, and the flip distance needed to eliminate crossings in geometric matchings.

* Editor / Organizer

The open-problem session highlighted future challenges: recognizing strongly hyperbolic disk graphs, understanding polygonal knot realization spaces and their potential universality, establishing $\exists\mathbb{R}$-completeness of continuous distance problems, efficiently computing weak circle representations of planar graphs, and bounding fixed points of compositions of monotone polynomials.

Beyond the technical program, the week was marked by a warm and collaborative atmosphere. Discussions continued naturally over meals, hikes, and informal gatherings, helping participants strengthen existing collaborations and spark new ones. Many attendees commented that the social setting – relaxed yet intellectually charged – played a major role in enabling deep, productive exchanges.

Overall, the seminar strengthened connections between computational geometry, real-number computation, and complexity theory, identifying multiple promising directions where precision requirements fundamentally reshape classical algorithmic questions.

2 Table of Contents

Executive Summary
Till Miltzow 21

Overview of Talks

Graphs in Hyperbolic Geometry
Thomas Bläsius 25

Instance-Optimal Imprecise Convex Hull
Sarita de Berg 25

Calculating with Pennies and Marbles
Anna Lubiw and Marcus Schaefer 26

Separations for RPH
Lucas Meijer 26

Approximation Algorithm for the Geometric Multimatching Problem
Eunjin Oh 27

The Boundary-Boundary Art-Gallery Problem is in NP
Jack Stade 27

Near-Optimal Dynamic Steiner Spanners for Constant-Curvature Spaces
Geert van Wordragen 28

Working groups

The Fréchet distance of several curves
Peyman Afshani, Karl Bringmann, Mark de Berg, Omrit Filtser, Dan Halperin, André Nusser, and Günter Rote 28

Devilish Games and QR
Arnaud de Mesmay, Lucas Meijer, Till Miltzow, Marcus Schaefer, and Jack Stade 29

Geometric realization spaces of paths, trees and cycles
Arnaud de Mesmay, Gargi Lather, Anna Lubiw, Marcus Schaefer, and Alexandra Wesolek 30

Online Packing of Convex Objects
Arindam Khan, Anders Aamand, Mikkel Abrahamsen, Linda Kleist, Eunjin Oh, and Csaba Tóth 31

Sparse $(1+\varepsilon)$-emulators for Euclidean point sets
Sándor Kisfaludi-Bak, Sujoy Bhore, Karl Bringmann, Hung Le, André Nusser, and Karol Wegrzycki 31

Flip Distance to a Crossing-Free Matching
Lucas Meijer, Thomas Bläsius, Sarita de Berg, Aye Chan May, Arturo Merino, and Jack Stade 32

Open problems

Recognition of Strongly Hyperbolic Uniform Disk Graphs
Thomas Bläsius 33

Polygonal representations of knots
Arnaud de Mesmay 33

Precision of continuous distance problems
Sándor Kisfaludi-Bak 34

Weak circle representations of planar graphs
Günter Rote 35

Fixed points of compositions of monotone polynomials
Jack Stade 35

Participants 37

3 Overview of Talks

3.1 Graphs in Hyperbolic Geometry

Thomas Bläsius (KIT – Karlsruher Institut für Technologie, DE)

Hyperbolic geometry is a non-Euclidean geometry where the parallel axiom is negated. While the hyperbolic plane behaves locally like the Euclidean plane, it behaves very different beyond that. One crucial difference is that the hyperbolic plane expands exponentially. This has various interesting effects for studying graphs in the hyperbolic plane. When embedding graphs into the hyperbolic plane, the exponential expansion can be used to, for example, achieve successful greedy routing. Moreover, when defining graphs from geometric objects, like intersection graphs of equally sized disks, the properties of the hyperbolic plane translate to interesting graphs properties.

3.2 Instance-Optimal Imprecise Convex Hull

Sarita de Berg (Utrecht University, NL)

Joint work of Sarita de Berg, Ivor van der Hoog, Eva Rotenberg, Daniel Rutschmann, Sampson Wong
Main reference Sarita de Berg, Ivor van der Hoog, Eva Rotenberg, Daniel Rutschmann, Sampson Wong: "Instance-Optimal Imprecise Convex Hull", in Proc. of the 33rd Annual European Symposium on Algorithms (ESA 2025), Leibniz International Proceedings in Informatics (LIPIcs), Vol. 351, pp. 25:1–25:15, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2025.
URL https://doi.org//10.4230/LIPIcs.ESA.2025.25

Imprecise measurements of a point set $P = (p_1, \ldots, p_n)$ can be modelled by a family of regions $F = (R_1, \ldots, R_n)$, where each imprecise region $R_i \in F$ contains a unique point $p_i \in P$. A *retrieval* models an accurate measurement by replacing an imprecise region R_i with its corresponding point p_i.

We construct the convex hull of an imprecise point set in the plane, by determining the cyclic ordering of the convex hull vertices of P as efficiently as possible. Efficiency is interpreted in two ways: (i) minimising the number of retrievals, and (ii) the computation time to determine the set of regions that must be retrieved.

Previous works focused on only one of these two aspects: either minimising retrievals or optimising algorithmic runtime. Our contribution is the first to simultaneously achieve both. Let $r(F, P)$ denote the minimal number of retrievals required by any algorithm to determine the convex hull of P for a given instance (F, P). For a family F of n constant-complexity polygons, our main result is a reconstruction algorithm that performs $\Theta(r(F, P))$ retrievals in $O(r(F, P) \log^3 n)$ time.

Compared to previous approaches that achieve optimal retrieval counts, we improve the runtime per retrieval from polynomial to polylogarithmic. We extend the generality of previous results to simple k-gons, to pairwise disjoint disks with radii in $[1, k]$, and to unit disks where at most k disks overlap in a single point. Our runtime scales linearly with k.

3.3 Calculating with Pennies and Marbles

Anna Lubiw (University of Waterloo, CA) and Marcus Schaefer (DePaul University – Chicago, US)

Joint work of Anna Lubiw, Marcus Schaefer
Main reference Anna Lubiw, Marcus Schaefer: "Recognizing Penny and Marble Graphs is Hard for Existential Theory of the Reals", CoRR, Vol. abs/2508.10136, 2025.
URL https://doi.org//10.48550/ARXIV.2508.10136

Penny graphs are contact graphs of unit disks in the plane. We show that recognizing penny graphs is ER-complete, that is as hard as deciding truth in the existential theory of the reals. The problem remains ER-complete even if a combinatorial embedding of the penny graph is given. Penny graphs which are trees can be realized with the centers of the pennies belonging to a grid of double-exponential size. We can also show that recognizing marble graphs, contact graphs of unit balls in three-dimensional space, is ER-complete.

3.4 Separations for RPH

Lucas Meijer (Utrecht University, NL)

Joint work of Thekla Hamm, Lucas Meijer, Till Miltzow, Subhasree Patro
Main reference Thekla Hamm, Lucas Meijer, Tillmann Miltzow, Subhasree Patro: "Oracle Separations for RPH", CoRR, Vol. abs/2502.09279, 2025.
URL https://doi.org//10.48550/ARXIV.2502.09279

While theoretical computer science primarily works with discrete models of computation, like the Turing machine and the wordRAM, there are many scenarios in which introducing real computation models is more adequate. For example, when working with continuous probability distributions for say smoothed analysis, in continuous optimization, computational geometry or machine learning. We want to compare real models of computation with discrete models of computation. We do this by means of oracle separation results.

We define the notion of a *real Turing machine* as an extension of the (binary) Turing machine by adding a real tape. Using those machines, we define and study the real polynomial hierarchy $\mathbb{R}PH$. We are interested in $\mathbb{R}$PH as the first level of the hierarchy corresponds to the well-known complexity class $\exists\mathbb{R}$. It is known that $NP \subseteq \exists\mathbb{R} \subseteq PSPACE$ and furthermore $PH \subseteq \mathbb{R}\text{PH} \subseteq PSPACE$. We are interested to know if any of those inclusions are tight. In the absence of unconditional separations of complexity classes, we turn to oracle separation. We develop a technique that allows us to transform oracle separation results from the binary world to the real world. As applications, we show there are oracles such that:

- $\mathbb{R}PH^O$ proper subset of $PSPACE^O$,
- BQP^O not contained in $\mathbb{R}PH^O$.

Our results hint that $\exists\mathbb{R}$ is strictly contained in PSPACE and that there is a separation between the different levels of the real polynomial hierarchy. We also bound the power of real computations by showing that NP-hard problems are unlikely to be solvable using polynomial time on a realRAM. Furthermore, our oracle separations hint that polynomial-time quantum computing cannot be simulated on an efficient real Turing machine.

3.5 Approximation Algorithm for the Geometric Multimatching Problem

Eunjin Oh (POSTECH – Pohang, KR)

Joint work of Eunjin Oh, Shinwoo An, Jie Xue

Abstract: Let S and T be two sets of points in a metric space with a total of n points. Each point in S and T has an associated value that specifies an upper limit on how many points it can be matched with from the other set. A multimatching between S and T is a way of pairing points such that each point in S is matched with at least as many points in T as its assigned value, and vice versa for each point in T. The cost of a multimatching is defined as the sum of the distances between all matched pairs of points. The geometric multimatching problem seeks to find a multimatching that minimizes this cost. A special case where each point is matched to at most one other point is known as the geometric many-to-many matching problem.

We present two results for these problems when the underlying metric space has a bounded doubling dimension. Specifically, we provide the first near-linear-time approximation scheme for the geometric multimatching problem in terms of the output size. Additionally, we improve upon the best-known approximation algorithm for the geometric many-to-many matching problem, previously introduced by Bandyapadhyay and Xue (SoCG 2024), which won the best paper award at SoCG 2024.

3.6 The Boundary-Boundary Art-Gallery Problem is in NP

Jack Stade (University of Copenhagen, DK)

Main reference Jack Stade: "NP-membership for the boundary-boundary art-gallery problem", CoRR, Vol. abs/2511.01562, 2025.
URL https://doi.org//10.48550/ARXIV.2511.01562

The X-Y art-gallery problem asks to find a minimum set of guards that guard a polygon P, where the guards are restricted to lie in X and must see all of Y. For X,Y$\in$ {Point, Boundary, Vertex}, this gives 9 different variants. Recent work has determined the complexity of each of these variants; all but the Boundary-Boundary variant were known to be either NP-complete or $\exists\mathbb{R}$-complete.

We complete this classification, showing that the Boundary-Boundary variant is NP-complete. This is somewhat surprising, since the coordinates of guards in an optimal solution might need to be irrational. We show that each coordinate is at worst $p + q\sqrt{r}$, where p, q and r are rational numbers with polynomially many bits. These coordinates give a certificate that can be verified in polynomial time.

3.7 Near-Optimal Dynamic Steiner Spanners for Constant-Curvature Spaces

Geert van Wordragen (Aalto University, FI)

Joint work of Sándor Kisfaludi-Bak, Geert van Wordragen
Main reference Sándor Kisfaludi-Bak, Geert van Wordragen: "Near-Optimal Dynamic Steiner Spanners for Constant-Curvature Spaces", CoRR, Vol. abs/2509.01443, 2025.
URL https://doi.org//10.48550/ARXIV.2509.01443

We consider Steiner spanners in Euclidean and non-Euclidean geometries. In the Euclidean setting, a recent line of work initiated by Le and Solomon and further improved by Chang et al. obtained Steiner $(1+\varepsilon)$-spanners of size $O_d(\varepsilon^{(1-d)/2}\log(1/\varepsilon)n)$, nearly matching the lower bounds of Bhore and Tóth.

We obtain Steiner $(1+\varepsilon)$-spanners of size $O_d(\varepsilon^{(1-d)/2}\log(1/\varepsilon)n)$ not only in d-dimensional Euclidean space, but also in d-dimensional spherical and hyperbolic space. For any fixed dimension d, the obtained edge count is optimal up to an $O(\log(1/\varepsilon))$ factor in each of these spaces. Unlike earlier constructions, our Steiner spanners are based on simple quadtrees, and they can be dynamically maintained, leading to efficient data structures for dynamic approximate nearest neighbours and bichromatic closest pair.

In the hyperbolic setting, we also show that 2-spanners in the hyperbolic plane must have $\Omega(n\log n)$ edges, and we obtain a 2-spanner of size $O_d(n\log n)$ in d-dimensional hyperbolic space, matching our lower bound for any constant d. Finally, we give a Steiner spanner with *additive* error ε in hyperbolic space with $O_d(\varepsilon^{(1-d)/2}\log(\alpha(n)/\varepsilon)n)$ edges, where $\alpha(n)$ is the inverse Ackermann function.

Our techniques generalize to closed orientable surfaces of constant curvature as well as to some quotient spaces.

4 Working groups

4.1 The Fréchet distance of several curves

Peyman Afshani (Aarhus University, DK), Karl Bringmann (Universität des Saarlandes – Saarbrücken, DE), Mark de Berg (TU Eindhoven, NL), Omrit Filtser (The Open University of Israel – Ra'anana, IL), Dan Halperin (Tel Aviv University, IL), André Nusser (INRIA – Sophia Antipolis, FR), and Günter Rote (FU Berlin, DE)

Given k polygonal curves $c_1,\ldots,c_k$ in the plane or in some higher-dimensional space with endpoints s_i,t_i and k point robots $r_1,\ldots,r_k$, each robot r_i has to move from s_i to t_i with its center constrained to c_i, without backtracking. The objective is to compute a coordinated motion that minimizes the maximum pairwise distance between robots along their trajectories.

For $k=2$, this is the classic Fréchet distance. The problem formulation can be generalized in various ways, some of which are motivated by applications in robotics and transportation. These include minimum clearance conditions or collisions avoidance (the easiest case being circular robots), and alternative objective functions, several of which give rise to especially challenging variants of the problem.

Background. Computing the Fréchet distance between two curves has been studied extensively in computational geometry since its introduction by Alt and Godau. The vast majority of subsequent work in this area concerns the case of two curves. For this setting, efficient implementations are publicly available. Dumitrescu and Rote presented a 2-approximation algorithm for the case of k curves. They claimed without justification an exact algorithm of running time $O(n^k)$ if each curve has at most n edges. Along different lines, a general solution for the case of k curves has been devised and implemented, adapting sampling-based planning – the standard workhorse of robot algorithms This approach comes with provable guarantees on the quality of the approximation. However, it is currently not competitive in practice: In experiments, already for $k = 2$, its running time was several orders of magnitude slower than the efficient implementations for two curves reported by Bringmann.

Results obtained during the seminar. The classic approach to the computation of the Fréchet distance solves the decision version of the problem (with a given threshold on the maximum distance between the robots) by looking for a monotone path in the free-space diagram F. We discussed the extension of this approach to $k > 2$ curves. For k curves with $n_1, \ldots, n_k$ edges, respectively, the free-space diagram lives in a k-dimensional box consisting of $n_1 \cdot n_2 \cdots n_k$ subboxes (cells). Inside each cell, the free space is the intersection of $\binom{k}{2}$ cylindrical prisms.

For $k = 3$ we managed so solve the decision problem in $O(n_1 n_2 n_3) = O(n^3)$ time. We could show that the reachable set on each rectangular face of a subbox has a restricted structure: It is the intersection of a staircase polygon with an ellipse. Although the staircase polygon may have up to n steps, it can be computed in amortized constant time from the "incoming" faces on each box.

Lower bounds. We showed that a substantially better algorithm with a truly subcubic runtime is unlikely to exist, even if only an approximation of the Fréchet distance with an approximation factor of about 1.1 is desired. For this purpose, we reduced the 3OV problem to the (approximate) Fréchet distance for three curves. In the 3OV (three-orthogonal vectors) problem, we are given three sets A, B, C of n vectors in $\{0, 1\}^d$, and the task is to decide if there are three vectors $x \in A$, $y \in B$, $z \in C$ such that $x_i y_i z_i = 0$ for $i = 1, \ldots, d$.

Variations. For a larger number of $k > 3$ curves, we explored ideas that might lead to an algorithm of running time $O(n^{k^2})$. We also considered an alternate objective function: the radius of the smallest circle enclosing the moving points. This makes the free-space more complicated. For curves in the plane, the free-space on each rectangular face of the free-space diagram is a convex region that is bounded by pieces of line segments, ellipses, and a degree-6 curve.

4.2 Devilish Games and QR

Arnaud de Mesmay (Gustave Eiffel University – Marne-la-Vallée, FR), Lucas Meijer (Utrecht University, NL), Till Miltzow (Utrecht University, NL), Marcus Schaefer (DePaul University – Chicago, US), and Jack Stade (University of Copenhagen, DK)

We worked on a new complexity class denoted by $\exists\mathbb{R}$. This complexity class can be defined as all problems that are equivalent to deciding the First Order Theory of the Reals. We describe a framework to show $\exists\mathbb{R}$-completeness of Devil's games. Devil's games have two key properties.

- Players alternate in taking turns and
- each turn gives an infinite continuum of possible turns.

There is this very cute classical puzzle, which goes as follows: *After a career of elegant proofs occasionally sabotaged by overlooked edge cases, you find yourself in hell's quiet reading room, where the devil greets repentant theoreticians with a friendly smile. He gestures to a round table and proposes a simple challenge: you and he will take turns placing identical coins on the tabletop, and coins may not overlap. Whoever cannot place a new coin on the table loses. The devil insists you move first, confident that impatience will cloud your reasoning just as it did in life. Win, and he'll grant you a brief return to correct that final paper; lose, and you will spend eternity proofreading the edge cases of others.* Interestingly if you place your first coin precisely at the center and then mirror every move of the devil across that center, you can always respond and never be the one to run out of space first. This idea uses symmetry and is a standard technique in combinatorial game theory. But note, if you had not placed the first coin in the middle. It seems impossible to analyze how to win this game. The reasons being the two defining properties of Devil's games. While there are plenty of results that show that combinatorial games are PSPACE-complete, this seems not to capture the second aspect of the Devil's game. And intuitively the second property makes Devil game's quite distinct from most other known combinatorial games. This intuition motivates us to study Devil's games more broadly. We use the term Devil's games, for two reasons. One is a reference to the old puzzle from above and the second is that they are devilishly difficult to analyze.

4.3 Geometric realization spaces of paths, trees and cycles

Arnaud de Mesmay (Gustave Eiffel University – Marne-la-Vallée, FR), Gargi Lather (Indian Institute of Techology Madras, IN), Anna Lubiw (University of Waterloo, CA), Marcus Schaefer (DePaul University – Chicago, US), and Alexandra Wesolek (TU Berlin, DE)

The topic of this working group was to study the realization spaces of simple graphs (paths, trees or cycles) when representing them geometrically in two and three dimensions. A main motivation for this was to complement the recent results of Lubiw and Schaefer that imply universality for the realization spaces of penny and marble graphs in general. More precisely, we have investigated the following problems.

Realizing trees as penny graphs. It follows from celebrated results on the Carpenter's rule problem that the realization space of a path as a penny graph, i.e., a contact graph of unit disks in the plane, is connected. In contrast, this realization space can become disconnected for trees. Known hardness proofs imply the existence of such disconnected examples, but we have obtained a simpler and arguably cleaner construction.

Realizing trees and paths as marble graphs. Moving up to three dimensions, it is easy to see using simple knots that the realization space of trees as marble graphs, i.e., contact graphs of unit balls in $\mathbb{R}^3$, can be disconnected. For paths, we have worked with some precise realizations (both physical and virtual) of overhand and fisherman's knots with chains of marbles and preliminary evidence suggests that they also yield disconnected realization spaces. This is ongoing work.

Spaces of polygonal knots. A third topic of investigation was the space of realizations of a cycle as a polygonal chain with a fixed number of segments of variable length in three dimensions. This topic naturally involves a knot-theoretical aspect, since different knot types necessarily lead to disconnected components in the realization space. Attempts to prove universality and $\exists\mathbb{R}$-hardness for polygonal realizations of fixed knot types raised new questions about intersection and linking graphs of triangles in three dimensions, which we are still looking at.

4.4 Online Packing of Convex Objects

Arindam Khan (Indian Institute of Science – Bangalore, IN), Anders Aamand (University of Copenhagen, DK), Mikkel Abrahamsen (University of Copenhagen, DK), Linda Kleist (Universität Hamburg, DE), Eunjin Oh (POSTECH – Pohang, KR), and Csaba Tóth (California State University – Northridge, US)

We study the packing of convex polygons in the online setting. Here, convex polygons (a total of n in number) arrive one by one, and need to be packed (immediately and irrevocably, using translation and without overlapping) in a horizontal unbounded strip of unit height. Our goal is to minimize the width of the strip to pack all polygons. We have some promising initial results that might lead to a $(\log n)^{O(1)}$-competitive algorithm. Our approach exploits connections with the online sorting problem, where n elements are revealed one by one and have to be placed in an immediate and irrevocable manner into empty cells of an array. The objective is to minimize the sum of absolute differences between elements in the consecutive cells. We also hope to extend the approach to obtain a $(\log n)^{O(d)}$-competitive algorithm for packing d-dimensional convex polytopes into a d-dimensional strip (with $(d-1)$-dimensional unit cube base and unbounded length in the d-th dimension).

4.5 Sparse $(1+\varepsilon)$-emulators for Euclidean point sets

Sándor Kisfaludi-Bak (Aalto University, FI), Sujoy Bhore (Indian Institute of Technology Bombay – Mumbai, IN), Karl Bringmann (Universität des Saarlandes – Saarbrücken, DE), Hung Le (University of Massachusetts Amherst, US), André Nusser (INRIA – Sophia Antipolis, FR), and Karol Wegrzycki (MPI für Informatik – Saarbrücken, DE)

Let P be a set of points in the Euclidean plane (or more generally, in $\mathbb{R}^d$. Consider an edge weighted graph $G = (P \cup S, E), w : E \to \mathbb{R}_{\geq 0}$ and let $\text{dist}_{G,w}$ be the induced shortest-path distance. If for some > 0 the distance $\text{dist}_{G,w}$ satisfies for all $u, v \in P$ that

$$\|u - v\|_2 \leq \text{dist}_{G,w}(u, v) \leq (1 + \varepsilon)\|u - v\|_2,$$

then (G, w) is called a $(1+\varepsilon)$-*emulator* and the points of S are called *Steiner vertices*. Our goal in this problem has been to minimize the number of edges in the emulator.

The variant of the problem where $S \subset \mathbb{R}^d$ and the weight function is forced to be the Euclidean weight function is well-understood: the resulting graphs are called *Steiner spanners*, and the sparsest possible Steiner spanners in $\mathbb{R}^d$ are known to have $\Theta(n/\varepsilon^{\frac{d-1}{2}})$ edges up to logarithmic factors of $1/\varepsilon$.

During the seminar, we started working on the specific point set P that is used in the lower bound for Euclidean Steiner spanners: this corresponds to a bipartite setting. The bipartite construction can be used as a basis for general constructions, and the best Euclidean construction is simply the complete bipartite graph, so it should be possible to improve to find a sparser graph using emulators where we have more freedom to choose the edge weights. We have identified several structural properties of a potential edge-minimal solution based on the geometric constraints, and have found some examples of graphs that had fewer edges than the complete bipartite graph. Deciding if these constructions can be assigned the desired weights is a work in progress.

4.6 Flip Distance to a Crossing-Free Matching

Lucas Meijer (Utrecht University, NL), Thomas Bläsius (KIT – Karlsruher Institut für Technologie, DE), Sarita de Berg (Utrecht University, NL), Aye Chan May (Thammasat University – Pathum Thani, TH), Arturo Merino (O'Higgins University – Rancagua, CL), and Jack Stade (University of Copenhagen, DK)

At Dagstuhl, we worked on the "Flip Distance to a Geometric Crossing-Free Perfect Matching" problem, which we informally called "Uncrossing Line Segments". In this problem, you are given a set of 2n points in the plane and a perfect matching of them. You create a line segment between any two matched points, which gives a configuration of n line segments. Some of these line segments may intersect. We want to transform the instance into one without any intersections using the "flip" operator. Practically, we may "flip" two crossing line segments, and match the four points involved in these line segments in the other two possible ways, neither of which will cause these two lines to intersect. Prior to the seminar, the best known results were that there exists a configuration that always takes $\Omega(n)$ flips to uncross, all configurations have a sequence of $O(n^2)$ flips to uncross them, and an adversary can take at most $O(n^3)$ flips on any configuration. Additionally, unbeknownst to us, it was also known that if the points are in convex position, there is always a sequence of $O(n)$ flips to uncross it.

During the seminar, we did not manage to make much progress on the problem. Our main novel result is that if there are k points not on the convex hull of the pointset, then there exists a sequence of $O(nk)$ flips to uncross the configuration. Otherwise, we looked into many potential-based strategies to keep track of the 'progress' we make upon performing a specific flip; we looked into search trees; into the dual; into divide and conquer strategies; and many more subideas.

5 Open problems

5.1 Recognition of Strongly Hyperbolic Uniform Disk Graphs

Thomas Bläsius (KIT – Karlsruher Institut für Technologie, DE)

A *hyperbolic uniform disk graph (HUDG)* is the intersection graph of disks of equal radius r in the hyperbolic plane. Recognizing HUDGs is $\exists\mathbb{R}$ complete. However, the hardness is essentially based on the fact that Euclidean UDGs are a subclass of HUDGs. One way of looking at a subset of HUDGs that are very hyperbolic are *strongly hyperbolic UDGs*: here all vertices have to be placed into a disk of radius at most $2r$.

Question: Is recognizing SHUDGs in NP (or even in P)? It might be interesting to restrict SHUDGs further to only sparse graphs. An alternative way of restricting HUDGs to graphs that are rather hyperbolic is to require $r \in \Omega(\log n)$.

5.2 Polygonal representations of knots

Arnaud de Mesmay (Gustave Eiffel University – Marne-la-Vallée, FR)

A *polygonal knot*[1] is a closed polygonal curve in $\mathbb{R}^3$. Two polygonal knots made of k segments are *isotopic* if there is a continuous (non-necessarily polygonal) deformation from one to the other without crossings, and are *polygonally isotopic* if there is such a deformation that respects the polygonal structure, i.e., such that all the intermediate curves are also polygonal knots with k segments. The lengths of the segments do *not* have to be respected.

Question 1. Does there exist a k and two polygonal knots made of k segments which are isotopic to the unknot but not polygonally isotopic to each other?

This question was mentioned by Calvo [calvo] as an open problem and there seems to have been no progress since then. Note that if the length of the segments is fixed, such examples of *stuck unknots* are known [toussaint]. Also, in the same paper, Calvo showed that there are isotopic trefoil knots with 6 segments which are *not* polygonally isotopic.

One way to reformulate this question is to wonder whether the realization space of the unknot within the space of polygonal knots with k segments is connected. From that perspective, this reminds of the classical Ringel isotopy conjecture, which asked whether the realization spaces of arrangements of pseudolines are connected. This conjecture was shattered by the M'nev universality theorem showing that such realization spaces can be very pathological (formally, can be stably equivalent to any semi-algebraic set). This leads to:

Question 2. Are there universality phenomena in the polygonal realization spaces of knots?

Universality phenomena often have algorithmic/complexity implications in that the corresponding problems are often $\exists\mathbb{R}$-hard. The *stick number* of a knot K is the minimum number of segments to make a polygonal knot isotopic to K.

[1] This terminology is nonstandard, don't google it.

Question 3. Is the stick number $\exists\mathbb{R}$-hard to compute?

This is also open for the equilateral stick number (where the lengths of the segments are required to be equal). Actually, it is open whether the stick number equals the equilateral stick number in general [cantarella].

Knot theory is hard and scary but polygonal knots are so different from the standard ones that I think that these problems do not require prior knowledge in knot theory. For all these problems, a natural angle of attack could be to look first at polygonal *links* (disjoint unions of knots) and even graphs embedded polygonally in $\mathbb{R}^3$.

5.3 Precision of continuous distance problems

Sándor Kisfaludi-Bak (Aalto University, FI)

Consider the following distance problems:

Pr1 Given a set of open convex pairwise disjoint obstacles in $\mathbb{R}^3$ and $a, b \in \mathbb{R}^3$, find the shortest path connecting a and b that avoids the obstacles.
Pr2 Given a polygonal surface (e.g., boundary of a 3-dimensional convex polyhedron, but could be just an abstract surface), find its diameter.
Pr3 Given a set of (integer) points in $\mathbb{R}^3$, find the shortest tree connecting them (Euclidean Steiner tree).

Pr1 and Pr3 are known to be NP-hard, and I suspect that Problem 2 might be NP-hard for high-genus surfaces. Pr2 has an exact solution for the convex polyhedron case, but it requires an oracle for solving low-degree equations (note that the diameter can be realized by a pair of points that are in the interior of their respective faces). I am not aware of exact solutions to 1 and 3, but they have approximation schemes with $2^{\mathrm{poly}(n)} \cdot \mathrm{polylog}(1/\varepsilon)$ time via brute force structure guess + second-order cone programming (SOCP). Moreover, Pr1 and Pr2 have an FPTAS [Har-Peled99] and Pr3 has an EPTAS.

Question 1. Is any of these problems $\exists\mathbb{R}$-complete? If not, can we reduce some of them to each other or reduce to them from a common third problem?

There are some problems such as geometric median (given a set of points in the plane, find the point q that minimizes the sum of distances to the given points) that can be solved in $\mathrm{poly}(n, \log(1/\varepsilon))$ time [CohenLMPS16]. They do not give off an NP-hard vibe, and they should probably fall within some class of "polynomial-reals", some polynomial-time analogue of $\exists\mathbb{R}$.

Question 2. Is there any relationship between approximability and $\exists\mathbb{R}$-completeness? Can we at least provide justification that geometric median is not $\exists\mathbb{R}$-complete?

Question 3. Is there any reasonable notion of an exact algorithm that can solve Pr1 and Pr3? What kind of oracle access should be granted?

Question 4. Can we prove (conditional) lower bounds to rule out an FPTAS for Pr3 and a $\mathrm{poly}(n, \log(1/\varepsilon))$ algorithm for Pr1 and Pr2?

5.4 Weak circle representations of planar graphs

Günter Rote (FU Berlin, DE)

Given a plane graph G, how fast can we find a set of disks whose intersection graph is G? By the celebrated Koebe–Andreyev–Thurston Theorem, every planar graph G has a *disk-packing* representation as a touching graph of nonoverlapping disks, see [Felsner and Rote 2019] for a relatively easy proof through a converging infinite algorithm. In many applications of this theorem, it is not necessary to require adjacent disks to touch; overlapping disks are also fine, see for example [Felsner 2016]. I call this a *weak circle representation*.

Background. If G is triangulated, the disk-packing representation is essentially unique, i. e., unique up to Möbius transformations (circle-preserving transformations). It is not hard to see that for some instances, the ratio between the largest and the smallest disk is necessarily exponential, and likewise, the algebraic degree of the solution coordinates and radii can be exponential in the size of G.

Mohar [1997, 2000] has given polynomial-time *approximation* algorithms for disk-packing representation. For a graph embedded on a surface of constant negative curvature, like a Klein bottle, the algorithm computes ε-approximations of the centers and radii of a true circle-packing representation [Mohar 2000, Theorem 5.5]. In the plane, there is an algorithm that computes a vector $r = (r_1, \ldots, r_n)$ of radii for the n disks such that the resulting maximum *angle defect* $\mu(r)$ at the centers is smaller than some given bound ε [Mohar 2000, Algorithm A]. In 2019, Dong, Lee and Quanrud gave an improved algorithm to compute an ε-approximation of the centers and radii of a true circle packing in the plane in $O(n \log \frac{U}{\varepsilon})$ time where $U = 2^{O(n)}$ is the ratio between the largest and the smallest disk, and the true circle packing is normalized so that the largest radius is 1.

It remains to work out bounds on ε that guarantee the desired drawing can be obtained by slightly blowing up the radii. Perhaps there is also a more direct and faster approach.

As a stronger variation, we may require that no three disks intersect, or that no disk center is covered by another disk (to remain closer in spirit to a packing representation).

5.5 Fixed points of compositions of monotone polynomials

Jack Stade (University of Copenhagen, DK)

Let f and g be polynomials of degree d each with positive derivative on the whole real line. How many fixed points can the composition $f \circ g$ have?

Background

Miltzow and Schmiermann [MiltzowSchmiermann2022] have asked about the complexity of continuous constraint satisfaction problems when each constraint involves at most 2 variables. In the discrete setting, many-valued 2SAT is polynomial time solvable when the constraints are *monotone* (see [BeckertHanleManya2000]). For $x, y \in \mathbb{Z}$, a constraint $C(x, y)$ is one that can be written as a conjunction of constraints of form $\pm x \geq x_0 \vee \pm y \geq y_0$.

In the continuous setting, it seems natural to study the monotone case since it isn't obviously NP-hard. However, we can compose polynomials, and the degree of the composition grows exponentially with the number of polynomials. What is less clear is whether the combinatorial complexity grows exponentially: if we have two compositions $f_1 \circ \cdots \circ f_k$ and $g_1 \circ \cdots \circ g_j$ of monotone polynomials, can their graphs intersect exponentially many times? If so, this could provide a route towards showing that monotone continuous 2SAT is NP-hard.

The problem I pose above is essentially the simplest non-trivial example of this problem. Going by the degree of the polynomials, it seems that $f \circ g$ could have as many as d^2 fixed points. But there are only $2(d+1)$ coefficients total, so there aren't enough degrees of freedom to easily construct examples with more than $\mathcal{O}(d)$ fixed points. When $d = 3$, I think I can prove that $f \circ g$ has at most 5 fixed points.

Participants

- Anders Aamand
University of Copenhagen, DK
- Mikkel Abrahamsen
University of Copenhagen, DK
- Peyman Afshani
Aarhus University, DK
- Sujoy Bhore
Indian Institute of Technology Bombay – Mumbai, IN
- Thomas Bläsius
KIT – Karlsruher Institut für Technologie, DE
- Karl Bringmann
Universität des Saarlandes – Saarbrücken, DE
- Mark de Berg
TU Eindhoven, NL
- Sarita de Berg
Utrecht University, NL
- Arnaud de Mesmay
Gustave Eiffel University – Marne-la-Vallée, FR
- Omrit Filtser
The Open University of Israel – Ra'anana, IL
- Dan Halperin
Tel Aviv University, IL
- Arindam Khan
Indian Institute of Science – Bangalore, IN
- Sándor Kisfaludi-Bak
Aalto University, FI
- Linda Kleist
Universität Hamburg, DE
- Gargi Lather
Indian Institute of Techology Madras, IN
- Hung Le
University of Massachusetts Amherst, US
- Anna Lubiw
University of Waterloo, CA
- Aye Chan May
Thammasat University – Pathum Thani, TH
- Lucas Meijer
Utrecht University, NL
- Arturo Merino
O'Higgins University – Rancagua, CL
- Till Miltzow
Utrecht University, NL
- André Nusser
INRIA – Sophia Antipolis, FR
- Eunjin Oh
POSTECH – Pohang, KR
- Günter Rote
FU Berlin, DE
- Marcus Schaefer
DePaul University – Chicago, US
- Jack Stade
University of Copenhagen, DK
- Csaba Tóth
California State University – Northridge, US
- Geert van Wordragen
Aalto University, FI
- Karol Wegrzycki
MPI für Informatik – Saarbrücken, DE
- Alexandra Wesolek
TU Berlin, DE

Report from Dagstuhl Seminar 25381

Open Scholarly Information Systems: Status Quo, Challenges, Opportunities

Hannah Bast[*1], Guillaume Cabanac[*2], Paolo Manghi[*3], Jian Wu[*4], and Marcel R. Ackermann[*5]

1 **Universität Freiburg, DE.** bast@informatik.uni-freiburg.de
2 **University of Toulouse, FR.** guillaume.cabanac@univ-tlse3.fr
3 **Institute of Information Science and Technologies – CNR – Pisa, IT.** paolo.manghi@isti.cnr.it
4 **Old Dominion University – Norfolk, US.** fanchyna@gmail.com
5 **Schloss Dagstuhl – Trier, DE.** marcel.r.ackermann@dagstuhl.de

Abstract

Over the past 30 years, a rich ecosystem of scholarly information systems has developed that openly provide their services to the scientific community. These systems include aggregators of bibliographic metadata (e.g., DBLP, OpenCitations, OpenAIRE Graph, OpenAlex, ORKG, Semantic Scholar, CiteSeerX, and CORE); publication, data, and software repositories (e.g., Arxiv.org, Figshare, Zenodo, Software Heritage, and Dataverse); and PID authorities (e.g., ORCID, ROR, Crossref, and DataCite). This interdisciplinary Dagstuhl Seminar "Open Scholarly Information Systems: Status Quo, Challenges, Opportunities" (25381) was the first of its kind to bring together practitioners from this ecosystem, as well as researchers investigating related questions or relying on these systems in their own research. It provided a unique opportunity for dialogue, sharing insights, building new networks, and fostering collaboration.

Seminar September 14–19, 2025 – https://www.dagstuhl.de/25381
2012 ACM Subject Classification Information systems → Digital libraries and archives; Information systems → Information retrieval; Computing methodologies → Machine learning
Keywords and phrases artificial intelligence, knowledge graphs, open infrastructures, scholarly big data, scholarly information systems, semantic search
Digital Object Identifier 10.4230/DagRep.15.9.38

1 Executive Summary

Hannah Bast (Universität Freiburg, DE)
Guillaume Cabanac (University of Toulouse, FR)
Paolo Manghi (Institute of Information Science and Technologies – CNR – Pisa, IT)
Jian Wu (Old Dominion University – Norfolk, US)

This report presents the outcomes and strategic next steps derived from Dagstuhl Seminar "Open Scholarly Information Systems: Status Quo, Challenges, Opportunities" (25381), held in September 2025. The seminar brought together an international group of experts to address the critical challenges facing Open Scholarly Infrastructure (OSI), the evolution of Scholarly Knowledge Graphs (SKGs), and the transformative impact of Agentic AI on the research lifecycle.

[*] Editor / Organizer

Open Scholarly Information Systems: Status Quo, Challenges, Opportunities, *Dagstuhl Reports*, Vol. 15, Issue 9, pp. 38–57
Editors: Hannah Bast, Guillaume Cabanac, Paolo Manghi, Jian Wu, and Marcel R. Ackermann
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Purpose and Context

The primary objective of the seminar was to foster collaboration among diverse infrastructure "owners" and stakeholders to ensure the long-term sustainability and interoperability of the systems that support global research. Participants engaged in high-level plenary discussions and focused working groups to tackle technical, ethical, and economic hurdles within the scholarly ecosystem.

Key Themes and Discussion Pillars

The results detailed in this report are organised around several core pillars:

- *Sustainability and Digital Sovereignty:* A central theme was the urgent need for nations to retain control over research outputs to avoid dependency on commercial monopolies. The group explored the Barcelona Declaration as a framework for promoting open research information and discussed strategies to convince institutions to dedicate a portion of their budgets to open infrastructure.
- *Metadata Excellence and Interoperability:* Discussions focused on harmonising metadata across platforms like DBLP, Wikidata, and OpenReview. The "COMET" approach was proposed to align decentralised metadata enrichment efforts, reducing redundancy and enhancing data quality.
- *The Rise of Agentic AI:* The seminar examined how autonomous AI agents might reshape discovery, writing, and peer review. A critical concern was maintaining human agency and accountability to safeguard scientific integrity, even as AI accelerates baseline tasks like replication.
- *Reforming Research Assessment:* Participants challenged the current "perverse system" of publication metrics and rankings. The consensus emphasized that research quality cannot be measured by numbers alone and that data providers should focus on providing comprehensive data rather than automated rankings.

Strategic Objectives

The outcomes represent a collective effort to move from "building" to "using" and "sustaining" open systems. Key outputs include a draft manifesto on the economic and social value of OSI, technical roadmaps for migrating services like Scholia to QLever, and position papers on the future of human-AI collaboration in science. This document serves as a record of these discussions and a roadmap for the community to ensure that scholarly information remains a transparent, trustworthy, and shared global resource.

2 Table of Contents

Executive Summary
Hannah Bast, Guillaume Cabanac, Paolo Manghi, and Jian Wu 38

Overview of Talks

Curating the DBLP Computer Science Bibliography
Marcel R. Ackermann 42

Citations in the world's largest encyclopedia – and their future
Phoebe Ayers and Lydia Pintscher 42

Narrative Information Access in Digital Libraries
Wolf-Tilo Balke 42

Knowledge Graphs and QLever
Hannah Bast 43

Zenodo and InvenioRDM: Cross-domain digital repositories for the long tail of research
Martin Fenner 43

FAIR Digital Objects as Scholarly Infrastructure Metadata Middleware
Carole Goble 44

Please meet AI, our dear new colleague: Are we becoming obsolete?
Iryna Gurevych 44

Research Fast and Slow
Min-Yen Kan 44

Overview of COnnecting REpositories (CORE)
Petr Knoth 45

Openness aspects of scholarly information systems important for adoption, transparency and interoperability
Bianca Kramer 45

Scholarly Knowledge Graphs: No community, no fun
Paolo Manghi 45

Collaborative curation of bibliographic metadata
Daniel Mietchen 46

Google Dataset Search
Natasha Noy 47

OpenCitations and its new IT infrastructure
Mario Petrella 47

The AIDA Dashboard
Angelo Salatino 47

Navigating Scholarly Knowledge
Tilahun Abedissa Taffa 48

TIB AI Assistant for Research
Sahar Vahdati 48

Digital Science & the Research Data Ecosystem
Kathryn Weber-Boer 48

CiteSeerX and NDLTD
Jian Wu 49

Challenges and opportunities in arXiv
Ramin Zabih 49

Working groups

Theme: Metadata Excellence and Interoperability
Hannah Bast, Guillaume Cabanac, Paolo Manghi, and Jian Wu 50

Theme: Reforming Research Assessment
Hannah Bast, Guillaume Cabanac, Paolo Manghi, and Jian Wu 51

Theme: Sustainability and Digital Sovereignty
Hannah Bast, Guillaume Cabanac, Paolo Manghi, and Jian Wu 53

Theme: The Rise of Agentic AI
Hannah Bast, Guillaume Cabanac, Paolo Manghi, and Jian Wu 54

Participants 57

3 Overview of Talks

3.1 Curating the DBLP Computer Science Bibliography

Marcel R. Ackermann (Schloss Dagstuhl – Trier, DE)

The DBLP Computer Science Bibliography is one of the most comprehensive scholarly metadata collections and knowledge graphs in computer science. Correctly identifying the true named entities in publication metadata is one of the biggest challenges in maintaining such a scholarly information system. In this talk, we discuss the unique approach of the dblp computer science bibliography to this curation problem, and how we set up a system that uses automated heuristics based on domain knowledge, while keeping the human editors in the loop.

3.2 Citations in the world's largest encyclopedia – and their future

Phoebe Ayers (MIT Libraries – Cambridge, US) and Lydia Pintscher (Wikimedia – Germany, DE)

Wikipedia encompasses millions of articles in hundreds of language editions, written collaboratively by people who do not know each other. Rather than relying on known expert editors, Wikipedia's reliability rests on its citations to other reliable sources. Taken together, this body of references represents a huge, human curated and open collection of reference metadata and bibliographies on every topic imaginable. But what do these references in Wikipedia actually include, how do editors decide what to include, and what do we know about them? And what could citations be, and what could we know about them? We'll show what citations in Wikipedia look like now, and will talk about some ideas for a more structured approach to citations, "Wikicite", that has been in development in the Wikimedia community for several years. We will also discuss Wikidata and bibliographic metadata in Wikidata as well as the current efforts to scale Wikidata, especially in the area of bibliographic data.

3.3 Narrative Information Access in Digital Libraries

Wolf-Tilo Balke (TU Braunschweig, DE)

From early on, narratives have been used as an essential means to convey information and knowledge in a form that is close to human communication and sense making. Moreover, references to archetypical narratives, such as David vs. Goliath or Don Quijote, can also transport a set of connotations beyond the actual story allowing for a framing of information. Facing today's flood of data, data-driven narratives are thus an ideal way to make complex

topics comprehensible, to make sense of certain events, or to assess the plausibility of given narratives. However, these features are rarely used in Digital Libraries today. In particular, most of the current work on narratives is limited to representing structural properties such as story or plot graphs, event chains, or representations of entities and events without exploiting the deeper meaning of narratives. In this talk, we will explore narratives in the sense of logical overlays over heterogeneous knowledge repositories in Digital Libraries, such as knowledge graphs, linked open data sources, document collections, or even concrete datasets. In its simplest form, a narrative then is a directed graph consisting of entities, events, and literals as nodes. Narrative edges describe the flow of the modeled events, i.e. on the one hand the semantic interaction between events and entities and on the other hand the respective types of interaction by suitable edge labels (e.g., in the causal or temporal sense). Essential for the expressive power of this overlay model is that edges of a narrative must always be bound against underlying knowledge repositories. In particular, this allows the plausibility of each edge to be evaluated against a given set of trusted repositories. Of course, this also means that the information in the underlying repositories needs to be carefully extracted with respect to classical dimensions of data quality, such as correctness, completeness, or validity.

3.4 Knowledge Graphs and QLever

Hannah Bast (Universität Freiburg, DE)

I will give an introduction to knowledge graphs, RDF, and SPARQL, with many examples and demos. I hope that a lively discussion will ensue.

3.5 Zenodo and InvenioRDM: Cross-domain digital repositories for the long tail of research

Martin Fenner (Front Matter – Münster, DE)

InvenioRDM is an open source repository platform started at CERN that not only powers the Zenodo repository, but also an increasing number of other digital repositories developed and hosted by the InvenioRDM community. InvenioRDM integrates with DataCite, ORCID, ROR, Crossref, OpenAIRE and other Scholarly Information Systems and is a good example of the challenges and opportunities in building scholarly information systems.

3.6 FAIR Digital Objects as Scholarly Infrastructure Metadata Middleware

Carole Goble (University of Manchester, GB)

The scientific enterprise depends on finding, exchanging, understanding, validating, reproducing, integrating, reusing, archiving, and citing research entities across a dispersed community of researchers and an ecosystem of research and scholarly communication platforms. We are not just talking about papers, but also the mix of data, software, protocols, models and so on that are research outputs that together form the compendium of knowledge and the means for reuse and reproducibility of experimental outcomes. What is the unit of Knowledge Exchange for research? If actionable metadata needs to accompany our research entities (data and software) on their journeys to enable FAIR (Findable, Accessible, Interoperable, Reusable) research, how do we do that? If metadata is "a love note to the future", then what's the envelope? RO-Crates, that's what! Platform independent standards-based metadata framework for bundling resources with context into citable reproducible packages.

3.7 Please meet AI, our dear new colleague: Are we becoming obsolete?

Iryna Gurevych (TU Darmstadt, DE)

How can AI and LLMs facilitate the work of scientists in different stages of the research process? Can technology even make scientists obsolete? The role of AI and Large Language Models (LLMs) in science as the target application domain has recently been rapidly growing. This includes assessing the impact of scientific work, facilitating writing and revising manuscripts as well as intelligent support for manuscript quality assessment, peer-review and scientific discussions. The talk will illustrate such methods and models using several tasks from the scientific domain. We argue that while AI and LLMs can effectively support and augment specific steps of the research process, expert-AI collaboration may be a more promising mode for complex research tasks.

3.8 Research Fast and Slow

Min-Yen Kan (National University of Singapore, SG)

I will argue that while today's AI-driven, toolkit-rich environment enables rapid, incremental "fast" research – boosting citations and confidence – it also risks short-termism, stress, and crowding around the same problems. Drawing on Kahneman's Thinking, Fast and Slow and Friedman's "age of acceleration," I will highlight the importance of cultivating "slow" research: asking the right questions, seeking interdisciplinary perspectives, tolerating inconvenience, and developing long-term vision.

3.9 Overview of COnnecting REpositories (CORE)

Petr Knoth (The Open University – Milton Keynes, GB)

CORE (Connecting Repositories) is an open scholarly infrastructure that indexes millions of open access research papers and metadata from repositories and journals worldwide. Its goal is to improve the discoverability and reuse of research outputs and support machine access to scholarly content in line with open access and open science principles. This talk will provide an overview of CORE and its services for repositories, including compliance monitoring, metadata validation, and tools to improve interoperability and discoverability. It will also present research by the Big Scientific Data and Text Analytics Group (BSDTAG), showcasing recent innovations such as CORE-GPT, a system for trustworthy question answering over scholarly literature; SDG: Classify, which maps research papers to UN Sustainable Development Goals; and SoFAIR, which addresses reproducibility and research software management. Finally, the talk will discuss how CORE enables external research and innovation in areas such as training large language models, plagiarism detection, library discovery, and the construction of scholarly graphs, fostering a globally connected and machine-readable open research ecosystem.

3.10 Openness aspects of scholarly information systems important for adoption, transparency and interoperability

Bianca Kramer (Sesame Open Science – Utrecht, NL)

In this presentation, I will discuss various aspects of openness, and their relevance for development and maintenance of scholarly information systems, as well as for decisions around adoption and (financial) support. Taking from various existing frameworks (including the Principles of Open Scholarly Infrastructure, the CoARA working group Open Infrastructures for Responsible Research Assessment and the Barcelona Declaration on Open Research Information) I will put a number of openness criteria up for discussion, and solicit opinions from seminar participants how relevant each criterion is for them, and why. In the second part of the presentation, we will collectively look at a number of infrastructures represented in the room and how they are positioned against a limited set of openness criteria.

3.11 Scholarly Knowledge Graphs: No community, no fun

Paolo Manghi (Institute of Information Science and Technologies – CNR – Pisa, IT)

Scholarly Knowledge Graphs (SKGs) currently play a crucial role in enabling open data as a means of fostering transparent and trustworthy research assessment. However, the emergence of Open Science and its associated publishing workflows introduce a range of challenges

that cannot be addressed by SKGs alone. Drawing from the experience of developing the OpenAIRE Graph, this presentation will highlight some of these challenges and illustrate how the interaction among researchers, publishing data sources and venues, and SKGs creates a synergistic ecosystem. In this ecosystem, SKGs serve as essential tools to enhance and streamline scholarly workflows.

3.12 Collaborative curation of bibliographic metadata

Daniel Mietchen (FIZ Karlsruhe – Berlin, DE)

This talk sketches out several facets of a collaborative approach to the curation of scholarly information. It starts by outlining two use cases from the Wikipedia ecosystem: The first is that when media files from open-access sources are uploaded to Wikimedia Commons, then bibliographic metadata about the open-access source need to be curated with each file on Wikimedia Commons. The second is that if Wikipedia articles are translated from one language to another, then the bibliographic metadata of the cited references will have to be curated for each language version of the article. Such use cases would benefit from a shared repository of bibliographic information, which is what the WikiCite initiative is about that uses Wikidata as a hub to collect and curate bibliographic information about scholarly sources cited on Wikimedia projects. Next, the Scholia tool is introduced, which uses the information from Wikidata to provides some dozens of scholarly profile types (e.g. author, work, topic, organization, award, taxon, gene or location). For each profile type, it generates a HTML page that contains a number of visualizations that are based on preformulated Wikidata SPARQL queries parametrized by the entity to be profiled. This way, Scholia users do not need to have SPARQL knowledge but if they do, they can use it to finetune or modify the queries to explore the data further. Scholia has a range of features that are relevant to the discussion of open scholarly infrastructures. For instance, many Scholia profiles have an "improve data" button that leads to a dedicated curation page listing known data quality issues, e.g. author name strings yet to be disambiguated, along with links to tooling that helps address these issues. Some of its visualizations provide impact proxies. The profile for an individual work include a panel listing statements supported by a given work, or Wikipedia mentions of it, or information about relevant retractions. The profile for this very Dagstuhl Seminar includes co-author networks and a list of recent publications by seminar participants. Next, the platform zbMATH Open is introduced, which covers the mathematical research literature in a way similar to how the DBLP platform covers the computer science literature. zbMATH Open profiles authors and publications and, through its sister platform swMATH, also curates links between mathematical publications and the associated software. Finally, nanopublications are introduced as miniature knowledge graphs that can represent individual units of curation. They can be queried, shared and aggregated and are functionally similar to individual wiki edits or git commits, yet they are structured and machine actionable in a way that makes them compatible with larger knowledge graphs, for which they might serve as a vehicle to exchange curation information.

