LLM RAG Reranker with Pairwise Comparison and Citation Context

Yexin Wu
yexinw3@illinois.edu

Yihang Jiao
yihangj3@illinois.edu

Abstract

In academic research, both citations and their accompanying contexts are crucial for discerning the influence and relevance of prior work. However, traditional citation-informed document embedding approaches— which generate fixed-size vector representations based solely on citation relationships—often neglect the rich contextual information surrounding citations. While these retrieval methods can identify a broad set of candidate papers, they typically fall short in distinguishing between works that are contextually pivotal and those that merely share superficial citation links. In this work, we introduce the Comparison-based LLM RAG Reranker, a novel method that assigns weights to candidate papers via pairwise comparisons, with a large language model serving as the adjudicator and augmented with citation context. To address the quadratic complexity inherent in exhaustive comparisons, we integrate the UCB Elimination algorithm, thereby enabling efficient reranking. Furthermore, we compile an extensive dataset of over 10,000 academic papers and more than 490,000 citation contexts, enriching the reranking process by incorporating in-link contextual information. Finally, we develop a RAG-powered LLM paper writing pipeline that, given a conceptual idea and a target section, automatically retrieves and organizes pertinent literature, thereby facilitating the generation of coherent and evidence-grounded academic text.

1 Introduction

The rapid growth of academic literature has made it increasingly difficult for researchers to identify, assess, and integrate prior relevant work into their own studies. As the volume of published material continues to grow, traditional literature review methods—relying heavily on manual searches and subjective evaluations—are increasingly insufficient.

In response to these challenges, there is a growing interest in leveraging advanced technologies to enhance the efficiency and effectiveness of literature reviews. Large Language Models (LLMs), such as ChatGPT, have demonstrated potential in automating aspects of the literature review process. However, their application is not without limitations. Studies have shown that while LLMs can assist in generating references and summarizing content, they are prone to issues like hallucinations—producing information that is not based on actual data—and may lack access to the most recent research developments. To address these limitations, integrating citation-context information has emerged as a promising approach. Citation-context is the information related to a citation in the citing document. It can include the section title("Related Work", "Methods", "Experimental Setup/Evaluation Dataset") and citation explanation("A similar approach is employed in xRAG (Cheng et al., 2024)"). In line with the definition provided by Jeong et al., we adopt the following definition. In a document d, we define a context c as a bag of words. The *local context* refers to the text immediately surrounding a citation or placeholder. When document d_1 cites document d_2, the local context associated with this citation is termed the *out-link context* with respect to d_1 and the *in-link context* with respect to d_2.

Building upon these insights, we introduce the Comparison-based LLM RAG Reranker, a novel method designed to assign weights to retrieved papers through pairwise comparisons, with an LLM serving as the adjudicator. Recognizing the computational challenges associated with exhaustive comparisons (C_n^2 for n candidates), we employ the Upper Confidence Bound (UCB) Elimination algorithm to enhance efficiency. Furthermore, we have curated a comprehensive dataset comprising over 10,000 academic papers and more than 490,000 citation contexts. By augmenting

Retrieval-Augmented Generation (RAG) outputs with in-link contexts, we aim to improve the accuracy and relevance of the reranking process.

In addition, we propose a RAG-powered LLM paper writing pipeline that assists researchers in drafting manuscripts. Given a conceptual idea and a target section, this pipeline retrieves pertinent literature and organizes the section content by leveraging the retrieval results. This approach not only streamlines the writing process but also ensures that the manuscript is grounded in the most relevant and up-to-date research, thereby enhancing the quality and credibility of scholarly communication.

Our main contributions are as follows:

- We introduce the Comparison-based LLM RAG Reranker, a method that assigns weights to retrieved papers through pairwise comparisons using LLM as the judge. To mitigate the quadratic complexity associated with exhaustive comparisons, we employ the UCB Elimination algorithm as described in (Mohankumar and Khapra, 2022), thereby efficiently reranking the candidate results.
- We present a comprehensive dataset consisting of over 10,000 academic papers and more than 490,000 citation contexts. By augmenting the RAG result with in-link contexts, we enhance the accuracy of the Comparison-based LLM RAG Reranker.
- We propose a RAG-powered LLM paper writing pipeline that, given a conceptual idea and a target section, retrieves relevant papers and organizes the section content by leveraging the retrieval results.

