SciEvent: Benchmarking Multi-domain Scientific Event Extraction

Anonymous EMNLP submission

Abstract

Scientific information extraction (SciIE) has primarily relied on entity-relation extraction in narrow domains, limiting its applicability to interdisciplinary research and struggling to capture the necessary context of scientific information, often resulting in fragmented or conflicting statements. In this paper, we introduce SciEvent[1], a novel multi-domain benchmark of scientific abstracts annotated via a unified event extraction (EE) schema designed to enable structured and context-aware understanding of scientific content. It includes 500 abstracts across five research domains, with manual annotations of event segments, triggers, and fine-grained arguments. We define SciIE as a multi-stage EE pipeline: (1) segmenting abstracts into core scientific activities—*background*, *methods*, *results*, and *conclusions*; and (2) extracting the corresponding triggers and arguments. Experiments with fine-tuned EE models, large language models (LLMs), and human annotators reveal a performance gap, with current models struggling in domains such as sociology and humanities. SciEvent serves as a challenging benchmark and a step toward generalizable, multi-domain SciIE.

1 Introduction

Scientific information extraction (SciIE) distills structured knowledge from unstructured scientific articles and supports key scientific applications such as literature review (Hong et al., 2021), paper recommendation (Ikoma and Matsubara, 2023), and knowledge discovery (Stavropoulos et al., 2023), especially in recent years as many domains are facing a publication deluge.

Existing works on SciIE generally follow an entity-relation extraction (ERE) paradigm that aims to extract isolated scientific concepts and connect them by identifying semantic relations, either binary (Luan et al., 2018; Zhang et al., 2024) or N-ary (Jain et al., 2020; Zhuang et al., 2022). Despite remarkable contributions made by prior studies, one major concern is that representing scientific content as disconnected entity-relation tuples may fragment the underlying narrative and even introduce conflicting statements, especially when synthesizing information across multiple publications. As shown in Figure 1, one paper may generate the tuple ⟨ "GPT-3.5-Turbo", "better than", "GPT-4-Turbo"⟩, while another produces the opposite. Without contextual cues such as task setup or evaluation criteria, these tuples alone fail to convey meaningful or reliable scientific insights.

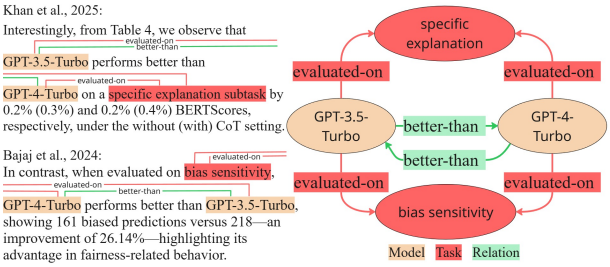


Figure 1: Conflicting statements in entity-relation extraction. ⟨GPT-3.5-Turbo, better than, GPT-4-Turbo⟩ vs. ⟨GPT-4-Turbo, better than, GPT-3.5-Turbo⟩

Inspired by the heavily context-dependent nature of scientific publications, we adopt an event extraction (EE) paradigm. This paradigm focuses on identifying triggers that best represent each event and extracting associated arguments, which are then assigned specific semantic roles. This enables a more structured and context-aware representation of important scientific information. Despite its potential for representing scientific information, a major limitation of existing EE efforts in the scientific domain is their narrow focus on specific fields, often resulting in the development of domain-specific EE schemas. For example, (Zhang et al., 2024) and (Jain et al., 2020) focus on machine learning, and (Kim et al., 2011) focus on bio-molecule area. Given the rapid growth of interdisciplinary research

[1]We will release the dataset and code upon publication of the paper.

in recent years, there is an increasing need for a unified scientific EE schema capable of generalizing across diverse scholarly domains.

To address this gap, we introduce SciEvent, a unified EE schema for scientific texts, along with a dataset featuring manually annotated events and fine-grained arguments drawn from diverse research abstracts. Building on this dataset, we define three SciIE tasks: (1) event segmentation, which involves dividing the text into spans that represent core scientific activities such as *background*, *methods*, *results*, and *conclusions*; (2) trigger identification, which aims to detect the key anchor of each scientific event; and (3) argument extraction, which focuses on identifying the arguments involved in each scientific activity and assigning them roles such as context, method, or result. Differing from conventional EE pipelines, we introduce event segmentation as a preliminary task, recognizing that events in scientific texts often span multiple sentences and lack clear boundaries. Additionally, trigger words in scientific texts—such as "show", "demonstrate", or "present"—are frequently shared across different event types. Without first segmenting the text into discrete events, it becomes challenging to accurately delineate event boundaries, increasing the risk of misinterpreting or misclassifying both triggers and their associated arguments.

SciEvent contains 500 abstracts from five diverse scientific domains, each fully annotated using an EE paradigm. To evaluate the challenges posed by this dataset, we assess the performance of fine-tuned EE models, tuning-free large language models (LLMs), and human annotators. The results demonstrate SciEvent's broad domain coverage and reveal that existing models consistently lag behind human performance. This gap highlights the limitations of current approaches and the absence of EE models capable of generalizing across scientific domains.

2 Related Work

Event Extraction Existing work on event extraction (EE) typically frames the task via two paradigms. One is trigger-argument extraction (Walker et al., 2006; Hsu et al., 2022; Lin et al., 2020), where the trigger serves as the event anchor, most clearly signaling the occurrence of an event, while the arguments represent entity mentions that participate in the event, each fulfilling distinct roles. The other one treats EE as a trigger-free template-filling task (muc, 1992; Du and Cardie, 2020a; Huang et al., 2021), aiming to extract event-relevant arguments and assigning them to specific roles within each event template. The latter mainly focuses on document-level EE (Du and Cardie, 2020a), while the former has been widely used in both sentence-level (Walker et al., 2006) and document-level EE (Li et al., 2021). Our benchmark follows the trigger-argument paradigm.

Regarding EE benchmarks, prior studies have largely focused on data in generic domains. Popular examples include newswire (Grishman and Sundheim, 1996; Nguyen et al., 2016; Doddington et al., 2004; Ebner et al., 2020; Song et al., 2015), Wikipedia (Li et al., 2021; Pouran Ben Veyseh et al., 2022), social media (Sharif et al., 2024; Wang and Zhang, 2017; Comito et al., 2019) and widely-used knowledgebases like FrameNet (Baker et al., 1998) and PropBank (Bonial et al., 2014). While some researchers have broadened the scope of EE to scientific literature, their efforts tend to center the biomedical domain, particularly emphasizing state changes and interactions between biomolecules such as genes and proteins (Kim et al., 2011; Pyysalo et al., 2012; Kim et al., 2013). Differing from prior work, we extend EE to encompass a broader range of scientific domains, creating a unified annotation schema designed to facilitate interdisciplinary information extraction.

Scientific Information Extraction Research on scientific information extraction (IE) primarily targets two main types of information: (1) citation-based analysis, which involves identifying either binary citation influence classification (Kunnath et al., 2020; N. Kunnath et al., 2021; Maheshwari et al., 2021) or multi-class citation intents (purpose) classification (Cohan et al., 2019; Jurgens et al., 2018), and (2) content-based analysis (Gupta and Manning, 2011; Tsai et al., 2013; Gábor et al., 2016; Pronesti et al., 2025), which primarily focuses on extracting scientific entities, supporting evidence, and semantic relationships among them, with the ultimate goal of building concept-centric knowledge graphs (Ma et al., 2022; Zhang et al., 2020; Sap et al., 2019). For example, SciERC (Luan et al., 2018), consists of 500 scientific abstracts annotated with scientific entities, their pairwise relations, and coreference clusters. SciREX (Jain et al., 2020) provides annotations across 438 full documents, covering four entity types: TASK, DATASET, METHOD, and METRIC. Beyond gen-

eral knowledge extraction, some studies further focus on specific research subjects. This line of work designs domain-specific event extraction tasks to capture fine-grained scientific activities (He et al., 2024,Kim et al., 2011, Huang et al., 2020, Björne et al., 2010). For example, various biomedical EE tasks have been proposed to investigate biological processes such as protein-protein and gene-disease interactions (Kim et al., 2013; Kim et al., 2011; Björne et al., 2010). Our work similarly focuses on scientific EE. However, differing from prior works targeting specific domain, we aim to design a unified schema for organizing general scientific activities across diverse scientific domains.

3 SciEvent Benchmark

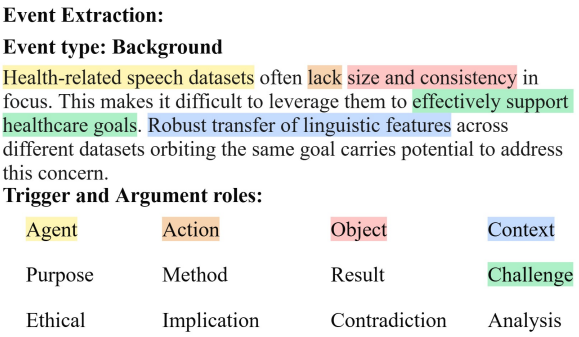


Figure 2: An example from SciEvent, each event in SciEvent is annotated with trigger and argument roles.

Data Collection To support cross-domain evaluation and capture diverse writing conventions, we selected publicly available, peer-reviewed English abstracts from 2023 to reflect contemporary language use. We selected five domains: natural language processing (NLP) from the Annual Meeting of the Association for Computational Linguistics (ACL)[2], social computing (SC) from Proceedings of the ACM on Human-Computer Interaction (CSCW)[3], medical informatics (MI) from Journal of Medical Internet Research (JMIR)[4], computational biology (CB) from Bioinformatics[5], and digital humanities (DH) from Digital Humanities Quarterly (DHQ)[6]. These domains were selected for their methodological diversity, different resource availability, relevance to interdisciplinary research, and representativeness of their respective fields. NLP and biomedical domains are well-studied and offer structured, technical abstracts, while SC and DH are underrepresented and characterized by more narrative, context-rich writing. To support document-level modeling, we retained abstracts with at least three sentences and two identifiable events, filtering out those that were too short to provide meaningful structure. In total, we collected 500 scientific abstracts—100 each in NLP, SC, and CB, 120 in DH, and 80 in MI. DH has fewer publications and shorter abstracts, so we extended the sampling range to 2021–2023 and included 120 abstracts to ensure sufficient coverage of domain variation. MI abstracts are longer and denser, so we selected 80 abstracts to balance event content comparability across all five domains.

