AI Declarations in Scholarly Articles under the Microscope

Aftab Anjum[1][0000−0001−7656−7255], Steffen Lemke[1][0000−0002−3506−7083], Isabella Peters[1,2][0000−0001−5840−0806], and Ralf Krestel[1,2][0000−0002−5036−8589]

[1] Kiel University, Germany
[2] ZBW – Leibniz Information Centre for Economics, Kiel, Germany

Abstract. The growing use of generative AI (GenAI) tools in academic writing has created opportunities and challenges. While these tools can improve writing efficiency, readability, and overall workflow, their use raises concerns about accuracy, reliability, and transparency. In response, many publishers have introduced policies that ask authors to clearly disclose any use of AI during manuscript preparation. A particularly nuanced measure in this regard is the *CEUR-Workshop GenAI usage taxonomy* introduced at the end of 2024. In this study, we conducted a large-scale analysis of CEUR-WS proceedings publications (N = 5,562), as a sample of recent computer science literature, in order to examine how authors report their use of GenAI tools and for which tasks these tools are used in this domain. To this end, we applied LLM-based entity and relationship extraction methods. We found that 58% of CEUR-WS articles published between 2024 and 2026 reported using at least one GenAI tool. The tools mentioned most frequently were *ChatGPT* and *Grammarly*, though tool preferences varied by country. Interestingly, an additional analysis of recent non-CEUR-WS publications by the authors from our sample suggests that in different publication venues authors almost never disclose any forms of AI usage; only 1.8% of such publications contained any GenAI usage declaration.

Keywords: AI usage · large language models · text classification · scientific writing · research integrity · bibliometrics · scholarly communication · academic publications

1 Introduction

Although the often passionate debates of the academic community about the use of generative AI (GenAI) in manuscript preparation are not likely to come to a conclusion anytime soon, it seems safe to assume that AI-powered tools such as *ChatGPT*, *Claude*, *Copilot*, or *Grammarly* are by now used by many researchers for a wide range of associated tasks, e.g., content drafting, spell checking, paraphrasing, translating, image creation, coding, and many more. The potential benefits of GenAI for manuscript preparation are easy to see, such as improved readability and clarity (in particular, concerning texts by non-native speakers of the respective target language), easier handling of large volumes

of textual information and overall higher efficiency in the writing process [6]. Therefore, many researchers have already embraced GenAI in their academic workflows and consider it a useful tool (but not more). However, it remains difficult to estimate the extent and gravity of the negative effects that a long-reaching reliance on AI tools in manuscript writing may have on the quality of research publications, e.g., through the introduction of factual inaccuracies, false references, the perpetuation of biases or unintentional plagiarism [6,4,16,28].

Recognizing these dangers, within academia there seems to be widespread agreement that the use of GenAI tools in research processes must come with the utmost openness and transparency [26,29]. This is, for instance, reflected by several guidelines published by scientific publishers in reaction to the public attention towards large language models (LLMs) sparked by the release of ChatGPT in November 2022 [11,25,34]. Through these guidelines, publishers urge authors to precisely describe any AI-based tools used during manuscript creation, for instance, in the Methods section, in an appendix, or right before the Reference section of the manuscript, as well as to publish relevant prompts. Further positions that most scholarly publishers appear to have quickly agreed on are that (1) AI tools should never be listed as manuscripts' authors, as this role involves rights and responsibilities only applicable to humans; (2) the outputs of AI tools should never be transferred to scholarly manuscripts without careful human review; and (3) that there are certain key research tasks, such as interpretation of results or drawing of conclusions, that may never be handed over to an AI tool [26].

To achieve even greater transparency, CEUR-WS.org; a free open-access publication service for workshop proceedings, mainly in computer science; introduced a structured disclosure requirement in late 2024. Submitters must now include a Declaration of Generative AI in their manuscripts, describing their GenAI use according to a taxonomy of 13 predefined tasks.[3] Additionally, the authors have to declare that they "take full responsibility for the publication". Although the taxonomy aids in standardizing how the use of GenAI is reported, the following exemplary statements from CEUR-WS publications reflect the various levels of detail and transparency the authors include in their declarations:

> "During the preparation of this work, the authors utilized Grammarly to check grammar and spelling, as well as to paraphrase and reword. They also used Julius AI for the data analysis. After using this tool/service, the authors reviewed and edited the content as needed [...]. No sections in the experimental code or manuscript have been created using Generative AI tool(s)/service(s)." [22]

> "During the preparation of this work, the authors used Grok 3 for rephrasing paragraphs in the introduction and discussion sections and Gemini for editing code. After using these tools, the author reviewed and edited the content as needed [...]" [37]

[3] CEUR-WS GenAI usage taxonomy https://ceur-ws.org/GenAI/Taxonomy.html

Despite the aforementioned measures to improve the traceability of AI usage in manuscript preparation processes, it remains unclear to what extent authors actually disclose their real AI use. In this study, we conduct a large-scale analysis of CEUR-WS proceedings using PDF Scraping and LLM-based entity and relation extraction to examine which GenAI tools authors report using and for which tasks. We analyze these patterns in relation to author metadata, including country affiliation. As a side question, we explore whether authors who disclose AI usage in CEUR-WS articles also do so in articles that they recently published at other venues, where they might not be explicitly requested to do so (although they arguably still should, as for the reasons mentioned above).

In this way, this study pursues four objectives. First, we provide an empirical estimation of how GenAI tools are currently used in computer science manuscript writing (while acknowledging the limitation that declared use does not necessarily equal actual use). Second, we evaluate which tasks defined in the CEUR-WS taxonomy are most commonly reported, offering an empirical perspective on the taxonomy's relevance. Third, we contribute a manually annotated dataset of 300 publications, labeled according to a structured annotation scheme that covers tool mentions, usage descriptions, and their mapping to CEUR-WS taxonomy roles, which we use to evaluate the quality of our LLM-based extraction pipeline. Fourth, by examining the behavior of the same authors in venues other than CEUR-WS, we explore whether structured disclosure requirements shape author behavior more broadly, or whether AI transparency is driven primarily by venue-specific policies. [4]

2 Related Work

Our related work section is split into related technical approaches to extract information from PDF documents and related work on the adoption and use of GenAI in the preparation and publication of scholarly literature.

2.1 Information Extraction from Scholarly Literature

Extracting information from scientific papers has been an active research field for many years now. With the rise of transformer-based architectures, new opportunities have arisen. Several datasets, models, and frameworks have been released to find different types of entities in scientific texts, such as tasks, models, metrics, and data sources, as well as the relationships between them.