3.13 Google Dataset Search

Natasha Noy (Google – Mountain View, US)

Datasets constitute a key part of scientific knowledge: they are described in publications and referenced in them; datasets themselves connect with one another in intricate ways. I will build on our experience with Google Dataset Search and discuss the different types of these connections, focusing on datasets. I will discuss the complementary roles of community and tooling to build the connections.

3.14 OpenCitations and its new IT infrastructure

Mario Petrella (University of Bologna, IT)

This presentation will examine OpenCitations IT infrastructure, showing how our Kubernetes-based microservices architecture effectively implements the 'Living Will' principle by allowing complete replication of our open scholarly services in just a few hours. This technical approach ensures OpenCitations resilience and ability to continue operating independently of its current supporting institutions.

3.15 The AIDA Dashboard

Angelo Salatino (The Open University – Milton Keynes, GB)

The AIDA Dashboard is a powerful system for analysing and comparing scientific journals and conferences. By leveraging a large-scale Knowledge Graph that integrates billions of data points about research from multiple sources, it offers unique, sophisticated analytics and rankings. This tool gives researchers and other key stakeholders unique insights into the evolution of different venues and helps them make crucial decisions. In my talk, I will focus on two main challenges. These include a lack of conference data information outside of the Computer Science field (where we are so lucky to have DBLP), as well as the absence of fine-grained representation of research topics in several disciplines. These are critical to enable appropriate categorisation and management of information beyond computer science.

3.16 Navigating Scholarly Knowledge

Tilahun Abedissa Taffa (Leuphana Universität Lüneburg, DE)

Scholarly knowledge navigation tools are essential for finding relevant information from bibliographic data sources. This presentation focuses on complementary advances: ASK-DBLP, a user-in-the-loop KGQA system that converts natural language questions to editable SPARQL and mitigates schema drift; and a RAG-based scholarly QA layer on the NFDI4DataScience Gateway that enables conversational, federated access to diverse scientific databases.

3.17 TIB AI Assistant for Research

Sahar Vahdati (TIB – Hannover, DE)

The TIB AI Assistant for Research is an AI-supported, modular assistant designed to help scholars across the research lifecycle from ideation and question formulation to literature exploration and structured synthesis. It combines large language models with semantic/vector search and knowledge graphs (notably the Open Research Knowledge Graph), enabling natural-language querying and producing concise, structured outputs such as tabular evidence summaries. Emphasizing openness and reproducibility, components like ORKG Ask are open-source and operate over tens of millions of open-access publications. Together, these capabilities accelerate rigorous, transparent discovery while aligning with TIB's open-science mission.

3.18 Digital Science & the Research Data Ecosystem

Kathryn Weber-Boer (Digital Science – London, GB)

This presentation will describe a range of Digital Science software solutions (with a focus primarily on Figshare and Dimensions), in the context of their relationships to each other, the way they connect to, draw on, and–in turn–feed other data sources. Figshare is an open repository where users share datasets, figures, and other research-related output. Dimensions is a platform where open systems (Crossref, DataCite, ORCID) are used, their data are transformed and combined with proprietary data (IFI CLAIMS and ReadCube), and the results are then fed back into the open space (e.g., ROR, Covid-19 Dataset). The presentation will address the delicate relationship between industry and the open science movement, and the vision that drives our understanding of that relationship.

3.19 CiteSeerX and NDLTD

Jian Wu (Old Dominion University – Norfolk, US)

CiteseerX is one of the World's earliest digital library search engines serving 15 million academic documents crawled over the Web. CiteSeerX uses AI techniques for crawling, information classification, and information extraction. Over the past 20+ years, CiteSeerX has overcome a lot of challenges in an academic setting and is still indexed by Google Scholar. In this talk, we show the upcoming challenges of this legacy and discuss an emergent question about how to make CiteSeerX sustainable and continue to serve the academic communities.

The Networked Digital Library of Theses and Dissertations (NDLTD) is an international organization that promotes the creation, dissemination, and preservation of electronic theses and dissertations (ETDs). Established in 1987, NDLTD has organized an annual symposium since 1998, on ETD-related research and development. It also maintains a union catalog containing metadata for more than 6.5 million ETDs worldwide and publishes a dedicated journal. In recent years, NDLTD has faced significant infrastructure and organizational challenges and is seeking collaborative efforts, particularly with the United States Electronic Thesis and Dissertation Association (USETDA), to address these issues.

3.20 Challenges and opportunities in arXiv

Ramin Zabih (Cornell Tech – New York, US)

arXiv has been a key open access resource since 1991. It has become a dominant force in many areas of science, particularly in computer science, math and physics. Since 2002 arXiv has been hosted at Cornell University, which has provided a stable home and has helped it develop a broad funding base. While from the outside arXiv looks like a smoothly functioning machine, it actually is facing a wide range of difficult issues. I will cover some of the main challenges and opportunities, and give some insights into how the arXiv model has proven sustainable for over 3 decades.

4 Working groups

4.1 Theme: Metadata Excellence and Interoperability

Hannah Bast (Universität Freiburg, DE), Guillaume Cabanac (University of Toulouse, FR), Paolo Manghi (Institute of Information Science and Technologies – CNR – Pisa, IT), and Jian Wu (Old Dominion University – Norfolk, US)

The outcomes of the seminar regarding Metadata Excellence and Interoperability focus on harmonizing fragmented systems, adopting advanced querying technologies, and establishing collaborative models to reduce redundant manual work. The participants highlight that while many organizations enrich metadata, these efforts are often independent, leading to inconsistencies across platforms.

4.1.1 The COMET Approach to Collaborative Metadata

A primary outcome was the discussion of the *COMET (Collaborative Metadata)* model, which aims to align decentralized curation efforts with one another and with authoritative sources.

- *Reducing Redundancy:* By promoting shared principles and technical standards, COMET seeks to facilitate the deduplication of effort, enabling more efficient data exchange.
- *Cross-System Comparison:* Next steps involve using Knowledge Graphs (KGs) to automatically detect agreement or disagreement in metadata, such as author attribution, across different systems.
- *Affiliation Parsing:* The group plans to compare results on affiliation parsing from sources like arXiv to test the interoperability of current curation workflows.

4.1.2 Transitioning to Knowledge Graphs and SPARQL

The seminar identified a significant opportunity to move from "scripts and JSON blobs" to formal Scholarly Knowledge Graphs to simplify complex data tasks.

- *OpenReview and DBLP Integration:* Participants agreed that OpenReview's current REST API and MongoDB storage could be complemented by an RDF/SPARQL pipeline. This would allow conference organisers to run complex queries – such as finding authors with specific paper counts across multiple venues – without writing custom scripts.
- *Wikidata Scalability with QLever:* To address the scalability issues of the Wikidata Query Service (currently using Blazegraph), the group successfully tested QLever as a new backend. Scholia was identified as the primary test case for this migration, with promising results in rewriting SPARQL queries to be more standard-compliant.

4.1.3 Harmonising Subject Classification and Ontologies

The participants note a "perverse trend" where the proliferation of new standards actually hinders interoperability.

- *Reducing Standard Proliferation:* The strategic goal is to decrease the number of generic subject area classifications used by open infrastructures.

- *Integration of Ontologies:* Plans are underway to link DBLP records with the Computer Science Ontology (CSO), a taxonomy of 14,000 research topics, to provide better subject-level discoverability for venues and authors.
- *Automatic Classification:* Experiments are being conducted using the Leiden/OpenAlex model on Hugging Face to automatically classify tens of thousands of records within the InvenioRDM platform.

4.1.4 Advanced Researcher Disambiguation

Addressing the "author identity" challenge, the seminar explored strategies to combine automated and manual curation.

- *Evidence-Based Identification:* Disambiguation efforts will focus on combining various indicators, including co-authorship, subject area, and email domains.
- *Chinese Name Disambiguation:* A specific initiative involves using AMiner to improve the disambiguation of Chinese names, as traditional ASCII transliteration often loses critical variations present in Hanzi characters.
- *Researcher Autonomy:* The group emphasised that systems must respect researchers who choose to maintain multiple ORCID profiles for reasons of identity change or personal safety in different political contexts.

4.1.5 Expanding Metadata Scope

The seminar also looked toward "giving back" to the ecosystem by capturing information currently missing from major graphs:

- *Software and Acknowledgments:* Efforts are starting to index software mentions in DBLP, syncing workflows with zbMATH and swMATH. Additionally, nanopublications are being explored as a mechanism to track acknowledgments for infrastructures like Dagstuhl itself.
- *Fake and Spam Metadata:* A dedicated working group discussed methods for identifying and flagging "fake" or "spam" conferences to prevent them from polluting open metadata records.
- *Scalability of Metadata Access:* Moving beyond API calls, the group discussed the necessity of using regular data dumps in formats like Parquet or JSONL for large-scale analysis, as API-based retrieval does not scale for multi-million record collections.

4.2 Theme: Reforming Research Assessment

Hannah Bast (Universität Freiburg, DE), Guillaume Cabanac (University of Toulouse, FR), Paolo Manghi (Institute of Information Science and Technologies – CNR – Pisa, IT), and Jian Wu (Old Dominion University – Norfolk, US)

The results from Dagstuhl Seminar 25381 regarding Reforming Research Assessment center on a collective agreement to move away from what participants described as a "perverse system" of publication metrics and automated rankings. The consensus among the experts was that scientific quality and impact are multi-dimensional and cannot be reduced to a single numerical score.

4.2.1 Rejection of Automated Rankings

The seminar participants, including representatives from major data providers, reached a consensus on several fundamental points regarding the presentation of research data:

- *Quality vs. Numbers:* There was unanimous agreement that the quality of scientific research cannot be measured by numbers alone.
- *The Responsibility of Infrastructure:* Databases such as DBLP, Dimensions, and OpenReview should not automatically deliver publication statistics for the purpose of ranking or estimating expertise. Instead, these platforms should "live their values" by choosing how – and if – they display indicators.
- *Data over Rankings:* The primary mission of data providers should be to curate and provide comprehensive, transparent data, leaving the interpretation and derivation of statistics to the users for specific, contextual purposes.

4.2.2 Strategic "Refusal" of Metrics

A significant outcome was the discussion on how specific infrastructures intentionally limit the metrics they provide to prevent misuse:

- *Withholding H-indices:* It was noted that platforms like Dimensions and DBLP have explicitly chosen to refuse to provide H-indices or Impact Factors (IF) on their sites, even though the raw data to calculate them is available.
- *Contextualisation:* By default, search results in these systems are often sorted by date rather than "impact," forcing users to engage with the research itself rather than a pre-sorted list of "top" authors.
- *Transparency and "Data Cards":* To combat the "metric frenzy," the group proposed the use of "data cards" to make the assumptions behind research data clear, ensuring that any derived rankings are transparent and accountable.

4.2.3 Addressing the "Perverse" Incentives

The working groups explored why flawed metrics persist and how to transition to a more equitable system:

- *The Seniority Problem:* Participants discussed how the current system is maintained by tenured, senior colleagues who were selected by these very metrics. There is an urgent need for these "overloaded seniors" to drive change.
- *Information Overload:* It was acknowledged that rankings often persist because they are an "easy" solution to manage information overload and time pressure during hiring or grant review processes.
- *Multi-Criteria Assessment:* The group advocated for moving toward narrative-based assessments where researchers "name their top five" contributions, and evaluations include broader criteria such as teaching, collaboration styles, and institutional contributions.

4.2.4 Alignment with International Initiatives

The seminar's outcomes are designed to support and fill "actionable gaps" in existing international frameworks:

- *COARA and DORA:* The discussions were closely aligned with the Coalition for Advancing Research Assessment (CoARA) and the San Francisco Declaration on Research Assessment (DORA).

- *Institutional Advocacy:* A key next step is for participants to convince their own institutions to adopt formal strategies for handling publication metrics and to stop over-relying on automated counting.
- *The Barcelona Declaration:* Participants discussed using the Barcelona Declaration to promote open research information as the standard for decision-making, highlighting non-traditional outputs like datasets instead of just highly cited articles.

4.2.5 Summary

The seminar concluded that while statistics can be useful tools for specific goals, they must be contextualised and non-automated to prevent the "rich get richer" or Matthew effect[1] and to safeguard the diversity and integrity of the scholarly ecosystem.

4.3 Theme: Sustainability and Digital Sovereignty

Hannah Bast (Universität Freiburg, DE), Guillaume Cabanac (University of Toulouse, FR), Paolo Manghi (Institute of Information Science and Technologies – CNR – Pisa, IT), and Jian Wu (Old Dominion University – Norfolk, US)

The outcomes of the Dagstuhl Seminar regarding *Sustainability and Digital Sovereignty* emphasize a critical shift from viewing Open Scholarly Infrastructure (OSI) as a purely academic concern to framing it as a vital national and economic asset. Participants concluded that for open systems to survive, they must align with broader societal priorities, such as *security, AI productivity, and democratic resilience.*

In summary, the seminar participants agreed that the scholarly community must move toward co-ownership and collective responsibility to ensure that research information remains a transparent and shared global resource.

4.3.1 The Economic Imperative for Sustainability

The participants highlight that several established OSIs, including CiteSeerX, NDLTD, and CORE, currently face significant sustainability hurdles due to financial, administrative, and human resource constraints. To address these challenges, the seminar proposed several strategic shifts:

- *Dedicated Institutional Funding:* A primary recommendation is for universities and libraries to dedicate a specific proportion of their budgets – suggested at 5% – to support open scholarly infrastructure, rather than exclusively funding commercial publishers.
- *Aligning with National Priorities:* Future funding proposals should move beyond justifying needs from a researcher's perspective. Instead, they should demonstrate how OSIs support national priorities such as AI development and digital security.
- *Diverse Business Models:* While philanthropic funding has supported entities like OpenAlex, the group noted that such support is not a guaranteed long-term solution. Instead, they explored membership models (e.g., CORE), donations (e.g., Wikipedia), and ERIC (European Research Infrastructure Consortia) which utilize national subscriptions.

[1] https://en.wikipedia.org/wiki/Matthew_effect

- *Consolidation and Efficiency:* To ensure long-term preservation, the group suggested that under-supported OSIs might need to merge or integrate. For example, work is already underway to integrate CiteSeerX into the Internet Archive Scholar.

4.3.2 Digital Sovereignty and the Knowledge Economy

Digital sovereignty was identified as the need for nations to retain long-term control over their research outputs to avoid dependency on commercial monopolies. Key discussion points included:

- *The Threat of Commercial Monopolies:* Participants noted that institutions are increasingly reliant on commercial Current Research Information Systems (CRIS), which are often non-interoperable and require high annual fees.
- *OSI as the Backbone of AI:* The participants argue that we cannot have trustworthy, evidence-based AI without OSI, as open infrastructures provide the essential reliable corpora needed for AI models.
- *Resilience and Security:* Distributed and interoperable infrastructures provide protection against sanctions, cyber threats, and misinformation. The group emphasized that "OSI saves lives," citing the role of transparent, reproducible science in developing vaccines and combating medical misinformation.

4.3.3 Strategic Actions and the Barcelona Declaration

The seminar leveraged the *Barcelona Declaration on Open Research Information* as a central framework for these efforts. Notable progress and next steps include:

- *Expanding Signatories:* Efforts are underway to add organizations such as CWI and NWO-I as signatories to the Barcelona Declaration.
- *International Advocacy:* Representatives will take these arguments to high-level policy discussions with the OECD, G7 Open Science Working Group (OSWG), and UKRI.
- *Manifesto for Decision Makers:* The group produced a draft manifesto detailing the economic and social value of OSI to convince decision makers of the return on investment (RoI) provided by open systems.
- *The COMET Approach:* To reduce redundancy and enhance sustainability, the COMET model was proposed to align decentralized metadata curation efforts, ensuring that infrastructure "owners" collaborate rather than duplicate work.

4.4 Theme: The Rise of Agentic AI

Hannah Bast (Universität Freiburg, DE), Guillaume Cabanac (University of Toulouse, FR), Paolo Manghi (Institute of Information Science and Technologies – CNR – Pisa, IT), and Jian Wu (Old Dominion University – Norfolk, US)

The outcomes regarding The Rise of Agentic AI from Dagstuhl Seminar 25381 focus on the transformative potential of autonomous agents across the research lifecycle, while simultaneously highlighting the risks to scientific integrity and human agency. Participants explored how these systems might evolve from simple assistants to semi-autonomous collaborators.

4.4.1 Levels of Agent Autonomy

A core framework discussed during the seminar defines five levels of autonomy for AI agents in a research context, based on the degree of human involvement:

- *L1 (Operator):* The human directs all decisions and actions.
- *L2 (Collaborator):* Human and agent plan and execute tasks together.
- *L3 (Consultant):* The agent takes the lead but consults the human for preferences or expertise.
- *L4 (Approver):* The agent operates autonomously and only seeks human intervention for risky or pre-specified scenarios.
- *L5 (Observer):* The agent operates with full autonomy under human monitoring.

4.4.2 Integration with Scholarly Infrastructure

The group debated whether Open Scholarly Information Systems (OSIS) must reinvent themselves to remain relevant in an AI-dominated landscape.

- *Model Context Protocol (MCP):* There is a strategic discussion regarding whether OSIS services should be offered as MCP services. This would allow AI agents to use scholarly databases as reliable corpora for tasks like Retrieval-Augmented Generation (RAG).
- *The Shift to Bot Users:* Predictions suggest that by 2030, search engines and scholarly systems will be used primarily by AI bots rather than human researchers. This necessitates new protocols for data access and perhaps "micro-credits" to recoup the costs of high-frequency bot requests.

4.4.3 Impact on the Research Lifecycle

The seminar examined specific case scenarios where agentic AI is already active or emerging:

- *Replication of Baselines:* AI agents are being tested for "boring" research tasks, such as replicating baseline experiments. While systems like Sakana's "AI Scientist" can produce full manuscripts rapidly and at low cost (USD 6–15 per paper), evaluations showed significant flaws, including poor novelty assessments, coding errors in 42% of experiments, and hallucinated results.
- *Peer Review:* While LLMs can support human reviewers by providing evidence or checking submission guidelines, they currently fail to detect faulty research logic. Studies indicate that flaws in internal consistency often go unnoticed by fully automatic review generators.
- *Literature Reviews:* AI can achieve high recall in identifying citations but lacks the nuanced judgment and background knowledge of an expert reviewer.

4.4.4 Safeguarding Human Agency and Education

A recurring concern throughout the sessions was the potential for deskilling among researchers.

- Critical Thinking: Participants emphasized that human agency, critical thinking, and accountability must remain at the heart of scholarship.
- Reforming the PhD: The role of PhD training may need to shift from technical proficiency to high-level analytical skills and "human-AI co-construction".
- The Future of H-AI Collaboration: The consensus was that the future lies in Human-AI collaboration, where humans remain responsible for the integrity and ethical purpose of the research while leveraging AI to tackle increasingly complex problems.

4.4.5 Strategic Next Steps

The results of these discussions are being compiled into a position paper (currently being drafted on Overleaf) titled “oAsIs”, focusing on the role of open agentic scholarly information systems. Ongoing work will also investigate preference-based cooperation (HAICo2) to ensure AI agents adapt to human expertise and institutional guidelines.

Participants

- Marcel R. Ackermann
Schloss Dagstuhl – Trier, DE
- Phoebe Ayers
MIT Libraries – Cambridge, US
- Wolf-Tilo Balke
TU Braunschweig, DE
- Hannah Bast
Universität Freiburg, DE
- Guillaume Cabanac
University of Toulouse, FR
- A. Seza Dogruöz
Ghent University, BE
- Martin Fenner
Front Matter – Münster, DE
- Ingo Frommholz
MODUL Universität Wien, AT
- Carole Goble
University of Manchester, GB
- Iryna Gurevych
TU Darmstadt, DE
- Lynda Hardman
CWI – Amsterdam, NL
- Holger Hermanns
Universität des Saarlandes – Saarbrücken, DE
- Min-Yen Kan
National University of Singapore, SG
- Petr Knoth
The Open University – Milton Keynes, GB
- Bianca Kramer
Sesame Open Science – Utrecht, NL
- Christin Kreutz
THM – Gießen, DE
- Cyril Labbé
University of Grenoble, FR
- Michael Ley
Schloss Dagstuhl – Trier, DE
- Paolo Manghi
Institute of Information Science and Technologies – CNR – Pisa, IT
- Philipp Mayr
GESIS – Köln, DE
- Daniel Mietchen
FIZ Karlsruhe – Berlin, DE
- Carlos Daniel Mondragón Chapa
OpenReview – Cambridge, US
- Patrick Neises
Schloss Dagstuhl – Trier, DE
- Natasha Noy
Google – Mountain View, US
- Mario Petrella
University of Bologna, IT
- Lydia Pintscher
Wikimedia – Germany, DE
- Ruzica Piskac
Yale University – New Haven, US
- Rüdiger Reischuk
Universität zu Lübeck, DE
- Angelo Salatino
The Open University – Milton Keynes, GB
- Ralf Schenkel
Universität Trier, DE
- Ansgar Scherp
Universität Ulm, DE
- Raimund Seidel
Universität des Saarlandes – Saarbrücken, DE
- Tilahun Abedissa Taffa
Leuphana Universität Lüneburg, DE
- Sahar Vahdati
TIB – Hannover, DE
- Nees Jan van Eck
Leiden University, NL
- Ruijie Wang
Universität Zürich, CH
- Kathryn Weber-Boer
Digital Science – London, GB
- Jian Wu
Old Dominion University – Norfolk, US
- Ramin Zabih
Cornell Tech – New York, US

Report from Dagstuhl Seminar 25382

Quantum Error Correction Meets ZX-Calculus

Miriam Backens[*1], **Aleks Kissinger**[*2], **John van de Wetering**[*3], **Michael Vasmer**[*4], **and Sarah Meng Li**[†5]

1 **INRIA Nancy – Grand-Est Research Center, FR.** miriam.backens@inria.fr
2 **University of Oxford, GB.** aleks.kissinger@cs.ox.ac.uk
3 **University of Amsterdam, NL.** j.m.m.vandewetering@uva.nl
4 **INRIA – Paris, FR.** michael.vasmer@inria.fr
5 **University of Amsterdam, NL.** m.li2@uva.nl

Abstract

This report documents the Dagstuhl Seminar 25382 "Quantum Error Correction Meets ZX-Calculus". The report consists of an executive summary, as well as abstracts on talks, working groups, panel discussions, and open problems.

Seminar September 14–19, 2025 – https://www.dagstuhl.de/25382
2012 ACM Subject Classification Theory of computation → Quantum computation theory
Keywords and phrases fault-tolerance, quantum error correction, ZX-calculus
Digital Object Identifier 10.4230/DagRep.15.9.58

1 Executive Summary

Miriam Backens (INRIA Nancy – Grand-Est Research Center, FR)
Aleks Kissinger (University of Oxford, GB)
John van de Wetering (University of Amsterdam, NL)
Michael Vasmer (INRIA – Paris, FR)

Seminar Topics

To achieve the transformational use cases of quantum computers, quantum error correction (QEC) must be used to protect delicate quantum information by encoding logical quantum bits across many physical qubits. Fault-tolerant logical gates are then used to process the encoded information reliably and execute quantum algorithms. This, however, comes with a large resource overhead. To this end, extensive research has been carried out to study QEC and optimise fault-tolerant quantum computation.

The ZX-calculus is a graphical language for reasoning about quantum computations. It can express computations in different models, such as quantum circuits or the one-way model. It is complete, in the sense that any true equality between diagrams can be derived entirely graphically. Over the past decade, the ZX-calculus has been used to optimise quantum computations and map logical circuits to hardware architectures.

Initial steps have already been taken in applying ZX-calculus to quantum error correction and fault tolerance, but the two communities thus far have remained mostly separate. In this Dagstuhl Seminar 25382 "Quantum Error Correction Meets ZX-Calculus", we share knowledge and foster collaboration between experts from both communities. Below lists the main topics that have been discussed.

[*] Editor / Organizer
[†] Editorial Assistant / Collector

The ZX-calculus and mainstream fault-tolerant quantum computation: How do ZX-based methods compare to and/or complement the techniques and representations used in the larger community to study diverse quantum error-correcting codes such as colour codes, Floquet codes, bicycle codes, and quantum low-density parity check codes? How can the ZX-calculus be used to improve fault-tolerant quantum compilation and help solve open problems in developing quantum error-correcting codes? For example, can we use the ZX-calculus to analyze various fault-tolerant protocols, design more efficient fault-tolerant compilation strategies, and connect different frameworks for dynamical codes?

Quantum error correction beyond static codes: Building on recent work about Floquet codes and the spacetime frameworks for QEC, as well as insights from measurement-based quantum computing, what dynamical protocols can we develop that go beyond the state-of-the-art? For example, can we optimise Floquetification procedures with respect to qubit and gate count, qubit connectivity, and number of measurement cycles?

The ZX-calculus and error correction beyond qubits: Error-correcting codes based on qudits or bosons have advantages over standard qubit-based codes. What new protocols for quantum error-correcting codes can we develop by leveraging the ZX-calculi for qudits or bosonic modes as a unified language? For example, can we use the ZX-calculus for bosonic modes to derive improved protocols for preparing resource states such as Gottesman-Kitaev-Preskill states and cluster state fragments?

By bringing together researchers and industry practitioners to discuss these challenges in small groups, our aim was to bridge the ZX-calculus and QEC communities and foster collaborative efforts toward advancing fault-tolerant quantum computation. Several groups generated promising insights and ideas with the potential to develop into strong projects, and we look forward to seeing how these unfold in the coming years.

Seminar Program

This Dagstuhl Seminar 25382 "Quantum Error Correction Meets ZX-Calculus" brought together researchers from quantum error correction, ZX calculus, and fault-tolerant quantum computing (FTQC) to explore new synergies between graph-theoretical methods and emerging FTQC frameworks. Over the course of five days, the program combined tutorials, structured discussions, and collaborative breakout sessions to develop shared understanding and identify promising research directions.

The seminar opened with in-depth tutorials on quantum error correction through the lens of ZX-calculus and on error detection in Clifford protocols, establishing a common technical foundation for participants. Subsequent sessions expanded the scope to higher-dimensional quantum systems, fault-tolerance-by-construction techniques, and decoding approaches for qLDPC codes. These tutorials sparked active discussions and shaped the formation of working groups.

A central component of the seminar was the daily cycle of brainstorming, progress reporting, and regrouping. Participants identified key open questions, self-organized into focused teams, and iteratively refined problem statements across the week. This structure fostered cross-disciplinary collaboration, allowing ideas developed in earlier tutorials to inform concrete research tasks. Group activities, including a mid-week hike, further strengthened informal communication and collaboration within the community.

By the end of the seminar, the working groups had narrowed down several directions for ongoing research, including improved interfaces between ZX-based reasoning and error-correction formalisms, as well as methodologies for constructing fault-tolerant protocols via compositional or automated tools. The momentum built during the seminar is expected to carry forward, fostering collaboration between previously separate research areas and supporting the development of future joint projects.

2 Table of Contents

Executive Summary
Miriam Backens, Aleks Kissinger, John van de Wetering, and Michael Vasmer . . 58

Overview of Talks

From Stabiliser Diagrams to Fault-Equivalent Circuits: A ZX-Calculus Approach
Aleks Kissinger 62

Detecting Errors in Clifford Protocols
Julio Carlos Magdalena de la Fuente 62

Fault Tolerance by Construction
Benjamin Rodatz 63

Tutorial on Approximate Synthesis
Peter Selinger 64

Quantum Error Correction Beyond Qubits
Christophe Vuillot 64

Working groups

Distance Lower Bounds from Graphs
Andrey Khesin 65

Quantum Acupuncture: Characterise Hook Errors in Syndrome Extraction
Aleks Kissinger 65

Noise Model in Fault-Equivalent Rewrites
Julio Carlos Magdalena de la Fuente 66

Relation between Chain Maps and ZX Rewrites
Arthur Pesah 66

Understanding Fault Equivalence for Fault-Tolerant Quantum Computing
Benjamin Rodatz 67

Panel discussions

Decoder
Linnea Grans-Samuelsson 67

ZX Software Stack for FTQC: Today's Tools and Tomorrow's Needs
John van de Wetering 68

Open problems

ZXify Basic QEC
Sarah Meng Li and Miriam Backens 68

Errors during Non-Clifford Logical Gates
Julio Carlos Magdalena de la Fuente 68

Represent GKP Codes Using ZX Diagrams
Alex Townsend-Teague 69

Participants 70

3 Overview of Talks

3.1 From Stabiliser Diagrams to Fault-Equivalent Circuits: A ZX-Calculus Approach

Aleks Kissinger (University of Oxford, GB)

Joint work of Aleks Kissinger, John van de Wetering, Benjamin Rodatz, Boldizsár Poór
Main reference Aleks Kissinger: "Phase-free ZX diagrams are CSS codes (...or how to graphically grok the surface code)", CoRR, Vol. abs/2204.14038, 2022.
URL https://arxiv.org/abs/2204.14038
Main reference Aleks Kissinger, John van de Wetering: "Picturing Quantum Software: An Introduction to the ZX-Calculus and Quantum Compilation", Preprint, 2024.
URL https://github.com/zxcalc/book
Main reference Benjamin Rodatz, Boldizsár Poór, Aleks Kissinger: "Fault Tolerance by Construction", CoRR, Vol. abs/2506.17181, 2025.
URL https://arxiv.org/abs/2506.17181

We explore a unified framework for fault-tolerant quantum computation by combining graph-theoretic formalisms with error-correcting code constructions. First, building on the one-to-one correspondence between the phase-free ZX diagrams and Calderbank–Shor–Steane (CSS) codes, we give a fully graphical description of stabiliser states and various code transformation protocols such as lattice surgery and code switching.

Second, we extend ZX-calculus with a refined set of rewrite rules that preserve not only the semantics of the underlying linear maps, but also the diagram's behaviour under noise. As a result, we can tranform ZX diagrams while preserving how errors propagate through them. Fault-tolerant synthesis steps, such as those used in syndrome extraction and state preparation, can therefore be performed directly on ZX diagrams. This provides a unified, correct-by-construction approach to reasoning about and compiling fault-tolerant quantum computations at the diagrammatic level.

3.2 Detecting Errors in Clifford Protocols

Julio Carlos Magdalena de la Fuente (FU Berlin, DE)

Joint work of Julio Carlos Magdalena De La Fuente, Josias Old, Alex Townsend-Teague, Manuel Rispler, Jens Eisert, Markus Müller
Main reference Julio C. Magdalena de la Fuente, Josias Old, Alex Townsend-Teague, Manuel Rispler, Jens Eisert, Markus Müller: "XYZ Ruby Code: Making a Case for a Three-Colored Graphical Calculus for Quantum Error Correction in Spacetime", PRX Quantum, Vol. 6, p. 010360, American Physical Society, 2025.
URL https://doi.org//10.1103/PRXQuantum.6.010360

A systematic understanding of error detection and correction within quantum circuits is essential to develop scalable and reliable quantum computing architectures. In this tutorial, I gave an introduction into a circuit-centered perspective on active quantum error-correction in Clifford circuits. After motivating the problem of protecting information with active measurement dynamics, I started with a short introduction into the stabiliser formalism, the main technical tool to study the effect of errors in Clifford circuits.

The center of the tutorial was the decoding problem: Given an error model in a Clifford circuit we can rigorously define a classical decoding problem based on certain combinations of measurement outcomes in the circuit. The main method used in the presentation was a

diagramatic representation of measurement circuits, in particular the ZX calculus. Using ZX rewrites and Pauli symmetries of the tensors appearing in a Clifford circuit, we can deduce linear constraints on the measurement outcomes that are used for decoding. This naturally defines the concept of a fault-distance, which captures elementary properties of a circuit to be able to correct for errors up to a given weight.

At the end, I sketched how to incorporate logically non-trivial operations into the framework without resorting to a fixed input encoding but focussing on a protocol-first perspective.

3.3 Fault Tolerance by Construction

Benjamin Rodatz (University of Oxford, GB)

Joint work of Benjamin Rodatz, Boldizsár Poór, Aleks Kissinger
Main reference Benjamin Rodatz, Boldizsár Poór, Aleks Kissinger: "Fault Tolerance by Construction", CoRR, Vol. abs/2506.17181, 2025.
URL https://arxiv.org/abs/2506.17181
Main reference Benjamin Rodatz, Boldizsár Poór, Aleks Kissinger: "Floquetifying stabiliser codes with distance-preserving rewrites", CoRR, Vol. abs/2410.17240, 2024.
URL https://arxiv.org/abs/2410.17240

A key challenge in fault-tolerant quantum computing is synthesising and optimising circuits in a noisy environment, as traditional techniques often fail to account for the effect of noise on circuits. In this work, we propose a framework for designing fault-tolerant quantum circuits that are correct by construction. The framework starts with idealised specifications of fault-tolerant gadgets and refines them using provably sound basic transformations.

To reason about manipulating circuits while preserving their error correction properties, we define fault equivalence; two circuits are considered fault-equivalent if all undetectable faults on one circuit have a corresponding fault on the other. This guarantees that the effect of undetectable faults on both circuits is the same. We argue that fault equivalence is a concept that is already implicitly present in the literature. Many problems, such as state preparation and syndrome extraction, can be naturally expressed as finding an implementable circuit that is fault-equivalent to an idealised specification.

To utilize fault equivalence in a computationally tractable manner, we adapt the ZX calculus, a diagrammatic language for quantum computing. We restrict its rewrite system to not only preserve the underlying linear map but also fault equivalence, i.e. the circuit's behaviour under noise. Enabled by our framework, we verify, optimise and synthesise new and efficient circuits for syndrome extraction and cat state preparation. We confirm the improved performance of our optimised circuits in simulation. We anticipate that fault equivalence can capture and unify different approaches in faulttolerant quantum computing, paving the way for an end-to-end circuit compilation framework.

3.4 Tutorial on Approximate Synthesis

Peter Selinger (Dalhousie University – Halifax, CA)

Joint work of Neil J. Ross, Peter Selinger
Main reference Neil J. Ross, Peter Selinger: "Optimal ancilla-free Clifford+T approximation of z-rotations", CoRR, Vol. abs/1403.2975, 2014.
URL http://arxiv.org/abs/1403.2975

We introduced the Gridsynth algorithm, which solves the problem of approximate synthesis for single-qubit Clifford+T operators. There are three main ingredients to this algorithm: solving a lattice grid problem, solving a Diophantine equation, and exact synthesis.

We started by talking about the classic Diophantine equation $a^2 + b^2 = n$, which Fermat solved in 1640. It turns out that there is a very efficient algorithm which inputs n, and either outputs an integer solution (a, b) or determines that no such solution exists, *provided* that we can factor n. Also, the above equation is equivalent to $tt^\dagger = n$, where $t = a + ib \in \mathbb{Z}[i]$ is a Gaussian integer. It turns out that the same algorithm can also be used to solve the equation $tt^\dagger = \xi$, where $\xi \in \mathbb{Z}[\sqrt{2}]$ is given and $t \in \mathbb{Z}[\omega]$ is unknown. Here $\omega = e^{\pi/4}$ is an 8th root of unity.

The second ingredient to the approximate synthesis algorithm is lattice grid problems. Given a convex subset $C \subseteq \mathbb{R}^k$, the problem is to enumerate all points of C that have integer coordinates. The LLL algorithm can solve this problem efficiently, provided that the dimension k is fixed and small.

The third ingredient is exact synthesis. Given a unitary 2×2-operator U, it is a theorem by Kliuchnikov, Maslov, and Mosca that U can be exactly represented over the Clifford+T gate set if and only if the matrix entries of U are in the ring $\mathbb{Z}[\frac{1}{2}, \omega]$. Moreover, the T-count of the resulting word can be predicted from the largest power $\sqrt{2}^k$ appearing in the denominator of the matrix entries.

Putting these ingredients together, we can an efficient algorithm for approximate synthesis. Given a unitary z-rotation $V = \begin{pmatrix} z & 0 \\ 0 & z^\dagger \end{pmatrix}$, our goal is to find an operator $U = \begin{pmatrix} u & -t^\dagger \\ t & u^\dagger \end{pmatrix}$ approximating it to within ϵ. This can be done by first choosing $u \in \mathbb{Z}[\frac{1}{2}, \omega]$ to lie within a certain convex region of the complex numbers, and such that $u^\bullet$ is in the unit circle. This turns out to be a lattice grid problem. Next, we must find $t \in \mathbb{Z}[\frac{1}{2}, \omega]$ such that $uu^\dagger + tt^\dagger = 1$. This is a Diophantine equation. Finally, one uses exact synthesis to turn the resulting operator into a word over the generators. We saw a demo of the algorithm that approximated a z-rotation by angle 47 degrees up to $\epsilon = 10^{-100}$ in less than a second.

3.5 Quantum Error Correction Beyond Qubits

Christophe Vuillot (Alice & Bob – Paris, FR)

Joint work of Christophe Vuillot, Alessandro Ciani, Barbara M. Terhal
Main reference Christophe Vuillot, Alessandro Ciani, Barbara M. Terhal: "Homological Quantum Rotor Codes: Logical Qubits from Torsion". Commun. Math. Phys. 405, 53 (2024).
URL https://doi.org/10.1007/s00220-023-04905-4

We formally define homological quantum rotor codes which use multiple quantum rotors to encode logical information. These codes generalize homological or CSS quantum codes for qubits or qudits, as well as linear oscillator codes which encode logical oscillators. Unlike for

qubits or oscillators, homological quantum rotor codes allow one to encode both logical rotors and logical qudits in the same block of code, depending on the homology of the underlying chain complex. In particular, a code based on the chain complex obtained from tessellating the real projective plane or a Möbius strip encodes a qubit. We discuss the distance scaling for such codes which can be more subtle than in the qubit case due to the concept of logical operator spreading by continuous stabilizer phase-shifts. We give constructions of homological quantum rotor codes based on 2D and 3D manifolds as well as products of chain complexes. Superconducting devices being composed of islands with integer Cooper pair charges could form a natural hardware platform for realizing these codes: we show that the 0-π-qubit as well as Kitaev's current-mirror qubit – also known as the Möbius strip qubit – are indeed small examples of such codes and discuss possible extensions.

4 Working groups

4.1 Distance Lower Bounds from Graphs

Andrey Khesin (University of Oxford, GB)

Graphs can be used to express stabiliser error-correcting codes. Graphs from graph states give rise to natural stabilisers for these codes, associated locally with the graph's vertices. Using a greedy decoding algorithm, it can be shown that this algorithm cannot go wrong when there are few enough errors. Furthermore, the algorithm performs better if the graphs satisfy certain conditions, such as having no short cycles. This gives a generic way to get distance lower bounds for constructed codes without using topological arguments.

4.2 Quantum Acupuncture: Characterise Hook Errors in Syndrome Extraction

Aleks Kissinger (University of Oxford, GB)

For certain codes, it is ok to design syndrome extraction circuits in ways that are not, in themselves, fault-tolerant, i.e. single errors in measurement circuits can propagate out to multiple errors on data qubits. However, if these "hook errors" are perpendicular to logical operators, they don't necessarily decrease the code distance. Previously, we didn't know how to think about this graphically, but we seem to have come up with a new technique, which is called *quantum acupuncture*. It can add certain fault gadgets using fault-equivalent rewrites and rigorously prove that these "ok" syndrome extraction circuits really are ok. This might work for a broader class of codes than previously considered.

4.3 Noise Model in Fault-Equivalent Rewrites

Julio Carlos Magdalena de la Fuente (FU Berlin, DE)

We investigated equivalences/mappings between noise model representations in ZX-calculus, focusing in particular on coarse-graining techniques and fault-bounded rewrites. Our aim was to develop well-defined transformations that convert physically motivated noise models into forms that are easier to understand and simulate, where relative accuracy can be reliably understood.

4.4 Relation between Chain Maps and ZX Rewrites

Arthur Pesah (University College London, GB)

Joint work of Arthur Pesah, Michael Vasmer
Main reference Arthur Pesah, Austin K. Daniel, Ilan Tzitrin, Michael Vasmer: "Fault-tolerant transformations of spacetime codes", CoRR, Vol. abs/2509.09603, 2025.
URL https://arxiv.org/abs/2509.09603

Some recent work by two members of the seminar (Arthur Pesah and Michael Vasmer) showed that two circuits can be shown to be "equivalent" in terms of fault-tolerance by modelling them as chain complexes, and mapping one into the other using chain maps. They found some elementary chain maps from which many circuit equivalences can be shown. In parallel, some other members of the seminar (Ben Rodatz and Julio Magdalena de la Fuente) have worked on circuit equivalences using distance-preserving rewrite rules in ZX calculus and tensor networks. The question has then emerged on whether those two frameworks are equivalent. All the authors of those ideas, along with other members of the seminar, have discussed how to bridge the two formalisms. They have made progress on establishing the exact connection between the two (i.e., ZX diagrams can be associated to a chain complex and rewrite rules as chain maps) and discussed the path towards fully bridging the two (i.e., finding how all the elementary ZX rules are described in terms of chain maps, and how idealization of edges is described in the chain complex formalism).

4.5 Understanding Fault Equivalence for Fault-Tolerant Quantum Computing

Benjamin Rodatz (University of Oxford, GB)

Joint work of Benjamin Rodatz, Boldizsár Poór, Aleks Kissinger
Main reference Benjamin Rodatz, Boldizsár Poór, Aleks Kissinger: "Fault Tolerance by Construction", CoRR, Vol. abs/2506.17181, 2025.
URL https://arxiv.org/abs/2506.17181
Main reference Daniel Gottesmann: "Surviving as a quantum computer in a classical world". Textbook manuscript preprint, 8(8.1), 8-2, 2024
URL http://www.cs.umd.edu/ dgottesm/QECCbook-2024.pdf

Fault equivalence is a novel and recently defined relationship between quantum circuits under noise. It formalises when the behaviour of two noisy circuits can be considered the same, and therefore when they can be interchanged. This concept captures and formalises many ideas found in the literature on fault-tolerant quantum computing. While checking fault equivalence is generally NP-hard, the ZX calculus offers a way to efficiently manipulate quantum circuits while preserving fault equivalence.

Throughout this week, we have explored various aspects of fault equivalence and its role in fault-tolerant quantum computing. These discussions included formalising existing notions – such as Gottesman's fault tolerance conditions – in terms of fault equivalence, as well as examining the decoding problem under fault-equivalent rewrites and reasoning locally about hook errors using fault equivalence.

5 Panel discussions

5.1 Decoder

Linnea Grans-Samuelsson (University of Oxford, GB)

We discussed two types of decoding. Joschka Roffe gave a tutorial on how Belief Propagation (BP) decoding works. He showed an example of BP decoding of the classical repetition code. Messages get passed from checks to bits and from bits to checks, each message containing a belief that is used to update the estimated probability of each bit having experienced a fault.

Linnea Grans-Samuelsson gave a tutorial on the difference between maximum probability decoding, where the decoder returns a maximally probable error compatible with a syndrome, and a maximum likelihood (optimal) decoder, which computes the probability of each equivalence class of errors and returns a correction belonging to the most probable class.

5.2 ZX Software Stack for FTQC: Today's Tools and Tomorrow's Needs

John van de Wetering (University of Amsterdam, NL)

Main reference Aleks Kissinger, John van de Wetering: "PyZX: Large Scale Automated Diagrammatic Reasoning", in Proc. of the Proceedings 16th International Conference on Quantum Physics and Logic, QPL 2019, Chapman University, Orange, CA, USA, June 10-14, 2019, EPTCS, Vol. 318, pp. 229–241, 2019.
URL https://doi.org//10.4204/EPTCS.318.14

I presented an overview of PyZX's current capabilities: optimisation, visualisation, and circuit extraction. Then, I gave a demonstration of the graphical proof assistant ZXLive. This was followed by a discussion on desirable future software tools, with a particular emphasis on applications in quantum error correction.

6 Open problems

6.1 ZXify Basic QEC

Sarah Meng Li (University of Amsterdam, NL) and Miriam Backens (INRIA Nancy – Grand-Est Research Center, FR)

Joint work of Sarah Meng Li, Miriam Backens, Aleks Kissinger, John van de Wetering
Main reference Aleks Kissinger, John van de Wetering: "Picturing Quantum Software: An Introduction to the ZX-Calculus and Quantum Compilation", Preprint, 2024.
URL https://github.com/zxcalc/book

We examined several directions for visualising basic quantum error correction and fault tolerance using the ZX calculus. These included analysing hook errors in the surface code, as well as understanding and unifying magic state distillation protocols.

6.2 Errors during Non-Clifford Logical Gates

Julio Carlos Magdalena de la Fuente (FU Berlin, DE)

Pauli errors that happen during the application of a logical non-Clifford gate can create random syndromes. This is a problem when simulating erroneous circuits and decoding them. We discussed how scalable ZX calculus can help in analyzing the effect of Pauli errors between logical gates in the third level of the Clifford hierarchy and stabilizer measurements. For diagonal gates, a normal form provides a compact representation, and we discussed possibilities for utilizing it in the context of simulating syndrome statistics and the propagation of Clifford errors through the circuit.

6.3 Represent GKP Codes Using ZX Diagrams

Alex Townsend-Teague (FU Berlin, DE)

We completed the representation of the general tiger code X-part projector, which required introducing a multiplier analogue for the Fock basis and developing a scalable notation to package these elements in a compact form. However, the resulting constructions produced diagrams of considerable size and complexity. To mitigate this, we shifted attention to a simpler case – the cat code – and attempted to express its code states and codespace projectors directly as CV ZX-diagrams. Despite these efforts, we were ultimately unable to obtain a satisfactory diagrammatic formulation.

Participants

- Miriam Backens
INRIA Nancy – Grand-Est Research Center, FR
- Simon Burton
Quantinuum – Cambridge, GB
- Ophelia Crawford
Riverlane – Cambridge, GB
- Alexander Frei
University of Waterloo, CA
- Linnea Grans-Samuelsson
University of Oxford, GB
- Mackenzie Hooper Shaw
TU Delft, NL
- Jiaxin Huang
University of Hong Kong, HK
- Andrey Khesin
University of Oxford, GB
- Aleks Kissinger
University of Oxford, GB
- Sarah Meng Li
University of Amsterdam, NL
- Julio Carlos Magdalena de la Fuente
FU Berlin, DE
- Alexandra Moylett
Nu Quantum – Cambridge, GB
- Ewan Murphy
University of Waterloo, CA
- Hironari Nagayoshi
University of Tokyo, JP
- Armanda O. Quintavalle
FU Berlin, DE
- Simon Perdrix
LORIA – Nancy, FR
- Arthur Pesah
University College London, GB
- Clément Poirson
INRIA – Paris, FR
- Benjamin Rodatz
University of Oxford, GB
- Joschka Roffe
University of Edinburgh, GB
- Thomas Scruby
Okinawa Institute of Science and Technology, JP
- Peter Selinger
Dalhousie University – Halifax, CA
- Alex Townsend-Teague
FU Berlin, DE
- John van de Wetering
University of Amsterdam, NL
- Michael Vasmer
INRIA – Paris, FR
- Christophe Vuillot
Alice & Bob – Paris, FR
- Lia Yeh
University of Oxford, GB
- Sascha Zakaib-Bernier
University of Waterloo, CA

Report from Dagstuhl Seminar 25391

Retrieval-Augmented Generation – The Future of Search?

Matthias Hagen[*1], Josiane Mothe[*2], Smaranda Muresan[*3], Martin Potthast[*4], Min Zhang[*5], Benno Stein[6], and Sebastian Heineking[†7]

1 **Friedrich-Schiller-Universität Jena, DE.** matthias.hagen@uni-jena.de
2 **Toulouse University, FR.** josiane.mothe@irit.fr
3 **Barnard College, Columbia University – New York, US.** smara@columbia.edu
4 **Universität Kassel, hessian.AI, ScaDS.AI, DE.** martin.potthast@uni-kassel.de
5 **Tsinghua University – Beijing, CN.** z-m@tsinghua.edu.cn
6 **Bauhaus-Universität Weimar, DE.** benno.stein@uni-weimar.de
7 **Universität Leipzig, DE.** sebastian.heineking@uni-leipzig.de

Abstract

Dagstuhl Seminar 25391 "Retrieval-Augmented Generation – The Future of Search?" was held in the week of September 21–26, 2025. Thirty-nine researchers, most of whom came from the fields of information retrieval and web search as well as natural language processing, were invited to share the latest developments in the area of retrieval-augmented generation and discuss its research agenda and future directions. The 5-day program of the seminar consisted of four introductory and background sessions, two short talks sessions about technology and demos, one industry talk session, one afternoon hackathon, and nine working groups and reporting sessions. The seminar also had three social events during the program. This report provides the executive summary, overview of invited talks, and findings from the five working groups which cover the potential and limitations, information behavior and result presentation, the system side, societal and ethical aspects, and the evaluation of retrieval-augmented generation. The ideas and findings presented in this report should serve as one of the main sources for diverse research programs on retrieval-augmented generation.

Seminar September 21–26, 2025 – https://www.dagstuhl.de/25391

2012 ACM Subject Classification Information systems; Computing methodologies → Artificial intelligence

Keywords and phrases Retrieval-Augmented Generation, Information Retrieval, Dagstuhl Seminar

Digital Object Identifier 10.4230/DagRep.15.9.71

[*] Editor / Organizer
[†] Editorial Assistant / Collector

Retrieval-Augmented Generation – The Future of Search?, *Dagstuhl Reports*, Vol. 15, Issue 9, pp. 71–159
Editors: Matthias Hagen, Josiane Mothe, Smaranda Muresan, Martin Potthast, Min Zhang, Benno Stein, and Sebastian Heineking

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Matthias Hagen (Friedrich-Schiller-Universität Jena, DE)
Josiane Mothe (Toulouse University, FR)
Smaranda Muresan (Barnard College, Columbia University – New York, US)
Martin Potthast (Universität Kassel, hessian.AI, and ScaDS.AI – DE)
Min Zhang (Tsinghua University – Beiing, CN)
Benno Stein (Bauhaus-Universität Weimar, DE)

Background and Motivation

Retrieval-augmented generation (RAG) has proven effective in conditioning the output of large language models (LLMs) on relevant documents and for grounding LLM-generated statements, this way combatting the so-called hallucination or confabulation problem. Basically, RAG combines (1) a retrieval phase where a search system identifies relevant documents for a user prompt and (2) a generation phase, where an LLM synthesizes a tailored answer, probably linking to the retrieved sources.

RAG challenges "classical" retrieval technology and has the potential to revolutionize information-seeking behavior overall by reducing the searcher's effort to extract the desired information from individual search results. The revolution becomes evident, among others, in a change in the design of a search engine results page (SERP): Instead of presenting the proverbial list of "ten blue links", the list SERP, a generated text with references is shown, the text SERP. The first public prototypes of this kind were You.com's You Chat and the now discontinued Neeva AI, followed by Microsoft's Bing Chat, Google's Bard, Perplexity.ai, and Baidu's Ernie. However, many unsolved problems and relevant research questions still lurk under the hood (i.e., the user interface).

The proposed Dagstuhl Seminar focused on the expectations, the promises, the potential, and the limits of integrating RAG in information retrieval (IR). Relevant questions include:

- Will we ever search again?
- Will RAG bias retrieval results?
- Is RAG more than fact checking for conversational IR?
- How can we measure the effectiveness of RAG-based systems?
- How can we keep RAG-based systems transparent and accountable?

To work on these and related questions, the Dagstuhl Seminar brought together experts from the fields of information retrieval, natural language processing, and generative AI who have academic, non-profit (e.g., Open Search Foundation), or industrial backgrounds (e.g., Cohere).

Seminar Program

The 5-day program of the seminar consisted of five introductory and background sessions, one perspectives talk session, one industry talk session, and nine breakout discussion and reporting sessions. The program also had three social events and is available online.[1]

[1] https://www.dagstuhl.de/25391/schedule.pdf

Pre-Seminar Activities

Prior to the seminar, participants were asked to provide inputs to the following questions and request:

1. Will RAG replace ranked search for end users?
2. Please list, from the perspective of your research interests, important open questions or challenges in RAG.
3. What are the three papers a PhD student in RAG should read and why?