Annotation Pipeline Our overall pipeline of annotation schema consist of (1) event segmentation, and (2) trigger-argument extraction. In the first step, we segment an abstract into four event types: *background*, *methods*, *results*, and *conclusions*, which are adopted from the most common aspects of scientific publications (Ripple et al., 2012). In the second step, each segment is annotated with a tuple representing the event trigger, and is enriched with role-specific arguments.

In prior event extraction work, particularly in newswire and broadcast domains, triggers like "attack" define clear and stable event frames, with roles such as "attacker" and "target" naturally grounded in the trigger's semantics. In scientific texts, however, single word triggers like "show" lack this clarity. Even after event segmentation and the event type (e.g., "Result") is known, the trigger alone does not specify what the event is about. Roles like "shower" or "shown item" are not meaningful on their own, as the event's meaning depends on the full proposition. For example, "showing a promising result" differs from "showing a methodological limitation". To capture this, we represent the trigger as a tuple of (Agent, Action, Object), anchoring the event in its core semantics.

To fully capture a scientific event, we then annotate its arguments. We defined nine argument roles: *Context*, *Purpose*, *Method*, *Results*, *Analysis*, *Challenge*, *Ethical*, *Implication*, and *Contradiction* [7]. Each role targets a specific dimension of scientific abstract, adapted from Core Scientific Concepts (Liakata et al., 2012) and inspired by scientific writing guides (Paltridge, 2002; Alley, 1996). While some argument roles share names with event types

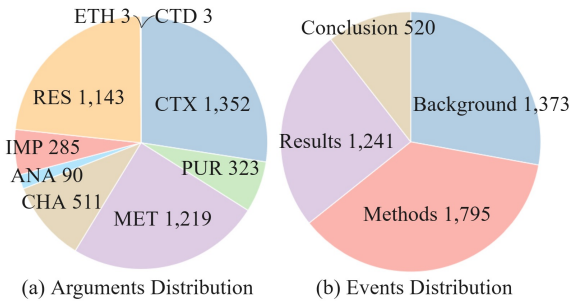
[2] https://aclanthology.org/events/acl-2023/
[3] https://dl.acm.org/toc/pacmhci/2023/7/CSCW1
[4] https://www.jmir.org/2023
[5] https://pubmed.ncbi.nlm.nih.gov/
[6] https://www.digitalhumanities.org/dhq/

[7] Detailed definitions in appendix B

(e.g., Methods, Results), they are not restricted to those events; for example, evaluation methods often appear within the Results event.

Annotation Quality We employed five graduate student annotators, all specializing in NLP and either native English speakers or PhD students. Each annotator had domain expertise in at least one of the five annotated domains, and collectively they covered all five diverse scientific areas. They were instructed to prepare gold annotations, beginning with event segmentation, followed by trigger identification and argument extraction. To evaluate the quality of event segment annotations, we randomly sampled 10 abstracts per domain and had two annotators independently annotate each. Intercoder reliability, measured by Cohen's Kappa (Cohen, 1960), is 0.83, showing strong agreement.

Given any extracted event, annotators then extract the trigger tuple and associated arguments within the event, following definitions and examples provided in the codebook[7]. Unlike event segmentation, assessing intercoder reliability for triggers and arguments is challenging due to span-level granularity and variability in trigger selection. Minor boundary differences can lead to large mismatches under strict metrics. To ensure quality, we conducted multiple annotation rounds with author-led review. To assess usability and estimate human upper bound, six additional untrained annotators received brief tutorials and annotated independently, providing a baseline for model comparison. All annotations were done using a custom-built interface shown in appendix B.1.



(a) Arguments Distribution (b) Events Distribution

Figure 3: Distribution of (a) argument roles and (b) event types across the dataset.

Data Analysis Using the above annotation pipeline, we construct a dataset of 500 annotated scientific abstracts containing 8,911 structured mentions, as shown in Table 1. Its broad domain coverage supports robust cross-domain analysis.

As shown in Figure 3, the most frequently annotated arguments are Context (CTX), Method (MET), and Results (RES), highlighting the dataset's emphasis on core components of scientific reporting. Rare arguments such as Contradictions (CTD) and Ethical (ETH) suggest that such aspects are rarely discussed in abstracts. The most common event types are Methods, consistent with typical abstract structure. Moreover, Figure 4 shows that argument types align well with event types—for example, Context appears predominantly in Background events, supporting the reliability and internal consistency of our annotations.

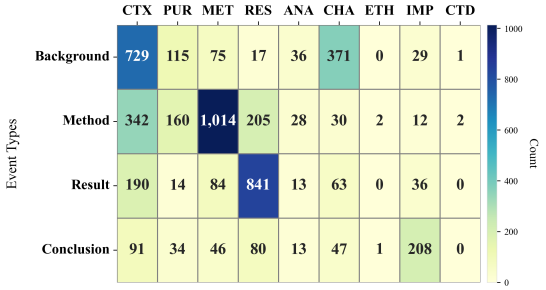


Figure 4: Distribution of argument across event types

4 Experiment and Evaluations Settings

We show the challenges of SciEvent by conducting comprehensive experiments on various state-of-the-art models. We define three tasks: Scientific Event Segmentation, SciEvent Trigger Identification, and SciEvent Argument Extraction.

Definition on Notations The input is a document represented as a sequence of sentences $D = \{S_1, \ldots, S_N\}$. Formally, the goal is to extract a set of scientific events $\mathcal{E} = \{E_1, \ldots, E_M\}$, where each event E_i is defined as: $E_i = (s_i, s_j, t_k, \text{Trigger}_i, \text{Arg}_i)$ Here, (s_i, s_j) denotes a contiguous sentence span in D, and $t_k \in \mathcal{T}$ is the event type. The trigger Trigger_i is a tuple of three core argument spans:

$$\text{Trigger}_i = (\sigma_{\text{agent}},\ \sigma_{\text{action}},\ \sigma_{\text{object}})$$

Each $\sigma \in D$ is a token span represented as (σ_s, σ_e) indicating the start and end token indices of the Agent, Action, and Object, respectively. The set Arg_i contains argument-role pairs:

$$\text{Arg}_i = \{((a_s, a_e), r_l) \mid 1 \le a_s \le a_e,\ r_l \in \mathcal{R}\}$$

where (a_s, a_e) denotes a token span and r_l is a semantic role from the predefined schema $\mathcal{R}$.

Task 1: Scientific Event Segmentation We define this task as segmenting the abstract into contiguous spans, where each segment is treated as an event and labeled with one of four event types.

Dataset	#Doc	#Mentions	Arg./Ent. Types	Avg Sent./Evt	Paradigm	Source	Domains
SciREX	438	8,592	4	-	ERE	Full paper	ML
SciERC	500	8,089	6	-	ERE	Abstract	Speech, ML, CV, AI
SemEval17	493	8,529	3	-	ERE	Paragraph	CS, MS, Physics
SemEval18	500	7,505	1	-	ERE	Abstract	CL
SciER	106	24,518	3	-	ERE	Full paper	ML
Genia2011	1,224	21,549	10	1	EE	Abstract/Full	BioMol
SciEvent (ours)	500	8,911	9	2.95	EE	Abstract	NLP, SC, CB, MI, DH

Table 1: Comparison of scientific IE datasets. Abbreviations: **Arg./Ent. Types** = Argument/Entity Types, **Avg Sent./Evt** = Average Sentence Per Event, **NLP** = Natural Language Processing, **SC** = Social Computing, **CB** = Computational Biology, **MI** = Medical Informatics, **DH** = Digital Humanities, **ML** = Machine Learning, **AI** = Artificial Intelligence, **CV** = Computer Vision, **CS** = Computer Science, **MS** = Material Science, **CL** = Computational Linguistics, **BioMol** = Biomolecular.

Formally, this task predicts a set of labeled sentence $\{(s_i, s_j, t_k)\}$, where $1 \leq i \leq j \leq N$ and $t_k \in \mathcal{T}$. We evaluate model predictions using *Exact Match* (EM) and *Intersection over Union* (IoU) metrics, adapted from span-based evaluation metrics in SemEval (Segura-Bedmar et al., 2013) and MUC-5 (Chinchor and Sundheim, 1993). A predicted tuple $(\hat{s}_i, \hat{s}_j, \hat{t}_k)$ is considered correct under:

- **Exact Matching (EM):** $(\hat{s}_i, \hat{s}_j) = (s_i, s_j)$ and $\hat{t}_k = t_k$
- **Intersection over Union (IoU):** $\frac{|\hat{s} \cap s|}{|\hat{s} \cup s|} > 0.5$ and $\hat{t}_k = t_k$, where $\hat{s}$ and s are the sets of sentence indices in the predicted and gold spans.

For both strategies, we report Precision (P), Recall (R), and F1-score (F1) over the set of predicted and gold event segments.

Task 2: SciEvent Trigger Identification This task involves extracting the trigger tuple (Agent, Action, Object) within each detected event span. As this is a document-level task, and scientific events often contain multiple candidate triggers, we treat this step separately to explicitly evaluate the model's ability to accurately locate the core semantic components of an event after its span and type have been identified.

Formally, for each detected event E_i, the model predicts: $\text{Trigger}_i = (\sigma_{\text{agent}}, \sigma_{\text{action}}, \sigma_{\text{object}})$. We use macro ROUGE-L (Lin, 2004) to evaluate predicted triggers, as it captures longest common subsequence (LCS) overlap, reflecting both lexical and structural alignment. Let C_i denote the concatenated text of the trigger, ordered as Agent, Action, and Object. We compute: $\text{ROUGE} - \text{L}(\hat{C}_i, C_i)$

Task 3: SciEvent Argument Extraction This task focuses on extracting arguments using a predefined set of roles $\mathcal{R}$. For each event E_i, the model predicts: $\text{Arg}_i = \{((a_s, a_e), r_l) \mid 1 \leq a_s \leq a_e, r_l \in \mathcal{R}\}$. Each argument is represented by a token span (a_s, a_e) and its role label r_l.

We decompose this task into two sub-tasks:

- **Argument Identification (Arg-I):** Predict the set of argument spans $\{(a_s, a_e)\}$ for each event.
- **Argument Classification (Arg-C):** Assign a role $r_l \in \mathcal{R}$ to each identified span.

we evaluate both Arg-I and Arg-C using F1 scores under two span matching strategies: Exact Match (EM) and Intersection over Union (IoU), with a threshold of 0.5.

A predicted argument $((\hat{a}_s, \hat{a}_e), \hat{r}_l)$ matches a gold argument $((a_s, a_e), r_l)$ in Arg-I sub-task if:

- **Exact Matching (EM):** $(\hat{a}_s, \hat{a}_e) = (a_s, a_e)$
- **Intersection over Union (IoU):** $\frac{|\hat{a} \cap a|}{|\hat{a} \cup a|} > 0.5$

where $\hat{a}$ and a denote the sets of token indices spanned by the predicted and gold arguments, respectively. If the role label of predicted argument also match with gold argument, i.e. $\hat{r}_l = r_l$, then predicted argument also matches a gold argument in Arg-C sub-task.

