Mining Inter-Document Argument Structures in Scientific Papers for an Argument Web

Florian Ruosch
University of Zurich, Switzerland

Cristina Sarasua
University of Zurich, Switzerland

Abraham Bernstein
University of Zurich, Switzerland

Abstract

In Argument Mining, predicting argumentative relations between texts (or spans) remains one of the most challenging aspects, even more so in the cross-document setting. This paper makes three key contributions to advance research in this domain. We first extend an existing dataset, the Sci-Arg corpus, by annotating it with explicit inter-document argumentative relations, thereby allowing arguments to be distributed over several documents forming an Argument Web; these new annotations are published using Semantic Web technologies (RDF, OWL). Second, we explore and evaluate three automated approaches for predicting these inter-document argumentative relations, establishing critical baselines on the new dataset. We find that a simple classifier based on discourse indicators with access to context outperforms neural methods. Third, we conduct a comparative analysis of these approaches for both intra- and inter-document settings, identifying statistically significant differences in results that indicate the necessity of distinguishing between these two scenarios. Our findings highlight significant challenges in this complex domain and open crucial avenues for future research on the Argument Web of Science, particularly for those interested in leveraging Semantic Web technologies and knowledge graphs to understand scholarly discourse. With this, we provide the first stepping stones in the form of a benchmark dataset, three baseline methods, and an initial analysis for a systematic exploration of this field relevant to the Web of Data and Science.

2012 ACM Subject Classification Computing methodologies → Information extraction; Computing methodologies → Language resources; Computing methodologies → Semantic networks; Information systems → Graph-based database models

Keywords and phrases Argument Mining, Large Language Models, Knowledge Graphs, Link Prediction

Digital Object Identifier 10.4230/TGDK.3.3.4

Category Research

Supplementary Material

Software (System): `https://gitlab.ifi.uzh.ch/DDIS-Public/midas/-/tree/main/bam`
archived at `swh:1:dir:b09e3e98962964553b846eda4014a48333f989b1`

Dataset: `https://gitlab.ifi.uzh.ch/DDIS-Public/midas/-/tree/main/data`
archived at `swh:1:dir:f8a081776dcaea3fd18c44d3dd0e8b9b4cd83800`

Software (Evaluation): `https://gitlab.ifi.uzh.ch/DDIS-Public/midas/-/tree/main/evaluation`
archived at `swh:1:dir:c808e145b2558a7353eb46ff0c34aede83ec2aac`

Funding This research was partially funded by the Swiss National Science Foundation (SNSF) through projects "CrowdAlytics" (contract 184994) and "Digital Deliberative Democracy" (contract 205975).

Acknowledgements The authors would like to thank the people at ARG-tech, particularly Chris Reed, for their helpful insights and the anonymous reviewers for their constructive feedback.

Received 2025-03-31 **Accepted** 2025-11-11 **Published** 2025-12-10

1 Introduction

The Argument Web [39] was postulated more than 15 years ago. Specified as a knowledge graph describing arguments in scientific publications, the Argument Web may serve several purposes for downstream tasks, such as fact-checking and misinformation detection [9] or text summarization [37]. As such, the Argument Web of Science [45] could be part of the Web of Data modeling relevant knowledge about the scientific process and its results. While there has been significant progress in the argument mining field, we still lack the necessary tools to generate the complete graph of linked arguments, as work on cross-document argumentative structure prediction has been limited, evidenced by a scarcity of corpora with such annotations [3]. Specifically, related work has largely focused on annotating intra-document content, neglecting the intricate and highly relevant relationships between scientific publications.

This paper addresses this crucial gap, primarily by introducing a novel dataset that comprehensively annotates inter-document argumentative relations in scientific publications. To construct this dataset, we extend the state-of-the-art Sci-Arg corpus [22], which describes arguments present in 40 computer graphics scientific publications. We identify and explicitly annotate inter-document argumentative relations, where one argument component from a paper attacks or supports a claim in another paper. Unlike intra-document relations, which describe argumentative links within a single paper, inter-document relations capture how argumentative components in one scientific publication explicitly support or attack those in another. These cross-document connections are crucial for understanding the broader argumentative discourse across the scientific literature and are inferred from citation data.

As a result, we produce a dataset that adds around 800 papers to the existing 40 annotated in Sci-Arg. To represent these links, we use the simplified *claim/premise model* [54] that has found widespread use [27]. These components are connected with relations such as *attack* or *support* to form an Argument Web. We also add and distinguish the notion of inter- (i.e., across document boundaries) and intra-document (i.e., within a single document) relationships. This dataset can serve as the stepping stone to larger datasets necessary for advancing the state of the art.

Furthermore, to demonstrate the utility and establish initial benchmarks for the new dataset, we investigate different methods to generate argumentative links: first, a rule-based model using the presence of discourse indicators to predict argumentative relations; second, a state-of-the-art argument miner [31] trained on the dataset; and third, Mistral [18], a pre-trained Large Language Model, used in two different ways: zero-shot classification of relations and few-shot with similar examples [28].

Then, we *compare the results of the methods for two different settings*: intra- and inter-document argumentative relation prediction. The former is to identify argumentative structures within a single document, while the latter also considers relations across document boundaries. To this end, we split the dataset into two disjoint subsets of intra- and inter-document relations. This allows for insights into how well the methods generalize across these two scenarios.

In this paper, we, hence, present three main contributions.

1. A novel, extended dataset of 836 papers based on the Sci-Arg corpus [22]. In addition to the original annotated arguments, we explicitly include and annotate inter-document relations at a document level. We publish these annotations using Semantic Web technologies (RDF, OWL), given their foundational role in the Argument Web vision.[1]
2. An evaluation of three diverse approaches for predicting inter-document argumentative relations. We find that a simple classifier based on the presence of discourse indicators outperforms modern neural methods for the given dataset. These results serve as a critical baseline and reference for future evaluations on the newly constructed dataset.

[1] An endpoint for SPARQL queries is available at `https://sciarg.ifi.uzh.ch/#/dataset/Sci-Arg/query`

3. A comparative analysis of these approaches in both inter- and intra-document settings. We identify notable differences in accuracy across systems, indicating varying performance capabilities depending on the relation scope (within vs. across documents). This analysis can serve as a template for future comparisons.

As such, this paper can act as a stepping stone for future studies in inter-document argument mining. Furthermore, our resulting annotations serve as a rich source of detailed, structured knowledge that can extend other Semantic Web data sources in the scientific publications domain. By publishing them in RDF, we facilitate a seamless integration with other Semantic Web sources.

The remainder of this paper is structured as follows. Section 2 presents the related work, and Section 3 introduces our methodology. In the ensuing Section 4, we describe our experiments and their results. Section 5 outlines possible use cases and applications. Then, Section 6 discusses limitations and future work. Finally, we conclude with Section 7.

2 Related Work

In this section, we present relevant related works. First, we cover existing argument mining corpora in the scientific domain. Then, we describe works relating to inter-document argumentative relations. Finally, we explore approaches to using the Semantic Web to support the scientific process.

2.1 Argument Mining Corpora in the Scientific Domain

Argument Mining for scientific papers has its roots in Argumentative Zoning [49]: classifying the relevant sentences from scientific articles into one of four categories (*Claim*, *Method*, *Result*, *Conclusion*). To the best of our knowledge, Sci-Arg [22] is the only dataset for Argument Mining in the scientific domain composed of fully annotated papers in English. This is unsurprising, given that annotating gold standard data is a resource-demanding task. The challenge is even exacerbated in the case of scholarly texts because they demand considerable domain expertise of the annotators [2].

For one of the few datasets [1] available, the authors annotated 60 abstracts based on a subset of the SciDTB corpus [57]. They add an argumentation layer containing 352 argumentative components, which are connected by 292 argumentative relations, resulting in partly annotated papers from computational linguistics. Since abstracts generally do not contain references to other documents, they cannot be used for inter-document relation prediction.

Another dataset consisting of 24 papers is presented in [20]. The articles from the domain of educational research have their introduction and discussion sections argument-annotated, notably including relations. However, this corpus has the reservation that it is comprised of publications in German, for which Argument Mining approaches are very limited. Instead, our work focuses on English data as it allows us to aim for a broader audience.

2.2 Inter-Document Argumentative Relations

While not intended to be purely argumentative, the Citation Typing Ontology (CiTO) [46] allows for the classification of the nature of in-text citations as factual or rhetorical relationships. The latter has more fine-grained types such as *supports* for positive and *refutes* for negative, which directly correspond to their argumentative equivalents. Since CiTO assigns types to

citations, these relationships inherently involve more than one document, making them inter-document. However, [46] presents no method for any (automated) annotations but solely an ontology. Furthermore, it also lacks a way to annotate intra-document relations since the entirety of CiTO targets citations.

The authors of [11] follow the relation-based Argument Mining paradigm [10]. Its goal is to predict the type of relations between text spans, such as attack, support, or neither. Using LSTMs [17] and GloVe embeddings [36], their approach is explicitly capable of classifying links between "any two texts," which also implies inter-document settings. The results indicate that their neural network methods improve over traditional classifiers.

The notion of "intertextual correspondence" [53] introduces another idea to connect annotated corpora (i.e., their documents) by exploiting relations, for example, of topical or temporal nature. This can also lead to multi-modal datasets for Argument Mining purposes. However, such links need to be identified manually first. The effectiveness of the approach is demonstrated by fusing a corpus of the debates in the US election of 2016 with commentary and reactions on the online platform Reddit. Such techniques could also be applied to the Sci-Arg dataset with other media or documents, since this inherently would result in inter-document relations.

In [35], the authors investigate the impact of content and context on analyzing argumentative relations. They find that systems that focus too much on the text of the argument components for relation prediction may easily be deceived and make wrong predictions. For example, they may rely on discourse indicators contained in the two text fragments to be analyzed, even though they are not adjacent, and, thus, the discourse indicator does not apply. Therefore, the importance of the context of argumentative units is asserted [35]: the position in the text and the surroundings (pre- and succeeding tokens). They show that systems only relying on the context instead of the content (i.e., component text span) may improve accuracy. Finally, the authors argue that systems dissecting content and context should also be more adequate for handling inter-document relation prediction.

2.3 Semantic Web and the Scientific Process

This work builds upon the vision of the Argument Web of Science [45]. While the notion of a scholarly knowledge graph is not novel, building one out of arguments extracted from scientific documents is. Given the prevalent role of argumentation in scientific communication [55], arguments are a suitable vehicle to represent scholarly information. Hence, we follow the *Argument Web* [39], which is based on Semantic Web knowledge representation technologies [8, 13].

The most prominent approach to a Knowledge Graph for Science has been described in [4]. Moving from document-centrism to a knowledge-based perspective would allow for a systematic organization of scholarly information. These efforts culminate in the Open Research Knowledge Graph (ORKG) [48, 5]: a representation that describes scientific papers in a structured way, facilitating question answering [33], paper comparisons, and visualizations. ORKG distinguishes itself from other large-scale efforts to represent scientific paper content, such as Microsoft Academic Graph (MAG) [56] (which is deprecated and succeeded by OpenAlex [38]), primarily in its granularity and explicit focus on structured, semantic content within papers. While MAG provided a vast bibliographic graph of papers, authors, and citations, ORKG aims to extract and structure the specific findings, methods, and results of scientific contributions, making the scientific content itself machine-readable and comparable. This deep semantic representation supports fine-grained analyses and direct comparisons of research outputs, unlike the primarily metadata-level focus of general bibliographic databases. However, ORKG uses a complex science ontology to represent the knowledge. While we share the goal of representing scientific information as a knowledge graph, our focus is different: we model arguments by means of their components (claims, premises) and two relationships connecting them (attacks, supports).