2 Related work

2.1 Retrieval Augmented Generation

Retrieval-Augmented Generation (RAG) is a framework that enhances the performance of large language models (LLMs) on knowledge-intensive tasks by integrating external non-parametric memory into the generation process. Introduced by Lewis et al., RAG combines a pre-trained sequence-to-sequence model with a neural retriever that accesses a dense vector index of a large corpus, such as Wikipedia. This architecture allows the model to retrieve relevant documents dynamically and generate responses conditioned on both the query and the retrieved documents, thereby improving accuracy and enabling the model to provide up-to-date information. RAG has been shown to outperform traditional parametric models and task-specific architectures in open-domain question answering and other knowledge-intensive tasks.

Subsequent research has expanded upon the RAG framework. Wu et al. provided a comprehensive survey of RAG techniques, discussing various retriever architectures, retrieval fusion methods, and training strategies. They highlighted the versatility of RAG in applications such as question answering, dialogue systems, and summarization, and addressed challenges like hallucination and knowledge updating.

In the context of academic writing, RAG-powered systems can assist researchers by retrieving pertinent literature and integrating relevant information into manuscripts. By leveraging a vast corpus of academic papers, these systems can provide comprehensive overviews, suggest citations, and ensure that the content reflects the most recent advancements in the field. Citation and citation context play a significant role in academic research and writing. Built on SciBERT (Beltagy et al., 2019), SPECTER (Cohan et al., 2020; Singh et al., 2022) adopts contrastive training objects to utilize citation information. The citation context refers to ”the particular passage or statement within the citing document containing the references”. IL-CiteR (Roy and Han, 2024) extract evidence spans from citation contexts to improve the accuracy of the local citation recommendation. However, in our automated academic writing setting, the retrieval process is conditioned solely on the primary concept presented in the abstract and the specific section being developed. Accordingly, we introduce a novel framework that leverages LLMs to incorporate citation contexts, thereby enhancing retrieval results and improving the overall quality of academic writing.

2.2 LLM as Reranker

The application of Large Language Models (LLMs) as rerankers in information retrieval tasks has garnered significant attention due to their ability to understand context and semantics deeply. Traditional retrieval systems often rely on lexical matching, which may not capture the nuanced relevance of documents to a given query. LLM-based rerankers address this limitation by reordering retrieved documents based on a more sophisticated understanding of the content.

Ma et al. introduced a zero-shot listwise document reranking approach using LLMs, where the model generates a reordered list of document identifiers based on their relevance to the query. This method demonstrated superior performance over pointwise approaches, particularly in scenarios lacking task-specific training data.

Chen et al. proposed an efficient reranking method that leverages attention patterns within LLMs. Their approach, termed in-context reranking (ICR), utilizes the attention weights assigned to documents when processing query tokens, enabling accurate reranking with significantly reduced computational overhead. This method outperformed traditional generative reranking models while requiring fewer resources.

In academic search systems, employing LLMs as rerankers can enhance the retrieval of relevant literature by considering the contextual relevance and quality of papers. This ensures that researchers are presented with the most pertinent studies, facilitating a more efficient and comprehensive literature review process.

2.3 Dueling Bandits Algorithm as Sampling Strategy

The dueling bandits algorithm is an extension of the traditional multi-armed bandit problem, tailored for scenarios where feedback is derived from pairwise comparisons rather than absolute rewards. In this framework, an agent selects two options (or "arms") to compare in each round, receiving binary feedback indicating which option is preferred. This approach is particularly advantageous in situations where obtaining absolute evaluations is challenging, but relative preferences are readily accessible. A seminal work in this domain is "The K-armed Dueling Bandits Problem" by Yue et al., which introduces a novel regret formulation and presents an algorithm achieving near-optimal regret bounds in the context of dueling bandits. Mohankumar and Khapra introduced the Active Evaluation framework, which leverages dueling bandit algorithms to minimize the number of necessary pairwise comparisons to evaluate NLG tasks. By actively selecting the most informative system pairs for comparison, this method can reduce human annotation efforts by up to 80%. Following the same idea, we apply deuling bandits algorithms in LLM reranking procedure.