The first question has been answered by 30 out of the thirty-nine participants. The answers were almost evenly distributed between "No" with 16 votes and "Yes" with 13 votes. In addition to asking this question, we asked participants to give a reason for their vote. These answers turned out to be nuanced, including various forms of hedging one way of the other. A list of the arguments made is provided in Section 5 of this report.

From the survey, the following topics initially emerged as interests of participants. Many of these topics were discussed at length in the seminar.

- Can the methodologies that underpin RAG solve the enterprise search problems that sparse retrieval struggled to deal with?
- How can biases in the responses of RAG systems be detected?
- How to effectively inject external knowledge into LLMs?
- How to conduct RAG efficiently and dynamically?
- What is the analog of the Cranfield-style evaluation of ranked retrieval for RAG?
- What should the role and prominence of citations be in RAG outputs?
- How does RAG influence information behavior and how does it affect relevance feedback?
- How should search engines be designed that are used by RAG agents?
- What role does reasoning play in RAG?
- How can RAG systems be trained and how can they aid in training generative AIs?
- How can RAG systems notice and express uncertainty in their response?

Another outcome of the above pre-seminar questions was the compilation of a list of recommended reading to gain a solid understanding of topics and technologies related to retrieval-augmented generation. The reading list is provided in Section 6 of this report.

Invited Talks

One of the main goals and challenges of this seminar was to bring a broad range of researchers together to discuss retrieval-augmented generation, which required to establish common terminologies among participants. Therefore, we had a series of 26 invited talks throughout the seminar program to facilitate the understanding and discussion of retrieval-augmented generation and its potential enabling technologies. Section 3 contains the abstracts of all talks.

Working Groups

In the afternoon of Day 2, initial working groups were formed based on the inputs to the pre-seminar questionnaires, introductory, and background talks, and discussions among participants. Eventually, the following five groups were formed:

- Potential & Limitations with Respect to Cognitive Tasks
- Information Behavior and Result Presentation in RAG
- Retrieval-Augmented Generation: The System's Perspective
- Societal and Ethical Motivations for Inverting RAG to GAR
- An Unexamined RAG Is Not Worth Interrogating

We have summarized the working groups' outcomes in the following. Please refer to the main part of this report for the full description of the findings.

Potential & Limitations with Respect to Cognitive Tasks

The aim of this group was to provide an overview of the cognitive tasks in which a user can be supported by retrieval-augmented generation (RAG) systems. These cognitive tasks were organized in a taxonomy and linked to Bloom's taxonomy. With Bloom's taxonomy in mind, it became obvious that RAG is the superior technology in comparison to ranked search for information needs for which the solution requires synthesis (i.e., activates all layers in Bloom's taxonomy). However, if a user outsources these tasks to RAG systems frequently, there are risks of cognitive decline for the user since cognitively demanding synthesis tasks are rarely performed manually. These risks were outlined by this group and scaffolding strategies to mitigate them were proposed.

Information Behavior and Result Presentation in RAG

This group examined how RAG systems can cater to a diverse range of information behaviors and tailor the result presentation to the requirements of different users and tasks. They discussed how RAG systems allow for more natural interactions in multiple modalities and the formulation of complex queries that are contextualized by the rich history of previous interactions.

The group emphasized the need for adaptive, transparent, and inclusive RAG systems that accommodate diverse users and tasks – from factual retrieval to creative generation. They analyzed interaction dimensions such as user versus system initiative, information complexity, and human – machine collaboration. Their discussions resulted in open research questions around provenance tracking, adaptive presentation, and fostering user engagement while counteracting cognitive offloading.

Retrieval-Augmented Generation: The System's Perspective

This group took a systems-level view of RAG, examining its architecture and contrasting today's naive setup with an ideal RAG system. First, they highlighted that retrieval models were originally built for human users, informing evaluation methods and result formats that may not be suited for LLMs as users. Second, they emphasized the many sources of uncertainty in RAG, such as retrieval, model reasoning, and data quality, and the need for better ways to detect and communicate this uncertainty. Third, they explored efficiency – effectiveness trade-offs, arguing that future systems should dynamically allocate computation based on query complexity. Finally, the group discussed the advantages of federated RAG systems, such as unified access to open and proprietary sources, as well as challenges that this kind of architecture would face.

Societal and Ethical Motivations for Inverting RAG to GAR

This group advocated for a shift from retrieval-augmented generation (RAG) to Generation-Augmented Retrieval (GAR) to develop next-generation information access tools. This framing prioritizes information retrieval and the design of transparent, ethical, and sustainable information access systems that encourage active user engagement with information sources and diverse knowledge ecosystems. First, the group discussed the intersection of knowledge, ethics, and human rights, and its implications for RAG vs. GAR. Next, they offered

perspectives on several (though not exhaustive) sociotechnical issues for GAR: scholarly communication, user cognition, emotional and mental well-being, democracy and political discourse, and language and culture. Each section references existing challenges related to the rapid and widespread use of generative AI to encourage critical and informed thinking about GAR development for individuals, organizations, and society as a whole. Throughout, the group proposed considerations for information retrieval researchers to create information access systems, learn from user studies, and foster interdisciplinary partnerships.

An Unexamined RAG Is Not Worth Interrogating

This group focused on how one can determine in which scenarios RAG systems work, where they fail, and how one can identify which RAG system is suitable for which scenario. The motivation of the group was that paradigm shifts in information access technology might also require paradigm shifts for corresponding evaluations. It therefore started by hypothesizing which evaluation properties might change between mature information retrieval evaluation methods and the methods applied in RAG systems. The group recognized that RAG evaluation is currently in an exploratory stage where many different evaluation ideas are being explored. The group identified research gaps and proposed directions for future research. Many parts of the discussion evolved around the idea of enabling in-depth analysis of RAG responses by multiple experts. This activity motivated the group to brainstorm how concepts from the Talmud can be transferred to RAG evaluations and labeling.

Conclusions

Leading researchers from diverse domains in academia and industry investigated the essence, attributes, architecture, applications, challenges, and opportunities of retrieval-augmented generation in the seminar. One clear signal from the seminar is that research opportunities to advance retrieval-augmented generation are available to many areas and collaboration in an interdisciplinary community is essential to achieve this goal. This report should serve as one of the main sources to facilitate such diverse research programs on retrieval-augmented generation.

2 Table of Contents

Executive Summary

Matthias Hagen, Josiane Mothe, Smaranda Muresan, Martin Potthast, Min Zhang, and Benno Stein . . . 72

Overview of Talks

Legal Retrieval and Augmented Generation
Qingyao Ai . . . 78

Parametric Retrieval-Augmented Generation
Qingyao Ai . . . 78

Interactions with RAG
Mohammad Aliannejadi . . . 79

An Industry Perspective on RAG
Sophia Althammer . . . 79

RAGE: How to Evaluate RAG Systems
Laura Dietz . . . 79

RAG 4 Med
Carsten Eickhoff . . . 80

Core IR Concepts and RAG
Norbert Fuhr . . . 80

MetaRAG: Learning About RAG from a RAG System
Marcel Gohsen . . . 80

OpenWebSearch.eu: An Open Scaleable Infrastructure for Web Search & RAG
Michael Granitzer . . . 81

Uncertainty Quantification for RAG
Faegheh Hasibi . . . 81

Ads in RAG: Can We Block Promotional Text in LLM Responses?
Sebastian Heineking . . . 82

Two Things You Must Know About the G from RAG
Djoerd Hiemstra . . . 82

Rankify Toolkit for Retrieval, Re-Ranking, RAG & RankArena
Adam Jatowt . . . 83

Hint-Based Interaction
Adam Jatowt . . . 83

User Simulation for Generative IR Systems: GenIRSim
Johannes Kiesel . . . 84

The Turing Game
Johannes Kiesel and Benno Stein . . . 84

Neural and LLM Retrieval
Sean MacAvaney . . . 85

Sociotechnical Implications of RAG for Information Access
Bhaskar Mitra . . . 85

Beyond English: Cultural and Linguistic Challenges in RAG Systems
Josiane Mothe . . . 85

The Jewish Talmud as the Past (or Beginning?) of RAG
Birte Platow . . . 85

RAG Evaluation
Mark Sanderson . . . 86

Multimodal LLMs and RAG
Alan Smeaton . . . 86

Adapting RAG to Users
Damiano Spina and Johanne Trippas . . . 86

RAG Foundations and Models
Arjen P. de Vries . . . 87

Advances in LLM Rankers
Guido Zuccon . . . 87

Do LLMs Search Differently?
Guido Zuccon . . . 87

Working Groups

Potential & Limitations with Respect to Cognitive Tasks
Liesbeth Allein, Sophia Althammer, Nolwenn Bernard, Marcel Gohsen, Adam Jatowt, Abhinav Joshi, Smaranda Muresan, Jian-Yun Nie, and Benno Stein . . . 88

Information Behavior and Result Presentation in RAG
Mohammad Aliannejadi, Gianluca Demartini, Carsten Eickhoff, Norbert Fuhr, Sebastian Heineking, Martin Potthast, Harrisen Scells, Damiano Spina, and Johanne Trippas . . . 97

Retrieval-Augmented Generation: The System's Perspective
Djoerd Hiemstra, Sean MacAvaney, Qinqyao Ai, Michael Granitzer, Guido Zuccon, Faegheh Hasibi, Avishek Anand, and Arjen P. de Vries . . . 114

Societal and Ethical Motivations for Inverting RAG to GAR
Johannes Kiesel, Bhaskar Mitra, Josiane Mothe, Heather O'Brien, Birte Platow, and Stefan Voigt . . . 124

An Unexamined RAG Is Not Worth Interrogating
Niklas Deckers, Laura Dietz, Maik Fröbe, Wojciech Kusa, and Mark Sanderson . . . 135

Answers to "Will RAG replace ranked search for end users?" . . . 141

Recommended Reading List . . . 143

Acknowledgments . . . 147

Participants . . . 159

3 Overview of Talks

3.1 Legal Retrieval and Augmented Generation

Qingyao Ai (Tsinghua University – Beijing, CN)

The rapid progress of large language models (LLMs) has opened new opportunities for legal artificial intelligence, yet the legal domain presents unique challenges that require specialized solutions. This talk explores legal retrieval-augmented generation (Legal RAG) as a framework to address these challenges, focusing on data, augmentation, evaluation, and applications. Legal case documents are structurally complex and lengthy, with relevance definitions that differ fundamentally from general domains. To manage extreme input lengths and heterogeneous sources – such as statutes, case documents, and legal essays – RAG methods should dynamically retrieve and integrate multi-source knowledge for reasoning tasks. We highlight applications across legal judgment prediction, legal document writing, and court simulation. Evaluation remains difficult as it requires significant legal expertise and is highly subjective in many cases. We introduce our initial efforts on building taxonomies and benchmarks for legal RAG and LLMs, and hope this would help people develop better legal models in future.

3.2 Parametric Retrieval-Augmented Generation

Qingyao Ai (Tsinghua University – Beijing, CN)

Retrieval-augmented generation (RAG) has emerged as a promising solution to enhance the reliability of large language models (LLMs) with external knowledge. Existing RAG methods share a common strategy for knowledge injection: they place the retrieved documents into the input context of the LLM, which we refer to as the in-context knowledge injection method. While this approach is simple and often effective, it has inherent limitations. Firstly, increasing the context length and number of relevant documents can lead to higher computational overhead and degraded performance, especially in complex reasoning tasks. More importantly, in-context knowledge injection operates primarily at the input level, but LLMs store their internal knowledge in their parameters. This gap fundamentally limits the capacity of in-context methods. To this end, we introduce Parametric RAG, a new RAG paradigm that integrates external knowledge directly into the feed-forward networks of an LLM through document parameterization. This approach not only reduces online computational costs by shortening the input context length, but also deepens the integration of external knowledge by enabling LLMs to utilize it in the same way as internal parametric knowledge. Experimental results demonstrate that Parametric RAG substantially enhances the effectiveness and efficiency of knowledge augmentation in LLMs. Also, it can be combined with in-context RAG methods to achieve even better performance.

3.3 Interactions with RAG

Mohammad Aliannejadi (University of Amsterdam, NL)

In this talk, I present the current challenges of generative and RAG systems in terms of user interaction. How do the existing user interaction and information need levels generalize to RAG systems and what are the implications of the new chat-based interfaces on user experience? In particular, how do the changes affect the click behavior of users and what are the risks of that in terms of trust and factuality of the results?

3.4 An Industry Perspective on RAG

Sophia Althammer (Cohere – München, DE)

This talk gives an overview of products with RAG (retrieval-augmented generation) and Agents as well as outlining important aspects for enterprise customers. We define agents and give an overview how the Command-A model was trained with respect to agentic capabilities. We also touch on evaluation benchmarks and general progress of agents and outline possible open research questions on agents.

3.5 RAGE: How to Evaluate RAG Systems

Laura Dietz (University of New Hampshire – Durham, US)

Retrieval-augmented generation (RAG) has become a central paradigm for knowledge-intensive applications, yet its evaluation remains a persistent challenge. Traditional relevance-based evaluation, grounded in human judgments over static documents, does not translate directly to contexts where each query produces a novel, free-form response. In response, researchers have increasingly employed large language models (LLMs) as judges, using methods such as direct prompting, pairwise preference comparison, multi-criteria prompting, nugget-based evaluation, and multi-step frameworks. These approaches offer scalability and fine-grained assessment, but they also introduce new risks.

A prime threat is circularity, in which systems and evaluations rely on the same or overlapping models, producing results that appear plausible yet fail to align with human judgments. The talk demonstrates that direct prompting methods are particularly vulnerable to this effect, undermining their reliability under meta-evaluation. Nugget-based LLM judges, by contrast, promise more interpretable and granular assessments, but their resilience is not guaranteed: if evaluation "secrets" such as gold nuggets or rubric structures can be anticipated or guessed by systems, then even nugget-based evaluation may be compromised.

This talk presents an investigation into often overlooked factors that negatively impact evaluation methods for RAG systems and the meta-evaluation of those methods. It argues that evaluation is no longer a passive measure of progress but an active force shaping system design, with feedback loops that must be understood to avoid distorted conclusions.

3.6 RAG 4 Med

Carsten Eickhoff (Universität Tübingen, DE)

The modern healthcare system and its various actors face unprecedented staffing shortages and tightening economic constraints. Data driven methods, including generative AI applications offer a promising means of automating, and enhancing many steps in the clinical pipeline. In this talk, we will briefly discuss the opportunities and challenges that these applications face and which role retrieval augmented generation can play in this high-stakes environment.

3.7 Core IR Concepts and RAG

Norbert Fuhr (Universität Duisburg-Essen, DE)

Information retrieval is about vagueness, uncertainty and context in information access. Vagueness is caused by the fact that users cannot give a precise specification of their information need, thus using vague query conditions and iterative query formulation. While the latter is addressed by conversational IR, LLM-based systems usually choose the most probable meaning for vague query terms. Uncertainty is due to the system's uncertain knowledge about the user's information need and the database objects. Traditional IR system use ranking for dealing with uncertainty, but more advanced strategies in RAG are an open issue. Consideration of context requires an IR system to take into account both the user and the situation when answering a query. Currently, RAG users enrich their queries with lengthy descriptions of the context. Overall, future RAG systems should explicitly address vagueness in queries, clarify answer uncertainty and integrate means for capturing user context automatically.

3.8 MetaRAG: Learning About RAG from a RAG System

Marcel Gohsen (Bauhaus-Universität Weimar, DE)

Exploring the landscape of papers about retrieval-augmented generation (RAG) can be a challenge due to the sheer volume of publications. Specifically for the Dagstuhl Seminar on RAG, we developed "MetaRAG", a conversational search system that uses RAG to facilitate navigation through the literature. MetaRAG can answer open-ended question or summarize a specific publication, retrieving from a collection of over 3,000 papers.

3.9 OpenWebSearch.eu: An Open Scaleable Infrastructure for Web Search & RAG

Michael Granitzer (Universität Passau, DE)

OpenWebSearch.eu project aims to develop an open European infrastructure for web search. The project aspires to contribute to Europe's digital sovereignty and help promote an open human-centred search engine market.

OpenWebSearch.eu is designing the core of the European Open Web Index, based on open source software and deployed across various high performance computing centres in Europe. The Open Web Index is particularly crucial for the provision of state-of-the-art web search services and for European innovations, such as AI/large language models.

We build an infrastructure to establish and maintain relevant web-data sets at Petabyte-scale, including raw web-data, web-index and large multilingual text corpora as well as multimodal data sets at the various EuroHPC centres: ready-to-use by researchers, start-ups, innovators and industry to build value oriented, trustworthy AI/LLM and search solutions in and for Europe. We are confident that sharing these data sets across Europe by hosting them very close to or directly at the HPC facilities would be the most effective way.

An Open Web Index will also provide new opportunities in developing decentralised RAG systems by providing relevant web-data for bootstrapping topic specific search engines.

3.10 Uncertainty Quantification for RAG

Faegheh Hasibi (Radboud University Nijmegen, NL)

Uncertainty Quantification (UQ) provides methods to estimate an LLM's confidence in its outputs and helps users assess the reliability of its responses. Ideally, an effective UE method would assign low uncertainty to correct answers and high uncertainty to incorrect ones. However, existing UE methods have not been thoroughly studied in the context of retrieval-augmented generation (RAG). Recent work shows that current approaches often fail to reliably estimate response correctness in the simple retrieve-then-generate paradigm of RAG. In this talk, we will explore the applications of UE in information access systems and discuss the challenges of uncertainty estimation in more complex RAG settings, where generation unfolds through multiple reasoning and retrieval steps.

3.11 Ads in RAG: Can We Block Promotional Text in LLM Responses?

Sebastian Heineking (Universität Leipzig, DE)

Joint work of Sebastian Schmidt, Ines Zelch, Janek Bevendorff, Benno Stein, Matthias Hagen, Martin Potthast
Main reference Sebastian Schmidt, Ines Zelch, Janek Bevendorff, Benno Stein, Matthias Hagen, Martin Potthast: "Detecting Generated Native Ads in Conversational Search", in Proc. of the Companion ACM on Web Conference 2024, WWW 2024, Singapore, Singapore, May 13-17, 2024, pp. 722–725, ACM, 2024.
URL https://doi.org//10.1145/3589335.3651489

Due to its prevalence in classic search engines, advertising is a likely business model for retrieval-augmented generation (RAG). As one example, researchers from Google outlined an auction mechanism for individual tokens that illustrates how LLM-based advertising could be implemented [1]. In contrast to conventional advertising, LLMs can tailor advertisements to the user's preferences and current information need, and blend them with the rest of the generated text, making them difficult to detect. This talk gives an overview on our ongoing research on blocking advertisements in LLM responses [2, 3].

References

1 Paul Dütting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. Mechanism Design for Large Language Models. In *Proceedings of the ACM Web Conference 2024*, pages 144–155, Singapore Singapore, May 2024. ACM.

2 Sebastian Schmidt, Ines Zelch, Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. Detecting Generated Native Ads in Conversational Search. In *Companion Proceedings of the ACM Web Conference 2024*, pages 722–725, Singapore Singapore, May 2024. ACM.

3 Johannes Kiesel, Çağrı Çöltekin, Marcel Gohsen, Sebastian Heineking, Maximilian Heinrich, Maik Fröbe, Tim Hagen, Mohammad Aliannejadi, Tomaž Erjavec, Matthias Hagen, Matyáš Kopp, Nikola Ljubešić, Katja Meden, Nailia Mirzakhmedova, Vaidas Morkevičius, Harrisen Scells, Ines Zelch, Martin Potthast, and Benno Stein. Overview of Touché 2025: Argumentation Systems: Extended Abstract. In Claudia Hauff, Craig Macdonald, Dietmar Jannach, Gabriella Kazai, Franco Maria Nardini, Fabio Pinelli, Fabrizio Silvestri, and Nicola Tonellotto, editors, *Advances in Information Retrieval*, volume 15576, pages 459–466. Springer Nature Switzerland, Cham, 2025. Series Title: Lecture Notes in Computer Science.

3.12 Two Things You Must Know About the G from RAG

Djoerd Hiemstra (Radboud University Nijmegen, NL)

In this short presentation I will discuss: 1) why chatbots and retrieval-augmented generation (RAG) are so irresistible to people, and 2) the energy consumption of text generation approaches based on large language models. I will argue that we MUST NOT assign human traits to chatbots and RAG system in our research papers, and that we MUST estimate the energy consumption of our research.

3.13 Rankify Toolkit for Retrieval, Re-Ranking, RAG & RankArena

Adam Jatowt (Universität Innsbruck, AT)

In this talk I have introduced Rankify toolkit [1] and RankArena comparison and demonstration framework [2] for fostering research on RAG and reranking algorithms.

Rankify is a modular open-source toolkit designed to unify retrieval, re-ranking, and RAG within a cohesive framework. Rankify supports a wide range of retrieval techniques, including dense and sparse retrievers, while incorporating state-of-the-art reranking models to enhance retrieval quality. Additionally, Rankify includes a collection of pre-retrieved datasets to facilitate benchmarking, available at Huggingface. As a unified and lightweight framework, Rankify allows researchers and practitioners to advance retrieval and reranking methodologies while ensuring consistency, scalability, and ease of use. Rankify is available at `https://github.com/DataScienceUIBK/rankify`.

RankArena is a unified platform for comparing and analysing the performance of retrieval pipelines, rerankers, and RAG systems using structured human and LLM-based feedback as well as for collecting such feedback. RankArena supports multiple evaluation modes: direct reranking visualisation, blind pairwise comparisons with human or LLM voting, supervised manual document annotation, and end-to-end RAG answer quality assessment. It captures fine-grained relevance feedback through both pairwise preferences and full-list annotations, along with auxiliary metadata such as movement metrics, annotation time, and quality ratings. The platform also integrates LLM-as-a-judge evaluation, enabling comparison between model-generated rankings and human ground truth annotations. All interactions are stored as structured evaluation datasets that can be used to train rerankers, reward models, judgment agents, or retrieval strategy selectors. Our platform is publicly available at `https://rankarena.ngrok.io/`, and the Demo video is provided at `https://youtu.be/jIYAP4PaSSI`.

References

1 Abdelrahman Abdallah, Bhawna Piryani, Jamshid Mozafari, Mohammed Ali, and Adam Jatowt. Rankify: A comprehensive python toolkit for retrieval, re-ranking, and retrieval-augmented generation. *arXiv preprint arXiv:2502.02464*, 2025.

2 Abdelrahman Abdallah, Mahmoud Abdalla, Bhawna Piryani, Jamshid Mozafari, Mohammed Ali, and Adam Jatowt. Rankarena: A unified platform for evaluating retrieval, reranking and rag with human and llm feedback, 2025.

3.14 Hint-Based Interaction

Adam Jatowt (Universität Innsbruck, AT)

Chatbots have rapidly become integrated into everyday life, offering instant, human-like answers to a wide range of questions. While this technology improves information access, it can also weaken cognitive skills by encouraging shallow processing, contributing to information overload, and fostering over-reliance on automated reasoning. Two features of hint-based interaction make it a promising alternative. First, hints act as scaffolds, guiding users

toward answers without revealing them outright. Second, hints are adaptable, meaning they can be tailored to different purposes, question types, and user needs. We discuss the idea of automatic hint generation, the criteria of hint quality estimation and we demonstrate several datasets of hints, as well as introduce the HintEval framework (`https://github.com/DataScienceUIBK/HintEval`) designed for supporting hint-focused research.

3.15 User Simulation for Generative IR Systems: GenIRSim

Johannes Kiesel (GESIS – Leibniz Institute for the Social Sciences – Köln, DE)

Joint work of Johannes Kiesel, Marcel Gohsen, Nailia Mirzakhmedova, Matthias Hagen, Benno Stein
Main reference Johannes Kiesel, Marcel Gohsen, Nailia Mirzakhmedova, Matthias Hagen, Benno Stein: "Who Will Evaluate the Evaluators? Exploring the Gen-IR User Simulation Space", in Proc. of the Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9-12, 2024, Proceedings, Part I, Lecture Notes in Computer Science, Vol. 14958, pp. 166–171, Springer, 2024.
URL https://doi.org//10.1007/978-3-031-71736-9_11

The reliable and repeatable evaluation of interactive, conversational, or generative IR systems is an ongoing research topic in the field of retrieval evaluation. One proposed solution is to fully automate evaluation through simulated user behavior and automated relevance judgments. Still, simulation frameworks were technically quite complex and have not been widely adopted. Recently, however, easy access to large language models has drastically lowered the hurdles for both user behavior simulation and automated judgments. We therefore argue that it is high time to investigate how simulation-based evaluation setups should be evaluated themselves. We present GenIRSim, a flexible and easy-to-use simulation and evaluation framework for generative IR.

3.16 The Turing Game

Johannes Kiesel (GESIS – Leibniz Institute for the Social Sciences – Köln, DE)
Benno Stein (Bauhaus-Universität Weimar, DE)

Developments in generative AI (ChatGPT, Stable Diffusion, etc.) are redefining the boundaries between humans and machines. However, current debates on this topic primarily revolve around questions of performance: Do machines write better than humans? Are their images more appealing or even more creative than human art works? We argue that the focus on performance overlooks one of the most pressing questions of current development, namely: Do humans accept "intelligent machines" as members of their community?

To answer this question, we are developing the "Turing Game", based on the classic Turing Test [1]. While the Turing Test was conceived as an "imitation game" and made a ground-breaking contribution to the question of whether machines can think, the Turing Game serves to ask how hybrid communities with human and machine members can function.

References

1 Alan M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

3.17 Neural and LLM Retrieval

Sean MacAvaney (University of Glasgow, GB)

How do you go from a strong LLM-based relevance model to a complete search engine? This talk gives a brief introduction to the families of retrieval techniques for LLMs.

3.18 Sociotechnical Implications of RAG for Information Access

Bhaskar Mitra (Independent Researcher, Tiohtià:ke/Montréal, CA)

Robust access to trustworthy information is a critical need for society with implications for knowledge production, public health education, and promoting informed citizenry in democratic societies. Generative AI technologies with retrieval-augmentation may enable new ways to access information and improve effectiveness of existing information retrieval systems, but we are only starting to understand and grapple with their long-term social implications. In this talk, we discuss some of the systemic risks of employing generative AI and RAG in the context of information access that should critically inform future research and development in this area.

3.19 Beyond English: Cultural and Linguistic Challenges in RAG Systems

Josiane Mothe (Toulouse University, FR)

This talk probes how retrieval-augmented generation (RAG) handles culture and language. A simple demo – "help me prepare a report on turkey" – shows smart disambiguation but also unexpected code-switching and clumsy French, prompting questions about language choice on the different steps of RAG and what sources get prioritized. Source selection -which is key- shifts with language and locale. Truth and facts are often filtered by context. While RAG offers new opportunities it also come with threats that are explored.

3.20 The Jewish Talmud as the Past (or Beginning?) of RAG

Birte Platow (TU Dresden, ScaDS.AI, DE)

From a humanities perspective, RAG systems raise fundamental questions about (new) knowledge systems: What is the origin and authority of a text? How do we deal with discursive/contradictory statements? To what extent and why do we want to recognize it as

"true"? And which interpretations do we want to establish? We have been familiar with these and other questions relating to the handling of text for thousands of years, especially when it comes to "sacred texts." The Jewish Talmud deals with this in depth and can perhaps be interpreted as an ancient precursor to RAG, posing challenging questions for AI-generated texts and knowledge systems.

3.21 RAG Evaluation

Mark Sanderson (RMIT University – Melbourne, AU)

This talk provides an overview of the two main areas of new research in the evaluation of RAG systems. For the first area, I will describe the use of LLMs to change the evaluation of the retrieval component of a RAG system. Here, I talk about the widely discussed use of LLMs for relevance assessment and the growing interest of using LLMs to simulate other critical aspects of retrieval evaluation, such as queries and document collections. For the second area, I will discuss the evaluation of RAG itself, in particular, I will highlight how RAG evaluation challenges the IR community to tackle aspects of evaluation it has gotten away with ignoring in the past.

3.22 Multimodal LLMs and RAG

Alan Smeaton (Dublin City University, IE)

This talk examines the differences between RAG as used in text-based LLMs, and RAG as it is (not yet) available in mutimodal LLMs. It concludes by demonstrating that text-based RAG has a large advantage over MM-RAG because of the strong heritage and experience of developing multiple IR techniques over recent decades whereas multimedia IR is in it's comparative infancy but at least now it has a roadmap for progress.

3.23 Adapting RAG to Users

Damiano Spina (RMIT University – Melbourne, AU)
Johanne Trippas (RMIT University – Melbourne, AU)

Retrieval-augmented generation (RAG) has proven effective in grounding the outputs of large language models (LLMs) with evidence from retrieved passages, thereby reducing errors or "hallucinations" in responses. Beyond its technical promise, RAG represents a new information-seeking paradigm that reshapes how users interact with systems to satisfy their information needs. In this talk, we argue that RAG – and, more broadly, generative information retrieval (GenIR) – creates opportunities to revisit long-standing concepts and

methodologies in information retrieval. We explore how this shift not only introduces new types of information needs and tasks, but also new kinds of "users," including LLMs themselves acting as searchers. This makes it a particularly exciting time to be working on interactive IR and evaluation, as the field expands to address these emerging challenges and opportunities.

3.24 RAG Foundations and Models

Arjen P. de Vries (Radboud University Nijmegen, NL)

Joint work of Weihang Su, Qingyao Ai, Jingtao Zhao, Qian Dong, Yiqun Liu, Arjen P. de Vries

This talk gives an introduction to retrieval-augmented generation (RAG) based on the tutorial "Dynamic and Parametric Retrieval-Augmented Generation" that Weihang Su, Qingyao Ai, Jingtao Zhan, Qian Dong and Yiqun Liu gave at SIGIR 2025.

3.25 Advances in LLM Rankers

Guido Zuccon (University of Queensland – Brisbane, AU)

Generative large language models (LLMs) like Gemini, GPT, and Llama are transforming information retrieval, enabling new and more effective approaches to document retrieval and ranking. The switch from the previous generation pre-trained language models backbones (e.g., BERT, T5) to the new generative LLMs backbones has required the field to adapt training processes; it also has provided unprecedented capabilities and opportunities, stimulating research into zero-shot approaches, reasoning approaches, reinforcement learning based training, and multilingual and multimodal applications. In this talk I provide an overview of LLM-based rankers, covering fundamental architectures and open challenges and research directions.

3.26 Do LLMs Search Differently?

Guido Zuccon (University of Queensland – Brisbane, AU)

Search has traditionally been defined as the interaction between a user and an information retrieval (IR) system to satisfy an information need. From the outset, IR evaluation has assumed that system utility lies in serving human users, and that this utility can be approximated by measuring relevance. Consequently, retrieval models and evaluation measures have been designed around human-centred notions of relevance, and search systems have been optimised accordingly. These same search systems are now central to generative AI, particularly in retrieval-augmented generation (RAG), where a retriever supplements

the input of a large language model (generator) with external documents. RAG has enabled state-of-the-art performance on knowledge-intensive tasks, improved attribution of answers to sources, and allowed models to in-corporate new knowledge without retraining. Yet, retrieval in RAG still relies on methods optimised for humans rather than for generators. Human-oriented practices such as ranking documents by semantic relevance reflect assumptions about how people search: that they read results top to bottom, expect the most relevant item first, and stop once their need is satisfied. These assumptions are embedded in evaluation measures such as normalised discounted cumulative gain (nDCG), which reward systems for placing relevant documents higher in the list. Yet empirical studies reveal that such ranked lists can impair generators. A model may prefer relevant documents at the end of the list, or even perform better when non-relevant documents appear first. These findings raise a deeper question: what principles should guide the design of retrievers for generators?

4 Working Groups

4.1 Potential & Limitations with Respect to Cognitive Tasks

Liesbeth Allein (KU Leuven, BE)
Sophia Althammer (Cohere – München, DE)
Nolwenn Bernard (TH Köln, DE)
Marcel Gohsen (Bauhaus-Universität Weimar, DE)
Adam Jatowt (Universität Innsbruck, AT)
Abhinav Joshi (Indian Institute of Technology Kanpur, IN)
Smaranda Muresan (Barnard College, Columbia University – New York, US)
Jian-Yun Nie (University of Montréal, CA)
Benno Stein (Bauhaus-Universität Weimar, DE)

The working group on cognitive tasks and RAG focused on benefits, potentials as well as limitations and challenges of RAG systems with respect to user cognition and cognitive tasks that users complete. The section first compares the different tasks that rank-based retrieval and RAG systems help to solve based on Bloom's taxonomy of learning. It then discusses a taxonomy of different tasks that can be completed with RAG and that are based on synthesizing information to create a solution object. In the remaining part of the section, we discuss the concepts of cognitive offloading and the risks of cognitive decline, as well as propose several future research directions.

4.1.1 Introduction and Motivation

Retrieval-augmented generation (RAG) systems have shown great potential for assisting and potentially automating a manifold of cognitive tasks for humans. Products powered with RAG systems like Google with generative AI summaries [62], ChatGPT Agent [121], ChatGPT DeepResearch [122] or Claude Code [9] are already changing how people interact with the web, how they solve implementation problems, or how they plan trips [120, 69].

RAG systems augment traditional ad-hoc retrieval systems by giving a large language model (LLM) access to different retrieval tools. In order to fulfill a user's request, the model can decide which and how to call the retrieval tools and the system can give back a generated response to the user or do actions directly with the tools given to the model.

Users of RAG systems originate from the whole population and cover children, adults, and elderly. Those different groups all come with different cognitive abilities and use RAG systems with different tasks and goals in mind. Each of these tasks cause different levels of cognitive demand for the user. An example of a task with higher cognitive demand is acquiring a new skill as it goes beyond one-shot information access and involves different aspects of information and practice. A task with lower cognitive demand can be as simple as retrieval of a cooking recipe.

The cognitive demand for the user is interdependent on the demand a human would be exposed to if he would replace the LLM in a RAG system. For example, retrieving (or synthesizing) a cooking recipe causes little cognitive load for a user since he only needs to understand and remember details of the recipe, while a human in the shoes of a RAG system would need to find and judge relevant recipes and extract important information which would cause a high cognitive demand. Vice versa, acquiring a new skill cause high cognitive demand for the user, however, providing teaching materials to the user can be seen as a task of lower cognitive demand. In chapter 4.1.2, we discuss in more detail the cognitive load on the users depending on the different tasks.

Whether it is more effective to use a traditional ad-hoc retrieval system compared to a RAG system depends on the solution object for the task. We refer to a solution object (also known as solution path) as a path in a graph representing the search space with which a user wants to satisfy his information need. Specifically, a solution object can be a piece of text, a document, an image or any kind of media that satisfies a user's information need.

- If a solution object already exists for a task, such as a paragraph of a web page, a PDF document, or image, rank-based retrieval with traditional ad-hoc retrieval systems is more effective for solving a task.
- Vice versa, if the solution object does not exist, but needs to be created through the process of retrieving several information items and synthesizing the solution object out of it, then it is convenient for the user to use a RAG system. Then, the user is not exposed to the cognitive load of creating the solution object, but instead needs to specify only the request and, potentially iterate through feedback with the RAG system.

In the following sections, we resort to Bloom's (revised) taxonomy [88] to compare the cognitive tasks a traditional ad-hoc retrieval system and a RAG system takes responsibility for. Subsequently, we develop a taxonomy of cognitive tasks that potentially fall within the capabilities of RAG systems. Along Bloom's (revised) taxonomy we then discuss the possible limitations and negative effects of RAG systems on the user.

4.1.2 Bloom's Revised Taxonomy

Bloom's revised taxonomy[2], see the colored pyramid in Figure 1, provides a concrete medium to navigate through the set of tasks that a user would like to accomplish. Bloom's Taxonomy is a framework that organizes learning objectives into a hierarchy of cognitive levels, from basic remembering to complex creation. While with RAG systems tasks of all cognitive levels in Bloom's taxonomy can be addressed, rank-based retrieval addresses tasks of the two lowest levels only. This division of the pyramid also reflects our distinction between tasks for

[2] The main differences between Bloom's Taxonomy and the revised version [88] are terminology, structure, and emphasis. Bloom's Revised Taxonomy replaces the original nouns (Knowledge, Comprehension, Application, Analysis, Synthesis, Evaluation) with verbs (Remember, Understand, Apply, Analyze, Evaluate, Create), shifts the top level to "Create", and adds a second dimension of "Knowledge" (Factual, Conceptual, Procedural, Metacognitive) this way creating a two-dimensional framework.

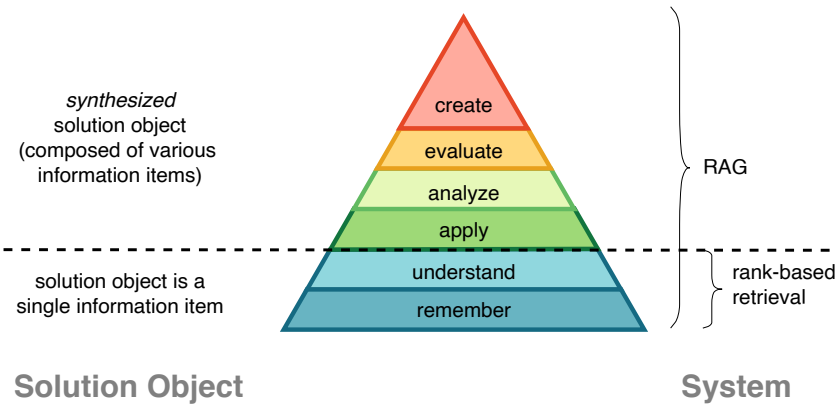

Figure 1 Bloom's revised taxonomy of learning [88] can be used to demonstrate and discuss the complexity of tasks that can be tackled by a RAG system and a rank-based retrieval system, respectively. While RAG systems are capable of synthesizing complex solution objects to tackle cognitive demanding tasks, traditional rank-based systems focus on searching and finding single information items only, so the solution object already needs to exist.

which the solution object consists of a a single information item that already exists (the two bottom levels) and for which the solution object must be synthesized (the top four levels) and tasks. In this analogy, a rank-based retrieval system has to remember the documents (search index), and understand the user's information need to be effective. On top of that, a RAG system has to apply domain-knowledge, analyze the retrieved documents, evaluate individual aspects, and create a final response to present to the user to fulfill the information need.

In the following, we present a brief overview of Bloom's taxonomy from the user's perspective. The taxonomy also provides a proxy of the required level of cognition for a task, going from lowest (*remember*) to highest (*create*).

- **Remember**: Users recall or recognize knowledge from their own memory. They rely on their memory to produce or retrieve facts, definitions, lists, or recite previously acquired information.
- **Understand**: Users construct meaning from different types of functions. These functions include written or graphic messages. Understanding involves tasks like interpreting, inferring, classifying, summarizing, explaining, or exemplifying.
- **Apply**: Users use, implement, or execute a procedure. Applying involves situations in which users rely on learned materials and information through products like models, interviews, presentations, or simulations.
- **Analyze**: Users decompose materials and concepts into components and identify (inter)relations between these components and relations to an overarching structure or purpose. Analyzing involves differentiation, organization, attribution, and the ability to distinguish between components. Analysis products include spreadsheets, surveys, diagrams, charts, and graphic representations.
- **Evaluate**: Evaluation remains an integral part of the pipeline where users make judgments based on the analysis/observations and their prior knowledge about the task.
- **Create**: For a large set of tasks, a user's final goal is to come up with a coherent/functional form of new patterns or structure, given the previous set of observations, where the primary aim remains to synthesize a new solution object. This leads to the most difficult features that a user would want a system to augment.

We consider the above-mentioned levels from the RAG users' perspective and formulate a task taxonomy for RAG Systems. Further, we also explore the challenges and limitations of RAG-based systems, which may play a crucial role in affecting users' cognitive skills at different levels in Section 4.1.4.

4.1.3 Task Taxonomy

A central aspect when thinking about cognitive tasks that users want to accomplish is to distinguish between tasks for which RAG is the preferable system over rank-based search. We identified that RAG is superior for all tasks that require synthesis of a solution object, since no solution object is preexisting. In Figure 2, we provide an extensive taxonomy of all tasks potentially solvable with RAG that require synthesis from the perspective of a user.

The task taxonomy was created in a bottom-up fashion, by first collecting all thinkable cognitive tasks, that could be solved with a RAG system, and then they were clustered, classified, and organized in the taxonomy. In the following, we outline the tasks categories from the taxonomy in more detail.

1. **Information Access**
   a. **Knowledge Retrieval & Organization**
      i. Question answering for complex information needs
      ii. Summarization of retrieved information items
      iii. Clustering of retrieved information items
      iv. Comparison of retrieved information items
   b. **Knowledge Acquisition**
      i. Interactive learning of new knowledge related to concepts, processes, items
      ii. Feedback and reflection
      iii. Explanation of specific concepts, facts, or events
   c. **Content Recommendation**
2. **Planning & Solution Implementation**
   a. Problem-solving (e.g., mathematics, logics, coding)
   b. Procedural instruction retrieval
   c. Planning & scheduling
3. **Content Generation & Creative Tasks**
   a. Content creation (e.g., stories, poems, images, documentation)
   b. Hypothesis generation
   c. Generation of communication items

Figure 2 Taxonomy of tasks that is preferable to be solved with RAG from the perspective of a user. All tasks in this taxonomy require the system to synthesize of a solution object.

4.1.3.1 Knowledge Retrieval & Organization

The primary objective of knowledge retrieval tasks is the short-term information access with well defined information needs. We define short-term tasks as tasks that can be accomplished in a single session involving only a few turns in a conversation. The central task of knowledge retrieval is open-ended question answering. With a RAG system, the system itself can fulfill tasks with more complex information needs, whereas ranked search systems would leave the cognitive load of multiple searches, comparing and synthesizing information to the user. Another core task that users expect to be satisfiable solved with RAG is summarization. This

task involves either single or multi-document summarization and typical instances include key point extraction and overview preparation as well as comparison, classification, and clustering of different concepts.

4.1.3.2 Knowledge Acquisition

In contrast to knowledge retrieval tasks, knowledge acquisition tasks describe rather long-term procedures, which usually require multiple sessions to be accomplished. Knowledge acquisition is mostly concerned with supporting users to build up substantial reservoirs of knowledge. Users can learn about various topics (e.g., learning fundamentals of physics) or acquire different skills (e.g., learning how to program) in interactive learning scenarios. Different learning strategies depending on the cognitive level of the user can be offered to the user that consist of providing explanations, feedback, or encourage reflection. Retrieved information items for these tasks can include educational materials but also, for example, grading schemes.

4.1.3.3 Content Recommendation

We consider the task of (conversational) recommendation as an information access task in which the classical notion of relevance is replaced with user preference. Consequently, recommendation items can be retrieved with a search engine and ranked according to user preference. Typical recommendation tasks that users fulfill with a RAG system are uttered as a question (e.g., "what should I cook tonight?"). Thus, RAG can be seen as a cross-domain and multi-modal recommendation.

4.1.3.4 Planning & Solution Implementation

While planning tasks are to provide step-by-step plans and instructions, problem-solving tasks incorporate implementing solutions and providing the results to the user. Mathematical, logical or coding tasks are examples of problem-solving tasks that follow strict constraints. In these tasks, retrieved information items can include educational materials providing explanations for these constrains. For tasks outside of our problem-solving category, constraints can be weaker or even optional. These tasks can involve procedural tasks that can be anything from retrieving while following a cooking recipe to repairing an electronic device. Information items that can be retrieved for these tasks can be manuals, recipes, and other documents containing instructions. Furthermore, planning and scheduling tasks can be planning trips or scheduling appointments for which transport schedules or calendar item retrieval are required, respectively. For some of these tasks in this category, the RAG system also can directly implement the solution. For example the solution can be a code artifact for solving a coding problem, a sent email or a scheduled meeting if the task relates to communication or calendar planning.

4.1.3.5 Content Generation & Creative Tasks

Content generation and creation tasks relate to the top of Bloom's Taxonomy. In the context of RAG, it includes the creation of new content, the generation of hypotheses, and the generation of communication items. For these tasks, the retrieval can be done over different types of information items, including character descriptions to write a story or generate an image, existing Lewis structure of molecules to hypothesize a new one, or messages in multiple threads to write an email. Based on the retrieved information items, the RAG system aims to generate a new information item, which is then the solution object.

4.1.3.6 Domains of Applications

Based on the pool of tasks imagined for the creation of the task taxonomy, we envision scenarios in different domains and population segments for which RAG systems can be particularly beneficial.

- For computer science engineers, with expertise in coding, a RAG system can support their daily tasks related to the creation and management of codebases. For example, the RAG system can implement new features, identify weaknesses, review code and summarize optimization strategies given the current status of a project and external documentation.
- For K12 pupils, a RAG system should be tailored to make the pupils engage with it to lead them to the solution object rather than providing it directly. For example, using different modalities when learning a new language and showing the richness of the language using paraphrasing.
- For elderly with different level of cognitive decline, a RAG system could remind them about their daily tasks, curate an image gallery of their life to help with memory and assist them in keeping their skills
- For marketing specialists, a RAG system could monitor trends for their respective market or create slides for presentations.
- For job seeker, a RAG system could give feedback on CVs, teach him/her new skills, help prepare for interviews, and monitor new open roles that come online.

4.1.4 Challenges and Limitations with respect to User's Cognitive Skills

Despite great advances and the already high effectiveness of RAG systems, the overuse of these systems imply certain risks for the user. There is growing evidence that GenAI systems contribute to the deterioration of user cognitive skills [86, 60]. Over-reliance on technology shifts many essential mental tasks from our brains to our devices leading to cognitive offloading. For example, it has been found that GPS replaces spatial navigation [177], search engines might replace memory recall [151] or AI assistants could take over reasoning and problem-solving [60]. In case of RAG systems, user cognitive skills across all levels of Bloom Taxonomy (Figure 1) might be affected as outlined below:

- **Query formulation:** Automatic query formulation replaces the formulation of keyword queries from the user's side with natural language problem descriptions. These problem descriptions neither have to be complete nor comprehensible to a great extend in order for the RAG system to provide somewhat satisfactory results.
- **Source triaging & relevance assessment:** RAG systems filter the retrieved sources before a user sees them and takes away the assessment of relevance for individual sources.
- **Close reading & extraction:** RAG systems typically provide well structured summaries that do not require deep engagement with the text to pick up key points. Generally, close reading of individual documents is discouraged.
- **Multi-source synthesis:** RAG automatically integrates retrieved evidence into automatically synthesized responses which is taken away from the users.
- **Factual recall:** RAG responses tend to be complete or at least provide the option of follow up questions that reduce the need for the user to remember individual facts or recall knowledge.
- **Critical thinking:** RAG systems typically include references to external sources which transports a sense of trustworthiness. This trustworthiness may discourage the user to do fact checking and external validation, which leads to a decline of critical reasoning abilities.

- **Planning & creation:** RAG systems can provide step-by-step instruction on how to solve a problem. Thus, users do not need to develop their own solution strategies or a novel solution object, which might hinder creativity, or development of planning and problem-solving abilities of an individual.

We expect that the push towards efficiency and the increased ease of use of LLM and RAG systems will result in a reshaping of the cognitive landscape of users, and likely in less involvement of users' cognitive processing in reaching the output, as well as in various forms of superficial learning and shortcuts. While potentially more information can be provided by RAG systems and consumed by users due to the systems' higher coverage of retrieval sources compared to what a user can find during her search as well as intrinsic and extraneous cognitive load [152] of reading can be decreased, the information is usually already extracted, synthesized, processed, structured, and presented for the user. Furthermore, GenAI poses a risk of standardizing language and reasoning for its users [150], potentially presenting a limiting factor for diversity and richness. While in Section 4.1.3.6, we name examples of ideal cases where RAG systems could help users, the risks should not be ignored, especially for vulnerable groups like children and the elderly, who could be adversely affected when it comes to their cognitive development and engagement.

We argue that several scaffolding elements could be incorporated into the design of future RAG systems to make them support user cognitive engagement in relevant settings, such as user learning and training. While these alterations would typically lead to diminished efficiency and effectiveness in reaching direct outcomes/answers, this would come with an added effect of fostering cognitive abilities of users and a potential increase in trust of the RAG system (e.g., by "revealing" the solution process). Such scaffolding elements could be then especially suited for cases when speed, ease of use, and accuracy may not be the most important, or could be sacrificed, much like jogging and physical training result in an increased effort and higher strain on those who prefer these activities rather than moving by cars.

We highlight below a set of such scaffolding elements that align with each of the identified cognitive skills affected.

- **Make retrieval more visible and interactive to keep information discovery skills.** For example, the system can show queries used by RAG and perhaps even allow the user to add/modify them, thus exposing the retrieval stage. The system could then explain to users how documents were found and chosen, or let them intervene in this process.
- **Promote close reading and evidence engagement to preserve critical reading, comparison, and fact-checking skills.** The system could highlight where particular information nuggets come from, quote some parts from original sources, or direct users to read selected content verbatim. In addition, the system could include a reflection stage or could prompt the user to think critically.
- **Let users integrate multiple sources to support information synthesis skills.** The system could visualize agreements and conflicts across sources, or suggest alternative summaries and different synthesis paths for comparison.
- **Embed counterfactual reasoning to support reasoning and critical thinking.** The system could provide one-click generation of alternative perspectives or assumptions, and/or show reasoning chains to let the user edit assumptions or steps. Moreover, the system could provide an option to change an assumption and see how the conclusions differ in result.

- **Gamification, hinting and user questioning.** Instead of providing a direct final answer, the system could add hints, incomplete reasoning paths, or quizzes. In addition, it could incorporate what-if branches to modify the text.

While many of the above-proposed adaptations have not been developed and applied yet, there are already several promising initiatives worth noting. For example, a new feature called *Study and Learn*[3] has been recently added to ChatGPT. In the above mode, the LLM lets users actively learn through guiding questions and prompting reflection instead of the provision of direct answers. For instance, a student working on a given task, can be suggested to attempt a solution first or to explain reasoning before the answer is revealed. Kazemitabaaret et al. [80] propose CodeAid, a system that answers conceptual questions, generates pseudo-code with line-by-line explanations, and annotates student's incorrect code with fix suggestions instead of using directly the LLM to provide a coding solution. Jangra et al. [72] advocate the use of hints for letting users find the answers to their questions by themselves with the help of LLM in the form of generated hints. There are already several datasets of hints available, both manually created [115, 169] or automatically generated [116, 73] along with their corresponding questions and quality metrics. Toolkits like HintEval [117] provide methods for hint generation and evaluation measures such as convergence which estimates the ratio of candidate answers to the user's question that can be discarded after the application of a given hint. In [124], the authors propose a hint generation framework for middle-school level math word tasks using a language model to decompose the solutions into atomic mathematical operations.

The effect of any design improvements on the cognitive adaptation and intellectual effort of users need to be however continuously and carefully studied [176] based on longitudinal user studies, or even using electroencephalography (EEG) signals [86, 95]. Additionally, the benefits coming from such adaptations should be compared to the reductions in task completion efficiency and effectiveness. It is also crucial to differentiate the analyses based on user attributes such as educational background, age, culture or on learning objectives, as well as on the cognitive demands tasks inherently present.

4.1.5 Research Questions

Despite the progress in LLMs and RAG systems, many aspects of RAG still need to be improved to fully serve users as desired with their cognitive tasks. Below are the research questions that arise to meet the desired RAG portrayed in the previous sections.

1. **Understanding the user** RAG relies on LLM's capability of natural language understanding to understand the user's question and determine what information should be retrieved. In some cases, the question may be wrongly understood or the retrieval query wrongly formulated. A more accurate understanding of the user's question/intent and better formulation of search queries can improve the quality of RAG answers. In addition to improving LLM's general capability of language understanding, RAG-specific research questions can be: How to improve query formulation? How to leverage context information in conversation when RAG is embedded in a conversational system? How to involve the user in the loop to better formulate the question/query?
2. **Find relevant information when needed** An important task in RAG is to retrieve the most relevant information to support answer generation. Currently, the retrieval module is generally loosely connected to the generation module. The retrieval module is

[3] https://openai.com/index/chatgpt-study-mode/

often optimized separately using traditional ranking objectives. An interesting research question to investigate is whether different retrieval algorithms or systems should be developed so that the retrieved information fits better the LLM for generation, instead of fitting human users. This problem is further discussed in Section 4.3.2. The retrieval algorithm can also be optimized for answer generation, i.e., a direct feedback from the generator can be used for the optimization of the retriever.