The idea behind Research Objects [6] contributes to Semantic Web efforts by defining a framework to preserve scientific workflows. It extends traditional workflow ontologies with metadata, annotations, provenance traces, and execution environments to improve reproducibility, reusability, and long-term preservation. A suite of ontologies provides structured descriptions of workflows and their evolution and aligns with other Semantic Web initiatives to structure scientific knowledge for better discovery and reuse. However, the focus is on Research Objects and their workflows, while this paper uses arguments as the core element of scientific knowledge representation.

Nanopublications [15, 21] aim to represent atomic units of knowledge with structured metadata and provenance information. This is to ensure trust, reproducibility, and interoperability. Hence, Nanopublications contribute to the Semantic Web's support for the scientific process by providing a provenance-centric format for publishing scholarly assertions, enhancing machine-readable scientific communication. However, Nanopublications focus on self-contained statements with metadata such as provenance and publication information. Our approach, in contrast, models arguments explicitly by capturing claims along with their attacking or supporting premises. This makes the logical relationship between statements a central feature.

In the context of knowledge representation and the Semantic Web, the Provenance Ontology (PROV-O) [26] primarily focuses on describing the origin, history, and derivation of entities and activities. While argument relations certainly possess provenance aspects (e.g., who asserted a claim, when, and based on what evidence), PROV-O's scope is distinct from the explicit modeling of types of argumentative connections (e.g., "supports", "attacks") between content units across documents. This is where the necessity of an Argument Web arises. Unlike general knowledge graphs that might link papers by co-authorship or topics, the Argument Web specifically aims to map the persuasive and confrontational dynamics across research. It addresses the limitations of single-document or citation-based views, particularly by capturing the complex inter-document support and attack relations that define scientific discourse. Our framework could be extended to integrate PROV-O for richer provenance tracking of the argumentative links themselves, but its core purpose lies in explicitly mapping the argumentative structure across documents, a functionality that goes beyond the scope of a pure provenance ontology. Our contribution lies in establishing the argumentative structure itself, which is a specialized form of inter-resource relationship not directly captured by general provenance models.

Lastly, SciHyp [52] introduces a fine-grained dataset representing hypotheses explicitly mentioned in scientific publications. It captures and structures scientific hypotheses from 479 computer science papers, categorized into relation-finding and comparative hypotheses. The dataset was created using a hybrid human-AI pipeline, combining experts, LLMs, and crowd refinements. Furthermore, SciHyp also extends existing hypothesis ontologies to better model their components. Through extensive evaluation, the authors demonstrate that LLMs can assist in hypothesis detection and component extraction, while also showing that human intervention remains crucial for precision. While SciHyp focuses on hypothesis-driven research, this paper aims to enhance the scientific process with a data-centered approach based on scientific arguments.

3 Methodology

This section explains the methodology behind our contributions. First, we lay the theoretical foundations of argumentative relations. Second, we describe the process of creating the new corpus by extending the Sci-Arg dataset [22] and list threats to the validity of this process and how we mitigated them. Finally, we lay out the details of the three approaches involved in generating the baseline for the new dataset, as well as how we evaluate them.

Table 1 Three examples of annotated relations taken from the Sci-Arg dataset [22].

Example	Component A	Relation	Component B
Intra-Doc	they tend to appear overly smooth and at times robotic	—attacks→	these methods do satisfy physical laws
Inter-Doc	Chadwick et al. 1989	—supports→	are computationally more expensive
Negative	its ease of implementation	—none→	SSD is a 3D transformation

3.1 Argumentative Relation Prediction

Predicting relations between argument components is the most complex and challenging part of the Argument Mining pipeline [27, 3]. The goal is to identify related pairs of argumentative components and to classify the nature of this link.

We follow Opitz and Frank [35] and define argumentative relation prediction to be an irreflexive function as follows:

$$f : C \times C \rightarrow R \quad \text{where } c_1 \neq c_2 \text{ for } (c_1, c_2) \in C \times C \tag{1}$$

C is the set of argumentative components. Each component $c \in C$ is associated with a specific document $doc(c)$ from which it originates. Components can be either a *text span* from within $doc(c)$ or a r*eference component* to an external document. For a reference component c_{ref}, $doc(c_{ref})$ refers to the document containing the citation. The range R of the function refers to the directed argumentative relation between the components. In this work, R is the set {attack, support, no-relation}, where "attacks" indicates a clash, "supports" indicates backing, and "no relation" signifies the absence of an argumentative link between the given components. Applying this function creates (typed) directed edges between argument component pairs, thus allowing us to conceive an argument graph. Table 1 provides examples of component pairs with the relation that connects them.

There are multiple ways to represent an argument, which influences the sets of C and R. In the context of this paper, we consider the following representation of an argument: the *claim/premise model* [54] with *attack* and *support* relations. In the context of this paper, the argumentative *roles* of components within C are further categorized as *claim* or *premise* (also called data, evidence, or reason in the literature [27]). A claim is a proposition about a point the arguer tries to make, and a premise is a statement about the validity of a claim. In the context of $f(\cdot)$, c_1 is typically the premise and c_2 the claim, especially for the support relation. However, there is no formal restriction on the type of components, as a claim may also attack another claim. This follows the conventions set by BAM, our Benchmark for Argument Mining [44], and is grounded in the fact that many other (more detailed) argument representations can be simplified to this model and, thus, allows for a unified evaluation (even if this incurs some information loss).

Building upon this general definition of argumentative relation prediction, we further categorize relations based on the origin and nature of their components, distinguishing between intra-document and inter-document argumentative relations. We use the following definitions:

▶ **Definition 1.** *An argumentative relation $f(c_1, c_2) \in R$ is **intra-document** if both components $c_1, c_2 \in C$ originate from the same document ($doc(c_1) = doc(c_2)$) and neither c_1 nor c_2 is a component representing a reference to an external document.*

▶ **Definition 2.** *An argumentative relation $f(c_1, c_2) \in R$ is **inter-document** if the two components $c_1, c_2 \in C$ do not originate from the same document ($doc(c_1) \neq doc(c_2)$) **OR** at least one of the components is a component representing a reference to an external document.*

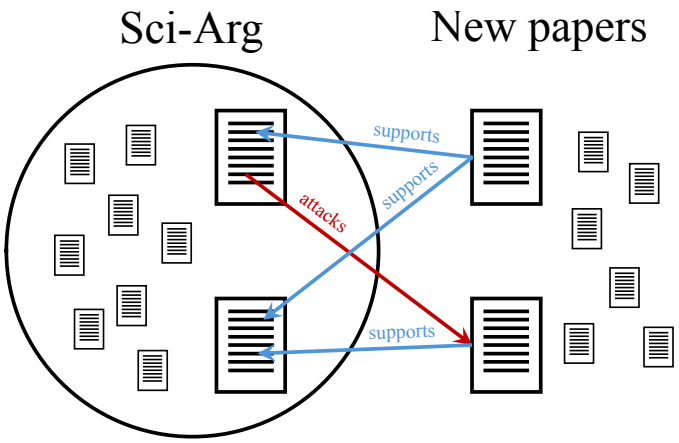

Figure 1 A visualization of the extension of the Sci-Arg corpus [22] with inter-document relations represented by the directed edges.

With the concept described in Definition 2, we can build an Argument Web as shown in Figure 1 since arguments can now be linked across multiple documents. Furthermore, using these definitions, we can form two (disjoint) subsets of relations for any dataset. It is important to clarify that this disjointness applies to the classification of argumentative relations based on the defined criteria of component origin and type (text span vs. reference to another document). This does not prevent the possibility of shared or overlapping content existing between documents that are involved in inter-document argumentative relations. Indeed, such shared content often provides the very basis for one paper to support or attack another. Our definitions categorize the type of argumentative relation established between components, rather than asserting that the underlying documents themselves are semantically or topically distinct. This ensures a clear distinction in how different types of argumentative interactions, within a document versus across documents, are formally represented and evaluated. Annotation examples from the Sci-Arg dataset [22] can be seen in Table 1 for the relation types to predict, as well as the two settings we distinguish.

3.2 Corpus Creation

We based the extended dataset on the freely available Sci-Arg corpus [22], which consists of 40 computer graphics papers fully argument annotated, including components and relations. Sci-Arg was annotated by one expert and three non-experts in five iterations. The reported Inter Annotator Agreement (IAA) in terms of the F_1-measure was reported with two criteria. For the strict version, span and type have to match exactly for components, and for relations, the direction also has to be correct. In the relaxed version, components only have to correspond in type and match half the span. This results in a higher IAA for the relaxed criteria: 72% for the components and 49% for relations. Meanwhile, the strictly measured IAA is 60% for the components and 35% for relations. The new annotations and the code for the automated processes are available in the online repository.[2]

Figure 1 visualizes the concept behind the extension of the dataset, which is represented by the circle, adding new documents outside of it. From the Sci-Arg annotations, we identified relations that involved components, which are references to other documents. That is, we determined triples consisting of an argumentative unit annotated in the dataset, a relation (type), and an

[2] `https://gitlab.ifi.uzh.ch/DDIS-Public/midas`

external paper acting as the respective endpoint. Crucially, for these inter-document relations, our annotation captures a document-level relationship, indicating that the referencing document supports or attacks the cited document as a whole, rather than specific claims or spans within it. Hence, we did not identify or annotate specific argumentative spans within the target (cited) papers. Instead, the argumentative relation type (supports or attacks) for these inter-document links was directly inherited from the original intra-document annotation in Sci-Arg. Specifically, if an argumentative component within one of the original 40 Sci-Arg papers was annotated with a certain relation (e.g., "supports") to another component, and that second component was identified as a citation to an external document, then that existing "supports" label was applied as the relation between the citing component and the cited document as a whole. Thus, the labels for inter-document relations originate solely from the high-quality, human-derived annotations within the original Sci-Arg corpus, ensuring their trustworthiness without requiring fine-grained annotation of the newly added documents. This process yielded 796 unique external document references from the initial 40 Sci-Arg papers that were part of annotated argumentative components. These references formed the basis for expanding our network of inter-document relations.

We assumed that if an in-text citation is annotated as an argumentative component, the original paper the citation is in reference to is meant to be used as the component. Therefore, we extracted all these components that contain a citation based on pattern matching (e.g., "Smith et al., 2000" or "[1]").[3] We then manually verified that all the identified components were citations in their context by looking through the 40 documents in the Sci-Arg corpus. This involved confirming the presence of standard citation formats (e.g., author-year mentions or numerical indices) that directly corresponded to entries in the papers' bibliographies, ensuring the component's accurate role as a citation to an external source.