LLM baselines We use LLM baselines for the above tasks: (1) meta-Llama-3.1-8B-Instruct (Llama) (Meta AI, 2024) , (2) Qwen2.5-7B-Instruct (Qwen) (Qwen Team, 2024), (3) DeepSeek-R1-Distill-Llama-8B (DeepSeek-R1) (DeepSeek-AI, 2025), and (4) GPT-4.1 (GPT) (OpenAI, 2025) as baseline models. For Task 1, we conducted preliminary research on zero-shot prompting, and we use the best prompt, adapted from Sharif et al., 2024. For Task 2 and 3, we use zero-shot and one-shot prompting, also preliminarily tested and adapted from Sharif et al., 2024.

Tuning-based baselines For Task 2 and 3, we also consider three supervised models: (1) DEGREE (Hsu et al., 2022), a data-efficient generative

approach to event argument extraction that leverages prompt-based learning for better generalization. (2) OneIE (Lin et al., 2020), a joint information extraction framework that simultaneously performs entity, relation, and event extraction using a unified representation. (3) EE_QA (Du and Cardie, 2020b), a transformer-based model that frames information extraction as a question-answering task, enabling contextualized argument extraction.

5 Experiment Results

Model	EM			IoU		
	P	R	F1	P	R	F1
DeepSeek-R1	31.26	34.13	32.63	58.97	64.38	61.56
Qwen	43.51	36.30	39.58	70.30	58.65	63.95
Llama	38.67	31.70	34.84	62.04	50.85	55.89
GPT	**59.07**	**62.96**	**60.95**	**82.98**	**88.45**	**85.63**

Table 2: Scientific event segmentation performance (%) on zero-shot LLMs using Exact Match (EM) and Intersection over Union (IoU) metrics, showing Precision (P), Recall (R), and F1-score

Scientific event segmentation We experiment Zero-shot LLMs on Task 1. As Table 2 presents, GPT-4.1 clearly outperforms all others by a wide margin, achieving 60.95% F1 under EM and 85.63% under IoU, indicating its strong ability to identify and segment coherent scientific spans. Qwen ranks second, while Llama and DeepSeek-R1 trail closely with modest differences. These results suggest that segmentation is best handled by higher-capacity models like GPT-4.1.

Methods	P	R	F1
Tuning-based models			
EEQA	**81.93**	34.57	45.05
DEGREE	64.56	63.49	56.85
OneIE	73.73	**79.40**	72.40
Zero-shot LLMs			
DeepSeek-R1	29.12	27.10	26.74
Qwen	43.84	55.25	47.57
Llama	54.88	61.07	55.83
GPT	65.38	72.73	67.57
One-shot LLMs			
DeepSeek-R1	41.81	41.94	40.72
Qwen	56.17	68.48	59.98
Llama	53.08	63.83	56.45
GPT	72.67	77.77	**74.05**

Table 3: ROUGE-L scores (%) for baseline models on the SciEvent trigger identification task, showing Precision (P), Recall (R), and F1.

SciEvent Trigger Identification We evaluate Task 2 with tuning-based models and LLMs using ROUGE-L metrics (Table 3). GPT-4.1 (One-shot) performs the best (F1: 74.05%). OneIE also performs competitively (F1: 72.40%). EEQA shows extremely high precision (P: 81.93%) but poor recall (R: 34.57%), suggesting over-conservative predictions. All LLMs improve with one-shot prompting, especially DeepSeek-R1 (F1: +13.98%), showing that in-context examples enhance LLM performance for scientific trigger identification.

Methods	Arg-I (IoU)			Arg-C (IoU)		
	P	R	F1	P	R	F1
Tuning-based models						
EEQA	32.09	33.77	32.91	25.85	27.20	26.51
DEGREE	**67.79**	19.13	29.84	**48.99**	13.83	21.57
OneIE	51.11	**56.29**	**53.57**	39.69	**43.71**	**41.61**
Zero-shot LLMs						
DeepSeek-R1	31.11	16.46	21.53	16.32	8.63	11.29
Qwen	35.68	26.41	30.35	17.58	13.01	14.96
Llama	24.37	24.90	24.63	11.68	11.93	11.80
GPT	43.03	55.56	48.50	30.40	39.25	34.26
One-shot LLMs						
DeepSeek-R1	42.62	17.67	24.98	19.59	8.12	11.48
Qwen	46.33	30.36	36.69	20.96	13.74	16.60
Llama	44.70	34.08	38.68	18.93	14.44	16.38
GPT	50.14	50.22	50.18	34.60	34.66	34.63

Table 4: IoU-based Precision (P), Recall (R), and F1-score (%) on baseline models for argument identification (Arg-I) and classification (Arg-C) tasks.

SciEvent Argument Extraction We evaluate Task 3 with tuning-based models and LLMs using argument identification (Arg-I) and classification (Arg-C) under the IoU metric (Table 4). OneIE achieves the highest scores (Arg-I: 53.57%, Arg-C: 41.61%), benefiting from its structured decoding and joint modeling framework. DEGREE exhibits high precision but low recall, indicating that the model consistently misses relevant arguments in scientific abstracts. Among LLMs, GPT-4.1 (one-shot) performs best (Arg-I: 50.18%, Arg-C: 34.63%), while other models perform notably worse, especially on argument classification (Arg-C around 15%). One-shot prompting offers a modest improvement over zero-shot settings.

Figure 5 reports IoU-based F1 scores for argument classification across argument roles. Among tuning-based and LLM-based models, OneIE and GPT-4.1 achieve the strongest performance across nearly all argument roles. Qwen achieves a spike on Contradiction, due to a few correct extractions, but shows worse performance overall. Across all models, Challenge, Result, and Method yield the highest F1 scores, due to their clearer lexical cues and more regular positioning in scientific abstracts.

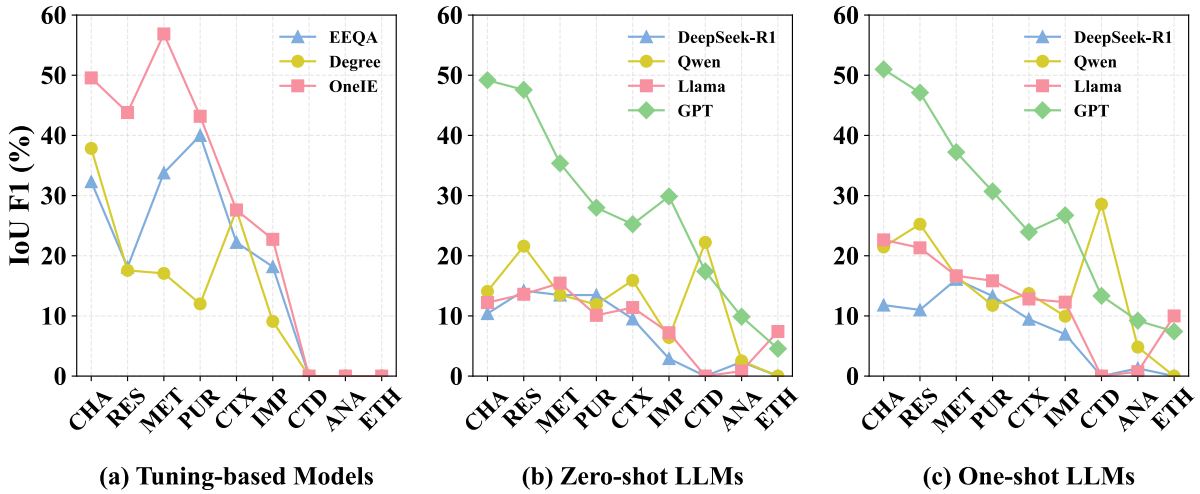


Figure 5: Intersection-over-Union (IoU) on Arg-C F1-scores (%) across different argument roles for various models on Analysis (ANA), Challenge (CHA), Context (CTX), Method (MET), Purpose (PUR), Result (RES), Ethical (ETH), Implication (IMP), Contradictions (CTD).

In contrast, arguments like Ethical, Contradiction, and Analysis remain challenging due to data sparsity and a lack of consistent lexical patterns.

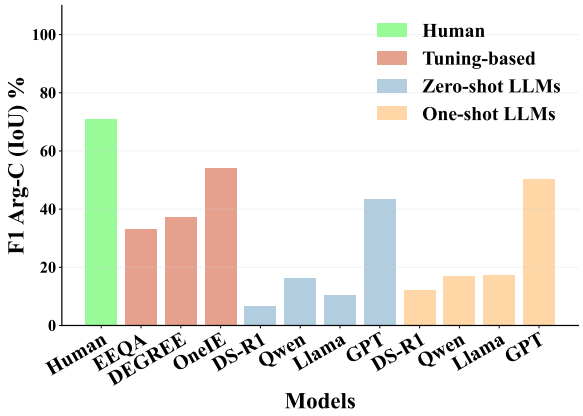


Figure 6: Human performance compared to all baselines on argument classification (Arg-C) using IoU F1 scores.

Human performance We compare model and human performance on argument classification. We do not report results for event segmentation, as the Cohen's kappa score of 0.83 (exact match) on a subset indicates consistently high agreement among annotators, suggesting that event segmentation is relatively unambiguous for humans. As shown in Figure 6, there is a substantial gap between human performance and the best model (20%). This highlights the challenge of multi-domain scientific event extraction and the value of SciEvent for advancing argument level scientific event extraction.

What is the impact of event type in SciEvent tasks? The arguments in Methods exhibit a notable gap: strong performance with supervision (OneIE, EEQA) but poor with zero-/one-shot LLMs on argument classification task (Figure 7). This finding suggests that arguments in the Method events are most demanding, due to event's complex structure, arguments' varied phrasing, and dependence on technical details, making performance poorer without supervision. Furthermore, Conclusion shows the lowest Arg-C performance for most models. EEQA performs better because its QA-based prompts help extract the implicit and interpretive content typical of Conclusion events.

What is the impact of scientific domains in SciEvent tasks? In the argument classification task (Figure 8), natural language processing and computational biology domains yield the highest F1 scores, benefiting from consistent linguistic patterns and clearer argument structures. In contrast, digital humanities and medical informatics present greater challenges, due to varied rhetorical styles and longer, denser abstracts, respectively.

