Towards AI-assisted Academic Writing

Anonymous ACL submission

Abstract

We present components of an AI-assisted academic writing system including citation recommendation and introduction writing. The system recommends citations by considering the user's current document context to provide relevant suggestions. It generates introductions in a structured fashion, situating the contributions of the research relative to prior work. We demonstrate the effectiveness of the components through quantitative evaluations. Finally, the paper presents qualitative research exploring how researchers incorporate citations into their writing workflows. Our findings indicate that there is demand for precise AI-assisted writing systems and simple, effective methods for meeting those needs.

1 Introduction

Scientific communication, including writing, is a necessary professional skill set. For example, The American Chemical Society guidelines for undergraduate education indicate that students must "learn how to communicate technical information . . . clearly and concisely, [i]n a scientifically appropriate style"' (American Chemical Society). Writing effective prose is a skill developed through practice and feedback. Although the vast majority of scholarly publications are written in English, most English-language authors are not L1 English speakers. The proficiency gap negatively affects productivity of non-L1 authors. For example, Flowerdew (1999) found that over two-thirds of L1 Cantonese academic authors, writing in English, felt disadvantaged relative to L1 English speakers. Even for L1 speakers, the precise nature of scholarly language takes time and practice to develop expertise. Morris (2023) interviewed scientists, who noted that their students were often "not strong writers." The respondents anticipated that assisted writing would "improve writing quality for a large number of students."

Writing is fundamentally a task of translating ideas into text (Flower and Hayes, 1981). Interactive writing systems guide authors through the writing process by destructuring the writing process and, most recently, generating fluent language. Borrowing from crowdsourcing, *Play Write* (Iqbal et al., 2018) "selfsourced" writing tasks through microtasks divorced from the document editor and delivered to the end user by an app. The tasks included outputs of typical NLP tasks such as summarization and grammar correction. Whereas older systems delegated tasks to the individual or crowd works, recent works incorporate LLMs as user-guided co-creators. For example, *Wordcraft* (Yuan et al., 2022) focused on story writing. The system provided affordances for rewriting, elaborating, and open text generation. *Sparks* (Gero et al., 2022a,b) used a LLM to suggest starter sentences intended to catalyze creative, compact writing for a general audience. Similar to users of Wordcraft, users of Sparks found value simply in generating narrative.

Computer-assisted writing is not a new concept; among others, Mahlow (2023) notes that AI-assisted writing is already commonplace. Modern LLMs are capable of generating text in scientific contexts comparable to expert human authors (e.g. Wang et al. (2019); Ali et al. (2023); Gao et al. (2023a)) although this depends on the context (c.f. Ruggeri et al. (2023)). Scientifically-grounded text generation is part of a larger adoption of AI in the sciences (Hope et al., 2022). In this paper, we present two affordances for generative text in scientific contexts: citation recommendation and introduction writing. We develop and evaluate these affordances in the context of user-facing AI-assisted writing. Finally, we present the results of qualitative research on how researchers incorporate citation recommendations into their workflows. The system and findings show that AI-assisted writing is capable of generating useful content for aca-

demic authors, and that richer *in situ* affordances can provide users with agency to craft more precise scholarly manuscripts.

2 Contextual citation recommendation

The citation recommendation task is typically framed as a recommender system that produces a ranked list of possible citations. Various approaches developed over time as machine learning methods evolved. Färber and Jatowt (2020) provide an overview of techniques that predate large language models (LLMs). Most techniques encode academic works into some semantic similarity (e.g. a topic model (Kataria et al., 2010) or an embedding (Beltagy et al., 2019)). Graph-based approaches (Ali et al., 2024) use the directed graph of citations as features or as a network for propagation of existing features.

Locating, copying, and formatting citations to include in the project takes time and effort. When performed concomitantly with writing, this context-switch between citation discovery can interrupt the user's writing flow.

We imagine *in situ* citation recommendation as a task which recommends citations given the user's context and focus. Here, the focus is the cursor (insertion point) in the active document, representing the desired location of the suggestion. The context is some substring of the document leading up to the insertion point. We envisioned multiple scenarios for suggestions, depending on how much context the author has, and what type of output they desire. Consider the known-item refinding task where the author knows of a specific work and wishes to cite it. Frequently, the author can recall details about the work that they wish to cite (Wildemuth and O'Neill, 1995; Bruce et al., 2004), although the known details might be incorrect. The author might recall these with a lower degree of precision (i.e. "about 5–10 years ago" or "at an NLP conference" or "from Yamada Hanako's lab"). Finally, the author might not recall any of the indexing details of the paper. Instead they might remember a summary of the contribution. These incomplete or incorrect semantic cues to the underlying item are opportunities for the system to use additional context and world knowledge for recommending citations.