Next, we matched the citations to their reference in the bibliographies in the papers. This was automated wherever possible. For the cases of ambiguity, occurring in approximately 10% of the instances, we determined the references by hand. These typically arose from multiple bibliography entries matching a single citation pattern, from unclear author-year combinations, or special characters not handled correctly during the content extraction. The issues were resolved by careful cross-referencing with the bibliography, the citing document's content, and, when necessary, external academic search engines, following a predefined set of internal guidelines to ensure consistency.[4] This way, we could assign the new documents to the annotated argument components.

Then, we resolved the references to their Digital Object Identifiers (DOI) using Crossref's API.[5] These API calls were successful for approximately 90% of the identified references. For the remaining cases without a direct DOI, we followed the guidelines detailed in the online repository to derive a DOI or, when none was available, another unique, persistent identifier (e.g., arXiv ID, official publisher URL) that ensured the consistent and unambiguous identification of the target paper. With these, we could now obtain the content of papers as PDFs, if accessible, from which we extracted information about their content and structure using Papermage [29]. We successfully obtained PDF content for 649 of the 796 new papers (approximately 81.5% of the newly added documents). From these PDFs, Papermage extracted key structural elements such as paragraphs, sentences, and headings, making their content programmatically accessible for potential future use, even though fine-grained annotation of these new documents was not within the scope of this initial extension.

[3] The full list of patterns is available in Appendix A.
[4] `https://gitlab.ifi.uzh.ch/ruosch/wp3/-/blob/main/data/guidelines.md`
[5] `https://api.crossref.org`

It is crucial to clarify that for the current scope of this dataset extension, the detailed full-text content extracted by Papermage from these additional 649 papers was not directly used to determine the argumentative relation types (supports/attacks) between documents. These relation types were derived by performing lookups in the original, human-annotated Sci-Arg data, where the citation acted as an argumentative component. Therefore, the content acquisition and extraction primarily served to identify and confirm the existence and structure of the cited documents and to build a resource for future, more granular analyses. While this approach might be considered "shallow retrieval" in terms of not using the full text for the initial relation labeling, it was a deliberate and necessary choice for the initial, broad-scale expansion of the Argument Web. It allowed us to efficiently establish a foundational network of document-level argumentative links. The availability of this extracted full text for a large portion of the new documents will be crucial for future work aimed at developing more sophisticated, content-based methods for fine-grained inter-document argument mining and automatic relation prediction, moving beyond simple citation-based links.

To finalize the new annotations for the intra- and inter-document argumentative relations, we needed to complete the triples with their predicate (relation type) and assign them to either of the two disjoint subsets. For this, we performed lookups in the original Sci-Arg data using the two component identifiers to get the relationship label connecting the two. Then, we added these triples to their corresponding subset based on Definitions 1 and 2: if both components were from the same document, they were classified as *intra*, and *inter*, otherwise.

As a result, we had a new dataset with two additional aspects. First, we added the distinctive notion of intra- and inter-document argumentative relations, producing annotations of these two disjoint subsets. This allows us to evaluate relation prediction approaches in the two settings independently, potentially leading to more in-depth analyses of how different methods work for them. Second, and more importantly, we augmented the original 40 papers by including an additional 796 new papers, extending the dataset to 836 total papers.

It is crucial to clarify that this expansion primarily involved identifying external papers cited by the original 40 and determining document-level relationships to them. Thus, while this represents a twentyfold increase in the number of documents, the detailed, claim-level annotations present in the original 40 papers were not replicated for the additional 796 documents. This targeted approach required substantially less effort per additional document, allowing us to broaden the scope of inter-document relations efficiently. Even though we do not currently make use of the full text of the additional papers for detailed internal annotations, their inclusion is vital. These 796 documents serve as essential nodes within the expanded inter-document argumentative graph, providing a richer, more representative context for studying cross-document argumentation. They enable the analysis and prediction of how arguments originating from the original 40 papers relate to a much wider body of scientific literature, making them valuable for understanding the broader Argument Web.

Additional statistics about the original and the extended dataset can be found in Table 2 (the class distributions are shown in Appendix F). The number of total relations remains unchanged in the extension because these 1996 relations represent all argumentative links identified within and originating from the original 40 Sci-Arg papers. Our extension primarily involved: 1) identifying which of these existing relations were inter-document (i.e., involved a reference to an external paper), and 2) adding the external documents themselves as nodes to the Argument Web to serve as endpoints for these newly categorized inter-document relations. We did not perform new fine-grained argumentative annotations within the 796 newly added papers. Instead, the total number of relations reflects the comprehensive set of argumentative links present in the original Sci-Arg corpus, now re-categorized and contextualized within a larger network of documents.

Table 2 The statistics of the original Sci-Arg [22] and its extension.

Number of	Sci-Arg	Extension
Total Relations	1996	1996
Intra-Doc Relations	n/a	1428
Inter-Doc Relations	n/a	568
Doc-Level Relations	n/a	1109
Documents	40	836

We determined 568 relations that go across document boundaries and 1428 within documents. If we only allow one relation per direction and type for each pair by consolidating relations where one paper is used to support or attack at multiple locations in another, we reduce the annotations to the document level. This way, we maintain a graph with 836 nodes and 1109 edges (i.e., the number of doc-level relations).

We publish these document-level annotations as Resource Description Framework (RDF) data since Semantic Web technologies provide the means to represent machine-readable data that is easy to integrate with other data despite heterogeneity.[1] This is particularly useful in this context, as there exists a variety of argument representation models with different levels of granularity [27]. To model the data, we used the CrowdAlytics ontology, which extends other ontologies and integrates ontological components to describe scientific hypotheses and arguments present in scientific publications [51]. Even though we produced an Argument Web, the inter-document relations are indeed only on the document level.

This document-level granularity for inter-document relations was a deliberate design choice, allowing us to scalably extend the dataset's scope to hundreds of external documents. Unlike the fine-grained, claim-level annotations within documents, identifying specific span-level argumentative connections across disparate documents proved prohibitively complex and resource-intensive for manual annotation, so we only resolved the references to the documents and did not determine the more fine-grained details of which exact proposition the reference is to. This is visualized in Figure 1 by the arrows representing the relations only pointing to the documents and lacking detail outside the circle representing the original corpus. The implication of this design is that our dataset provides a foundational macro-level view of argumentative connections between papers, capturing their overall support or attack roles. While this work focuses on these document-level interactions, future research can leverage this foundation to develop automated methods for identifying more granular, span-level inter-document argumentative links.

For consistency, we thus make this difference in the relations evident by altering their labels slightly. *Supports* becomes *is used as support in*, and *attacks* turns into *is used as attack in*. This emphasizes the difference between argument component pairs where both are well-specified propositions as opposed to those consisting of one text span and one reference to a different document. Semantic Web technologies facilitate the alignment of the two different granularities of annotations in the same dataset.

Threats to Validity

Several threats to validity impact our corpus creation. First, manual processes were crucial for verifying identified citation components within their context and for resolving ambiguous references in bibliographies. However, such manual intervention inherently introduces the risk of human error, misinterpretation, or inconsistency, which could lead to incorrect associations between argumentative components and external papers, thereby compromising the accuracy of

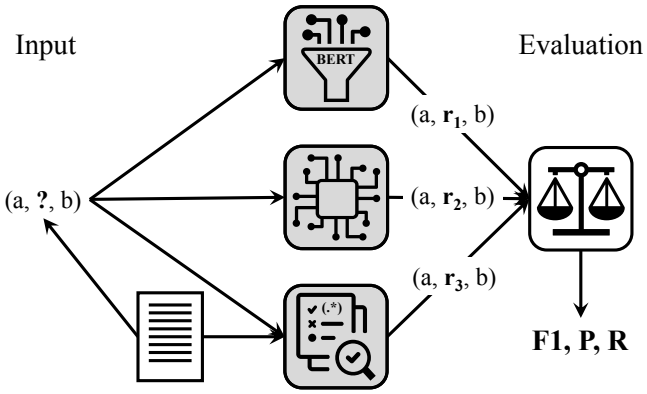

Figure 2 Overview of the three approaches for predicting the argumentative relation for two given argumentative units.

the inter-document relations. To address this, we systematically went through all 40 documents of the Sci-Arg corpus for manual verification and resolved ambiguities by hand, aiming for precision in these critical steps. For consistency and reproducibility, these manual steps are detailed in the internal annotation guidelines.[4]

Second, our reliance on pattern matching for extracting components containing citations, while efficient, carries the threat of incompleteness if non-standard citation formats were missed, or inaccuracy if text was incorrectly identified as a citation, potentially leading to an incomplete or flawed set of inter-document relations. We sought to mitigate this by providing a comprehensive list of patterns in the Appendix, allowing for transparency and future refinement.

Third, challenges encountered during the Digital Object Identifier (DOI) resolution process using Crossref's API, and the subsequent efforts to manually derive DOIs or obtain paper content, meant that not all referenced papers were accessible or fully processable. This limitation means the dataset may not encompass all intended inter-document relations, or the full text content crucial for future work leveraging Papermage was not uniformly obtained across all added documents. We addressed this by following specific guidelines detailed in our online repository for deriving DOIs or other unique identifiers when automated methods failed. Crucially, our underlying assumption that an in-text citation annotated as an argumentative component refers to the entire original paper simplifies potentially more nuanced argumentative connections. This design choice directly impacts the granularity of our inter-document relations, which are modeled only at the document level. Consequently, while we clarified this distinction in the annotations by altering relation labels (e.g., "supports" to "is used as support in"), the dataset does not capture the precise argumentative unit (e.g., a specific proposition) within the external document that is being supported or attacked, thereby limiting the depth of fine-grained argumentative analysis possible across documents. We explicitly acknowledged this limitation by making the difference in relations evident through altered labels and noting that Semantic Web technologies facilitate the alignment of these different granularities within the same dataset.

3.3 Automated Relation Prediction

This subsection describes the three approaches used to create the baseline for the new dataset. Furthermore, we also explain the evaluation methods that we apply. An overview is shown in Figure 2: the inputs on the left-hand side, the three approaches in the middle, and the evaluation on the right. The first two approaches only receive the two components to predict the relation, while the bottom one also has access to the full text. They are all evaluated by the same methodology to produce comparable results in the F1-score.

3.3.1 Rule-Based Approach

The first approach for predicting argumentative relations is rule-based on the presence of discourse indicators. It is depicted as the bottom of the three gray center squares in Figure 2. Discourse indicators have already been successfully employed as an Argument Mining technique [47] and have been shown to be a dependable method for argument relation prediction [25]. To keep this method simple, we take the 24 expressions from [24, p. 128] (cf. Appendix B) as the set of discourse indicators.

In order to predict the relation for a given pair of argumentative components (text spans), we first check if they occur in the same sentence in the original text. For this, we have to give the system access to the document and, thus, the context of the argumentative units, since the discourse indicator may occur outside of their boundaries. If we fail one of the two checks, we assume the pair to be not related and predict *no relation*. Otherwise, we construct a triple of the two argument components and derive the relation type from the nature of the discourse indicator. Some signify *support* (e.g., "because" or "since") and others *attack* (e.g., "but" or "despite").

It is important to note that the annotation guidelines for Sci-Arg [23] only mention discourse indicators as cues for annotating components but not for relations. Hence, no correlation is expected ex-ante for the relation prediction.

3.3.2 Argument Miner

As the argument miner, we use the system described and implemented in [31]. Depicted as the top of the three gray squares in Figure 2, it combines pre-trained transformers (variants of BERT [12]) with recurrent architectures (GRU, CRF, LSTM) for mining arguments and their structures from natural language texts.