How does removal of domain affect performance? We compare the argument classification performance of the OneIE model under the Exact Match (EM) setting using the full training set versus ablated training sets (Figure 9). Removing a domain from training data leads to a noticeable drop in its corresponding performance, confirming that domain-specific knowledge contributes directly to accurate argument classification. The largest declines are observed in Digital Humanities and Computational Biology, indicating that these domains contain more unique or specialized linguistic patterns that are not easily generalized from

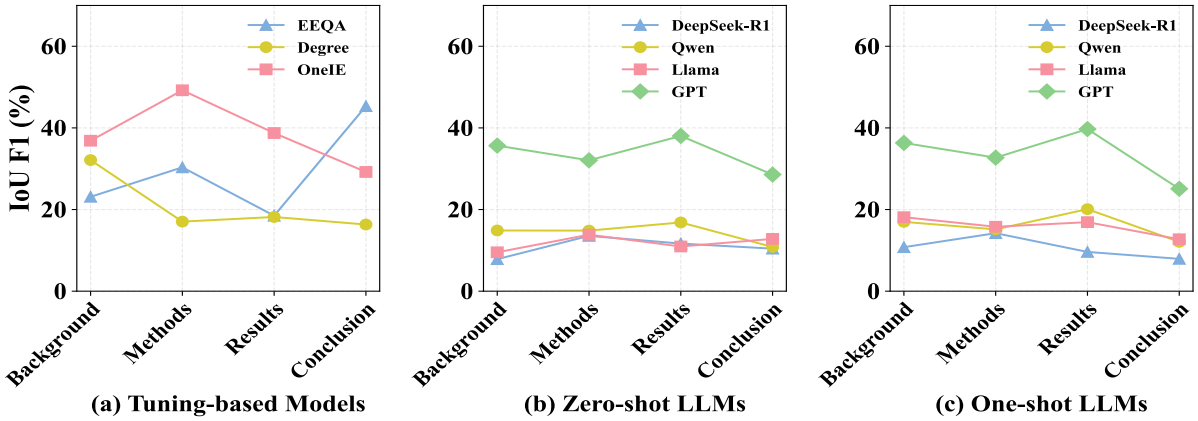


Figure 7: Comparison of Intersection-over-Union (IoU) on Arg-C F1-scores (%) across different event types for various models on *background*, *methods*, *results*, and *conclusions* events.

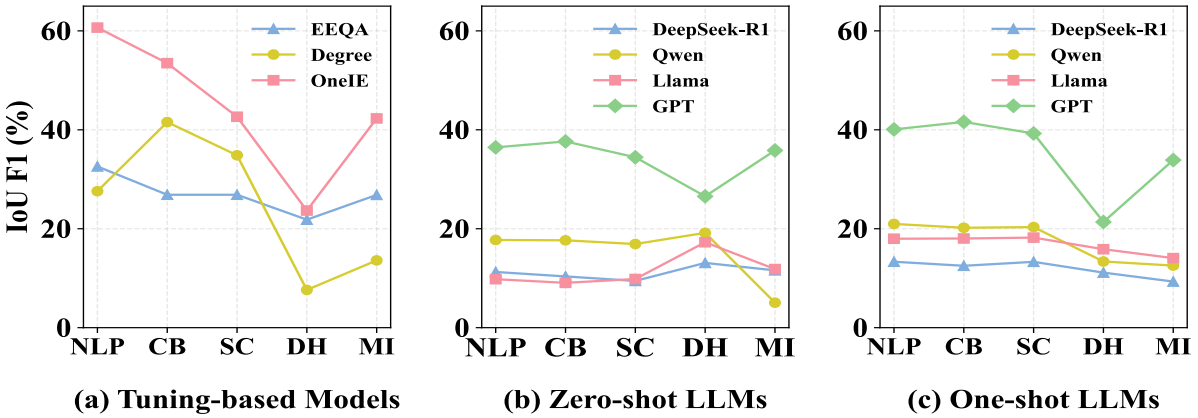


Figure 8: Comparison of Intersection-over-Union (IoU) on Arg-C F1-scores (%) across different academic domains for various models on Natural Language Processing (NLP), Computational Biology (CB), Social Computing (SC), Digital Humanities (DH), and Medical Informatics (MI).

other domains. In contrast, Medical Informatics shows relatively smaller drop, suggesting better generalizability or partial overlap with language patterns present in the other domains.

6 Conclusion

In this paper, we introduce SciEvent, a novel benchmark for SciIE across multiple domains. By framing scientific texts as a sequence of universal events and corresponding fine-grained argument roles, SciEvent provides a unified and domain-independent structure for representing scientific information. Specifically, We developed an annotation pipeline comprising event segmentation, trigger identification, and argument extraction, and defined three corresponding tasks: Scientific Event Segmentation, SciEvent Trigger Identification, and SciEvent Argument Extraction. Our benchmark covers five diverse domains with manual annotations, enabling robust evaluation of EE. Experiments on diverse state-of-the-art tuning-based EE systems and tuning-free LLMs show clear performance gaps (∼20%) between model predictions and human annotations, especially on argument extraction tasks. SciEvent serves as a challenging and realistic benchmark for advancing multi-domain scientific EE.

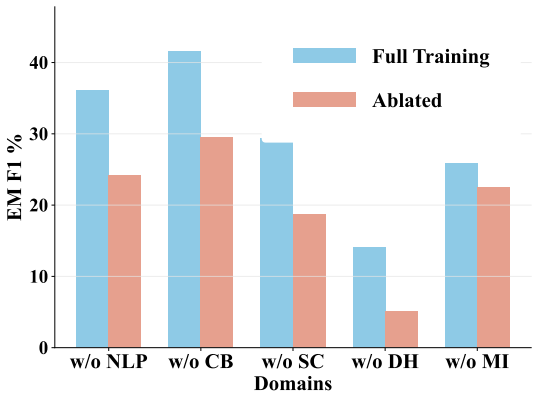


Figure 9: Arg-C F1-scores reported for full training versus training with one domain removed for OneIE under Exact Match (EM).

Limitations

One limitation of our work is the potential for data contamination in large language models, as our dataset is constructed from recent publications (mostly from 2023, and 2021–2023 for Digital Humanities), which may overlap with LLM pretraining corpora. This could inflate model performance and should be considered when interpreting results. Additionally, SciEvent is built on abstracts only, which, while concise and widely available, may omit key discourse elements found in full papers limiting applicability to document-level information extraction. In future work, we plan to extend SciEvent to include full papers to better support comprehensive scientific IE, and also consider more event types and arguments roles since the full paper can contain more information such as Assumptions.

Ethical Considerations

We provide details about compensation rate for annotators. We recruited eleven graduate students in total and provided a compensation rate of $12.80 per hour. This rate applied to both gold-standard annotation and human performance baseline annotations.

References

1992. *Fourth Message Understanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*.

Michael Alley. 1996. *The Craft of Scientific Writing*, 3rd edition. Springer, New York.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Jari Björne, Filip Ginter, Sampo Pyysalo, Jun'ichi Tsujii, and Tapio Salakoski. 2010. Scaling up biomedical event extraction to the entire PubMed. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 28–36, Uppsala, Sweden. Association for Computational Linguistics.

Claire Bonial, Julia Bonn, Kathryn Conger, Jena D. Hwang, and Martha Palmer. 2014. PropBank: Semantics of new predicate types. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3013–3019, Reykjavik, Iceland. European Language Resources Association (ELRA).

Nancy Chinchor and Beth Sundheim. 1993. MUC-5 evaluation metrics. In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. In *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 3586–3596, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Carmela Comito, Agostino Forestiero, and Clara Pizzuti. 2019. Bursty event detection in twitter streams. *ACM Trans. Knowl. Discov. Data*, 13(4).

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Xinya Du and Claire Cardie. 2020a. Document-level event role filler extraction using multi-granularity contextualized encoding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8010–8020, Online. Association for Computational Linguistics.

Xinya Du and Claire Cardie. 2020b. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077, Online. Association for Computational Linguistics.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Sonal Gupta and Christopher Manning. 2011. Analyzing the dynamics of research by extracting key aspects of scientific papers. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1–9, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Kata Gábor, Haïfa Zargayouna, Isabelle Tellier, Davide Buscaldi, and Thierry Charnois. 2016. Unsupervised relation extraction in specialized corpora using sequence mining. In *Proceedings of the XVIth Symposium on Intelligent Data Analysis (IDA 2016)*, pages 237–248, Stockholm, Sweden. Springer.

Song He, Xin Peng, Yihan Cai, Xin Li, Zhiqing Yuan, WenLi Du, and Weimin Yang. 2024. ZSEE: A dataset based on zeolite synthesis event extraction for automated synthesis platform. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1791–1808, Mexico City, Mexico. Association for Computational Linguistics.

Zhenzhen Hong, Logan Ward, Kyle Chard, Ben Blaiszik, and Ian Foster. 2021. Challenges and advances in information extraction from scientific literature: a review. *JOM*, 73(10):3383–3400.

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. DEGREE: A data-efficient generation-based event extraction model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1890–1908, Seattle, United States. Association for Computational Linguistics.

Kung-Hsiang Huang, Sam Tang, and Nanyun Peng. 2021. Document-level entity-based extraction as template generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5257–5269, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kung-Hsiang Huang, Mu Yang, and Nanyun Peng. 2020. Biomedical event extraction with hierarchical knowledge graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1277–1285, Online. Association for Computational Linguistics.

Tomoki Ikoma and Shigeki Matsubara. 2023. Paper recommendation using citation contexts in scholarly documents. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 710–716, Hong Kong, China. Association for Computational Linguistics.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7506–7516, Online. Association for Computational Linguistics.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. 2011. Overview of Genia event task in BioNLP shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 7–15, Portland, Oregon, USA. Association for Computational Linguistics.

Jin-Dong Kim, Yue Wang, and Yamamoto Yasunori. 2013. The Genia event extraction shared task, 2013 edition - overview. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 8–15, Sofia, Bulgaria. Association for Computational Linguistics.

Suchetha Nambanoor Kunnath, David Pride, Bikash Gyawali, and Petr Knoth. 2020. Overview of the 2020 WOSP 3C citation context classification task. In *Proceedings of the 8th International Workshop on Mining Scientific Publications*, pages 75–83, Wuhan, China. Association for Computational Linguistics.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online. Association for Computational Linguistics.

Maria Liakata, Suraj Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out (WAS 2004)*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Yubo Ma, Zehao Wang, Mukai Li, Yixin Cao, Meiqi Chen, Xinze Li, Wenqi Sun, Kunquan Deng, Kun Wang, Aixin Sun, and Jing Shao. 2022. MMEKG: Multi-modal event knowledge graph towards universal representation across modalities. In *Proceedings of the 60th Annual Meeting of the Association*

for Computational Linguistics: System Demonstrations, pages 231–239, Dublin, Ireland. Association for Computational Linguistics.

