Report from Dagstuhl Seminar 25461

Security and Privacy of Large Language Models

Stephan Günnemann[*1], Pavel Laskov[*2], Emil Lupu[*3], Vera Rimmer[*4], Advije Rizvani[†5], and Qianying Liao[†6]

1 **TU München – Garching, DE.** s.guennemann@tum.de
2 **Universität Liechtenstein – Vaduz, LI.** pavel.laskov@uni.li
3 **Imperial College London, GB.** e.c.lupu@imperial.ac.uk
4 **KU Leuven, BE.** vera.rimmer@kuleuven.be
5 **Universität Liechtenstein – Vaduz, LI.** advije99@gmail.com
6 **KU Leuven, BE.** qianying.liao@kuleuven.be

Abstract

Large Language Models (LLMs) have rapidly evolved from experimental systems capable of generating simple text to powerful general-purpose tools that solve exam-level problems, assist human experts, summarize complex documents, and write code, leading to their widespread deployment in production environments to improve productivity. Their rapid adoption has outpaced scientific understanding of the associated security and privacy risks, creating a growing gap between real-world use and trust in LLM-based applications. Unlike traditional machine learning models, LLMs are inherently general-purpose, operate at unprecedented scale, are often proprietary, and are accessed through interactive dialog interfaces. This makes their behavior difficult to predict and introduces novel attack surfaces and emerging phenomena such as hallucinations. Existing methods and analyzes for conventional machine learning are insufficient to cope with these aspects. So, this Dagstuhl Seminar "Security and Privacy of Large Language Models" (25461) aimed to initiate a systematic scientific discussion on the security and privacy of language models by bringing together researchers from AI, security, privacy, and natural language processing to address three central questions: how safe and secure are LLMs in adversarial environments; to what extent users' privacy may be at risk when interacting with LLMs; and what other categories of societal impact may arise from the rapid advancement of LLM technologies. The seminar sought to systematize existing knowledge, identify high-risk applications and prevalent attack vectors, discuss defense strategies and deployment challenges, and outline directions for future research in LLM security and privacy.

Seminar November 9–14, 2025 – https://www.dagstuhl.de/25461
2012 ACM Subject Classification Security and privacy → Software and application security; Security and privacy → Human and societal aspects of security and privacy; Computing methodologies → Natural language processing
Keywords and phrases Large Language Models, Artifician Intelligence, Security and Privacy
Digital Object Identifier 10.4230/DagRep.15.11.87

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

Security and Privacy of Large Language Models, *Dagstuhl Reports*, Vol. 15, Issue 11, pp. 87–113
Editors: Stephan Günnemann, Pavel Laskov, Emil Lupu, and Vera Rimmer

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

1 Executive Summary

Stephan Günnemann (TU München – Garching, DE)
Pavel Laskov (Universität Liechtenstein – Vaduz, LI)
Emil Lupu (Imperial College London, GB)
Vera Rimmer (KU Leuven, BE)

Overview

Large Language Models (LLM) have rapidly evolved from simple toys capable of generating stories about unicorns to powerful tools that can solve exam-level problems, advise human experts, summarize texts, and write code, among other tasks. The remarkable advances in LLM capabilities over the past few years have led to their rapid adoption and deployment in production environments without allowing to the time necessary to understand, analyze, and mitigate the resulting security risks. As a result, the security and privacy aspects are not well understood in the scientific community, causing a concerning lack of trust in the widely deployed LLM applications. This Dagstuhl Seminar "Security and Privacy of Large Language Models" (25461) was an attempt to bring together leading researchers in AI, security, privacy, and NLP to systematize knowledge, assess systemic risks, identify mitigation challenges, and map future directions in this area.

LLMs are qualitatively different from more traditional machine learning models. They are intended to be "general purpose" and handle a broad variety of tasks. While an image classifier trained to distinguish cats from dogs cannot even distinguish cars from airplanes, a language model is expected to be useful on any question expressible natural language. This makes it much more challenging to predict the context in which they will be deployed, the tasks they will be given and how they will perform. There are also many other differences. For example, these models have hundreds of billions to trillions of parameters and are often private to companies that sell their services. This makes it more difficult to understnd how they work, a challenging question already even when models are open. Loss functions optimized by LLMs are extremely complex, with local minima being a tangible threat to their generalization capabilities – a threat that also results in the widely recognized "hallucination" phenomena. Last, but not least, the learning components of LLMs are not directly exposed to end users but accessed via a dialog interface (and its own learning mechanisms) that itself can lead to "intriguing properties".

This Dagstuhl Seminar aimed to bring together researchers from a diverse set of backgrounds to discuss research directions that could lead to the scientific foundation for privacy and security of LLMs.

Goal of the Seminar

The seminar focused on three main themes of discussion, consistently with the issues highlighted above:

- How secure are language models? When an adversary is present and interacts with these models, can they cause the model to behave incorrectly? Which mitigations are effective?
- How private are language models? Can an adversary interact with the model to violate the privacy of users? What are the possible privacy risks?
- What other categories of societal impact may arise from the rapid advancement and deployment of LLMs within information and institutional systems?

In addition to these themes, the seminar also provided an open environment for the discussion of related topics. Emerging research in this area is often exploratory and rapidly evolving. So, organisers deliberately encouraged participants to bring different perspectives from the academic and the industry environments.

Overall Organization and Schedule

The structure and schedule of the seminar combined the advantages of conventional conference formats with the specific traditions of Dagstuhl Seminars in the following schedule:

Schedule	Activities
Day 1	Seminar introduction, short self-introductions by participants, six keynote presentations
Day 2	Three keynote presentations, a brain-storming session on challenges and open problems, the building of working groups, one breakout session
Day 3	Two breakout sessions, one keynote presentation, social event
Day 4	Reporting from breakout sessions, a podium discussion of societal impact of LLM security
Day 5	A discussion of future research on LLM security, summary of results

In what follows, the report presents an overview of the contributed talks, a summary of the main results that emerged from the seminar discussions, and detailed reports from the formed working groups.

2 Table of Contents

3 Overview of Talks

3.1 Is AI Security Doomed?

Ilia Shumailov (AI Sequrity Company – London, GB)

This talk examines emerging security risks in modern large language model (LLM) systems beyond modellevel performance, with a focus on system opacity and prompt injection attacks. We first move past the abstraction of simple inference interfaces (e.g., model.generate()) to analyze the complex, layered machine learning pipelines underlying contemporary LLM deployments, highlighting how pre-trained models, third-party components, and opaque dependencies introduce untrusted origins and systemic vulnerabilities. We then turn to prompt injection attacks, which have proven both pervasive and practically exploitable, and critically assess the current landscape of proposed mitigations. Despite significant research attention, effective and deployable defenses remain limited. The talk concludes by outlining promising directions for system-level defenses that aim to mitigate prompt injection risks in realistic LLM-based applications.

3.2 Inducing bias and measuring consistency in LLM responses

Lujo Bauer (Carnegie Mellon University – Pittsburgh, US)

This talk explores a subtle but powerful threat to user autonomy arising from prompt optimization services for large language models (LLMs). While such services aim to reduce the burden of prompt engineering, they also introduce a new and largely overlooked attack surface: untrusted prompt providers can deliberately manipulate prompts to bias model outputs. We demonstrate that minor synonym-level modifications to prompts can significantly influence LLM behavior, increasing the likelihood that models mention targeted concepts – such as brands, political entities, or nations – by up to 78%. Through a user study, we show that these adversarially perturbed prompts are indistinguishable from benign prompts to human users, successfully steer model responses, and increase user attention to the targeted concepts without raising suspicion. The talk highlights the practical feasibility and societal implications of prompt-based manipulation and discusses mitigation strategies, including the need for warnings and safeguards when relying on prompts from untrusted sources.