The prediction of argumentative relations between components (text spans) is modeled as a sequence classification problem. For the pairs of components, one of the following three classes is to be chosen: *Support*, *Attack*, *NoRelation*. Furthermore, one component can be related to multiple other components, since each combination is classified independently. Given its top-performing results in [31], we opted for the uncased SciBERT [7] as the transformer to compute the embeddings. The pooled representation of the sequence to be classified is fed into a linear layer with a softmax that produces a distribution over the target classes. These results can be used to predict the argumentative relationship type between the two given components. The details regarding fine-tuning, exact parameters, and data splits can be found in Appendix C.

3.3.3 Large Language Model

We use a quantized version of Mistral-7B [18] as the large language model in two different ways, shown as the middle of the three gray squares representing the approaches in Figure 2. Our choice of Mistral-7B was driven by the constraint of running the model on a single GPU. This approach improves the reproducibility and generalizability of our work, as it makes the methodology accessible to researchers with more limited computational resources. It also significantly speeds up computation, which is essential for future large-scale experiments. While we also experimented with several other LLMs (variants of GPT [34] and Llama [14], among others) that fit this criterion, we ultimately selected to only include Mistral-7B as it consistently showed superior performance on our specific task.

The first is naive zero-shot classification, whereby the model is prompted with the task phrased as a multiple-choice problem. This involves predicting the argumentative relation from the set *attacks*, *supports*, or *none* between two argumentative components, given as text spans. Mistral is then asked to respond with a single word, i.e., the classification of the relation. The detailed prompts can be found in Appendix D, including the class distribution of the in-context examples.

Since argumentative relation prediction is a complex task [3], we also applied a more sophisticated few-shot learning method, utilizing in-context learning based on similar examples [28]. We use the same template as in the zero-shot classification, but enrich it with examples of triples formed by component pairs and their connecting relation. These examples are dynamically generated by identifying the five most semantically similar components in the training set by deriving their embeddings with Sentence-BERT [42] and applying cosine similarity. We select the five most semantically similar examples for in-context learning because they are hypothesized to provide the most relevant contextual cues and linguistic patterns for the LLM to learn from, facilitating better generalization. This approach aims to maximize the effectiveness of the limited in-context examples by ensuring their direct applicability to the target prediction, thereby improving the model's ability to discern complex argumentative relationships. We then retrieve the ground truth relations for these five semantically similar pairs to create the in-context examples. The final prompt is composed of the task description, followed by these curated examples, and finally, the components for which the relation is to be predicted. We hypothesize that providing the LLM with information about relations from semantically similar components will help it to predict the relation correctly.

Finally, for a fair comparison to the other two methods, we will evaluate the LLM approach with and without providing the context of the argumentative components. To this end, the LLM will receive the full sentences from which the components are taken in the prompt, alongside the argumentative component spans themselves. This allows the model to leverage a more complete understanding of the surrounding discourse, which is known to be a crucial factor in argumentative reasoning and provides a direct way to measure the impact of context on LLM performance for this task.

3.3.4 Evaluation

For the evaluation of the three approaches, we employ the routines implemented by BAM [44], our benchmark for Argument Mining. It is represented by the scale in the square on the right in Figure 2. It splits the evaluation into the four stages (*sentence classification*, *boundary detection*, *component identification*, and *relation prediction*) of the Argument Mining pipeline described in [27], from which we only use the last step.

To evaluate the argumentative relation prediction, BAM treats it as a binary classification (retrieved or missed) of triples *(subject, predicate, object)* and applies the F1-score [50]. While argumentative relations in our dataset consist of three classes ("attacks", "supports", and "no relation"), BAM's relation prediction evaluation frames the problem as identifying relevant/irrelevant and retrieved/missed triples. This means, for each potential (subject, object) pair, the system's task is to determine if a specific "attacks" or "supports" triple exists and, if so, to correctly identify it. The "no relation" case is implicitly handled as the absence of such a positively identified argumentative triple. The subject and object are represented by the identifiers of the corresponding argumentative component (i.e., text span) as given in the ground truth annotations of the Sci-Arg dataset. The predicate is the label of the argumentative relationship between the two. By comparing the ground truth of the test set and the predictions, we get a result for the relation prediction score represented by F1 between zero and one for each approach, where bigger signifies better. The code involved in the evaluation is available in the online repository.[2]

Table 3 The results of the experiments on the overall dataset. We report the F1-score as measured by BAM [44], the precision, and the recall. Also, we indicate if the method had access to the context.

Approach		Context	F1	Precision	Recall
Rule-Based		✓	0.437	0.554	0.372
Argument Miner	[31]		0.251	0.410	0.187
Zero-Shot LLM	[18]		0.113	0.061	0.802
Zero-Shot LLM	[18]	✓	0.119	0.065	0.850
Few-Shot LLM	[18]		0.109	0.141	0.095
Few-Shot LLM	[18]	✓	0.132	0.099	0.217

4 Results

This section discusses the experiments conducted to create the baseline and their results. We first look at the outcome of the approaches on the overall dataset. Then, we examine the differences when they are applied in the intra- and inter-document settings.

The code of the experiments and the results are available in the online repository.[2] All statistical test procedures and their results ($\alpha < 0.05$) can be found in Appendix E.

4.1 Overall Dataset

Table 3 shows the results for the approaches on the overall dataset. A class-based evaluation can be found in Appendix F. Furthermore, we show more detailed statistics on the distribution of the misclassified samples per relation type of each system in Appendix G.

With a relation prediction score of 0.437, the naive approach, which leverages only the presence of discourse indicators, clearly – and statistically significantly – outperforms the other more sophisticated neural methods. This may be surprising, but it can be put into perspective with the following three points. First, discourse indicators have been shown to work well for predicting argument relations [25]. Second, it was the only technique that had access to not only the content but also to the context of the argument components. This has been noted to contribute to accuracy positively [35]. Still, the only influence the access to the context had on the approach was that the discourse indicator could be contained in either the sentence holding the components or in the components themselves. Finally, even though the rule-based approach achieved the highest score, the result is nowhere near where it could be, as the 0.437 score clearly indicates significant room for improvement.

Scoring 0.251, the transformer-based Argument Mining system [31] came in second place and considerably ahead of the LLM approaches. This result is consistent with the outcome for TRABAM in the original showcase of BAM [44]. Even with the training on the dataset, it appears the system still fails to predict most of the relations for argument component pairs correctly.

The results of the LLM-based approaches are nuanced. Both the zero-shot and few-shot methods show a statistically significant improvement in their mean F1-scores when provided with context (p=0.008 and p=0.012, respectively). However, the effect size of this improvement is markedly different. For the zero-shot LLM, the improvement is negligible (d=-0.173), while for the few-shot LLM, the effect is medium (d=-0.618). This finding corrects our initial hypothesis, as it indicates that the information from semantically similar examples *is* effective at improving accuracy, but only when combined with the contextual information from the surrounding sentences.

When we examine the performance with context more closely, the few-shot LLM achieves the best F1-score among the LLM variants (0.132). However, this improvement comes with a trade-off: while its recall increases substantially (0.095 → 0.217), its precision decreases (0.141 → 0.099).

Table 4 The results of the experiments on the intra- and inter-document subsets. We report the F1-score as measured by BAM [44], the precision, and the recall.

Approach		Context	F1	Precision	Recall
Rule-Based	Intra	✓	0.430	0.514	0.380
	Inter		0.432	0.649	0.348
Argument Miner	Intra		0.238	0.360	0.184
	Inter		0.345	0.502	0.276
Zero-Shot LLM	Intra		0.083	0.044	0.718
	Inter		0.249	0.145	0.970
Zero-Shot LLM	Intra	✓	0.090	0.048	0.782
	Inter		0.252	0.146	0.987
Few-Shot LLM	Intra		0.091	0.131	0.075
	Inter		0.136	0.136	0.144
Few-Shot LLM	Intra	✓	0.113	0.085	0.182
	Inter		0.165	0.118	0.294

This behavior indicates that providing context and few-shot examples enables the model to identify more potential relations but at the cost of generating more false positives. This finding confirms that while LLMs are sensitive to contextual information, they still struggle to accurately and precisely discern the correct relations, particularly without extensive fine-tuning.

A key observation from our class-based results (cf. Appendix F) is the significant disparity in performance between "supports" and "attacks" relations, where models, particularly the LLM-based approach, consistently exhibit much higher F1-scores for "supports". This phenomenon can be attributed to a combination of factors. Firstly, the inherent class imbalance in our dataset, as detailed in Subsection 3.2 (Table 2), means "supports" relations are substantially more frequent than "attacks". This naturally biases models towards the majority class. Secondly, identifying "attacks" relations is often an intrinsically harder task in argument mining. They frequently involve more nuanced linguistic cues, require deeper semantic understanding of contradiction or refutation, and may manifest in more diverse textual patterns compared to expressions of support. The confusion matrices (Appendix G) further illuminate this, revealing instances where models tend to conflate "attacks" with "no relation" or even incorrectly classify them as "supports", highlighting the challenge in accurately discerning these critical dissenting links.

4.2 Intra- Versus Inter-Document Setting

The results for the intra- and inter-document settings are shown in Table 4, revealing that all systems achieve higher scores for the latter. We make pairwise comparisons to see whether these numbers are statistically significantly different for F1.

This is neither the case for the rule-based nor for the few-shot LLM approach. The two represented the two different ends of the scale of the results. The former is at the top (0.430 and 0.432), and the latter is at the bottom (0.091 and 0.136). However, they do not exhibit statistically significantly different scores for the argumentative relation prediction in the intra- and the inter-document settings.

Still, the results of the rule-based approach give insights. They indicate that discourse indicators in scientific papers are a simple but reliable signal for predicting argumentative relations in the two situations. This is shown in both settings by the higher precision than recall of the method, confirming the findings in [24].

Meanwhile, the transformer-based argument miner [31] achieves a distinctly higher score in the inter-document (0.345) than in the intra-document (0.238) setting. The difference has statistical significance, suggesting that it is better at predicting this type of argumentative relationship. The precision and recall values, with the former being higher than the latter, also indicate that the AM system is able to pick up on the cues in the content of the components. However, it fails to capture most of the argumentative relations.

For the zero-shot LLM approach, the statistically significant difference in the relation prediction scores is even larger: 0.083 and 0.249, respectively. The very high recall values stand out for both settings, indicating that it catches many of the argumentative relations to predict, but also produces a lot of garbage annotations, considering the very low precision. The analysis for the few-shot LLM approach is the same as that for the overall dataset. It is inadequate for predicting argumentative relations either way, with some of the lowest F1, precision, and recall scores across the board. Since these systems are based on black-boxes that are the neural network architectures, we can only speculate about the reasons for the disparities. For some LLMs, we have knowledge about the data involved in their pre-training, while others may be completely closed off. However, given that their training involved coherent and, most likely, academic text, we can assume that their capabilities involve recognizing argumentative components such as claims and premises but struggle to pick up on cues for the more complex task of argumentative relation prediction. More investigations in this direction are clearly necessary, also when taking the difficulties of effective prompting into account.