2.1 Implementation

Our system recommends citations from two sources. First, a user's writing project typically contains one or more files with citations expressed as structured content, e.g. BibTeX. Second, the system contains a local database of scholarly works: a copy of the OpenAlex corpus (Priem et al., 2022). Each record in this corpus includes the work's author(s), title, abstract, date, publication venue, citation count, and so forth. We used a language detection classifier to exclude works that appeared to be written in a language other than English. Because the mode of citation count is zero, we also excluded uncited works. Since the experiments documented herein were performed, the current implementation of the system retains recent uncited works in the database to allow them to be surfaced. After filtering, our database copy had 60.3 million rows out of the original 263.3 million rows.

As an interactive system, reducing response latency is critical to user perception and satisfaction. The system uses a highly scalable approximate nearest neighbor search (Sun et al., 2024) index for rapid retrieval of similar records in an embedding space. We chose the SPECTER2 embedding (Singh et al., 2023), a multi-format embedding developed specifically to represent scientific documents. SPECTER2 was trained on data from 23 different fields, not limited to computer science. SPECTER2 embeddings outperformed existing models on retrieval tasks. Our system concatenates each paper's title and abstract (if available in the OpenAlex record), projecting this text into the SPECTER2 embedding space.

In addition to works available within the user's BibTeX files, the system needs to find novel candidates from the index that satisfy the user's intent. We implement this recommender as a Retrieval Augmented Generation (RAG) system (Gao et al., 2023b). To retrieve a set of relevant citations, the system queries the index of existing works. Recall that the works are represented by a vector embedding of the title and abstract. The system takes advantage of LLMs observed behavior of "hallucinating" nonexistent facts or concepts (Ji et al., 2023). Essentially, we prompt the LLM to fabricate a likely citation and then use that to find real citations. To do this, the system supplies the LLM with a prompt (see Appendix A.1) containing the previous, current, and subsequent sentences from the user's content. The current sentence contains a special token which indicates to the system where in the sentence the citation is desired. The prompt instructs the LLM to fabricate the title and abstract of a paper that satisfies the user's context. Note that

the system does not care if the LLM's generated citation exists. Rather, the fabricated citations are used as queries into the index of existing works. The fabricated title and abstract are embedded using the SPECTER2 model, which creates a vector used to query the nearest-neighbor index. As implemented, at most 10 nearest neighbors are returned.

Although each result could be ranked by its distance to the query vector in embedding space, we apply an additional layer of scoring. Each result retrieved from the index is formatted into a new prompt (Appendix A.2. These results are formatted as JSON objects. Each result is also given a unique, short hexadecimal string as a "key" property. Keys are constructed rather than using ordinal numbers (1, 2, ...) or letters (A, B, ...) to avoid label bias (Reif and Schwartz, 2024). Some LLMs also exhibit order bias (Shi et al., 2024); we did not evaluate this in our study. The prompt instructs the LLM to output the key that matches "best citation to support [the] claim." Rather than using the key as output, the system runs model inference and collects the model's *scores* for each of the keys in the input. A model's output score for each key is, to an approximation, the log probability of outputting that key to complete the input (prompt). The results are then ranked by their respective scores.

We also implemented pairwise comparison to score suggestions. Qin et al. (2024) showed that LLMs can be used to rank by presenting pairwise choices and having the LLM choose one of the items. This method differs from the scoring method described above. The model is prompted to choose the item from a pair of items that best matches the prompt. By combining pairwise ranks, one can determine a total ranking. By focusing the model's attention on a smaller number of targets, adverse effects from irrelevant targets are avoided.[1] Constructing the total ordering requires many pairwise comparisons. Although some techniques for reducing the quantity of comparisons exist (Bradley and Terry, 1952; Chen et al., 2013), we discarded this method due to the substantial increase in inference time, favoring the scoring method above.

The online citation recommender system allows the user to request a set of citation suggestions by right-clicking in the text editor. The client sends a substring of text adjacent to the insertion point, as well as the contents of BibTeX files. The latter include structured data about publications the author intends to cite.

[1] c.f. Cuconasu et al. (2024), where noise improves quality.

2.2 Evaluation

To assess the efficacy of our citation recommendation system, we evaluated the LLM's performance on the task of retrieving ground truth citations extracted from existing papers. The evaluation dataset was created from papers in S2ORC, a corpus of over 81 million papers spanning STEM disciplines (Lo et al., 2020). We uniformly sampled 0.1% of papers from this corpus, then filtered to papers that include at least 10 sentences that include citations that existed in OpenAlex prior to September 2023 (our cutoff date). This ensured that the system would have access to titles and abstracts for these citations and would be able to use them as distractors in our evaluations. Five citation-containing sentences were randomly sampled from each qualifying paper, resulting in a dataset of 1015 sentences.

For each sentence, we gathered the necessary inputs to run the suggestion citation prompt described in Section 2.1. This includes the target sentence's surrounding context and titles and abstracts of n possible citations, for $n \in \{3, 5, 10\}$.