Himanshu Maheshwari, Bhavyajeet Singh, and Vasudeva Varma. 2021. SciBERT sentence representation for citation context classification. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 130–133, Online. Association for Computational Linguistics.

Meta AI. 2024. Introducing Llama 3.1: Our Most Capable Models to Date.

Suchetha N. Kunnath, David Pride, Drahomira Herrmannova, and Petr Knoth. 2021. Overview of the 2021 SDP 3C citation context classification shared task. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 150–158, Online. Association for Computational Linguistics.

Kiem-Hieu Nguyen, Xavier Tannier, Olivier Ferret, and Romaric Besançon. 2016. A dataset for open event extraction in English. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1939–1943, Portorož, Slovenia. European Language Resources Association (ELRA).

OpenAI. 2025. Introducing GPT-4.1 in the API.

Brian Paltridge. 2002. Thesis and dissertation writing: an examination of published advice and actual practice. *English for Specific Purposes*, 21(2):125–143.

Amir Pouran Ben Veyseh, Javid Ebrahimi, Franck Dernoncourt, and Thien Nguyen. 2022. MEE: A novel multilingual event extraction dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9603–9613, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Massimiliano Pronesti, Joao Bettencourt-Silva, Paul Flanagan, Alessandra Pascale, Oisin Redmond, Anya Belz, and Yufang Hou. 2025. Query-driven document-level scientific evidence extraction from biomedical studies.

Sampo Pyysalo, Tomoko Ohta, Makoto Miwa, Han-Cheol Cho, Jun'ichi Tsujii, and Sophia Ananiadou. 2012. Event extraction across multiple levels of biological organization. *Bioinformatics*, 28(18):i575–i581.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

AM Ripple, JG Mork, JM Rozier, and LS Knecht. 2012. Structured abstracts in medline: Twenty-five years later. `https://structuredabstracts.nlm.nih.gov/Structured_Abstracts_in_MEDLINE_Twenty-Years_Later.pdf`. Accessed: 2014-09-17.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: an atlas of machine commonsense for if-then reasoning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

Omar Sharif, Joseph Gatto, Madhusudan Basak, and Sarah Masud Preum. 2024. Explicit, implicit, and scattered: Revisiting event extraction to capture complex arguments. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12061–12081, Miami, Florida, USA. Association for Computational Linguistics.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ERE: Annotation of entities, relations, and events. In *Proceedings of the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98, Denver, Colorado. Association for Computational Linguistics.

Petros Stavropoulos, Ioannis Lyris, Natalia Manola, Ioanna Grypari, and Haris Papageorgiou. 2023. Empowering knowledge discovery from scientific literature: A novel approach to research artifact analysis. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 37–53, Singapore. Association for Computational Linguistics.

Chen-Tse Tsai, Gourab Kundu, and Dan Roth. 2013. Concept-based analysis of scientific literature. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, page 1733–1738, New York, NY, USA. Association for Computing Machinery.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. Web Download. LDC Catalog No. LDC2006T06.

Zhongqing Wang and Yue Zhang. 2017. A neural model for joint event detection and summarization. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 4158–4164. AAAI Press.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020. Aser: A large-scale eventuality knowledge graph. In *Proceedings of The Web Conference 2020*, WWW '20, page 201–211, New York, NY, USA. Association for Computing Machinery.

Qi Zhang, Zhijia Chen, Huitong Pan, Cornelia Caragea, Longin Jan Latecki, and Eduard Dragut. 2024. SciER: An entity and relation extraction dataset for datasets, methods, and tasks in scientific documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13083–13100, Miami, Florida, USA. Association for Computational Linguistics.

Yuchen Zhuang, Yinghao Li, Junyang Zhang, Yue Yu, Yingjun Mou, Xiang Chen, Le Song, and Chao Zhang. 2022. ReSel: N-ary relation extraction from scientific text and tables by learning to retrieve and select. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 730–744, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Detailed data analysis

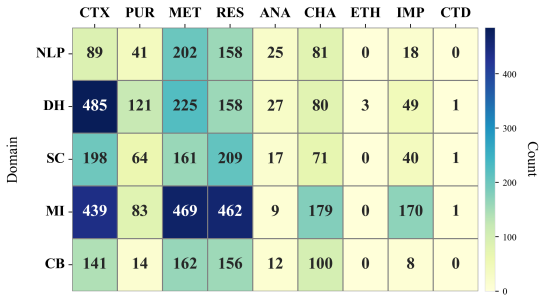


Figure 10: Distribution of argument types across all domains

We provided more details and insights into our SciEvent data. Figure 10 shows domain-wise distribution of argument types. DH focus on context, more than other domains. MI have nearly equal focus in context, method and result. Through all domains, context, method and results are the most commonly discussed arguments. Ethical, contradiction are really mentioned at all, and ethical only present in DH domain, showing its difference from other domains. It is also worth noting that DH abstracts are shorter than average (156.25 tokens vs. 207.72), while MI abstracts are considerably longer (359.41 tokens). To balance domain-specific content across the dataset, we sampled 60 DH abstracts and 40 MI abstracts accordingly.

B codebook details

B.1 Annotation Tool

We deploy our annotation tool on Render[8]. Figure 11 shows our annotation interface.

B.2 Event Type Definition

- **Background/Introduction:** Briefly outlines the context, motivation, and problem being addressed. It highlights the research gap and the paper's objectives or research questions.
- **Method/Approach:** Summarizes the methodologies, frameworks, or techniques used to conduct the study, including experimental setups, algorithms, datasets, or analytical tools.
- **Results/Findings:** Reports the main outcomes of the research, emphasizing key data, trends, or discoveries. Focuses on what was achieved or learned.
- **Conclusions/Implications:** Discusses the significance of the findings, their impact on the field, potential applications, and how they address the initial problem or research gap. May include recommendations or future research directions.

B.2.1 Trigger Definition

- **Action:** The most representative verb or verb phrase in the event, including auxiliary verbs like *am*, *is*, *are*, *have*, and *has*.
- **Agent:** The entity responsible for initiating or performing the Action, such as a person, system, method, or organization.
- **Object:** The entity that receives, is affected by, or is the focus of the Action, such as a concept, result, or entity being acted upon.

B.3 Argument Definition

- **Context:** Provides foundational or situational information of the event.
- **Purpose:** Defines the purpose or aim of the event.
- **Method:** Techniques, tools, methodology, or frameworks used in the event.
- **Results:** Observations or outputs of the event.

[8] https://render.com/

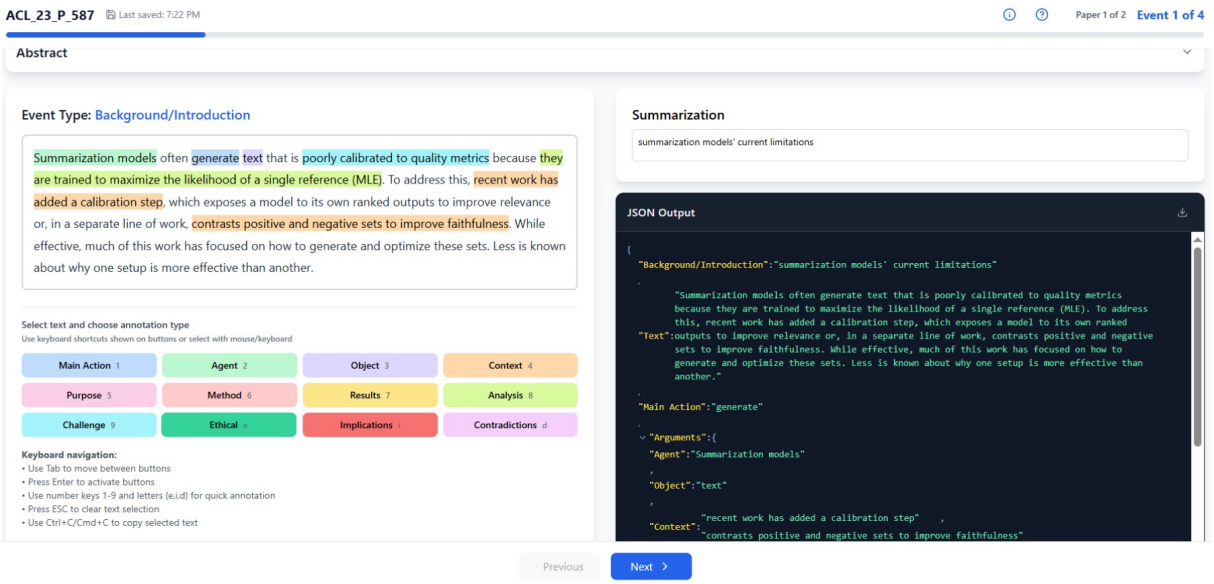


Figure 11: Annotation Tool Interface

- **Analysis:** Interpretation or explanation of other arguments.
- **Challenge:** Constraints or weaknesses of the context, method, or results.
- **Ethical:** Ethical concerns, implications, and justifications of the event.
- **Implication:** Broader applicability, significance, or potential for future research.
- **Contradiction:** Disagreements with existing knowledge.

B.4 Additional Annotation Rules

- **Annotate by Breaking Down Sentences:** Please annotate segments of a sentence (a part of a sentence) instead of a full sentence if different segments of the sentence can be fit into different arguments.
- **Passive Tense:** In a passive tense structure: Something (Agent) + is done (Passive Verb) + by Someone/Something (Object).
- **Indirect Object:** There is no direct object sometimes; you should leave the Object empty.
- **Entire Part of Sentence as Object:** In the following structure, the entire clause is the <Primary Object>: <Agent> + <Actions like: show, demonstrate, illustrate, prove, found, explain, indicate, conclude, etc.> + that / what / who / which / where / when / how / whether + clause.
- **Text that Fits Multiple Arguments:** If a text span can fit into multiple <Arguments>, follow this order of importance: **Results** > Purpose > Method > Analysis > Implication > Challenge > Contradiction > Context > Ethical. Results is the most important, and Ethical is the least.
- **Abbreviation:** You should use both the original term and its abbreviation, when both are given next to each other, e.g., Chain-of-Thought (CoT), not only Chain-of-Thought or CoT.

C Prompts

In this section, we present the prompt designs for each task. We include the Zero-Shot prompt for *Scientific Abstract Segmentation* and for *SciEvent Trigger Identification & Argument Extraction*, as well as the One-Shot prompt for *SciEvent Trigger Identification & Argument Extraction*.