For both the argument miner and the LLM, the results may seem counterintuitive when comparing the relation prediction scores to those of the naive rule-based approach. The components in the inter-document setting tend to contain less semantic and syntactic information, as at least one of them is a reference to another document: a citation. In most cases, that does not give any details about the referenced document apart from possibly author names and the publication year. Neither of which carries significant information without more semantic context. Therefore, the only conclusion is that for the inter-document setting, the complementary component of the pair (i.e., not the citation) contains particularly useful information for predicting the relation. Surprisingly, this would also imply that the cues for the inter-document argumentative relation prediction from only one component can be leveraged more effectively than those from two components in the intra-document setting, resulting in higher scores for the former.

The LLMs show a varied response to the addition of context. The few-shot LLM with context shows no statistically significant difference between intra- and inter-document performance. In contrast, the zero-shot LLM with context exhibits a large and statistically significant difference between intra- and inter-document F1 scores. This indicates that while context helps both approaches, the few-shot learning method, when combined with context, is better able to adapt and apply its learned knowledge more consistently to both intra- and inter-document relationships, reducing the performance gap between the two, as was also the case without context.

5 Use Cases for an Argument Web

The Argument Web [40] and, hence, our extended dataset and the developed approaches for inter-document argumentative relation prediction, provide a foundational layer for a range of applications, particularly within the scientific domain. By explicitly modeling how scientific arguments interact across different papers, our work enables a deeper understanding and more efficient navigation of scholarly discourse.

Firstly, our approach can significantly enhance scholarly recommendation systems. Moving beyond traditional keyword matching or citation networks, our dataset allows for the development of systems that suggest papers based on their argumentative relationship to a user's current reading.

This means a researcher could be proactively recommended not only supporting evidence but also crucial counter-arguments or alternative perspectives, fostering a more critical and comprehensive literature review process. Such systems could also help researchers build a better understanding of a topic's argumentative landscape, rather than just its content.

Secondly, these insights are crucial for the automated construction of scientific knowledge graphs. By identifying and classifying explicit "supports" and "attacks" relations between documents, our methods facilitate the population of knowledge graphs with argumentative links. This enriches the semantic representation of scientific discourse, allowing for complex queries that trace the evolution of ideas, identify the provenance of specific claims, or map the full spectrum of evidence surrounding a hypothesis. Such structured argumentative knowledge can serve as a backbone for advanced AI applications in scientific discovery.

In the evolving landscape of LLMs and conversational AI, the explicit modeling of argumentative structures, as enabled by our dataset extension, gains particular significance. Our annotations provide a crucial resource for training LLMs to generate more accurate, reasoned, and evidence-backed responses. By understanding support and attack relations across documents, LLMs can help with:

1. Enhance Factuality: Ground their generated content in verifiable evidence by identifying supporting arguments from established literature, thereby combating hallucinations with a belief graph [19].
2. Improve Reasoning: Generate more coherent and logically structured arguments by mimicking observed patterns of claims, premises, and their interconnections.
3. Synthesize Debates: Effectively summarize complex scientific discussions by identifying the core arguments for and against specific theories or findings from multiple sources.
4. Provide Justifications: Equip chatbots and conversational agents with the ability to offer transparent justifications for their answers, referencing supporting evidence or acknowledging counter-arguments.
5. Detect Controversies and Bias: Recognize areas of scientific disagreement or identify potential biases in presented arguments by mapping where attacks are concentrated or support is lacking.

Ultimately, our work contributes foundational data structured by explicit inter-document relations, establishing a basis for future research. This data can empower LLMs to move beyond mere text generation toward more sophisticated, critically aware, and trustworthy engagement with scientific knowledge by providing access to the evidentiary chain of scientific claims.

Furthermore, the ability to identify cross-document argumentative relations can greatly facilitate automated literature review, survey generation, and summarization [37]. Researchers often spend considerable time synthesizing arguments and counter-arguments across a vast body of scientific literature. Our approach could help automate this process by providing an "argument map" for a given research question, highlighting key supporting evidence, summarizing different positions on controversial topics, or even generating preliminary argumentative outlines for review articles, thereby drastically reducing manual effort.

Beyond general scientific discourse, Argument Mining and the Argument Web approach hold significant promise for applications in healthcare [30]. Clinicians and researchers in medicine constantly grapple with vast amounts of evidence for and against various therapies, diagnostic methods, and treatment protocols. An Argument Web in the clinical space could provide structured, machine-readable evidence for decision-making, allowing users to trace the supporting and attacking arguments for specific medical recommendations or interventions across countless research papers, clinical trials, and guidelines. While the current dataset is from computer graphics, the methodology for extracting and linking inter-document arguments is potentially transferable, suggesting a powerful tool for evidence-based medicine, systematic reviews, and even supporting

clinical guideline development by explicitly mapping the underlying argumentative landscape of medical knowledge. We believe this area represents a particularly impactful future direction for the Argument Web.

Finally, by comprehensively mapping the argumentative landscape of scientific fields, our work provides a robust framework based on explicit inter-document support and attack relations for subsequent analysis aimed at quantitatively identifying research gaps, key disagreements, and emergent scientific controversies across a field. The explicit identification of "attacks" relations, particularly when concentrated on specific arguments or findings, can pinpoint areas of ongoing debate, unresolved issues, or even fundamental assumptions that lack robust support. Conversely, the absence of strong support relations for a new claim might indicate a research gap. This capability can guide future research directions by highlighting critical areas for further investigation or areas where a particular line of argument has been consistently refuted, enabling the scientific community to focus efforts more effectively.

6 Limitations and Future Work

The main limitation of this work is that all investigations and evaluations were conducted on a dataset that only represents a specific domain of scientific papers. Therefore, whether the insights generalize well or at all to other natural language texts remains to be evaluated. In addition, while we know the domain of the initial dataset is computer graphics [22], we do not have any information about the newly added papers. Since we did not find an automated way to get the topics for them, we leave it as future work to see if and how many new topics have been added to the dataset.

Another constraint was already pointed out in the description of the corpus creation: the inter-document relations are only annotated on the document level, and, therefore, there is a lack of detail. The new annotations are valuable information for evaluating methods for predicting argumentative relations. Still, *anchoring* the relations (i.e., identifying the exact location) in the new documents is vital. This goes hand in hand with bringing these annotations to the same level of detail as the Sci-Arg corpus. Furthermore, this would enable an iterative process by further extending this dataset with newly identified inter-document relations for the additional papers, and so forth.

Curating these annotations is very time-consuming, even more so for documents as specialized as scientific papers, since they require extensive domain knowledge. Therefore, the question of how human annotations can be reduced to facilitate the tasks for the annotators should be considered. This exploration is left as future work.

Furthermore, we imposed some limitations on the argument model in the evaluation. They stem from the way the evaluation is set up using BAM, because only in doing so could we produce comparable results for the different Argument Mining approaches. This also entailed a degree of information loss since we did not incorporate all available information from the original dataset. For example, we did not distinguish between *own* or *background* claims, and for relations, we left out the *semantically-same* and *parts-of-same*.

To enhance the robustness and generalizability of argumentative relation prediction, particularly for the challenging "attacks" class, future work should prioritize targeted strategies. Given the observed class imbalance and the inherent difficulty of the task, approaches such as cost-sensitive learning, advanced data augmentation for minority classes, or specialized prompt engineering for LLMs that emphasizes adversarial reasoning could prove beneficial. Furthermore, a detailed analysis of the misclassified instances, guided by the inter-class confusion matrices, will be instrumental. This fine-grained error analysis can help in identifying specific linguistic or structural

patterns that currently mislead models, giving rise to the development of more discriminative features or targeted training strategies to reduce confusion between "attacks" and other relation types. Improving the detection of "attacks" is crucial for tasks like identifying scientific controversies and understanding dissenting viewpoints.

Moreover, there are further approaches that could be applied to argumentative relation prediction that we did not include in our work. The explanations for this are twofold. First, we aim to provide a baseline on the dataset. Second, we wanted to use off-the-shelf methods, which also facilitate the reproducibility of our results. Our main contribution remains the new dataset that we generated and published as machine-readable data. Using it, we can improve the baseline by putting out a shared task or a challenge and getting more methods involved.

While the groundwork has been laid in [35] with their analysis of content and context for argument relations, a detailed study of what works and what does not remains pending. We can only surmise why the simple, rule-based approach outperforms the more sophisticated approaches so clearly. Examining the shortcomings of the transformer-based and the LLM methods should shed some light on the matter.

Another aspect that we left out in this work is investigating whether the contents of the documents alone can be used to predict the document-level argumentative relations. With the increasing context windows of LLMs [43], they might be capable of correctly identifying the relations between documents without relying on detailed component annotations. They could construct Argument Graphs from document contents alone. However, as made evident in the results of the LLM approach in our work, the relation prediction remains challenging for them. Thus, we hypothesize that this is no trivial task and leave it for future work.

To further facilitate the adoption and application of our work by the broader research community, future efforts can include the development of a dedicated use cases website, accompanied by comprehensive documentation and tutorials. This resource could aim to provide practical guidance and showcase various applications of the extended dataset and the developed argumentative relation prediction approaches.

7 Conclusions

This work addressed the challenging task of mining inter-document arguments in scientific papers. We focused on predicting argumentative relations, such as *attacks* and *supports*, between argument components (*claims*, *premises*). In this context, we extended an existing dataset [22] by explicitly annotating it with inter-document argumentative relations. Then, we explored three automated argumentative relation prediction approaches and evaluated them on the original dataset and the newly annotated inter-document relations.

This work aligns with the ultimate goal of constructing an Argument Web [39] of Science. Our contributions include the creation of a new dataset with explicitly annotated inter-document argumentative relations. We published it using Semantic Web technologies, extending its size from the original 40 papers to over 800. This endeavor hopes to contribute to advancing the dissemination of scholarly discourse.

Furthermore, our analysis of the baseline results indicates that a simple rule-based classifier leveraging the presence of discourse indicators outperforms neural methods for argumentative relation prediction. This finding emphasizes the effectiveness of exploiting the linguistic features of the context of components in predicting argumentative relations, which has previously also been noted in [35]. Furthermore, we observed statistically significant differences in accuracy between the intra- and inter-document settings for the evaluated approaches. This highlights the importance of distinguishing between these two.

In summary, our efforts mark a step forward in understanding and harnessing the complex web of arguments in scientific papers. By providing an extended, well-analyzed dataset for intra-document Argument Mining, we hope to bootstrap the effort for large-scale datasets that pave the way to the Argument Web of Science [45], which in turn could serve as a major pillar for the Web of Data with respect to science, the scientific process and discourse, and its results.

References

1 Pablo Accuosto and Horacio Saggion. Transferring knowledge from discourse to arguments: A case study with scientific abstracts. In Benno Stein and Henning Wachsmuth, editors, *Proceedings of the 6th Workshop on Argument Mining*, pages 41–51, Florence, Italy, August 2019. Association for Computational Linguistics. doi:10.18653/v1/W19-4505.

2 Titipat Achakulvisut, Chandra Bhagavatula, Daniel E. Acuna, and Konrad P. Körding. Claim extraction in biomedical publications using deep discourse model and transfer learning. *CoRR*, abs/1907.00962, 2019. arXiv:1907.00962.