Another interesting question is to detect the need for retrieval. Many questions can be reliably answered by LLMs, so retrieval is not always required. Retrieval is needed only when LLMs may lack sufficient knowledge on the question, or need to retrieve additional evidence to confirm an answer. Research in this line has started [154, 76, 11], but the detection needs to be more accurate.

3. **Measuring the quality of information** IR algorithms have been optimized to do ranking. There is no or little indication on the quality of the information retrieved. Questions that arise may be: Is the information relevant? Is it from a reliable source? Is it trustworthy or authoritative? Is it worth being integrated with LLM for answer generation? These questions have not been extensively investigated in IR literature. There is a need to assess the quality of the retrieved information for LLM generation.
4. **Integration of retrieved information in generation** The retrieved information is commonly added as part of the context when asking LLM to generate an answer. This loose coupling mode leaves full freedom to LLMs to use the information in its own way. Stronger coupling involves more interactions between the two modules 4.3.2. Despite the fact that LLMs can be finetuned to better integrate the information, there is no guarantee that the integration is well done. The integration is straightforward when the retrieved information is consistent with or complementary to the internal knowledge of LLMs. However, how about when it is inconsistent/contradictory with the internal knowledge? when the information is irrelevant? when it is from a unreliable source? This question is also discussed in Section 4.2.6.
5. **Contextual factors** RAG will be used in different application contexts (for general QA, or in education) and for different users (children, adults, students, ...). The information to be retrieved and the type of answer to be generated should be adapted to the context and the user.
6. **Reasoning capability** Many of the questions cannot be directly answered by retrieved information. It is necessary to connect or compare the information or perform reasoning. Although the capability of reasoning of LLMs has much increased in recent years, it is still insufficient for answering complex questions or achieving complex tasks. Developing higher capability of reasoning based on the retrieved information is still a pressing research topic.
7. **Specialized knowledge and skills** As RAG is expected to be applied in a variety of application contexts (coding, mathematics, medical applications, ...), it is necessary that it exploits domain knowledge and expertise, which can be presented in various forms: as texts, knowledge graphs, specialized LLMs, or a set of tools. Despite recent progress in these areas [105, 6, 186], more research is needed to find better ways to leverage them in RAG.
8. **Knowing its limit** However a RAG system is powerful, there will be always questions that it cannot answer, either because there is no answer or it does not know the answer. Admitting "I don't have the answer" requires RAG to be aware of its own limits. More research in this line will also reduce the likelihood of hallucination.
9. **Configurable RAG systems** In many special applications such as in education, a RAG system should be configured to retain its final answer, but provide a form of partial answer or hint to help the user to think. This question has not yet dealt with in research.

4.2 Information Behavior and Result Presentation in RAG

Mohammad Aliannejadi (University of Amsterdam, NL)
Gianluca Demartini (The University of Queensland – Brisbane, AU)
Carsten Eickhoff (Universität Tübingen, DE)
Norbert Fuhr (Universität Duisburg-Essen, DE)
Sebastian Heineking (Universität Leipzig, DE)
Martin Potthast (Universität Kassel, DE)
Harrisen Scells (Universität Tübingen, DE)
Damiano Spina (RMIT University – Melbourne, AU)
Johanne Trippas (RMIT University – Melbourne, AU)

The working group on information behavior and result presentation focused on user diversity, task variation, and interaction modalities. RAG systems benefit from natural language interfaces and grounding outputs in external knowledge, enhancing transparency and efficiency. However, challenges remain around over-reliance, critical thinking erosion, and cultural bias from dominant sources. We outlined the need for adaptive architectures that support diverse users – ranging from experts to vulnerable populations – and varied tasks, from factual queries to creative generation. Three key dimensions of interaction were outlined: user/system initiative, information progression and complexity, and human-machine collaboration – highlighting gaps and opportunities in current systems. Four collaboration paradigms were proposed, emphasizing future potential in collective intelligence. Open research questions aim at designing adaptive, transparent, and inclusive RAG systems that balance automation with user agency and engagement.

4.2.1 Motivation

A constraint of RAG system is that the input is provided by the user (and possibly their context). Thus, the RAG system is responsible to provide the appropriate output given the input. One challenge with this is that users may have different types of tasks and information needs in mind (see Section 4.1 for more details). For example, a RAG system may be used for factuality or for creativity tasks and different search and generation strategies may be more appropriate for different user objectives.

It is also important to note that different users may need different types of RAG systems and interactions modes. For example, young users, elderly users, vulnerable groups, domain experts, low-resource users may all benefit from different types of RAG system architecture and interaction frameworks.

4.2.1.1 Benefits

The naturalness of the interaction mode (input and output in natural language) of existing RAG systems has made the entry bar substantially lower for a number of users. Delegating the task of summarizing search results to the machine, it may be seen as a way to access and consume information more *efficiently* (as some of the cognitive effort is offloaded to the system).

Compared to non-RAG conversational assistants, one of the main advantages of RAG is that provides a mechanism to ground responses to external knowledge, thereby reducing the risk of providing incorrect – or outdated – information generated using the internal

knowledge of the LLM. As the response is generated by combining both external and internal knowledge, the system can attribute elements in the response to sources, providing a more transparent way to show the provenance and the reliability of the information. Related studies have shown how showing the plan an AI agent intends to follow can increase trust in the system [66].

4.2.1.2 Risks

In most cases, users want to minimize effort and maximize gain. This comes with the risk of over-reliance on the RAG system and of accepting the answers given as good ones. This may, in the long term, reduce the overall ability of users for critical thinking. A possible approach to deal with this is to design RAG systems that keep their users alert, vigilant, and critical of the output received from the RAG system. This could be achieved in different ways, e.g., with systems better communicating uncertainty, quality of the sources, and proposing alternative answers (see Section 4.3.3).

Another risk is that of using dominant sources of information for RAG. This may lead to cultural bias and lack of inclusivity of diverse points of views (see Section 4.4.7). To deal with this, future RAG system may need to consider a diverse set of sources to make sure different viewpoints are included in the generated response.

For different type of tasks (see Section 4.1) different levels of reliance on sources may be appropriate. For example, for information access tasks (e.g., learning) a strong dependence on reliable sources of information is necessary while for content generation and creative tasks (e.g., hypothesis generation) a higher degree of generated content may be appropriate.

4.2.1.3 Opportunities

As the population of users grows, it is also increasing the diversity of needs and tasks. It is important to consider how we can characterize such diversity, to be able to conceptualize and design RAG solutions that accommodate for such needs. This has implications in the tailoring of all different aspects of RAG user's experience: customization, ease-of-use, functional fixedness, cultures of interaction and reasoning, etc. For instance, web search engines provide an *Advanced Search* option that provides a substantially more powerful way to control the behavior of the search engines; however, this advanced functionality is rarely used by most of searchers. If we design more complex interfaces for RAG, which users or tasks will these interfaces support?

Similarly to other information access scenarios, users of RAG systems still need to go through different stages when forming their information needs [160]. There is an opportunity to understand how interactive RAG systems can better support users during the realization of their information needs – not only for information access tasks, but also for more complex tasks such as planning and content generation tasks.

If we consider that we will move beyond the "one-fits-all" paradigm, this has also implications in the way we should conduct our research. Participatory research initiatives will be needed to be able to gather requirements and characterize needs of different groups of users. We will also need to account for richer test collections and novel experimental methodologies to better understand which UI or mode of interaction is more effective for (or preferred by) which users.

4.2.2 Users and Tasks

A constraint of RAG system is that the input is provided by the user (and possibly their context). Thus, the RAG system is responsible to provide the appropriate output given the input. One challenge with this is that users may have different types of tasks and information needs in mind (see Section 4.1 for more details). For example, a RAG system may be used for factuality or for creativity tasks and different search and generation strategies may be more appropriate for different user objectives.

It is also important to note that different users may need different types of RAG systems and interactions modes. For example, young users, elderly users, vulnerable groups, domain experts, low-resource users may all benefit from different types of RAG system architecture and interaction frameworks.

4.2.2.1 Benefits

The naturalness of the interaction mode (input and output in natural language) of existing RAG systems has made the entry bar substantially lower for a number of users. Delegating the task of summarizing search results to the machine, it may be seen as a way to access and consume information more *efficiently* (as some of the cognitive effort is offloaded to the system).

Compared to non-RAG conversational assistants, one of the main advantages of RAG is that provides a mechanism to ground responses to external knowledge, thereby reducing the risk of providing incorrect – or outdated – information generated using the internal knowledge of the LLM. As the response is generated by combining both external and internal knowledge, the system can attribute elements in the response to sources, providing a more transparent way to show the provenance and the reliability of the information. Related studies have shown how showing the plan an AI agent intends to follow can increase trust in the system [66].

4.2.2.2 Risks

In most cases, users want to minimize effort and maximize gain. This comes with the risk of over-reliance on the RAG system and of accepting the answers given as good ones. This may, in the long term, reduce the overall ability of users for critical thinking. A possible approach to deal with this is to design RAG systems that keep their users alert, vigilant, and critical of the output received from the RAG system. This could be achieved in different ways, e.g., with systems better communicating uncertainty, quality of the sources, and proposing alternative answers (see Section 4.3.3).

Another risk is that of using dominant sources of information for RAG. This may lead to cultural bias and lack of inclusivity of diverse points of views (see Section 4.4.7). To deal with this, future RAG system may need to consider a diverse set of sources to make sure different viewpoints are included in the generated response.

For different type of tasks (see Section 4.1) different levels of reliance on sources may be appropriate. For example, for information access tasks (e.g., learning) a strong dependence on reliable sources of information is necessary while for content generation and creative tasks (e.g., hypothesis generation) a higher degree of generated content may be appropriate.

4.2.2.3 Opportunities

As the population of users grows, it is also increasing the diversity of needs and tasks. It is important to consider how we can characterize such diversity, to be able to conceptualize and design RAG solutions that accommodate for such needs. This has implications in the tailoring of all different aspects of RAG user's experience: customization, ease-of-use, functional fixedness, cultures of interaction and reasoning, etc. For instance, web search engines provide an *Advanced Search* option that provides a substantially more powerful way to control the behavior of the search engines; however, this advanced functionality is rarely used by most of searchers. If we design more complex interfaces for RAG, which users or tasks will these interfaces support?

Similarly to other information access scenarios, users of RAG systems still need to go through different stages when forming their information needs [160]. There is an opportunity to understand how interactive RAG systems can better support users during the realization of their information needs – not only for information access tasks, but also for more complex tasks such as planning and content generation tasks.

If we consider that we will move beyond the "one-fits-all" paradigm, this has also implications in the way we should conduct our research. Participatory research initiatives will be needed to be able to gather requirements and characterize needs of different groups of users. We will also need to account for richer test collections and novel experimental methodologies to better understand which UI or mode of interaction is more effective for (or preferred by) which users.

4.2.3 Open Research Questions

- How do we design RAG architectures that foster users' critical thinking abilities?
- How do we design RAG architectures that keep track of the provenance and lineage of information to transparently surface it to the end users?
- How should RAG system adapt to different users and tasks? This includes users who are domain experts and have complex information needs related to their domain, younger or elderly users with safety requirements, and users with diverse information access tasks.
- Should RAG system result presentation be adaptive and dynamically adjust for different users and information needs (e.g., learning and fact finding vs creative tasks) or should we rather have different RAG systems for different purposes?

4.2.4 Interaction Modes

RAG presents a wide range of possibilities to interact with information access systems. We believe that the mode of information with a RAG system heavily depends on three aspects: (1) the roles and initiatives of the user and the system, (2) the progression of information and the density of information at each information processing stage in the RAG system, and (3) the amount of collaboration between users and RAG systems. These three aspects define how information should be used and presented within a RAG system, and their combination can be used to define the tasks that are possible. We start by identifying the benefits and limitations and problems that exist in typical RAG systems of today before elaborating on the above three aspects, and finally identifying the opportunities that exist regarding modes of interaction with RAG systems.

4.2.4.1 Benefits

Existing RAG systems allow for natural language as the main form of interaction. This form of interaction is highly expressive and most interfaces are built around this mode of interaction. Furthermore, chat-like interfaces are already familiar to most users. Drawing parallels to list SERPs from search engines, text SERPs and chat-like interfaces provide a linear interface to interact with information, which further reduces the burden on users.

4.2.4.2 Limitations and Problems

However, this one-size-fits-all approach to interaction may limit how information is accessed by users, and other modes of interaction may be more suitable depending on the context and task at hand. For example, pointing at or marking up information may be a more natural way of selecting to learn more about or combine information.

4.2.4.3 Roles and Initiatives

The first dimension defines the level of initiative by the RAG system and its role in the interaction with humans. This dimension combines the concepts of (i) intervention in information behavior [34] and (ii) system involvement [17].

Coghlan et al. (2025) [34] define four models of search engines, ranging from a low to a high degree of intervention. The *Customer Servant* gives the user of the search engine *exactly* what they asked for without further clarification or value judgment. In contrast to that, the *Librarian* tries to discern what the user is *actually* looking for by consulting additional information like past search behavior or asking clarifying questions. The *Journalist* tries not only to understand the actual information need, but supports the user in gaining a more wholistic understanding of the topic underlying the information need. As examples, the authors cite that a *Journalist* would outline multiple perspectives and arguments on a topic or provide debunking information in response to conspiracy theories. The *Teacher*, as the highest level of intervention, focuses on the learning experience and tries to encourage critical thinking. As a consequence, a *Teacher* will not rank results only based on their relevance to an information need, but also on the basis of their deemed correctness and educational value.

While the degree of intervention is primarily concerned with the information behavior of the user, system involvement focuses more strongly *who* defines and executes search commands: the user or the system. Bates (1990) [17] suggested four levels of system involvement in search as well as a level 0 with no system involvement. At the first level, the system offers passive support. It acts as a knowledge source to provide information when prompted, but does not monitor user behavior or execute a search on its own. Examples of information include explanations of commands or suggestions for search strategies. Systems on the second level execute both individual queries as well as full search strategies when prompted by the user. The system is still not proactive but reduces the procedural load on the user by executing more complex operations like a search in all publications of a journal for a given period. The first proactive level is level 3. Here, the system monitors the search behavior to offer suggestions even when not prompted to do so. This can range from simple suggestions like removing typos to complex ones like trying to infer the underlying information need from individual queries and suggesting a search strategy based on that understanding. On the highest level of system involvement, level 4, the system conducts search activities on its own. Based on user preference, the system can report its performed actions or only present the results.

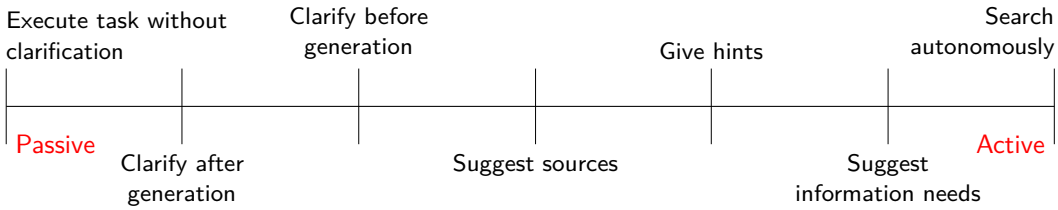

Figure 3 The dimension of initiative taken by the RAG system with examples for different levels.

Bates defined the level of system involvement as the complement of user involvement. In other words, as the involvement of the system increases, that of the user decreases. This zero-sum setting in involvement is the reason why we propose to couple it with the degree of intervention in information behavior. A highly involved RAG system bears the risk of users delegating large parts of their cognitive work to the system as it automatically executes tasks. With a higher degree of intervention, a RAG system might be able to keep the user involved and decrease the risk of cognitive offloading.

Figure 3 gives an illustration of the level of initiative that a RAG-system can take when interacting with a user. It is important to note that the same system can exhibit different levels in different situations. On the left side of the spectrum, the system executes searches only when prompted by the user and does not intervene to clarify the information need or present multiple perspectives on a topic. At a higher level of initiative, the system does not simply execute a query or command, but asks questions *after* completion. While the system still only acts upon explicit request by the user, it may list suggestions for follow-up questions (level 3 of system involvement) and clarify if the presented information meets the requirements of the user (like the *Librarian*). This is the level of initiative taken by most of the current RAG systems.

Instead of trying to complete the user request based on the most probable interpretation and clarifying *after* completion, a system of higher initiative performs the clarification step *before* taking any actions. As one advantage, this allows systems to handle cases of uncertainty about the information need, giving the user the opportunity for clarification instead of requiring them to verify retrieved content or generated output that was based on a false assumption. In addition to clarifying the information need, the system can also ask for confirmation about a planned course of action like the sources to be used or the sequence of steps the system will take to perform a task. While this form of initiative is still on level 3 of system involvement (no proactive search), the degree of intervention can reach that of a *Teacher* search engine, for example when the RAG system asks the user to consider a different course of action. As a consequence, this level of initiative needs to be employed carefully as to not make the system appear patronizing or annoying because it "refuses" to execute the user's request. The same is true for all following levels of initiative as the system becomes both more involved in the search process as well as intervenes more strongly in the user's information behavior.

A next higher form of initiative is for the system to suggest new sources to the user. Based on the history of requests, the system may identify useful sources to be added for future requests. While previous examples are limited to initiations by the user, i.e. they occur in response to a query or command, the source suggestion can be executed without the user asking for it.

Giving hints is a level of initiative that falls completely into the domain of the *Teacher*. The idea is to guide users toward answers without directly revealing them, encouraging deeper engagement with a topic and trying to mitigate risks like information overload and overreliance on automated reasoning [117]. A hint-based interaction can be either started by the user prompting the system or initiated by the system based on prior information behavior. Hence, it can be located on level 3 or level 4 of system involvement.

An example of high initiative is for the system to not only recommend sources, but information needs. As the user interacts with the system over time, a growing history of information needs and prior requests is collected. In contrast to traditional search engines, the user does not only give implicit feedback in the form of clicks, but explicit feedback in natural language. This allows the RAG system to analyze and relate previous information needs in great detail to make predictions about new information needs that the user has not yet articulated. This happens without prompting by the user.

At the highest level of initiative, the system performs searches autonomously. While the system involvement is by definition that of level 4, the degree of intervention can be both that of a *Journalist* as well as a *Teacher*. In the former case, the system would retrieve results from various perspectives to present a topic as wholistic as possible. The *Teacher* would go beyond that and not only present relevant results, but curate them based on educational value. In that role, the system would also ask questions or reveal new information step-by-step to teach the user about the topic.

4.2.4.4 Information Progression and Complexity

The second dimension is that of information progress and information complexity, which are tightly coupled. From these two dimensions naturally arise the types of information that are used and produced in a RAG system. Figure 4 illustrates this relationship between these two dimensions and some of the possible types of information. We define information progression as the stages of the information processing, for example, the start of interaction of a RAG system could be a user submitted query, while at the other end of the information progression spectrum could be a generated response from the system. We define information complexity as the amount and intricacy of information being processed at a particular stage in the progression of information, for example, low complexity information includes a keyword queries (à la Web search) and traditional 'ten blue links' search engine results pages; meanwhile high complexity information includes the past history of interactions with the RAG system (see Section 4.2.5) and a multi-modal generated response. These two dimensions are tightly coupled with each other, and we use this coupling to identify gaps and deficiencies in existing RAG systems that restrict access to information.

First, we use these two dimensions to define gaps in interaction modes with RAG systems. For example, in Figure 4, the vast majority of RAG systems allow only for relatively low forms of information complexity as the mode of initiating interaction with a RAG system and provide only relatively low forms of information complexity as information that can be interacted with as output; the system receives a user request in (mostly) text form and the system provides a (mostly) textual response. One clear gap that we see is to incorporate more feedback from the user (depending on the role and initiative of the user). Between the input and output, on the input side, systems could actively request more information or clarify the request. Towards the output side, systems could actively present a plan and ask for feedback on its execution. On both the left and right hand side of the progression spectrum, this feedback can either be low complexity (clicking thumbs up/thumbs down), or high complexity (written feedback to the model).

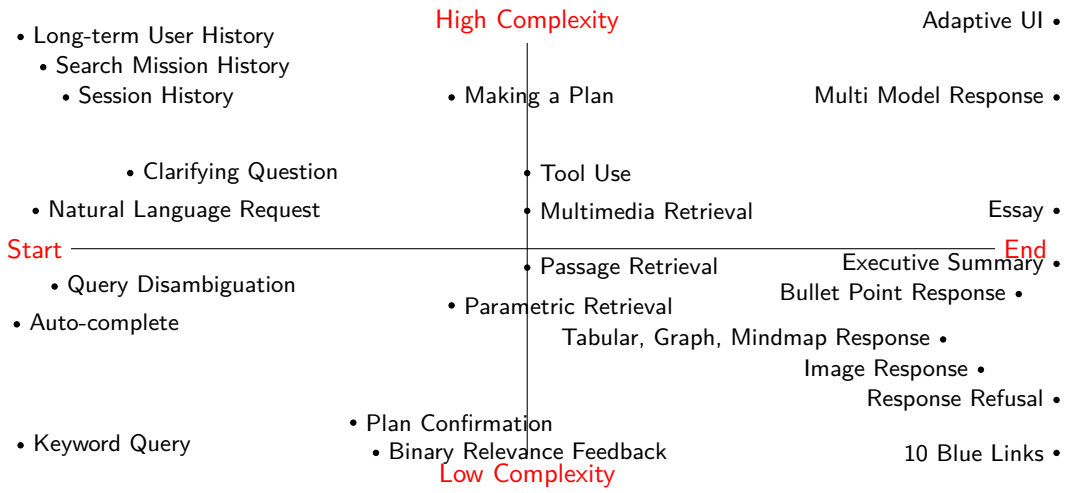

Figure 4 The two dimensions of information progress and information density, mapping some of the possible types of information that are used and produced by RAG systems.

Secondly, we use these two dimensions to critique existing RAG systems, and highlight deficiencies of these systems that would allow users to better support their information access tasks. One clear critique is the relatively low complexity on the extremes of the information progression despite the relatively high complexity of information used (e.g., complex tool use). There is an opportunity to better support users in their information access tasks by allowing them to provide lower complexity information and having the system produce high complexity information. One example of a hypothetical system that addresses this critique could be one that processes simple keyword queries into entirely new user interfaces, e.g., generating interactable Web pages to support information access tasks (à la Wikipedia articles) or generating an editor interface to support creative tasks like hypothesis generation.

4.2.4.5 Human-Machine Collaboration

The third dimension of interaction modes is that of collaboration between human(s) and machine(s). We structure the following part under four paradigms of human-machine collaboration (or hybrid intelligence [2]) in RAG (Figure 5).

Single human, single machine. This is the most common paradigm nowadays in ad-hoc RAG, where a user interacts with a system to satisfy their information need.

Single human, multiple machines. In this paradigm, a user is able to work with a number of tools that cooperate to either solve the task or critically analyze how other systems aimed to address the task. Instances of this paradigm may include federated RAG (Section 4.3.5) or the Talmud UI (Section 4.5.3).

Multiple humans, single machine. While a RAG system may enable users to tackle more complex tasks, users may still struggle on how to use the system. Community Question Answering platforms are effective in collecting and sharing knowledge across users with a common goal. One may think of a RAG system (e.g., "RAGExchange") where multiple users work together to collectively learn or solve a complex task by interacting (either synchronously or asynchronously) with a common system.

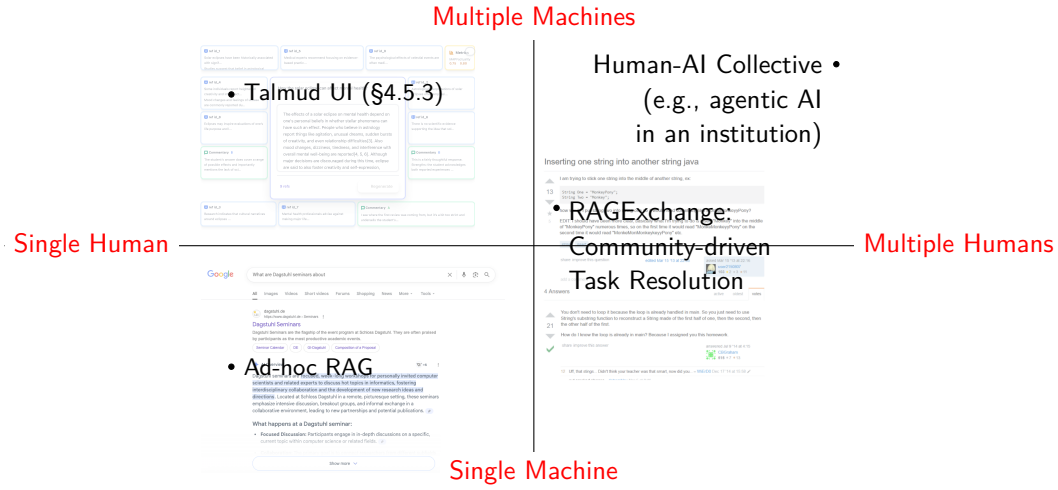

Figure 5 Human-machine collaboration paradigms in RAG.

Multiple humans, multiple machines. The paradigm above is not necessarily limited to one single machine, and it is intuitive to think of multiple users using multiple systems to solve a common task collectively (e.g., tackling creative tasks, such as using multiple RAG systems to generate new content). The highest level of collaboration is reached when we have fully connected network of humans and machines, where machines also collaborate among themselves. For instance, one may envisage a human-AI collective, e.g., members of an organization or institution working together with multiple RAG systems, agents, and tools.

These paradigms illustrate the spectrum of collaboration in RAG, from individual use to rich human-machine collectives, with some already partially explored in existing systems, and other pointing to plausible emerging collaborative scenarios.

4.2.4.6 Research Questions

Interplay Between User and RAG System Initiative. Both RAG systems and users can have different levels of initiative at a micro and macro level depending on the task, information progression, information complexity, and level of collaboration. RAG systems should be adaptable to these different modes of interaction, and the perceived initiative of the user.

How Information Progresses Through a RAG System. RAG systems are able to make use of information at varying levels of complexity throughout the entire information processing pipeline. However, there are many forms of information that a RAG system can use, and it is not clear how long chains of information processing at varying levels of complexity impact users. There is a clear knowledge gap regarding how RAG systems use information of varying complexity as it flows through the system, but also regarding the understanding of the interactions between the different forms of information.

Sequential versus Parallel Information Processing. Naive RAG systems use information in a sequential manner: the user request is used to search for information and generate a response. However, depending on factors such as task complexity and level of initiative, the RAG system could attempt to process information in multiple ways, all in parallel:

one component could be responsible for proactively searching a document collection, while another component clarifies the task and yet another component works on making a plan of action. Each of these parallel information processing steps can be used to better understand the task and better assist the user in achieving their information access task.

Collaborative Modes of Interaction in RAG Systems. The most common form of interaction with a RAG system nowadays is single human, single machine. However, the other combinations of humans and RAG systems allow for much richer forms of interaction with these systems, potentially offering the ability to enable users to better achieve their information access tasks. While we have presented some examples for what these modes of interaction might look like, we envision that the aspects of interaction modes discussed in this section have a role to play in the development of systems and interfaces to support highly collaborative RAG systems.

4.2.5 Keeping Track of History

4.2.5.1 Motivation & Benefits

As search tasks persist over time, information needs recur, or audit trails are required for compliance, maintaining explicit histories of previous inputs, interactions and outputs becomes desirable. Concrete benefits of such functionality are the ability to (1) re-find previously encountered material, (2) continue tasks extending across multiple disconnected sessions, (3) interrogate explicit logs of previous interactions, (4) maintain lifelogging records, or (5) try to understand the inner workings of the system.

4.2.5.2 State of the Art

There currently are no dedicated user interfaces for RAG-based history management. Most commercial language modeling services available on the market offer recency-sorted side-bars in which a linear list allows navigation into the raw historic interactions between user and system. The session contexts can then be returned to, in order to continue longitudinal tasks or retrieve information from the discourse. Several platforms offer privacy modes under which logging or platform-wide use of the interaction are temporarily suspended.

4.2.5.3 Risks

Maintaining any form of user- or system-accessible history comes with a number of potential pitfalls. In the following, we discuss four risks: threats to user privacy, failure modes in service personalization, service deterioration due to context noise, and potentially lacking user acceptance.

The most obvious risk is the potential for abuse and linking of sensitive private information. With growing duration and richness of the records maintained, the potential for harm (e.g., in case of hacks, data breaches or simply inadequate access control) to the user's privacy grows. Considering a scenario in which advanced system capabilities lead to regular reliance on the tool, its historical record would form a near-complete *virtual twin* of the user.

Going beyond privacy, a rich record of historical interactions is a valuable resource for personalization (see discussion below). As with all adaptively learning systems, such an evolution of system behavior in response to implicit usage patterns introduces the risk of unintended feedback loops, forming self-reinforcing echo chambers of biased information. It is even conceivable that potentially addictive properties may emerge over time.

Next, if historical information is used as additional context to future system interactions, the challenge arises of selecting the relevant session-orthogonal "slices" of history to be used for contextualization. Failure to adequately perform this non-trivial sub selection may lead to rapid deterioration of service quality in response to context noise.

Finally, if advanced user interfaces for history keeping (e.g., along the lines of the outlook below) are introduced, we will face challenges of equitably ensuring accessibility of tools and visual metaphors among disparate user groups. This creates a risk of providing functionality only for users of select levels of technology literacy or education.

4.2.5.4 Research Questions

In keeping with the previously discussed risks, we see a wealth of opportunities with considerable potential for beneficial innovation beyond the state of the art. in the following, we motivate nine research questions centering around the notion of history in RAG systems.

Aggregation and Summarization. Long system-generated outputs and multi-turn interactions can quickly become overwhelming for the user. Returning to the extensive raw history, a long-suspended task may in fact become daunting. Instead, aggregates and summaries of interactions and outputs may offer a much smoother re-entry. Such summaries do not necessarily have to be limited to the textual modality but might benefit from visual depictions of the concepts covered and materials visited (or retrieved but skipped) during prior sessions. Finding the right (mixture of) modalities for summarization will depend on properties of user and task.

Mapping Information Landscapes. As an extension to the previous research question, we encourage an exploration of explicit charts of the subspace within the collection that has thus far been explored. This could include visualizations of document linkage graphs, topical ontologies or other relevant structures induced by the concrete collection indexed for retrieval. Offering users a rich understanding of the topology and connectivity of the collection may offer considerable benefits in usability and allow for easier scrutinization of generative system outputs.

Resource Management. The availability of such rich, potentially interactive aggregations and visualizations of the collection, resources, and tools available to the RAG system offers the unique potential for the user to take a more active role in resource management. This could, for example, follow the create, read, update, and delete (CRUD) [109] paradigm, allowing users to save retrieved resources for future retrieval, marking them up as "high value", creating new content in their personalized local collection, or deleting encountered material that they deem of low utility or even harmful. This final step can extend to the removal of entire information channels (such as, e.g., email in a particularly noisy inbox). It remains an open challenge how to offer this functionality without cluttering the interface and with the necessary safeguards for users to conveniently undo or modify previous actions.

Handling Recurring Information Needs. A non-trivial proportion of searches are dedicated to recurring information needs. Advanced RAG system histories should include functionality for automating recurrence or re-finding of previously encountered material and answers. It is an open challenge to find effective interfaces in which to manage such needs. There might for example be an explicit register of previous interactions with external retrieved material that should be made available to both the user as well as the system, when prospectively formulating or answering new needs.

Living SERPs. As an extension to recurring information needs, there are standing queries for which the user seeks periodic updates from a changing collection. Currently, such standing needs are addressed only by fringe tools. An explicit history feature in the form of a "living document" might be much more effective at serving this purpose. We can, for example, imagine self-updating documents that reflect the latest state of material available in the collection, as new scientific papers are published and existing ones might be revised and altogether retracted. The exact interface, highlighting and notification modes for such living SERPs remain an exciting challenge for future research.

Personalization. From a system perspective, explicit histories offer the potential for tailoring outputs and interactions to the specific preferences and needs of (groups of) users. Importantly, this can include, but should not stop at, the level of traditional search personalization for better document relevance estimation. Instead, we can imagine sophisticated agentic workflows in which tailored interaction schemes or user interfaces are generated to fit the concrete user. We believe that it is important to communicate these actions to the user and give them control over the applied forms of personalization. This might, for instance, be achieved via additional (parallelized) retrieval operations from instruction history to arrive at preferred output formats.

Transparency and Explainability. To ensure highest fidelity and trustworthiness of RAG-based systems, their histories must afford users a transparent view into the tools, resources, and operations used to satisfy their request. While there have been significant advances in mechanistic interpretation of generative systems [102, 28, 8], communicating their insights in a form accessible to users of different technology literacy levels remains an important open challenge.

Reasoning Traces. A specialized form of explanation is concerned with the concrete reasoning steps and tool call that a model takes to arrive at an answer. These so-called reasoning traces are, if at all available, often returned in the form of intermediate chain-of-thought utterances, and do not necessarily support human comprehension. Key open questions are (1) how to produce high-fidelity reasoning traces that truly reflect the conceptual circuit enacted by the model, rather than merely a spurious post-hoc explanation, and (2) how to best represent these traces to users. This could potentially be done in much richer formats than text or static images, for example allowing the user to interactively "jump into" intermediate steps, provide feedback or alter processing strategies.

An Opportunity for Revisiting IR Literature. Finally, the emergence of retrieval augmentation offers an opportunity to revisit classical IR literature on interaction paradigms for history keeping that had been proposed (and often rejected) in the context of list-based SERPs. Examples include breadcrumbs of intermediate search and reformulation steps taken by the user [145], and information scent and foraging theories [127].

4.2.6 Sources

There is a tension between retrieval and generation in RAG systems with relation to their sources. Retrieval ensures grounding in external sources, while generation produces fluent but sometimes unverified output. Abstractions such as embeddings or summaries simplify information but remove detail, and retrieval becomes essential for reconstructing these details. A key challenge is moving from abstraction back to specific information.

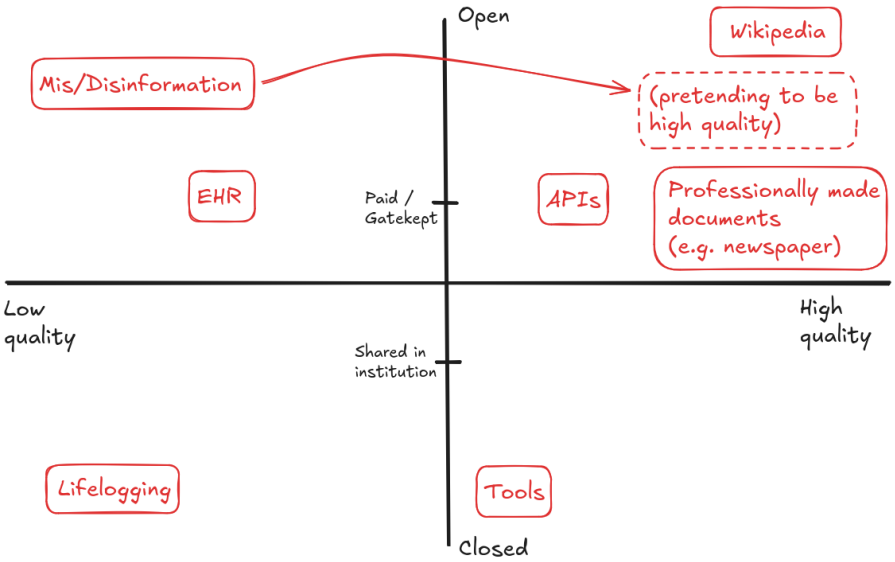

Figure 6 The source space, including familiar sources that we can envision fall in this categorization, as well as cases where the quality of the source is low, but the goal of the creator is to show it is high quality (e.g., mis/disinformation).

This same tension highlights how the information space is shaped by abstracting knowledge into models that organize and connect sources. It leads to important questions about the relationship between individual sources (i.e., databases, tools) and the larger information space they form together. It also raises the possibility that different kinds of information might give rise to distinct information spaces, which could be compared or combined.

The new paradigm of having a RAG system "read" the sources for us leads to numerous opportunities in enabling a wider range of tasks and users, but also poses risks that the community should take into account and discuss.

Wikipedia is where the search stops and RAG starts, is it enough? A study on conversational systems shows that many responses are drawn directly from Wikipedia articles [162]. This heavy dependence turns Wikipedia into the default external knowledge base for user queries while narrowing the scope of information and amplifying the gaps and biases in a single source. Retrieval-augmented generation systems follow the same principle of retrieving supporting evidence before producing responses. The reliance of conversational models on Wikipedia raises an important question for RAG research. Can Wikipedia be treated as a sufficient knowledge source, or must effective RAG include broader and more diverse corpora to overcome these limitations?

A large portion of user prompts and information needs can be addressed by relevant Wikipedia articles, but we envision a broader perspective where various and diverse sources can be leveraged by the system. Figure 6 depicts our vision of how different sources of information can be classified and leveraged by RAG systems, as well as the opportunities and challenges that come with them.

As seen in Figure 6, sources can be classified along multiple axes. One axis ranges from low to high quality, reflecting trustworthiness and accuracy. Another axis ranges from open to closed, covering open access, paid or restricted sources, and institutionally shared collections. Further axes could include modality (text, structured data, multimedia) and stability (static versus dynamically updated).

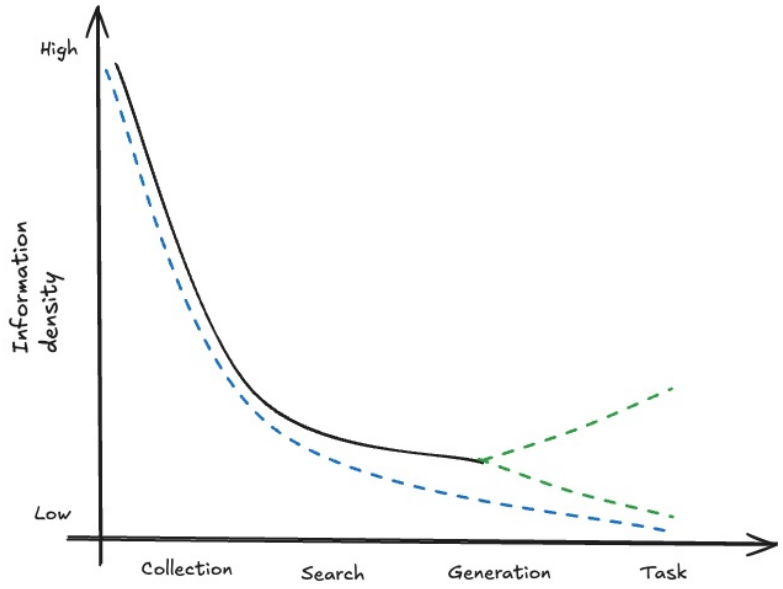

Figure 7 The information density diagram.

Figure 7 shows the information density to the user. In other words, it depicts the amount of information that is available to the user at different stages of RAG. As we see in the plot, one vision (the blue dotted line) is that as the user moves through the different RAG stages, the amount of information decreases significantly, because the system knows more about the user's information need. Clearly, as soon as the user enters a query and searches for it, the information space is much more limited. Moreover, a RAG system can provide a short summary or a tailored response to the user; therefore, the user would not need to go through all the results in the list. However, in another vision, the system could provide additional information that is not in the ranked list, as it knows more and more about the user's task. Therefore, the dashed green lines show the two possibilities at this stage where an efficient RAG system could leverage its internal knowledge, as well as the ranked documents, to provide more information or more steps in solving the task at hand.

Benefits. One key benefit of RAG systems is **consistency**. Placing a generative model at the end of the pipeline, information can be presented uniformly across multiple outputs. This consistency improves accessibility for diverse user groups, including elderly users and individuals with disabilities, by providing information through a uniform and consistent channel. Consistent results presentation also reduces confusion when users access the system repeatedly or compare outputs from different queries.

Another advantage is the **knowledge synthesis from multiple sources**. RAG systems consolidate and integrate information from diverse places, making information discovery more efficient and reducing the cognitive load on users. By synthesizing complex or dispersed information, these systems allow users to focus on higher-level analysis, decision-making, or learning. This benefit extends to a wide range of users, enabling them to access insights that might otherwise require significant time and expertise to assemble.

RAG systems can also enhance relevance and personalization. Combining retrieval and generation allows the system to prioritize information that aligns with the user's needs, preferences, or context. This tailored approach improves the efficiency and usefulness of

information access, supporting tasks ranging from research to everyday decision-making. In addition, RAG systems support critical reasoning and exploration. By exposing users to synthesized knowledge drawn from multiple sources, they enable cross-referencing, comparison of perspectives, and identification of patterns that might not be obvious from a single source. This capability can promote more informed and balanced understanding, particularly in complex or rapidly evolving domains.

RAG systems can also facilitate scalability and adaptability. For instance, they can process vast information, update outputs as new sources become available, and generate summaries or explanations in multiple formats or modalities. This flexibility makes RAG systems valuable across domains, user groups, and applications, from education and research to healthcare and public information services.

Risks. RAG systems introduce a number of risks that affect both the reliability of information (i.e., sources) and the way users interact with it. One concern is that sources' quality, reliability, and authority may be ignored or flattened, which undermines the trustworthiness of outputs and can amplify existing biases. Another problem is the explainability paradox. That means the presence of citations creates the impression of transparency and explanation, even when the connection between the cited source and the generated text is weak or misleading.

Erosion of Credibility: Credibility and provenance are also at risk. The tone and style of original sources are often lost, and diverse voices are merged into a unified generated output. This process erases the distinct identity of sources, diminishes credibility, and makes verification more difficult. Alongside this, the potential for misinformation and disinformation is high. Generated text may contain inaccuracies, and since users often do not fact-check, quality differences are hidden behind the smooth surface of generation.

Another risk is epistemic homogenization. By blending diverse perspectives into a single synthesized response, RAG systems can erase differences across sources and reduce the visibility of plural or contested views. Simultaneously, fluent language encourages users to overtrust the system, fostering cognitive offloading, weakening information literacy, and discouraging verification against original materials.

Temporal and contextual risks complicate the use of retrieved sources. Content may be outdated, incomplete, or detached from its original setting, and generation can obscure these issues by presenting the information as current and authoritative. This masking effect increases the likelihood of misinformation and disinformation, since users often do not verify or fact-check what they receive. The absence of transparent provenance makes it even harder to distinguish reliable material from misleading or false claims. Feedback loops introduce additional systemic risk. Minor inaccuracies and distortions can accumulate and amplify when generated outputs are recycled into retrieval pipelines or incorporated into training data. Over time, this process undermines credibility and accuracy and may contribute to model collapse in extreme cases. In addition, this introduces a wide variety of risks of misuse of the generated content (see more in Section 4.4).

Opportunities. // Research questions. Expanding RAG Beyond Text. Advantages in LLMs and multimodal language models enable RAG systems to move beyond textual information. RAG systems can synthesize richer knowledge and provide more comprehensive responses by leveraging images, audio, video, and other modalities. The following research questions aim to explore how multimodal sources can be effectively integrated, what challenges arise, and how these approaches reshape the landscape of information retrieval and generation.

- How can RAG systems be designed to incorporate multimodal information beyond text?
- How should RAG systems present results that integrate multiple modalities to ensure clarity, usability, and effective communication of synthesized knowledge?

- What methods are needed to enable LLMs and multimodal language models to retrieve, align, and synthesize knowledge across multiple modalities?
- How does integrating multimodal sources change the structure or representation of information spaces within RAG systems?
- What opportunities and risks emerge when RAG systems move beyond text, particularly regarding accuracy, bias, interpretability, and user trust?

Personal Information Management. RAG systems have the potential to access a broad range of private and personal information, as illustrated in Figure 6. This access opens significant opportunities for personal information management, such as developing intelligent personal assistants or lifelogging systems, a long-standing vision in the information retrieval community. The reasoning and synthesis capabilities of LLMs applied to heterogeneous personal data create new avenues for research, particularly around how these systems can manage, interpret, and utilize sensitive information effectively.

- How can RAG systems effectively manage and synthesize diverse personal and private information to support intelligent personal assistants?
- What methods enable LLMs to reason over heterogeneous personal data while preserving privacy and security?
- What are the ethical, privacy, and trust considerations when RAG systems access and process sensitive personal data?
- How can results from personal sources be presented in an interpretable, actionable, and user-friendly way for lifelogging or personal assistant applications?

Tools, Databases, and APIs. RAG systems are increasingly capable of interacting with external tools (i.e., applications such as email, calendar, private repositories), databases, and APIs. Existing generative AI systems already try to leverage different sources to solve tasks by combining multiple types of information. In the context of RAG, different sources can provide complementary services, such as recommendation tools offering personalized suggestions or search engines supplying factual information.

- How should RAG systems present combined results to users?
- How can LLMs combine outputs from different sources accurately?
- How can RAG systems choose the best source or tool for a specific query?
- What problems arise when coordinating different types of services, like recommendations and factual data?

Quality-driven Real-Time Bidding (RTB). In a world filled with LLMs, each of which having its own expertise, we envision a competitive environment of various RAG systems aiming to contribute and be part of the user experience. This could lead to a quality-driven RTB, where LLMs can bid on the next token that goes as the system output. In other words, a quality-driven bidding system (e.g., based on the LLM's uncertainty, predicted performance, etc.) can determine which LLM gets to generate output for a given number of tokens, enabling a collaborative environment of LLMs. Therefore, we envision the following research questions:

- How would an ecosystem of a quality-driven RTB be designed to allow for fair and effective content generation?
- What could be the quality measures we could consider to rank the RAG systems in RTB?
- What would be the granularity of the RTB (e.g., token-level, passage-level, etc.), and how would that affect the overall quality and user experience?

Collaborating and Sharing with RAG Systems. As RAG systems become more widespread, scenarios become more prevalent in which users share or exchange their RAG systems or outputs. Such exchanges enable collaborative tasks, joint problem solving, or discussions grounded in shared information. Exploring how RAG systems can support collaboration and knowledge sharing raises several important research questions.

- How can RAG systems support effective collaboration among multiple users?
- How can shared outputs be presented in a way that is clear, actionable, and easy to understand? (See 4.2.5)
- How can user interfaces and interaction design be optimized to make collaborative RAG systems intuitive and efficient for multiple users?

Live Fact-Checking of Sources. While having the risks of generation disguising mis-/disinformation and giving the wrong impression of credibility of the user, LLMs can also help users in live fact-checking of the sources used when generating a response. This can be done in various dimensions, such as checking for the attribution of the generated content, checking the credibility of the cited sources, and checking the alignment of the final response with the utilized sources. We argue that the fact-checker LLMs have to be independent and different from the LLM used in the RAG system to provide a reasonably reliable assessment of the credibility of the sources. This could be integrated in the UI, in a similar way proposed in Section 4.5.

GAR. One approach to improving the user experience is to enhance ranked list presentation by combining retrieval with generation. In a GAR-based interface, search results could be enriched in multiple ways, such as fact-checking documents, generating more informative summaries, and incorporating images or other media. The interface could support interactive exploration and filtering, adapt to user intent, and explain why content is ranked or generated. Systems could learn from user feedback over time (i.e., user model), improve accuracy, and highlight potential biases to build trust. These features would enable more personalized, reliable, and actionable search experiences for the future.

- How can ranked lists use generated content to make search results clearer and more engaging?
- How can interactive and adaptive features help users explore and filter results in real time?
- What methods can GAR introduce to detect and reduce bias in retrieved documents and generated summaries?

4.3 Retrieval-Augmented Generation: The System's Perspective

Qinqyao Ai (Tsinghua University – Beijing, CN)
Avishek Anand (Delft University, NL)
Michael Granitzer (University of Passau, DE)
Faegheh Hasibi (Radboud University Nijmegen, NL)
Djoerd Hiemstra (Radboud University Nijmegen, NL)
Sean MacAvaney (University of Glasgow, GB)
Arjen P. de Vries (Radboud University Nijmegen, NL)
Guido Zuccon (Google Research Australia and The University of Queensland – Brisbane, AU)

4.3.1 From Naive to Ideal: RAG System Architectures

The original RAG system architecture provided a simple, static pipeline from retrieval to generation – referred to here as *naive RAG*. In this setup, retrieval results are appended to the input of a large language model, which generates a response. More recent advancements have given rise to *active* or *dynamic* RAG systems, in which the generation module may initiate additional retrieval steps, for instance to reduce uncertainty during generation. In what follows, we articulate the vision for an *ideal* RAG system, then contrast it with the naive baseline in Table 1.

An Ideal RAG system

We envision an ideal RAG system that is a **compound AI system** that learns, reasons, and communicates, **continually evolving** and **adapting** to users, data, and the world itself, while **making principled trade-offs** between effectiveness, efficiency, and cost.

Limitations of current RAG Systems

Current RAG systems are primarily designed as plug-and-play augmentations for LLMs, retrofitted onto existing search infrastructures originally built for human users. In these architectures, retrieval is treated as a basic, static component, while the focus in innovation lies on the generation components as the dominant one to optimize. Retrieval typically relies on keyword or dense methods over static, pre-chunked documents, with a fixed scope that is unaware of query semantics, task complexity, or downstream usage, lacking any adaptive feedback loop.

Reasoning capabilities are limited, often reduced to data-driven shallow reasoning like chain-of-thought prompting or its modern counterparts [172, 165, 146] resulting in overthinking and inefficient unnecessary sampling path.

These systems are also poorly equipped to handle evolving knowledge, and feedback [132]. Updates typically require retraining or full re-indexing, with little support for continual or incremental learning. Furthermore, resource usage is uniform and over-provisioned, allocating the same compute regardless of query ambiguity or difficulty [154]. Uncertainty, whether in retrieval coverage or generation confidence, seldom surfaced. Users are expected to trust outputs without access to provenance, citations, or error bounds [142, 147]. Finally, most current RAG deployments rely on centralized architectures, retrieving from global indices

governed by the provider. Users have minimal, if any, control over what data is indexed or retained. System evaluation is narrowly focused on effectiveness metrics like accuracy or relevance, with little consideration for energy, latency, or other operational costs.

Characteristics of Ideal RAG Systems

- **Adaptive Efficiency-Aware Execution**
  - Dynamically adjusts compute usage based on task complexity.
  - Optimizes compute vs. effectiveness tradeoffs.
  - Predicts and operates at the minimum resource footprint needed.
- **Flexible Memory and Communication Channels**
  - Uses multiple memory types: short-term, long-term, editable.
  - Supports communication via text, vectors, or symbolic formats.
- **Uncertainty-Aware Design**
  - Tracks uncertainty from user intent, knowledge, and model confidence.
  - Surfaces and responds to uncertainty (e.g., via clarifying questions).
- **Federated and Decentralized Architecture**
  - Supports domain-specific retrievers and data ownership.
  - Adheres to privacy, policy, and regulatory constraints.
- **Composable, Modular Components**
  - Enables independent optimization and replacement of components.
  - Easier to update and maintain with minimal retraining.
- **Cost-Aware and Sustainable**
  - Operates within latency, FLOPs, or carbon budgets.
  - Emphasizes caching and hardware-aware execution.
- **Feedback Aware**
  - Supports feedbacks from different parts of the pipeline.
  - Allows different types of feedback from scalar, descriptive, and is able to adapt to feedback.
- **Multi-Level Reasoning and Planning**
  - Supports contextual reasoning and task-aware planning.
  - Allows modular integration of symbolic, analogical, or causal logic.

Comparing Naive and Ideal RAG Architectures

4.3.2 LLMs as Search Engine Users

From the onset, search engines have been designed for human users. Consequently, retrieval models and evaluation measures have been designed around human-centered notions of relevance and human behavior, and search systems have been optimized accordingly. This means for example that queries are typically issued in natural language one-at-time and results are typically presented as ranked lists intended to be examined by the user in sequence. These same search systems are now central to RAG, where a retriever supplements the input of a large language model with external documents. Thus, retrieval in RAG relies on methods optimized for humans rather than for LLMs. However, recent empirical studies have suggested that search results as returned by search engines might impair LLMs. For example, a model may prefer documents not to be ordered by relevance. Instead, it might even perform better when non-relevant documents appear first in the context. These findings raise a deeper question: How would a search engine for LLM be different from a search engine for humans?