Zero-Shot Scientific Abstract Segmentation Prompt

You are a strict extraction assistant. Never explain, never repeat, only extract in the required format.
 # # # Abstract: # # #
{abstract}
 # # # Extraction Rules: # # #

- Copy full, continuous sentences from the abstract. No changes, summaries, or guessing allowed.
- Each sentence must belong to only one section.
- Sections must use continuous text spans. No skipping around.
- If no content fits a section, output exactly <NONE>.
- No explanations, no extra text, no format changes.

 # # # Section Definitions: # # #

- **Background**: Problem, motivation, context, research gap, or objectives.
- **Method**: Techniques, experimental setups, frameworks, datasets.
- **Results**: Main findings, discoveries, statistics, or trends.
- **Implications**: Importance, impact, applications, or future work.

 # # # Exact Output Format: # # #

```
[Background]: <EXACT TEXT or <NONE>>

[Method]: <EXACT TEXT or <NONE>>

[Results]: <EXACT TEXT or <NONE>>

[Implications]: <EXACT TEXT or <NONE>>
```

Zero-Shot SciEvent Trigger Identification & Argument Extraction Prompt

You are an expert argument annotator. Given a part of a scientific abstract, you need to identify the key trigger for the event (the main verb or action that signals an important research activity) and annotate the abstract with the corresponding argument components related to this trigger. Extractions should capture complete phrases around this key trigger and be organized in a single JSON format, containing only what is explicitly stated in the text without adding any interpretation.

 # # # Abstract Segment to Analyze:

{abstract}

 # # # Argument Components to Extract:

Action: What is the SINGLE most representative trigger (verb or verb phrase) in the segment?
Agent: Who or what is performing this Action?
Object:

- **Primary Object:** What is directly receiving or affected by the Action?
- **Secondary Object:** What is a secondary entity also receiving the Action?

Context: What provides foundational or situational information of the event?
Purpose: What is the purpose or aim of the event?
Method: What techniques, tools, approaches, or frameworks are used in the event?
Results: What are the outcomes, observations or findings of the event?
Analysis: What are the interpretations or explanations of other arguments?
Challenge: What are the constraints or weaknesses of the event?
Ethical: What are the ethical concerns, justifications or implications of the event?
Implications: What is the broader significance or potential for future applications/research?
Contradictions: What are the disagreements with existing knowledge?

 # # # Extraction Rules:

1. Extract complete phrases, not just single words.
2. Only extract elements that are explicitly present. Mark missing elements as ["<NONE>"].
3. Use the exact text from the abstract.
4. Break down sentences when different parts fit different arguments.
5. NEVER use the same span of text for multiple arguments - each piece of text must be assigned to exactly one argument type. However, multiple text spans can be part of the same argument (e.g., ["text span 1", "text span 2".....] can be used for a single argument type) if different parts of the text contribute to the same argument.
6. If text could fit multiple arguments, prioritize in this order: Results > Purpose > Method > Analysis > Implication > Challenge > Contradiction > Context > Ethical

 # # # Output Format:

```
{
 "Action": "EXACT TEXT or <NONE>",
 "Agent": ["EXACT TEXT or <NONE>"],
 "Object": {
  "Primary Object": ["EXACT TEXT or <NONE>"],
  "Secondary Object": ["EXACT TEXT or <NONE>"]
 },
 "Context": ["EXACT TEXT or <NONE>"],
 "Purpose": ["EXACT TEXT or <NONE>"],
 "Method": ["EXACT TEXT or <NONE>"],
 "Results": ["EXACT TEXT or <NONE>"],
 "Analysis": ["EXACT TEXT or <NONE>"],
 "Challenge": ["EXACT TEXT or <NONE>"],
```

```
 "Ethical": ["EXACT TEXT or <NONE>"],
 "Implications": ["EXACT TEXT or <NONE>"],
 "Contradictions": ["EXACT TEXT or <NONE>"]
}
```

 # # # IMPORTANT INSTRUCTIONS:

- You MUST return ONLY ONE JSON structure.
- NO explanation text, thinking, or commentary before or after the JSON.
- NEVER repeat the JSON structure.
- ALL fields must use arrays with ["<NONE>"] for missing arguments.
- Follow the EXACT format shown in the template.
- ONLY extract arguments that are explicitly present in the text. DO NOT hallucinate or add any information not found in the abstract.

 # # # Output (JSON only)

One-Shot SciEvent Trigger Identification & Argument Extraction Prompt

You are an expert argument annotator. Given a part of a scientific abstract, you need to identify the key trigger for the event (the main verb or action that signals an important research activity) and annotate the abstract with the corresponding argument components related to this trigger. Extractions should capture complete phrases around this key trigger and be organized in a single JSON format, containing only what is explicitly stated in the text without adding any interpretation.
 # # # Abstract Segment to Analyze:
{abstract}
 # # # Argument Components to Extract:
Action: What is the SINGLE most representative trigger (verb or verb phrase) in the segment?
Agent: Who or what is performing this Action?
Object:

- **Primary Object:** What is directly receiving or affected by the Action?
- **Secondary Object:** What is a secondary entity also receiving the Action?

Context: What provides foundational or situational information of the event?
Purpose: What is the purpose or aim of the event?
Method: What techniques, tools, approaches, or frameworks are used in the event?
Results: What are the outcomes, observations or findings of the event?
Analysis: What are the interpretations or explanations of other arguments?
Challenge: What are the constraints or weaknesses of the event?
Ethical: What are the ethical concerns, justifications or implications of the event?
Implications: What is the broader significance or potential for future applications/research?
Contradictions: What are the disagreements with existing knowledge?
 # # # Extraction Rules:

1. Extract complete phrases, not just single words.
2. Only extract elements that are explicitly present. Mark missing elements as ["<NONE>"].
3. Use the exact text from the abstract.
4. Break down sentences when different parts fit different arguments.
5. NEVER use the same span of text for multiple arguments - each piece of text must be assigned to exactly one argument type. However, multiple text spans can be part of the same argument (e.g., ["text span 1", "text span 2".....] can be used for a single argument type) if different parts of the text contribute to the same argument.
6. If text could fit multiple arguments, prioritize in this order: Results > Purpose > Method > Analysis > Implication > Challenge > Contradiction > Context > Ethical

Here is a one-shot example of a complete abstract:
Background/Introduction Event
For abstract: "Second language acquisition (SLA) research has extensively studied cross-linguistic transfer, the influence of linguistic structure of a speaker's native language [L1] on the successful acquisition of a foreign language [L2]. Effects of such transfer can be positive (facilitating acquisition) or negative (impeding acquisition). We find that NLP literature has not given enough attention to the phenomenon of negative transfer."
Output:

```
{
 "Action": "has extensively studied",
 "Agent": ["Second language acquisition (SLA) research"],
 "Object": {
  "Primary Object": ["cross-linguistic transfer"],
  "Secondary Object": ["<NONE>"]
 },
 "Context": ["Effects of such transfer can be positive (facilitating acquisition) or negative (impeding acquisition)"],
 "Purpose": ["<NONE>"],
 "Method": ["<NONE>"],
 "Results": ["<NONE>"],
 "Analysis": ["<NONE>"],
 "Challenge": ["We find that NLP literature has not given enough attention to the phenomenon of negative transfer"],
 "Ethical": ["<NONE>"],
 "Implications": ["<NONE>"],
 "Contradictions": ["<NONE>"]
}
```


Methods/Approach Event

For abstract: "To understand patterns of both positive and negative transfer between L1 and L2, we model sequential second language acquisition in LMs. Further, we build a Multilingual Age Ordered CHILDES (MAO-CHILDES) — a dataset consisting of 5 typologically diverse languages, i.e., German, French, Polish, Indonesian, and Japanese — to understand the degree to which native Child-Directed Speech (CDS) [L1] can help or conflict with English language acquisition [L2]."
Output:


```
{
 "Action": "model",
 "Agent": ["we"],
 "Object": {
  "Primary Object": ["sequential second language acquisition in LMs"],
  "Secondary Object": ["<NONE>"]
 },
 "Context": ["<NONE>"],
 "Purpose": ["To understand patterns of both positive and negative transfer between L1 and L2"],
 "Method": ["we build a Multilingual Age Ordered CHILDES (MAO-CHILDES)"],
 "Results": ["<NONE>"],
 "Analysis": ["a dataset consisting of 5 typologically diverse languages, i.e., German, French, Polish, Indonesian, and
Japanese"],
 "Challenge": ["<NONE>"],
 "Ethical": ["<NONE>"],
 "Implications": ["<NONE>"],
 "Contradictions": ["<NONE>"]
}
```


Results/Findings Event

For abstract: "To examine the impact of native CDS, we use the TILT-based cross lingual transfer learning approach established by Papadimitriou and Jurafsky (2020) and find that, as in human SLA, language family distance predicts more negative transfer. Additionally, we find that conversational speech data shows greater facilitation for language acquisition than scripted speech data."
Output:


```
{
 "Action": "use",
 "Agent": ["we"],
 "Object": {
  "Primary Object": ["the TILT-based cross lingual transfer learning approach"],
  "Secondary Object": ["<NONE>"]
 },
 "Context": ["<NONE>"],
 "Purpose": ["To examine the impact of native CDS"],
 "Method": ["<NONE>"],
 "Results": ["as in human SLA, language family distance predicts more negative transfer", "conversational speech data
shows greater facilitation for language acquisition than scripted speech data"],
 "Analysis": ["<NONE>"],
 "Challenge": ["<NONE>"],
 "Ethical": ["<NONE>"],
 "Implications": ["<NONE>"],
 "Contradictions": ["<NONE>"]
}
```


Conclusions/Implications Event

For abstract: "Our findings call for further research using our novel Transformer-based SLA models and we would like to encourage it by releasing our code, data, and models."
Output:

```
{
 "Action": "call for",
 "Agent": ["Our findings"],
 "Object": {
  "Primary Object": ["further research"],
  "Secondary Object": ["<NONE>"]
 },
 "Context": ["<NONE>"],
 "Purpose": ["<NONE>"],
 "Method": ["using our novel Transformer-based SLA models"],
 "Results": ["<NONE>"],
 "Analysis": ["<NONE>"],
 "Challenge": ["<NONE>"],
 "Ethical": ["<NONE>"],
 "Implications": ["we would like to encourage it by releasing our code, data, and models"],
 "Contradictions": ["<NONE>"]
}
```

Output Format:

```
{
 "Action": "EXACT TEXT or <NONE>",
 "Agent": ["EXACT TEXT or <NONE>"],
  "Object": {
  "Primary Object": ["EXACT TEXT or <NONE>"],
  "Secondary Object": ["EXACT TEXT or <NONE>"]
 },
 "Context": ["EXACT TEXT or <NONE>"],
 "Purpose": ["EXACT TEXT or <NONE>"],
 "Method": ["EXACT TEXT or <NONE>"],
 "Results": ["EXACT TEXT or <NONE>"],
 "Analysis": ["EXACT TEXT or <NONE>"],
 "Challenge": ["EXACT TEXT or <NONE>"],
 "Ethical": ["EXACT TEXT or <NONE>"],
 "Implications": ["EXACT TEXT or <NONE>"],
 "Contradictions": ["EXACT TEXT or <NONE>"]
}
```

IMPORTANT INSTRUCTIONS:

- You MUST return ONLY ONE JSON structure.
- NO explanation text, thinking, or commentary before or after the JSON.
- NEVER repeat the JSON structure.
- ALL fields must use arrays with ["<NONE>"] for missing arguments.
- Follow the EXACT format shown in the template.
- ONLY extract arguments that are explicitly present in the text. DO NOT hallucinate or add any information not found in the abstract.
- Carefully study the one-shot examples to understand how arguments should be correctly annotated from the text.