The n candidate citations included the ground truth citation and $n - 1$ distractor citations. We chose distractors in three different ways to test the system under varying difficulty. From least to most difficult, distractors were chosen uniformly randomly from:

- all papers in the evaluation dataset (sample of S2ORC)
- the ground-truth citation's nearest neighbors in SPECTER2 embedding space
- the references of the source paper containing the test sentence, excluding the ground truth reference

We employ Precision at k (P@k) and mean reciprocal rank (MRR) as evaluation metrics. Because the randomly chosen set of distractors is domain agnostic, we expect a paper chosen from S2ORC at random to be unrelated to the test sentence. The two more difficult distractor sources include papers that are semantically related. In the *nearest neighbors* condition, one of the distractors could be a reasonable substitute for the ground truth citation, particularly for well-known results.

Table 1 shows the results. As expected, the ground truth citation tends to rank higher against

Distractor Type	n	MRR	p@1	p@3	p@5
Random	3	0.755	0.612		
Random	5	0.549	0.333	0.665	
Random	10	0.320	0.124	0.348	0.500
Nearest neighbors	3	0.661	0.428		
Nearest neighbors	5	0.506	0.254	0.661	
Nearest neighbors	10	0.300	0.110	0.327	0.523
References	3	0.676	0.462		
References	5	0.496	0.261	0.641	
References	10	0.308	0.109	0.326	0.519

Table 1: Retrieval metrics for 1,015 contextual citation retrieval cases with n targets.

randomly selected distractors when compared to distractors drawn from the semantic space or from the manuscript's references. However, the distractor source has less effect on precision. In a live system that uses this method, the user would need to choose from multiple suggestions rather than having the system propose only the top-ranked item.

3 Writing introductions

3.1 Generating introductions

We frame the introduction writing task as a mapping from the manuscript and references to a small number of paragraphs. The related work in the introduction should act like a microscope: canonical works coarsely orient the reader to a subfield; important recent works provide fine adjustment to the specific research track. Upon this foundation, the introduction builds the case for the specific contribution of the manuscript that follows. Our prompt chain follows this paradigm in three steps.

First, the system uses an LLM to identify novel claims from the author's manuscript relative to other works that the author cited. It assumes that the author already documented references in their BibTeX files at this time. For each reference, the system looks up the corresponding record in the OpenAlex database, retaining only those where a title and abstract are available. These references are split into two groups: *canonical* and *recent*. The *canonical* references were published more than Y years ago while *recent* were published within the last Y years. As in other systems, our system uses the title and abstract as a rough substitute for the work itself (Li and Ouyang, 2024). To perform the relative comparison, the system then extracts paragraphs from the author's current work. Each paragraph is then combined with each of the references to form tuples of (paragraph, title, abstract). The prompt (Appendix B.1) acts as a binary classifier that confounds relevance and novelty. The LLM assess if the each paragraph's content is related to the abstract of the author's paper *and* it is novel relative to the abstract a cited paper. The idea is to use this filter to find the work's novel contributions for incorporation into the introduction.

Each paragraph then receives one or more votes from the binary classifier. The system filters out paragraphs with low support. The remaining paragraphs, assumed to discuss novel results, are then passed to a simple summarization prompt (Appendix B.2. Although current LLMs have long context lengths, at the time of our experiments, the token limit was smaller, and hence the (possibly many) novel paragraphs needed to be reduced into a shorter text.

Finally, the system combines the canonical works, recent works, and summary of novel contributions into the written introduction section using the prompt in Appendix B.3. Example output of running the prompt chain on this submission is provided in Appendix C.

3.2 Evaluation

We evaluate the generated introductions using text metrics and by prompting an LLM. Our evaluation dataset is a subset of papers from the [United States] National Bureau of Economic Research[2] (NBER). We extracted the introduction from 14 NBERs papers. For text evaluations, we use ROUGE (Lin, 2004) which is a recall-based metric and often used in the context of summarization. The average ROUGE score across the papers is