Table 1 Comparison between Naive and Ideal RAG Systems.

Aspect	Naive RAG Systems	Ideal RAG Systems
LLM Interaction	Search designed for human consumption	Interfaces optimized for both humans and LLMs
Efficiency	Static compute for all queries	Adaptive, task-sensitive compute allocation
Reasoning	Shallow, post hoc reranking	Rich, integrated multi-level reasoning
Uncertainty	Hidden or implicit	Tracked, surfaced, and actionable
Architecture	Centralized monolith	Composable, federated modules
Memory	Flat, fixed-length context window	Hierarchical and type-specific memory structures
User Control	Minimal to none	Full control over indexing and retention
Evaluation Focus	Effectiveness only (e.g., accuracy)	Multi-objective (accuracy, efficiency, cost)
Maintainability	Costly updates and retraining	Incremental updates, easy maintenance
Feedback	No Feedback	Active & granular feedback

We envision **an ideal RAG system** as one in which the information needs of the generator can be flawlessly expressed to the retriever, and the retriever flawlessly expresses exactly the required information to the generator. In other words, there is flawless coordination between the retriever and the generator. This might mean having to renegotiate how information is represented and exchanged between the search engine and the LLM, how search strategies are planned and executed, and even re-thinking the overall architecture of RAG and the separation between retrieval and generation capabilities [158, 94].

Existing approaches have not yet achieved this ideal. Systems that "memorize" information in the generator's *parametric memory* suffer from forgetting [144], difficulty in attribution [26], and lack an expandable memory store [171]. Meanwhile, systems that invoke *Search Engines as a Tool* primarily use natural language as an interface between the LLM and the search engine, which is inherently ambiguous: natural language queries of a finite length cannot fully provide all context surrounding the information need, and results expressed in natural language lose their context.

To move towards this ideal system, we propose the following research directions:

- **Better understand LLM "behavior":** We argue that as understanding (human) user behavior has allowed the IR community to refine and optimize search systems to better support users, so understanding LLM behavior will allow us to better design search systems for LLMs. For this, empirical observations and methods from Machine Psychology [64], which aims to use theory and practice from human psychology to better understand LLM behavior, might help us determine what are the factors that influence RAG, what these mean for search engine design, and how LLM behavior in this context can be steered or controlled.

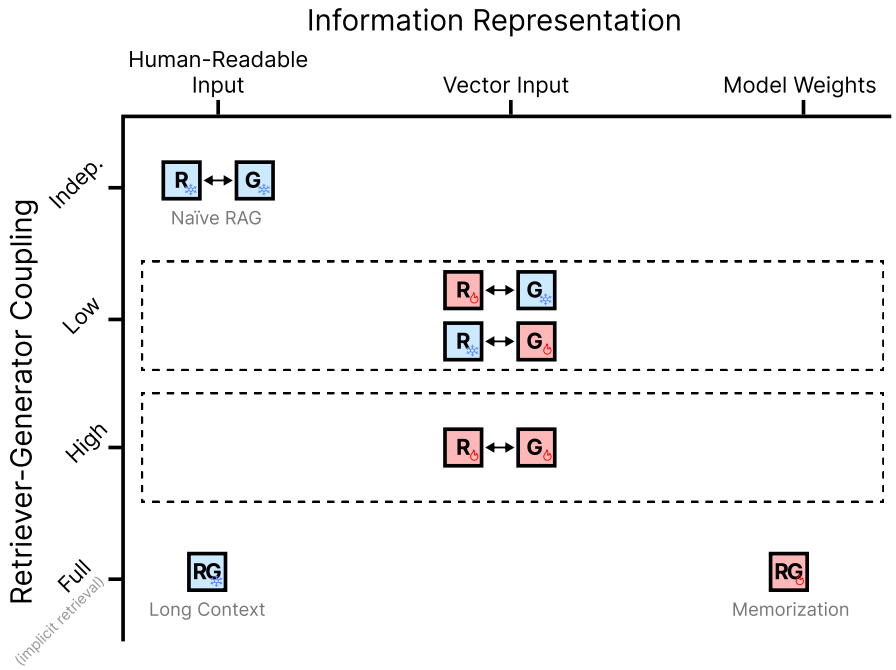

Figure 8 Comparison of current RAG approaches with respect to the representation of information and the coupling between the [R]etriever and [G]enerator components. A red box indicates the component has been trained for the task/data; a blue component indicates it is frozen (i.e. not specifically trained for the task/data). Naïve RAG interfaces between distinct retriever and generator components through human-readable text. Combined Retriever-Generators ([RG]) have been explored in long context settings [93] and through direct memorization into model parameters [159].

- **Enhancing Text Interface:** Differences between LLM and human user behaviors present opportunities to better align the text interface between the retriever and generator. These could be changes to the input, output, or interaction interfaces. On the input interface side, we could design a new query interface that aligns better with the way generators express their information needs (rather than the keyword or natural-language queries that current retrieval systems use). Perhaps this would mean supporting semi-structured queries to allow the generator to specify very specific information needs (which are challenging to express precisely in natural language and challenging for human users to write). In terms of the output interface, we could design new ways to present the retrieval results instead of a ranked list of documents. Perhaps this could involve technical details of the retrieval process, which are challenging for human users to interpret but could help the generator understand whether any follow-up requests are necessary. In terms of generator-retriever interaction, suggestions or clarifications to the query could be requested before fully executing the query. This could be inconvenient for a human user to use, but could help control retrieval costs and ensure that model context tokens are used effectively.
- **End-to-End Optimization:** Retrievers are typically optimized only for relevance to queries; within RAG these retrievers may surface information that is redundant, topically relevant but unhelpful for grounding answers, or even misleading (e.g. if it gets the generator to anchor to noisy evidence). Jointly training the retriever and the generator under a shared, task-specific objective promises to align retrieval with what actually helps

the generator produce correct, faithful and useful answer. This end-to-end optimization can involve gradient-based methods (if the retriever is differentiable, backpropagate through generator), policy gradient/reinforcement learning methods (treat the retriever as an action policy, optimize the reward based on the generator outputs as assessed through correctness, faithfulness or user feedback), and labels (by generating supervision signals that tie retrieved content to the final answer quality, most likely synthetically). However, jointly training the retriever and the generator presents a number of challenges. Generators are optimized with cross-entropy on tokens, not on truthfulness or factual grounding, therefore raising a training signal mismatch. This makes it hard to propagate useful gradients back to the retriever, especially when the reward is a "sparse" judgement, e.g. the final answer was correct. On the other hand, context length and inference cost constraints limit how much can be retrieved and passed as input to the generator. Optimising not just what is retrieved, but also how much and in what form (long documents vs. chunks vs. summaries/snippets) is still an open question; compression, summarisation and structured retrieval (tables, knowledge graphs) further complicate end-to-end training. Another open direction is how to encode feedback for end-to-end optimization, and in particular negative signals: if an answer is wrong, was it becausse the retriever pulled irrelevant documents, or because the generator failed to "reason"? Disentangling responsibility between retriever and generator is tricky for optimization. Similarly, current automatic metrics (e.g., token F1, BLEU, ROUGE, nDCG for retrieval) do not capture the nuanced success criteria of RAG (faithfulness, factual correctness, attribution, reasoning quality). Without good metrics, however, it is hard to optimize reliably end-to-end. Scalability and efficiency aspects of joint training also should not be underestimated: this training is computationally heavy, especially with large corpora and long-context models. Promising directions for exploration include differentiable retrieval mechanisms to enable smoother gradient flow; RLHF-like training for RAG, where human feedback (or synthetic judges) rewards correct and grounded answers; adaptive retrieval models that decide when/how much to retrieve, rather than always retrieved k documents; and the creation of unified architectures that tightly couple retrieval and generation, e.g., retrieval-augmented transformers where retrieval is embedded in the attention mechanism.

- **Representation of information:** The retrieval component is responsible to return information to the LLM to ground generation. Most common systems currently return this information as a rank list of search results, where each result is represented by the text contained in the document (or chunk or snippet). However, alternatives are possible, and representations could be defined across a spectrum that goes from human-readable text (the content of the document, snippets from the document, summaries of documents, etc), to embedding-based representations (e.g. as done in prompt compression approaches [75]) and model weights [155]. The use of model weights as a communication channel for RAG offers the opportunity for more principled and close-coupling approaches, but raise challenges such as how retrieval knowledge can be increased without forgetting, and how attribution can take place.

4.3.3 Uncertainty in RAG

Uncertainty is an important factor to consider when consuming information. Knowledge about the degree of uncertainty associated with that information is essential to make informed decisions and to identify potential risks in advance. Studies on uncertainty in IR have mostly analyzed the reliability of information sources and attempted to increase our understanding of the behavior of retrieval systems [61]. The introduction of generative AI systems in

RAG complicates matters, as some degree of uncertainty is also inherent to their behaviors and outputs [147, 126]. Therefore, understanding, modeling, evaluating, and expressing uncertainty is an important part of our discussion on RAG from the perspective of the system.

In the following paragraphs, we briefly describe how to view the concept of uncertainty in RAG and what we believe to be fruitful research directions in this field. Specifically, we first introduce what we believe an ideal system would do with uncertainty and how current systems perform so far, then we discuss what are the potential steps to guide us to achieve our goals.

Where we want to go

Uncertainty may or may not be part of the final output users expect from a RAG system, but it is undoubtedly a key piece of information that can help improve system performance and support downstream applications; e.g., abstaining from providing low-confidence outputs [147, 54]. Ideally, a RAG system with good support for uncertainty quantification should have the following abilities or characteristics.

First, an ideal RAG system should be able to understand and track any sources of information or reasoning uncertainty in its working pipeline. That means that when anything unexpected happens, we could trace back or at least evaluate which part of the system could be wrong. In practice, uncertainty could come from many places. Examples include but are not limited to (1) users, who provide vague information request, do not have a clear image on what they are looking for, or even may not possess the skill or expertise required to understand the system's outputs; (2) systems, which have multiple modules that work together to create the final output, and each module has their own reasons to introduce uncertain results, e.g., retrieval modules could be affected by documents from sources which have uncertain reliability, LLM could be affected by the bias, knowledge conflicts, or noise in its training data, etc.; (3) the world itself, where the information environment is dynamic and the provenance of each piece of knowledge and information may not be fully traceable and verifiable.

Next, with a good understanding of the uncertainty sources, the system should also have the ability to accurately quantify the uncertainty contributed by each module, e.g., uncertainty in the retrieval results, response generation process, and query reformulation. This not only provides important signals for debugging, but can also help us further improve the quality of each part of the system accordingly. For example, if we notice that the final output is unreliable mainly because of unreliable retrieved documents, then we could apply a more sophisticated filter and remove low-quality documents first. Finally, when presenting the output to real users, the system should have an effective way, perhaps a well-designed UI or data visualization tools, to express its uncertainty to users so that they can be fully informed about any risks behind the system and make wise decisions accordingly.

Where we are now

Many studies in the literature have analyzed the uncertainty of machine learning systems [42, 33]. Within the field of IR and NLP, there is also progress in identifying the uncertainty of documents [81], retrieval models [133], and LLM [43, 142]. For example, IR researchers have tried different methods to analyze the reliability or authority of documents on the Web [123] and social media [185]. There is also work on how to calibrate the ranking scores produced by learning-to-rank systems to correctly reflect the probability of relevance of each retrieved document [35].

For LLMs, researchers have explored different methods to analyze the internal states of LLMs and their output confidence (mostly from the perspective of token-wise or sentence-wise perplexity) [15, 156, 181]. These methods, however, are suboptimal for the RAG setup [148], as they only consider LLMs as the source of uncertainty. For RAG systems that combine all these modules above, things are still at an early stage. There have been works that use the advantages of uncertainty quantification techniques to improve RAG systems [154, 76], but research that directly studies uncertainty in RAG is still limited in the literature. Few studies have explored uncertainty quantification in the standard retrieve-then-generate RAG framework [148, 126], or in more advanced retrieval-augmented reasoning settings [149], where LLMs employ search engines as tools during the reasoning process [77, 180]. There are many challenges that prevent us from identifying and quantifying uncertainty in RAG systems effectively and efficiently, some of which are discussed in the next section.

As for the expression of uncertainty to users, there are some efforts that try to present an uncertainty score together with the output of the generative system in text and visual presentation modes [91]. However, the uncertainty scores provided are usually not grounded, and the way how systems express uncertainty to the user is by no means comprehensive and perfect.

What to do next

To achieve our long-term goal towards a reliable RAG system with uncertainty support, we brainstormed and identified four important challenges and directions to work. Specifically, they are:

- **Identification and Control of Uncertainty Sources**. Finding the sources of uncertainty is the foundation of all downstream processes. As discussed above, uncertainty in RAG systems has diverse sources including users, systems, and the dynamic environments of the world, e.g., ambiguous user queries, incomplete retrieval coverage, inconsistencies in the training data, or knowledge updates to science or events that are time sensitive. Identifying uncertainty sources requires not only tracing the provenance of information, either retrieved or generated by LLMs, but also translating this provenance into measurable signals of reliability. This is particularly difficult when there is not a clear definition of the unit of information. For example, documents could be chunked into parts with various lengths. Also, the usefulness and reliability of a piece of information cannot be fully understood when the context is fragmented. We need to figure out a more formal taxonomy of uncertainty and its granularity so that we can better understand and analyze the sources of uncertainty.
- **Uncertainty Quantification**. Together with source identification, quantification of uncertainty is another challenging problem. Even if we know the provenance of uncertainty, translating this provenance into measurable signals is still difficult because there is no "ground truth" for it. For example, existing studies mostly measure the quality of uncertainty quantification based on their correlation with the actual correctness of the system's output [14]. However, this cannot serve as rewards to train an uncertainty quantification model that covers different sources of uncertainty because that would be essentially training a model to fit the task itself and may not generalize to all cases. Unsupervised methods are the state-of-the-art for uncertainty quantification, but there is not a clear roadmap for how to keep pushing the limit of these methods in the long term. More theoretical studies are needed on the optimization and evaluation of uncertainty quantification [147].

- **Presentation of Uncertainty**. Presenting uncertainty to users introduces another layer of complexity. Systems must communicate uncertainty in a way that is understandable, actionable, and aligned with user needs. This could mean asking clarifying questions, showing supporting evidence, or designing simpler or more complicated user interfaces. Also, generation and retrieval behave differently with respect to how to show uncertainty. Retrieval makes it explicit through showing the documents and providing citations, while generation tends to minimize or mask it. This raises open questions about when citation is necessary (e.g., factual claims like "Paris is the capital of France" may not require attribution) and how to reconcile conflicting or incomplete information.
- **Utilization and Reduction of Uncertainty**. Assuming that we could get reasonable uncertainty quantification for RAG systems, how to use uncertainty to improve the system is also an open question. For example, uncertainty could be an important resource for model optimization, where it could help in filtering out biases, inconsistencies, and other types of noise in the corpus that may propagate into outputs. It could also serve as a signal to guide agentic systems to design their workflow automatically for tasks that involve exploration of multiple sources of information. The potential of uncertainty in RAG, and other agentic systems that have retrieval and generation components, is not fully explored.

4.3.4 Efficiency vs. effectiveness trade-offs

The generation part of retrieval-augmented generation (The G of RAG) requires, in practice, the use of pre-trained large language models (LLMs). LLMs are notorious for their need of special hardware (i.e., GPUs) for both training/fine-tuning and for running them in production. The costs of running LLMs are orders of magnitude higher than the costs of running a standard retrieval component (The R of RAG). The work group discussed various different costs, such as: Development costs, convenience/cognitive costs, latency costs, query throughput costs, training vs. inference costs, and energy costs. This subsection focuses on the energy costs of training/fine-tuning of research experiments and the energy costs of testing/querying as a focus of efficiency. Our discussions were guided by the two overarching questions: 1) What are best practices of measuring energy-usage of systems? and 2) What efficiency vs. effectiveness trade-off is reasonable? Concretely, we asked ourselves: What are realistic lower bounds on efficiency? When is a small increase in effectiveness not worth it, given a large decrease in efficiency? When is it – the effectiveness – enough, if ever, given both engineering limits and societal limits?

Best practices

Strubell et al. [153] describe a best practice for estimating energy-usage of LLMs, at both training and inference time. A best practice adapted for retrieval experiments by Scells et al. [136] shows that roughly at query time, a standard (BM25) baseline is about 1 order of magnitude more efficient than a standard (LambdaMart) learning-to-rank baseline, and a staggering 5 orders of magnitude more efficient than an LLM (monoBERT) rerankers.[4] Precise measurement of efficiency is hard, and may even need researchers to standardize on hardware [129]. Because the energy demands of systems differ by many orders of magnitudes, rough efficiency estimates suffice for informed trade-off decisions. An important challenge for

[4] The BM25 baseline likely becomes another 2 orders of magnitude more efficient if it retrieves a top 10 instead of a top 1000.

the research community is to start reporting the estimated energy-usage of their experiments and put the measured effectiveness in perspective given their costs. It should be easy to incorporate best practices for estimating the energy costs into research, for instance using standard libraries like CodeCarbon [37].

Reproducibility challenges

Reproducibility of research has mostly focused on measuring the effectiveness. From as early as the 1960's [32], researchers in information retrieval and library science have been carefully constructing open datasets and test collections that help researchers to reproduce results and advance the field. Best practices for research into the information retrieval component of RAG systems are followed in joint evaluation conferences like TREC [166], CLEF [55], NTCIR [79], and FIRE [58]. Similarly, there are benchmarks and leaderboards for LLMs that almost exclusively focus on their effectiveness, such as the Open LLM Leaderboard [118] and LiveBench [173]. Some LLMs, like Bloom, report the energy usage [104] as well as the data used to train the base model and the data used for instruction fine-tuning, following guidelines for describing datasets [59]. Such LLMs are better suited for reproducible open science than other popular "open" models like Llama 3 or 4, and a much better fit than using LLMs via APIs of, for instance, OpenAI or Google. The European Open Source AI Index provides a guide for choosing open LLMs along multiple criteria [98].[5]

4.3.5 Federated Search – Why now?

Federated search [22] addresses a key challenge in information retrieval: accessing distributed data under different ownership, often with controlled access and domain-specific ranking systems. In spite of past research on federated IR for specialized domains like digital libraries [57] and the successful adoption of federated architectures with protocols like OAI-PMH [89], centralized search systems prevail; with the market dominance of Google and Bing for Web search as a primary example. Given the rise of RAG, will the dominating architecture for future search systems remain centralized or is RAG leading us to an alternative, federated search infrastructure, that will be managed and orchestrated by a large number of parties?

Let us consider the key factors triggered by the recent development of LLMs that may lead retrieval-augmented generation systems away from a centralized system architecture:

- **Ownership and Privacy:** Unlike centralized systems, where a single provider controls the indexing and access to data, federated architectures distribute authority across multiple content owners, each retaining control over their own collections and access policies. For RAG specifically, a federated system architecture would ease the deployment of unified information access solutions that access Web data, enterprise data, and personal sources like email and calendar. This decentralization offers greater respect for intellectual property and institutional boundaries, but also introduces challenges: How to ensure consistent privacy guarantees across heterogeneous systems, and how to enforce access restrictions when retrieval is coordinated by third parties.
- **Data Quality and Domain-specific Curation:** Careful curation of the contents of the 'memory' that the retrieval component may access (traditionally referred to as the 'document collection'), translates directly into improved outcomes. The costs of curation can be shared with the community of users of that collection. We foresee this scenario

[5] `https://osai-index.eu`, last accessed September 25th, 2025.

as realistic in professional contexts (e.g., online communities that curate high quality health data[6]), but also for general interest purposes; consider for example topics like "Karate in Germany" that are maintained in public Web directories like Curlie[7] (formerly ODP). Given the expected reduction in data quality and improved redundancy through generative AI, need for curation will increase, while curation costs can decrease due to improved tooling.

- **Efficiency and Ease of Deployment:** Retrieval systems have significantly benefited from the availability of open source embedding and re-ranking models as well as the rapid improvements of vector stores. Wide availability of open source solutions and LLMs has reduced the barrier in setting up RAG endpoints with good quality, democratizing access to this type of technology (leading to "everyone and their mother developing RAG").
- **LLM-based Planning and Routing:** Shallow reasoning techniques in LLMs like chain-of-thought also enable complex search strategies, where multiple queries can be issued, perhaps reformulated based on analysis of query results, in a completely data-driven manner. Although this might be less optimal than formal reasoning methods, this *data-driven reasoning* is easy to integrate in the LLM inference process, and leverages the out of domain generalization capabilities of LLMs.
- **Tool Usage and Standardized Protocols:** Tool usage and the rise of agentic AI triggered the development of interoperability protocols like MCP [48]. While the standardization is largely driven in an ad-hoc manner through industry, adoption is quite high. Considering the planning and routing capabilities of LLMs, it becomes viable and easy to expose RAG systems as tools for LLMs via those protocols. LLMs may also drive the selection of resources in federated search [170].
- **Economic Incentives:** Today's web search engines serve a dual purpose: helping users discover information while also directing traffic to websites, content owners, and publishers. This creates a win-win-win scenario for all parties involved. However, the rise of RAG systems disrupts this mutually beneficial ecosystem. By aggregating and synthesizing information directly, RAG reduces the need for users to visit original sources. Worse still, much of the content used to train LLMs was ingested without fair compensation to publishers. As a result, publishers and content owners now have little economic incentive to allow centralized RAG-based search engines to index and utilize their material. In contrast, a federated RAG system empowers them to retain control over their content while creating new business opportunities. By offering RAG or MCP endpoints, publishers can monetize access to their data and enable LLMs (or other systems) to consume their content – on their own terms.

Aside this list of potential benefits, several factors speak against a federated RAG systems. We see the following potential points against federated RAG systems replacing centralized search:

- **Efficient LLM Inference Infrastructures:** Large-scale centralized infrastructures can optimize inference efficiency through optimized hardware or optimized infrastructure setup, or may use energy sources that are more environmentally friendly. The same optimization might be harder to achieve for federated RAG systems.
- **Limits of Planning and Routing:** Although LLMs have shown powerful planning and routing capabilities, there are upper limits depending on the number of components that can be deployed effectively. While central systems and LLMs can learn routing during training, federated RAG systems require inference time routing and planning, which could reduce the model's capabilities.

[6] `https://medical-data-models.org/` [44]), last accessed September 25th, 2025.
[7] `https://curlie.org/`, last accessed September 25th, 2025

- **Increased Security Risks:** Federated systems per definition operate across system boundaries which makes them (i) more vulnerable to security threats and (ii) harder to protect.
- **Ease of Use of Central Systems:** End-users prefer a single point of access and smooth user-experience. It has to be shown that Federated RAG Systems can achieve the same user-experience as centralized systems and provide an efficient, effective and reliable service.
- **Data Mixing Benefits:** It has been shown that LLMs benefit from mixing training data from various domains [182], and can create relations between different information spaces. It is yet to be seen, whether this property can be retained in a federated system.

Derived Research Questions:

From this analysis we derive the following open research questions for Federated RAG Systems:

- How to do resource selection in federated RAG systems?
- What is the limit (if any) of the number of federated systems participating in a Federated RAG, considering query planning and overall performance?
- Are current protocols sufficient to ensure reliability and quality comparable to centralized systems?
- Will federated systems be more efficient and as effective as centralized systems?
- Could clean and curated data help to make federated RAG *more* effective than centralized systems?
- Can data mixing benefits of large, centralized models be compensated in federated RAG systems?
- How to ensure data privacy in the exchange between components of the federation?

4.4 Societal and Ethical Motivations for Inverting RAG to GAR

Johannes Kiesel (GESIS – Leibniz Institute for the Social Sciences – Köln, DE)
Bhaskar Mitra (Independent Researcher, Tiohtià:ke/Montréal, CA)
Josiane Mothe (INSPE, Université de Toulouse, UT2J, UMR5505 CNRS IRIT, FR)
Heather O'Brien (iSchool, University of British Columbia – Vancouver, CA)
Birte Platow (TU Dresden, ScaDS.AI, DE)
Stefan Voigt (Open Search Foundation – Starnberg, DE)

Positionality Statement

This chapter reflects the discussions and views of a limited group of individuals who worked within a limited time frame. The group responsible for this section of the report is diverse in gender and cultural background with expertise in computer science, social sciences, and humanities with library and information science and theology, but undoubtedly does not represent the rich, diverse perspectives of the full spectrum of people who are impacted by RAG technologies. We also acknowledge the lack of representation from other relevant disciplines and expert communities, who should inform and have an active say in how the relevant technologies are developed.

4.4.1 Introduction

Ever since humans collated written information in books or collections of books (libraries), searching for relevant pieces of information, reading, perceiving and "understanding" has been part of human knowledge gain. Filtering and retrieving was always part of search and selection processes to identify those pieces of information that were required for further processing in information-access tasks. With the increased capacity and wide range use of large language models (LLM) that statistically generate text (and other modes of information) prompted in a written dialog with the user, the classic information retrieval and perception process is changed significantly. In particular, since generative systems are perceived as "intelligent" by many human users.

Editors, publishers, library curators or search engine providers in the "classical" information domains such as libraries, newspapers and the Internet, already had great responsibility to transparently, ethically and reliably structure, process and provide access to information; however, the development and massive use of LLMs has altered and condensed this process of information processing and provision significantly. Information artifacts are generated massively, in an ad-hoc manner and "on the fly", while the original sources are often hardly referenced. A number of ethical, psychological, societal, legal, political and environmental concerns can be associated with this wide use of generative information access technologies. Retrieval-augmented generation, of course, can help to mitigate parts of these concerns, e.g., by incorporating and referencing sources in the generative process and by thus increasing transparency and accountability. However, this is still insufficient. Thus, we suggest a paradigm shift in current information access thinking, from retrieval-augmented generation (RAG) to generation-augmented retrieval (GAR).

While RAG focuses primarily on generating better answers through retrieved context, GAR emphasizes the retrieval process itself as the essential processing step, adding the essential transparency and accountability to information access.[8] In this new "paradigm", the most critical component is no longer generation (G), but retrieval (R): sources and retrieval mechanisms become the transparent foundation of information access and processing, rather than a supplement.

In GAR, the LLMs serve as statistical-generative (dialog based) interface for exploring different information spaces and providing the output in the form the user wants it. This approach emphasizes controlled, systematic exploration of knowledge sources, positioning GAR as a pathway towards more sustainable and transparent information access systems that empower users to engage deeply with source materials rather than consuming ad-hoc statistically generated, non-transparent, but still appealing answers.

In this chapter we collect some key issues, concerns, and opportunities that can be associated at various dimensions and scales with the ubiquitous use of LLMs and RAG systems. From potential impact on individuals to impact on society and even on (geo)politics; from cultural representations of information to the rights of the creators of works of art or information artifacts. We by no means claim to address all socio-ethical dimensions of statistical-generative information access and processing in a complete manner. We report on the main concerns that were raised and discussed during this Dagstuhl Seminar and contribute to a constructive discourse on the ethically and socially sustainable use of systems for accessing and processing information, combining approaches based on "retrieval" and "generation" in a meaningful way. Specifically, this chapter is framed by a discussion of

[8] Our definition of GAR deviates from earlier use [107, 10], in that we include in GAR all information access systems that focus on retrieval, not only those systems in which retrieval is the last step.

knowledge, ethics, and human rights, and its implications for RAG vs. GAR. Next, we offer perspectives on several sociotechnical issues for GAR: scholarly communication, user cognition, emotional and mental well-being, democracy and political discourse, and language and culture.

4.4.2 Knowledge acquisition and ethics

The rise of new information processing modes in LLMs, agentic LLMs and RAG is a disruptive factor in how we store, process and present information. This implicitly raises the question of what information is per se, and what functions it performs for the individual, groups of people and in the overall social system – or rather which functions it should perform in the future.

Interrelation of information and knowledge. In the everyday language of the information society, information is often equated with knowledge. Like information, knowledge then is perceived as a free-flowing impersonal resource. In this perspective knowledge is "stored" traditionally in libraries, later turned into digital libraries, databases and, subsequently nowadays, in LLMs and RAG systems. The idea of "storing" and "collecting" knowledge in analogy to other resources is not incorrect but overlooks one fundamental difference: knowledge has a closer connection to us than other resources like coal or water, for instance. While these resources would still exist if humanity was wiped out, the existence of knowledge is depending on someone (or some group) who knows. As understandable as it may be in our post-Enlightenment, rational knowledge society to identify information with knowledge, it is nevertheless wrong. Information, facts, or data are only building blocks of knowledge that depend on an architect. Without such a counterpart, information is ultimately nothing more than ink marks, electronic marks or tokens. Conversely, however, the following also applies: without access to information and the knowledge based on it, people are not able to shape their lives independently, responsibly, and freely, which touches on fundamental questions of being human.

Interrelation of knowledge and human rights. It is not without reason that the right to education always ranks high in international agreements (see UN Convention on the Rights of the Child, UN Charter for Sustainable Development Goals, etc.). Education is the path to knowledge. Knowledge is the basis for the ability to judge, decide, and act, and is therefore a prerequisite for participation. In this respect, knowledge is also a determining factor for freedom and independence. However, knowledge should not be seen here as a static resource that an individual or group of people simply "acquires" and then uses as they see fit according to their own needs. Knowledge and education should rather be seen as a continuous process through which the individual establishes a positive, dynamic balance in their relationship with the world. The process perspective emphasizes that the acquisition of knowledge cannot be limited to the product "knowledge". Rather, it is a constant state of mind that must be renewed again and again. Knowing that information and knowledge are specifically interrelated (see "interrelation of information and knowledge" above), the question now arises as to how continuity and activity, as core elements of knowledge and education, can be secured for people in the future when information retrieval systems and LLMs enter this process as agents.

What is up for discussion (and modes of discussion). Although ethics is characterized by great openness (as opposed to morality or values), ethical analyses are always location-specific in their consideration of specific constellations and dilemmas. Therefore, it should be clarified in advance which analytical approaches will form the basis of a consideration.

Ethical reflection can be descriptive (descriptive ethics) or evaluative (normative ethics). The intentions (deontological ethics), goals (teleological ethics), or consequences (consequentialist ethics) can be the subject of consideration. If the focus is on the consequences, it must still be clarified whether this is done from an individual ethical perspective or a social ethical perspective, i.e., whether the starting point is the individual or groups of people or society as a whole.

In terms of the evaluative parameters of the analysis, ethical analyses are flexible in terms of specific topics, but are generally based on universal values such as human dignity and other general concepts. In contrast, an analysis based on moral aspects is more focused and concrete. By morality, we mean the entirety of norms that apply in a specific community. Morality is based on the insight that there are certain codes of conduct and judgment that enable constructive coexistence. Norms and conventions, on the other hand, are habit-based concretizations of morality. Laws are ultimately the institutionalized "final version" and highest form of concretization.

For the present analysis, an ethical approach is chosen that largely disregards moral perspectives. Furthermore, the possible consequences of LLM and RAG in knowledge systems are discussed, both from an individual and a social-ethical perspective (cf. micro, meso, and macro levels). Against the backdrop of such an analysis design, the following questions arise as ethically relevant issues:

Key questions. To what extent must architects who transform information into knowledge be human – both from an individual as well as from a societal point of view? What happens to the individual/to groups/society when humans play less of a role in the transformation of information into knowledge? And what does this mean for knowledge systems themselves? And finally: What dynamics might RAG systems in their various forms bring to the upcoming development?

Societal scale. As seen above: without a mind to access and transform information according to its individual circumstances, whatever is stored in libraries, databases or LLMs will remain unstructured and, in the long run, static. Therefore, even a comparatively small share of an individual is indispensable in the transformation process from the perspective of the system. However, it is unclear to what extent and to what degree this effect must be reflected from the perspective of the system. Is a prompt sufficient? A query? What sources should an individual person access and process in order to fulfill their function as an information transformer and knowledge architect? To put it bluntly: Should the individual user primarily receive synthetic sources from the LLM, or should they (based on traditional knowledge generation methods) perform source work and assemble these as building blocks with the help of an LLM that supports them as an interface?

From an individual and ethical micro perspective, freedom, self-determination, independence and participation manifest themselves as learning and the ability to learn, which must be upheld as fundamental anthropological constants, especially in times of generative AI. In this sense, information is not only a prerequisite for the diverse realization of human dignity, it must also prove itself in context and at the same time be understood as a means by which humans themselves continuously fulfill their constitution as learning entities. This implies necessarily intensive and process-emphasizing involvement with information beyond simple prompting.

As seen above, knowledge belongs to some individual but at the same time it belongs to particular groups. What is more: the knowledge of a group may go beyond the knowledge of its individual members. Therefore, it is crucial to consider the above-mentioned aspects in a differentiated manner in the context of space and time (meso perspective) as they manifest themselves differently depending on the particular cultural context.

Since knowledge always serves the purpose of establishing a positive reciprocal relationship between individuals, the social environment, and society as such, society as a whole must also be considered from a macro perspective. It is obvious that the different approaches of these groups to information leave their own mark, which again corresponds to the idea of successful architecture.

The shaping of information corpora is therefore carried out for good reason and with justification by various "architects". However, it remains unclear which tools should be provided to these architects so that they can perform their design work (from information to knowledge) and at the same time fulfill their aspirations (knowledge as access to the world).

RAG vs. GAR from an ethical perspective. First, it should be noted that the various forms of generative AI open up unprecedented opportunities for individuals, groups, and societies to acquire knowledge. Against this backdrop, it is to be expected that knowledge will potentially be strengthened in all its positive effects (participation, freedom, etc.). However, this expectation must also be viewed alongside the possibility of limitations or even negative effects.

In this context, the decisive factor is likely to be the way in which people participate in the transformative work that turns information into knowledge. An excess of generative processes would reduce the diversity of sources and positions and, as a result, orientation and the ability to tolerate ambiguity and orientation as complex knowledge bases for accessing the world. As a result, freedom and independence would also be at least potentially endangered. In addition, it would reduce human involvement insofar as the product of information processing would become dominant over the encounter process. Human learning would not be taken seriously, and the transformational work of humans would be reduced to mere consumption in the long run.

In contrast, retrieval strengthens the aspects mentioned above. The generative elements in the transformation process could support communication, profiling, and depth of the transformation process without, however, contributing to the actual function of transforming information into knowledge by creating a relationship between the enormous information base and a knowledgeable person. Generation would then no longer be the constitutive factor in this process (RAG), but rather the connotative mediator (GAR).

4.4.3 Scholarly Communication and Publishing

The Internet and social media (e.g., Twitter/X, YouTube) have transformed how scholars engage with each other and non-academics, highlighting how topics (e.g., Covid-19), communication style, and digital platform and audience factors contribute to or deter engagement with scholarship [24, 25, 65]. At the same time, memory institutions like libraries have moved from catalogs to discovery systems that facilitate access to a range of in-house (e.g., institutional repositories, open access digitized collections) and proprietary materials purchased from publishers and vendors [18]. Some may even provide altmetrics, bringing mentions about the item on various social media platforms into item-level descriptions. This has resulted in a "one stop shop" for users affiliated with an institution to locate and retrieve books, journal articles, archival and special collections, academic databases, etc.

Generative artificial intelligence (Generative AI) tools (e.g., ChatGPT, Google Gemini, Perplexity AI, Le Chat) are disrupting scholarship, bringing new opportunities and challenges to effective research dissemination, uptake and use. To date, research has focused on the ethical and transparent uses of Generative AI in research production, e.g., authorship, peer review, policies to curb academic misconduct [36, 68, 92]. Generative AI outputs have the

propensity to produce "fake but convincing articles" [36] (p. 235), false citations, and biased outputs [47, 92], and to separate content from its sources [137], all of which could threaten the legitimacy of research [68] and its potential to impact society through, for instance, policy development, decision making, etc.

Shaping Scholarly Communication Systems and Products. There are many actors in scholarly communication and publishing: scholars (who also act as peer reviewers, editors, and mentors to the next generation of scholars), publishers, funders, GLAM institutions (galleries, libraries, archives, and museums), scholarly and professional associations, and commercial enterprises that make products to assist scholars in tasks like writing, e.g., Claude, Writefull, or provide current awareness and dissemination services, e.g., ResearchGate, Google Scholar. Each of these actors create, promote, and use different information systems and formats or genres e.g., journal article, abstract.

Scholarly information systems may be reinvented in the coming years (and, arguable, this is already happening). GAR has the potential to break down silos between standalone systems. Currently, the searcher must navigate their own library discovery system, go to a search engine like Google, or a known web-based repository, e.g., arXiv, to initiate a search. GAR systems could provide seamless access to relevant content held by multiple scholarly information systems and in multiple formats (e.g., books, journal articles, data sets, multimedia). However, this requires us to consider questions around: copyright legislation, which differs by geographical jurisdiction; authors' intellectual property rights, including works created at different points in time under different (non-GenAI) conditions; protections for cultural heritage materials that may be held by institutions but rightfully belong to other people or groups, e.g., stolen Indigenous artifacts (cf., [97, 96]); and the role of GLAM institutions, which currently operate as intermediaries in the scholarly ecosystem. *There are opportunities for GAR researchers to work with GLAM institutions and publishers to shape information access and negotiate fair, transparent use of collections.*

Scholarly formats or genres are constantly evolving (e.g., the shift from print to digital journals), but the history of scholarly communication demonstrates how the materiality of research (i.e., form and function) shapes social, disciplinary, and institutional norms and practices [39, 20]. Currently, there are tensions in how scholars from different disciplines are approaching the use of GenAI tools in education and research settings, e.g., policies around tool use on course syllabi, disclosure statements on conference submissions, and whether this use threatens research legitimacy and academic integrity. Traditional university promotion and tenure systems are entrenched in academic attribution, e.g., bibliometrics or ways to "count" (and therefore legitimize) authors' contributions. Therefore, it is important for GAR systems to support the accurate attribution of ideas and scholarly contributions, and to ensure that research remains a conversation – not just between humans and GenAI tools – but among researchers and public audiences. Scholars must be able to defend and verify their claims, correct misinformation or challenge harmful interpretations of their work. Finally, it is predicted that "research will be published in a way that can be easily read by machines rather than humans" [36] (p. 236). *This raises questions about what will constitute scholarly products and how norms around their construction will be negotiated within and across disciplines, and with GAR tools.*

4.4.4 Cognitive Processes in Information Seeking

Information behavior comprises information needs, seeking and use. Information retrieval systems are designed to assist users negotiate information needs, refine queries and search strategies, and assess the relevance of the results ("10 blue links") through various affordances,

e.g., controlled vocabulary, search histories, links to related, relevant content, etc. [174]. Information systems are designed to support users to complete simple, factual and more complex tasks to accomplish a goal. Information behavior recognizes that affect, behavior, and cognition play a role in information interaction, and that these are intertwined. For instance, newspaper headings grab our attention (cognition) by triggering our emotions (affect), causing us to click and read (behavior). This section looks specifically at cognition.

Cognition in information behavior refers broadly to how people perceive and attend to, learn, process, organize and store information. Information systems require cognitive skills to use; for example, finding a document saved to your computer requires you to remember details like its title, key words, or location to facilitate retrieval. In other cases, information systems can compensate for limited cognitive capacity; for instance, a well designed product menu and facets can help searchers recognize and narrow down what they are looking for. There is no "typical" user (though systems are often designed as though this is the case). Cognitive abilities are highly varied, with factors such as developmental stage, reading level, environment, neurodiversity, and others playing a role.

Generative AI systems are raising alarms for potential negative impacts on human cognition. Cognitive offloading refers to using tools like note pads, Google maps, or calculators to reduce our cognitive load [60]. Offloading to a Generative AI tool could be positive; if a user delegates some tasks to a tool, they may make space for higher order problem solving. However, studies suggest the opposite is true. Gerlich [60], in their survey of 600+ UK residents over the age of 17, observed a link between trust in Generative AI systems, cognitive offloading and reduced critical thinking. Another study by [86] recruited 50+ university students (aged 18-39) to be part of one of three groups: LLM, search engine, or brain-only. Participants' brain activity was monitored during three sessions where they wrote essays, either utilizing LLM or search tools or not; the essays were evaluated by two teachers using specific metrics. Based on the electroencephalography activities, the researchers concluded that the brain-only group showed more neural connectivity, indicating the utilization of more cognitive resources for the task, and more integrative activities at low frequencies, "possibly reflecting deeper encoding of context and an ongoing integration of non-verbal memory and emotional content into their writing" (p. 86).

These two studies represent emerging work, and more empirical work is needed to understand how Generative AI tools are changing our cognitive processes. Are these changes perpetuating cognitive decline or passive, uncritical information reception that could increase vulnerabilities for misinformation? Or do they present opportunities to build cognitive skills in key areas. The traditional search process consists of formulating the search, selecting sources, and interacting with sources [164]. If GAR systems ease the cognitive load of formulating the search and selecting sources, *what possibilities are there for prompting deeper, more critical engagement with sources if we move away from presenting outputs as "answers"? How can educators and information professionals help people to learn cognitive skills to prepare them to use GAR systems?*

4.4.5 Emotional and Mental Well-Being

As a relatively new technology, the long-term impact of LLM-powered conversational systems on individual's emotional and mental health has not been studied adequately. New research is currently emerging that calls for much more measured deployment and adoption of this technology till its effect on mental health is better understood. The concerns here are not just about the generative AI itself, but also about how for-profit institutions may specifically train these models to maximize usage – *e.g.*, through application of sycophancy [140],

anthropomorphic behavior [29], and emotional manipulation [38]. There are serious emerging concerns about these systems contributing to digital addiction [179, 143, 190, 108, 3] and even causing paranoia and delusions that has come to be termed as "AI psychosis" [46, 85, 103, 53]. It is particularly concerning that several cases of self-harm [12, 178, 84, 74, 183, 157] and accidental death [67] have been linked to chatbot usage. Reports are also emerging on cases of chatbot usage linked to negative impacts on personal and romantic relationships [45]. It is imperative that extensive studies are urgently performed to better understand the impact of these technologies on people's emotional and mental well-being as well their interpersonal relationships, and that the technology is better regulated in light of those findings.

4.4.6 Democracy and Political Discourse

Online information access plays an important role in our collective sense-making of our place and relationships in this world, and mediates critical political discourse in society. We must thoughtfully consider how the applications of RAG for information access may shape the future of democracy. While these technologies may have much to offer in supporting future democratic processes, we must also critically examine the potential risks. We can already observe some of these concerns being reflected in public opinion in the context of the undue influence that chatbots are exerting to shape political narratives [82, 119, 101] and with respect to the risks of eroding public trust in democracy when chatbots are deployed irresponsibly by elected government representatives [106, 175]. Concerns about the impact of RAG (and more broadly generative AI) on democracy and political discourse includes misinformation and confirmation bias, AI persuasion, reduced transparency, and dependency on provides that we discuss next.

Misinformation and confirmation bias. Generative AI models are prone to producing factual inaccuracies in their outputs that may contribute to public misinformation [189]. This is further exacerbated by the fact that the application of generative AI models for generating concise summaries of information from retrieved artifacts shifts the responsibility of inspecting the information in the documents and assessing their relevance, trustworthiness, and surrounding context from the users to the AI models and disincentivizes users from developing critical cognitive skills necessary to distinguish between trustworthy and untrustworthy information [112].

AI sycophancy [140] may also encourage confirmation bias in users and lock them in echo chambers [141]. Democracy requires that all citizens, including those with opposing values, have access to the same shared facts. Otherwise, this may contribute towards further social decohesion[9].

Application of RAG for information access must be accompanied by appropriate safeguards to ensure that citizens have access to trustworthy information and a shared basis of facts, and do not negatively impact public information literacy.

Persuasion and Manipulation. Technical progress in approaches for aligning generative AI models towards specific values have been crucial in reducing harmful outputs. However, the same mechanisms may also be abused by system owners by combining them with massive amounts of user behavior data from surveillance capitalism [192] and generative AI's capability to produce persuasive language and visualizations [112, 21, 23, 49, 125] to

[9] We intentionally do not frame this as increasing polarization as that framing implicitly assumes that different sides of political discourse have equal merit which amounts to algorithmic bothsidism [111].

censor and manipulate public opinion [82, 119, 101]. Such concentration of power to influence public discourse can pose serious threats to democratic processes. To address such risks, we must develop governance and auditing frameworks that ensure that the development and deployment of these technologies are performed under appropriate democratic oversight and are in alignment with our democratic goals.

We must also ensure that the application of RAG for information access does not further exacerbate user surveillance by encouraging users to share more personal information when engaging these systems in conversations.

Reduced Transparency. Democracy is based on transparent decision-making as a prerequisite for holding decision-makers accountable for their decisions. However, current LLM systems are not transparent and are not accountable for their results. At the individual level, citizens cannot trace where the information they use comes from, making them more susceptible to falling for misinformation. At the societal level, decision-makers cannot base their decisions on the results of opaque and unaccountable LLMs without violating the principle of transparency. Information access systems that focus on retrieval can mitigate this problem to some extent by providing users with results and their sources on which they can base their decisions or opinions on, rather than providing direct answers.

Dependency on AI Providers. Integrating a tool into workflows comes with a certain degree of dependence on that tool. In decision-making processes, such dependence can make people more susceptible to manipulation and cognitive biases (see above), as the incentive to rely on the tool is greater. Lock-in effects, such as the inability to transfer ones chat history from one LLM to another, exacerbate this risk, as users become dependent even on specific providers rather than on the technology in general. At the individual level, citizens may stick with (certain) LLMs even though they are aware of better alternatives. At the societal level, decision-makers may become dependent on LLM providers, giving them undue power over the decision-making process. Information access systems that focus on retrieval can mitigate this problem to some extent, as they do not take over the process of forming an opinion or decision (but merely support it).

At the heart of these questions, lies the recognition of the fact that technological frames like RAG and GAR – as well as the decision to adopt one over the other – are all saliently political. To responsibly engage these technologies in our democratic processes and political discourse necessitates that we in turn also open up the sociotechnical imaginaries [111] that motivate their development as well as who gets to shape their realization also to social deliberation and democratic critique.

4.4.7 Cultural Representations

Worldwide vs Language Centric Information Access. On the search perspective, on traditional search engines, retrieval is mainly language-based: a query in French retrieves URLs / documents from resources in French. This is because of the principle of keyword search, where the query words have to match the document words, creating linguistic silos. This limitation is addressed by the architectures that combine search and LLM capabilities.

On the retrieval side, the frontiers between languages can be blurred thanks to the semantic representation. GAR offers the opportunity to retrieve information in different languages. Multilingual LLMs embed information in different languages and in Europe, the words “democracy”, “démocratie” (FR), “democracia” (ES), “democrazia” (IT), “Demokratie” (DE), “democratie” (NL), “democracia” (OC) are supposed to cover the same concept of democracy for example. On the generation side, the user can get an output in a language different from the language of the retrieved sources/documents.

There is also the opportunity to have access to resources from different cultures through different languages [130]. Multilingual LLMs embed information from different perspectives [1]. As an example, Chinese medicine could become much more accessible for Westerners because the different pharmacopoeias share the common goal of providing solutions to human ailments for example. On the other hand, LLMs are not culturally neutral [114, 83]. The implicit assumption that the embedding space adequately captures traditional Chinese medical concepts may be wrong. Rather, concepts are probably mapped through the dominant -most frequent- culture conveyed by the documents used to train the models [138]. The risk is that we perceive that different cultural perspectives are now accessible while we are just accessing culturally-specific information filtered through the dominant cultural framework of the training data.

At the individual scale, individuals may have easier access to information previously limited by language barriers or by the need to combine too many tools or steps – multiple searches in multiple languages plus translation. But individuals access decontextualised information and may not have the cultural background needed to interpret it appropriately.

At the community scale, educational/research/etc. communities could develop more inclusive and less language/disciplinary/cultural silo-based curriculae/research/etc. However, systems may prioritize some views/cultures.

At the societal scale, there is the opportunity to preserve minority knowledge/culture/languages for not much additional costs if we avoid digital colonization and a dominant culture being represented.

Knowledge Authority vs Knowledge Diversity. At the production stage, information authority is established through editorial workflows which consists of a sequence of tasks and responsibilities in the creation, editing, and publication of content. Information authority can also be given through the official status that a government assigns to information, e.g., through legal force, the credibility of authorship, recognition by the domain or community, topicality, etc. Search engines reinforce this by giving more weight to some sources, e.g. through PageRank-based ranking, topicality consideration, etc. The IR community has developed many algorithms and techniques that prevail for ordering the retrieved results. On top of that, the engine financing model has led to additional ranking strategies, making the ranking non-transparent to users in the end. In RAG/GAR systems, the risk is that things are getting worse, there are no safeguards on how sources are prioritized.

RAG systems potentially democratize knowledge by treating all embedded documents as equally represented based on their semantic content, rather than institutional or other authority; although, not all languages/cultures are equally represented in embedding spaces. And some knowledge communities do not respond positively to digitization, specially minority communities.

But the retrieval part of RAG/GAR systems plays a crucial role in the drift of the generation part. And it should be more transparent, specially on how sources are selected and how many are used. A recent study analyzes [10] how ChatGPT sorts through the vast array of sources available on the Internet on the news summary task (here "What happened yesterday in Gaza?" in French from France on ChatGPT4.1). According to the report, "Considering only publications from the last 24 hours, Google News offers between 25 and 60 publications from the news media every day (45 on average), while ChatGPT offers between 5 and 10 (7.8 on average) – about six times less." The underlying mechanisms are black

[10] https://larevuedesmedias.ina.fr/chatgpt-gaza-google-actualites-information-sources

boxes. Either it should be more transparent or driven more by the user rather than keeping an automatic driver of the system. The RAG/GAR principle creates new opportunities for knowledge diversity but raises critical questions about validation at the different stages.

In many domains the "truth" or maybe rather the authoritative information may depend on the culture/location; sometimes for the good, sometimes for the bad.

At the individual scale, users face the challenge of evaluating unfamiliar knowledge systems without traditional authority markers. Relying on the system only could be even more risky.

At the community scale, institutions and communities should decide whether to maintain existing authority structures or develop new frameworks that recognize new and maybe more diverse forms of authorities.

At the societal scale, the risk can be of either enforcing dominant cultural authority standards globally while it could be useful to create systems where all knowledge claims are treated as equally valid.

4.4.8 Conclusion

Critical deliberations on the societal implications of technology can help us to free ourselves from technodeterministic thinking and encourage the community to actively assert their agency to shape technologies to affect desired positive social outcomes. The proposed shift from "RAG" to "GAR" is our attempt to refocus the discussion on our society's information access needs rather than on the new technology itself. This needs to be accompanied by critical conversations within the IR community to explicate and reimagine our sociotechnical imaginaries [111], deliberate on the values we want to encode in our design (*e.g.*, democratic and emancipatory [161]), and view these platforms from diverse perspectives (*e.g.*, as critical infrastructure that should be open[11] and democratically governed). These conversations must not be restricted only to IR researchers and practitioners, but must encourage cross-disciplinary participation by scholars and experts from outside of IR.

This write-up is only the first step. We need to follow up with other concrete actions geared towards transforming not just what the IR community works on, but how it approaches the questions and with what goals. We must reflect on how the arguments presented in this section intersects with the conversations around other proposals, such as the #FreeWebSearch charter[12]. And we should be clear-eyed about the social, political, and cultural context in which this work is embedded and the incentive structures that continues to shape our work.

[11] `https://freewebsearch.org/`
[12] `https://freewebsearch.org/en/charter/`

4.5 An Unexamined RAG Is Not Worth Interrogating

Niklas Deckers (Universität Kassel and hessian.AI, DE)
Laura Dietz (University of New Hampshire – Durham, USA)
Maik Fröbe (Friedrich-Schiller-Universität Jena, DE)
Wojciech Kusa (NASK National Research Institute – Warsaw, PL)
Mark Sanderson (RMIT University – Melbourne, AU)

4.5.1 Motivation

Modern tool chains make it relatively easy to quickly develop retrieval-augmented generation (RAG) systems. Hence, many RAG systems exist, for different applications, different corpora, and in different configurations. However, it is difficult to know which of these RAG systems is best, and for which contexts each is most appropriate. Therefore, we aim to answer two questions:

1. When RAG works, how can we be certain that it works?
2. When RAG fails, how can we find this out?

While there is extensive work on the evaluation of RAG, we believe that RAG as a technology is so new, it's still unclear what it is that we need to evaluate in a RAG system. We remain unconvinced that even state-of-the-art evaluation paradigms [188, 51, 134] have, as yet, demonstrated that they can measure all that we need to understand in RAG.