SemEval 2017 [2] and SemEval 2018 [10] provided paragraph level datasets with 500 paragraphs annotated for tasks, methods, and materials. Luna et al. 2018 [19] introduced the SCI-ERC dataset, which covers 500 abstracts and includes annotations for entity recognition, relation extraction, and co-reference resolution. SCIREX [15] and DMDD [24] addressed document level extraction.

[4] Code, dataset, additional results and all prompts used in this study are publicly available in https://github.com/aftabcau/Gen_AI.

More recently, GSAP [23] provided fine-grained NER annotations for 100 full ML papers, covering ten types of entities and capturing informal and nested mentions.

In terms of model or framework design on entity extraction, early work mainly used rule-based systems [13,17] and models built with hand-crafted features. But with the rise of deep learning, sequence-labeling models that combine bi-directional LSTMs with CRFs [33] became widely used because they can capture local context and character-level patterns. However, these models struggled with entities related to research that are often complex or nested. Transformer-based pre-trained language models [18,9] changed the field significantly. Beltagy et al. introduced SciBERT [3] and showed that pre-training on large scholarly text corpora greatly improves performance compared to general-domain models such as BERT [7], for tasks such as NER and relation extraction. Similarly, DyGIE++ [31] used dynamic span graphs to update entity representations using relation information on datasets such as SciERC. Furthermore, span-based models such as SpERT [9] and the PURE framework [36] showed that lighter architectures can achieve similar performance with lower computational costs.

The field is also exploring how large language models can perform few-shot or zero-shot information extraction at the document and paragraph level [32,35,8,12]. Instead of relying on task-specific supervised training, these methods use prompting and in-context learning to identify entities and relations, making them particularly attractive for domains where labeled data are scarce or expensive to obtain.

2.2 Analyses of AI Use in Scholarly Writing

Compared to the plethora of recent research articles discussing the use of AI in research processes from a consultative, cautionary, or normative point of view, there is relatively little empirical evidence on the actual prevalence of AI assistance in the creation of scholarly publications. Miller et al. [20] used AI detection software to analyze 390 MEDLINE-abstracts of randomized controlled trials indexed between January 2020 and March 2023, finding that the prevalence of texts with a high probability of being AI-generated had increased from 21.7% to 36.7% during the study period. They conclude that the use of AI-tools in scholarly publishing has been increasing in recent years, even before tools like ChatGPT achieved widespread adoption. Raman [27] performed a semi-manual review of 1,226 articles published between November 2022 and July 2023 that mentioned generative AI in their acknowledgments, coding the articles based on the respective kind of attribution. They found Biomedical and Clinical Sciences to be the most prevalent fields among their set, followed by Information and Computing Sciences, but overall observed a widespread use of technologies like ChatGPT across diverse fields of research. Mishra et al. [21] surveyed 226 medical and paramedical researchers from 59 countries on their perceptions and use of LLMs related to academic work, among other aspects. They found 16.4% of their respondents to have used LLMs previously in publications, most prevalently to fix grammatical errors or improve formatting. However, most of those

respondents did not acknowledge such use in their publications. In a similar vein, Chemaya and Martin [4] surveyed 271 academics about whether they consider it necessary to report ChatGPT use in manuscript writing, with most respondents holding the opinion that AI use for minor textual improvements, such as fixing grammar mistakes, does not require to be reported. However, considering more substantial uses of ChatGPT, they found more disagreement among their respondents about the perceived necessity of reporting. Tang and Eaton [30] performed a bibliometric analysis on articles containing a phenomenon they describe as "AIGC (AI-generated content) footprints", i.e., fairly clear traces of the use of generative AI that were most likely overlooked by the respective articles' authors. In their small-scale investigation, they found that articles containing such footprints in most cases had been published by single authors from countries in Central Asia, South Asia, or Africa and in academic journals that were not indexed by Web of Science, Scopus, and the DOAJ. In their large-scale analysis of 5.2 million articles from 5,114 journals, He and Bu [14] estimated that only a minuscule share of about 0.1% of recent research papers explicitly disclosed AI use, revealing a substantial gap between AI disclosure and actual AI content proportions as estimated by applying a combination of four different AI detection methods. They conclude that the disclosure policies issued by many academic journals largely fail in promoting transparency about AI use in academic writing. Similar conclusions can be drawn from the analysis of Ans et al. [1], who in their analysis of 2,046 recent empirical articles from leading medical education journals found only 2.5% of them including an AI disclosure statement.

3 Dataset Construction

In the following, we describe how the dataset was built and how the construction process was evaluated.

3.1 PDF Scraping

To extract information about GenAI use in scholarly publications, together with associated author metadata, we built the framework illustrated in Figure 1. Our dataset consists of 5,562 PDF files scraped from CEUR-WS proceedings published between 2024 and 2026 (2024: N=601; 2025: N=4,850; 2026: N=111). We only included articles that contained a Declaration on Generative AI, excluding papers where no such declaration was present. The share of articles reporting GenAI use increased across the three years: 32.9% in 2024 (198 out of 601), 61.0% in 2025 (2,957 out of 4,850), and 65.8% in 2026 (73 out of 111). Note that the 2026 figures reflect a partial year, as data collection was conducted in early 2026.

After scraping the PDF files, we used Marker[5] to parse the PDFs into text (Markdown) files. After performing some data cleaning, such as removing equations and HTML code, we passed each article's first page of text to an LLM

[5] https://github.com/cuuupid/cog-marker

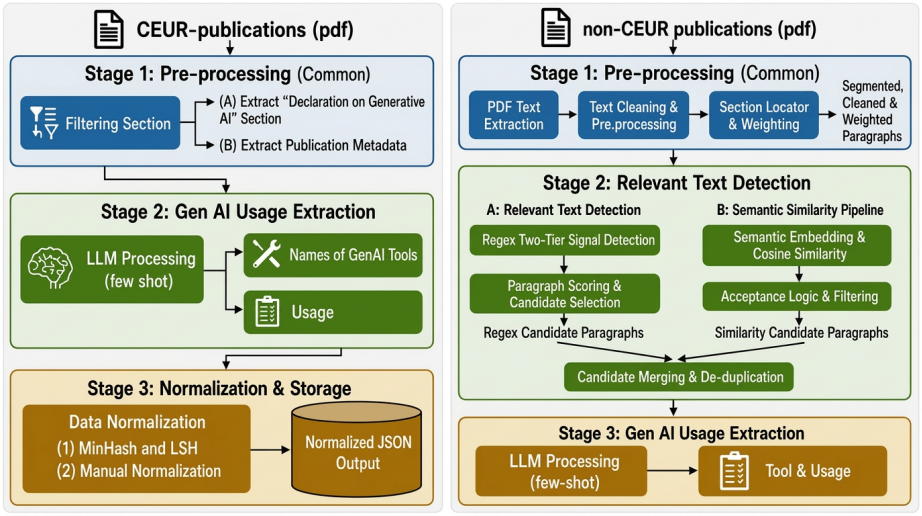


Fig. 1: Framework for extracting metadata and GenAI declarations.