[2] https://www.nber.org/research/data

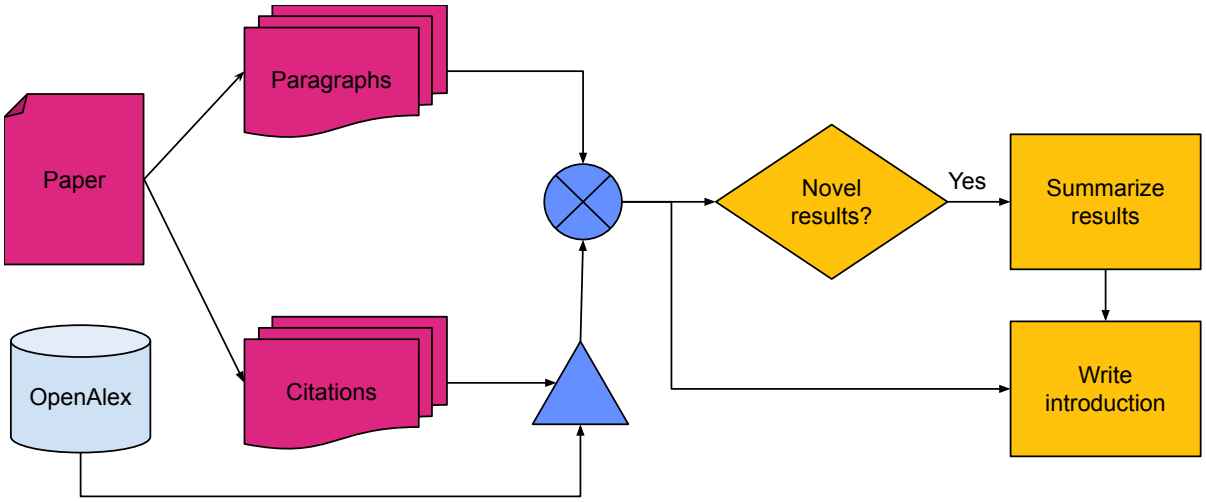

Figure 1: Flowchart from paper and citations to written introduction.

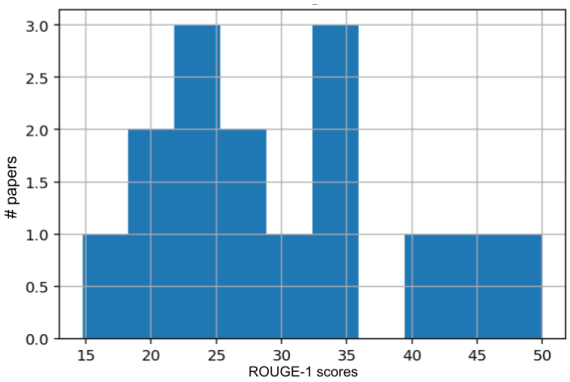

Figure 2: (Supplement) Distribution of ROUGE-1 scores for the generated introductions.

29.9, the distribution of scores is shown in Figure 2.

For LLM-based evaluations, we consider evaluating the introductions based on the claims made in the generated introductions in comparison to the original introductions. Specifically, we used a prompt (Appendix B.4) to extract 3–5 claims from each of the generated texts. From 14 generated introductions, we extracted 52 claims. Then, we prompt the LLM to verify whether the claim from the generated introduction entails from the full original introduction. Appendix B.5 includes the full prompt used. We consider two versions of the LLM based evaluation: (1) we ask the LLM for a simple "yes" or "no" response for the prompt, (2) we consider the log-likelihood scores for the "yes" and "no" response tokens and normalize them to determine the probability that the generated claim is entailed from the the original. Figure 3 presents the probability scores for whether claims from the generated introduction entail from the original introduction. Of the 52 claims extracted from the generated introductions, 47 of them are entailed from the original introduction indicating a high degree of precision. In general, we find that the generated introductions score highly when the original introduction section hews closely to a single topic. Table 2 compares a generated paragraph from NBER 20209 (Borovička et al., 2014) with entailment score 0.983 versus a lower-performing paragraph from NBER 22392 (Nakamura et al., 2016) with entailment score 0.279. The higher-scoring generated passage captures the main concepts from the original work, while the lower-scoring passage hones in on natural disasters, which is not the theme of the original work. The LLM correctly identified the paragraph as not entailing the original. It seems reasonable to expect that self-critique (Madaan et al., 2023) can be used to improve generated introductions in the future.

4 Opportunities

As part of a study on citation verification, we conducted semi-structured interviews with academic authors about their experiences citing related work. Six researchers (5 self-identifying as men; 1 woman) from the research division of a large technology company participated. All researchers are experienced academic authors (mean h-index 26, $\sigma = 13$). Their research domains include subfields of computer science including quantum computing, virtual reality, biomedical imaging, natural language processing, and responsible artificial intelligence.

Original excerpt	Generated excerpt
NBER 20209, entailment 0.983 It has been known, at least since the path-breaking work of Arrow, that asset prices reflect a combination of stochastic discounting and probability distributions. . . . We demonstrate in Section 5 that a Perron–Frobenius approach leads naturally to the construction of a martingale component . . .	This research explores alternative solutions by investigating the application of Perron–Frobenius theory to construct a probability distribution from Arrow prices. . . .
NBER 22392, entailment 0.279 However, just because the inhabitants of some locations have higher incomes than others does not mean there is a large causal effect of moving to these locations. . . . Distinguishing between selection and direct causal effects of locations is challenging. Large, exogenous relocation shocks are few and far between . . . We shed new light on the role of location in shaping economic outcomes by studying the consequences of a true "natural" experiment. On January 23, 1973, a long-dormant volcano erupted unexpectedly on the Westman Islands . . .	The impact of natural disasters on economic development is a topic of considerable interest to economists. While the immediate consequences of natural disasters are often well documented, the long-term impacts are less well understood. In particular, the impact on intergenerational mobility and the transmission of risk across generations is a critical yet underexplored area of research. <cit.> have shown that children who spend more time in a low-poverty environment have better long-term outcomes. This suggests that the effects of natural disasters may not only be felt by those who experience them directly, but also by future generations.