Output (JSON only)

Methods	Arg-I (EM)			Arg-C (EM)		
	P	R	F1	P	R	F1
Tuning-based models						
EEQA	14.26	15.01	14.63	11.59	12.20	11.88
DEGREE	**44.97**	12.69	19.79	**34.23**	9.66	15.07
OneIE	32.03	**35.27**	**33.57**	25.38	**27.95**	**26.61**
Zero-shot LLMs						
DeepSeek-R1	10.33	5.46	7.15	6.23	3.30	4.31
Qwen	9.59	7.10	8.16	5.08	3.76	4.33
Llama	7.01	7.17	7.09	3.73	3.81	3.77
GPT	17.84	23.03	20.10	13.37	17.27	15.07
One-shot LLMs						
DeepSeek-R1	13.28	5.51	7.79	7.08	2.93	4.15
Qwen	13.98	9.16	11.07	7.24	4.74	5.73
Llama	13.02	9.93	11.27	6.55	5.00	5.67
GPT	25.75	25.79	25.77	19.38	19.41	19.4

Table 5: EM-based Precision (P), Recall (R), and F1-score (%) on baseline models for argument identification (Arg-I) and classification (Arg-C) tasks.

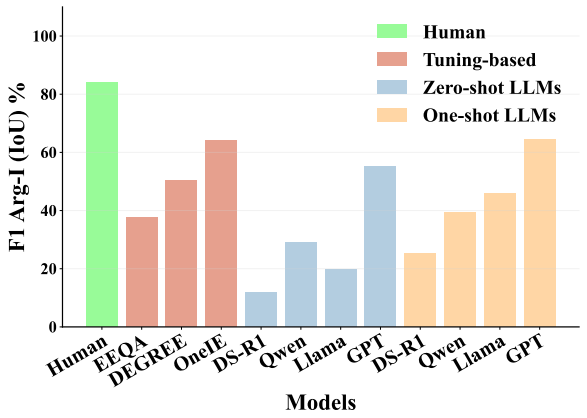


Figure 12: Performance comparison of various methods on argument identification (Arg-I) using IoU F1 scores. Methods are grouped by type: Human baseline, tuning-based models, zero-shot LLMs, and one-shot LLMs.

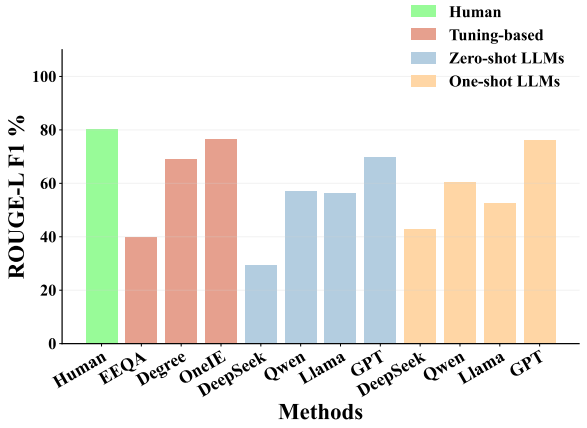


Figure 13: Performance comparison of various methods on ROUGE-L F1 scores. Methods are grouped by type: Human baseline, tuning-based models, zero-shot LLMs, and one-shot LLMs.

D Domain-wise Arguments Distribution Analysis

We also report the distribution of argument types across scientific domains (Figure 10). While all domains emphasize Results, Digital Humanities (DH) is a notable exception, being dominated by Context arguments. Among STEM domains—Natural Language Processing (NLP), Computational Biology (CB), and Medical Informatics (MI)—Method arguments are the most prevalent, reflecting their methodological focus. In contrast, DH and Social Computing (SC) place more emphasis on Context and Results, respectively, aligning with the rhetorical nature of these fields. Notably, MI contains the highest number of arguments overall, likely due to the length of its abstracts, even though fewer were annotated to balance domain coverage.

E SciEvent Argument Extraction with EM Metrics and detailed Human Performance Comparison

In Section 5, we analyzed argument extraction under the IoU metric and examined the human–model performance gap for argument classification (Arg-C) using IoU. Here, we complement that analysis by reporting results under the EM metric for argument extraction, as well as argument identification (Arg-I) and trigger identification with ROUGE-L human–model gaps, to provide a more comprehensive evaluation. As shown in Table 5, OneIE remains the best-performing model, while DEGREE continues to exhibit high precision but low recall. Among LLMs, GPT-4.1 consistently achieves the best performance, and one-shot prompting again improves results across all LLMs. Overall, the findings remain consistent—switching from IoU to EM does not alter the relative comparison between models, but EM results in lower scores for all models due to its stricter matching criteria.

Figure 12 shows the Arg-I performance gap between humans and models, which closely mirrors the Arg-C results. The gap remains around 20%, highlighting the need for multi-domain scientific EE models. In contrast, Figure 13 reveals a smaller gap in ROUGE-L scores for trigger identification, indicating that this task is considerably easier and most models perform well. Nevertheless, since argument extraction is the core challenge, there remains significant room for improvement in addressing multi-domain scientific EE.

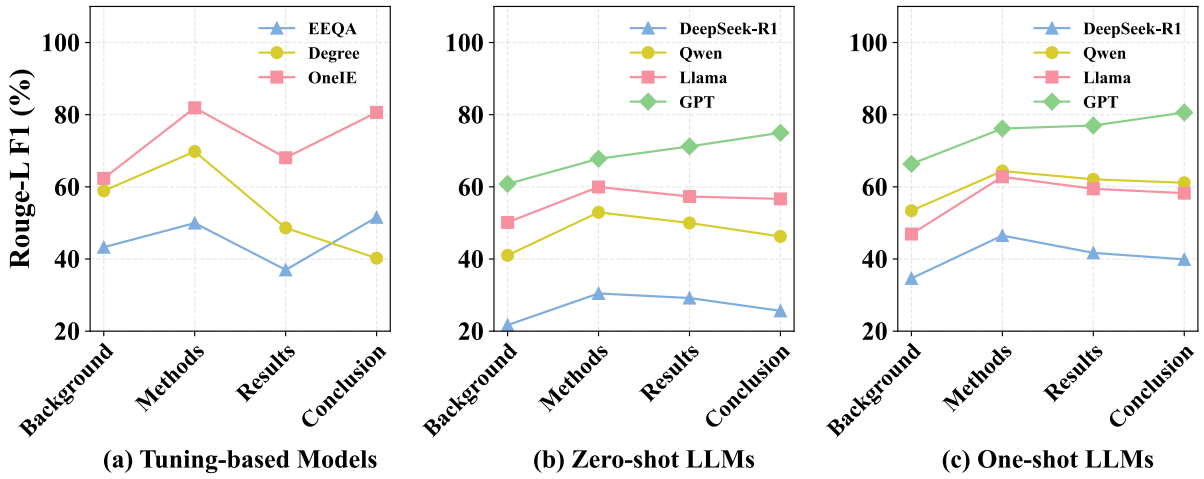


Figure 14: Comparison of Rouge-L F1 scores (%) across different event types for various models on Background, Methods, Results, and Conclusion sections of scientific papers.

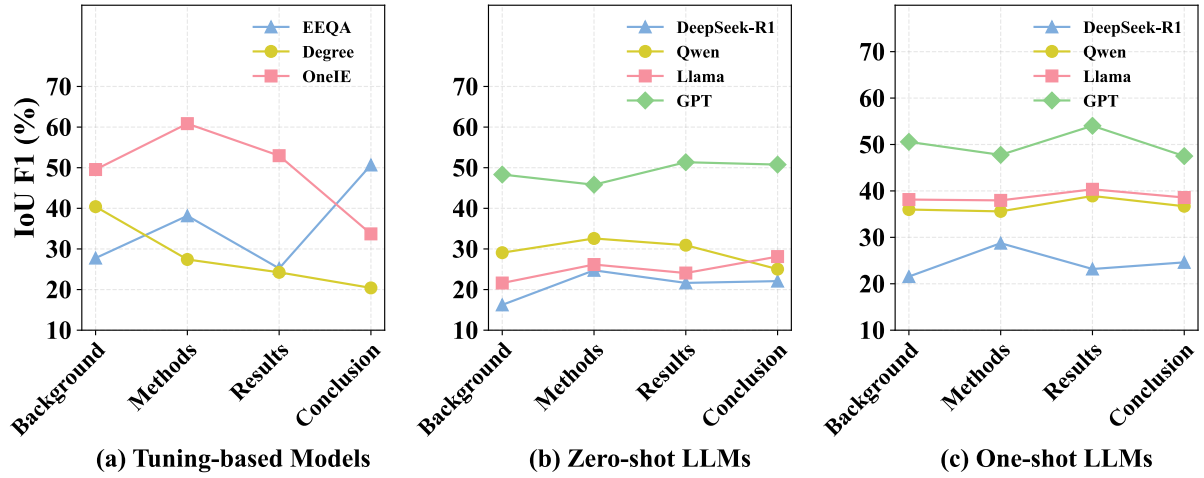


Figure 15: Comparison of Intersection-over-Union (IoU) on Arg-I F1-scores (%) across different event types for various models on Background, Methods, Results, and Conclusions sections.

F Trigger and Argument Identification by Event Types and Domains

Results for trigger identification and argument identification are presented by event type and domain, providing supplementary detail to the analysis in Section 5 and offering deeper insight into how event types and domains impact SciEvent performance.