(llama-3-70b-instruct) to extract metadata including authors' names, affiliations, cities, countries, ORCIDs, and email addresses. To ensure consistency across metadata fields, we normalized the data using MinHash and locality-sensitive hashing (LSH), which group semantically similar mentions based on token overlap. Afterwards, we performed manual normalization by constructing a dictionary mapping standardized names to their observed variants. For example, country mentions such as "USA," "U.S.A.," and "United States of America" were grouped and normalized to the standardized country name "United States". We manually evaluated this step by spot checking the extracted metadata for 50 randomly selected papers. Since this information is present in a structured form at the beginning of each PDF, the LLM had no trouble correctly extracting these metadata. Therefore, a more thorough evaluation was not indicated.

3.2 Extraction of GenAI Declarations

To extract GenAI declaration text from each article, we developed a two-method rule-based pipeline. The first method uses a set of 20 regular expression heading variants to handle the considerable variation in how authors titled this section, covering forms such as "Declaration on Generative AI", "GenAI Declaration", and "AI Tools Disclosure", as well as common OCR artefacts. The second method is a line-based fallback, applied when the first method finds no match, which scans the full document line by line to locate and extract the declaration. Both methods apply the same validation logic to classify each extracted declaration as positive (GenAI was used) or negative (GenAI was not used). The extracted declaration text was then processed by an LLM (llama-3-70b-instruct) in a few-shot setting and fine-tuned BERT model to identify the specific GenAI

tools used and the tasks for which they were used, following the CEUR-WS taxonomy. The resulting outputs were combined with the extracted author metadata and stored in JSON format, the whole process can be seen in Figure 1.

We again used llama-3-70b-instruct as LLM and fine-tuned BERT to extract GenAI tools and their contribution roles, i.e., for which tasks the GenAI tools were used. The complete prompt including the few-shot samples used to query the LLM can be found at https://github.com/aftabcau/Gen_AI. Similarly, as with the extracted metadata, extracted GenAI tool mentions and contribution role mentions (i.e., for which tasks the tools were used) were normalized using MinHash and LSH. Afterwards, we performed manual normalization by constructing a dictionary mapping standardized names to their observed variants.

Annotated data: To evaluate the quality of the LLM extraction, we randomly sampled 300 publications and had them independently annotated by two annotators with backgrounds in computer science and AI. Before annotation, both annotators completed a training session based on shared guidelines available in the code repository. Among the 300 publications sampled, 159 declared use of GenAI tools and 141 declared no use.

The annotation scheme consists of three entity types and two relation types, as summarized in Table 1. Annotators first identified *TOOL* entities (the name of a GenAI tool) and *USAGE* entities (the phrase describing how the tool was used). Each *USAGE* was then mapped to a predefined label from the CEUR-WS taxonomy (named contribution role), forming the *TAXONOMY_ROLE* entity. Two relations connect these entities: *USED_FOR* links each *TOOL* to its *USAGE*, and *MAPS_TO* links each *USAGE* to its *TAXONOMY_ROLE*. For our further analyzes, only annotations where both annotators agreed were retained. Inter-annotator agreement was measured using Cohen's κ [5], with individual scores per label reported in Table 1. The overall weighted Cohen's κ is 0.81, which indicates substantial agreement, confirming the reliability of the annotation. The annotated data statistics can be seen in Table 1.

Table 1: Statistics for Human-Annotated Data

Type	**Label**	**Count**	**Cohen's κ**
Entity	*TOOL*	246	0.94
	USAGE	309	0.85
	TAXONOMY_ROLE	309	–
Relation	*USED_FOR*	412	0.76
	MAPS_TO	309	0.75

Table 2: Performance on GenAI Tool & Usage Detection (LLM vs. BERT)

	LLM (Few-shot)			BERT		
Task	**P**	**R**	**F1**	**P**	**R**	**F1**
T1: AI Tool Detection	1.00	1.00	1.00	1.00	1.00	1.00
T2: Tool (E)	0.89	0.95	0.92	0.94	0.96	0.95
T2: Tool (P)	0.89	0.95	0.92	0.98	1.00	0.99
T2: Usage (E)	0.78	0.80	0.79	0.86	0.91	0.88
T2: Usage (P)	0.90	0.93	0.92	0.92	0.98	0.95
T3: `used_for` (E)	0.74	0.78	0.75	0.68	0.79	0.73
T3: `used_for` (P)	0.82	0.87	0.84	0.70	0.79	0.74

E = Exact match, P = Partial match

3.3 Evaluation of Extraction Method

To evaluate the performance of both the LLM and BERT model in detecting GenAI tools and their contribution roles in scholarly papers, we employed a three-task evaluation framework. Performance is measured using precision, recall, and F1 score, computed under two matching criteria: exact match and partial match. Token-level predictions and ground truth labels are first converted into entity spans by identifying contiguous sequences marked with ”B-“ (begin) and ”I-“ (inside) tags. An exact match is registered when a predicted entity span matches a gold (true) entity in both its span boundaries and entity type. In contrast, a partial match is counted when a predicted and a gold entity share the same type and exhibit any degree of span overlap, even if their boundaries do not align precisely. This dual evaluation strategy enables a comprehensive assessment of the model's ability to both identify entity boundaries and correctly classify entity types. Task 1 was a binary classification at the publication level, determining whether a publication mentions any AI tool at all; a simple yes or no decision per publication. Task 2 was a multi-label classification task covering both tool entity and usage entity extraction. For tool entities, the model had to identify which specific tools were mentioned in a paper (e.g., ChatGPT, Grammarly, DeepL), while for usage entities it had to extract the corresponding free-text descriptions of how those tools were used (e.g., "paraphrasing and rewording" or "text translation"). Task 3 evaluated the used_for relation, which links each identified tool entity to its corresponding usage entity.