3 Methodology

Figure 1 demonstrates our proposed framework. which streamlines literature retrieval, ranking, and integration within an LLM-powered academic writing pipeline. The process begins with a given research concept and a target manuscript section, followed by retrieving N relevant documents from a pre-constructed document embedding database.

To ensure the most relevant sources are prioritized, the LLM RAG Reranker employs pairwise comparisons to assign weights to the retrieved documents. Two selection strategies are used: an exhaustive pairwise comparison approach and an Upper Confidence Bound (UCB) Elimination (Mohankumar and Khapra, 2022) method that iteratively filters out less relevant documents based on statistical confidence bounds. The detailed algorithm is illustrated in Appendix A.1.

The reranked documents are then fed into the LLM writing module, which generates coherent and contextually appropriate content for the specified section by leveraging the ranked literature. The generated text undergoes further evaluation using standard evaluation metrics or LLM-based assessment to ensure the quality and relevance of the synthesized content. This framework enhances the efficiency and accuracy of integrating prior research, addressing the challenges posed by the exponential growth of academic literature.

In-link context is selected according to the target section and the BERT similarity between the context and the research idea.

3.1 Pairwsie-Comparison-based LLM RAG Reranker

Motivation for Pairwise Comparisons. Incorporating all retrieved documents into a single LLM prompt can lead to context overflow, surpassing the model's maximum input length and resulting in truncated or incomplete processing. Moreover, such an approach may obscure the rationale behind ranking decisions, reducing interpretability. To mitigate these issues, our reranker employs pairwise comparisons, evaluating documents two at a time. This method not only aligns with the LLM's input constraints but also enhances the transparency of the decision-making process, as each comparison provides a clear basis for preference. Recent studies (Qin et al., 2024) have demonstrated that pairwise ranking prompting effectively reduces task complexity and improves alignment with human