Table 2: Comparison of two generated introduction paragraphs with high and low entailment scores relative to the original text.

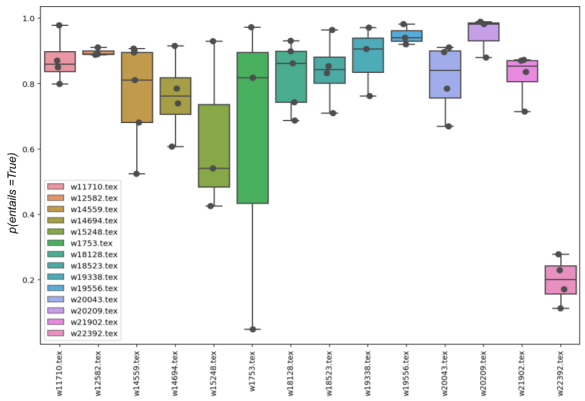

Figure 3: Distribution of scores for whether claims from the generated introductions *entail* the original introduction based on an LLM.

The semi-structured interviews covered the following topics: participants' current approaches to find and validate references, if their approaches would change with unlimited time and resources, how their approach differs depending on citation type, and imagined capabilities of an ideal support tool for citation verification and recommendation. We performed inductive thematic analysis of the interviewees' statements. We performed three rounds of coding to create themes, resolving disagreements through conversation among two authors.

Time constraints limit the validation process. Nearly all participants raised the concern of careful validation. That is, they needed to understand specifically how the citation was relevant. However, several participants mentioned time constraints influencing their decision to cite works. Although every participant indicated that they sometimes cited papers that they had fully read, they also noted instances where they cited papers they had not entirely read. They employed skimming strategies while engaged in the literature review process in order to find more precisely related works.

Participants suggested various affordances for a tool to support the validation process. For example, one participant suggested finding the specific claims in the suggested citation that were related to the author's citing text. Going to the original source was important because some participants remarked that papers' claims can be misrepresented by citing authors, or the abstract did not accurately reflect the paper's results. In interfaces for scholarly readers, existing systems such as *Relatedly* (Palani et al., 2023) provide affordances similar to those suggested by the participants. The system we presented in this work only surfaces paper metadata such as title and abstract, so incorporating additional sensemaking affordances as part of the user's workflow will support more rigorous citation suggestion.

Surrounding text must be accurately scoped. Participants also stressed the importance of having nuanced enough statements to accurately represent

the paper [P1,P2]. They recognized that inaccurate corresponding text is often the consequences of human error or time constraints, rather than bad faith actions. Therefore, P2 expressed interest in support for rewriting existing text spans to better represent the cited paper. Our work finds a reference from a text span. Future work could also improve an existing span to better represent the reference.

Community norms impact reference choices. Some participants felt pressure to cite "the right" source because peer reviewers would easily identify gaps in the related work. However, the precision of those citations varied depending on the field and relevance to the author's work. Several suggested that the situating citations might be more interchangeable than the more recent works.

Importance of contextualization within the broader literature. Participants reflected that although a given citation may be relevant, it may not be sufficient [P3,P4]. For instance, multiple references may be needed if the statement is multifaceted and nuanced, or if the statement is broad and requires a set of references. This idea of sufficiency extended to the reference set of entire sections, as P1 expressed concerns about misrepresenting sub-fields when merging or combining subsections of a related work.

Part of the challenge of building a good reference set is understanding the broader trends of the overall field. Participants expressed interest in a tool that bridges relevant but separate streams of the literature, whether it be similar methods and theories from a different field or differing methods and theories from a similar field [P1,P2,P3]. The challenge of becoming aware of and fully encapsulating these different strands motivated their wish for a tool with a broad sense of the literature. These reflections suggest that reference selection must be valid on multiple levels, with each individual reference accurately represented in the close text and the set of references sufficient in representing the overall literature. Our tool focuses on the former, and there is a rich opportunity for future work in the latter.

5 Conclusion

As a highly developed, precise form of communication, the skill set of academic writing takes time to develop. The writing process requires focus, yet can be disrupted by related tasks such as the curation of related work. The qualitative research showed that even experienced authors have nuanced procedures for identifying and citing prior work. Rather than fully replace academic authors, it seems more likely that writing assistants will continue to proliferate, capturing a rich design space (Lee et al., 2024). In this paper, we presented two affordances for academic writing framed in the context of a live authoring experience: suggesting citations in the context of the document, and writing an introduction section. Quantitative evaluation shows that these methods are capable of generating content that augments the author's writing process.