Figure 14 presents ROUGE-L scores for trigger identification by event type, where the Conclusion event achieves the highest performance. This is likely due to its shorter and simpler structure, offering fewer candidate verbs, making trigger extraction easier. The performance trends for other event types are similar to those discussed in Sections 5 on argument classification. Figure 15 reports IoU scores for argument identification, which closely mirror the argument classification results but show an overall performance increase of about 20%, due to a looser matching metric. Figure 16 shows some difference in the Medical Informatics (MI) domain compared to argument classification. MI exhibits lower trigger identification performance, due to longer texts containing more verbs, which increases ambiguity and makes trigger extraction more difficult. Figure 17 again shows a 20% performance boost across all models, due to the looser IoU matching metric, while preserving trends consistent with those observed in argument classification.

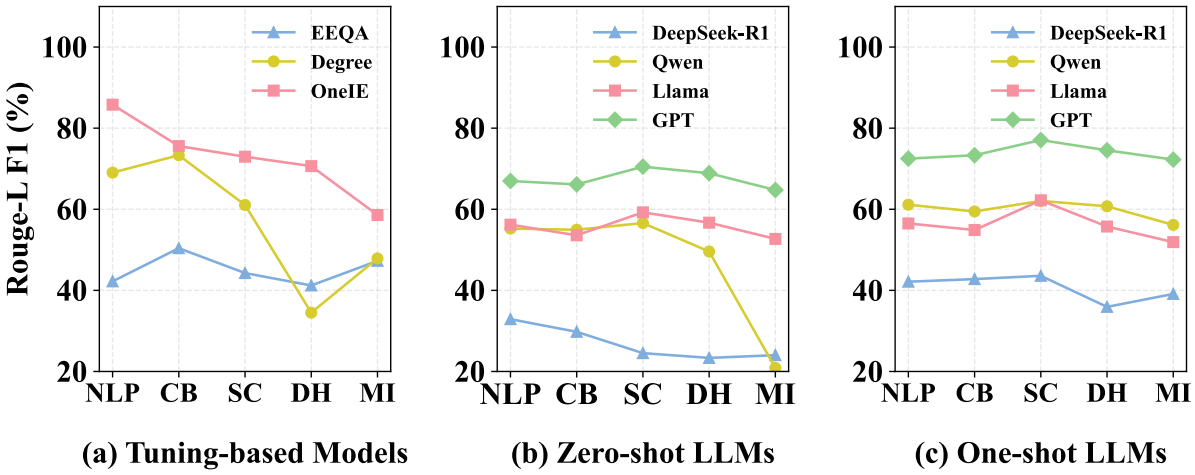


Figure 16: Comparison of Rouge-L F1 scores (%) across different academic domains for various models on Natural Language Processing (NLP), Computational Biology (CB), Social Computing (SC), Digital Humanities (DH), and Medical Informatics (MI).

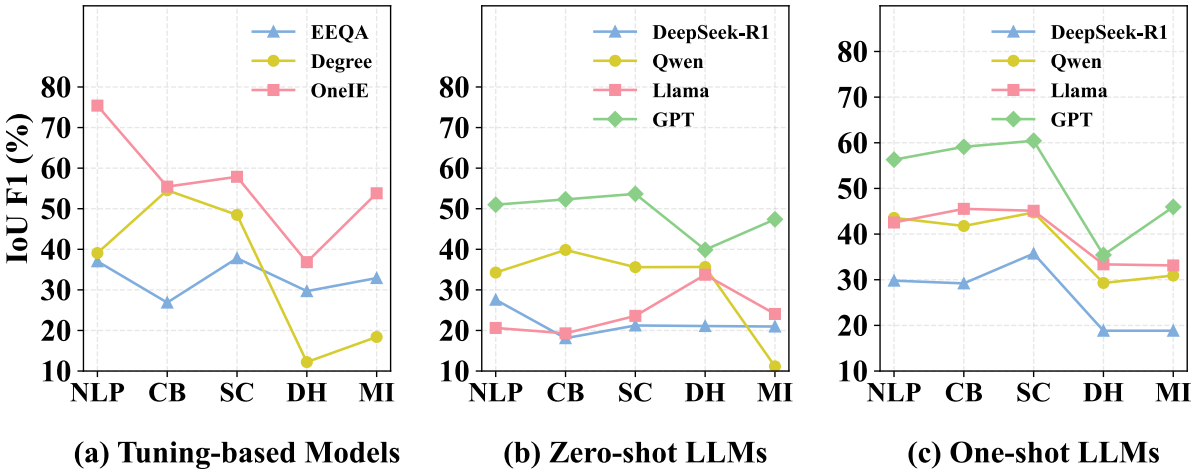


Figure 17: Comparison of Intersection-over-Union (IoU) on Arg-I F1-scores (%) across different academic domains for various models on Natural Language Processing (NLP), Computational Biology (CB), Social Computing (SC), Digital Humanities (DH), and Medical Informatics (MI).

G Effects of removal of domains on each tuning-based model

We present domain ablation results for DEGREE and EEQA under the EM setting in Figure 18 and Figure 19, respectively. For DEGREE, removing a domain consistently leads to performance drops, similar to OneIE, though the impact is generally smaller. This suggests DEGREE benefits from domain-specific training but is somewhat more resilient, possibly due to its generative nature. In contrast, EEQA shows minimal sensitivity to domain removal. This may be because its QA-based design relies more on question formulation and span selection, making it less dependent on domain-specific linguistic patterns.

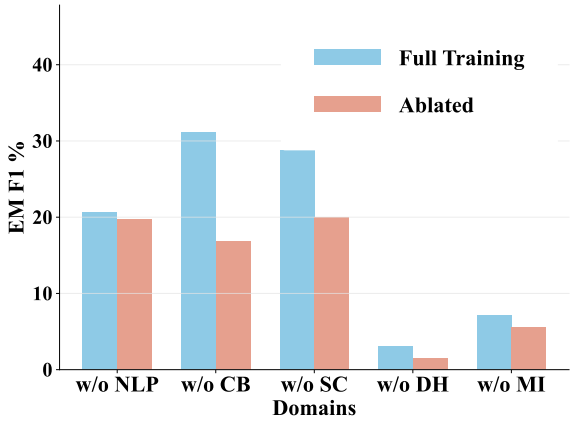


Figure 18: F1-scores reported for full training versus training with one domain removed for DEGREE under Exact Match (EM).

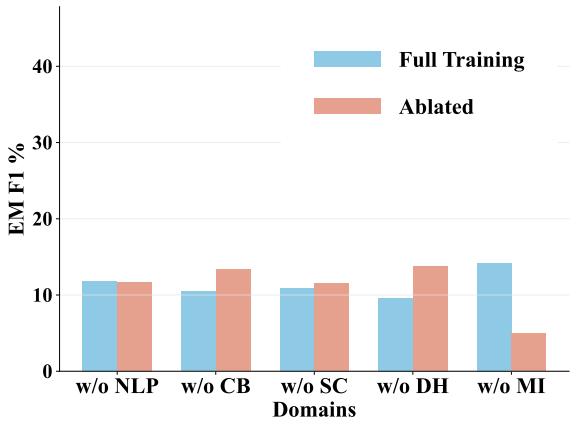


Figure 19: F1-scores reported for full training versus training with one domain removed for EEQA under Exact Match (EM).

H Detailed example of SciEvent dataset

We show one detailed example of SciEvent dataset, including event segmentation and event extraction in Figrue 20

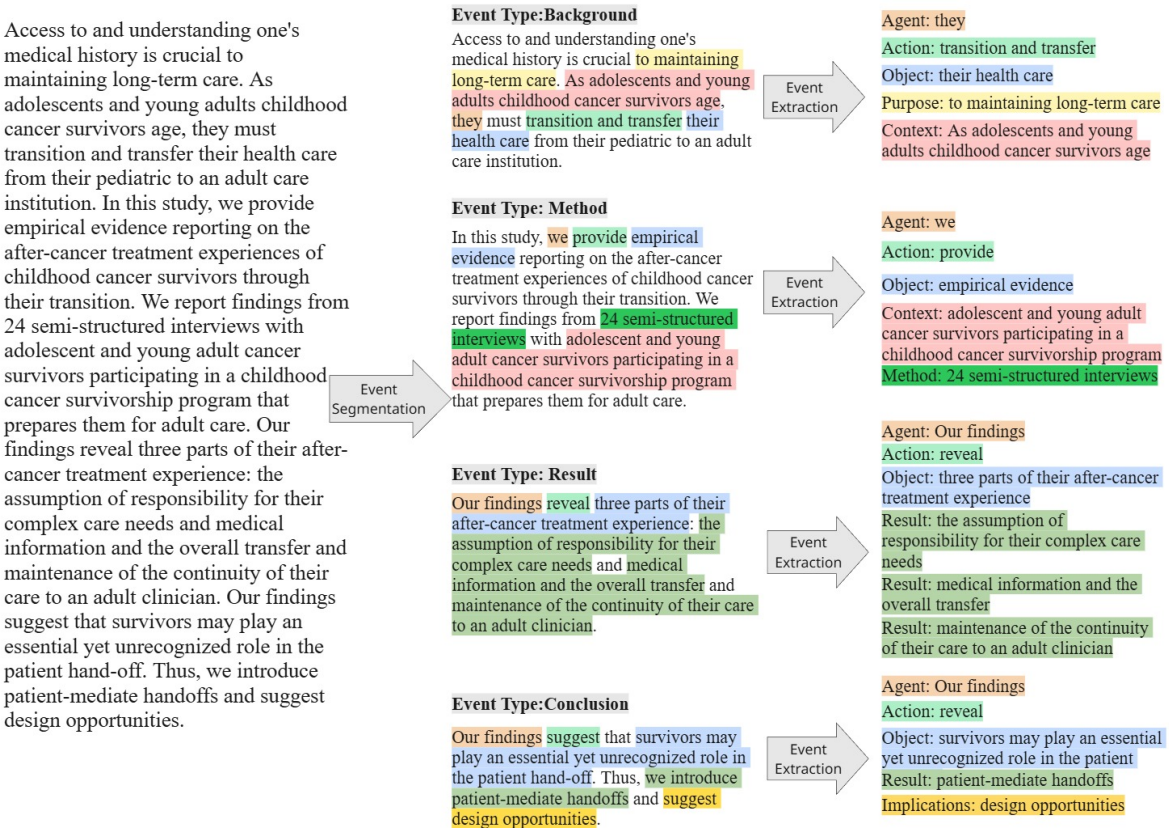


Figure 20: Full event extraction example from SciEvent, including event segmentation and event extraction, where trigger is a tuple including Agent, Action and Object.