The results are shown in Table 2. To ensure a fair comparison, the BERT model was trained on 170 samples, validated on 30 samples, and evaluated on a held-out test set of 100 samples. The same 100 test samples were then used to measure the LLM performance, ensuring that both models are evaluated under identical conditions on unseen data. We can see that Task 1 (binary classification) for tool detection performs very well, with an F1-score of 1.00 for both models, showing that both systems accurately identify whether a scholarly paper men-

tions any GenAI tools or not. For Task 2, BERT outperforms the LLM on both tool and usage entity extraction under exact and partial match. The consistent improvement from exact to partial match for both models confirms that some predictions capture the correct content but with slightly imprecise boundaries. For Task 3, the USED_FOR relation is evaluated using micro-averaged precision, recall, and F1, computed at the individual relation instance level across all tool-paper pairs in the test set. Here the LLM slightly outperforms BERT; under exact match the LLM achieves F1 of 0.75 vs 0.73 for BERT, and under partial match 0.84 vs 0.74. The recall of 0.78 under exact match and 0.87 under partial match for the LLM USED_FOR relation indicates that the model misses some tool-to-usage links, likely because correctly associating each tool with its specific usage description in academic writing is challenging due to variations in writing style and phrasing across papers. The higher recall under partial match confirms that many of these missed cases involve slight wording differences in the usage description that are too strict to pass exact matching. Overall, BERT is stronger at precise entity extraction while the LLM has a slight edge in relation linking, likely because the LLM leverages broader language understanding to interpret how tools are described and linked to their usage in academic writing. In this work, the BERT model was not trained on the MAPS_TO relation, as our primary focus was to evaluate how well BERT can extract tool and usage entities directly from text. However, it is worth noting that in our annotated dataset, the same text span is labeled both as a usage entity and as a predefined contribution category from the CEUR-WS taxonomy. This means the dataset is designed to support both directions of exploration; the MAPS_TO relation is included in the annotations so that future work can also train and evaluate a BERT model for mapping usage entities to predefined contribution categories without requiring any additional annotation effort. For the LLM, we did evaluate the maps_to relation and it achieved an F1 score of 0.85 under exact match, demonstrating that the LLM is capable of correctly assigning usage descriptions to their predefined contribution categories in most cases.

3.4 Declarations in Non-CEUR-WS Publications

To answer our side question of whether CEUR-WS authors also disclosed GenAI usage in their other recent publications, we used the Scopus and OpenAlex databases to identify additional publications by researchers who also contributed to the articles scraped from CEUR-WS Proceedings (i.e., our main analysis dataset). Importantly, we considered all authors listed on the CEUR-WS papers when constructing the author set. To ensure accurate author-level matching, we first removed duplicate author records using ORCID identifiers whenever available, as ORCID provides a unique and reliable identifier. Records sharing the same ORCID were merged into a single author entry, and missing metadata were completed from other matching records. For authors without an ORCID, we applied normalized name matching (case-insensitive) as a best-effort strategy. When identical names appeared with different institutional affiliations and no ORCID was available, these cases were conservatively treated as the same

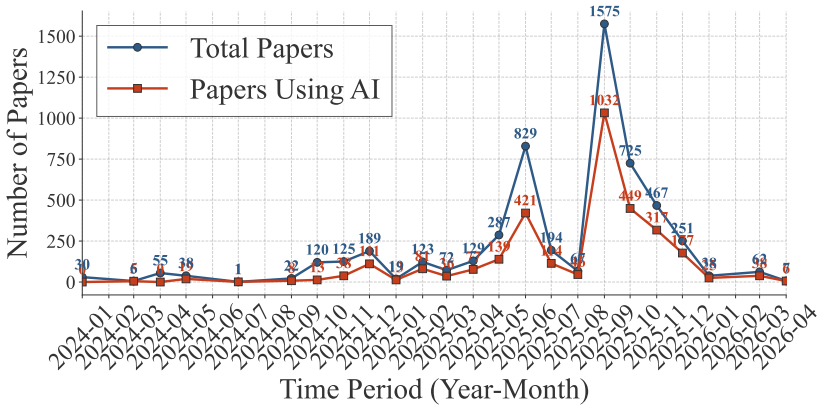


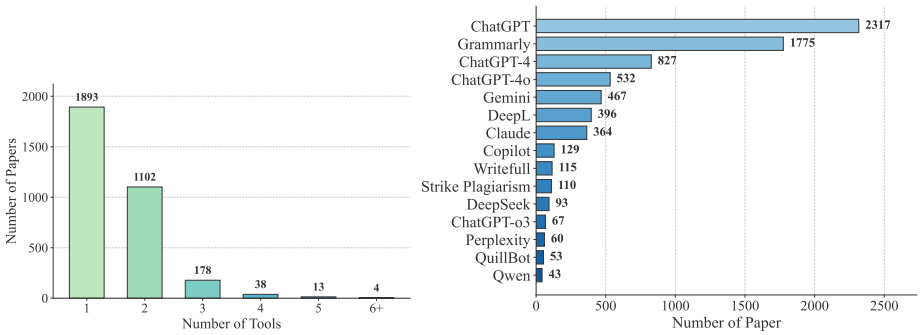
Fig. 2: Declared use of GenAI tools over time

author. To evaluate the potential impact of this conservative matching decision, we performed an additional ambiguity analysis on the non-CEUR publication dataset. Among 902 unique author mentions, 105 authors had an ORCID identifier available, while the remaining authors were identified using their names and affiliations. We identified only five cases where identical author names from different affiliations were merged. All five cases occurred within the same country and, therefore, do not affect our country-level analysis based on institutional affiliations. We then employed a retrieval process combining authors' names and, where available, their ORCID identifiers to obtain publications from 2022 onward. For each publication, we attempted to retrieve the full text through OpenAlex and Scopus. In cases where full texts were not directly available, we additionally used the "Find Full Text" feature in *Zotero*. Using this combined workflow, we collected 15,409 publications, which we then analyzed for GenAI usage disclosures using the same procedure applied to the CEUR-WS dataset in stages 2 and 3.

4 Analysis of AI Use in CEUR-WS Publications

Using our LLM-based framework illustrated in Figure 1, we examined both CEUR-WS and non-CEUR publications for statements about the (non-)use of GenAI. We identified 5,562 CEUR-WS publications that made such statements. 58.26% of the papers report using GenAI tools, with a total of 129 unique tools identified. Most GenAI-using papers reportedly employ a single tool, on average 1.51 tools were declared per paper with a maximum of nine tools. These publications represent 2,865 institutions from 120 countries, indicating that GenAI is widely used in scholarly writing around the world.