6 Limitations

The system, studies, and participants described herein were only evaluated on English-language documents and queries, although five of the six participants were fluent in a language other than English. The OpenAlex corpus includes non-English documents, but we excluded those from our database. Finally, citation suggestion is an inherently biased task. Simple filters such as citation count prevent the discovery of "sleeping beauties" (van Raan, 2004), while heuristics such as the venue's impact factor may obscure novel ideas that have not made it into mainstream publication. Systems that take diverse viewpoints into account, and present them to authors in an interpretable fashion, will help diffuse novel ideas into scientific discourse.

7 Acknowledgments

Removed for review.

References

Rohaid Ali, Oliver Y. Tang, Ian David Connolly, Patricia L. Zadnik Sullivan, J. H. Shin, Jared S. Fridley, Wael Asaad, Deus Cielo, Adetokunbo A. Oyelese, Curtis E. Doberstein, Ziya L. Gokaslan, and Albert E. Telfeian. 2023. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. *Neurosurgery*, 93:1353–1365.

Zafar Ali, Guilin Qi, Irfan Ullah, Adam A. Q. Mohammed, Pavlos Kefalas, and Khan Muhammad. 2024. GLAMOR: Graph-based LAnguage MOdel embedding for citation Recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, page 929–933, New York, NY, USA. Association for Computing Machinery.

American Chemical Society. Professional Skills & Competencies - American Chemical Society

— acs.org. https://www.acs.org/education/policies/acs-approval-program/guidelines/professional-skills.html. [Accessed 7 January 2025].

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

Jaroslav Borovička, Lars P Hansen, and José A Scheinkman. 2014. Misspecified recovery. Working Paper 20209, National Bureau of Economic Research.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs. *Biometrika*, 39:324–335.

Harry Bruce, William Jones, and Susan Dumais. 2004. Keeping and re-finding information on the web: What do people do and what do they need? *Proceedings of the American Society for Information Science and Technology*, 41(1):129–137.

Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. 2013. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, page 193–202, New York, NY, USA. Association for Computing Machinery.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for RAG systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 719–729, New York, NY, USA. Association for Computing Machinery.

Linda S. Flower and J. R. Hayes. 1981. A cognitive process theory of writing. *College Composition & Communication*.

John Flowerdew. 1999. Writing for scholarly publication in English: The case of Hong Kong. *Journal of Second Language Writing*, 8(2):123–145.

Michael Färber and Adam Jatowt. 2020. Citation recommendation: approaches and datasets. *International Journal on Digital Libraries*, 21(4):375–405.

Catherine A. Gao, Frederick M. Howard, Nikolay S. Markov, Emma C. Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T. Pearson. 2023a. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *npj Digital Medicine*, 6(1):75.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *ArXiv*, abs/2312.10997.

Katy Gero, Vivian Liu, and Lydia Chilton. 2022a. Sparks: Inspiration for science writing using language models. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, pages 83–84, Dublin, Ireland. Association for Computational Linguistics.

Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022b. Sparks: Inspiration for science writing using language models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, DIS '22, page 1002–1019, New York, NY, USA. Association for Computing Machinery.

Tom Hope, Doug Downey, Oren Etzioni, Daniel S. Weld, and Eric Horvitz. 2022. A computational inflection for scientific discovery. *Communications of the ACM*, 66:62–73.

Shamsi T. Iqbal, Jaime Teevan, Daniel J. Liebling, and Anne Loomis Thompson. 2018. Multitasking with Play Write, a mobile microproductivity writing tool. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, UIST '18, page 411–422, New York, NY, USA. Association for Computing Machinery.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12).

Saurabh Kataria, Prasenjit Mitra, and Sumit Bhatia. 2010. Utilizing context in generative Bayesian models for linked corpus. *Proceedings of the AAAI Conference on Artificial Intelligence*, 24(1):1340–1345.

Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A. Alghamdi, Tal August, Avinash Bhat, Madiha Zahrah Choksi, Senjuti Dutta, Jin L.C. Guo, Md Naimul Hoque, Yewon Kim, Simon Knight, Seyed Parsa Neshaei, Antonette Shibani, Disha Shrivastava, Lila Shroff, Agnia Sergeyuk, Jessi Stark, Sarah Sterman, Sitong Wang, Antoine Bosselut, Daniel Buschek, Joseph Chee Chang, Sherol Chen, Max Kreminski, Joonsuk Park, Roy Pea, Eugenia Ha Rim Rho, Zejiang Shen, and Pao Siangliulue. 2024. A design space for intelligent and interactive writing assistants. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Xiangci Li and Jessica Ouyang. 2024. Related work and citation text generation: A survey. In *Proceedings of the 2024 Conference on Empirical Meth-*

ods in Natural Language Processing, pages 13846–13864, Miami, Florida, USA. Association for Computational Linguistics.

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

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

Cerstin Mahlow. 2023. Writing tools: Looking back to look ahead. *Preprint*, arXiv:2303.17894.