3.3 Phishing in the LLM Era: Challenges and Opportunities

Giovanni Apruzzese (Reykjavik University, IS)

Every day, our inboxes are flooded with unsolicited emails, ranging between annoying spam to more subtle phishing scams. Unfortunately, despite abundant prior research efforts proposing solutions that achieve near-perfect accuracy and despite numerous organizations employing phishing education campaigns, the reality is that countering malicious emails still remains an unsolved dilemma – which has been exacerbated by the advent of LLM.

This talk will cover two recent publications (AsiaCCS'25, and AISec'25) which present novel results featuring the application of LLMs in the phishing-email context. First, I will show the effectiveness that LLM-generated emails, crafted with naive OSINT techniques, can have against 18k employees of three different organizations. Then, I will focus on the research side, and elucidate various "open problems" that affect academic literature on phishing-email detection – such as the excessive reliance on datasets containing emails exchanged up to 20 years ago. Finally, I will present how LLMs can be used defensively, i.e., to generate large-scale datasets of phishing emails which could resemble those sent by real attackers, and which could be used to benchmark existing (and future) detection methods – including, of course, those based on LLMs!

3.4 LLM Security & Privacy: A Perspective

Giovanni Cherubin (Microsoft Research – Cambridge, GB)

When developing security and privacy for "traditional" supervised Machine Learning (ML), the research community has matured a lot of wisdom. Arguably, this is in part attributable to the deep theoretical understanding of ML itself, which was developed by scientists such as T. Covert, P. Hart, C. J. Stone, A. Chervonenkis, V.Vapnik, and L. Valiant in the second half of 1900. Thanks to this, security & privacy for traditional ML flourished, with successes (e.g., privacy-preserving ML), as well as with the uncovering of negative results (e.g., impossibility of preventing adversarial examples in some domains). Conversely, developing a similarly rich understanding of auto-regressive models, such as Large Language Models (LLMs) seems out of reach for the foreseeable future. This makes securing this new learning pattern an extremely complex endeavor.

In this invited talk, we focus on the main security threat to LLMs, prompt injection attacks. We look at the defence paradigms to date, and show how each of them has inherent limitations; in conjunction, we highlight gaps and promising avenues in this field. Finally, we look at how traditional ML and LLMs behave differently also when it comes to privacy; this suggests that applying to LLMs our largely mature understanding of ML privacy is far from immediate, and can lead to mistakes. We hope that this analysis can help the research community prioritize their focus over the coming decade.

3.5 Capture the Narrative: Social media manipulation wargaming for cyberliteracy and research

Hammond Pierce (UNSW – Sydney, AU)

This talk presents Capture the Narrative, an educational competition designed to expose and study the mechanics of AI-driven information manipulation in online social platforms. The primary intention of the competition is to raise awareness of how coordinated automated agents can shape public discourse, while providing participants with hands-on experience in adversarial thinking, social media dynamics, and cyber-security challenges. Participants formed teams to design and deploy bots that operated on a simulated social media platform during a fictional election campaign, with the goal of amplifying or suppressing narratives to influence public opinion and polling outcomes. The competition unfolded over several weeks, during which millions of posts were generated and interacted with under controlled conditions. The outcome demonstrated that even simple, well-coordinated automated strategies can dominate online narratives and meaningfully affect collective perceptions, ultimately altering the simulated election result. The talk reflects on the lessons learned from the competition, including the effectiveness of narrative manipulation, the vulnerabilities of social platforms, and the value of experiential approaches for understanding AI-enabled influence operations.

3.6 LLM for malware detection and classification

Lorenzo Cavallaro (University College London, GB)

Large Language Models (LLMs) are emerging as a powerful paradigm for malware detection and classification, offering a flexible alternative to traditional signature-based and heuristic approaches. By analyzing source code, decompiled binaries, and system call traces, LLMs can reason about program semantics and malicious intent, enabling accurate classification of malware families and even zero-shot detection of previously unseen threats. This talk surveys recent advances in applying LLMs to malware analysis, including code-level reasoning over malicious logic, the use of decompilation tools to bridge binary and source-level analysis, and the classification of system call sequences to distinguish benign from malicious behavior. We also discuss context-driven approaches for domain-specific threats, such as Android malware and script-based attacks, as well as the role of LLMs in explainable malware analysis through natural-language descriptions of observed behavior. While modern transformer-based models have demonstrated significant performance gains on complex and evolving malware families, important challenges remain, including context window limitations, structural complexity of real-world malware, and adversarial evasion techniques. The talk concludes by outlining open research directions for scaling, hardening, and operationalizing LLM-based malware detection systems.

3.7 Impossibility Results in AI Safety

Eric Wong (University of Pennsylvania – Philadelphia, US)

What kinds of defenses are possible for safety research? We will discuss two impossibility results, one theoretical and one empirical. Theoretically, we will demonstrate how rule-following can always be attacked with token-level jailbreaks as an architectural inevitability of attention. Empirically, we will demonstrate how a dedicated attacker aiming to misuse a model can use decomposition to hide their intent and make their queries indistinguishable from benign queries. We will end with some optimistic discussion on how these studies inform potential ways forward.

References

1 Xue, Anton, Avishree Khare, Rajeev Alur, Surbhi Goel, and Eric Wong. "Logicbreaks: A framework for understanding subversion of rule-based inference." arXiv preprint arXiv:2407.00075 (2024).
2 Brown, Davis, Mahdi Sabbaghi, Luze Sun, Alexander Robey, George J. Pappas, Eric Wong, and Hamed Hassani. "Benchmarking Misuse Mitigation Against Covert Adversaries." arXiv preprint arXiv:2506.06414 (2025).
3 Guardieiro, Vitoria, Adam Stein, Avishree Khare, and Eric Wong. "Instruction Following by Boosting Attention of Large Language Models." arXiv preprint arXiv:2506.13734 (2025).

3.8 AI-based Traffic Modeling for Network Security and Privacy: Challenges Ahead

*Dinil Mon Divakaran (Institute for Infocomm Research, A*STAR – Singapore, SG)*

This talk reflected on the advancements in network traffic analysis using machine learning (ML) and discussed the challenges that remain in effectively securing networks and protecting user privacy.

AI-based traffic models are developed to address various challenging problems in network security and privacy (NetS&P). Deep Learning (DL) models are particularly valuable as network data is enormous (a 1 Gbps link can generate hundreds of GBs per day) and the feature space is vast (representation of a short browsing session may require 2,000 features).

Some of the common and relevant NetS&P tasks where AI models are developed are i) Detection of network anomalies, ii) Classification of attacks into different types, iii) Website fingerprinting attack on user privacy, and iv) IoT device identification. There are also other problems of interest, such as LLM token inference attack, de-anonymization attack on Tor networks, countering of network fingerprinting attacks on user privacy, evading network-based censorship, etc. Research on these problems have helped us identify the features from (encrypted) network traffic that are useful for modeling.

Despite these advancements, there are significant challenges in the domain of traffic modeling and analysis that need to be addressed to effectively secure our networks from evolving threats and attacks.

Data Challenge: Lack of high-quality, labeled real-world network traffic data, primarily due to the privacy risks associated with leaking sensitive user information. Existing open datasets are shown to have artifacts that affect the evaluations, consequently creating hurdles for objective evaluations and generalization of AI models developed.