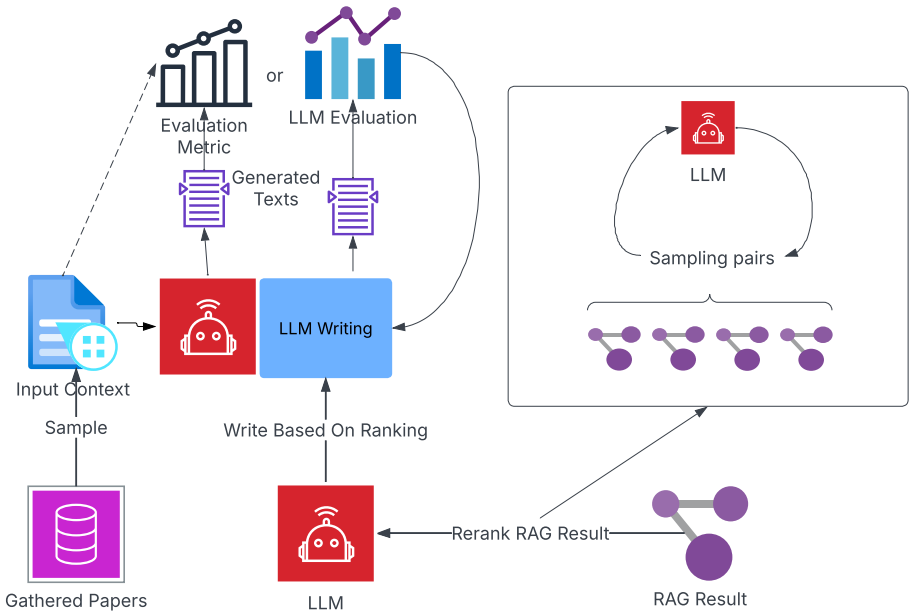

Figure 1: **Our proposed LLM RAG Reranker and RAG-powered LLM paper writing pipeline.** Proposed framework integrating an LLM RAG Reranker and a RAG-powered LLM paper writing pipeline. Starting with a research concept and specified section, the system retrieves N relevant documents from a document embedding database. The LLM RAG Reranker then assigns weights to these documents through pairwise comparisons, utilizing two strategies: an exhaustive approach comparing all pairs, and an Upper Confidence Bound (UCB) Elimination method that iteratively discards less relevant documents based on calculated confidence bounds. The reranked papers are then fed into the LLM writing module, which synthesizes coherent and contextually relevant content for the specified section.

judgment in ranking tasks.

Challenges of Pairwise Comparison: Computational Complexity. A notable limitation of exhaustive pairwise comparisons is their computational complexity, which scales quadratically with the number of documents $O(N^2)$. This can become computationally prohibitive as the size of the document set increases. Following Mohankumar and Khapra's work on NLG evaluation with the dueling bandits algorithm, we integrate the Upper Confidence Bound (UCB) Elimination algorithm, which strategically reduces the number of necessary comparisons. By focusing on the most informative pairs, the UCB Elimination method iteratively filters out less relevant documents, thereby optimizing the reranking process without compromising accuracy

Upper Confidence Bound (UCB) Elimination Strategy. The UCB Elimination algorithm operates by maintaining confidence bounds for each document's relevance score. In each iteration, it selects document pairs with the highest potential for informative comparison, based on their upper confidence bounds. Documents that consistently demonstrate lower relevance are progressively eliminated from consideration. This approach ensures that computational resources are concentrated on evaluating the most promising documents, effectively balancing exploration and exploitation in the reranking process.

By integrating pairwise comparisons with the UCB Elimination strategy, our framework enhances the efficiency and interpretability of literature selection, facilitating the incorporation of high-quality, relevant sources into the academic writing pipeline.

3.2 In-link Context Augmented RAG Reranking

Challenges in Citation Relevance Assessment. LLMs may encounter difficulties in accurately determining the relevance of a paper for citation purposes when provided solely with the paper's title and abstract. This limitation arises because titles and abstracts often lack comprehensive contextual information, making it challenging to assess how the paper has been utilized or referenced in existing literature.

Leveraging In-link Contexts for Improved Reranking To address this challenge, we inte-

grate in-link contexts into the reranking mechanism. As defined in Section 1, in-link contexts refer to the textual environments within existing papers where a candidate paper has been cited. By analyzing these contexts, the model gains insights into the specific ways a candidate paper has been employed in scholarly discourse, thereby enhancing the accuracy of relevance assessments.

Implementation in the Reranking Process In practice, the reranking process involves augmenting the candidate papers' metadata with their corresponding in-link contexts. The LLM RAG Reranker then evaluates these enriched representations, performing pairwise comparisons to assign relevance weights more effectively. This method not only improves the precision of the reranking process but also ensures that the selected literature is contextually aligned with the research at hand.

By incorporating in-link citation contexts, our framework addresses the limitations of relying solely on titles and abstracts, thereby refining the selection of pertinent literature and enhancing the overall quality of the academic writing pipeline.

4 Evaluation

4.1 Evaluation Data Collection

We collect more than 10,000 articles using the ArXiV API. The data set is a curated data set of academic articles, incorporating textual content, figures, tables, and metadata to facilitate research on scholarly document analysis. The raw LaTeX files of the papers were collected from ArXiv[1] using the ArXiv API, in accordance with the papers' licenses, which are primarily CC0 or CC-BY 4.0, permitting redistribution and sharing.