In Figure 2 we can see the temporal distribution of publications for our sample of CEUR-WS publications. The blue line gives the total number of papers that contain a declaration of GenAI usage (yes/no), while the red line shows



(a) GenAI tools declared per paper (b) Most often declared GenAI tools

Fig. 3: Statistics of declared GenAI tools usage

the number of papers declaring that GenAI was used in the process of writing the papers. The observed temporal pattern aligns closely with the CEUR-WS policy timeline. A GenAI disclosure guideline was introduced in December 2024, initially as a non-mandatory recommendation. A formal policy took effect on 1 January 2025, followed by mandatory disclosure enforcement from 1 June 2025. Before these policies, AI declarations were almost absent (e.g., 6 papers in February 2024 and 1 in May 2024). A gradual increase appears in late 2024 and early 2025, when disclosure was voluntary. After June 2025, both the number of declarations and the reported use rose sharply, showing a clear effect of policy enforcement. We see a strong effect in August 2025, with 1,575 papers declaring AI use and about 65% reporting active GenAI use (1,032 papers), followed by about 62% in September 2025 and 71% in November 2025. The April 2026 data (7 papers) is likely incomplete and not a real decline.

We also analyzed which GenAI tools are most commonly used in the community of CEUR-WS authors, according to their declarations. Figure 3b shows the top 15 GenAI tools declared. ChatGPT and its versions are the most widely declared tools by researchers followed by Grammarly, Gemini, DeepL, and Claude. The high number of mentions for different versions of ChatGPT (ChatGPT, ChatGPT-4, ChatGPT-4o, and ChatGPT-o3, with a total of 3,743 mentions) highlights OpenAI's popularity among researchers. Grammarly ranks second, reflecting that traditional writing support tools still hold an important place alongside newer GenAI tools.

Figure 3a shows the number of GenAI tools reportedly used in each research paper that declared to have used GenAI tools. The data indicates that 59% of papers report to use only one GenAI tool, followed by 34% of papers that declare two tools. The reported use of more tools is much less common. Only four papers reported using six or more GenAI tools.

Table 3 presents the five most popular GenAI tools for each of the contribution roles defined by the CEUR-WS GenAI usage taxonomy. Although the

Table 3: Contribution Roles with top-5 Declared GenAI Tools (# Papers)

Grammar & spelling check (2750)	**Paraphrase and reword (1078)**	**Improve writing style (793)**	**Text translation (315)**	**Drafting content (162)**
Grammarly (1137)	ChatGPT (658)	ChatGPT (355)	DeepL (123)	ChatGPT (68)
ChatGPT (978)	Grammarly (375)	Grammarly (188)	ChatGPT (113)	Claude (29)
ChatGPT-4 (457)	ChatGPT-4 (115)	ChatGPT-4 (109)	Claude (30)	ChatGPT-4 (27)
ChatGPT-4o (274)	Gemini (95)	AI unspec. (101)	Grammarly (27)	ChatGPT-4o (17)
Gemini (178)	ChatGPT-4o (74)	ChatGPT-4o (75)	ChatGPT-4o (25)	Gemini (14)
Code assistance (122)	**Formatting assistance (110)**	**Generate images (104)**	**Plagiarism detection (117)**	**Content enhancement (70)**
ChatGPT (32)	AI unspec. (29)	ChatGPT (17)	Strike Plag. (110)	Claude (17)
claude (19)	ChatGPT (25)	X-AI-IMG (15)	Grammarly (4)	ChatGPT (14)
ChatGPT-4 (19)	Gemini (17)	ChatGPT-4 (15)	ChatGPT-4 (3)	ChatGPT-4 (11)
AI unspec. (17)	Claude (15)	Gemini (10)	ChatGPT (1)	AI unspec. (9)
Gemini (16)	ChatGPT-4 (14)	DALL-E (10)	QuillBot (1)	Gemini (9)
Generate literature review (63)	**Abstract drafting (56)**	**Citation management (36)**	**Peer review simulation (28)**	**Fact-checking (13)**
Claude (15)	Claude (22)	ChatGPT (17)	Gemini (8)	ChatGPT (4)
ChatGPT (13)	ChatGPT (11)	ChatGPT-4 (5)	Claude (7)	Claude (2)
Gemini (11)	Gemini (10)	Grammarly (4)	ChatGPT-4o (6)	ChatGPT-4 (1)
Perplexity (9)	ChatGPT-4o (9)	ChatGPT-4o (3)	ChatGPT (6)	Copilot (1)
SciSpace (8)	ChatGPT-4 (9)	AI unspec. (2)	ChatGPT-4 (5)	Gemini (1)

CEUR-WS GenAI usage taxonomy only proposes 13 contribution roles, our data analysis indicated that these 13 contribution roles are not covering all tasks mentioned by authors. Specifically, the 13 categories do not capture instances in which authors reported to have used GenAI tools to write or debug code during their research. Our data revealed that 122 papers allegedly used GenAI tools for programming-related tasks and declared it in the GenAI declaration. Hence, we introduce a new contribution role named "code assistance". Similarly, another role for "fact-checking" is included, as 13 papers reported using GenAI for this purpose.

Table 4 illustrates the GenAI declaration rates for countries with at least 50 total publications. In this analysis, Country-level statistics were calculated using the whole-counting on author affiliations. Each paper was counted once for every country represented in the authors' affiliations, while multiple affiliations within the same country were counted only once. The data suggests a broad spectrum of adoption, with rates ranging from 16.0% to 77.8%. Norway shows the highest declaration rate (77.8%), followed closely by India (71.7%), Brazil (71.4%), and Japan (69.4%). Other countries with high engagement (above 65%) include Sweden, Austria, China, Belgium, and the United States. A secondary group demonstrates moderate declaration levels between 55% and 63%; this includes Germany, the Netherlands, and the United Kingdom. In contrast, lower declaration rates (54.5%-51.5%) are observed in Italy, Canada, and Poland. Whereas, Greece (42.5%) and Kazakhstan (16.0%) show the lowest rates among the nations.