Practical Deployment: Feature extraction for ML inference is challenging as it demands real-time per-packet processing at network speeds of up to hundreds of Gbps. Programmable data planes presents an opportunity to tackle this.

Explainability: Cybersecurity solutions need to provide human-consumable explanations for the predicted results to build trust between different stakeholders, to assist analyst decide on the next steps, to improve model performance as well as to help build cost-effective solutions.

Model Convergence: Is it time to move beyond multiple task-specific model, to build a foundation model for network traffic analysis? An effective foundation network model must possess key properties, including flexible representation of traffic to deal with different formats and granularity of packet captures at different networks, robustness to missing and noisy labels during fine-tuning, and explainability for various downstream tasks.

References

1 Dinil Mon Divakaran, "Traffic Modeling for Network Security and Privacy: Challenges Ahead," in *Proc. AAAI Workshop on AI for Cyber Security*, 2026.

3.9 Safety of Large Language Models Beyond English

Wojciech Kusa (NASK – Warsaw, PL)

As Large Language Models (LLMs) continue to evolve, ensuring their safety across multiple languages has become a critical concern. While LLMs demonstrate impressive capabilities in English, their safety mechanisms may not generalize effectively to other languages, leading to disparities in toxicity detection, bias mitigation, and harm prevention. In this talk, I will present a systematic review examining the multilingual safety of LLMs by synthesizing findings from recent studies that evaluate their robustness across diverse linguistic and cultural contexts beyond the English language. I will discuss the methodologies used to assess multilingual safety and identify challenges such as dataset availability and evaluation biases. Based on this analysis, I will highlight gaps in multilingual safety research and provide recommendations for future work. I will also show recent results in multilingual refusal alignment, illustrating where smaller language models struggle with following instructions in low-resource language settings.

3.10 LLM Security – An industrial and an end-users' perspective

Kathrin Grosse (IBM Research Europe – Zürich, CH)

Joint work of Javier Carnerero Cano, Marc Stoecklin, Ian M. Molloy, Vinayaka Pandit, Mark Purcell

LLMs are widely used in both industry and by end users. Yet, on both ends of these spectrums, there are very different challenges, which we discuss in this talk. On the one hand, companies must ensure the availability, confidentiality, and integrity of systems on a level that supports both governance and auditability. Due to the complexity of LLM-based programs, this requires many aspects when, for example, properly implementing access control. Orthogonally, to ensure governance and safety of these systems, proper testing suits are needed. We review ARES, IBM's open source library. On the other hand, end-users face different dilemmas when interacting with LLMs. They are often unaware of pitfalls and engage, potentially unaware, in security-relevant behavior. Also, about a third of the end-users admitted to jailbreaking their chatbots. However, this behavior is not malicious but motivated by play and curiosity about the limitations of the new, end-user-friendly technology. Lastly, we briefly discuss AI security incident reporting as a possible solution to the open questions raised in both areas.

References

1 Kathrin Grosse, Nico Ebert: Prevalence of Security and Privacy Risk-Inducing Usage of AI-based Conversational Agents. CoRR abs/2510.27275 (2025)

4 Working Groups

During a match-making session on Day 2, the interests expressed by the participants were consolidated into a set of six working groups, each addressing a distinct theme related to the security, privacy, and safety of LLMs and AI-enabled systems.

The topics discussed in the subsequent break-out sessions were tailored to the specific interests of the participants in each working group and were selected through informal topic-selection discussions during the seminar. The following themes and underlying open questions are representative of the discussions that emerged across the groups:

- **AI system capabilities certification.** How can the capabilities and limitations of AI systems be systematically characterized and certified? What forms of assurance or guarantees are meaningful for complex, adaptive systems such as LLMs?
- **LLM manipulation on humans.** To what extent can LLMs influence, persuade, or manipulate users, intentionally or unintentionally? How should such risks be modeled, measured, and mitigated in real-world deployments?
- **Privacy risks of LLMs.** What new privacy threats arise from training, fine-tuning, and interacting with LLMs? How can risks such as memorization, inference, and data leakage be systematically assessed and reduced?
- **Attack surface of LLMs.** How can the attack surface of LLM-based and AI-enabled systems be decomposed and analyzed at the system level, including components such as prompts, tools, agents, external services, and data pipelines?
- **LLM safety and security benchmarks.** What benchmarks are needed to meaningfully evaluate the safety and security of LLMs beyond traditional accuracy metrics? How can benchmarks remain relevant in the face of rapidly evolving models and attack techniques?

- **Practical implications and deployment challenges.** How do all these concerns translate to real-world applications, and what trade-offs arise between security, privacy, safety, usability, and performance?
- **Societal impact of LLMs.** How may the widespread deployment of LLMs affect the functioning of society, including economic structures, legal systems, academic peer review, and democratic processes? Are the existing AI regulatory frameworks able to govern their development and deployment? In parallel, how might LLMs themselves become tools within legal and institutional systems, and what risks arise from such integration?
- **Concentration of AI capabilities.** How does the growing concentration of cutting-edge LLMs and AI infrastructure within a small number of large technology companies affect research, innovation, and societal power dynamics? What mechanisms could help democratize access to models, data, and compute so that academia and broader society can meaningfully participate in the development and governance of advanced AI systems?

While the seminar could not fully address all of the questions outlined above, identifying and articulating them was itself an important outcome of the event. The discussions highlighted these issues as central challenges for the research community and for society at large, helping to frame a broader agenda for future work in the area of LLM safety and security.

Among the posed themes, six were explored in greater depth through dedicated working groups, each of which produced a written report summarizing their findings and perspectives. In the remainder of this document, the working group reports on the following topics are presented:

- AI system capabilities certification;
- LLM manipulation of humans;
- Privacy risks of LLMs;
- Responsible usage of LLMs in peer review;
- Modeling the attack surface of LLM- and AI-enabled systems;
- LLM safety and security benchmarks.

4.1 AI System Capabilities Certification

Tobias Callies (Universität der Bundeswehr München, DE)
Jonathan Petit (Qualcomm – Boxborough, US)
Mathilde Aliénor Raynal (EPFL Lausanne, CH)
Eric Wong (University of Pennsylvania – Philadelphia, US)

Models are being released or deployed at a record pace. Nowadays, users are exposed to chatbots on most websites, while developers must decide which pre-trained model to deploy or fine-tune. This explosion of models makes it difficult to decide which one to use (from a user or developer standpoint), or more importantly, which one to trust. Indeed, how can a user (or an AI agent) be aware of the security, privacy, and safety capabilities of the AI system she interacts with? We propose a solution to increase the transparency and trust in AI systems. More specifically, the proposed solution should aim to:

- Increase transparency for the user about the AI system/model capabilities using state-of-the-art Human-Computer-Interaction techniques.

- Create accountability (for the model deployer).
- Enable "model shopping" and specialized models.

As we noted that existing certification regimes helped improving the reliability and cybersecurity of certified devices (e.g., RED certification of wireless devices, U.S Cyber Trust Mark, ISO 21434 for automotive), we devised a certification process that would output a AI trustworthiness label.

4.1.1 Methodology

This section briefly explains the methodology followed by the group.

Definition. First, we defined the term "capability" of an AI system. As the focus of this seminar is LLM security, we focused on security-related capabilities such as "prompt injection protection".