3 Khalid Al Khatib, Tirthankar Ghosal, Yufang Hou, Anita de Waard, and Dayne Freitag. Argument mining for scholarly document processing: Taking stock and looking ahead. In Iz Beltagy, Arman Cohan, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Keith Hall, Drahomira Herrmannova, Petr Knoth, Kyle Lo, Philipp Mayr, Robert M. Patton, Michal Shmueli-Scheuer, Anita de Waard, Kuansan Wang, and Lucy Lu Wang, editors, *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 56–65, Online, June 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.sdp-1.7.

4 Sören Auer, Viktor Kovtun, Manuel Prinz, Anna Kasprzik, Markus Stocker, and Maria Esther Vidal. Towards a knowledge graph for science. In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, WIMS '18, New York, NY, USA, 2018. Association for Computing Machinery. doi:10.1145/3227609.3227689.

5 Sören Auer, Allard Oelen, Muhammad Haris, Markus Stocker, Jennifer D'Souza, Kheir Eddine Farfar, Lars Vogt, Manuel Prinz, Vitalis Wiens, and Mohamad Yaser Jaradeh. Improving access to scientific literature with knowledge graphs. *Bibliothek Forschung und Praxis*, 44(3):516–529, 2020. doi:10.1515/bfp-2020-2042.

6 Khalid Belhajjame, Jun Zhao, Daniel Garijo, Matthew Gamble, Kristina M. Hettne, Raúl Palma, Eleni Mina, Óscar Corcho, José Manuél Gómez-Pérez, Sean Bechhofer, Graham Klyne, and Carole A. Goble. Using a suite of ontologies for preserving workflow-centric research objects. *J. Web Semant.*, 32:16–42, 2015. doi:10.1016/J.WEBSEM.2015.01.003.

7 Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics. doi:10.18653/v1/D19-1371.

8 Tim Berners-Lee, James Hendler, Lassila, and Ora. The Semantic Web. *Scientific American*, 284(5): 34–43, 2001.

9 Elena Cabrio and Serena Villata. Five years of argument mining: a data-driven analysis. In Jérôme Lang, editor, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 5427–5433. ijcai.org, 2018. doi:10.24963/IJCAI.2018/766.

10 Lucas Carstens and Francesca Toni. Towards relation based argumentation mining. In Claire Cardie, editor, *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 29–34, Denver, CO, June 2015. Association for Computational Linguistics. doi:10.3115/v1/W15-0504.

11 Oana Cocarascu and Francesca Toni. Identifying attack and support argumentative relations using deep learning. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1374–1379, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi:10.18653/v1/D17-1144.

12 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1423.

13 John Domingue, Dieter Fensel, and James A. Hendler, editors. *Handbook of Semantic Web Technologies*. Springer, 2011. doi:10.1007/978-3-540-92913-0.

14 Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel

Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul

Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. doi:10.48550/arXiv.2407.21783.

15 Paul Groth, Andrew Gibson, and Jan Velterop. The anatomy of a nanopublication. *Inf. Serv. Use*, 30(1-2):51–56, 2010. doi:10.3233/ISU-2010-0613.

16 Steffen Herbold. Autorank: A python package for automated ranking of classifiers. *J. Open Source Softw.*, 5(48):2173, 2020. doi:10.21105/JOSS.02173.

17 Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997. doi:10.1162/NECO.1997.9.8.1735.

18 Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *CoRR*, abs/2310.06825, 2023. doi:10.48550/arXiv.2310.06825.

19 Nora Kassner, Oyvind Tafjord, Ashish Sabharwal, Kyle Richardson, Hinrich Schuetze, and Peter Clark. Language models with rationality. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14190–14201, Singapore, December 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.emnlp-main.877.

20 Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. Linking the thoughts: Analysis of argumentation structures in scientific publications. In Claire Cardie, editor, *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 1–11, Denver, CO, June 2015. Association for Computational Linguistics. doi:10.3115/v1/W15-0501.

21 Tobias Kuhn, Albert Meroño-Peñuela, Alexander Malic, Jorrit H. Poelen, Allen H. Hurlbert, Emilio Centeno Ortiz, Laura I. Furlong, Núria Queralt-Rosinach, Christine Chichester, Juan M. Banda, Egon L. Willighagen, Friederike Ehrhart, Chris T. A. Evelo, Tareq B. Malas, and Michel Dumontier. Nanopublications: A growing resource of provenance-centric scientific linked data. In *14th IEEE International Conference on e-Science, e-Science 2018, Amsterdam, The Netherlands, October 29 - November 1, 2018*, pages 83–92. IEEE Computer Society, 2018. doi:10.1109/ESCIENCE.2018.00024.

22 Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. An argument-annotated corpus of scientific publications. In Noam Slonim and Ranit Aharonov, editors, *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi:10.18653/v1/W18-5206.

23 Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Kai Eckert. Annotating arguments in scientific publications, 2018. URL: https://data.dws.informatik.uni-mannheim.de/sci-arg/annotation_guidelines.pdf.

24 John Lawrence and Chris Reed. Combining argument mining techniques. In Claire Cardie, editor, *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 127–136, Denver, CO, June 2015. Association for Computational Linguistics. doi:10.3115/v1/W15-0516.

25 John Lawrence, Jacky Visser, and Chris Reed. Harnessing rhetorical figures for argument mining. *Argument Comput.*, 8(3):289–310, 2017. doi:10.3233/AAC-170026.

26 Timothy Lebo, Satya Sahoo, and Deborah McGuinness. PROV-O: The PROV Ontology. https://www.w3.org/TR/prov-o/, 2013. W3C Recommendation 30 April 2013. Accessed: 2025-07-30.

27 Marco Lippi and Paolo Torroni. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Techn.*, 16(2):10:1–10:25, 2016. doi:10.1145/2850417.

28 Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In Eneko Agirre, Marianna Apidianaki, and Ivan Vulić, editors, *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics. doi:10.18653/v1/2022.deelio-1.10.

29 Kyle Lo, Zejiang Shen, Benjamin Newman, Joseph Chang, Russell Authur, Erin Bransom, Stefan Candra, Yoganand Chandrasekhar, Regan Huff, Bailey Kuehl, Amanpreet Singh, Chris Wilhelm, Angele Zamarron, Marti A. Hearst, Daniel Weld, Doug

Downey, and Luca Soldaini. PaperMage: A unified toolkit for processing, representing, and manipulating visually-rich scientific documents. In Yansong Feng and Els Lefever, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 495–507, Singapore, December 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.emnlp-demo.45.

30 Tobias Mayer, Elena Cabrio, Marco Lippi, Paolo Torroni, and Serena Villata. Argument mining on clinical trials. In Sanjay Modgil, Katarzyna Budzynska, and John Lawrence, editors, *Computational Models of Argument - Proceedings of COMMA 2018, Warsaw, Poland, 12-14 September 2018*, volume 305 of *Frontiers in Artificial Intelligence and Applications*, pages 137–148. IOS Press, 2018. doi:10.3233/978-1-61499-906-5-137.

31 Tobias Mayer, Elena Cabrio, and Serena Villata. Transformer-based argument mining for healthcare applications. In Giuseppe De Giacomo, Alejandro Catalá, Bistra Dilkina, Michela Milano, Senén Barro, Alberto Bugarín, and Jérôme Lang, editors, *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2108–2115. IOS Press, 2020. doi:10.3233/FAIA200334.

32 Peter Bjorn Nemenyi. *Distribution-free multiple comparisons*. PhD thesis, Princeton University, 1963.

33 Allard Oelen, Mohamad Yaser Jaradeh, and Sören Auer. ORKG ASK: a neuro-symbolic scholarly search and exploration system. In Daniel Garijo, Anna Lisa Gentile, Anelia Kurteva, Andrea Mannocci, Francesco Osborne, and Sahar Vahdati, editors, *Joint Proceedings of Posters, Demos, Workshops, and Tutorials of the 20th International Conference on Semantic Systems co-located with 20th International Conference on Semantic Systems (SEMANTiCS 2024), Amsterdam, The Netherlands, September 17-19, 2024*, volume 3759 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3759/paper7.pdf.

34 OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave

Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. arXiv:2303.08774.

35 Juri Opitz and Anette Frank. Dissecting content and context in argumentative relation analysis. In Benno Stein and Henning Wachsmuth, editors, *Proceedings of the 6th Workshop on Argument Mining*, pages 25–34, Florence, Italy, August 2019. Association for Computational Linguistics. doi:10.18653/v1/W19-4503.

36 Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi:10.3115/v1/D14-1162.

37 Georgios Petasis and Vangelis Karkaletsis. Identifying argument components through TextRank. In Chris Reed, editor, *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 94–102, Berlin, Germany, August 2016. Association for Computational Linguistics. doi:10.18653/v1/W16-2811.

38 Jason Priem, Heather A. Piwowar, and Richard Orr. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *CoRR*, abs/2205.01833, 2022. doi:10.48550/arXiv.2205.01833.

39 Iyad Rahwan, Fouad Zablith, and Chris Reed. Laying the foundations for a world wide argument web. *Artif. Intell.*, 171(10-15):897–921, 2007. doi:10.1016/J.ARTINT.2007.04.015.

40 Iyad Rahwan, Fouad Zablith, and Chris Reed. Laying the Foundations for a World Wide Argument Web. *Artificial Intelligence*, 171(10-15):897–921, 2007. doi:10.1016/j.artint.2007.04.015.

41 Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics, 2019. doi:10.18653/V1/D19-1410.

42 Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. Classification and clustering of arguments with contextualized word embeddings. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy, July 2019. Association for Computational Linguistics. doi:10.18653/v1/P19-1054.

43 Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton-Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. Code llama: Open foundation models for code. *CoRR*, abs/2308.12950, 2023. doi:10.48550/arXiv.2308.12950.

44 Florian Ruosch, Cristina Sarasua, and Abraham Bernstein. BAM: benchmarking argument mining on scientific documents. In Amir Pouran Ben Veyseh, Franck Dernoncourt, Thien Huu Nguyen, Walter Chang, and Viet Dack Lai, editors, *Proceedings of the Workshop on Scientific Document Understanding co-located with 36th AAAI Conference on Artificial Intelligence, SDU@AAAI 2022, Virtual Event, March 1, 2022*, volume 3164 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2022. URL: https://ceur-ws.org/Vol-3164/paper5.pdf.

45 Florian Ruosch, Cristina Sarasua, Chris Reed, and Abraham Bernstein. Toward the Argument Web of Science. In *Computational Models of Argument - Proceedings of COMMA 2024, Hagen, Germany, 18-20 September 2024*, Frontiers in Artificial Intelligence and Applications, 2024.

46 David M. Shotton. Cito, the citation typing ontology. *J. Biomed. Semant.*, 1(S-1):S6, 2010. URL: http://www.jbiomedsem.com/content/1/S1/S6.

47 Christian Stab and Iryna Gurevych. Identifying argumentative discourse structures in persuasive essays. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar, October 2014. Association for Computational Linguistics. doi:10.3115/v1/D14-1006.

48 Markus Stocker, Allard Oelen, Mohamad Yaser Jaradeh, Muhammad Haris, Omar Arab Oghli, Golsa Heidari, Hassan Hussein, Anna-Lena Lorenz, Salomon Kabenamualu, Kheir Eddine Farfar, Manuel Prinz, Oliver Karras, Jennifer D'Souza, Lars Vogt, and Sören Auer. Fair scientific information with the open research knowledge graph. *FAIR Connect*, 1:19–21, 2023. 1. doi:10.3233/FC-221513.