Table 4: Reported GenAI Usage Rates by Country

Country	Total	AI	Rate %	Country	Total	AI	Rate %
NOR	54	42	77.8	MEX	60	36	60.0
IND	283	203	71.7	SVK	83	49	59.0
BRA	70	50	71.4	PRT	56	33	58.9
JPN	121	84	69.4	FRA	223	128	57.4
SWE	84	58	69.0	FIN	77	44	57.1
AUT	147	100	68.0	GBR	240	135	56.2
CHN	131	89	67.9	CZE	80	45	56.2
BEL	93	63	67.7	IRL	52	29	55.8
USA	342	231	67.5	ITA	1056	575	54.5
ESP	268	176	65.7	CAN	70	37	52.9
CHE	69	45	65.2	POL	136	70	51.5
DZA	66	42	63.6	UKR	1094	507	46.3
DEU	574	360	62.7	GRC	113	48	42.5
NLD	154	94	61.0	KAZ	81	13	16.0

GenAI use is declared for a variety of roles (tasks) in research work (see Table 3). Nearly half of all GenAI tool use (49.44%) is declared for grammar and spelling checks, while paraphrasing (19.38%) and improving writing style (14.25%) are the next most common tasks. Together, these three writing-related functions account for 83% of all declared GenAI use. This shows that researchers primarily report to use GenAI as an advanced writing assistant to refine and rephrase text rather than to generate content. Tasks such as generating abstracts, literature reviews, content enhancement, or drafting content are not so often declared to be done with the help of GenAI tools and their combined usage is less than 6% of the total. In addition, fact-checking, peer review simulation, and citation management are the least declared, suggesting that researchers still trust human judgment for critical analysis and validation more or at least don't declare their use. In terms of tools mentioned, Grammarly dominates grammar and spelling checks with 1,137 declarations, while ChatGPT and its variants (ChatGPT-4, ChatGPT-4o) dominate across multiple tasks, such as paraphrasing, improving writing style, and drafting content. In contrast, text-translation and plagiarism-detection are allegedly mainly performed with tools such as DeepL and Strike Plagiarism. Among formatting assistance, content enhancement, and citation management, the tool labeled "AI unspecified" is reported. This category mainly represents edge cases where authors state that they used GenAI for these tasks but do not specify which tool was used for each contribution or role.

Figure 4 shows the reported use of individual generative AI tools across countries. For this analysis, only countries with at least 100 publications in the dataset were included, so that the comparison is based on a sufficient number of papers. ChatGPT is the most frequently declared generative AI tool in almost every country, usually making up about half or more of total declarations.

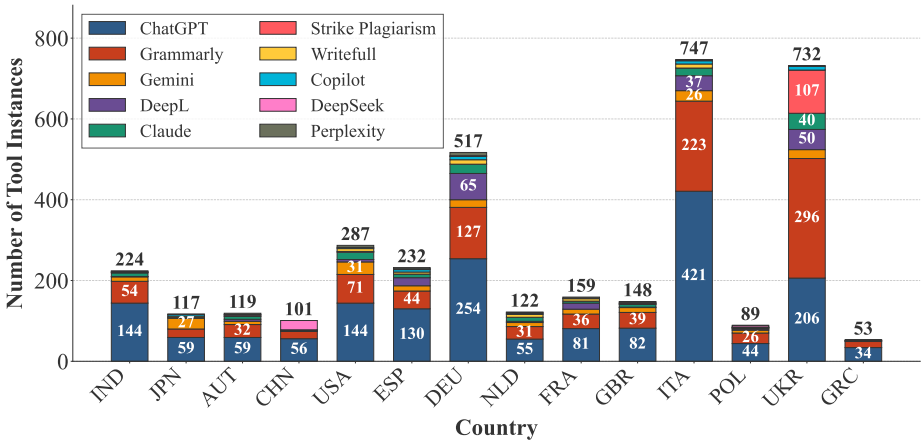


Fig. 4: Declared GenAI tools by country

Its relative alleged use is especially high in India and Greece (around 64%), as well as in Italy and Spain (around 56%) and the United Kingdom (around 55%). This suggests that many researchers depend heavily on one main tool. In countries like Germany (about 49%), Japan (about 50%), and the Netherlands (about 45%), the share is slightly lower, which indicates researchers there are using a wider mix of tools. Grammarly is the second most reported tool in most countries, usually making up about 20–30% of declared usage. However, Ukraine is different, where Grammarly (about 40%) is allegedly used more than ChatGPT (about 28%). Ukraine also shows higher reported use of Strike Plagiarism (about 15%), which suggests a stronger focus on improving writing and checking originality and which might be due to the cooperation of Strike Plagiarism with the Ministry of Education and Science of Ukraine.[6] There are also some regional differences in other tools. For example, DeepL is declared more in countries like Germany (about 13%) and France (about 9%), likely because of language translation needs. In China, DeepSeek has a higher share (about 23%). Other tools such as Gemini, Claude, Perplexity, Copilot, and Writefull are reported much less, usually below 10%, meaning they play a smaller role. These findings suggest that while certain GenAI tools such as Grammarly and ChatGPT are widely used across all countries, some tools are more commonly used in specific countries.

5 Comparison with Non-CEUR-WS Publications

To identify AI usage declarations in Non-CEUR-WS publications, we developed a multi-stage framework (Figure 1). First, raw PDF files were parsed using

[6] From September 2020 the tool was made available to all Ukrainian universities and researchers, https://strikeplagiarism.com/en/news-media-01.html

PyMuPDF (fitz) as the main extractor and pdfplumber as backup. The extracted text was cleaned by removing references, formulas, URLs, and citations. Documents were then segmented into sections and assigned position labels (FRONT, BODY, BACK, FINAL). End sections received higher priority, as AI declarations usually appear in acknowledgments, disclosures, or funding sections rather than abstracts or introductions. Next, each paragraph was evaluated using a two-level rule-based detection system. The first level used a selected list of strong AI tool names such as ChatGPT, Grammarly, and GPT-4. The second level used patterns related to manuscript tasks such as grammar correction, paraphrasing, translation, drafting, and other manuscript preparing activities. A paragraph was selected as a candidate only when it matched both a tool name and a task pattern. This reduced false positives from papers that only discussed AI as a research topic. The highest-scoring candidate paragraphs from each paper were then stored in JSON format.

As a second approach, we designed a semantic similarity pipeline to complement the rule-based method and capture declarations that may not follow fixed wording. Using the same pre-processing steps and section heading mentioned in the first above, each paragraph was encoded with the all-MiniLM-L6-v2 sentence embedding model. Paragraph vectors were compared with more than 180 manually written seed sentences across 17 usage categories using cosine similarity. Paragraphs with similarity above 0.45 were retained as candidates. Finally, the outputs from both the rule-based and semantic similarity pipelines were merged, duplicates removed, and candidate paragraphs passed to the LLM to extract AI declarations.