Meredith Ringel Morris. 2023. Scientists' perspectives on the potential for generative AI in their fields. *arXiv preprint arXiv:2304.01420*.

Emi Nakamura, Jósef Sigurdsson, and Jón Steinsson. 2016. The gift of moving: Intergenerational consequences of a mobility shock. Working Paper 22392, National Bureau of Economic Research.

Srishti Palani, Aakanksha Naik, Doug Downey, Amy X. Zhang, Jonathan Bragg, and Joseph Chee Chang. 2023. Relatedly: Scaffolding literature reviews with existing related work sections. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

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

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2024. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1504–1518, Mexico City, Mexico. Association for Computational Linguistics.

Yuval Reif and Roy Schwartz. 2024. Beyond performance: Quantifying and mitigating label bias in LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6784–6798, Mexico City, Mexico. Association for Computational Linguistics.

Federico Ruggeri, Mohsen Mesgar, and Iryna Gurevych. 2023. A dataset of argumentative dialogues on scientific papers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7684–7699, Toronto, Canada. Association for Computational Linguistics.

Lin Shi, Chiyu Ma, Wenhua Liang, Weicheng Ma, and Soroush Vosoughi. 2024. Judging the judges: A systematic study of position bias in LLM-as-a-judge. *Preprint*, arXiv:2406.07791.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5548–5566, Singapore. Association for Computational Linguistics.

Philip Sun, David Simcha, Dave Dopson, Ruiqi Guo, and Sanjiv Kumar. 2024. Soar: improved indexing for approximate nearest neighbor search. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Anthony F. J. van Raan. 2004. Sleeping Beauties in science. *Scientometrics*, 59(3):467–472.

Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. 2019. PaperRobot: Incremental draft generation of scientific ideas. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1980–1991, Florence, Italy. Association for Computational Linguistics.

Barbara M. Wildemuth and Ann L. O'Neill. 1995. The "known" in known-item searches: Empirical support for user-centered design (research note). *College & Research Libraries*, 56(3):265–281.

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, IUI '22, page 841–852, New York, NY, USA. Association for Computing Machinery.

Prompts templates are processed using the Jinja[3] templating library. Line breaks shown here may not match the line breaks used in the text prompt.

A Prompts for suggesting citations

A.1 Citation fabrication

```
You are an expert at suggesting relevant scientific papers.

I will provide some sentences from a paper that I am writing. In the sentences,
I will place a token CITE-HERE where I need to cite a relevant paper. Your task
is to make up the title and abstract of a paper that you think would be relevant
to this context. Give your output in JSON format with values for keys "title"
and "abstract".

SENTENCES: {{ previous_sentence }} {{ masked_sentence }} {{ next_sentence }}

Now, make up the title and abstract of a paper that I should cite at the CITE-HERE token.

Answer:
```

A.2 Citation scoring

```
You are the editor at a prestigious scientific journal. The author of a paper asks
you to recommend the best citation to support their claim. You are given a set of
citations of papers in JSON format. Each citation includes a key in the "key" field,
the paper title in the "title" field, and the paper abstract in the "abstract" field.
You are also given an extraction of the paper, which indicates the location of the
desired citation with the string "CITE-HERE".

Select the best citation from the list of citations that best supports the context
of the extraction and give the value of the corresponding "key" field. Only give me
the value, nothing else.

EXTRACTION
{{ previous_sentence }} {{ masked_sentence }}  {{ next_sentence }}

CITATIONS
[{% for c in citations %}
{
  'key': {{ c['key'] }},
  'title': {{ c['title'] }},
  'abstract': {{ c['abstract'] }},
}
{% endfor %}

The key of the citation that best fits this extraction is:
```

[3] https://jinja.palletsprojects.com/en/stable/

B Prompts for writing introductions

B.1 Determining claims

Extracts claims from the author's manuscript and compares them with existing work.

```
Your task is to determine if a paragraph from a scientific paper discusses
a novel result. You are given the abstract of the paper, abstract of related
paper, and a paragraph from the body of the paper. You answer YES if and only
if the paragraph's content is related to the abstract of this paper, and it
is novel relative to the abstracts of related papers.

ABSTRACT OF THIS PAPER
{{ abstract }}

ABSTRACT OF A RELATED PAPER
{{ ref_chunk[1].abstract }}

PARAGRAPH FROM THIS PAPER
{{ ref_chunk[0] }}

QUESTION
Q: Does the paragraph from this paper show a novel result worth mentioning
in the introduction? Respond YES or NO and explain your answer in one
sentence.
A:
```

B.2 Summarizing claims

Summarizes claims extracted using the previous prompt.