49 Simone Teufel, Jean Carletta, and Marc Moens. An annotation scheme for discourse-level argumentation in research articles. In Henry S. Thompson and Alex Lascarides, editors, *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117, Bergen, Norway, June 1999. Association for Computational Linguistics. URL: https://aclanthology.org/E99-1015.

50 C. J. van Rijsbergen. *Information Retrieval*. Butterworth, 1979.

51 Rosni Vasu, Cristina Sarasua, and Abraham Bernstein. SciHyp: A Fine-grained Dataset Describing Hypotheses and Their Components from Scientific Articles (Version V1) [Dataset], December 2023. Zenodo. https://doi.org/10.5281/zenodo.10298218.

52 Rosni Vasu, Cristina Sarasua, and Abraham Bernstein. Scihyp: A fine-grained dataset describing hypotheses and their components from sci-

entific articles. In Gianluca Demartini, Katja Hose, Maribel Acosta, Matteo Palmonari, Gong Cheng, Hala Skaf-Molli, Nicolas Ferranti, Daniel Hernández, and Aidan Hogan, editors, *The Semantic Web - ISWC 2024 - 23rd International Semantic Web Conference, Baltimore, MD, USA, November 11-15, 2024, Proceedings, Part III*, volume 15233 of *Lecture Notes in Computer Science*, pages 134–152. Springer, 2024. doi:10.1007/978-3-031-77847-6_8.

53 Jacky Visser, Rory Duthie, John Lawrence, and Chris Reed. Intertextual correspondence for integrating corpora. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL: https://aclanthology.org/L18-1554.

54 Douglas Walton. Argumentation theory: A very short introduction. In Guillermo Ricardo Simari and Iyad Rahwan, editors, *Argumentation in Artificial Intelligence*, pages 1–22. Springer, 2009. doi:10.1007/978-0-387-98197-0_1.

55 Douglas Walton and Nanning Zhang. The epistemology of scientific evidence. *Artif. Intell. Law*, 21(2):173–219, 2013. doi:10.1007/S10506-012-9132-9.

56 Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, 1(1):396–413, February 2020. doi:10.1162/qss_a_00021.

57 An Yang and Sujian Li. SciDTB: Discourse dependency TreeBank for scientific abstracts. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 444–449, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi:10.18653/v1/P18-2071.

A Regular Expressions for Identifying Citations

We used the following regular expressions to identify argumentative components that contain citations.

```
^\[.*?\]$
```

This matches square brackets with any content; for example, "[Smith et al., 2000]" or "[1]."

```
^.*?\[?(18|19|20)\d{2}(a|b|c|d|e)?\]?$
```

This matches patterns ending in years (1800–2099) plus optionally a letter or brackets; for example, "1992a" or "[2005]."

```
^\d?\d$
```

This matches matches one or two digits; for example, "1" or "13."

```
^\d?\d-\d?\d$
```

This matches matches digit rangs, "1–2" or "9–11."

```
^.{0,10} ?(5|6|7|8|9|0|1|2)\d$
```

This matches abbreviations and the last two digits of a year (50-29, space optional); for example, "EiHe92" or "RSB 23."

B Discourse Indicators for Rule-Based Argumentative Relation Prediction

The discourse indicators used for the rule-based argumentative relation prediction are shown in Table 5. The type for the predicted relation is also indicated for each column.

Table 5 The list of discourse indicators used for the rule-based argumentative relation prediction taken from [24, p. 128].

Supports	Attacks
because	however
therefore	but
after	though
for	except
since	not
when	never
assuming	no
so	whereas
accordingly	nonetheless
thus	yet
hence	despite
then	
consequently	

Table 6 The dataset splits for the fine-tuning of the transformer-based argument miner [31].

Set	Items
Train	A03, A04, A05, A06, A07, A08, A12, A13, A14, A15, A17, A19, A20, A22, A23, A24, A25, A28, A32, A33, A35, A36, A38, A39
Val	A02, A11, A21, A31
Test	A01, A09, A10, A16, A18, A26, A27, A29, A30, A34, A37, A40

C Fine-Tuning of the Transformer-Based Argument Miner

We fine-tuned the transformer-based argument miner on the argumentative relation annotations of the Sci-Arg corpus [22]. To this end, we used the uncased SciBERT [7] with the following parameters. The maximum sequence length was set to 128, the batch size per GPU to 32, the learning rate was $2e-5$, and we trained for three epochs. The system used an Adam optimizer with an epsilon of $1e-8$. All other parameters used the default values, as specified in the API of the argument miner. For the data splits, we followed [22] with only the validation set being our own choice. The exact subsets are shown in Table 6, where `AXX` denotes the identifier of one of the 40 papers in the dataset.

D Prompts for the Large Language Model

Here, we show the prompts for the two approaches using Mistral [18] for the argumentative relation prediction.

For the zero-shot, the prompt looked as follows. x and y are replaced with the components to be predicted.

```
Classify the type of argumentative relationship between two given text spans.  The
relationship can be from the list [none, supports, attacks].  Reply with only one word.
COMPONENT A: x
COMPONENT B: y
RELATION:
```

Table 7 The abbreviations used for the different approaches.

Approach	Abbreviation
Rule-Based	diam
TRABAM	trabam
Zero-Shot LLM	llmam_zero
Few-Shot LLM	llmam_prompt
Zero-Shot LLM with Context	llmam_zero_context
Few-Shot LLM with Context	llmam_prompt_context

For the few-shot, the prompt looked as follows. x and y are replaced with the components to be predicted. a and b are replaced with the components of the example pair, and r with their relation. Furthermore, the $5\times\{\dots\}$ indicates that this block is repeated five times with five different example pairs and their relations. The examples pairs are drawn from the 66,668 relations in the test set of the data based on the semantic similarity to the components to be classified. This similarity was computed with Sentence Transformer [41]. The class distribution of these in-context examples was as follows:

- noRel: 62,226 examples
- supports: 3,644 examples
- attacks: 798 examples

```
Classify the type of argumentative relationship between two given text spans.  The
relationship can be from the list [none, supports, attacks].  Reply with only one word.
Examples:
5×
{COMPONENT A: a
COMPONENT B: b
RELATION: r}
COMPONENT A: x
COMPONENT B: y
RELATION:
```

E Statistical Significance Testing

Below, we report the p-values where we claim to have found significant differences for the results shown in Table 3 and in Table 4. For the comparison of the accuracy on the overall dataset for all systems, we show the mean rank differences in Table 8 and for the comparison of the intra- and inter-document setting in Table 9. Bold numbers indicate statistically significant differences ($p < 0.05$). The testing for the statistical significance of the results was conducted with Autorank [16]. For brevity, the different approaches use abbreviations which are explained in Table 7. Furthermore, the suffix indicates whether subsets ("-intra" for intra-document, "-inter" for inter-document) or the entirety of the dataset ("-all") was used. The detailed reports on the conducted tests for statistical significance, including all procedures and assumptions testing, are partially shown below, but all are available in the code repository.[2]

Table 8 Pairwise absolute mean rank differences from the Nemenyi test [32]. Values in bold indicate a significant difference ($p < 0.05$, based on a critical distance of 2.176).

Population	Rule-Based	TRABAM	Zero-Shot LLM	Few-Shot LLM	Zero-Shot LLM (with Context)	Few-Shot LLM (with Context)
Rule-Based		1.000	**4.167**	**4.000**	**3.167**	**2.667**
TRABAM	1.000		**3.167**	**3.000**	2.167	1.667
Zero-Shot LLM	**4.167**	**3.167**		0.167	1.000	1.500
Few-Shot LLM	**4.000**	**3.000**	0.167		0.833	1.333
Zero-Shot LLM (with Context)	**3.167**	2.167	1.000	0.833		0.500
Few-Shot LLM (with Context)	**2.667**	1.667	1.500	1.333	0.500	

Table 9 The p-values of the paired t-test for the comparison of the results in the intra- versus inter-document setting.

	p-Value
Rule-Based	0.977
TRABAM	**0.003**
Zero-Shot LLM	**0.000**
Few-Shot LLM	0.149
Zero-Shot LLM with Context	**0.000**
Few-Shot LLM with Context	0.131

Overall Dataset for All Systems

The statistical analysis was conducted for 6 populations with 12 paired samples. The family-wise significance level of the tests α=0.050. We failed to reject the null hypothesis that the population is normal for all populations (minimal observed p-value=0.139). Therefore, we assume that all populations are normal. We applied Bartlett's test for homogeneity and reject the null hypothesis (p=0.039) that the data is homoscedastic. Thus, we assume that our data is heteroscedastic. Because we have more than two populations and the populations are normal but heteroscedastic, we use the non-parametric Friedman test as omnibus test to determine if there are any significant differences between the mean values of the populations. We use the post-hoc Nemenyi test to infer which differences are significant. We report the mean value (M), the standard deviation (SD) and the mean rank (MR) among all populations over the samples. Differences between populations are significant, if the difference of the mean rank is greater than the critical distance CD=2.176 of the Nemenyi test. We reject the null hypothesis (p=0.000) of the Friedman test that there is no difference in the central tendency of the populations llmam_zero-all (M=0.113±0.032, SD=0.034, MR=5.167), llmam_prompt-all (M=0.109±0.041, SD=0.044, MR=5.000), llmam_zero_context-all (M=0.119±0.035, SD=0.037, MR=4.167), llmam_prompt_context-all (M=0.132±0.028, SD=0.030, MR=3.667), trabam-all (M=0.251±0.049, SD=0.052, MR=2.000), and diam-all (M=0.437±0.067, SD=0.073, MR=1.000). Therefore, we assume that there is a statistically significant difference between the median values of the populations. Based on the post-hoc Nemenyi test, we assume that there are no significant differences within the following groups: llmam_zero-all, llmam_prompt-all, llmam_zero_context-all, and llmam_prompt_context-all; llmam_zero_context-all, llmam_prompt_context-all, and trabam-all; trabam-all and diam-all. All other differences are significant.

Inter- Versus Intra-Document

For each of the four comparisons, we had two populations with 12 paired samples. Since BAM [44] reports the mean of the samples, we conducted the paired t-test, the p-values of which are shown in Table 9. As we fail to reject the null hypothesis for normal distribution for all populations,

Table 10 The p-values of the Wilcoxon Signed-Rank Test comparison for different settings ($N = 12$ per population).

Comparison	M_1 (SD_1)	Mdn_1	M_2 (SD_2)	Mdn_2	W	p
diam-intra vs. diam-inter	0.430 (± 0.083)	0.421	0.432 (± 0.135)	0.424	39.0	1.000
trabam-intra vs. trabam-inter	0.238 (± 0.059)	0.215	0.345 (± 0.100)	0.351	1.0	**0.001**
llmam_zero-intra vs. llmam_zero-inter	0.083 (± 0.028)	0.084	0.249 (± 0.065)	0.257	0.0	**0.000**
llmam_prompt-intra vs. llmam_prompt-inter	0.091 (± 0.052)	0.092	0.136 (± 0.079)	0.139	18.0	0.110
llmam_zero_context-intra vs. llmam_zero_context-inter	0.090 (± 0.031)	0.093	0.252 (± 0.069)	0.256	0.0	**0.000**
llmam_prompt_context-intra vs. llmam_prompt_context-inter	0.113 (± 0.043)	0.121	0.165 (± 0.082)	0.176	17.0	0.092
llmam_prompt-all vs. llmam_prompt_context-all	0.109 (± 0.044)	0.104	0.132 (± 0.030)	0.128	11.0	**0.027**
llmam_zero-all vs. llmam_zero_context-all	0.113 (± 0.034)	0.108	0.119 (± 0.037)	0.114	3.0	**0.002**

we also report the p-values of the more robust Wilcoxon Signed-Rank Test in Table 10. They yield the same results with respect to the statistical significance of the differences between the populations.