Using a curated selection of survey papers from diverse areas within Artificial Intelligence (AI)—including Natural Language Processing (NLP), Computer Vision (CV), Bioinformatics, and Robotics—we employ a breadth-first search (BFS) algorithm to collect cited papers. For each paper in the BFS queue, we traverse the abstract syntax tree generated from its LaTeX source to extract citation reference keys and the associated citation context surrounding the macro command $\backslash cite$. Subsequently, we utilize the corresponding .bib file to map each reference key to its associated metadata.

Based on the title and author information obtained from the metadata, we query the arXiv API to retrieve the cited paper. In instances where a paper is not available in the arXiv database, we calculate a string similarity score between the title provided in the .bib file and that of the retrieved paper, retaining only those with a similarity score exceeding 0.9. Additionally, we enforce a temporal constraint, including only those cited papers published between January 1, 2018, and December 31, 2024. Papers that satisfy these criteria are subsequently appended to the end of the BFS queue.

A detailed statistical overview of the collected dataset is presented in Table 1.

Table 1: **Data statistics for Paper-10k.**

Dataset	#Paper	#Citation context
Paper-10k	10,395	497,210

4.2 Evaluation Metrics

RAG system. To assess the efficacy of our RAG system, we employ precision and recall as primary evaluation metrics. Precision measures the proportion of retrieved documents that are relevant, while recall quantifies the proportion of relevant documents successfully retrieved. For the reranking component, we evaluate performance by computing the Mean Reciprocal Rank (MRR), which reflects the average rank position of the positive document across multiple queries. $MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$ where set $Q = Ref \cap RAG$, Ref is the reference set and RAG is the retrieval set.

Paper writing. We employ three types of evaluation to assess both retrieval accuracy and the final text quality:

1. **Coverage and Hit Position.** We measure how often the retrieved references contain relevant information and at what ranks they appear.

2. **Factual Consistency.** We use two automated approaches:

   (a) **SummaCConv (L)** (**??**) to measure factual consistency between source texts (i.e., the original abstract plus references) and the generated content. We report the average scores of:

   - Original→Gold
   - Original→Generated
   - Gold→Generated

[1] https://arxiv.org/ We thank ArXiv for providing open access interoperability.

Table 2: **Precise and Recall score of Embedding-based RAG system**. All sections (avg. 34.1 citations), Introduction (avg. 11.1 citations), Related Work (avg. 19.8 citations).

	All		Intro		Related Work	
@K	Precise	Recall	Precise	Recall	Precise	Recall
5	0.335	0.083	0.197	0.129	0.261	0.093
10	0.269	0.122	0.150	0.181	0.202	0.142
20	0.203	0.170	0.108	0.240	0.150	0.197
50	0.132	0.251	0.064	0.335	0.095	0.291
100	0.091	0.329	0.042	0.423	0.064	0.381

Here, *Original* is the input file, *Gold* is our ground-truth introduction by academic authors, and *Generated* is the LLM-produced introduction. When the Original→Gold score is close to Original→Generated, and the Gold→Generated score is high, the writing quality is demonstrated to be high.

(b) **T5-based NLI (?)** to check whether the generated introduction is entailed by the source. We treat the paper's abstract plus references as the "premise" and the generated introduction as the "hypothesis," computing a binary entailment score.

3. **Semantic Similarity.** We compute a sentence-transformer (all-MiniLM-L6-v2) based cosine similarity between the gold introduction and the generated introduction. A high cosine similarity suggests a strong overlap in key points.

5 Experiments

5.1 Embedding-based RAG System

In our study, we employed the SPECTER2 (Singh et al., 2022) model to encode scientific documents. SPECTER2 is a state-of-the-art scientific document embedding model capable of generating task-specific embeddings when paired with appropriate adapters. e processed all collected papers by inputting their titles and abstracts into SPECTER2 to obtain dense vector representations.

For efficient similarity search within these embeddings, we utilized the FAISS (Douze et al., 2024) library. FAISS is designed for rapid similarity search and clustering of dense vectors, making it well-suited for handling large-scale datasets. By indexing the document embeddings with FAISS, we facilitated swift retrieval of relevant documents.