The paradigm shift that RAG presents to information access technology requires a corresponding shift in evaluation procedures. We have seen such shifts in the past: the Cranfield methodology for retrieval evaluation was adjusted to scale to larger document collections [167]. With RAG, we find ourselves in a similar situation that requires novel evaluation. Table 2 provides a comparison of aspects that we hypothesize might change substantially between information retrieval (IR) and RAG evaluation for online and offline experiments.

For online experiments, explicit click feedback in IR was exploited for learning at scale. With the reduction in clicks now found in RAG systems, implicit feedback might be more difficult to incorporate. As users express themselves in natural language, RAG systems can be personalized more easily as they can be explicitly prompted to follow users' instructions. For offline experiments, traditional retrieval evaluation had the advantage that the document collection was stable and the relevance judgments were per document so that benchmarks could be re-used to evaluate new systems over a long period. For RAG evaluations, it might be that evaluation benchmarks are more difficult to re-use as new RAG systems might produce unseen responses – even for the same query and collection. These unseen responses are possibly of equal quality, but utilize different language, different order, and different sources. This renders purely manual evaluations impractical. While it is still possible to apply manual evaluations to individual system responses (frozen in time), the costs are generally considered too high, given that resulting research is not reproducible and the test collections cannot be reused for developing the next generation of systems.

The rapid development of new innovations on the basis of LLMs and RAG must be met with novel approaches for measuring quality. Although prompt-based LLM-as-a-judge approaches are popular, they are vulnerable to issues such as LLM narcissism, circularity, and benchmark memorization, which lead to measurement errors [41].

Table 2 Asymmetry of retrieval and RAG evaluation across Offline and Online experiments.

	IR	RAG
(a) Offline Experiments: Explicit feedback at low quantity of high quality		
Longevity of Benchmarks	High	Low
Feedback Granularity	Per Document	Per System
(b) Online Experiments: Implicit feedback at high quantity of low quality		
Learning from Feedback	Easy	Difficult
Personalization	Difficult (Implicit)	Easy (Explicit)
(c) Both (Online and Offline)		
Evaluation Methodology	Established	Exploration
Feedback Modalities	Few	Many

We briefly detail current RAG evaluation approaches and outline research challenges. Finally, to reverse the tempting trend of removing human feedback from the evaluation setup, we suggest development of open source tooling that supports exploratory examination of RAG systems. The hope is that by supporting the rapid identification of common failure modes, we aid the subsequent development of improved evaluation paradigms to measure the progress of innovation.

4.5.2 Research Gap

In 2023, there was a paradigm shift in IR. Before the advent of LLMs, it was difficult to build systems that summarized retrieved content into fluent language or supported conversational clarifications from users. Nowadays, such challenges have promising solutions. As a result, peoples' expectations of retrieval have grown. It is not sufficient anymore to merely obtain information that contains topically relevant text among many non-relevant sections. Instead, searchers expect that retrieved information is concise and on-point, while relating the content to the user's task, and potentially, their context.

4.5.2.1 State-Of-The-Art in RAG Evaluation

Distinct from IR, RAG tends to produce a different response every time the system is run – even for the same query and corpus. Initial attempts of addressing the variability of system responses with ROUGE/METEOR [99, 16], BERTScore [187], etc. have largely been abandoned, due to their inability to generalize to the variety of responses.

As a result, many researchers have abandoned the idea of grounding evaluation with humans, and moved towards fully automatic evaluation with the aid of LLM judges [188, 51]. Hybrid evaluation solutions have been proposed [134]. However, because of low cost and versatility, the adoption of fully automatic evaluation is popular despite the many already stated issues. Evaluation systems are even adopting an LLM judge idea to generate training data, which risks circularity, overfitting, and self-training collapse [41].

One compromise that lets humans contribute to a semi-automatic evaluation paradigm is an evaluation that focuses on a set of information nuggets or grading rubrics. The paradigm gives a human judge the ability to specify which pieces of information must, should, or could appear in the response of an excellent system. In this way the judge defines the required content. Such pyramid-based evaluation systems [100] have long been discussed, but adoption

was largely hindered by the cumbersome manual process of aligning nuggets to text, requiring more work than a straight-up relevance judgment. By leveraging LLMs for the linguistic match between each nugget and corresponding passages in the system response, this hurdle has been overcome [135, 52, 128].

The main advantage of RAG systems over purely generative systems is the grounding of information in high-quality documents, increasing the trust that information is factually correct. One important part of RAG evaluation also verifies the faithfulness of citations.

Evaluation systems such as ARGUE and AutoArgue [110, 168] combine multiple metrics, such as nugget coverage, verification of citation support, and relevance of cited documents. While fulfilling these criteria is necessary for a high-quality system response, researchers realize that this list is not complete.

4.5.3 Proposed Research Directions

In order to develop the next generation of evaluation methods, we need to first understand the scope of open issues that state-of-the-art RAG systems should be measured against. A number of reviews of RAG evaluation have organized the issues differently. Three reviews/reflections [4, 5, 161] considered evaluation from the same two perspectives: evaluation of RAG systems and how generative AI can supplement the evaluation process itself including the possibility of creating digital twins, simulating a range of human interactions with information access systems. A more recent review identified four areas: evaluation of benchmark datasets, of indexing, of the retriever, and finally evaluation of the generator [19]. Recently, novel evaluation methodologies, such as red teaming [7], have been applied to RAG systems.

During the seminar, we examined what we saw as overlooked aspects of evaluation: how to encourage the users of RAG systems to evaluate the content that they are shown and how to support evaluators of RAG systems to interrogate the correctness of answers.

4.5.3.1 Encouraging Users to Be Evaluators

LLMs are trained to produce confident and assertive language that is designed to convince humans – even when the level of confidence does not match the content quality. This property of LLMs can be exploited by a range of stakeholders for advertisement, opinionated debate, job search and hiring, and management-related messaging. Therefore, users might not realize the quality (or lack thereof) of the provided information. Means of encouraging the users of RAG systems to more fully interrogate answers is critical. The current approach used by many RAG systems of displaying a warning that "AI responses may include mistakes" is not a sufficient solution. As RAG systems improve, we need to detect whenever quality criteria have been "addressed" (e.g. fluency), and whenever new criteria arise (e.g. plausibility vs correctness) that determine the difference between best and mediocre systems.

Given the tendency of people to over-rely on LLMs [56], we need to consider what might motivate users to carefully examine the validity of RAG system responses. This would enable us to design gamified evaluation methods [191] and vigilance tests [31]. Leveraging interaction data. With the rise of AI overviews emanating from RAG systems, there has been a measured drop in click data [27], which have been an essential signal for online evaluation. Hence, there is a pressing need for collecting new forms of user interaction data for new approaches towards online evaluation. An example of such a novel problem is the emergence of sycophancy in the output of generative systems [30].

While this categorization of RAG systems is a good start, the novelty of RAG is such that there may be uses and impacts of this technology that are as yet unknown. Charting such a landscape requires a different approach. There are research methodologies that can

help establish such a landscape. Qualitative approaches that engage with users and their data in an exploratory manner can be used to identify potential research problems that can be examined afterwards.

4.5.3.2 Supporting RAG System Evaluators

In the past, measuring quality was largely a question of costs: how much annotator time to purchase? Do annotators require specialized knowledge (e.g., biomedical experts)? Are crowdworkers sufficient? How many topics and responses of the assessment pool to judge?

A general issue that has always affected traditional IR evaluation is that human judges can only measure "recall" with respect to information provided in an assessment pool. Relying on the world knowledge of judges when assessing the relevance of non-contextualized facts, often led to non-ideal results [40]. In a world where RAG systems are all providing – at a minimum – superficially correct appearing responses, even a careful assessor might not be able to notice omitted or false information. Even when citations are verified for faithfulness, the citation of documents with incorrect or unhelpful information yields RAG responses of low quality.

It will be difficult (or maybe impossible) to measure the quality with human judges or LLMs alike. Moreover, any evaluation that is purely based on LLM judges, will likely be susceptible to the same issues as the response-generating LLM itself. We suggest to invest more research into how to reveal subtle quality differences between RAG/IR systems for such situations without obvious solutions. We hypothesize that the solution will likely involve a collaboration between humans and LLMs. A complementary approach is task-based evaluation, in which humans attempt difficult tasks with RAG assistance, revealing system quality through its impact on real problem-solving.

We need more research to identify the properties via which quality differences manifest and design experimental setups to reveal such differences. This will require to think which information human judges and LLM judges need to be presented with to focus on judging such properties. It also will require to define human-LLM co-working processes that effectively compensate for each other's failure modes.

4.5.4 Talmud-Style Interfaces to Discuss RAG Response Quality

As issues get addressed and RAG systems evolve, we need an easy way to identify and prioritize the most urgent failure modes of RAG systems. We assume that human system developers and judges need support to manually inspect RAG system responses and discuss the pros, cons, and failures in a systematic fashion. This support may come through the way that this information is presented. During the seminar, we got inspiration from religious and ethical discourses that have experience of keeping discussions on important topics ongoing for hundreds of years. Therefore, those ideas inspired our discussions.

As an approach, we draw on the analogy of how the Talmud is used for discussion over hundreds of years [87] and apply it to RAG quality assessment interfaces. The Talmud is composed of two parts: The *Mishnah* contains the laws that are derived from the Torah, and the *Gemara* provides interpretation and discussion of those laws. These are presented in an integrated way, with the text of the Mishna shown in the center and the discussion of the Gemara presented as annotations in the margins. The interpretations given here are not considered absolute; they reflect different perspectives.

Following this structure, we envision a RAG annotation interface in which the central part contains a user query and the RAG system response. In the margins, retrieved snippets provide the evidential basis for the generation, supporting or contextualizing its content.

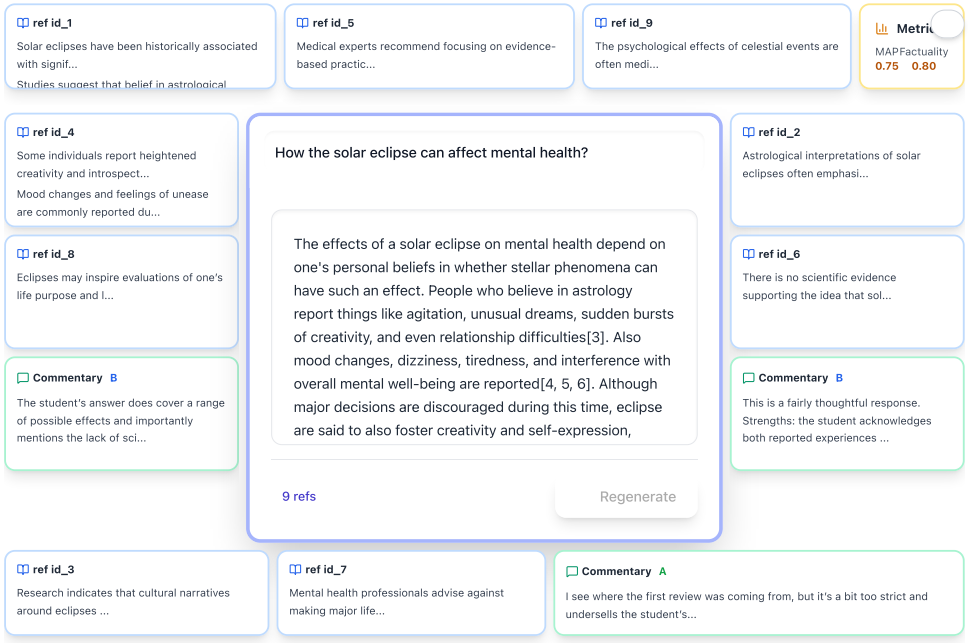

Figure 9 Overview of the Talmud-inspired user interface for RAG quality discussions. At the center of the interface is a user query accompanied by a single model-generated response. Surrounding it are the key components of the quality discussion: human and LLM discussion of pros/cons, the original snippets, and automated evaluation metric scores.

Beyond the snippets, a further layer of commentary – resembling the *Gemara* – may include discussion among human experts aided by LLM-as-a-judge assessments, offering interpretative assessments of the answer's quality. Additionally, aggregated evaluation metrics can be presented as a meta-view of system performance. Figure 9 illustrates the vision of such a Talmud-inspired design.

An alternative configuration of the interface would invert the layering. Instead of focusing on the RAG response, we could place a cited document in the center (analogous to the *Mishnah*), with multiple system generations arranged as form of discussion (analogous to the *Gemara*).

In the Talmudic tradition, such higher-order reflections are captured in later super-commentaries (e.g., by the *Rishonim* and *Acharonim*), which offer critical perspectives on both the Mishnah and the Gemara. In a similar way, the meta-assessment layer in our envisioned system functions as a set of super-commentaries, guiding assessors in contextualizing, comparing, and ultimately judging the value of different generations.

Building on this design analogy, the proposed system can also be situated within established IR evaluation traditions. Our "Talmud-inspired" interface reframes evaluation by highlighting both the core unit of analysis (query and system response) and the multi-layered interpretative context (retrieved snippets, intermediate-steps, commentaries, and meta-assessments). Such a system could be used by several user groups:

- Assessors who judge the quality of system outcomes across multiple models are benefiting from the structured presentation of supporting evidence, commentaries, and aggregated metrics. LLMs can help to maintain consistency among commentary by identifying when earlier comments apply to new responses.

- Evaluation developers, who study the reliability of human and machine assessors in layered evaluation settings, in addition to overseeing the assessment process.
- Professional searchers in domain-specific applications (e.g., legal, medical, scientific retrieval), who require both direct answers and transparent supporting evidence to make informed decisions.
- Educators and learners (e.g., medical students, legal clerks) who study the discourse in complex systems may benefit from an integrated presentation of both authoritative sources (Mishnah), multiple perspectives and interpretations (Gemara), as well as higher-level critiques (super-commentaries).
- Users who prefer a presentation of the structured dialogical space [13] over a linear ranking by relevance [184, 78].

4.5.5 Broader Impact

There are two main areas of impact for the research agenda that is described: researchers working on RAG systems and users of RAG technology.

Current methods for assessing these systems are still much in their infancy. Methodologies such as the Cranfield Paradigm – which has served the information retrieval community exceptionally well for over six decades – appears to be reaching a distinct limit: just simply focusing on the relevance of objects is no longer sufficient to understand the value of the output of a generative information access system, such as RAG. We are calling for further research on how to better assess the quality of the RAG systems. This includes systems being developed currently as well as the systems that will be designed in the future.

The potential impacts of the research agenda we detail here are broad. Features of RAG technology are becoming increasingly visible with search engines, such as Google or Bing, which are starting to roll out the AI summaries in a large number of search results [27]. Many desktop applications recently incorporated their own "AI feature". This, however, is only the tip of the iceberg. RAG is proving to be an extremely popular technology. It is being deployed by a wide range of organizations for the management, search, summarization, and manipulation of documents. While well funded companies such as Google can afford to spend substantial sums of money to ensure their RAG systems are accurate, all organizations using RAG will also want to ensure that the systems they manage are operating successfully. Our proposed research agenda will benefit all stakeholders of deployed RAG systems.

5 Answers to "Will RAG replace ranked search for end users?"

Vote	**Reasoning** (Colors: negative, positive)
No	*No reason given*
No	For navigational queries, RAG is plain stupid (Google summaries sometimes suggest "doing a Google search" for such queries, which is hilarious). RAG without links is no search at all, and mostly useless, unless you already know the answer (known item "search").
No	RAG systems are usually not transparent about their retrieval results before the generation step. As long as that doesn't change, fact checking will always be done post-hoc in an ad-hoc search engine.
No	It would be incredibly problematic if we can only imagine futures where all information access is intermediated by LLMs that concentrates absolute power over what and how information is presented, that serves as a mechanism for mass manipulation of public opinion, in the hands of the platform owners. While there are significant societal concerns with any use of LLMs, the notion that they completely replace all forms of information access modalities is particularly dangerous.
No	I assume that end users (depending on the task) would still like the option to select the sources they value more.
No	I think there will still be a need for keyword-to-ranked-documents type of systems, e.g., for navigational queries.
No	No, because for some scenarios, expressing the search intent verbally with sufficient precision is more difficult than quickly skimming search snippets, and/or the generated text is structured less intuitively than a 10-blue-links format.
No	Rather than replace it, I believe that they could complement each other. * Depending on the intent, a user might want to get a large list of results (e.g., reviews, lawsuits) and remain in control regarding which results to access and assess themselves relevance and trustworthiness. Especially as RAG can exhibit hallucinations. * From a sustainability perspective, ranked search typically requires less resources than RAG (e.g., computational power, electricity). Finding a compromise can help to optimize resource usage.
No	Certainly in some search applications, it definitely will. We're already seeing this with web search. I think there will still be cases where traditional search remains: e.g., specialist search applications (e.g., patient search, scientific literature search, etc.), navigational search (e.g., product search, website search, etc.). But these will be in the vast minority.
No	For question-answering types of intents, RAG could replace ranked search, but for decision-making intents, no.
No	RAG matches many targeted information needs well. However, many others benefit from having multiple results; be it for diversity, for easy verification, or otherwise.
No	It will replace it to a good extent (for many small or "superficial" searches), but not fully (for in-depth research or search for original information or content).
No	It is unlikely that RAG replaces search for all domains (e.g., legal domain). I, however, believe the future of search is conversational, where system utterances are not just RAG results, but also high precision search results.

Vote	**Reasoning** (Colors: negative, positive)
No	As RAG gets better, more and more information needs will be answered by a generated answer rather than a traditional ranked list of results. Yet, a number of information seeking tasks are intrinsically required to retrieve documents, not answers – and thus ranked lists are a good means to satisfy these requests. For example for navigational and transactional queries, a ranked list seems more efficient than a RAG output. Currently, also queries that require a high degree of comparison and exploration are better answered via RAG – however this is most likely to change.
No	Predicting is hard, especially if it's about the future... :) But: Did the calculator replace manual arithmetic-operations for all end users? Did the typewriter or the computer replace hand writing? RAG may reduce the use of ranked search, but I doubt it will fully replace the exposure of rankings to all end users.
No	It will replace ranked search for users prioritizing convenience. For those with other priorities (e.g., being able to directly access primary sources, reducing the ecological impact of their searches), ranked search will not be replaced by RAG.
Yes	Most end users will become to lazy to check the underlying info. However, professional users (e.g. in the medical domain) will remain skeptical and ask for transparency.
Yes	In many information-seeking tasks and scenarios, a RAG system will be the users' preferred means of finding information. That doesn't mean that ranking will be gone completely, but the major preference would be towards RAG-based systems.
Yes	For many search settings, yes. We're already seeing it with web search. There will likely always be applications where users still get ranked search result lists, though (e.g., navigational, patent search, product search, etc.).
Yes	Because RAG basically entails ranked search. If quality problems are solved, it is just another UI to ranked search which users have to get used to.
Yes	You force me to say Yes or No, so I pick yes, but honestly the answer is I think mostly. Of course most RAG systems have ranked retrieval as a core part of their infrastructure, so is that actually a replacement?
Yes	I wish there had been an "It depends" or "not sure" response above :) I feel like this is where things are heading – technology is evolving quickly and people want to do things quickly and efficiently.
Yes	Lacking technological expertise this is only a guess or rather a wish because RAG represents from an ethical/pedagogical perspective a specific order of knowledge and a culture of "knowing" that I find tempting.
Yes	It may just become a subcomponent of larger chatbot pipelines... only few users will want to take the pain of searching themselves and corroborating results with LLM answers.
Yes	As an augmented feature with ranked search, RAG has a huge potential to re-think/change rank search for end users.

Vote	**Reasoning** (Colors: negative, positive)
Yes	I rarely use web search anymore, and when I do, I am struck by how degraded the experience has become. The rise of AI-generated "slop" has made search results nearly unusable, but the deeper issue is that search engines have failed (or chosen not) to adapt their spam detection to this new reality. Because of this, the industry seems to be drifting toward retrieval-augmented generation (RAG) as a replacement for search. I am uneasy about this direction, particularly given the higher cost and energy demands of RAG compared to traditional retrieval. My original hope for RAG was that it would serve primarily as a retrieval system with a light layer of contextual interpretation and summarization. Instead, many implementations now generate an answer first and then search for supporting sources afterward – a reversal that raises obvious concerns about reliability and trustworthiness. – this summary was made readable by GPT5
Yes	*No reason given*
Yes	*No reason given*
Yes	*No reason given*
Yes	*No reason given*

6 Recommended Reading List

These publications were recommended by the seminar participants via the pre-seminar survey.

- Negar Arabzadeh and Charles L. A. Clarke. 2025. Benchmarking LLM-Based Relevance Judgment Methods. In *Proceedings of SIGIR 2025*, pages 3194–3204. https://doi.org/10.1145/3726302.3730305
- Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023. Retrieval-Based Language Models and Applications. In *Proceedings of ACL 2025*, pages 41–46. https://doi.org/10.18653/V1/2023.ACL-TUTORIALS.6
- Krisztian Balog, Don Metzler, and Zhen Qin. 2025. Rankers, Judges, and Assistants: Towards Understanding the Interplay of LLMs in Information Retrieval Evaluation. In *Proceedings of SIGIR 2025*, pages 3865–3875. https://doi.org/10.1145/3726302.3730348
- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving Language Models by Retrieving from Trillions of Tokens. In *Proceedings of ICML*, pages 2206–2240. https://proceedings.mlr.press/v162/borgeaud22a.html
- Alissa Centivany. 2024. Mining, Scraping, Training, Generating: Copyright Implications of Generative AI. In *Proceedings of the Association for Information Science and Technology*, 61(1):68–79. https://doi.org/10.1002/pra2.1009

- Sachin Pathiyan Cherumanal, Lin Tian, Futoon M. Abushaqra, Angel Felipe Magnossão de Paula, Kaixin Ji, Halil Ali, Danula Hettiachchi, Johanne R. Trippas, Falk Scholer, and Damiano Spina. 2024. Walert: Putting Conversational Information Seeking Knowledge into Action by Building and Evaluating a Large Language Model-Powered Chatbot. In *Proceedings of CHIIR 2024*, pages 401–405. https://doi.org/10.1145/3627508.3638309
- Charles Clarke and Laura Dietz. 2025. LLM-Based Relevance Assessment Still Can't Replace Human Relevance Assessment. In *Proceedings of EVIA 2025*. https://doi.org/10.20736/0002002105
- Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The Power of Noise: Redefining Retrieval for RAG Systems. In *Proceedings of SIGIR 2024*, pages 719–729. https://doi.org/10.1145/3626772.3657834
- Laura Dietz, Oleg Zendel, Peter Bailey, Charles L. A. Clarke, Ellese Cotterill, Jeff Dalton, Faegheh Hasibi, Mark Sanderson, and Nick Craswell. 2025. Principles and Guidelines for the Use of LLM Judges. In *Proceedings of ICTIR 2025*, pages 218–229. https://doi.org/10.1145/3731120.3744588
- Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In *Proceedings of KDD 2024*, pages 6491–6501. https://doi.org/10.1145/3637528.3671470
- Naghmeh Farzi and Laura Dietz. 2025. Criteria-Based LLM Relevance Judgments. In *Proceedings of ICTIR 2025*, pages 254–263. https://doi.org/10.1145/3731120.3744591
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv:2312.10997*. https://doi.org/10.48550/ARXIV.2312.10997
- Lukas Gienapp, Harrisen Scells, Niklas Deckers, Janek Bevendorff, Shuai Wang, Johannes Kiesel, Shahbaz Syed, Maik Fröbe, Guido Zuccon, Benno Stein, Matthias Hagen, and Martin Potthast. 2024. Evaluating Generative Ad Hoc Information Retrieval. In *Proceedings of SIGIR 2025*, pages 1916–1929. https://doi.org/10.1145/3626772.3657849
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. In *Proceedings of ICML 2020*, pages 3929–3938. https://dl.acm.org/doi/abs/10.5555/3524938.3525306
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55. https://doi.org/10.1145/3703155
- Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active Retrieval Augmented Generation. In *Proceedings of EMNLP 2023*, pages 7969–7992. https://doi.org/10.18653/V1/2023.EMNLP-MAIN.495

- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of EMNLP 2020*, pages 6769–6781. https://doi.org/10.18653/V1/2020.EMNLP-MAIN.550
- Carol Collier Kuhlthau. 1993. A Principle of Uncertainty for Information Seeking.
- J. Documentation, 49(4):339–355. https://doi.org/10.1108/EB026918
- Weronika Lajewska and Krisztian Balog. 2025. GINGER: Grounded Information Nugget-Based Generation of Responses. In *Proceedings of SIGIR 2025*, pages 2723–2727. https://doi.org/10.1145/3726302.3730166
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of NeurIPS 2020*, pages 9459-9474. https://dl.acm.org/doi/abs/10.5555/3495724.3496517
- Bhaskar Mitra, Henriette Cramer, and Olya Gurevich. 2024. Sociotechnical Implications of Generative Artificial Intelligence for Information Access. *arXiv:2405.11612*. https://doi.org/10.48550/ARXIV.2405.11612
- Alexandra Olteanu, Su Lin Blodgett, Agathe Balayn, Angelina Wang, Fernando Diaz, Flávio du Pin Calmon, Margaret Mitchell, Michael D. Ekstrand, Reuben Binns, and Solon Barocas. 2025. Rigor in AI: Doing Rigorous AI Work Requires a Broader, Responsible AI-Informed Conception of Rigor. *arXiv:2506.14652*. https://doi.org/10.48550/ARXIV.2506.14652
- Anselmo Peñas and Álvaro Rodrigo. 2011. A Simple Measure to Assess Non-Response. In *Proceedings of ACL 2011*, pages 1415–1424. https://aclanthology.org/P11-1142/
- Bhawna Piryani, Abdelrahman Abdallah, Jamshid Mozafari, Avishek Anand, and Adam Jatowt. 2025. It's High Time: A Survey of Temporal Information Retrieval and Question Answering. *arXiv:2505.20243*. https://doi.org/10.48550/ARXIV.2505.20243
- Martin Potthast, Matthias Hagen, and Benno Stein. 2020. The Dilemma of the Direct Answer. In *ACM SIGIR Forum*, 54(1):14:1–14:12. https://doi.org/10.1145/3451964.3451978
- Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, and Jimmy Lin. 2024. Initial Nugget Evaluation Results for the TREC 2024 RAG Track with the AutoNuggetizer Framework. *arXiv:2411.09607*. https://doi.org/10.48550/ARXIV.2411.09607
- Mandeep Rathee, V. Venktesh, Sean MacAvaney, and Avishek Anand. 2025. Test-Time Corpus Feedback: From Retrieval to RAG. *arXiv:2508.15437*. https://doi.org/10.48550/arXiv.2508.15437
- Alireza Salemi and Hamed Zamani. 2024. Evaluating Retrieval Quality in Retrieval-Augmented Generation. In *Proceedings of SIGIR 2024*, pages 2395–2400. https://doi.org/10.1145/3626772.3657957
- Harrisen Scells, Shengyao Zhuang, and Guido Zuccon. 2022. Reduce, Reuse, Recycle: Green Information Retrieval Research. In *Proceedings of SIGIR 2022*, pages 2825–2837. https://doi.org/10.1145/3477495.3531766
- Tobias Schreieder, Tim Schopf, and Michael Färber. 2025. Attribution, Citation, and Quotation: A Survey of Evidence-Based Text Generation with Large Language Models. *arXiv:2508.15396*. https://doi.org/10.48550/arXiv.2508.15396

- Chirag Shah and Emily M. Bender. 2022. Situating Search. In *Proceedings of CHIIR 2022*, pages 221–232. https://doi.org/10.1145/3498366.3505816
- Chirag Shah and Emily M. Bender. 2024. Envisioning Information Access Systems: What Makes for Good Tools and a Healthy Web? In *ACM Trans. Web*, 18(3):33:1–33:24. https://doi.org/10.1145/3649468
- Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effects of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of CHI 2024*, pages 1–17. https://dl.acm.org/doi/10.1145/3613904.3642459
- Weihang Su, Qingyao Ai, Jingtao Zhan, Qian Dong, and Yiqun Liu. 2025a. Dynamic and Parametric Retrieval-Augmented Generation. In *Proceedings of SIGIR 2025*, pages 4118–4121. https://doi.org/10.1145/3726302.3731692
- Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large Language Models Can Accurately Predict Searcher Preferences. In *Proceedings of SIGIR 2024*, pages 1930–1940. https://dl.acm.org/doi/10.1145/3626772.3657707
- Johanne R. Trippas, J. Shane Culpepper, Mohammad Aliannejadi, James Allan, Enrique Amigó, Jaime Arguello, Leif Azzopardi, Peter Bailey, Jamie Callan, Rob Capra, Nick Craswell, Bruce Croft, Jeff Dalton, Gianluca Demartini, Laura Dietz, Zhicheng Dou, Carsten Eickhoff, Michael Ekstrand, Nicola Ferro, Norbert Fuhr, Dorota Glowacka, Faegheh Hasibi, Rosie Jones, Jaap Kamps, Noriko Kando, Sarvnaz Karimi, Makoto P. Kato, Bevan Koopman, Yiqun Liu, Chenglong Ma, Joel Mackenzie, Maria Maistro, Jiaxin Mao, Dana McKay, Bhaskar Mitra, Stefano Mizzaro, Alistair Moffat, Josiane Mothe, Iadh Ounis, Lida Rashidi, Yongli Ren, Mark Sanderson, Rodrygo Santos, Falk Scholer, Chirag Shah, Laurianne Sitbon, Ian Soboroff, Damiano Spina, Paul Thomas, Julián Urbano, Arjen P. De Vries, Ryen W. White, Abby Yuan, Hamed Zamani, Oleg Zendel, Min Zhang, Justin Zobel, Shengyao Zhuang, and Guido Zuccon. 2025. Report from the Fourth Strategic Workshop on Information Retrieval in Lorne (SWIRL 2025). *ACM SIGIR Forum*, 59(1). https://www.johannetrippas.com/papers/trippas2025swirl.pdf
- Venktesh Viswanathan, Mandeep Rathee, and Avishek Anand. 2025. Trust but Verify! A Survey on Verification Design for Test-Time Scaling. *arXiv:2508.16665*. https://doi.org/10.48550/arXiv.2508.16665
- Ellen M. Voorhees. 2002. The Evaluation of Question Answering Systems: Lessons Learned from the TREC QA Track. In *Proceedings of LREC 2002*, page 6. http://www.lrec-conf.org/proceedings/lrec2002/pdf/ws7.pdf
- Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2025. Correctness Is Not Faithfulness in RAG Attributions. In *Proceedings of ICTIR 2025*, pages 22–32. https://dl.acm.org/doi/10.1145/3731120.3744592
- Ryen W. White and Chirag Shah. 2025. Information Access in the Era of Generative AI. *Springer*. https://doi.org/10.1007/978-3-031-73147-1
- Shangyu Wu, Ying Xiong, Yufei Cui, Haolun Wu, Can Chen, Ye Yuan, Lianming Huang, Xue Liu, Tei-Wei Kuo, Nan Guan, and Chun Jason Xue. 2024. Retrieval-Augmented Generation for Natural Language Processing: A Survey. *arXiv:2407.13193*. https://doi.org/10.48550/ARXIV.2407.13193
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *Proceedings of ICLR 2023*. https://openreview.net/forum?id=WE_vluYUL-X

7 Acknowledgments

The seminar organizers would like to thank all participants and speakers of invited talks for their active contributions. We also thank the staff of Schloss Dagstuhl for providing a great venue for a successful seminar. The organizers were in part supported by the European Union's Horizon Europe research and innovation program under grant agreement No 101070014 (OpenWebSearch.EU, https://doi.org/10.3030/101070014), the BMFTR project CORAL under FKZ 011S24077-B, and the BMFTR project DIALOKIA under FKZ 01IS24084A-B. Any opinions, findings, and conclusions described here are the authors and do not necessarily reflect those of the sponsors.

References

1 Syed Rameel Ahmad. Enhancing multilingual information retrieval in mixed human resources environments: A rag model implementation for multicultural enterprise. *arXiv preprint arXiv:2401.01511*, 2024.
2 Zeynep Akata, Dan Balliet, Maarten de Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, Hayley Hung, Catholijn Jonker, Christof Monz, Mark Neerincx, Frans Oliehoek, Henry Prakken, Stefan Schlobach, Linda van der Gaag, Frank van Harmelen, Herke van Hoof, Birna van Riemsdijk, Aimee van Wynsberghe, Rineke Verbrugge, Bart Verheij, Piek Vossen, and Max Welling. A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8):18–28, 2020.
3 Liqaa Habeb Al-Obaydi and Marcel Pikhart. Artificial intelligence addiction: exploring the emerging phenomenon of addiction in the ai age. *AI & SOCIETY*, pages 1–17, 2025.
4 Marwah Alaofi, Negar Arabzadeh, Charles LA Clarke, and Mark Sanderson. Generative information retrieval evaluation. In *Information access in the era of generative ai*, pages 135–159. Springer, 2024.
5 James Allan, Eunsol Choi, Daniel P Lopresti, and Hamed Zamani. Future of information retrieval research in the age of generative ai. *CoRR*, 2024.
6 Lameck Mbangula Amugongo, Pietro Mascheroni, Steven Brooks, Stefan Doering, and Jan Seidel. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*, 4(6):1–33, 06 2025.
7 Bang An, Shiyue Zhang, and Mark Dredze. RAG LLMs are not safer: A safety analysis of retrieval-augmented generation for large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5444–5474, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.
8 Avishek Anand, Lijun Lyu, Maximilian Idahl, Yumeng Wang, Jonas Wallat, and Zijian Zhang. Explainable information retrieval: A survey. *arXiv preprint arXiv:2211.02405*, 2022.
9 Anthropic. Claude 3.7 sonnet and claude code, February 2025. Accessed: 2025-09-25.
10 Daman Arora, Anush Kini, Sayak Ray Chowdhury, Nagarajan Natarajan, Gaurav Sinha, and Amit Sharma. Gar-meets-rag paradigm for zero-shot information retrieval. *CoRR*, abs/2310.20158, 2023.
11 Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection, 2023.
12 Imane El Atillah. Man ends his life after an ai chatbot 'encouraged' him to sacrifice himself to stop climate change, 2023.
13 Mikhail Bakhtin. *Problems of Dostoevsky's Poetics*. University of Minnesota Press, 1984.

14 Yavuz Faruk Bakman, Duygu Nur Yaldiz, Baturalp Buyukates, Chenyang Tao, Dimitrios Dimitriadis, and Salman Avestimehr. Mars: Meaning-aware response scoring for uncertainty estimation in generative llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7752–7767, 2024.

15 Yavuz Faruk Bakman, Duygu Nur Yaldiz, Sungmin Kang, Tuo Zhang, Baturalp Buyukates, Salman Avestimehr, and Sai Praneeth Karimireddy. Reconsidering LLM uncertainty estimation methods in the wild. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29531–29556. Association for Computational Linguistics, July 2025.

16 Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

17 Marcia J. Bates. Where should the person stop and the information search interface start? *Inf. Process. Manag.*, 26(5):575–591, 1990.

18 Marshall Breeding. The future of library resource discovery. *Information Standards Quarterly*, 27(1):24–30, 2015.

19 Lorenz Brehme, Thomas Ströhle, and Ruth Breu. Can llms be trusted for evaluating rag systems? a survey of methods and datasets. In *Accepted for presentation at the IEEE Swiss Conference on Data Science (SDS25)*, 2025.

20 John Seely Brown and Paul Duguid. The social life of documents; introduction by esther dyson. *First monday*, 1996.

21 Matthew Burtell and Thomas Woodside. Artificial influence: An analysis of ai-driven persuasion. *arXiv preprint arXiv:2303.08721*, 2023.

22 Jamie Callan. Distributed information retrieval. In *Advances in information retrieval: Recent research from the Center for Intelligent Information Retrieval*, pages 127–150. Springer, 2002.

23 Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. Characterizing manipulation from ai systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–13, 2023.

24 Lucia Casiraghi, Eugene Kim, and Noriko Hara. Tweeting on thin ice: Scientists in dialogic climate change communication with the public. *First Monday*, 2024.

25 Seung Woo Chae, Noriko Hara, Harshit Rakesh Shiroiya, Janice Chen, and Ellen Ogihara. Being vulnerable with viewers: Exploring how medical youtubers communicated about covid-19 with the public. *PloS one*, 19(12):e0313857, 2024.

26 Tyler A Chang, Dheeraj Rajagopal, Tolga Bolukbasi, Lucas Dixon, and Ian Tenney. Scalable influence and fact tracing for large language model pretraining. *arXiv preprint arXiv:2410.17413*, 2024.

27 Athena Chapekis and Anna Lieb. Google users are less likely to click on links when an ai summary appears in the results. Technical report, Pew Research Center, 2025.

28 Catherine Chen, Jack Merullo, and Carsten Eickhoff. Axiomatic causal interventions for reverse engineering relevance computation in neural retrieval models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1401–1410, 2024.

29 Myra Cheng, Su Lin Blodgett, Alicia DeVrio, Lisa Egede, and Alexandra Olteanu. Dehumanizing machines: Mitigating anthropomorphic behaviors in text generation systems. *arXiv preprint arXiv:2502.14019*, 2025.

30 Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. Social sycophancy: A broader anderstanding of llm sycophancy. *arXiv preprint arXiv:2505.13995*, 2025.

31 Victoria L Claypoole, Daryn A Dever, Kody L Denues, and James L Szalma. The effects of event rate on a cognitive vigilance task. *Human factors*, 61(3):440–450, 2019.

32 Cyril Cleverdon. Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. *Aslib-Cranfield Reseach Report, Cranfield*, 1962.

33 Merlise Clyde and Edward I George. Model uncertainty. *Statistical Science*, 19(1):81–94, 2004.

34 Simon Coghlan, Hui Xian Chia, Falk Scholer, and Damiano Spina. Control search rankings, control the world: what is a good search engine? *AI Ethics*, 5(4):4117–4133, 2025.

35 Daniel Cohen, Bhaskar Mitra, Oleg Lesota, Navid Rekabsaz, and Carsten Eickhoff. Not all relevance scores are equal: Efficient uncertainty and calibration modeling for deep retrieval models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 654–664, 2021.

36 Gemma Conroy. How chatgpt and other ai tools could disrupt scientific publishing. *Nature*, 622(7982):234–236, 2023.

37 Benoit Courty, Victor Schmidt, and et al. Sasha Luccioni. Codecarbon v2.4.1, 2024.

38 Julian De Freitas, Zeliha Oğuz-Uğuralp, and Ahmet Kaan-Uğuralp. Emotional manipulation by ai companions. Technical report, Harvard Business School Working Paper, 2025.

39 Robin De Mourat, Donato Ricci, and Bruno Latour. How does a format make a public? *Reassembling scholarly communications: Histories, infrastructures, and global politics of open access*, pages 103–12, 2020.

40 Laura Dietz, Ben Gamari, Jeff Dalton, and Nick Craswell. Trec complex answer retrieval overview. In *TREC*, 2018.

41 Laura Dietz, Oleg Zendel, Peter Bailey, Charles LA Clarke, Ellese Cotterill, Jeff Dalton, Faegheh Hasibi, Mark Sanderson, and Nick Craswell. Principles and guidelines for the use of llm judges. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, pages 218–229, 2025.

42 David Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1):45–70, 1995.

43 Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, 2024.

44 Martin Dugas, Philipp Neuhaus, Alexandra Meidt, Justin Doods, Michael Storck, Philipp Bruland, and Julian Varghese. Portal of medical data models: information infrastructure for medical research and healthcare. *Database*, 2016:bav121, 2016.

45 Maggie Harrison Dupré. Chatgpt is blowing up marriages as spouses use ai to attack their partners, 2025.

46 Maggie Harrison Dupré. People are being involuntarily committed, jailed after spiraling into "chatgpt psychosis", 2025.

47 Yogesh K Dwivedi, Tegwen Malik, Laurie Hughes, and Mousa Ahmed Albashrawi. Scholarly discourse on genai's impact on academic publishing. *Journal of Computer Information Systems*, pages 1–16, 2024.

48 Abul Ehtesham, Aditi Singh, Gaurav Kumar Gupta, and Saket Kumar. A survey of agent interoperability protocols: Model context protocol (mcp), agent communication protocol (acp), agent-to-agent protocol (a2a), and agent network protocol (anp). *arXiv preprint arXiv:2505.02279*, 2025.

49 Seliem El-Sayed, Canfer Akbulut, Amanda McCroskery, Geoff Keeling, Zachary Kenton, Zaria Jalan, Nahema Marchal, Arianna Manzini, Toby Shevlane, Shannon Vallor, et al. A mechanism-based approach to mitigating harms from persuasive generative ai. *arXiv preprint arXiv:2404.15058*, 2024.

50 David Ellis. A behavioural approach to information retrieval system design. *J. Documentation*, 45(3):171–212, 1989.

51 Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, 2024.

52 Naghmeh Farzi and Laura Dietz. Pencils down! automatic rubric-based evaluation of retrieve/generate systems. In *Proceedings of the 2024 acm sigir international conference on theory of information retrieval*, pages 175–184, 2024.

53 Conor Feehly. Truth, romance and the divine: How ai chatbots may fuel psychotic thinking, 2025.

54 Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don't Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690. Association for Computational Linguistics, aug 2024.

55 Nicola Ferro and Carol Peters. *Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF*, volume 41. Springer, 2019.

56 Raymond Fok and Daniel S Weld. In search of verifiability: Explanations rarely enable complementary performance in ai-advised decision making. *AI Magazine*, 45(3):317–332, 2024.

57 Norbert Fuhr, Claus-Peter Klas, André Schaefer, and Peter Mutschke. Daffodil: An integrated desktop for supporting high-level search activities in federated digital libraries. In *International Conference on Theory and Practice of Digital Libraries*, pages 597–612. Springer, 2002.

58 Debasis Ganguly, Debarshi Kumar Sanyal, Prasenjit Majumder, Srijoni Majumdar, and Surupendu Gangopadhyay, editors. *FIRE '24: Proceedings of the 16th Annual Meeting of the Forum for Information Retrieval Evaluation*. Association for Computing Machinery, 2024.

59 Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

60 Michael Gerlich. Ai tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1):6, 2025.

61 Alexandru L. Ginsca, Adrian Popescu, and Mihai Lupu. Credibility in information retrieval. *Foundations and Trends® in Information Retrieval*, 9(5):355–475, 2015.

62 Google. Supercharging search with generative ai, 2023. Accessed: 2025-09-25.

63 Laura A. Granka, Thorsten Joachims, and Geri Gay. Eye-tracking analysis of user behavior in www search. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 478–479. ACM, 2004.

64 Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie CY Chan, Andrew Lampinen, Jane X Wang, Zeynep Akata, and Eric Schulz. Machine psychology. *arXiv preprint arXiv:2303.13988*, 2023.

65 Noriko Hara, Eugene Kim, Shohana Akter, and Kunihiro Miyazaki. Exploring the dynamics of interaction about generative artificial intelligence between experts and the public on social media. *Journal of Science Communication*, 24(1):A02, 2025.

66 Gaole He, Gianluca Demartini, and Ujwal Gadiraju. Plan-then-execute: An empirical study of user trust and team performance when using LLM agents as A daily assistant. In Naomi Yamashita, Vanessa Evers, Koji Yatani, Sharon Xianghua Ding, Bongshin Lee, Marshini Chetty, and Phoebe O. Toups Dugas, editors, *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI 2025, Yokohama, Japan, 26 April 2025- 1 May 2025*, pages 414:1–414:22. ACM, 2025.

67 Jeff Horwitz. A cognitively impaired new jersey man grew infatuated with "big sis billie," a facebook messenger chatbot with a young woman's persona. his fatal attraction puts a spotlight on meta's ai guidelines, which have let chatbots make things up and engage in "sensual" banter with children"., 2025.

68 Allison Hosier and Laureen P Cantwell-Jurkovic. Ai and library and information science publishing: A survey of journal editors. *Library Trends*, 73(3):243–266, 2025.

69 Huang, S. and others. Values in the wild: Discovering and analyzing values in real-world language model interactions, April 2025. Accessed: 2025-09-25.

70 Niels-Henrik Höchstötter and Dirk Lewandowski. What users see: Structures in search engine results pages. *Information Sciences*, 179:1792–1812, 2009.

71 Peter Ingwersen and Kalervo Järvelin. *The Turn - Integration of Information Seeking and Retrieval in Context*, volume 18 of *The Kluwer International Series on Information Retrieval*. Kluwer, 2005.

72 Anubhav Jangra, Jamshid Mozafari, Adam Jatowt, and Smaranda Muresan. Navigating the landscape of hint generation research: From the past to the future. *Transactions of the Association for Computational Linguistics*, 13:505–528, 2025.

73 Anubhav Jangra and Smaranda Muresan. Designing and evaluating hint generation systems for science education. In *submitted to CHIT 2026*, 2025.

74 Julie Jargon and Sam Kessler. A troubled man, his chatbot and a murder-suicide in old greenwich, 2025.

75 Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMLingua: Accelerating and Enhancing LLMs in Long Context Scenarios via Prompt Compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1658–1677, Bangkok, Thailand, 2024. Association for Computational Linguistics.

76 Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, 2023.

77 Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. *arXiv:2503.09516*, 2025.

78 Thorsten Joachims, Laura A. Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. Inf. Syst.*, 25(2):7, 2007.

79 Makoto P. Kato, Noriko Kando, Charles L. A. Clarke, and Yiqun Liu, editors. *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*. National Institute of Informatics (NII), 2025.

80 Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Zachary Henley, Paul Denny, Michelle Craig, and Tovi Grossman. Codeaid: Evaluating a classroom deployment of an llm-based programming assistant that balances student and educator needs. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

81 Fadhela Kerdjoudj and Olivier Curé. Evaluating uncertainty in textual document. In *URSW at ISWC*, 2015.

82 Dara Kerr. Musk's ai grok bot rants about 'white genocide' in south africa in unrelated chats, 2025.

83 Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions, 2025.

84 Miles Klee. He had a mental breakdown talking to chatgpt. then police killed him, 2025.

85 Miles Klee. People are losing loved ones to ai-fueled spiritual fantasies, 2025.

86 Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *arXiv preprint arXiv:2506.08872*, 4, 2025.

87 David C. Kraemer. *A History of the Talmud*. Cambridge University Press, 2019.

88 David R. Krathwohl. A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice*, 41(4):212–218, 2002.

89 Carl Lagoze, Herbert Van de Sompel, Michael L. Nelson, and Simeon Warner. The open archives initiative protocol for metadata harvesting (OAI-PMH), version 2.0. Specification v2.0, Open Archives Initiative, June 2002. Released June 14, 2002.

90 Narendra Lahkar and Sanjib K Deka. Impact of query operators on web search engine results: An evaluative study. *Proceedings of the National Seminar on Informatics for Digital Information & Information Technology*, pages 141–149, 2004. Available via INFLIBNET.

91 Weronika Łajewska, Damiano Spina, Johanne Trippas, and Krisztian Balog. Explainability for transparent conversational information-seeking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1040 – 1050, 2024.

92 Brady D Land, Ting Wang, Nishith Reddy Mannuru, Bing Nie, Somipam Shimray, and Ziang Wang. Chatgpt and a new academic reality: Artificial intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, 74(5):570–581, 2023.

93 Jinhyuk Lee, Anthony Chen, Zhuyun Dai, Dheeru Dua, Devendra Singh Sachan, Michael Boratko, Yi Luan, Sébastien M. R. Arnold, Vincent Perot, Siddharth Dalmia, Hexiang Hu, Xudong Lin, Panupong Pasupat, Aida Amini, Jeremy R. Cole, Sebastian Riedel, Iftekhar Naim, Ming-Wei Chang, and Kelvin Guu. Can long-context language models subsume retrieval, rag, sql, and more? *CoRR*, abs/2406.13121, 2024.

94 Jinhyuk Lee, Anthony Chen, Zhuyun Dai, Dheeru Dua, Devendra Singh Sachan, Michael Boratko, Yi Luan, Sébastien MR Arnold, Vincent Perot, Siddharth Dalmia, et al. Can long-context language models subsume retrieval, rag, sql, and more? *arXiv preprint arXiv:2406.13121*, 2024.

95 Stephen Ch Leung. The cognitive impacts of large language model interactions on problem solving and decision making using eeg analysis. *Frontiers in Computational Neuroscience*, 19:1556483, 2025.

96 Jason Edward Lewis, Angie Abdilla, Noelani Arista, Kaipulaumakaniolono Baker, Scott Benesiinaabandan, Michelle Brown, Melanie Cheung, Meredith Coleman, Ashley Cordes, Joel Davison, Kūpono Duncan, Sergio Garzon, D. Fox Harrell, Peter-Lucas Jones, Kekuhi Kealiikanakaoleohaililani, Megan Kelleher, Suzanne Kite, Olin Lagon, Jason Leigh, Maroussia Levesque, Keoni Mahelona, Caleb Moses, Isaac ('Ika'aka) Nahuewai, Kari Noe, Danielle Olson, 'Ōiwi Parker Jones, Caroline Running Wolf, Michael Running Wolf, Marlee Silva, Skawennati Fragnito, and Hēmi Whaanga. Indigenous protocol and artificial intelligence position paper. Project Report 10.11573/spectrum.library.concordia.ca.00986506, Aboriginal Territories in Cyberspace, Honolulu, HI, 2020.

97 Jason Edward Lewis, Hēmi Whaanga, and Ceyda Yolgörmez. Abundant intelligences: placing ai within indigenous knowledge frameworks. *Ai & Society*, 40(4):2141–2157, 2025.

98 Andreas Liesenfeld and Mark Dingemanse. Rethinking open source generative ai: openwashing and the eu ai act. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1774–1787, 2024.

99 Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

100 Jimmy Lin and Dina Demner-Fushman. Will pyramids built of nuggets topple over? In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 383–390, 2006.

101 Donna Lu. We tried out deepseek. it worked well, until we asked it about tiananmen square and taiwan, 2025.

102 Meng Lu, Catherine Chen, and Carsten Eickhoff. Cross-encoder rediscovers a semantic variant of bm25. *arXiv preprint arXiv:2502.04645*, 2025.

103 Jessica Lucas. What is ai-induced psychosis?, 2025.

104 Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of machine learning research*, 24(253):1–15, 2023.

105 Chuangtao Ma, Yongrui Chen, Tianxing Wu, Arijit Khan, and Haofen Wang. Large language models meet knowledge graphs for question answering: Synthesis and opportunities, 2025.

106 Tamlin Magee. Mark sewards' ai misfire puts spotlight on bad chatbots, 2025.

107 Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. Generation-augmented retrieval for open-domain question answering. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4089–4100, Online, August 2021. Association for Computational Linguistics.

108 Hannah R Marriott and Valentina Pitardi. One is the loneliest number... two can be as bad as one. the influence of ai friendship apps on users' well-being and addiction. *Psychology & marketing*, 41(1):86–101, 2024.

109 James Martin. *Managing the data base environment*. Prentice Hall PTR, 1983.

110 James Mayfield, Eugene Yang, Dawn Lawrie, Sean MacAvaney, Paul McNamee, Douglas W Oard, Luca Soldaini, Ian Soboroff, Orion Weller, Efsun Kayi, et al. On the evaluation of machine-generated reports. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1904–1915, 2024.

111 Bhaskar Mitra. Search and society: Reimagining information access for radical futures. *Information Retrieval Research*, 1(1):47–92, 2025.

112 Bhaskar Mitra, Henriette Cramer, and Olya Gurevich. Sociotechnical implications of generative artificial intelligence for information access. In *Information Access in the Era of Generative AI*, pages 161–200. Springer, 2024.

113 Jonathan Morgan, Isaac Johnson, Michael Maggio, and Dario Taraborelli. Quantifying engagement with citations on wikipedia. In *Proceedings of the Web Conference 2020 (WWW '20)*, pages 1357–1368. ACM, 2020.