Process. Then, we specified a three-step certification process:

- Step 1 (Define): Specify AI system task(s) and harm(s). Tasks define the responsibilities and actions an AI system is expected to perform. Harms define and categorize potential threats to protection-worthy assets.
- Step 2 (Measure): Specify the evaluation process and parameters. For example, use a population of AI agents that is representative of the target population. The objective is to assess the probability of being exposed to the harm(s) defined in Step 1. We proposed a badge system that would enable users to understand the risk of using the system under test.
- Step 3 (Update): Any respectable certification process must be updated to reflect the ecosystem.

4.1.2 Next Steps

After presenting to the entire seminar's participants, we gathered extensive feedback and interest in writing a concept paper. We will work towards that goal.

4.2 On LLM manipulation of humans

Lujo Bauer (Carnegie Mellon University – Pittsburgh, US)
Rebekka Burkholz (CISPA – St. Ingbert, DE)
Lorenzo Cavallaro (University College London, GB)
Andreas Happe (TU Wien, AT)
Pavel Laskov (Universität Liechtenstein – Vaduz, LI)
Nils Lukas (MBZUAI – Abu Dhabi, AE)
Hammond Pearce (UNSW – Sydney, AU)
Lea Schönherr (CISPA – Saarbrücken, DE)

Humans have always sought to influence one another, toward both positive and negative ends. Political leaders campaign for the support of their citizens. Suitors emphasise their most attractive qualities. Scammers and criminals attempt to manipulate victims into surrendering valuable assets.

Historically, such influence campaigns were constrained by the inherent capabilities and resources of the influencing party. A single scammer could reach only a small number of targets at a time. Large-scale influence required large organisations, often at the scale of governments (e.g., mass propaganda, nation-level education policy).

With the rise of generative AI (GenAI), these constraints are fundamentally changing. Influencers can now scale their campaigns both in breadth (tailoring highly customised messages to many more targets) and in depth (conducting extensive background research on individuals or groups autonomously). Further, GenAI-based influence needs not be bounded by human limitations such as tiredness, empathy, or guilt. In our discussion, we examine a taxonomy of GenAI-based influence, pose open questions for the research community, and offer our perspective on what the future may hold for autonomous influence campaigns.

In interactions between humans and an LLM-driven influencer, communication can unfold in multiple ways. It may follow an intentional solicitation (e.g., a user engaging with a chatbot) or occur without the human ever realising that an AI system is involved. Influence can also take place over multiple turns or within a single exchange. Depending on the application, the agent may be granted access to additional tools and information that amplify its influence – for example, by retrieving external information or personal data about the human target.

There is therefore a large design space to consider when thinking about the many ways humans may be influenced by LLMs. We organise our thoughts into four questions for future research:

4.2.1 Q1: On Influencing Humans – What are the properties of content that best influence humans in this setting?

We already know some common patterns that shape how people respond. Messages that appear trustworthy, personalised, and context-aware can lower suspicion and increase engagement. Emotional triggers such as fear, sympathy, outrage, or excitement can drive rapid decisions. Creating a sense of urgency or stress can cause people to act before they have time to reflect. Decomposing a questionable request into small, seemingly harmless steps (a "foot-in-the-door" strategy) can make it easier for someone to agree without noticing the cumulative effect. Understanding these patterns in the context of GenAI is crucial so that people, platforms, and policymakers can better recognise and resist manipulation.

4.2.2 Q2: LLM-based Capabilities – What could LLMs enable in this setting that is infeasible/uneconomical for pure human/non-LLM influencers?

LLMs enable unprecedented scalability, allowing attackers and other influencers to dramatically increase the number of people they can target, including in semi-autonomous ways. This efficiency can make small or low-value operations economically viable, because generating customised content and interactions becomes extremely cheap. LLMs can also deepen the quality and speed of reconnaissance, rapidly gathering and analysing large volumes of unstructured data to tailor messaging, timing, and strategy to individual targets or micro-groups. Even partial automation can substantially increase reach and impact. Unlike humans, LLMs do not become tired, demotivated, remorseful, or empathetic; if misused, they can sustain persistent, high-volume activity without emotional or cognitive limits.

4.2.3 Q3: LLM-powered Influencers – How is an LLM-powered "influencer" best designed to enable Q1/Q2?

Designing an LLM-powered "influencer" raises difficult technical, ethical, and legal questions. We must define what counts as "influence" and how to measure it – for example, changes in attitudes, short-term engagement (clicks, shares), or longer-term behaviour. We must also account for who is being influenced, the context of the interaction, the time and effort required to influence them, and realistic limits on how far automated systems can shape beliefs and actions.

"Good" or prosocial influencers should reject mis- and disinformation and favour honest, transparent communication. Their design must incorporate robust legal and ethical safeguards that respect autonomy and rights, avoid covert manipulation, and support informed, voluntary choices. By contrast, "bad" influencers (such as those deployed by scammers or state actors with malicious intent) will not be constrained by such norms, and may optimise solely for desired, potentially harmful outcomes. Understanding these unconstrained designs is essential for anticipating threat models and developing effective mitigations and defenses.

4.2.4 Q4: Mitigation – How can LLM-powered influencers be mitigated?

Mitigating the risks of LLM-powered influencers is challenging because it requires understanding and measuring their impact on human cognition and behaviour. While LLMs sometimes exhibit recognisable patterns – such as repetitive phrasing, overly consistent tone, or specific stylistic signatures – these signals are imperfect and may diminish as models improve.

Addressing these risks will likely require a combination of technical, institutional, and societal responses. On the technical side, stronger methods are needed for detecting and tracing influence, including behavioural signals, metadata standards, platform-level transparency mechanisms, and clearer audit trails. Techniques such as watermarking, provenance tracking, and prompt extraction (e.g., retrieving a model's system prompt or high-level intentions) may play supporting roles. On the societal side, increasing information diversity, investing in public education, and strengthening digital literacy can all help build resilience to manipulation. Finally, any defensive approach must explicitly account for the dual-use nature of LLM influencers: the same tools can be deployed by both "good" and "bad" actors, which complicates which interventions are feasible, effective, and acceptable.

4.3 Privacy risks of LLMs

Giovanni Cherubin (Microsoft Research – Cambridge, GB)
Ana-Maria Cretu (EPFL – Lausanne, CH)
Stephan Günnemann (TU München – Garching, DE)
Qianying Liao (KU Leuven, BE)
Vera Rimmer (KU Leuven, BE)
Ali Satvaty (University of Groningen, NL)

Large Language Models (LLMs) introduce numerous privacy challenges, some of which were known in traditional Machine Learning (ML) systems, others that are entirely novel. We discuss ways in which personal data of individuals can be misused in the lifecycle of LLMs,

ranging from data leakage and sensitive inferences to LLMs influencing the behavior of users. We also map out how LLMs differ from traditional ML applications, and how those differences can scale up existing threats or introduce new privacy harms.

4.3.1 Data leakage from LLMs

LLMs risk leaking the data of users on which they were trained. There are four different stages in the LLM lifecycle where they ingest data.

In the **pre-training stage**, LLMs are trained on vast amounts of publicly available personal data. It is debatable whether the leakage of personal data about a user that was published with their consent constitutes a privacy violation. However, in many cases personal data is online without the consent of the user, or the user may decide to remove their online data after it was scraped and used in LLM training; this is a great source for concern. A further concern is the potential increase in the ability to profile the users or infer sensitive information about them, as compared to not having been trained at all on their data.

In the **fine-tuning stage**, LLMs are trained on smaller and domain-specific datasets (e.g., medical notes, transaction data, customer chats). Fine-tuning on private data can amplify privacy exposure by making the model more aware of individuals or groups, and may lead to increased leakage about individuals due to the smaller dataset size.