Inputs: `novel_results`, a list of text chunks from a paper.

```
You are a scientist writing up the results of your work. The following
paragraphs contain information about your results. Summarize the key
results in a few sentences.

{% for result in novel_results %}
  {{ result | trim }}
{% endfor %}

Now summarize the results in a few sentences.
```

B.3 Composing introduction

Final step in the prompt chain to compose the introduction section. Inputs:

Field name	Description
`title`	Manuscript title
`results`	Summary of experimental results
`[genesis_references]`	List of canonical references
`[recent_references]`	List of recent references

```
Given a list of related work, and the results of a paper, write the
introduction section for that paper. Refer to any of the REFERENCE
papers using the id in that REFERENCE.

PAPER TITLE: {{ title }}
```

```
FUNDAMENTAL PAPERS IN THIS FIELD:

{% for ref in genesis_references -%}
REFERENCE #{{ loop.index }}:
{% if ref.title is not none %}Title: "{{ ref.title }}"{% endif %}
{% if ref.abstract is not none %}Abstract: "{{ ref.abstract }}"{% endif %}
{% endfor -%}

RECENT RESULTS THAT THIS PAPER BUILDS ON:

{% for ref in recent_references -%}
REFERENCE #{{ loop.index + len(genesis_references) }}:
{% if ref.title is not none %}Title: "{{ ref.title }}"{% endif %}
{% if ref.abstract is not none %}Abstract: "{{ ref.abstract }}"{% endif %}
{% endfor -%}

RESULTS: {{ results }}

Now write the paper introduction. Cite references from both the FUNDAMENTAL PAPERS and
the RECENT RESULTS. When you cite a reference, use the reference number in brackets.
Begin and end your introduction with three single quotes (''').
{%- endsection -%}

INTRODUCTION:
```

B.4 Claim extraction (for evaluation)

Inputs: `num_claims`, the number of claims to extract from `introduction`, the text of the introduction.

```
Your job is to answer questions about the following introduction section of an
academic paper:

{{introduction}}

List {{num_claims}} novel claims that the introduction section makes. Output only
one sentence per claim and do not add additional explanation. Do not add any additional
formatting. Begin each claim with a number.
```

B.5 Entailment (for evaluation)

This prompt was used for evaluation.

```
You will be given a generated paragraph (hypothesis) and the text corresponding
to the original introduction (context). Your goal is to evaluate whether the
generated paragraph (hypothesis) entails the original introduction.

It is alright if the generated paragraph does not encompass the core of the
original introduction, as long as the generated paragraph does not introduce
any new claims.

Start your answer with "yes" or "no" depending on whether the generated
paragraph entails the original introduction and then provide a brief 1-2
sentence explanation for why.

GENERATED PARAGRAPH (hypothesis):
```

```
{{gen_intro_para}}

Original introduction (context):
{{orig_intro}}

Use the following format:
<yes or no depending on whether hypothesis entails the context>

Explanation:
```

C Generated output

*This **AI-generated text** is the system's output using the introduction writing method, applied to this manuscript. The system did not have access to the human-authored introduction at the time of generation.*

The proliferation of academic literature presents challenges for researchers seeking to efficiently locate and incorporate relevant prior work into their writing. Citation recommendation systems aim to alleviate this burden by suggesting relevant citations given a user's writing context. Early approaches to this task relied on representing academic works using topic models (Kataria et al., 2010) or embedding techniques (Beltagy et al., 2019), while graph-based methods leveraged citation networks to identify relevant work (Ali et al., 2024). While these methods can identify broadly relevant papers, they often fail to capture the nuanced ways in which researchers incorporate citations within their writing.

A key challenge in citation recommendation is the need for systems to understand not only the topical relevance of a citation, but also its specific role within the author's argument. Färber and Jatowt (2020) provide a comprehensive overview of pre-LLM techniques for citation recommendation, highlighting the limitations of existing methods in capturing this contextual information. More recent work has explored the use of large language models (LLMs) for citation recommendation, leveraging their ability to process and understand natural language (Gao et al., 2023b). These LLM-based approaches have shown promise in improving the accuracy and relevance of citation recommendations.

In this paper, we present a novel approach to AI-assisted academic writing that combines contextual citation recommendation with automated introduction generation. Our system leverages the power of LLMs to provide researchers with in-situ citation suggestions, recommending citations that are both topically relevant and appropriate for the specific context of their writing.

We demonstrate the effectiveness of our approach through a quantitative evaluation on a dataset of academic papers. Our results show that our system is able to accurately identify relevant citations, even when presented with limited context. Furthermore, we find that our system is able to generate coherent and informative introductions that accurately reflect the key contributions of a research paper.

Beyond simply identifying relevant citations, our system is designed to support the broader process of academic writing. Recognizing that researchers often struggle to synthesize and integrate prior work into their own writing, we incorporate a novel introduction generation component that automatically generates structured introductions, situating the contributions of a research paper within the context of prior work. This component leverages LLMs to identify the novel contributions of a paper, summarize related work, and compose a coherent introduction that effectively sets the stage for the research presented in the manuscript.