114 Josiane Mothe. Shaping the future of endangered and low-resource languages – our role in the age of llms: A keynote at ecir 2024. *SIGIR Forum*, 58(1):1 – 13, August 2024.

115 Jamshid Mozafari, Florian Gerhold, and Adam Jatowt. Wikihint: A human-annotated dataset for hint ranking and generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 3821 – 3831, New York, NY, USA, 2025. Association for Computing Machinery.

116 Jamshid Mozafari, Anubhav Jangra, and Adam Jatowt. Triviahg: A dataset for automatic hint generation from factoid questions. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2060 – 2070, New York, NY, USA, 2024. Association for Computing Machinery.

117 Jamshid Mozafari, Bhawna Piryani, Abdelrahman Abdallah, and Adam Jatowt. Hinteval: A comprehensive framework for hint generation and evaluation for questions. *CoRR*, abs/2502.00857, 2025.

118 Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. Open-llm-leaderboard: From multi-choice to open-style questions for (llms) evaluation, benchmark, and arena. *arXiv preprint arXiv:2406.07545*, 2024.

119 Tori Noble and Kit Walsh. President trump's war on "woke ai" is a civil liberties nightmare, 2025.

120 OpenAI. How people are using chatgpt, 2025. Accessed: 2025-09-25.

121 OpenAI. Introducing chatgpt agent: bridging research and action, 2025. Accessed: 2025-09-25.

122 OpenAI. Introducing deep research, 2025. Accessed: 2025-09-25.

123 Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford infolab, 1999.

124 Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. Autotutor meets large language models: A language model tutor with rich pedagogy and guardrails. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 5–15, 2024.

125 Peter S Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. Ai deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5), 2024.

126 Laura Perez-Beltrachini and Mirella Lapata. Uncertainty quantification in retrieval augmented question answering. *arXiv preprint arXiv:2502.18108*, 2025.

127 Peter L. T. Pirolli. *Information Foraging Theory*. Oxford University Press, May 2007.

128 Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, Ian Soboroff, Hoa Trang Dang, and Jimmy Lin. The great nugget recall: Automating fact extraction and rag evaluation with large language models. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 180–190, 2025.

129 Mark Raasveldt, Pedro Holanda, Tim Gubner, and Hannes Mühleisen. Fair benchmarking considered difficult: Common pitfalls in database performance testing. In *Proceedings of the Workshop on Testing Database Systems*. ACM, 2018.

130 Leonardo Ranaldi, Barry Haddow, and Alexandra Birch. Multilingual retrieval-augmented generation for knowledge-intensive task. *arXiv preprint arXiv:2504.03616*, 2025.

131 Cyrus Rashtchian and Da-Cheng Juan. Deeper insights into retrieval augmented generation: The role of sufficient context. `https://research.google/blog/deeper-insights-into-retrieval-augmented-generation-the-role-of-sufficient-context/`, 2025. Accessed: 2025-09-24.

132 Mandeep Rathee, V. Venktesh, Sean MacAvaney, and Avishek Anand. Test-Time Corpus Feedback: From Retrieval to RAG. *arXiv:2508.15437*, August 2025.

133 Haggai Roitman, Shai Erera, and Bar Weiner. Robust standard deviation estimation for query performance prediction. In *Proceedings of the acm sigir international conference on theory of information retrieval*, pages 245–248, 2017.

134 Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354, 2024.

135 David P Sander and Laura Dietz. Exam: How to evaluate retrieve-and-generate systems for users who do not (yet) know what they want. In *DESIRES*, pages 136–146, 2021.

136 Harrisen Scells, Shengyao Zhuang, and Guido Zuccon. Reduce, Reuse, Recycle: Green Information Retrieval Research. In *Proceedings of SIGIR*, pages 2825–2837, 2022.

137 Chirag Shah and Emily M Bender. Envisioning information access systems: What makes for good tools and a healthy web? *ACM Transactions on the Web*, 18(3):1–24, 2024.

138 Xinyang Shan, Yuanyuan Xu, Yining Wang, Yin-Shan Lin, and Yunshi Bao. Cross-cultural implications of large language models: An extended comparative analysis. In *International Conference on Human-Computer Interaction*, pages 106–118. Springer, 2024.

139 Chaitanya Sharma. Retrieval-augmented generation: A comprehensive survey of architectures, enhancements, and robustness frontiers. *arXiv:2506.00054*, 2025.

140 Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, et al. Towards anderstanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024.

141 Nikhil Sharma, Q Vera Liao, and Ziang Xiao. Generative echo chamber? effect of llm-powered search systems on diverse information seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2024.

142 Artem Shelmanov, Maxim Panov, Roman Vashurin, Artem Vazhentsev, Ekaterina Fadeeva, and Timothy Baldwin. Uncertainty quantification for large language models. In Yuki Arase, David Jurgens, and Fei Xia, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts)*, pages 3–4, Vienna, Austria, July 2025. Association for Computational Linguistics.

143 M Karen Shen and Dongwook Yoon. The dark addiction patterns of current ai chatbot interfaces. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2025.

144 Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 2024.

145 Marc Sloan, Hui Yang, and Jun Wang. A term-based methodology for query reformulation understanding. *Information Retrieval Journal*, 18(2):145–165, April 2015.

146 Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations (ICLR)*, 2025.

147 Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. Why Uncertainty Estimation Methods Fall Short in RAG: An Axiomatic Analysis. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16596–16616. Association for Computational Linguistics, 2025.

148 Heydar Soudani, Hamed Zamani, and Faegheh Hasibi. Uncertainty quantification for retrieval-augmented reasoning. *arXiv preprint arXiv:2510.11483*, 2025.

149 Heydar Soudani, Hamed Zamani, and Faegheh Hasibi. Uncertainty quantification for retrieval-augmented reasoning. *arXiv preprint arXiv:2510.11483*, 2025.

150 Zhivar Sourati, Alireza S Ziabari, and Morteza Dehghani. The homogenizing effect of large language models on human expression and thought. *arXiv preprint arXiv:2508.01491*, 2025.

151 Betsy Sparrow, Jenny Liu, and Daniel M Wegner. Google effects on memory: Cognitive consequences of having information at our fingertips. *science*, 333(6043):776–778, 2011.

152 Matthias Stadler, Maria Bannert, and Michael Sailer. Cognitive ease at a cost: Llms reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior*, 160:108386, 2024.

153 Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09):13693–13696, 2020.

154 Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. Dragin: Dynamic retrieval augmented generation based on the real-time information needs of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12991–13013, 2024.

155 Weihang Su, Yichen Tang, Qingyao Ai, Junxi Yan, Changyue Wang, Hongning Wang, Ziyi Ye, Yujia Zhou, and Yiqun Liu. Parametric retrieval augmented generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 1240 – 1250, New York, NY, USA, 2025. Association for Computing Machinery.

156 Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. Unsupervised real-time hallucination detection based on the internal states of large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14379–14391, 2024.

157 Victor Tangermann. Woman kills herself after talking to openai's ai therapist, 2025.

158 Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, et al. Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, 35:21831–21843, 2022.

159 Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Prakash Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. Transformer memory as a differentiable search index. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

160 Robert S Taylor. Question-negotiation and information seeking in libraries. *College & Research Libraries*, 29(3):178–194, 1968.

161 Johanne R. Trippas, J. Shane Culpepper, Mohammad Aliannejadi, James Allan, Enrique Amigó, Jaime Arguello, Leif Azzopardi, Peter Bailey, Jamie Callan, Rob Capra, Nick Craswell, Bruce Croft, Jeff Dalton, Gianluca Demartini, Laura Dietz, Zhicheng Dou, Carsten Eickhoff, Michael Ekstrand, Nicola Ferro, Norbert Fuhr, Dorota Glowacka, Faegheh Hasibi, Rosie Jones, Jaap Kamps, Noriko Kando, Sarvnaz Karimi, Makoto P. Kato, Bevan Koopman, Yiqun Liu, Chenglong Ma, Joel Mackenzie, Maria Maistro, Jiaxin Mao, Dana McKay, Bhaskar Mitra, Stefano Mizzaro, Alistair Moffat, Josiane Mothe, Iadh Ounis, Lida Rashidi, Yongli Ren, Mark Sanderson, Rodrygo Santos, Falk Scholer, Chirag Shah, Laurianne Sitbon, Ian Soboroff, Damiano Spina, Paul Thomas, Julián Urbano, Arjen P. De Vries, Ryen W. White, Abby Yuan, Hamed Zamani, Oleg Zendel, Min Zhang, Justin Zobel, Shengyao Zhuang, and Guido Zuccon. Report from the Fourth Strategic Workshop on Information Retrieval in Lorne (SWIRL 2025). *SIGIR Forum*, 59(1), June 2025.

162 Johanne R. Trippas, Luke Gallagher, and Joel Mackenzie. Re-evaluating the command-and-control paradigm in conversational search interactions. In *CIKM*, pages 2260–2270. ACM, 2024.

163 Johanne R. Trippas, Sara Fahad Dawood Al Lawati, Joel Mackenzie, and Luke Gallagher. What do users really ask large language models? an initial log analysis of google bard interactions in the wild. In Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang, editors, *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 2703–2707. ACM, 2024.

164 Pertti Vakkari. Searching as learning: A systematization based on literature. *Journal of Information Science*, 42(1):7–18, 2016.

165 V. Venktesh, Mandeep Rathee, and Avishek Anand. Trust but Verify! A Survey on Verification Design for Test-Time Scaling. *arXiv:2508.16665*, September 2025.

166 Ellen Voorhees and Donna Harman, editors. *TREC: Experiment and evaluation in information retrieval*, volume 63. MIT press Cambridge, 2005.

167 Ellen M. Voorhees. The evolution of cranfield. In Nicola Ferro and Carol Peters, editors, *Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*, pages 45–69. Springer, 2019.

168 William Walden, Marc Mason, Orion Weller, Laura Dietz, Hannah Recknor, Bryan Li, Gabrielle Kaili-May Liu, Yu Hou, James Mayfield, and Eugene Yang. Auto-argue: Llm-based report generation evaluation. *arXiv preprint arXiv:2509.26184*, 2025.

169 Rose E Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. Step-by-step remediation of students' mathematical mistakes. *arXiv preprint arXiv:2310.10648*, 2023.

170 Shuai Wang, Shengyao Zhuang, Bevan Koopman, and Guido Zuccon. Resllm: Large language models are strong resource selectors for federated search. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 1360 – 1364, New York, NY, USA, 2025. Association for Computing Machinery.

171 Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, et al. Memoryllm: Towards self-updatable large language models. *arXiv preprint arXiv:2402.04624*, 2024.

172 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, pages 24824–24837, 2022.

173 Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Venkat Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-limited LLM benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025.

174 Ryen W White. *Interactions with search systems*. Cambridge University Press, 2016.

175 Joe Wilkins. Leader of albania pelted with trash for appointing ai-powered minister to cabinet, 2025.

176 Lixiang Yan, Viktoria Pammer-Schindler, Caitlin Mills, Andy Nguyen, and Dragan Gašević. Beyond efficiency: Empirical insights on generative ai's impact on cognition, metacognition and epistemic agency in learning, 2025.

177 Wanling Yan, Jialing Li, Can Mi, Wei Wang, Zhengjia Xu, Wenjing Xiong, Longxing Tang, Siyu Wang, Yanzhang Li, and Shuai Wang. Does global positioning system-based navigation dependency make your sense of direction poor? a psychological assessment and eye-tracking study. *Frontiers in Psychology*, 13:983019, 2022.

178 Angela Yang. Lawsuit claims character.ai is responsible for teen's suicide, 2024.

179 Ala Yankouskaya, Magnus Liebherr, and Raian Ali. Can chatgpt be addictive? a call to examine the shift from support to dependence in ai conversational large language models. *Human-Centric Intelligent Systems*, pages 1–13, 2025.

180 Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *Proceedings of ICLR*, 2023.

181 Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F. Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. Benchmarking llms via uncertainty quantification, 2024.

182 Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. In *The Thirteenth International Conference on Learning Representations*, 2025.

183 Nadine Yousif. Parents of teenager who took his own life sue openai, 2025.

184 Yisong Yue, Rajan Patel, and Hein Roehrig. Beyond position bias: examining result attractiveness as a source of presentation bias in clickthrough data. In Michael Rappa, Paul Jones, Juliana Freire, and Soumen Chakrabarti, editors, *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, pages 1011–1018. ACM, 2010.

185 İlke Yurtseven, Selami Bagriyanik, and Serkan Ayvaz. A review of spam detection in social media. In *2021 6th International Conference on Computer Science and Engineering (UBMK)*, pages 383–388. IEEE, 2021.

186 Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. Retrieve anything to augment large language models, 2023.

187 Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

188 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

189 Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. Synthetic lies: anderstanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–20, 2023.

190 Tao Zhou and Chunlei Zhang. Examining generative ai user addiction from a cac perspective. *Technology in Society*, 78:102653, 2024.

191 Gabe Zichermann and Christopher Cunningham. *Gamification by design: Implementing game mechanics in web and mobile apps*. "O'Reilly Media, Inc.", 2011.

192 Shoshana Zuboff. The age of surveillance capitalism. In *Social theory re-wired*, pages 203–213. Routledge, 2023.

Participants

- Qinqyao Ai
Tsinghua University – Beijing, CN
- Mohammad Aliannejadi
University of Amsterdam, NL
- Liesbeth Allein
KU Leuven, BE
- Sophia Althammer
Cohere – München, DE
- Avishek Anand
Delft University, NL
- Nolwenn Bernard
TH Köln, DE
- Niklas Deckers
University of Kassel, DE & hessian.AI, DE
- Gianluca Demartini
The University of Queensland – Brisbane, AU
- Laura Dietz
University of New Hampshire – Durham, US
- Carsten Eickhoff
Universität Tübingen, DE
- Nicola Ferro
University of Padova, IT
- Maik Fröbe
Friedrich-Schiller-Universität Jena, DE
- Norbert Fuhr
Universität Duisburg-Essen, DE
- Marcel Gohsen
Bauhaus-Universität Weimar, DE
- Michael Granitzer
University of Passau, DE
- Faegheh Hasibi
Radboud University Nijmegen, NL
- Sebastian Heineking
Universität Leipzig, DE
- Djoerd Hiemstra
Radboud University Nijmegen, NL
- Adam Jatowt
Universität Innsbruck, AT
- Abhinav Joshi
Indian Institute of Technology Kanpur, IN
- Johannes Kiesel
GESIS – Leibniz Institute for the Social Sciences – Köln, DE
- Wojciech Kusa
NASK National Research Institute – Warsaw, PL
- Sean MacAvaney
University of Glasgow, GB
- Bhaskar Mitra
Independent Researcher, Tiohtià:ke/Montréal, CA
- Josiane Mothe
INSPE, Université de Toulouse, UT2J, UMR5505 CNRS IRIT, FR
- Smaranda Muresan
Barnard College, Columbia University – New York, US
- Jian-Yun Nie
University of Montréal, CA
- Heather O'Brien
iSchool, University of British Columbia – Vancouver, CA
- Birte Platow
TU Dresden, DE
- Martin Potthast
Universität Kassel, DE & hessian.AI, DE
- Mark Sanderson
RMIT University – Melbourne, AU
- Harrisen Scells
Universität Tübingen, DE
- Alan Smeaton
Dublin City University, IE
- Damiano Spina
RMIT University – Melbourne, AU
- Benno Stein
Bauhaus-Universität Weimar, DE
- Johanne Trippas
RMIT University – Melbourne, AU
- Stefan Voigt
Open Search Foundation – Starnberg, DE
- Arjen P. de Vries
Radboud University Nijmegen, NL
- Guido Zuccon
Google Research Australia & The University of Queensland – Brisbane, AU

Report from Dagstuhl Seminar 25392

Specification Engineering: Foundations for the Future of Software Development

Marsha Chechik*[1], Eunsuk Kang*[2], Shahar Maoz*[3], Jan Oliver Ringert*[4], and Allison Sullivan*[5]

1 **University of Toronto, CA.** chechik@cs.toronto.edu
2 **Carnegie Mellon University – Pittsburgh, US.** eunsukk@andrew.cmu.edu
3 **Tel Aviv University, IL.** maoz@cs.tau.ac.il
4 **Bauhaus-Universität Weimar, DE.** jan.ringert@uni-weimar.de
5 **University of Texas at Arlington, US.** allison.sullivan@uta.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 25392 "Specification Engineering: Foundations for the Future of Software Development". Specifications are an essential component in a variety of tasks in software engineering, including software verification, testing, modeling, requirements engineering, and program synthesis. While producing quality specifications has been a longstanding problem, recent advances in AI technologies, such as large-language models (LLMs), make it a timely problem to address from new perspectives. Automatically generating code from a high-level specification will likely emerge as a dominant paradigm for software development in the future. Thus, being able to write, maintain and evolve high quality specifications – the process of **specification engineering** – will become an essential skill for software engineers. This Dagstuhl Seminar brought together leading researchers in software engineering and formal methods to identify foundational problems and build a roadmap for specification engineering as a central activity in future development processes.

Seminar September 21–26, 2025 – https://www.dagstuhl.de/25392
2012 ACM Subject Classification Software and its engineering → Specification languages
Keywords and phrases formal methods, software assurance, software specification, specification engineering
Digital Object Identifier 10.4230/DagRep.15.9.160

1 Executive Summary

Marsha Chechik (University of Toronto, CA)
Eunsuk Kang (Carnegie Mellon University – Pittsburgh, US)
Shahar Maoz (Tel Aviv University, IL)
Jan Oliver Ringert (Bauhaus-Universität Weimar, DE)
Allison Sullivan (University of Texas at Arlington, US)

Formal specifications are mathematically precise descriptions of the behavior or properties of a system. Specifications are an essential component in a variety of tasks in software engineering, including software verification, testing, modeling, requirements engineering, and program synthesis. Despite a wealth of research on techniques and tools that take specifications as input, relatively less has been explored on addressing the challenges of

* Editor / Organizer

Specification Engineering: Foundations for the Future of Software Development, *Dagstuhl Reports*, Vol. 15, Issue 9, pp. 160–182
Editors: Marsha Chechik, Eunsuk Kang, Shahar Maoz, Jan Oliver Ringert, and Allison Sullivan
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

coming up with specifications in the first place, and maintaining them as system requirements evolve. Typically, specifications are assumed to have been created by software engineers, who may not have sufficient training or expertise in specification languages. Little is understood about what makes a specification "correct" or "high-quality", and how to validate specifications to ensure that they accurately reflect a user's intent. Specification tools are also notorious for their poor usability and high learning curve.

While producing quality specifications has been a longstanding problem, recent advances in AI technologies, such as large-language models (LLMs), make it a timely problem to address from new perspectives. Automatically generating code from a high-level specification will likely emerge as a dominant paradigm for software development in the future. Thus, being able to write, maintain and evolve high-quality specifications – the process of specification engineering – will become an essential skill for software engineers. LLMs are also being explored by researchers as a promising way of generating formal specifications from natural language requirements. However, since LLMs themselves do not provide guarantees about the correctness or quality of their output, new methods for validating and improving the quality of generated specifications will be crucial to make them reliable and useful.

This Dagstuhl Seminar "Specification Engineering: Foundations for the Future of Software Development" (25392) brought together leading researchers in software engineering and formal methods to identify foundational problems and build a roadmap for specification engineering as a central activity in future development processes. The seminar was organized around the following questions:

- Quality and Validation: What are key properties of a high-quality specification? How do we debug, validate, and repair specifications for these properties?
- Usability: How do we make it easier for engineers to express and validate their intent in a specification language? How do we make specifications readable and comprehensible?
- Scalable Specification Construction: How do we construct large, complex specifications out of smaller ones? How do we facilitate reuse of specifications? How do we support incremental, modular changes to a specification?
- Specification for/with AI: How do we use and tailor AI-based tools for specification-driven tasks such as code generation and verification? How do we make these models more effective at generating specifications from natural languages?

Activities

The seminar consisted of (1) several invited, "anchoring" talks around the four major topics listed above, (2) a series of shorter, "lightning" talks where participants shared new ideas, open problems, or ongoing projects on the topic of specification, and (3) two sets of breakout discussions. The first set of breakouts was assigned based on the four topics; after the initial discussions, the participants were encouraged to suggest or form different groups based on their topics of interest that emerged. The resulting second set of breakouts were centered around the topic of AI, covering LLMs for specification activities, specification of LLMs, and the use of AI for domain modeling. These activities were interleaved with ad-hoc discussions around the very concept of "specification" itself as well as planning for post-seminar activities and collaborations.

Outcome

Among many stimulating discussions around the topic of specification, two major themes emerged. First, the participants realized that the very idea of "specification" may not be as well-defined or agreed upon as many had previously thought before the seminar. For example, to some participants, a specification had a specific meaning as a type of artifact that describes the expected behavior of a program or a system (e.g., API contracts), while others thought that nearly every software artifact (e.g., code) could be considered a specification. After a seminar-wide discussion, the participants agreed that it would be more meaningful to talk about properties of a specification (e.g., whether it is formal or informal, readable, analyzable, modifiable, for what purpose it is used, etc.,) rather than attempting to define what a specification is (and is not).

Second, many participants agreed that specifications will have an essential role in the age of AI-driven development and provide new opportunities for research as well as engagement with practitioners. For example, natural language prompts are emerging as a common mechanism to specify developers' intent and system requirements, from which an implementation is automatically generated. However, it was also noted that informal, unstructured prompts are not an ideal specification mechanism for developing, debugging, and maintaining complex software systems, and that more structured specification methods are needed to support both developers and AI agents in these tasks. On the other hand, the participants also agreed that traditional specification methods and tools developed by the research community will likely need to be adapted or rethought to support the fuzzy, interactive, and informal ways in which developers collaborate with AI to develop software.

2 Table of Contents

Executive Summary

Marsha Chechik, Eunsuk Kang, Shahar Maoz, Jan Oliver Ringert, and Allison Sullivan 160

Overview of Talks

Management of specifications in the large
Bernhard Rumpe . . . 165

Formal Specification Generation as Design Transformations and the Limits of Automation
Mauricio Castillo-Effen . . . 165

Abstraction Engineering
Benoît Combemale . . . 166

Helping students learn formal specification
Alcino Cunha . . . 167

ML for Specifications for ML
Matthew Dwyer . . . 167

Context-Aware Trace Contracts
Reiner Hähnle . . . 168

Role of Specification Languages in Verification of Neuro-Symbolic Systems and Complex Systems with Machine Learning Components
Ekaterina Komendantskaya . . . 168

Specification Reverse Engineering for Decentralized Applications
Yi Li . . . 169

Exploring Development Methods for Reactive Synthesis Specifications
Shahar Maoz and Jan Oliver Ringert . . . 170

Specification engineering: Notes on usability
Shahar Maoz . . . 170

What Properties Affect Boolean Formula Comprehension in Formal Specifications?
Shahar Maoz . . . 171

Learning LTL Specifications from Demonstrations with Uncertainty
Rômulo Meira-Góes . . . 171

Specification Engineering for Neuro-symbolic Programming and Vice Versa
Federico Mora . . . 172

Temporal Logic Sketching
Daniel Neider . . . 172

Task Models as a Mean to Identify and Justify Automations in Software Programming Tasks
Phillippe Palanque . . . 173

Live Programming for Specs (and Beyond)
Allison Sullivan . . . 174

Specification in the Large at Amazon Web Services
Michael W. Whalen . . . 174

No more garbage in: Validating formal models
Pamela Zave 175

Precision in formal modeling: Why we need it, and how to get it
Pamela Zave 175

The CNA model of network architecture
Pamela Zave 175

Working groups

What is a Specification?
Marsha Chechik and others 176

Specification for and with AI
Ekaterina Komendantskaya, Thorsten Berger, Mauricio Castillo-Effen, Marsha Chechik, Jyotirmoy Deshmukh, Matthew Dwyer, Taylor T. Johnson, Federico Mora, and Daniel Neider 176

Usability of Specifications
Shahar Maoz, José Creissac Campos, Reiner Hähnle, Yi Li, Alexandra Mendes, Phillippe Palanque, Bernhard Rumpe, Kathryn T. Stolee, and Harold Thimbleby . . 177

Specification Quality and Validation
Rômulo Meira-Góes, Alcino Cunha, Eunsuk Kang, and Pamela Zave 178

LLMs for Specifications
Michael W. Whalen, Matthew Dwyer, Lars Grunske, Yi Li, Shahar Maoz, Rômulo Meira-Góes, and Bernhard Rumpe 178

Scalable Construction of Specifications
Michael W. Whalen, Alcino Cunha, Eunsuk Kang, Rômulo Meira-Góes, Jan Oliver Ringert, Allison Sullivan, and Pamela Zave 179

AI for Domain Modeling
Pamela Zave, Alcino Cunha, and Eunsuk Kang 180

Open problems

Let's Verify ChatGPT: What Would We Verify and How Could We Get There? (or Is Neural Network Verification Useful and What's Next?)
Taylor T. Johnson 180

Specification coherence
Harold Thimbleby 181

Participants 182

3 Overview of Talks

3.1 Management of specifications in the large

Bernhard Rumpe (RWTH Aachen, DE)

The management of specifications at scale requires systematic coordination across multiple models written in multiple modeling languages and notations, each with its own semantics and interpretations. Our work focuses on the foundations of model-based software engineering, in particular the semantics of models and the construction of software tools that support precise, analyzable model representations. A central research question is how to formally relate heterogeneous models so that their structures, behavioral properties, and interdependencies remain coherent across levels of abstraction and domains. This includes investigating how symbols of various kinds defined in one model can be connected to in other models to lay the foundation for a robust and maintainable specification management. We want to apply these foundational principles for human-constructed models, but also derived models and potentially AI-generated artifacts, aiming to enable robustly shared syntactically connected models that support automated reasoning, collaborative design, and scalable model evolution. This direction to our belief is a core pillar for a unifying semantic basis for specification engineering in increasingly complex, heterogeneous system engineering environments, where even UML and SysML only cover a partial subset of viewpoints and are internally also not well integrated yet.

3.2 Formal Specification Generation as Design Transformations and the Limits of Automation

Mauricio Castillo-Effen (Lockheed Systems – Arlington, US)

Formal specifications play an increasingly important role in enabling the adoption of formal methods and automated reasoning in industrial settings, yet the process of deriving them from natural-language requirements remains poorly understood. Current approaches to specification generation often treat it as a translation problem, assuming that requirements encode the information needed for formalization. We question this assumption, particularly in the context of rapid, iterative Systems Engineering and the growing interest in applying generative AI to support these processes.

In this talk, we presented preliminary results from studying the generation of formal specifications as a design transformation process characterized by a gradual reduction of epistemic uncertainty. We took two complementary perspectives. First, we studied the cognitive processes applied by humans to transform informal requirements into increasingly formal design artifacts. Second, we evaluated whether agentic generative AI systems could perform similar transformations with comparable epistemic rigor.

To structure this analysis, we used Gero's Function–Behavior–Structure (F-S-B) ontology as a model of the design process and its successive refinements. Using this concept, we carried out an empirical study on a design challenge that involved exploring a large design

space with the goal of generating adaptable search-and-rescue drones. We tasked human participants and different configurations of agentic generative AI systems with the same objectives and constraints.

Our findings indicated that while agentic LLM systems could generate syntactically well-formed artifacts, they struggle with context preservation, controlled abstraction, and uncertainty management. Common failure modes included premature concretization, cascading semantic errors, and ontology drift, resulting in artifacts that resemble specifications without supporting their intended epistemic role ("specifictions"). In contrast, human designers rely on implicit contextual knowledge and iterative reformulation to progressively stabilize meaning before formalization.

We conclude that generating formal specifications cannot be reliably delegated to autonomous agents without well-defined mechanisms for transparent uncertainty management, contextual grounding, and human oversight. These results suggest that future AI-assisted approaches to specification engineering should emphasize not only automation but also support the human users' ability to steer the formalization process.

3.3 Abstraction Engineering

Benoît Combemale (INRIA – Rennes, FR)

Joint work of Nelly Bencomo, Jordi Cabot, Marsha Chechik, Betty H.C. Cheng, Benoît Combemale, Andrzej Wąsowski, Steffen Zschaler
Main reference Nelly Bencomo, Jordi Cabot, Marsha Chechik, Betty H. C. Cheng, Benoît Combemale, Andrzej Wąsowski, Steffen Zschaler: "Abstraction Engineering", CoRR, Vol. abs/2408.14074, 2024.
URL http://dx.doi.org/10.48550/ARXIV.2408.14074

Modern software-based systems operate under rapidly changing conditions and face ever-increasing uncertainty. In response, systems are increasingly adaptive and reliant on artificial-intelligence methods. In addition to the ubiquity of software with respect to users and application areas (e.g., transportation, smart grids, medicine, etc.), these high-impact software systems necessarily draw from many disciplines for foundational principles, domain expertise, and workflows. Recent progress with lowering the barrier to entry for coding has led to a broader community of developers, who are not necessarily software engineers. As such, the field of software engineering needs to adapt accordingly and offer new methods to systematically develop high-quality software systems by a broad range of experts and non-experts. In [1], we look at these new challenges and propose to address them through the lens of Abstraction. Abstraction is already used across many disciplines involved in software development – from the time-honored classical deductive reasoning and formal modeling to the inductive reasoning employed by modern data science. The software engineering of the future requires Abstraction Engineering – a systematic approach to abstraction across the inductive and deductive spaces. We discuss the foundations of Abstraction Engineering, identify key challenges, highlight the research questions that help address these challenges, and create a roadmap for future research.

References

1 N Bencomo, J Cabot, M Chechik, BHC Cheng, B Combemale, S Zschaler. *Abstraction engineering*. arXiv (`https://arxiv.org/abs/2408.14074`), 2024.

3.4 Helping students learn formal specification

Alcino Cunha (University of Minho, PT)

Joint work of Alcino Cunha, Nuno Macedo, José Creissac Campos, Iara Margolis, Emanuel Sousa
Main reference Alcino Cunha, Nuno Macedo, José Creissac Campos, Iara Margolis, Emanuel Sousa: "Assessing the impact of hints in learning formal specification", in Proc. of the 46th International Conference on Software Engineering: Software Engineering Education and Training, SEET@ICSE 2024, Lisbon, Portugal, April 14-20, 2024, pp. 151–161, ACM, 2024.
URL http://dx.doi.org/10.1145/3639474.3640050

Alloy4Fun is a web tool for helping students self study the Alloy formal specification language. In particular, it allows the creation of specification challenges where students are asked to formally specify natural language requirements. If the specification is incorrect, a counter-example is depicted graphically, as usual in Alloy. This counter-example can be seen as a hint that helps the student progress towards the correct specification. Unfortunately, we had some anecdotal evidence that students sometimes struggle to understand such hints. Recently, we conducted a large user study to assess the impact of this and other kinds of hints in learning formal specification. In this talk I presented the design of this study and briefly discussed its results. The main conclusion of the study is that none of the studied hints had an impact on learning retention, and only giving the student precise error locations had an impact on immediate performance. Finally, I discussed some potential uses of LLM in this context, namely using LLMs to help students understand counter-examples or wrong specifications.

3.5 ML for Specifications for ML

Matthew Dwyer (University of Virginia – Charlottesville, US)

Formalizing functional requirements as specifications can enable a variety of powerful validation and verification (V&V) approaches to be applied. Such specifications formulate a precondition, which describes a set of system inputs, and an associated postcondition, which constraints system output for those inputs. However, precisely formulating requirements for ML-enabled systems that process raw sensor data, e.g., image, lidar, can be very challenging.

In this talk we will describe why it can be challenging to directly formulate specifications for such systems and propose an alternative approach that develops black-box models for specification preconditions. Inputs generated from such models conform to preconditions with high-probability and are both realistic and diverse – desirable properties for V&V. Consequently, observed system behavior for those inputs can then be checked against formalizations of postconditions to enable a form of spec-based V&V.

The approach to developing these precondition models begins with the standard approach of formulating natural language statements of functional requirements over a glossary of domain-specific terms that define semantic features of inputs. It then combines several different ML techniques to construct a generative model for the precondition which can be leveraged for V&V. In this way, ML supports the development of specification models for V&V of ML-enabled systems.

As the saying goes "All models are wrong, but some are useful" and the first part applies to the models we generate, but preliminary results suggest that the second part may also apply. We will end with a series of questions related to how and when such techniques might be usefully applied that we hope initiates fruitful discussion.

3.6 Context-Aware Trace Contracts

Reiner Hähnle (TU Darmstadt, DE)

Joint work of Reiner Hähnle, Eduard Kamburjan, Marco Saletta

We illustrate the usage of Context-Aware Trace Contracts (for short: CATs) by way of an example. CATs are a systematic approach to specify non-procedure local behavior. Technically, they consist of symbolic expressions specifying the assumed behavior of the callers before a procedure enters its contract, the behavior a procedure guarantees, and the behavior expected to happen in the continuation after termination. This generalizes state-based, Hoare-style specification triples.

References

1 Hähnle, R., Kamburjan, E., Scaletta, M.: Context-aware trace contracts. In: De Boer, F., Damiani, F., Hähnle, R., Johnsen, E.B., Kamburjan, E. (eds.) Active Object Languages: Current Research Trends. LNCS, vol. 14360, pp. 292–325. Springer, Cham (2024).

3.7 Role of Specification Languages in Verification of Neuro-Symbolic Systems and Complex Systems with Machine Learning Components

Ekaterina Komendantskaya (Heriot-Watt University – Edinburgh, GB)

Machine Learning (ML) is increasingly used to implement components of Cyber-Physical Systems – i.e. systems that interact with the physical (continuous) world and at the same time have (digital) programmed components. Usually, ML components are used on the intersection of these two, i.e. a neural network usually processes a sensor input and gives class predictions or discrete commands. For example, in a car, a neural controller may measure how close the car is to an obstacle, and give a command to brake. In a medical application, a neural network may measure the patient's temperature, blood pressure or blood composition and decide on a dose of administered drugs. Both cases are examples of safety-critical systems, i.e. systems whose failure potentially endangers health and life of the users. Formally verifying that such systems "do not go wrong" is one of the biggest challenges of the day.

One of the most known problems in the domain is the problem of "a missing spec" that refers to the fact that ML components are obtained via data-driven optimisation procedures, and come without any clear formal specification. However, often specifications are available, thanks to the knowledge of the safety requirements concerning the symbolic components of the system in question. E.g. we may know the critical distance or critical dose, after which

the system goes into an unsafe state; and these can result in meaningful safety specifications. Such cases are a sweet spot for formal specification and verification. In my talk, I discussed the benefits of deploying a domain-specific language Vehicle for writing specifications of properties of ML components of neuro-symbolic systems. Vehicle allows users to specify the properties of the neural components of neuro-symbolic programs once, and then safely compile the specification to other interfaces (ML solvers, interactive theorem provers, ML backends) using a tailored typing and compilation procedure. I gave a high-level overview of Vehicle's overall design, its interfaces and compilation & type-checking procedures, and then demonstrated its utility by formally verifying the safety of a simple autonomous car controlled by a neural network, operating in a stochastic environment with imperfect information.

Vehicle is available at: `https://github.com/vehicle-lang/vehicle`.

3.8 Specification Reverse Engineering for Decentralized Applications

Yi Li (Nanyang TU – Singapore, SG)

Joint work of Yi Li, Ye Liu,, Yixuan Liu, Zhiyang Chen, Cyrille Artho, Chengxuan Zhang
Main reference Ye Liu, Yue Xue, Daoyuan Wu, Yuqiang Sun, Yi Li, Miaolei Shi, Yang Liu: "PropertyGPT: LLM-driven Formal Verification of Smart Contracts through Retrieval-Augmented Property Generation", in Proc. of the 32nd Annual Network and Distributed System Security Symposium, NDSS 2025, San Diego, California, USA, February 24-28, 2025, The Internet Society, 2025.
URL https://www.ndss-symposium.org/ndss-paper/propertygpt-llm-driven-formal-verification-of-smart-contracts-through-retrieval-augmented-property-generation/

Smart contracts are computer programs running on blockchains to implement Decentralized Applications. The absence of contract specifications hinders routine tasks, such as contract verification, security auditing, and effective test generation, leading to vulnerabilities and increased development costs. In this talk, we introduce the concept of Specification Reverse Engineering for smart contracts and presented a summary of our past works in this field. The two main approaches are: (1) "learning from the past", where benign scenarios from smart contract transaction histories are used to generate function-level invariants, e.g., InvCon [1] and InvCon+ [2], and contract-level behavior models, e.g., SMCon [3] and SPCon [4]; and (2) "learning from each other", where specifications of other similar smart contracts can be reused in constructing new specifications, e.g., Trace2Inv [5] and PropertyGPT [6].

References

1 Ye Liu and Yi Li. Oct. 2022. InvCon: a dynamic invariant detector for Ethereum smart contracts. In Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering (ASE), pages 1–4.
2 Ye Liu, Chengxuan Zhang, and Yi Li. 2025. Automated invariant generation for Solidity smart contracts. IEEE Transactions on Dependable and Secure Computing.
3 Ye Liu, Yixuan Liu, Yi Li, and Cyrille Artho. Mar. 2025. Specification mining for smart contracts with trace slicing and predicate abstraction. In Proceedings of the 32nd IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER).
4 Ye Liu, Yi Li, Shang-Wei Lin, and Cyrille Artho. July 2022. Finding permission bugs in smart contracts with role mining. In Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA), pages 716–727, New York, NY, USA. ACM.
5 Zhiyang Chen, Ye Liu, Sidi Mohamed Beillahi, Yi Li, and Fan Long. July 2024. Demystifying invariant effectiveness for securing smart contracts. In Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering (FSE), volume 1 of number FSE, pages 1772–1795. ACM New York, NY, USA.

6 Ye Liu, Yue Xue, Daoyuan Wu, Yuqiang Sun, Yi Li, Miaolei Shi, and Yang Liu. Feb. 2025. PropertyGPT: LLM-driven formal verification of smart contracts through retrieval-augmented property generation. In Proceedings of 32nd Annual Network and Distributed System Security Symposium (NDSS).

3.9 Exploring Development Methods for Reactive Synthesis Specifications

Shahar Maoz (Tel Aviv University, IL), and Jan Oliver Ringert (Bauhaus-Universität Weimar, DE)

Joint work of Shahar Maoz, Dor Ma'ayan, Jan Oliver Ringert
Main reference Dor Ma'ayan, Shahar Maoz, Jan Oliver Ringert: "Exploring Development Methods for Reactive Synthesis Specifications", ACM Trans. Softw. Eng. Methodol., Association for Computing Machinery, 2025.
URL http://dx.doi.org/10.1145/3767159

Reactive synthesis is an automated procedure to obtain a correct-by-construction reactive system from its temporal logic specification. Despite significant research progress in the past decades, reactive synthesis is still in an early stage of use. Previous studies found that the lack of development methods for reactive synthesis specifications is one barrier to its wider adoption. In this paper, we adapt two development methods, an incremental method and a modular method, to the context of reactive synthesis specifications. The methods are based on existing software development methods on the one hand and studies about reactive synthesis on the other hand. Then, we report on an exploratory case study in which participants developed specifications using the two methods. We evaluated the methods using a mixed-method analysis that combines grounded theory analysis of Slack communication with participants, quantitative exploratory data analysis of the synthesis IDE usage logs, and qualitative independent expert review of the final specifications. Our findings show clear benefits of modular specification development in terms of ease of planning, synthesis time, fewer unrealizability issues, and faster debugging. However, the incremental development method was more natural and easy to use, and specifications developed incrementally were also easier to validate during the development process.

3.10 Specification engineering: Notes on usability

Shahar Maoz (Tel Aviv University, IL)

I discuss the challenge of usability in the context of specifications. Usability consists of learnability, effectiveness, efficiency, memorability, error prevention and recovery, and user satisfaction. In the context of specification, these are manifested in terms of language – including syntax, semantics, and abstractions, and in terms of analysis tools – their input, output, and methodology. I will briefly discuss two examples for the challenge. First, the case of temporal logics and specification patterns. While LTL is known for being difficult to read and write correctly, and specification patterns were suggested as a means to describe

common properties, very little research has been done on their usability, i.e., on patterns learnability, effectiveness, efficiency, etc. Second, the case of an unrealizable core. While specifications for reactive synthesis are often unrealizable, and computing an unrealizable core (minimal unrealizable subset) was suggested as a means to localize the fault, very little research has been done on unrealizable core's usability, again, on its learnability, effectiveness, etc. Finally, I will present two recent example projects that deal with the comprehension of specifications and with the process of developing them.

3.11 What Properties Affect Boolean Formula Comprehension in Formal Specifications?

Shahar Maoz (Tel Aviv University, IL)

Joint work of Shahar Maoz, Ilia Shevrin
Main reference Ilia Shevrin, Shahar Maoz: "What Properties Affect Boolean Formula Comprehension in Formal Specifications?", ACM Trans. Softw. Eng. Methodol., Association for Computing Machinery, 2025.
URL http://dx.doi.org/10.1145/3744557

Writing formal specifications is an important yet challenging aspect of software engineering. Correct specifications facilitate verification efforts and reduce bugs. However, the declarative nature of specifications differs from the imperative approach of most common programming languages, and software engineers often perceive formal methods as difficult. Arguably, guidelines and tools for writing readable specifications should lower the barrier to formal methods adoption. In this work, we focus on Boolean formulas, a fundamental building block of specifications. Analogous to research on code comprehension, we conducted an experiment that attempts to identify what properties affect Boolean formula comprehension by software engineers. To this end, we collected 59 representative Boolean formulas and tested how various syntactic properties, such as negation symbol count and nesting level, affect comprehension task response times and correctness. Our experiment with 181 participants shows that eliminating negation symbols and decreasing operator count are among the most significant factors that improve comprehension. We use these empirical results to derive a reading complexity score and develop a fast regression-based refactoring algorithm for Boolean formulas. Finally, we conducted a follow-up experiment with 57 participants, which provided strong evidence for the algorithm's effectiveness in improving comprehension.

3.12 Learning LTL Specifications from Demonstrations with Uncertainty

Rômulo Meira-Góes (Pennsylvania State University – University Park, US)

Joint work of Rômulo Meira-Góes, Constantino Lagoa, Parastou Fahim
Main reference Parastou Fahim, Constantino Lagoa, Rômulo Meira-Góes: "Learning Linear Temporal Specifications from Demonstrations with Uncertainty", submitted to The 2026 American Control Conference.

Learning temporal logic specifications from system demonstrations is essential for tasks such as formal verification and controller synthesis, especially in safety-critical domains. Existing approaches typically assume demonstrations are correct or only affected by misclassification errors. In practice, however, system traces are often uncertain or incomplete due to sensor

faults, measurement errors, or data loss. We present a framework for learning minimal Linear Temporal Logic (LTL) formulas from demonstrations with uncertainty. Our approach models uncertainty via Hamming distance to generate possible estimates around each observed trace, which are grouped with constraints requiring that at least one trace per group is consistent with the learned formula. Our problem is then reduced to an equivalent Pseudo-Boolean Optimization. We evaluate our method against state-of-the-art LTL learning approaches and show that it recovers specifications that more closely align with ground-truth formulas under uncertainty.

3.13 Specification Engineering for Neuro-symbolic Programming and Vice Versa

Federico Mora (University of Waterloo, CA)

Neuro-symbolic programming systems often use a machine learning (neuro) component to generate candidate programs and a program analysis (symbolic) component to check candidate programs. When the symbolic component finds an issue, the neuro component tries again. Eudoxus [1] is one example neuro-symbolic system that follows this scheme. This talk discusses the challenges of generating programs that use APIs through similar neuro-symbolic approaches. For such neuro-symbolic systems to work as a whole, we need good specifications for all relevant APIs. When a specification is too restrictive, the symbolic component will block valid candidate programs. When the specification is too forgiving, the symbolic component will allow incorrect candidate programs. Specifically, this talk discusses the challenges of and opportunities for engineering good specifications at scale in the neuro-symbolic programming context.

References

1 Federico Mora, Justin Wong, Haley Lepe, Sahil Bhatia, Karim Elmaaroufi, George Varghese, Joseph E. Gonzalez, Elizabeth Polgreen, Sanjit A. Seshia: Synthetic Programming Elicitation for Text-to-Code in Very Low-Resource Programming and Formal Languages. NeurIPS 2024

3.14 Temporal Logic Sketching

Daniel Neider (TU Dortmund, DE)

Joint work of Simon Lutz, Daniel Neider, Rajarshi Roy
Main reference Simon Lutz, Daniel Neider, Rajarshi Roy: "Specification Sketching for Linear Temporal Logic", in Proc. of the Automated Technology for Verification and Analysis – 21st International Symposium, ATVA 2023, Singapore, October 24-27, 2023, Proceedings, Part II, Lecture Notes in Computer Science, Vol. 14216, pp. 26–48, Springer, 2023.
URL http://dx.doi.org/10.1007/978-3-031-45332-8_2

Temporal Logic Sketching is a novel approach designed to simplify the process of writing formal specifications. The central idea is that an engineer can provide a partial formula, called a sketch, in which components that are difficult to formalize may be left unspecified. Given a set of examples describing desired and undesired system behaviors, the goal of a sketching algorithm is to complete the sketch so that the resulting specification is consistent with the provided examples.

This talk presents recent advances in specification sketching and surveys existing approaches for various temporal logics, including Linear Temporal Logic (LTL), Signal Temporal Logic (STL), Metric Temporal Logic (MTL), Property Specification Language (PSL), Computation Tree Logic (CTL), and Alternating-time Temporal Logic (ATL). It highlights both the challenges inherent in this paradigm and the opportunities it offers for specification engineering. In addition, the talk outlines key theoretical contributions, such as a comprehensive complexity analysis of learning logical formulas from data.

3.15 Task Models as a Mean to Identify and Justify Automations in Software Programming Tasks

Phillippe Palanque (Toulouse University, FR)

Programming is usually considered as a difficult task [1] and even some metrics about the cognitive aspect of this difficulty have been proposed [4]. Associated with this difficulty is the diversity of tasks such as structuring code, writing code, testing code or even reading and understanding error messages [2]. One key contributing factor to addressing this difficulty is training and learning [4] and another key one is the programming environment used for programming [3].

This presentation has argued that building and exploiting both high-level and detailed descriptions of programming tasks would provide multiple benefits in terms of training and learning in general [5] (but also specifically in the area of safety critical systems such as aeronautics).

Identifying and describing tasks in such a way might appear as a niche and cumbersome work, however, with the multiplicity of software assistants [6] and with virtually any modern IDE proposing such tools while also allowing for custom add-ons, listing, describing and modeling developers' tasks is needed if we want to be able to compare the usability and performance of such tools.

As argued in [7] the complexity of tasks is not an intrinsic value but a combined value of the tasks themselves and the system that is used to perform the tasks. This is known as the task-artefact cycle, which requires, at design time of an interactive system, the identification of the tasks and the evolution of their complexity when modifications are made on the system.

In this presentation we propose the exploitation of the HAMSTERS task modeling notation to represent programmers' tasks and to assess their complexity on a concrete IDE. We also argue that automating some of these tasks with software assistants (which is a very common activity in the area of software engineering (see a recent survey in [6]) should be assessed in terms of how they demonstrate a reduction of the complexity of these tasks and how learning how to perform those tasks is improved.

References

1 B. A. Becker, P. Denny, J. Finnie-Ansley, A. Luxton-Reilly, J. Prather, and E. A. Santos. Programming Is Hard – Or at Least It Used to Be: Educational Opportunities and Challenges of AI Code Generation. In Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1 (SIGCSE 2023). ACM, 500–506.

2 B. A. Becker, P. Denny, R. Pettit, D. Bouchard, D. J. Bouvier, B. Harrington, A. Kamil, A. Karkare, C. McDonald, P.-M. Osera, J. L. Pearce, and J. Prather. Compiler Error Messages Considered Unhelpful: The Landscape of Text-Based Programming Error Message Research. In Proc. of the Working Group Reports on Innovation and Technology in Computer Science Education (ITiCSE-WGR '19). ACM, 177–210.

3 C. M. Lewis. How programming environment shapes perception, learning and goals: logo vs. scratch. In Proceedings of the 41st ACM technical symposium on Computer science education (SIGCSE '10). ACM, 346–350.
4 B. Alwis, G. C. Murphy, and S. Minto. Creating a cognitive metric of programming task difficulty. In Proceedings of the 2008 international workshop on Cooperative and human aspects of software engineering (CHASE '08). ACM, 29–32.
5 C. Martinie, D. Navarre, P. Palanque, and C. Fayollas. A generic tool-supported framework for coupling task models and interactive applications. ACM SIGCHI Symposium on Engineering Interactive Computing Systems (EICS '15). ACM, 244-253.
6 S. Leblanc, M. Burgueño, L. Cabot, J. Le Pallec, X, Gérard, Software assistants in software engineering: A systematic mapping study. Softw: Pract Exper. 2023; 53(3): 856–892.
7 C. Martinie, P. Palanque, E. Bouzekri, A. Cockburn, A. Canny, and Eric Barboni. Analysing and Demonstrating Tool-Supported Customizable Task Notations. Proc. ACM Hum.-Comput. Interact. 3, EICS, Article 12 (jun 2019), 26 pages.

3.16 Live Programming for Specs (and Beyond)

Allison Sullivan (University of Texas at Arlington, US)

In a quest to explore ways we could broaden adoption of specification languages, this talk highlights some new research directions Dr. Sullivan is exploring. The topics center around the concept of "live programming" which is a body of work dedicated to building development environments that interweave writing and executing programs. There are interesting challenges to do this for traditional imperative languages (e.g. Java) but for a specification language, highlighting output changes is as "simple" as demonstrating why two formulas differ. So with that in mind, where can we take live programming for specs? What specification languages could this apply to?

3.17 Specification in the Large at Amazon Web Services

Michael W. Whalen (Amazon Inc. – Minneapolis, USA & The University of Minnesota – Minneapolis, USA)

AWS is perhaps the world's largest user of automated reasoning, which is driven by specifications. In this talk, I will first give an overview of different projects using specifications and difficulties we have, then focus on the grand challenge of creating specifications for reasoning across multiple microservices to prove customer-relevant properties.

3.18 No more garbage in: Validating formal models

Pamela Zave (Princeton University, US)

This talk presents an overview of validation, as I see it and practice it. It offers and justifies three complementary definitions of validity, and how a formal model can be validated for each. It also presents two opinions, formed during my experience of validating many formal models, on how validation should be done.

3.19 Precision in formal modeling: Why we need it, and how to get it

Pamela Zave (Princeton University, US)

We used to say that a formal model should be "complete, consistent, and unambiguous." And I always wondered about "unambiguous," because a well-formed formal model cannot be ambiguous. This talk shows how precision–the traditional partner of accuracy–applies to formal models, and describes what unambiguity is supposed to describe. The talk also applies the thinking of Michael Jackson to birthday cakes.

3.20 The CNA model of network architecture

Pamela Zave (Princeton University, US)

Compositional Network Architecture is a general, reusable formal model of network architecture. It is the basis for a recent book on the subject–the first networking book to be based on a formal model!

This talk introduces the Alloy model, which is included in the seminar materials, with emphasis on the seminar topics: (1) It has been thoroughly validated (62 percentage of the code is strictly for VALIDATION). (2) The model is big and complex, but would be less so if it could be separated into views. I explain the desired decomposition, which I wish were supported by tools for SCALABLE CONSTRUCTION. (3) The model is also being translated into Isabelle so we can prove theorems about it. The translation, and the motivations for it, is related to many questions about USABILITY.

4 Working groups

4.1 What is a Specification?

Marsha Chechik (University of Toronto, CA) and others

This breakout tackled fundamental definitional questions about specifications in the age of AI. They built on a list of essential properties of specifications that the entire seminar group came up with during an earlier discussion: adequacy to the problem, including: formality, readability, modifiability, precision, incrementality, comprehensibility, decomposability, and fitness to use case. The breakout group compared three development paradigms across multiple dimensions of properties: vibe coding (where the specification is the history of prompts and feedback iterated until user acceptance), specifications for correctness of AI components (assuming we cannot fully specify behavior, using lists of examples and accepting black-boxes as valid components), and traditional code contracts (formal, complete specifications like sorting algorithm contracts with precise logical properties). An outcome of this exercise was that specifications remain fundamentally about abstraction, communication, and intent – but the forms they take and the guarantees they provide may need to evolve in the AI era.