In the **in-context learning stage**, private data can be embedded directly in prompts or system instructions. This data may leak in subsequent user interactions, and it is particularly hard to prevent this from happening.

To limit the risks of data leakage, the ML privacy community has built a solid understanding of the various attacks and defenses for traditional ML (supervised learning). Yet, this research is not always directly applicable to LLMs, as they differ from classic ML in several respects. For example, LLMs are trained as general purpose models, which increases the attack surface. Further, it is harder to audit their privacy, as this requires to contextually determine the "privacy unity", e.g., whether to protect individual documents, all the data belonging to a user, or specific secrets that are shared across users.

Beyond training data leakage, there is a risk of data exfiltration attacks in the **AI agentic stage**, where the LLM can access user files and integrate with external systems. The boundaries between stored data, model memory, and ephemeral inference become increasingly blurred. This complicates GDPR-style risk assessment, transparency obligations, and expectations about how long data persists, whether deletion is respected, and what constitutes a privacy unit in LLMs.

4.3.2 Privacy harms from interactions between a user and the LLM

Privacy harms may arise from sharing information with a system that mimics trustworthiness. For example, users increasingly disclose intimate details to LLM-powered chatbots, far more than they would to friends or doctors. Based on these interactions, LLM-powered chatbots can create internal profiles of users on inferred attributes, preferences, vulnerabilities, or political leanings. These inferences can persist across conversations (e.g., ChatGPT's memory feature), shape future responses, and influence user's behavior. Even if the model does not exploit this information, profiling raises questions about autonomy, manipulation, and privacy as freedom from interference (e.g., Article 12 GDPR, UN privacy rights).

4.3.3 LLMs as privacy attackers

At scale, LLMs can also function as powerful tools to help attackers violate people's privacy, by exploiting linkages and contextual inferences. LLMs can aggregate data from web browsing tools, API calls, and user history, enabling attacks that combine inference, linkage, and manipulation. LLMs can also synthesize facts from multiple data sources to infer information about individuals from public fragments, e.g., link a user's Linkedin to anonymous postings or guessing medical, financial, or political attributes. Unlike search engines, LLMs generate new spans of text, making the disclosure harder to track and validate, especially since ground truth is unavailable for many inferred secrets. Even if the inferred information is "public", the act of connecting, aggregating, or contextualizing can scale the privacy harms that previously required a significant human effort.

4.3.4 LLMs as the privacy defender

Paradoxically, LLMs may be a better tool to defend privacy than traditional ML systems, because of their ability to understand natural language and context; crucially, this sets them apart from standard privacy techniques (e.g., DP mechanisms). One example is text anonymization, which typically requires removing personally identifiable information and paraphrasing to remove secrets. LLMs could enforce personal-data minimization, detect exfiltration attempts, refuse risky actions, or provide dynamic privacy explanations to users. However, realizing this defensive potential requires solving major open problems: defining what privacy means in generative contexts, creating measurable risk metrics, distinguishing intended vs. non-intended behavior, and ensuring that safeguards cannot be bypassed.

4.3.5 Conclusions

These discussions highlight that privacy in LLMs goes beyond traditional ML concerns: from data leakage at various training stages, to inference-based profiling and large-scale aggregation. Interestingly, LLMs can also offer novel means to enhance privacy, e.g., via more contextually aware anonymization. Further research is essential to adequately describe privacy threats, clarify acceptable use contexts, prevent misuse, and to harness the potential privacy benefit that LLMs may offer.

4.4 Responsible Usage of LLMs in Peer Review

Giovanni Apruzzese (Universität Liechtenstein – Vaduz, LI)
Amir Houmansadr (University of Massachusetts Amherst, US)
Marc Juarez (University of Edinburgh, GB)
Florian Kerschbaum (University of Waterloo, CA)
Wojciech Kusa (NASK – Warsaw, PL)
Fabio Pierazzi (University College London, GB)
Konrad Rieck (BIFOLD – Berlin, DE & TU Berlin, DE)
Christian Wressnegger (Karlsruher Institut für Technologie, DE)

Large language models (LLMs) are increasingly embedded in scientific research practices. Their use for writing assistance, programming, and exploratory analysis is now widespread and well documented [1, 2, 3, 4, 5]. More recently, LLMs have begun to influence peer review, raising concerns regarding accountability, fairness, quality, and policy compliance [6, 7]. Empirical analyses have revealed that, following the release of ChatGPT, the reviews submitted for ICLR'23 included a substantially higher number of words typically associated to LLMs [8]. In response, venues and publishers have adopted policies ranging from explicit prohibitions on LLM use [9] to controlled LLM integration into the review process [10], as well as revising their respective authorship and reviewing guidelines [11, 12].

This working group examined "responsible LLM usage in peer review" from a security-oriented perspective. The guiding assumption was that irresponsible LLM usage (e.g., delegating substantial intellectual tasks without human oversight, or violating explicit venue policies) can compromise the integrity of the scientific process. The intention was to discuss how security mechanisms, whose real-world objective is more often to discourage rather than complete prevention, can help in addressing this issue of "irresponsible" usage of LLM in peer review [13, 14].

4.4.1 Discussed Problems

The discussion mostly concerned the detectability of LLM-generated content, particularly LLM-written reviews. Participants broadly agreed that they had encountered such reviews and that these were often unhelpful or incorrect. The group discussed several detection strategies:

- Canary-based approaches, such as including hidden prompts in papers to reveal whether reviewers submit papers to LLMs, have been proposed in prior work [15]. However, such approaches rely on security through obscurity and are easily circumvented once known.
- Automated detectors and watermarking techniques (both academic and commercial [16, 17, 18, 19]) were also discussed. These methods were viewed as fundamentally unreliable due to false positives (including flagging "benign" uses, such as language polishing) and ease of evasion through minor editing or prompt engineering.
- More elaborate strategies, such as finding the prompt that generates a given review, were deemed impractical, due to the intrinsic difficulty in finding a prompt without falling into subjective choices, as well as raising potential confidentiality issues (since the process involves submitting the prompt to an LLM alongside the reviewed paper – which can be done by the authors of the paper, but not by, e.g., PC members during a confidential peer-review process).

The group therefore converged on the conclusion that robust detection of LLM-generated reviews is not currently feasible.

The group also questioned whether LLM usage in scientific contexts is intrinsically problematic. Opinions varied considerably. In the context of peer review, three use cases were examined:

- First, using LLMs for language polishing or grammar was viewed by some as not necessary, while others argued that improved clarity benefits authors and reviewers alike [5].
- Second, generating reviews with LLMs – either partially or entirely – was considered unacceptable by many participants, as reviewing is meant to be done by "peers" (i.e., humans); moreover, reliance on LLMs is disrespectful towards other reviewers who spend considerable time for their reviewing duties. Others, however, argued that what really matters is the quality of a review, rather than the process by which reviews are produced: if LLMs can lead to better reviews, their usage would be justified.
- Third, the use of LLMs to analyse the submitted reviews (an application evaluated by the organizers of ICLR'25 [10]) was viewed as potentially beneficial. However, some concerns were still raised about the practical utility of LLM feedback and the difficulty of objectively measuring improvements in review quality.

The group also discussed LLM usage in writing scientific articles. Using LLMs for editing and stylistic refinement was generally considered acceptable. However, generation of content from scratch raises concerns such as hallucinated references and factual inaccuracies [20]. This problem is aggravated by the fact that identifying such errors is challenging for reviewers, who are not typically expected to verify every citation and its surrounding context. Participants emphasized the need for automated mechanisms to support reference checking.