Rule-Based: Inter- Versus Intra-Document

The statistical analysis was conducted for 2 populations with 12 paired samples. The family-wise significance level of the tests α=0.050. No check for homogeneity was required because we only have two populations. We use the t-test to determine differences between the mean values of the populations and report the mean value (M) and the standard deviation (SD) for each population. We failed to reject the null hypothesis (p=0.977) of the paired t-test that the mean values of the populations diam-intra (M=0.430±0.062, SD=0.083) and diam-inter (M=0.432±0.101, SD=0.135) are equal. *Therefore, we assume that there is no statistically significant difference between the mean values of the populations.*

Argument Miner: Inter- Versus Intra-Document

The statistical analysis was conducted for 2 populations with 12 paired samples. The family-wise significance level of the tests α=0.050. No check for homogeneity was required because we only have two populations. We use the t-test to determine differences between the mean values of the populations and report the mean value (M) and the standard deviation (SD) for each population. We reject the null hypothesis (p=0.003) of the paired t-test that the mean values of the populations trabam-intra (M=0.238±0.044, SD=0.059) and trabam-inter (M=0.345±0.075, SD=0.100) are equal. *Therefore, we assume that the mean value of trabam-inter is significantly larger than the mean value of trabam-intra with a large effect size (d=-1.299).*

Zero-Shot LLM: Inter- Versus Intra-Document

The statistical analysis was conducted for 2 populations with 12 paired samples. The family-wise significance level of the tests α=0.050. No check for homogeneity was required because we only have two populations. We use the t-test to determine differences between the mean values of the populations and report the mean value (M) and the standard deviation (SD) for each population. We reject the null hypothesis (p=0.000) of the paired t-test that the mean values of the populations llmam_zero-intra (M=0.083±0.021, SD=0.028) and llmam_zero-inter (M=0.249±0.049, SD=0.065) are equal. *Therefore, we assume that the mean value of llmam_zero-inter is significantly larger than the mean value of llmam_zero-intra with a large effect size (d=-3.323).*

Prompted LLM: Inter- Versus Intra-Document

The statistical analysis was conducted for 2 populations with 12 paired samples. The family-wise significance level of the tests α=0.050. No check for homogeneity was required because we only have two populations. We use the t-test to determine differences between the mean values of the populations and report the mean value (M) and the standard deviation (SD) for each population. We failed to reject the null hypothesis (p=0.149) of the paired t-test that the mean values of the populations llmam_prompt-intra (M=0.091±0.039, SD=0.052) and llmam_prompt-inter (M=0.136±0.059, SD=0.079) are equal. *Therefore, we assume that there is no statistically significant difference between the mean values of the populations.*

Zero-Shot LLM with Context: Inter- Versus Intra-Document

The statistical analysis was conducted for 2 populations with 12 paired samples. The family-wise significance level of the tests α=0.050. No check for homogeneity was required because we only have two populations. We use the t-test to determine differences between the mean values of the populations and report the mean value (M) and the standard deviation (SD) for each population. We reject the null hypothesis (p=0.000) of the paired t-test that the mean values of the populations llmam_zero_context-intra (M=0.090±0.023, SD=0.031) and llmam_zero_context-inter (M=0.252±0.052, SD=0.069) are equal. *Therefore, we assume that the mean value of llmam_zero_context-inter is significantly larger than the mean value of llmam_zero_context-intra with a large effect size (d=-3.022).*

Few-Shot LLM with Context: Inter- Versus Intra-Document

The statistical analysis was conducted for 2 populations with 12 paired samples. The family-wise significance level of the tests α=0.050. No check for homogeneity was required because we only have two populations. We use the t-test to determine differences between the mean values of the populations and report the mean value (M) and the standard deviation (SD) for each population. We failed to reject the null hypothesis (p=0.131) of the paired t-test that the mean values of the populations llmam_prompt_context-intra (M=0.113±0.032, SD=0.043) and llmam_prompt_context-inter (M=0.165±0.061, SD=0.082) are equal. *Therefore, we assume that there is no statistically significant difference between the mean values of the populations.*

Zero-Shot LLM: Context Versus No Context

The statistical analysis was conducted for 2 populations with 12 paired samples. The family-wise significance level of the tests α=0.050. No check for homogeneity was required because we only have two populations. We use the t-test to determine differences between the mean values of the populations and report the mean value (M) and the standard deviation (SD) for each population. We reject the null hypothesis (p=0.008) of the paired t-test that the mean values of the populations llmam_zero-all (M=0.113±0.026, SD=0.034) and llmam_zero_context-all (M=0.119±0.028, SD=0.037) are equal. *Therefore, we assume that the mean value of llmam_zero_context-all is significantly larger than the mean value of llmam_zero-all with a negligible effect size (d=-0.173).*

Few-Shot LLM: Context Versus No Context

The statistical analysis was conducted for 2 populations with 12 paired samples. The family-wise significance level of the tests α=0.050. No check for homogeneity was required because we only have two populations. We use the t-test to determine differences between the mean values of the populations and report the mean value (M) and the standard deviation (SD) for each population.

Table 11 The results of the experiments on the overall dataset. We report the class-based F1-score, precision, and recall.

Approach		F1	Precision	Recall
Rule-Based	Supports	0.346	0.929	0.212
	Attacks	0.342	0.318	0.371
Argument Miner	Supports	0.309	0.882	0.187
	Attacks	0.196	0.708	0.114
Zero-Shot LLM	Supports	0.842	0.792	0.899
	Attacks	0.010	1.000	0.005
Zero-Shot LLM + Context	Supports	0.863	0.790	0.950
	Attacks	0.010	0.286	0.005
Few-Shot LLM	Supports	0.182	0.936	0.101
	Attacks	0.020	0.667	0.010
Few-Shot LLM + Context	Supports	0.380	0.899	0.241
	Attacks	0.010	0.400	0.005

Table 12 The results of the experiments on the intra-document subset. We report the class-based F1-score, precision, and recall.

Approach		F1	Precision	Recall
Rule-Based	Supports	0.329	0.888	0.202
	Attacks	0.376	0.382	0.370
Argument Miner	Supports	0.308	0.865	0.187
	Attacks	0.185	0.694	0.107
Zero-Shot LLM	Supports	0.770	0.700	0.855
	Attacks	0.010	1.000	0.005
Zero-Shot LLM + Context	Supports	0.802	0.704	0.930
	Attacks	0.010	0.400	0.005
Few-Shot LLM	Supports	0.140	0.876	0.076
	Attacks	0.020	0.800	0.010
Few-Shot LLM + Context	Supports	0.329	0.830	0.205
	Attacks	0.010	0.400	0.010

We reject the null hypothesis (p=0.012) of the paired t-test that the mean values of the populations llmam_prompt-all (M=0.109±0.033, SD=0.044) and llmam_prompt_context-all (M=0.132±0.022, SD=0.030) are equal. *Therefore, we assume that the mean value of llmam_prompt_context-all is significantly larger than the mean value of llmam_prompt-all with a medium effect size (d=-0.618).*

F Class-Based Performance

Table 11 shows the class-based performance for the four approaches on the overall dataset. Tables 12 and 13 show the class-based results for the intra- and inter-document subset, respectively. The class distribution is shown in Table 14.

Table 13 The results of the experiments on the inter-document subset. We report the class-based F1-score, precision, and recall.

Approach		F1	Precision	Recall
Rule-Based	Supports	0.375	1.000	0.231
	Attacks	0.024	0.012	1.000
Argument Miner	Supports	0.429	1.000	0.273
	Attacks	0.000	0.000	0.000
Zero-Shot LLM	Supports	0.988	0.998	0.979
	Attacks	0.000	0.000	0.000
Few-Shot LLM	Supports	0.253	1.000	0.145
	Attacks	0.000	0.000	0.000
Zero-Shot LLM + Context	Supports	0.993	0.998	0.988
	Attacks	0.000	0.000	0.000
Few-Shot LLM + Context	Supports	0.470	1.000	0.307
	Attacks	0.000	0.000	0.000

Table 14 The distribution of the classes for the overall dataset and the intra- and inter-document subsets.

Class	Overall	Intra	Inter
Supports	1592	1025	567
Attacks	404	403	1
Total	1996	1428	568

G Statistics of Misclassified Samples

Table 15 shows the distributions of the misclassified samples per relation type for the different approaches. Each column indicates the real class of the samples, and each row shows the assigned relation type by the approaches. Since the zero-shot LLM approach did not only produce valid labels but also a variety of responses, Tables 15c and 15d have an additional row (*Various*) where these predictions are pooled.

H Technical Infrastructure and Computational Budget

All experiments were performed on a machine with eight NVIDIA GeForce RTX 4090 GPUs and one AMD EPYC 9124 3.0GHz 16-core CPU. Only a single GPU was used at a time for each experiment run, where applicable. Any experiment not leveraging the GPU was performed on the CPU: this only applies to the Argument Mining using the presence of discourse indicators.

The computational budget for all experiments was approximately 120 GPU hours.

Table 15 Distribution of the misclassified samples per relation type for all six approaches evaluated. The gold standard relation is shown in the columns, and the predicted system relation is shown in the rows. Subfigure (a) is the rule-based approach, (b) the argument miner, (c) the zero-shot LLM, (d) the zero-shot LLM with context, (e) the few-shot LLM, and (f) the few-shot LLM with context.

(a) Rule-based approach.

System \ Gold	Attacks	Supports	No Relation
Attacks	–	0.257	0.440
Supports	0.102	–	0.560
No Relation	0.898	0.743	–

(b) Argument miner [31] approach.

System \ Gold	Attacks	Supports	No Relation
Attacks	–	0.015	0.134
Supports	0.112	–	0.866
No Relation	0.888	0.985	–

(c) Zero-shot LLM approach.

System \ Gold	Attacks	Supports	No Relation
Attacks	–	0.000	0.000
Supports	0.935	–	0.911
No Relation	0.005	0.087	–
Various	0.060	0.913	0.089

(d) Zero-shot LLM approach with context.

System \ Gold	Attacks	Supports	No Relation
Attacks	–	0.063	0.004
Supports	0.998	–	0.953
No Relation	0.000	0.051	–
Various	0.002	0.886	0.043

(e) Few-shot LLM approach.

System \ Gold	Attacks	Supports	No Relation
Attacks	–	0.001	0.034
Supports	0.028	–	0.966
No Relation	0.972	0.999	–

(f) Few-shot LLM approach with context.

System \ Gold	Attacks	Supports	No Relation
Attacks	–	0.002	0.017
Supports	0.107	–	0.983
No Relation	0.893	0.998	–