Across 15,409 Non-CEUR publications, only 203 papers (1.32%) explicitly declared AI use, while 80 papers stated no AI use and 15,126 papers provided no declaration. In total, 58 unique AI tools were identified, with most AI-declaring papers reporting only one tool (median = 1, maximum = 6). ChatGPT was the most frequently reported tool and dominated several roles, particularly improving writing style, paraphrasing, drafting content, and image generation, while Grammarly was most common for grammar and spelling correction. Overall, writing-related support was the alleged main use of AI, as the top three contribution roles accounted for 72.1% of all role mentions, indicating that AI was primarily used for language polishing rather than core research tasks. Geographically, the highest number of AI declarations came from Ukraine, Germany, and Italy, followed by China, the United Kingdom, and the United States. Due to paper space limitations, we were not able to include all figures and tables for Non-CEUR publications, as well as some additional analyzes of CEUR publications. Therefore, we created an online application that presents the full findings and extended analyzes results.[7]

[7] https://36fjntm9huhwqw3cl3tlhy.streamlit.app/

6 Conclusions

Against the background of the increasing usage of GenAI tools in scholarly writing and the demand to transparently declare if and how they were used, we set out to better understand who declares GenAI use in their publication, which GenAI tools are popular among authors of papers published in CEUR-WS proceedings, and for what tasks they are used in the preparation of scholarly manuscripts. Authors of approximately 60% of the published papers have declared the use of at least one GenAI tool for manuscript preparation, and ChatGPT and Grammarly are by far the most popular GenAI tools used for tasks ranging from grammar and spelling checks over generating images to citation management and peer review simulation. In addition, we investigated GenAI adoption patterns across countries, identifying Norway, India, Australia, and the United States as having the highest reported usage rates. Country-specific preferences also emerged; for instance, DeepL is declared more in Germany compared to other countries, while Gemini is more common in the United States, as well as DeepSeek in China. Furthermore, we found that explicitly declaring the (non-) use of GenAI in scholarly works seems not to be a widely adopted part of good scholarly practice yet, since the authors of articles published in CEUR-WS proceedings did not show this behavior in their other recent publications, where only 1.8% of papers declared AI usage, while the remaining 98.2% did not. However, a limitation of our non-CEUR publication analysis is that author matching without ORCID identifiers relied on normalized name matching, which may have merged different individuals with identical names, which may have introduced some publications from authors not actually part of the CEUR-WS dataset. More general limitations of our analysis are (1) that CEUR-WS proceedings are focused on computer science, hence our results cannot safely be generalized to other domains; and (2) that declared GenAI usage does not necessarily accurately reflect actual GenAI usage. Reported usage shares might constitute underestimations of actual usage shares, especially concerning use cases for which authors might deem GenAI use less accepted by the scholarly community, like content generation (as compared to, for instance, spell-checking; see also [34,14]). In future continuations of this study, the complementary use of AI detection tools might be a promising way to enable comparisons between reported and detected GenAI usage. Overall, the findings emphasize the (normative) function of publisher guidelines on AI disclosure in promoting research integrity and good scholarly practice.

GenAI Declaration. We used ChatGPT and Claude Sonnet 4.6 as coding assistance tools, especially for regex formulation.After using these tools, we reviewed and edited the content as needed and take full responsibility for the article's content.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ans, M., Maggio, L.A., Algodi, H., Costello, J.A., Driessen, E., Oswald, K., Lingard, L.: The Presence and Nature of AI-Use Disclosure Statements in Medical Education Journals: A Bibliometric Study. Perspectives on Medical Education **15**(1) (2026). https://doi.org/10.5334/pme.2431, ubiquity Press
2. Augenstein, I., Das, M., Riedel, S., Vikraman, L., McCallum, A.: SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 546–555. ACL (2017). https://doi.org/10.18653/v1/S17-2091
3. Beltagy, I., Lo, K., Cohan, A.: SciBERT: A pretrained language model for scientific text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019). pp. 3615–3620. ACL (2019). https://doi.org/10.18653/v1/D19-1371
4. Chemaya, N., Martin, D.: Perceptions and detection of AI use in manuscript preparation for academic journals. PLOS ONE **19**(7), e0304807 (2024). https://doi.org/10.1371/journal.pone.0304807, Public Library of Science
5. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960). https://doi.org/10.1177/001316446002000104, Sage Publications
6. Dergaa, I., Chamari, K., Zmijewski, P., Saad, H.B.: From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. Biology of Sport **40**(2), 615–622 (2023). https://doi.org/10.5114/biolsport.2023.125623, Termedia
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the conference of the North American chapter of the association for computational linguistics: human language technologies (NAACL-HLT 2019). pp. 4171–4186. ACL (2019). https://doi.org/10.18653/v1/N19-1423
8. Dong, X., Zhao, D., Meng, J., Guo, B., Lin, H.: Syract: zero-shot biomedical document-level relation extraction with synergistic rag and cot. Bioinformatics **41**(7), btaf356 (2025)
9. Eberts, M., Ulges, A.: Span-based joint entity and relation extraction with transformer pre-training. In: ECAI 2020 - 24th European Conference on Artificial Intelligence. Frontiers in Artificial Intelligence and Applications, vol. 325, pp. 2006–2013. IOS Press (2020). https://doi.org/10.3233/FAIA200321
10. Gábor, K., Buscaldi, D., Schumann, A.K., QasemiZadeh, B., Zargayouna, H., Charnois, T.: SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers. In: Proceedings of the 12th International Workshop on Semantic Evaluation. pp. 679–688. ACL (2018). https://doi.org/10.18653/v1/S18-1111
11. Ganjavi, C., Eppler, M.B., Pekcan, A., Biedermann, B., Abreu, A., Collins, G.S., Gill, I.S., Cacciamani, G.E.: Publishers' and journals' instructions to authors on use of generative artificial intelligence in academic and scientific publishing: bibliometric analysis. BMJ **384**, e077192 (2024). https://doi.org/10.1136/bmj-2023-077192, British Medical Journal Publishing Group
12. Ge, Y., Das, S., Guo, Y., Sarker, A.: Retrieval augmented generation based dynamic prompting for few-shot biomedical named entity recognition using large language models. Research Square pp. rs–3 (2025)