The group also discussed LLM usage by students, particularly PhD candidates. Participants agreed that LLM use in graded coursework (e.g., for exams involving use of internet-capable devices, or for homework) was problematic. However, greater concern was expressed about "undeclared" usage of LLMs for writing research documents, given that supervisors and co-authors remain accountable for any errors introduced. The importance of educating early-career researchers about the pros-and-cons of LLMs in scientific work was strongly emphasized.

Disclosure of LLM usage in peer-review contexts generated further disagreement. From the author side, some venues explicitly require authors to disclose LLM usage (e.g., SaTML'26, or S&P'26); however, it is unclear if authors truly comply with such requirements (participants reported accounts of clearly LLM-written content in papers they reviewed which had not been acknowledged). From the reviewing side, disclosure may also undermine trust in reviews, even when LLMs are used responsibly. In either case, while disclosure could increase transparency, it is uncertain how to ensure that researchers comply with disclosure requirements.

4.4.2 Possible Approaches

Despite divergent views, a shared conclusion emerged: the quality, correctness, and accountability of scientific content matter more than the tools used to produce it. Based on this principle, several recommendations were formulated.

- Reviewers should remain engaged beyond submitting their own reviews by critically examining other reviews and correcting factual errors when possible. If LLMs are (responsibly) used, any time savings should be reinvested in enhancing the overall quality of the peer-review, which goes beyond the mere act of submitting a review, since it also encompasses discussions with other reviewers.

- The community should also place greater value on high-quality reviewing by strengthening incentive and recognition mechanisms, as well as pointing out cases of "policy violations" wherein some researchers (whose identity should not be disclosed) received some penalty due to illegitimate usage of LLMs.
- Finally, more education, especially for young researchers, is needed to foster a nuanced understanding of LLM capabilities, limitations, and responsibilities.

4.4.3 Conclusion

Overall, responsible LLM usage in peer review cannot be guaranteed via detection alone. Instead, preserving review quality requires clearer norms, better incentives, targeted education, and accountability mechanisms that discourage irresponsible behavior while acknowledging the practical role LLMs may play.

References

1 A. Birhane, A. Kasirzadeh, D. Leslie, and S. Wachter, "Science in the age of large language models," *Nature Reviews Physics*, vol. 5, no. 5, pp. 277–280, 2023.

2 M. Binz, S. Alaniz, A. Roskies, B. Aczel, C. T. Bergstrom, C. Allen, D. Schad, D. Wulff, J. D. West, Q. Zhang *et al.*, "How should the advancement of large language models affect the practice of science?" *Proceedings of the National Academy of Sciences*, vol. 122, no. 5, p. e2401227121, 2025.

3 J. Clusmann, F. R. Kolbinger, H. S. Muti, Z. I. Carrero, J.-N. Eckardt, N. G. Laleh, C. M. L. Löffler, S.-C. Schwarzkopf, M. Unger, G. P. Veldhuizen *et al.*, "The future landscape of large language models in medicine," *Communications medicine*, vol. 3, no. 1, p. 141, 2023.

4 M. Naddaf, "How are researchers using AI? Survey reveals pros and cons for science," *Nature*, 2025, https://www.nature.com/articles/d41586-025-00343-5.

5 M. Geng, C. Chen, Y. Wu, Y. Wan, P. Zhou, and D. Chen, "The impact of large language models in academia: from writing to speaking," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 19 303–19 319.

6 J. Zou, "ChatGPT is transforming peer review – how can we use it responsibly?" *Nature*, vol. 635, no. 8037, pp. 10–10, 2024, https://www.nature.com/articles/d41586-024-03588-8.

7 G. O'Brien, "How scientists use large language models to program," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–16.

8 W. Liang, Z. Izzo, Y. Zhang, H. Lepp, H. Cao, X. Zhao, L. Chen, H. Ye, S. Liu, Z. Huang *et al.*, "Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews," in *International Conference on Machine Learning*. PMLR, 2024, pp. 29 575–29 620.

9 CVPR, "CVPR 2025 Reviewer Guidelines," 2025, https://cvpr.thecvf.com/Conferences/2025/ReviewerGuidelines.

10 N. Thakkar, M. Yuksekgonul, J. Silberg, A. Garg, N. Peng, F. Sha, R. Yu, C. Vondrick, and J. Zou, "Can LLM feedback enhance review quality? a randomized study of 20k reviews at ICLR 2025," *arXiv:2504.09737*, 2025.

11 ACM, "ACM Policy on Authorship," 2025, https://www.acm.org/publications/policies/new-acm-policy-on-authorship.

12 – – , "ACM Peer Review Policy Frequently Asked Questions," 2025, https://www.acm.org/publications/policies/peer-review-faq.

13 V. M. Bier, "Choosing what to protect," *Risk Analysis: An International Journal*, vol. 27, no. 3, pp. 607–620, 2007.

14 L. A. Gordon and M. P. Loeb, "The economics of information security investment," *ACM Transactions on Information and System Security (TISSEC)*, vol. 5, no. 4, pp. 438–457, 2002.

15 M. G. Collu, U. Salviati, R. Confalonieri, M. Conti, and G. Apruzzese, "Publish to perish: Prompt injection attacks on LLM-assisted peer review," *arXiv preprint arXiv:2508.20863*, 2025.
16 Pangram, "AI Detection that actually works," 2025, https://www.pangram.com/.
17 M. Naddaf, "Major AI conference flooded with peer reviews written fully by AI," *Nature*, 2025, https://www.nature.com/articles/d41586-025-03506-6.
18 R. Zhang, S. S. Hussain, P. Neekhara, and F. Koushanfar, "REMARK-LLM: A robust and efficient watermarking framework for generative large language models," in *33rd USENIX Security Symposium (USENIX Security 24)*, 2024, pp. 1813–1830.
19 T. Lee, S. Hong, J. Ahn, I. Hong, H. Lee, S. Yun, J. Shin, and G. Kim, "Who wrote this code? watermarking for code generation," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 4890–4911.
20 P. Esau, A. Cui, and A. Adam, "GPTZero finds over new 50 hallucinations in ICLR 2026 submissions," 2025, https://gptzero.me/news/iclr-2026/.

4.5 Modeling the Attack Surface of LLM- and AI-enabled Systems

Hyrum Anderson (CISCO Systems – San Jose, US)
Javier Carnerero Cano (IBM Research – Dublin, IE)
*Dinil Mon Divakaran (Institute for Infocomm Research, A*STAR – Singapore, SG)*
Mario Fritz (CISPA – Saarbrücken, DE)
Emil C. Lupu (Imperial College London, GB)
Irina Nicolae (Apple – Paris, FR)
Tim Van hamme (KU Leuven, BE)

The rapid integration of large language models (LLMs) into different applications across various enterprise sectors has led to new and complex security challenges, as LLMs have inherent weaknesses such as non-deterministic output, hallucinations, and privacy leakages. Although recent research has focused largely on model-level issues such as prompt injection, the field lacks a systematic characterization of threats that emerge when LLMs are embedded in real applications, particularly with respect to implications such as business risks, motivation and effort estimates, both from an attacker and defender perspective. In this work, we attempt to fill this gap by (i) proposing a risk assessment framework for modeling the attack surface of LLM systems and (ii) mapping existing defense strategies to their roles and impacts from a system-level viewpoint.