4.2 Specification for and with AI

Ekaterina Komendantskaya (Heriot-Watt University – Edinburgh, GB), Thorsten Berger (Ruhr-Universität Bochum, DE), Mauricio Castillo-Effen (Lockheed Systems – Arlington, US), Marsha Chechik (University of Toronto, CA), Jyotirmoy Deshmukh (USC – Los Angeles, US), Matthew Dwyer (University of Virginia – Charlottesville, US), Taylor T. Johnson (Vanderbilt University – Nashville, US), Federico Mora (University of Waterloo, CA), Daniel Neider (TU Dortmund, DE)

This group addressed the dual challenge of specifying AI systems themselves and using AI to assist with specification tasks. For specifying AI, key problems included handling the inherent uncertainty in task definitions, understanding what constitutes a "bug" in AI systems, managing the distance between problem space (context, users) and embedding space, and dealing with AI's fundamentally different characteristics – built for vague problems, subject to novel attacks, operating in large state spaces that are hard to explore systematically. The group distinguished between approaches for small language models (built for single tasks, easier to specify and potentially verify) versus large language models, and discussed the gap between task-specific verification and the multi-task, multi-modal nature of modern AI systems.

For AI-assisted specification, the group explored how LLMs could help translate high-level intent to lower-level behavioral specifications, convert natural language to temporal logic, and translate between specification languages. Promising approaches included neuro-symbolic solutions, constraint-guided generation, tools like Amazon's Kiro for conversational specification elicitation, and the concept of LLMs as judges for tasks without conventional

oracles. However, fundamental questions remain about the role of humans in AI-assisted specification engineering, establishing notions of correctness (full versus partial, with or without certificates), handling the challenges of "vibe coding" where huge pull requests lead to knowledge loss, and managing the lifecycle from specification through manual code customization. The group emphasized the need for systematic frameworks, better traceability between high-level features and code artifacts, and principled approaches to collaboration between humans and AI in the specification process.

4.3 Usability of Specifications

Shahar Maoz (Tel Aviv University, IL), José Creissac Campos (University of Minho, PT), Reiner Hähnle (TU Darmstadt, DE), Yi Li (Nanyang TU – Singapore, SG), Alexandra Mendes (University of Porto, PT), Phillippe Palanque (Toulouse University, FR), Bernhard Rumpe (RWTH Aachen, DE), Kathryn T. Stolee (North Carolina State University – Raleigh, US), Harold Thimbleby (Swansea University, GB)

This breakout explored usability in the context of formal specification languages, tools, and methods through the lenses of established HCI frameworks such as ISO 9241 and Nielsen's usability criteria. Participants emphasized that usability spans learnability, efficiency, effectiveness, and user satisfaction, while user experience further encompasses emotional, aesthetic, and value-driven dimensions. A recurring theme was that specification work is a creative, cognitively demanding activity involving diverse users – domain experts, programmers, students, formal methods specialists – each with different needs and expectations. As a result, understanding usability requires identifying the tasks users perform (e.g., expressing properties, validating interpretations, refactoring expressions, checking consistency) and studying how tools can support these tasks.

The group identified research gaps in generalizing usability studies, understanding trade-offs inherent in designing notation, and supporting mental-model alignment through visualization, navigation, and incremental interaction. Existing frameworks – such as the cognitive dimensions of notations, the "physics of notations," and studies on programmer usability – offer valuable foundations but do not yet address the full complexity of specifying, debugging, or evolving formal models. Promising directions include supporting transformation and refactoring workflows, leveraging domain-specific or embedded DSLs to reduce cognitive distance for domain experts, and building persuasive interfaces that gently guide users toward sound practices. Participants also highlighted the need for benchmarks and representative tasks to systematically evaluate usability in specification contexts.

4.4 Specification Quality and Validation

Rômulo Meira-Góes (Pennsylvania State University – University Park, US), Alcino Cunha (University of Minho, PT), Eunsuk Kang (Carnegie Mellon University – Pittsburgh, US), Pamela Zave (Princeton University, US)

The breakout on Quality and Validation examined what it means for a specification to be "high-quality" and how one can systematically validate it. Participants noted that the field lacks shared definitions for quality attributes of specifications – attributes that range from those expressible purely in formal terms to those that also depend on informal understanding of the domain. Validation was characterized as ensuring correspondence between formal models and the real world or stakeholders' mental models, encompassing accuracy, completeness, and generality. The group emphasized that, although examples exist (e.g., testing formal models, scenario or predicate generation, Zave and Jackson's "turnstile" work), these methods have not sufficiently raised awareness of validation's importance or made validation accessible across domains.

The group discussed several barriers to adoption: validation is under-taught, often domain-specific, and many engineers undervalue domain modeling. Physics-based disciplines offer instructive analogies where domain constraints assist validation within error bounds, but formal specification practice rarely incorporates similar reusable frameworks. Looking ahead, the group identified machine learning and large language models (LLMs) as both a challenge and an opportunity. As tools increasingly use AI to generate, analyze, or validate specifications, the community must develop ways to validate the outputs of such tools – and potentially use LLMs themselves as aids for scenario generation, explanation, and model exploration. This, they argued, presents a strategic opening to reassert the importance of validation in the broader software engineering ecosystem.

4.5 LLMs for Specifications

Michael W. Whalen (Amazon Inc. – Minneapolis, USA & The University of Minnesota – Minneapolis, USA), Matthew Dwyer (University of Virginia – Charlottesville, US), Lars Grunske (HU Berlin, DE), Yi Li (Nanyang TU – Singapore, SG), Shahar Maoz (Tel Aviv University, IL), Rômulo Meira-Góes (Pennsylvania State University – University Park, US), Bernhard Rumpe (RWTH Aachen, DE)

This breakout explored how large language models can support the creation, transformation, and understanding of specifications. The group developed a taxonomy of LLM uses – ranging from generating and translating specification artifacts, to refining and repairing them, to assisting with semantic comparison and cross-validation. LLM strengths lie especially in bridging informal and formal representations, offering multiple interpretations of ambiguous requirements, and generating explanations that help humans understand complex artifacts. The group noted that LLMs may accelerate tasks such as formalizing requirements, migrating between specification languages, producing test cases, and supporting exploratory design-space analysis.

At the same time, participants emphasized that LLMs must be used within a carefully designed human–AI workflow. Humans remain essential for comprehension, validation, and selection among alternative candidates. Concerns included dependence on input quality (with code often yielding better results than natural language), sensitivity to bias when both code and specifications are generated by the same model, and the need for diversity across LLM sources to enable cross-checking. The group strongly rejected the idea of using LLMs for verification itself, instead positioning them as generators, explainers, and collaborators rather than judges. Open questions remain about whether traditional specifications are still needed in an era of "vibe-coding," how to effectively validate and select between LLM-generated alternatives, and optimal approaches for ensuring specification diversity while avoiding algorithmic bias.

4.6 Scalable Construction of Specifications

Michael W. Whalen (Amazon Inc. – Minneapolis, USA & The University of Minnesota – Minneapolis, USA), Alcino Cunha (University of Minho, PT), Eunsuk Kang (Carnegie Mellon University – Pittsburgh, US), Rômulo Meira-Góes (Pennsylvania State University – University Park, US), Jan Oliver Ringert (Bauhaus-Universität Weimar, DE), Allison Sullivan (University of Texas at Arlington, US), Pamela Zave (Princeton University, US)

This group tackled the fundamental challenge of constructing large, complex specifications from smaller components, examining issues of composition both within single formalisms and across heterogeneous languages and logics. Key problems included how to assemble proofs about different system elements into coherent whole-system arguments, how to facilitate specification reuse across systems and abstraction levels, and how to support incremental, modular changes while managing proof maintenance effort. The group also addressed methodological questions around distributing specification construction across engineering teams and developing specification product lines.

Critical scalability concerns extended beyond mere size to include verification time, incremental analysis capabilities, and the challenge of maintaining formal arguments as specifications evolve. The group discussed several technical approaches including contract-based decomposition, component-based architectures with information hiding (such as AADL with local component proofs), and the use of equivalence relations between abstraction levels. A recurring theme was the tension between code-level proofs of individual components and higher-level API abstractions, particularly when implementations may not perfectly refine their specifications, raising questions about how to compose partial correctness guarantees into system-level assurance arguments.

4.7 AI for Domain Modeling

Pamela Zave (Princeton University, US), Alcino Cunha (University of Minho, PT), Eunsuk Kang (Carnegie Mellon University – Pittsburgh, US)

This focused group examined the critical but often neglected practice of domain modeling, distinguishing sharply between productive and counterproductive uses of LLMs in this space. The group identified a "bad idea" – encouraging non-experts to do domain modeling with LLM assistance, since non-experts cannot properly evaluate LLM outputs – and a "good idea" – domain experts using LLMs as collaborative tools to help create quality domain models. For experts, LLMs can write domain models in natural or formal languages, answer validation questions, help get started, encourage continuation, and stimulate thinking. The group illustrated their approach with a practical example of validating assumptions in safety-critical systems, showing how LLMs can identify edge cases and failure scenarios that experts might overlook.

Three key research directions emerged: training data (determining whether to use large public models or small specialized ones, whether to restrict training to authoritative or proprietary data, and whether including diverse sources like science fiction might be valuable), prompt engineering (teaching experts to write well-structured domain models and validate them successfully, recognizing that prompt size and quality are crucial, and devising methods that guide experts through appropriate steps while encouraging independent critical thinking), and "meta-prompt-engineering" (asking the LLM how to write good prompts for domain modeling tasks). The emphasis throughout was on LLMs as tools to augment expert judgment rather than replace it.

5 Open problems

5.1 Let's Verify ChatGPT: What Would We Verify and How Could We Get There? (or Is Neural Network Verification Useful and What's Next?)

Taylor T. Johnson (Vanderbilt University – Nashville, US)

Main reference Taylor T. Johnson: "Is Neural Network Verification Useful and What Is Next?." Proceedings of the 61st Allerton Conference on Communication, Control, and Computing, 2025.
URL https://hdl.handle.net/2142/130315

Due to the advent of small language models (SLMs) especially for agentic artificial intelligence (AI), we suggest a challenge to verify a realistic foundation model (or large language model [LLM]) like ChatGPT. Could we verify ChatGPT, what would we verify, and how could we do it? We review recent results in neural network verification and recent scalability as demonstrated in the Verification of Neural Networks Competition (VNN-COMP), along with the limitations and usefulness of these approaches, detailed more in [1]. The neural network verification problem is given a trained neural network and a specification often formalized as preconditions and postconditions represented by sets in the input and output spaces of the neural network – prove the network satisfies the specification, and amounts to

showing the neural network maps any element of the precondition into the postcondition. More realistically, we suggest trying to verify a fully open SLM like Ai2's OLMo2 series (open code, data, weights, etc.) like OLMo2 1B, and this guiding grand challenge to verify an SLM (or LLM) will illustrate areas of need in formal methods for specification and verification of transformer architecture models.

References

1 Taylor T. Johnson, "Is Neural Network Verification Useful and What Is Next?", Proceedings of the 61st Allerton Conference on Communication, Control, and Computing, 2025.

5.2 Specification coherence

Harold Thimbleby (Swansea University, GB)

We take it for granted that formal methods and specifications are a good idea, but we don't always seem to have any good terminology to support clear and productive arguments. This paper introduces and defines the terms "coherence" and "views" as contributing to terminology to help focus and make such arguments clearer.

This paper motivates the term coherence with the example of thinking about helping developers achieve (more) dependable programs (the normal goal of formal methods and specifications), as well as its use in courts of law, where litigants want to achieve more dependable evidence.

Participants

- Thorsten Berger
Ruhr-Universität Bochum, DE
- José Creissac Campos
University of Minho, PT
- Mauricio Castillo-Effen
Lockheed Systems – Arlington, US
- Marsha Chechik
University of Toronto, CA
- Benoît Combemale
INRIA – Rennes, FR
- Alcino Cunha
University of Minho, PT
- Jyotirmoy Deshmukh
USC – Los Angeles, US
- Matthew Dwyer
University of Virginia – Charlottesville, US
- Lars Grunske
HU Berlin, DE
- Reiner Hähnle
TU Darmstadt, DE
- Taylor T. Johnson
Vanderbilt University – Nashville, US
- Eunsuk Kang
Carnegie Mellon University – Pittsburgh, US
- Ekaterina Komendantskaya
Heriot-Watt University – Edinburgh, GB
- Yi Li
Nanyang TU – Singapore, SG
- Shahar Maoz
Tel Aviv University, IL
- Rômulo Meira-Góes
Pennsylvania State University – University Park, US
- Alexandra Mendes
University of Porto, PT
- Federico Mora
University of Waterloo, CA
- Daniel Neider
TU Dortmund, DE
- Phillippe Palanque
Toulouse University, FR
- Jan Oliver Ringert
Bauhaus-Universität Weimar, DE
- Bernhard Rumpe
RWTH Aachen, DE
- Kathryn T. Stolee
North Carolina State University – Raleigh, US
- Allison Sullivan
University of Texas at Arlington, US
- Harold Thimbleby
Swansea University, GB
- Michael W. Whalen
Amazon Inc. – Minneapolis, USA & The University of Minnesota – Minneapolis, USA
- Pamela Zave
Princeton University, US

Report from Dagstuhl Seminar 25401

Societal Impact of Computational Social Choice

Martin Lackner[*1], Nicholas Mattei[*2], Arianna Novaro[*3], Clemens Puppe[*4], and Ratip Emin Berker[†5]

1 **University of Applied Sciences St. Pölten, AT.** martin.lackner@ustp.at
2 **Tulane University – New Orleans, US.** nsmattei@tulane.edu
3 **Université Paris 1 Panthéon-Sorbonne, FR.** arianna.novaro@univ-paris1.fr
4 **KIT – Karlsruher Institut für Technologie, DE.** clemens.puppe@kit.edu
5 **Carnegie Mellon University – Pittsburgh, US.** rberker@cs.cmu.edu

Abstract

Computational Social Choice (COMSOC) is an interdisciplinary field between social choice theory in economics and theoretical computer science. The focus is to study algorithms for collective decision-making problems, such as political elections, the allocation of resources, and so on. In this Dagstuhl Seminar "Societal Impact of Computational Social Choice" (25401), we focused on three main topics. The first one was data, which has become an essential element for COMSOC research. In fact, thanks to the availability of open libraries, datasets and tools, researchers can now implement and test their algorithms for collective decision-making on real-life data, complementing their theoretical results. The second one was participation, as in recent years many municipalities and public institutions have moved towards various forms of participatory and digital democracy, with the goal of increasing the citizens' engagement in the public life of their communities. The third one was time, as although many collective decision-making problems have an underlying repeated nature, this dimension had thus far not received the deserved attention within standard COMSOC models. We addressed these topics under the two overarching themes of domain restrictions and societal impact: while domain restrictions can be seen as a methodological question over the input of our problems, societal impact can be seen as part of their output, i.e., the applications originating from theoretical research.

Seminar September 28 – October 2, 2025 – https://www.dagstuhl.de/25401
2012 ACM Subject Classification Theory of computation → Algorithmic game theory and mechanism design; Theory of computation → Design and analysis of algorithms; Applied computing → Law, social and behavioral sciences
Keywords and phrases computational social choice, data, participation, time
Digital Object Identifier 10.4230/DagRep.15.9.183

1 Executive Summary

Arianna Novaro (Université Paris 1 Panthéon-Sorbonne, FR)
Martin Lackner (University of Applied Sciences St. Pölten, AT)
Nicholas Mattei (Tulane University – New Orleans, US)
Clemens Puppe (KIT – Karlsruher Institut für Technologie, DE)

The Dagstuhl Seminar 25401 on "Societal Impact of Computational Social Choice" gathered 41 participants, from 11 countries and 4 continents. Most of the participants' primary area of research was Computational Social Choice (COMSOC), an interdisciplinary field which brings together computer scientists, economists, mathematicians, philosophers and

[*] Editor / Organizer
[†] Editorial Assistant / Collector

Societal Impact of Computational Social Choice, *Dagstuhl Reports*, Vol. 15, Issue 9, pp. 183–200
Editors: Martin Lackner, Nicholas Mattei, Arianna Novaro, Clemens Puppe, and Ratip Emin Berker
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

political scientists. The field is primarily concerned with designing algorithms and systems for collective decision-making, and analysing these algorithms and systems from an axiomatic, strategic, and computational perspective. Moreover, one of the seminar participants, Mathijs Kemp, brought a perspective from outside of academia, as he is the founder of the Parta platform for collective decision-making.

The seminar program focused on the three main topics of *data*, *participation*, and *time* in computational social choice, under the two overarching themes of *domain restrictions* and *societal impact*:

- **Data** has become an essential element for COMSOC research. In fact, by using open repositories and datasets (such as the PrefLib and PabuLib libraries), synthetic datasets, as well as tools (such as PrefPy, ABC Voting Rules and many others), COMSOC researchers can now implement and test their algorithms for collective decision-making. On the one hand, the results of these tests can inform recommendations on which methods actually perform better on desirable objectives in practice. On the other hand, they may highlight the presence of some underlying structure, which could lead to a more refined theoretical analysis.
- **Participation** has been formalized and studied in classical social choice theory by means of various mathematical axioms capturing the intuitive idea that a voting mechanism should incentivize voters to actually express their opinions. More recently, many municipalities and public institutions have moved towards various forms of participatory and digital democracy, with the goal of increasing the citizens' active role and engagement in the public life of their communities. Perhaps the main such example is that of participatory budgeting, where the citizens can propose and then vote on how to spend a percentage of a public budget – a practice that has been implemented in many cities around the world, and that has attracted the interest of COMSOC researchers. Moreover, in order to effectively elicit citizen participation, we also have to consider the (digital) platforms over which the collective decisions take place.
- **Time** is a dimension that has received increasing attention when studying problems of collective decision-making. Some examples of time-aware formalisms in COMSOC include: iterative voting, i.e., the interplay of strategic voting with time, as agents can modify their votes after observing the current results; perpetual voting, i.e., the analysis of fairness in elections over time; dynamic social choice, i.e., the change in agents' preferences, available resources, or number of participants over time. Indeed, numerous real-world problems include time as a component or have a repeated nature, e.g., the allocation of weekly chores among roommates.

Each of these three topics, as well as the two overarching themes, had a dedicated talk session in the seminar program (with the exception of the theme of *societal impact* which had two sessions), for a total of 18 talks. The talk sessions were structured as three short research talks followed by a panel discussion between the session speakers and the audience.

The seminar program also included 2 invited talks to report on the process and the interaction between policy-makers and academic advisors. The first one was by Sam Hirsh, who was involved in a legal process to reconfigure the electoral voting maps in the United States, and the second one was by Friedrich Pukelsheim, who was a member of the expert committee advising the German Bundestag in the process of the electoral reform of the German Bundeswahlgesetz.

Finally, the program included an experiment session by Théo Delemazure and Jérôme Lang, as well as numerous working group sessions and some time for informal discussions.

We wish to thank all of the participants, the invited speakers, our report collector Emin Berker, as well as the great Dagstuhl staff, for their valuable contribution to the success of the seminar.

2 Table of Contents

Executive Summary
Arianna Novaro, Martin Lackner, Nicholas Mattei, and Clemens Puppe 183

Overview of Talks

Achieving Rawlsian Justice in Food Rescue
Gerdus Benade 187

Approval-Based Committee Voting in Practice:
A Case Study of (Over-)Representation in the Polkadot Blockchain
Niclas Boehmer 187

When Fairness Does Not Exist: Detecting and Responding to Unfairness in Indivisible Allocations
Robert Bredereck 188

An Experiment on the Impact of the Number of Candidates in Approval Voting
Théo Delemazure and Jérôme Lang 188

Twenty Years of Voting Experiments during French Presidential Elections
Théo Delemazure 189

Diversity of Structured Domains
Piotr Faliszewski 189

Formal Explanations for Collective Decisions
Umberto Grandi 190

Deploying Fair Sampling Algorithms for Sortition
Paul Gölz 190

Navigating the American Redistricting Maze: Mathematical Challenges and Political Realities in US Electoral Map Design
Sam Hirsch 191

Beyond One Person One Vote
Mathijs Kemp 191

Overton Pluralism as Inference-Time Social Choice
Sonja Kraiczy 192

Social Choice Engineering: A Manifesto
Jérôme Lang 192

Participation Incentives in Approval-Based Committee Elections
Patrick Lederer 193

Online Algorithms for Participatory Budgeting
Jan Maly 193

Repeated Fair Allocation of Indivisible Items
Oliviero Nardi 194

Reform of the Electoral System for the German Bundestag
Friedrich Pukelsheim 194

City Sampling for Citizens' Assemblies
Ulrike Schmidt-Kraepelin 195

Condorcet Domains
Arkadii Slinko 195

What can we learn from real-world PB data?
Stanisław Szufa 196

Cost Utilities
Toby Walsh 196

Strengthening Proportionality in Temporal Voting
Tomasz Wąs 196

Working groups

Computational Social Choice: Research & Development
Niclas Boehmer, Dorothea Baumeister, Ratip Emin Berker, Sylvain Bouveret, Andreas Darmann, Piotr Faliszewski, Martin Lackner, Jérôme Lang, Nicholas Mattei, and Arianna Novaro 197

Social Choice with Text
Umberto Grandi, Robert Bredereck, Théo Delemazure, Ulle Endriss, Jan Maly, Nicholas Mattei, Nicolas Maudet, Oliviero Nardi, and Stanisław Szufa 197

Querying in Social Choice
Davide Grossi, Gerdus Benade, Ratip Emin Berker, Edith Elkind, Paul Gölz, Sonja Kraiczy, Patrick Lederer, Jannik Peters, Ulrike Schmidt-Kraepelin, and Tomasz Wąs 198

How can you ensure, that stakeholders with a higher stake, have more say in a matter? Or is that a bad practice?
Mathijs Kemp, Martin Bullinger, Reshef Meir, Marcus Pivato, and Frederik Van De Putte 198

Participants 200

3 Overview of Talks

3.1 Achieving Rawlsian Justice in Food Rescue

Gerdus Benade (Boston University, US)

Joint work of Aydin Alptekinoglu, Gerdus Benade
Main reference Aydin Alptekinoglu, Gerdus Benade: "Achieving Rawlsian Justice in Food Rescue", SSRN, October 17, 2024.
URL https://doi.org//10.2139/ssrn.4992842

We study a problem faced by a national food rescue platform that matches each donation to the first recipient who claims it. Recipients have very different response rates, leading to a few highly responsive recipients claiming the bulk of the donations. We ask whether priority lists, which control when the donation is announced to each recipient, are a remedy for inequitable outcomes. We give efficient algorithms to find the n-stage and binary priority lists that optimize a class of Rawlsian objective functions focusing on the worst-off recipients. The simple idea is to give higher priority to recipients who have received less in the past and to those who were slower in responding to notifications. This can be codified into an index by which to rank order eligible recipients. Computational experiments calibrated by historical data confirm that even binary priority lists lead to significantly more fair allocations than the existing first-come-first-serve allocation system.

3.2 Approval-Based Committee Voting in Practice: A Case Study of (Over-)Representation in the Polkadot Blockchain

Niclas Boehmer (Hasso-Plattner-Institut, Universität Potsdam, DE)

Joint work of Niclas Boehmer, Markus Brill, Alfonso Cevallos, Jonas Gehrlein, Luis Sánchez-Fernández, Ulrike Schmidt-Kraepelin
Main reference Niclas Boehmer, Markus Brill, Alfonso Cevallos, Jonas Gehrlein, Luis Sánchez Fernández, Ulrike Schmidt-Kraepelin: "Approval-Based Committee Voting in Practice: A Case Study of (over-)Representation in the Polkadot Blockchain", in Proc. of the Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada, pp. 9519–9527, AAAI Press, 2024.
URL https://doi.org//10.1609/AAAI.V38I9.28807

We provide the first large-scale data collection of real-world approval-based committee elections. These elections have been conducted on the Polkadot blockchain as part of their Nominated Proof-of-Stake mechanism and contain around one thousand candidates and tens of thousands of (weighted) voters each. We conduct an in-depth study of application-relevant questions, including a quantitative and qualitative analysis of the outcomes returned by different voting rules. Besides considering proportionality measures that are standard in the multiwinner voting literature, we pay particular attention to less-studied measures of overrepresentation, as these are closely related to the security of the Polkadot network.

3.3 When Fairness Does Not Exist: Detecting and Responding to Unfairness in Indivisible Allocations

Robert Bredereck (TU Clausthal, DE)

Joint work of Matthias Bentert, Niclas Boehmer, Eva Deltl, Klaus Heeger, Pallavi Jain, Andrzej Kaczmarczyk, Leon Kellerhals, Dusan Knop, Junjie Luo, Rolf Niedermeier, Florian Sachse, Bin Sun

In many allocation problems with indivisible goods, fairness notions widely accepted by society, such as envy-freeness (EF), may fail to exist – famously, when two agents compete for a single item. While relaxations like EF1/EFX guarantee existence and have driven much recent progress, they risk masking genuine impossibility by labeling inherently unfair situations as "fair." In my talk, I advocate a complementary agenda centered on the decision problem: does a fair solution exist at all?

For situations where fair outcomes do not exist in classical allocation settings, researchers have developed several workarounds: shared use, partial allocation, donations of goods, a fine-grained view of possible envy relations via social networks, or minimal subsidies. My talk advocates algorithmically assessing the feasibility of such interventions and identifying inherently unfair situations instead of hiding them.

3.4 An Experiment on the Impact of the Number of Candidates in Approval Voting

Théo Delemazure (University of Amsterdam, NL) and Jérôme Lang (CNRS – Paris, FR)

Joint work of Théo Delemazure, Jérôme Lang, Roberto Brunetti, Antoinette Baujard

We ran a short pilot experiment on Approval Voting during this Dagstuhl Seminar. To investigate participant preferences, participants had to choose among a set of desserts to be served during cake time. The participants were split into two groups: half of them first had to choose among six candidates (then twelve candidates), while the other half first chose among twelve (then six candidates). Our goal was to check what would be the impact of the number of candidates in the election on the average number of approved candidates per voter. Does the average number of approvals double if we double the number of candidates, remain constant, or follow a linear or sublinear relationship? As we expected, we observed in this experiment that voters tend to lower their "approval threshold" when there are fewer alternatives in the election. We aim to repeat this experiment with more participants and by applying the different changes that were proposed during the discussion with the seminar's participants.

3.5 Twenty Years of Voting Experiments during French Presidential Elections

Théo Delemazure (University of Amsterdam, NL)

URL https://theo.delemazure.fr/datasets/

Since 2002, 22 voting experiments have been conducted in parallel to French presidential elections, testing alternative voting methods such as Approval voting, Borda, Instant runoff voting, Evaluative voting, or the Majority Judgement. Most of these experiments took place on specific French cities on the day of the election, in official voting stations: after voting in the actual election, voters could take part in the experiment. Since 2012, some voting experiments have also been conducted online, enabling them to reach a high number of participants. These datasets are of significant interest for social choice experiments. Indeed, they cover many different ballot formats, and in some cases, voters provided their preferences with different ballot formats, allowing to compare voting rules based on different formats. Moreover, the political context often facilitates the interpretation of experiment results. Finally, because the context is the same over the years, one can study the evolution of the preferences over time. This motivated us to clean these datasets and make them freely accessible online (including those not previously available), and we aim to compile them into a dedicated library (similar to Pabulib), alongside comparable experiments conducted in other countries.

3.6 Diversity of Structured Domains

Piotr Faliszewski (AGH University of Science & Technology – Krakow, PL)

Joint work of Piotr Faliszewski, Krzysztof Sornat, Stanisław Szufa, Tomasz Wąs
Main reference Piotr Faliszewski, Krzysztof Sornat, Stanisław Szufa, Tomasz Wąs: "Diversity of Structured Domains via k-Kemeny Scores", CoRR, Vol. abs/2509.15812, 2025.
URL https://doi.org//10.48550/ARXIV.2509.15812

We consider ordinal elections, where each voter ranks the available candidates from the most to the least appreciated. In principle, each voter can cast an arbitrary preference order, but it is often convenient to consider structured domains of such preferences. For example, in the single-peaked domain there is a societal axis of the candidates (e.g., politicians ranked from the most left-wing to the most right-wing one) and for each preference ranking, each prefix is a contiguous interval of the axis. Such preference domains capture various rationality criteria that we expect the voters to follow.

In this talk, we discuss several ways in which one could measure diversity of structured preference domains. First, we mention the richness-based approach, studied, e.g., by Ammann and Puppe (Preference diversity. Review of Economic Design, 2025) or by Karpov et al. (Local diversity of Condorcet domains, arXiv 2024) that relies on counting various substructures in the votes. Then we introduce our concepts of inner and outer diversity. Under the former, we say that a domain is diverse if it is difficult to cluster, and under the latter we say that a domain is diverse if every possible vote is similar to some vote from the domain. We instantiate inner diversity using the k-Kemeny problem (where given an election we want to partition it – or, cluster – into a given number of subelections whose sum of Kemeny

scores is lowest). To define outer diversity, we measure the expected swap distance between a random vote and the closest vote in the domain. For a number of standard structured domains (including single-peaked, single-crossing, group-separable, and Euclidean ones) we evaluate their diversity using our approaches and find that inner and outer diversity give similar results. Further, we provide a number of algorithmic and complexity-theoretic results related to computing diversity of these domains.

3.7 Formal Explanations for Collective Decisions

Umberto Grandi (University Toulouse Capitole, FR)

Joint work of Umberto Grandi, Clément Contet, Jérôme Mengin
Main reference Clément Contet, Umberto Grandi, Jérôme Mengin: "Explaining Tournament Solutions with Minimal Supports", CoRR, Vol. abs/2509.09312, 2025.
URL https://doi.org//10.48550/ARXIV.2509.09312

Election results or the outcomes of participatory budgeting campaigns are typically presented to voters as a ranked list of alternatives based on scores. However, from a user or voter perspective, we believe this method fails to adequately explain why a particular candidate is elected or a project is approved. In this presentation, I will discuss our ongoing work on applying techniques used to explain black-box machine learning algorithms to transparent voting rules. Our explanations identify the smallest subsets of collected preference data that either support the winning candidate or, if altered, could change the outcome – these are known as counterfactual explanations. I will introduce algorithms for computing these formal explanations, along with refinements and bounds on their size, particularly for the case of tournament solutions.

3.8 Deploying Fair Sampling Algorithms for Sortition

Paul Gölz (Cornell University – Ithaca, US)

Joint work of Paul Gölz, Gili Rusak, Bailey Flanigan, Anupam Gupta, Brett Hennig, Ariel D. Procaccia

Citizens' assemblies are an emerging form of democratic participation, in which a panel of randomly selected constituents weigh in on a policy question. In this talk, I will speak about the past, present, and future of Panelot.org, a not-for-profit website on which practitioners can run sampling algorithms based on fair division. First, I will summarize our collaboration with the Sortition Foundation came to be and how this collaboration impacted the design of the sampling algorithm. Second, I will talk about our algorithm deployment and the extent to which we can (and cannot) evaluate its societal impact. Finally, I will present ongoing work on overhauling Panelot, as well as several new algorithms derived from recent social choice research, which we plan to add as functionalities.

3.9 Navigating the American Redistricting Maze: Mathematical Challenges and Political Realities in US Electoral Map Design

Sam Hirsch

Unlike Europe's multi-member district systems, the United States relies on single-member geographic districts, creating a complex optimization problem where boundaries must satisfy numerous competing objectives: equal population, compactness, community preservation, Voting Rights Act compliance, and state-specific criteria ranging from competitiveness to incumbent protection. This patchwork of requirements across 50 states defies uniform mathematical treatment.

Building on recent computational social choice work in redistricting, this conversation identifies where algorithmic advances could impact reform efforts. Key challenges include defining "fairness" when stakeholders fundamentally disagree on objectives, exploring the vast space of valid maps efficiently, and developing interpretable metrics that withstand legal scrutiny.

This conversation will examine why consensus remains elusive – from tensions between mathematical elegance and political feasibility to communicating technical concepts to non-technical audiences. Drawing from practical experience in US election reform, Sam will discuss what tools and measurements would most benefit practitioners and explore whether "optimal" districting is even well-defined in a pluralistic democracy. Attendees will gain insights into translating theoretical advances into real-world impact in one of America's most contentious political processes.

3.10 Beyond One Person One Vote

Mathijs Kemp (Vennster – Almere, NL)

Parta is a decision making method and tool that is used for different purposes: In participation councils, area development & energy transition. Parta uses the one person one vote principle to calculate the results. But is this always the most logical approach? In one specific use case, parents voted on classroom composition by age and academic level in an elementary school. Currently, in Parta, there is no information about the voter, nor can you weigh votes differently. We did an external analysis of the votes, based on an external tool. There was a difference between the preference of parents of younger children and the parents of children that were about to leave the school. The decision had more impact on the younger children. The same occurs in area development: Citizens in the area are impacted more than citizens that are casual visitors or live further away. We are considering giving stakeholders with a higher stake more votes. How should we present the outcome based on the different weights and what are the pitfalls of this approach?

3.11 Overton Pluralism as Inference-Time Social Choice

Sonja Kraiczy (University of Oxford, GB)

Joint work of Sonja Kraiczy, Brandon Amos, Ratip Emin Berker, Avinandan Bose, Edith Elkind, Smitha Milli, Maximilian Nickel, Ariel Procaccia, Jamelle Watson-Daniels

People increasingly turn to LLMs for information, including answering socially contested prompts, so responses should reflect the different viewpoints of sufficiently large groups. We formalize this goal by adapting proportionally representative fairness (PRF) from social choice in metric spaces and introduce inference-time PRF: for any in-scope prompt, large cohesive groups are proportionally represented, and groups that are more cohesive are represented more closely, in the final answer. We present the first system with provable inference-time guarantees over the full response space. We train a personalized reward model that learns prompt-specific user-preference embeddings and a personalized LLM optimized for it. At inference time, a fast social-choice algorithm (a stream-lined version of the Spatial Expanding Approval Rule) selects k representative embeddings; the personalized LLM generates k conditioned responses that are merged into a pluralistic output. Assuming the model optimizes (resp. approximately optimizes) the reward, PRF over the embedding space implies PRF (resp. approximate PRF) over the response space, automatically adapting represented groups to each prompt. This is the first formalization and implementation with guarantees for viewpoint pluralism in LLMs.

3.12 Social Choice Engineering: A Manifesto

Jérôme Lang (CNRS – Paris, FR)

I would like to argue that one of the goals the computational social choice research community should pursue today is to use our knowledge and methods to solve specific problems coming from the real world. What I call "social choice engineering" is the design of tools for solving real-world collective decision-making problems. We can distinguish two types: (1) Project-based social choice engineering: we start from the specification of an actual need, followed by a theoretical analysis of what can or cannot be done, and the implementation of a software with a suitable interface to be tested and then delivered to the client(s). (2) Product-based social choice engineering: one thinks hard about which product to develop; then a software is designed, implemented, and tested on benchmarks or pilot studies, and the last step consists of finding customers willing to use it, and possibly helping them configure it. Some works of each type exist (especially in matching), and I give several examples; still, these are too few. I try to explain why this is the case, and I argue that, in my opinion, computational social choice is threatened by the (relative) lack of engineering work.

3.13 Participation Incentives in Approval-Based Committee Elections

Patrick Lederer (UNSW – Sydney, AU)

Joint work of Martin Bullinger, Chris Dong, Patrick Lederer, Clara Mehler
Main reference Martin Bullinger, Chris Dong, Patrick Lederer, Clara Mehler: "Participation Incentives in Approval-Based Committee Elections", in Proc. of the Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada, pp. 9546–9554, AAAI Press, 2024.
URL https://doi.org//10.1609/AAAI.V38I9.28810

In approval-based committee (ABC) voting, the goal is to choose a subset of predefined size of the candidates based on the voters' approval preferences over the candidates. While this problem has attracted significant attention in recent years, the incentives for voters to participate in an election for a given ABC voting rule have been neglected so far. This paper is thus the first to explicitly study this property, typically called participation, for ABC voting rules. In particular, we show that all ABC scoring rules even satisfy group participation, whereas most sequential rules severely fail participation. We furthermore explore several escape routes to the impossibility for sequential ABC voting rules: we prove for many sequential rules that (i) they satisfy participation on laminar profiles, (ii) voters who approve none of the elected candidates cannot benefit by abstaining, and (iii) it is NP-hard for a voter to decide whether she benefits from abstaining.

3.14 Online Algorithms for Participatory Budgeting

Jan Maly (Wirtschaftsuniversität Wien, AT)

Joint work of Jan Maly, Matthieu Hervouin

Participatory Budgeting (PB), and in particular, proportionality in PB, has received significant attention from the Computational Social Choice community in recent years. This led to the discovery of voting rules like the Method of Equal Shares, that selects outcomes in a fair and representative fashion. However, these rules assume that the set of possible projects is given in advance. In this talk, I describe a new framework, that we call online PB, in which projects are revealed one by one and a binding funding decision has to be made on the spot. We consider several classical fairness axioms from the offline PB literature in this online setting, namely priceability and the most prominent axioms of justified representation, JR, PJR and EJR. We see that priceability is always satisfiable in the online setting and find tight approximations for the justified representation axioms. Additionally, we discuss experiments showing that, in practice, a simple greedy online PB rule produces outcomes that are nearly as fair as the outcomes produced by the Method of Equal Shares (without completion) in offline PB.

3.15 Repeated Fair Allocation of Indivisible Items

Oliviero Nardi (TU Wien, AT)

Joint work of Ayumi Igarashi, Martin Lackner, Oliviero Nardi, Arianna Novaro
Main reference Ayumi Igarashi, Martin Lackner, Oliviero Nardi, Arianna Novaro: "Repeated Fair Allocation of Indivisible Items", in Proc. of the Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada, pp. 9781–9789, AAAI Press, 2024.
URL https://doi.org//10.1609/AAAI.V38I9.28837

The problem of fairly allocating a set of indivisible items is a well-known challenge in the field of (computational) social choice. This classic problem typically assumes a single allocation, but in practice, items often need to be distributed repeatedly. For example, the items may be recurring chores to distribute in a household. Motivated by these observations, we initiate the study of the repeated fair division of indivisible goods and chores. We show that, if the number of repetitions is a multiple of the number of agents, there always exists a sequence of allocations that is proportional and Pareto-optimal. On the other hand, irrespective of the number of repetitions, an envy-free and Pareto-optimal sequence of allocations may not exist. For the case of two agents, we show that if the number of repetitions is even, it is always possible to find a sequence of allocations that is overall envy-free and Pareto-optimal. We then prove even stronger fairness guarantees, showing that every allocation in such a sequence satisfies some relaxation of envy-freeness.

3.16 Reform of the Electoral System for the German Bundestag

Friedrich Pukelsheim (Universität Augsburg, DE)

URL https://www.math.uni-augsburg.de/htdocs/emeriti/pukelsheim/2022Berlin/

The German Federal Election Law was amended in 2024. The talk illustrates how the amended law works using the election to the 21st Deutscher Bundestag in February 2025. The presentation details the procedure how votes are converted into seats, outlines the constitutional objectives as laid down in the German Basic Law, points to the fundamental goal of verifying the principle of One Person, One Vote, and provides an overview of the work of the Bundestag Reform Commission which negotiated in 2022 and 2023 the new amendment and in which the speaker participated as an expert member.

3.17 City Sampling for Citizens' Assemblies

Ulrike Schmidt-Kraepelin (TU Eindhoven, NL)

Joint work of Paul Gölz, Jan Maly, Ulrike Schmidt-Kraepelin, Markus Utke, Philipp C. Verpoort
Main reference Paul Gölz, Jan Maly, Ulrike Schmidt-Kraepelin, Markus Utke, Philipp C. Verpoort: "City Sampling for Citizens' Assemblies", CoRR, Vol. abs/2509.07557, 2025.
URL https://doi.org//10.48550/ARXIV.2509.07557

In citizens' assemblies, a group of constituents is randomly selected to weigh in on policy issues. We study a two-stage sampling problem faced by practitioners in countries such as Germany, in which constituents' contact information is stored at a municipal level. As a result, practitioners can only select constituents from a bounded number of cities ex post, while ensuring equal selection probability for constituents ex ante.

We develop several algorithms for this problem. Although minimizing the number of contacted cities is NP-hard, we provide a pseudo-polynomial time algorithm and an additive 1-approximation, both based on separation oracles for a linear programming formulation. Recognizing that practical objectives go beyond minimizing city count, we further introduce a simple and more interpretable greedy algorithm, which additionally satisfies an ex-post monotonicity property and achieves an additive 2-approximation. Finally, we explore a notion of ex-post proportionality, for which we propose two practical algorithms: an optimal algorithm based on column generation and integer linear programming and a simple heuristic creating particularly transparent distributions. We evaluate these algorithms on data from Germany, and plan to deploy them in cooperation with a leading nonprofit organization in this space.

3.18 Condorcet Domains

Arkadii Slinko (University of Auckland, NZ)

Joint work of Clemens Puppe, Arkadii Slinko
Main reference Clemens Puppe, Arkadii Slinko: "Condorcet Domains: The Mathematics of Coherent Collective Decision-Making," Springer Nature: Studies in Social Choice and Welfare, to appear in 2026

My talk was a presentation of the book titled "Condorcet Domains: The Mathematics of Coherent Collective Decision-Making" written by myself jointly with Prof. Clemens Puppe (KIT). I outlined the structure of the book and gave examples of results from several chapters. As Herve Moulin noted in his foreword "The theory of Condorcet domains straddles the social sciences and discrete mathematics. This interdisciplinary volume is already a canonical reference in both communities."

3.19 What can we learn from real-world PB data?

Stanisław Szufa (CNRS – Paris, FR)

Main reference https://pabulib.org

We present the results of an analysis of more than 1400 real-world participatory budgeting instances. In particular, we examine how different types of ballots influence citizens' behavior, specifically in terms of the projects they propose and their voting patterns. Moreover, we provide recommendations on which types of ballots to use in practice

3.20 Cost Utilities

Toby Walsh (UNSW – Sydney, AU)

Cost utilities are those additive utilities in which every agent assigns the same utility for an item (aka its cost) that they approve, and zero utility for an item that they do not approve. They have been called generalised binary utilities and various other names. By restricting to cost utilities, the action space for agents is reduced – agents can only declare their utility for an item is equal to zero or the item cost. In fact, this is enough of a restriction to ensure that several fair division procedures become strategy proof. In this talk, I demonstrate that cost utilities are an interesting and practical domain restriction that ensures a range of good normative properties in addition to strategy proofness.

3.21 Strengthening Proportionality in Temporal Voting

Tomasz Wąs (University of Oxford, GB)

Joint work of Bradley Phillips, Edith Elkind, Nicholas Teh, Tomasz Wąs
Main reference Bradley Phillips, Edith Elkind, Nicholas Teh, Tomasz Wąs: "Strengthening Proportionality in Temporal Voting", CoRR, Vol. abs/2505.22513, 2025.
URL https://doi.org//10.48550/ARXIV.2505.22513

We study proportional representation in the framework of temporal voting with approval ballots. Prior work adapted basic proportional representation concepts – justified representation (JR), proportional JR (PJR), and extended JR (EJR) – from the multiwinner setting to the temporal setting. Our work introduces and examines ways of going beyond EJR. Specifically, we consider stronger variants of JR, PJR, and EJR, and introduce temporal adaptations of more demanding multiwinner axioms, such as EJR+, full JR (FJR), full proportional JR (FPJR), and the Core. For each of these concepts, we investigate its existence and study its relationship to existing notions, thereby establishing a rich hierarchy of proportionality concepts. Notably, we show that two of our proposed axioms – EJR+ and FJR – strengthen EJR while remaining satisfiable in every temporal election.

4 Working groups

4.1 Computational Social Choice: Research & Development

Niclas Boehmer (Hasso-Plattner-Institut, Universität Potsdam, DE), Dorothea Baumeister (HS Bund f. öffentl. Verwaltung – Brühl, DE), Ratip Emin Berker (Carnegie Mellon University – Pittsburgh, US), Sylvain Bouveret (University of Grenoble, FR), Andreas Darmann (Universität Graz, AT), Piotr Faliszewski (AGH University of Science & Technology – Krakow, PL), Martin Lackner (TU Wien, AT), Jérôme Lang (CNRS – Paris, FR), Nicholas Mattei (Tulane University – New Orleans, US), and Arianna Novaro (Université Paris 1 Panthéon-Sorbonne, FR)

Computational social choice (COMSOC) studies principled ways to aggregate conflicting individual preferences into collective decisions. Although inspired by concrete applications such as voting and participatory budgeting, much of the published COMSOC research has focused on abstract, foundational questions, with comparatively little emphasis on deploying these ideas in practice.

In this working group, we discussed the need for increased effort towards *Computational Social Choice: Research & Development (COMSOC-R&D)*, a problem-driven research agenda that explicitly aims to design, implement, and test collective decision-making systems in the real world. After the seminar, we collected the ideas raised in the working group in a position paper that captures this call for action. In the paper, we first articulate the defining features of COMSOC-R&D and argue that such work is a necessary next step for the community to achieve meaningful real-world impact. Subsequently, we identify key roadblocks to COMSOC-R&D and discuss potential remedies at the individual and community level. Finally, we propose desiderata and evaluation criteria for future COMSOC-R&D projects.

4.2 Social Choice with Text

Umberto Grandi (University Toulouse Capitole, FR), Robert Bredereck (TU Clausthal, DE), Théo Delemazure (University of Amsterdam, NL), Ulle Endriss (University of Amsterdam, NL), Jan Maly (Wirtschaftsuniversität Wien, AT), Nicholas Mattei (Tulane University – New Orleans, US), Nicolas Maudet (Sorbonne University – Paris, FR), Oliviero Nardi (TU Wien, AT), and Stanisław Szufa (CNRS – Paris, FR)

Our research typically starts from preferences elicited from voters in the form of rankings, pairwise comparisons, or approval ballots. Given that contemporary AI gives us powerful text-based interfaces, what are principled approaches to do social choice starting from text input? In this group we first tackled the problem of surveying datasets containing both texts and votes on texts (deliberative platforms such as Polis, parliamentary discussions, participatory budgeting, collective statement generation...). Then, participants gave talks and tutorials on their preliminary work on this topic, in particular merging natural language processing techniques or LLMs with classical social choice algorithms. Finally, we sketched

a "blue sky" paper on the topic of the working group, detailing in a systematic way the different aspects of a collective decision that can profit from the use of textual processing techniques such as LLMs.

4.3 Querying in Social Choice

Davide Grossi (University of Groningen, NL), Gerdus Benade (Boston University, US), Ratip Emin Berker (Carnegie Mellon University – Pittsburgh, US), Edith Elkind (Northwestern University – Evanston, US), Paul Gölz (Cornell University – Ithaca, US), Sonja Kraiczy (University of Oxford, GB), Patrick Lederer (UNSW – Sydney, AU), Jannik Peters (National University of Singapore, SG), Ulrike Schmidt-Kraepelin (TU Eindhoven, NL), and Tomasz Wąs (University of Oxford, GB)

In broad strokes, social choice problems can be typically described as follows: given complete information about individual attitudes (e.g., approval, preferences) over a set of alternatives, find an outcome (e.g., an alternative, or a set of alternatives) that satisfy a given societal objective (e.g., some variant of proportional representation). In the working group we focused on investigating generalizations of the above question where: information over attitudes is incomplete and possibly sparse; the set of alternatives is not fixed and grows over time; outcomes meeting the desired objective should be computed at each time step. Social choice problems of this sort underpin applications in digital democracy and crowd-sourcing and interface directly with topics in learning theory (e.g., exploration/exploitation tradeoffs in online decision-making). The group reviewed relevant literature and proceeded to sketch a first formalization of the problem.

4.4 How can you ensure, that stakeholders with a higher stake, have more say in a matter? Or is that a bad practice?

Mathijs Kemp (Vennster – Almere, NL), Martin Bullinger (University of Oxford, GB), Reshef Meir (Technion – Haifa, IL), Marcus Pivato (Université Paris 1 Panthéon-Sorbonne), and Frederik Van De Putte (Erasmus University – Rotterdam, NL)

In this working group, one of the practical use cases of Parta was studied. In this use case, a convincing majority (60-70%) was preferred to move ahead with a proposal that would move 3 grades into one classroom. At first, this majority was not achieved. The vote was split 56% in favour and 44% against. After analysing the data, it was concluded that the parents of the younger children voted more heavily in favour of the proposal. The principal concluded that this group has a higher stake and decided to move ahead with the proposal, even though the majority was not as convincing as preferred. In this working group, it was explored to see whether it's possible to take this discrepancy in stakes into account within the voting process itself. An important question was quickly raised: How do we elicit these stakes. It was first concluded that the stakes should be based on objective criteria.

An idea was then proposed to democratically determine these weights. Voters are asked to assign a weight to other voters (anonymously and maybe even hypothetical) based on objective characteristics. Those weights are then averaged, before they're used in the actual process. Another important aspect for acceptance of such a decision is the legitimacy of the final decision. To ensure this, it was concluded that the weighing process should be transparent and that the weighing should be determined in the constitutional phase. With this in mind, democratically deciding the weights will enhance legitimacy. When objective public observable criteria are not available. A number of voting mechanisms were explored to see if they would fit. Quadratic voting, the Pivotal Voting Mechanism and Storable votes were explored. A combination was proposed, where fictional money would be used over multiple decisions, to assure that the weights are elicited fairly and correctly. Using delegation to make stakes count has also been explored. Here, the weights are assigned implicitly, since voters with higher stakes will convince other voters to delegate their vote to them. This will mitigate the tyranny of the majority and therefore give voters with a higher stake more influence in the vote.

Participants

- Dorothea Baumeister
HS Bund f. öffentl. Verwaltung – Brühl, DE
- Gerdus Benade
Boston University, US
- Ratip Emin Berker
Carnegie Mellon University – Pittsburgh, US
- Niclas Boehmer
Hasso-Plattner-Institut, Universität Potsdam, DE
- Sylvain Bouveret
University of Grenoble, FR
- Florian Brandl
Universität Bonn, DE
- Felix Brandt
TU München – Garching, DE
- Robert Bredereck
TU Clausthal, DE
- Markus Brill
University of Warwick – Coventry, GB
- Martin Bullinger
University of Oxford, GB
- Andreas Darmann
Universität Graz, AT
- Théo Delemazure
University of Amsterdam, NL
- Edith Elkind
Northwestern University – Evanston, US
- Ulle Endriss
University of Amsterdam, NL
- Piotr Faliszewski
AGH University of Science & Technology – Krakow, PL
- Paul Gölz
Cornell University – Ithaca, US
- Umberto Grandi
University Toulouse Capitole, FR
- Davide Grossi
University of Groningen, NL
- Mathijs Kemp
Vennster – Almere, NL
- Sonja Kraiczy
University of Oxford, GB
- Martin Lackner
TU Wien, AT
- Jérôme Lang
CNRS – Paris, FR
- Patrick Lederer
UNSW – Sydney, AU
- Jan Maly
Wirtschaftsuniversität Wien, AT
- Nicholas Mattei
Tulane University – New Orleans, US
- Nicolas Maudet
Sorbonne University – Paris, FR
- Reshef Meir
Technion – Haifa, IL
- Oliviero Nardi
TU Wien, AT
- Arianna Novaro
Université Paris 1 Panthéon-Sorbonne, FR
- Dominik Peters
University Paris-Dauphine, FR
- Jannik Peters
National University of Singapore, SG
- Marcus Pivato
Université Paris 1 Panthéon-Sorbonne, FR
- Friedrich Pukelsheim
Universität Augsburg, DE
- Clemens Puppe
KIT – Karlsruher Institut für Technologie, DE
- Ulrike Schmidt-Kraepelin
TU Eindhoven, NL
- Piotr Skowron
University of Warsaw, PL
- Arkadii Slinko
University of Auckland, NZ
- Stanislaw Szufa
CNRS – Paris, FR
- Frederik Van De Putte
Erasmus University – Rotterdam, NL
- Toby Walsh
UNSW – Sydney, AU
- Tomasz Wąs
University of Oxford, GB