13. Gerner, M., Nenadic, G., Bergman, C.M.: Linnaeus: a species name identification system for biomedical literature. BMC bioinformatics **11**(1), 85 (2010). https://doi.org/10.1186/1471-2105-11-85
14. He, Y., Bu, Y.: Academic journals' AI policies fail to curb the surge in AI-assisted academic writing. Proceedings of the National Academy of Sciences **123**(9), e2526734123 (2026). https://doi.org/10.1073/pnas.2526734123, National Academy of Sciences
15. Jain, S., van Zuylen, M., Hajishirzi, H., Beltagy, I.: SciREX: A challenge dataset for document-level information extraction. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7506–7516. ACL (2020). https://doi.org/10.18653/v1/2020.acl-main.670
16. Kacena, M.A., Plotkin, L.I., Fehrenbacher, J.C.: The Use of Artificial Intelligence in Writing Scientific Review Articles. Current Osteoporosis Reports **22**(1), 115–121 (2024). https://doi.org/10.1007/s11914-023-00852-0, Springer
17. Kluegl, P., Toepfer, M., Beck, P.D., Fette, G., Puppe, F.: UIMA Ruta: Rapid development of rule-based information extraction applications. Natural Language Engineering **22**(1), 1–40 (2016). https://doi.org/10.1017/s1351324914000114, Cambridge University Press
18. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020). https://doi.org/10.1093/bioinformatics/btz682, Oxford University Press
19. Luan, Y., He, L., Ostendorf, M., Hajishirzi, H.: 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. pp. 3219–3232 (2018). https://doi.org/10.18653/v1/D18-1360
20. Miller, L.E., Bhattacharyya, D., Miller, V.M., Bhattacharyya, M.: Recent Trend in Artificial Intelligence-Assisted Biomedical Publishing: A Quantitative Bibliometric Analysis. Cureus **15**(5), e39224 (2023). https://doi.org/10.7759/cureus.39224, Springer Nature
21. Mishra, T., Sutanto, E., Rossanti, R., Pant, N., Ashraf, A., Raut, A., Uwabareze, G., Oluwatomiwa, A., Zeeshan, B.: Use of large language models as artificial intelligence tools in academic research and publishing among global clinical researchers. Sci Rep **14**(1), 31672 (Dec 2024). https://doi.org/10.1038/s41598-024-81370-6, Nature Publishing Group
22. Murgia, E., Landoni, M., Huibers, T., Pera, M.S.: Empowered or lost? how search tools and social networks shape teen learning in an ai world. In: Proceedings of the 15th Italian Information Retrieval Workshop (IIR 2025). CEUR Workshop Proceedings, vol. 4026, pp. 7–19. CEUR-WS.org (2025)
23. Otto, W., Zloch, M., Gan, L., Karmakar, S., Dietze, S.: GSAP-NER: a novel task, corpus, and baseline for scholarly entity extraction focused on machine learning models and datasets. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 8166–8176. ACL (2023). https://doi.org/10.18653/v1/2023.findings-emnlp.548
24. Pan, H., Zhang, Q., Dragut, E., Caragea, C., Latecki, L.J.: DMDD: A large-scale dataset for dataset mentions detection. Transactions of the Association for Computational Linguistics **11**, 1132–1146 (2023). https://doi.org/10.1162/tacl_a_00592, mIT Press
25. Perkins, M.: Generative AI policies for academic publishers (2023). https://doi.org/10.6084/m9.figshare.24124860.v1, dataset

26. Perkins, M., Roe, J.: Academic publisher guidelines on AI usage: A ChatGPT supported thematic analysis. F1000Research **12**, 1398 (2024). https://doi.org/10.12688/f1000research.142411.2, F1000 Research Ltd
27. Raman, R.: Transparency in research: An analysis of ChatGPT usage acknowledgment by authors across disciplines and geographies. Accountability in Research **32**(3), 277–298 (2025). https://doi.org/10.1080/08989621.2023.2273377, taylor & Francis
28. Semrl, N., Feigl, S., Taumberger, N., Bracic, T., Fluhr, H., Blockeel, C., Kollmann, M.: AI language models in human reproduction research: exploring ChatGPT's potential to assist academic writing. Hum Reprod **38**(12), 2281–2288 (2023). https://doi.org/10.1093/humrep/dead207, Oxford University Press
29. Tang, A., Li, K.K., Kwok, K.O., Cao, L., Luong, S., Tam, W.: The importance of transparency: Declaring the use of generative artificial intelligence (AI) in academic writing. Journal of Nursing Scholarship **56**(2), 314–318 (2024). https://doi.org/10.1111/jnu.12938, Sigma Theta Tau International
30. Tang, G., Eaton, S.E.: A Rapid Investigation of Artificial Intelligence Generated Content Footprints in Scholarly Publications. Journal of Scholarly Publishing **55**(3), 337–355 (2024). https://doi.org/10.3138/jsp-2023-0079, University of Toronto Press
31. Wadden, D., Wennberg, U., Luan, Y., Hajishirzi, H.: Entity, relation, and event extraction with contextualized span representations. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019). pp. 5784–5789. ACL (2019). https://doi.org/10.18653/v1/D19-1585
32. Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., Wang, G., Guo, C.: GPT-NER: Named entity recognition via large language models. In: Findings of the Association for Computational Linguistics: NAACL 2025. pp. 4257–4275. Association for Computational Linguistics, Albuquerque, New Mexico (Apr 2025). https://doi.org/10.18653/v1/2025.findings-naacl.239, https://aclanthology.org/2025.findings-naacl.239/
33. Yan, R., Jiang, X., Dang, D.: Named entity recognition by using XLNet-BiLSTM-CRF. Neural Processing Letters **53**(5), 3339–3356 (2021). https://doi.org/10.1007/s11063-021-10547-1, Springer Nature
34. Yoo, J.H.: Defining the Boundaries of AI Use in Scientific Writing: A Comparative Review of Editorial Policies. J Korean Med Sci **40**(23) (2025). https://doi.org/10.3346/jkms.2025.40.e187, The Korean Academy of Medical Sciences
35. Zhao, L., Kang, L., Guo, Q.: Zero-shot document-level biomedical relation extraction via scenario-based prompt design in two-stage with llm. Computational Biology and Chemistry **123**, 108978 (2026). https://doi.org/https://doi.org/10.1016/j.compbiolchem.2026.108978, https://www.sciencedirect.com/science/article/pii/S1476927126001039
36. Zhong, Z., Chen, D.: A frustratingly easy approach for entity and relation extraction. In: Proceedings of the conference of the North American chapter of the association for computational linguistics: human language technologies (NAACL-HLT 2021). pp. 50–61. ACL (2021). https://doi.org/10.18653/v1/2021.naacl-main.5
37. Žust, M., Grobelnik, M., Grobelnik, A.M., Sittar, A.: Towards ai-powered real-time negotiation agent. In: Joint Proceedings of the ESWC Workshops and Tutorials co-located with 22nd Extended Semantic Web Conference (ESWC 2025). CEUR Workshop Proceedings, vol. 3977. CEUR-WS.org (2025)