4.5.1 Impact-oriented LLM Application Threat Modeling Framework

We outline an impact-oriented threat-modeling framework for LLM applications that connects business risks to system-level vulnerabilities. The framework integrates four elements: (i) a business-impact analysis that identifies critical assets and evaluates the consequences of their compromise; (ii) an attack-surface modeling process that derives concrete system vulnerabilities from high-level impact scenarios; (iii) a characterization of threat sources, spanning highly capable adversaries (e.g., nation-state actors) to insiders and accidental actors; and (iv) a lightweight risk-scoring method that combines scenario likelihood with impact severity. Together, these components provide a structured approach for assessing and prioritizing risks in LLM-enabled systems.

4.5.2 Mitigating the Risks

Risk-mitigation decisions can be guided by two complementary system-level perspectives. The weakest-link view assesses risk along the easiest attack path, revealing the objectives and vectors an adversary is most likely to exploit. The full attack-surface view aggregates risk across all feasible paths, highlighting overall exposure and capturing cases where multiple comparable or cumulative attack routes create material risk even when no single path dominates.

The proposed framework formalizes how impacts, vulnerabilities, threat sources, and risk metrics interrelate, providing a structured basis for prioritizing defenses and guiding secure system design. In doing so, it offers a path for organizations to assess, communicate, and mitigate the real-world risks introduced by LLM integration. More broadly, the framework is intended to equip practitioners with practical guidance for securing LLM-enabled systems and to help researchers identify and prioritize high-impact directions in LLM security and privacy.

4.6 LLM Safety & Security Benchmarks

Maksym Andriushchenko (ELLIS – Tübingen, DE)
Marc Fischer (Snyk – Zürich, CH)
Kathrin Grosse (IBM Research Europe – Zürich, CH)
Daphne Ippolito (Carnegie Mellon University – Pittsburgh, US)
Andrew Paverd (Microsoft Research – Cambridge, GB)
Advije Rizvani (Universität Liechtenstein – Vaduz, LI)
Ilia Shumailov (AI Sequrity Company – London, GB)

This working group investigated the challenges in evaluating the trustworthiness (specifically security and safety) of current state-of-the-art large language models (LLMs), AI applications, defenses meant to augment them, and attacks considered to break them. This spans the robustness of models to jailbreaks, prompt injections by users, indirect prompt injections via tools or third party actors as well as ways to defend against these. We outline current issues and summarize our recommendations for further benchmarks and evaluations.

The ubiquitous use of large language models in virtually all areas of life and to dynamically drive the behavior of computer programs ("AI Applications" and "AI Agents") raises questions about their trustworthiness and how to establish it. Key among these concerns are safety and security, e.g.

Safety Entails, among other things, the alignment of the underlying AI model and the overall AI application, i.e., whether a model will refuse to engage in hate speech or to explain or enact dangerous practices (e.g. "How to build a Bomb?"). While these properties are often established at training time, this initial alignment is the often circumvented by jailbreaks [9] and/or direct prompt injections [8].

Security Assesses whether the AI model or AI application can engage in dangerous or destructive behavior by hijacking it's behavior, e.g. leaking or deleting information, in some way. This is typically achieved via indirect prompt injections.

Typically, these properties are best measured using benchmarks [10, 11]. In this work, we outline the current challenges and shortcoming of these benchmarks, our suggestions to overcome these along with open challenges.

4.6.1 Benchmarking AI Security & Safety

An enduring cycle of incremental advancements in AI defences characterizes the current landscape of AI security literature [2, 1, 5]. This research typically follows a predictable three-part loop: the specification of a system evaluation, the development of attacks against that system, and the subsequent creation of defenses [3]. There is the cycle of defenses and attacks each getting better as newly devised attacks are tailored to exploit the specific mechanisms of defenses [2, 1, 5], or evaluations that showed the performance of defense were flawed [4]. In either case, new defenses are proposed to remove the uncovered mechanisms. There is also the larger cycle of researchers realizing that the specification of the problem itself is flawed [3], and the attacks and defenses being created are insufficiently relevant to real-world systems [6, 7]. This all perpetuates a reactive and iterative sequence of "Specification → Attack → Defence → ... → Attack → Defence."

We argue that the non-determinism in LLMs and the fact that advancements in LLMs are typically interoperable with their predecessors lead to a rapid and low-reproducibility development cycle for AI Security & Safety compared to other areas of Cyber Security. LLMs are increasingly being incorporated into the specification of the problem (e.g., the LLM underlying the agent in an agentic system), the attacks and defense methods, and the evaluation of attack success (e.g., LLM-as-judge methods). The behaviour of LLMs is non-deterministic–the same prompt can yield different generations, and small changes to the prompt can result in significant changes to the generations. This means, all the components which rely on LLMs are also non-deterministic in their behaviour. Moreover, it is usually trivial to swap out LLM x with its successor LLM $x + 1$, or with an entirely different LLM. This means that:

- Updated LLMs may be inherently less vulnerable to attacks without the need for post-hoc defenses. They may have also been trained on the evaluation scenario, resulting in inflated confidence in scenarios being "solved."
- Attack and defense development becomes faster when there are so many "easy" parameters that can be changed that may significantly improve attack/defense success.

This paradigm is further complicated by the "noisy" nature of evaluating the efficacy of these attacks and defenses, which are often assessed in limited and simplistic "Rosetta Stone" environments, for example with attacks that only exceed with low probability. Consequently, many proposed attacks and defenses offer only short-term value, quickly losing their relevance as they are over-fitted to environments that are challenging and time-consuming to build and maintain. At the same time, environment data is easily incorporated into training future models, invalidating the assessment.

We argue that the only way to break this inefficient cycle, the research community must shift its focus towards two more fundamental goals: the creation of attacks that can generally break existing defenses and the development of defenses that are secure by design. We contend that progress in either of these areas is contingent upon incentivizing the development of meaningfully diverse, realistic, and dynamic evaluation environments. Such environments are critical for rigorously assessing the generalizability of defensive measures and the true potency of attack vectors, thereby fostering more robust and lasting advancements in AI Security & Safety.

Some proposed attacks only succeed with a low probability, limiting their practical impact. Consequently, many of these meticulously crafted attacks and defenses offer only short-term value, quickly losing their relevance as they are overfitted to static evaluation environments that are both challenging and resource-intensive to build and maintain. The problem is exacerbated by the fact that data from these evaluation environments can be easily incorporated into the training of future models, thereby invalidating the very assessments they were designed for. This is a critical issue, as companies are incentivized to "train on the evals," creating a situation where the benchmarks for security become moving targets that are constantly being gamed.

4.6.2 Strong Attacks, Secure by Design

To break this inefficient and unsustainable cycle, the research community must fundamentally shift its focus toward two more foundational goals: the creation of attacks that can generally break all existing defenses and the development of defenses that are secure by design. Progress in either of these ambitious directions is contingent upon a critical prerequisite: the development of diverse, realistic, and dynamic evaluation environments. Such environments are paramount for rigorously assessing the generalizability of defensive measures and the true potency of attack vectors, thereby fostering more robust and lasting advancements in AI security.

Upon deeper investigation into the complexities of this challenge, it becomes apparent that the evaluation of an AI system is not a monolithic task. A fundamental tension exists between evaluating the system as a whole versus assessing its individual components. Each component possesses unique semantics, constraints, and its own input-output domains, making a holistic evaluation incredibly difficult. The problem is compounded by the fact that it is often impossible to decompose a single model inference into its sub-components for individual evaluation. This challenge is further magnified when considering the trade-offs between evaluation time and attack efficiency. A thorough evaluation may require significant computational resources and time, which may not be practical in rapidly evolving threat landscapes.

4.6.3 Dynamic Efficient Benchmarks

Furthermore, the nature of attacks themselves is context-dependent. For instance, the relevance of "imperceptible" attacks from the old adversarial example literatures, which are subtle manipulations to inputs that are not easily detectable by humans, is not a universal constant. For many attacks, imperceptibility is a property of the context in which the model appears, such as within the structure of a markdown document or as invisible text. This highlights the limitations of evaluations conducted in sterile, context-free environments. Similarly, many current evaluations operate with restricted attack budgets, limiting the number of dimensions an attacker can manipulate. In reality, attackers may have unrestricted budgets, allowing for far more sophisticated and multi-faceted attacks. A crucial insight is that many successful attacks are, in fact, functionally exploiting the defenses themselves; they target the interpretive nature of the defenses, turning their own logic against them. This is particularly true for adaptive attacks, which are tailored to the specific defenses in place. The need for more robust evaluation is also underscored by the increasing prevalence of AI in multi-modal domains, including audio and vision, which present new and more continuous domains for potential attacks. At the same time, such expensive attacks, compared with the need to evaluate across various and wide settings, can make the benchmarking and

development of new attacks prohibitively expensive. In an adhoc exporation we found many datapoints in exsisting benchmarks to be highly correlated (succeeding or failing jointly). We thus call for benchmarks to provide represenative core sets for quick evaluation and encourage the exploration of methods to determine these automatically.

The current paradigm also struggles with the comparative evaluation of different defense systems. While it is understood that defenses need to be evaluated using strong, adaptive attacks, these attacks are, by their very nature, specific to a particular defense and environment. This specificity makes it incredibly challenging to compare the relative strengths and weaknesses of different defensive strategies in a generalized manner. Without the ability to conduct meaningful comparative evaluations, the community risks investing in defenses that are only effective against a narrow and perhaps irrelevant set of attacks.

4.6.4 Recommendations

To conclude this working group report, we outline our recommendations based on the above discussion. Above all we encourage the development of secure/safe-by-design approaches and strong adaptive attacks along with efficient, comprehensive, dynamic, interchangeable evaluation environments. In particular, benchmarks need to be dynamic in the sense of (i) multiple adaptive versions of attacks are run to deal with the stochasticity of LLMs, (ii) while the scaffolding of benchmarks is static the concrete content changes dynamically to avoid being trained on by foundation models, (iii) to further mitigate overfitting, new benchmarks are continually developed and published. To facilitate this, we call for the reuse of standardized eval hardnesses in which attacks and defenses can be deployed in a plug-and-play manner. We find datapoints within existing benchmarks to be highly correlated. Thus, we need to be able to efficiently discover the most insightful datapoints and run the evals on them. While it is crucial to test full AI applications and AI agents, often the key vulnerability can be broken down into a single jailbreak of prompt injection. We believe that benchmarks that highlight both these pathological cases as well as the full environments, can help make progress at a rapid scale. We further encourage investigation into quantifying attack budgets with respect to cost, time and imperceptibility constraints and welcome arguments to the required economics of AI safety and security, though we caution to artificially limit benchmarks with these arguments.

Our goal with these measures is to drive the creation of safe and secure AI systems by making it easy to evaluate newly proposed systems against the current state-of-the-art adaptive and generic attacks.

References

1 Nicholas Carlini. Cutting through buggy adversarial example defenses: fixing 1 line of code breaks Sabre. arXiv preprint arXiv:2405.03672, 2024.

2 Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14, 2017.

3 Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In *Advances in Neural Information Processing Systems*, volume 33, pages 1633–1645, 2020.

4 Maura Pintor, Luca Demetrio, Angelo Sotgiu, Ambra Demontis, Nicholas Carlini, Battista Biggio, and Fabio Roli. Indicators of attack failure: Debugging and improving optimization of adversarial examples. In *Advances in Neural Information Processing Systems*, volume 35, pages 23063–23076, 2022.

5 Reza Shokri et al. Bypassing backdoor detection algorithms in deep learning. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 175–183, 2020.

6 Giovanni Apruzzese, Hyrum S. Anderson, Savino Dambra, David Freeman, Fabio Pierazzi, and Kevin Roundy. "real attackers don't compute gradients": bridging the gap between adversarial ML research and practice. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 339–364, 2023.

7 Kathrin Grosse, Lukas Bieringer, Tarek R. Besold, and Alexandre M. Alahi. Towards more practical threat models in artificial intelligence security. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 4891–4908, 2024.

8 Sahar Abdelnabi, Kai Greshake, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. In *AISec@CCS*, pages 79–90. ACM, 2023.

9 Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How Does LLM Safety Training Fail? In *NeurIPS*, 2023.

10 Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents. In *NeurIPS*, 2024.

11 Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David A. Forsyth, and Dan Hendrycks. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. In *ICML*, 2024.

Participants

- Hyrum Anderson
CISCO Systems – San Jose, US
- Maksym Andriushchenko
ELLIS – Tübingen, DE
- Giovanni Apruzzese
Reykjavik University, IS
- Lujo Bauer
Carnegie Mellon University – Pittsburgh, US
- Rebekka Burkholz
CISPA – St. Ingbert, DE
- Tobias Callies
Universität der Bundeswehr München, DE
- Javier Carnerero Cano
IBM Research – Dublin, IE
- Lorenzo Cavallaro
University College London, GB
- Giovanni Cherubin
Microsoft Research – Cambridge, GB
- Ana-Maria Cretu
EPFL – Lausanne, CH
- Dinil Mon Divakaran
Institute for Infocomm Research, A*STAR – Singapore, SG
- Marc Fischer
Snyk – Zürich, CH
- Mario Fritz
CISPA – Saarbrücken, DE
- Kathrin Grosse
IBM Research Europe – Zürich, CH
- Stephan Günnemann
TU München – Garching, DE
- Andreas Happe
TU Wien, AT
- Amir Houmansadr
University of Massachusetts Amherst, US
- Daphne Ippolito
Carnegie Mellon University – Pittsburgh, US
- Marc Juarez
University of Edinburgh, GB
- Florian Kerschbaum
University of Waterloo, CA
- Wojciech Kusa
NASK – Warsaw, PL
- Pavel Laskov
Universität Liechtenstein – Vaduz, LI
- Nils Lukas
MBZUAI – Abu Dhabi, AE
- Emil C. Lupu
Imperial College London, GB
- Irina Nicolae
Apple – Paris, FR
- Andrew Paverd
Microsoft Research – Cambridge, GB
- Hammond Pearce
UNSW – Sydney, AU
- Jonathan Petit
Qualcomm – Boxborough, US
- Fabio Pierazzi
University College London, GB
- Liao Qianying
KU Leuven, BE
- Mathilde Aliénor Raynal
EPFL – Lausanne, CH
- Konrad Rieck
BIFOLD – Berlin, DE & TU Berlin, DE
- Vera Rimmer
KU Leuven, BE
- Advije Rizvani
Universität Liechtenstein – Vaduz, LI
- Ali Satvaty
University of Groningen, NL
- Lea Schönherr
CISPA – Saarbrücken, DE
- Ilia Shumailov
AI Sequrity Company – London, GB
- Tim Van hamme
KU Leuven, BE
- Eric Wong
University of Pennsylvania – Philadelphia, US
- Christian Wressnegger
Karlsruher Institut für Technologie, DE