Table 2 presents the precision and recall metrics of our embedding-based RAG system across different sections of the papers. As observed, increasing the value of @K leads to a decrease in precision but an increase in recall. This trade-off suggests that while more documents are retrieved (higher recall), a smaller proportion of them are relevant (lower precision). To leverage the higher recall at larger @K values, a re-ranking strategy can be employed. Re-ranking involves applying a more precise model or heuristic to the initially retrieved set to improve the relevance of the top-ranked documents.

Table 3: **MRR score of RAG results on Related Work section**. In the strategy (Strat.) column, All denotes comparing all pairs. UCB denotes using UCB Elimination as the dueling bandit algorithm. UCB (In-link) denotes using UCB Elimination while augmenting the retrieved papers with in-link context.

		Related Work		
@K	Strat.	MRR (RAG)	MRR (Rerank)	Comp
5	All	0.485	**0.584**	10
	UCB		0.528	9.6
	UCB (In-link)		0.572	9.1
10	All	0.433	0.458	45
	UCB		0.454	10.3
	UCB (In-link)		**0.470**	11.2
20	All	0.413	**0.427**	190
	UCB		0.389	48.3
	UCB (In-link)		0.395	44.3
50	All	0.341	-	(1225)
	UCB		0.102	68.1
	UCB (In-link)		0.151	72.3
100	All	0.304	-	(4950)
	UCB		0.116	73.8
	UCB (In-link)		0.132	69.2

5.2 Reranking Performance

We evaluated three re-ranking strategies: *All Pairs Comparison* (All), which involves exhaustive pairwise comparisons; *UCB Elimination* (UCB), utilizing the Upper Confidence Bound (UCB) elimination algorithm; and *UCB Elimination with In-link Context* (UCB in-link), which augments UCB elimination with in-link contextual information. Our experiments were conducted on 20 samples, focusing on the Related Work sections.

Table 3 summarizes the Mean Reciprocal Rank (MRR) scores for the initial RAG retrieval and after applying each re-ranking strategy, along with the number of comparisons (#Comp) performed. The results indicate that:

- All Pairs Comparison: This approach enhances MRR but is computationally intensive, as the number of comparisons grows quadratically with the number of documents (i.e., C_n^2). This renders it impractical for large N.

- UCB Elimination: This strategy significantly reduces the number of comparisons while maintaining competitive MRR scores, especially for smaller @K values.
- UCB Elimination with In-link Context: Incorporating in-link context further improves MRR, even at higher @K values where UCB Elimination falls short, suggesting that additional contextual information enhances re-ranking effectiveness.

5.3 Academic Writing Pipeline

Overview. We adopt a Retrieval-Augmented Generation pipeline to produce academic introductions for a target paper. Specifically:

1. **Gathering Relevant Passages.** For each candidate paper in the top portion of the reranked list, we summarize its abstract or relevant sections.
2. **LLM-based Composition.** We prompt a large language model (e.g., Llama-2-13B) to write an "Introduction" section. The prompt includes the target paper's title and abstract, as well as short summaries (or key sentences) of the top-ranked references. We instruct the LLM to cite all references explicitly.
3. **Refinement.** The output is then optionally refined to ensure readability and consistency. For these experiments, however, we use the raw model generations for evaluation.

5.4 Results of Academic Writing

Academic Writing Quality. We run the pipeline on a set of 20 papers with their top-100 retrieved references. Our system composes the introduction sections, and we compare them to the gold references (the original or author-provided introduction).

The factual consistency evaluation with SummaCConv yields:

- Original→Gold: 0.65
- Original→Generated: 0.64

These results indicate that our automatically generated introductions remain relatively close in factual content to the gold text, though slightly lower than gold itself (as expected).

Using the T5-based NLI approach, we find that in many cases, the "Gold→Generated" scores reach or exceed 0.9 for entailment, demonstrating that the generated text rarely contradicts the source. Additionally, the semantic similarity between the gold and generated introduction is around 0.87 in one representative example, suggesting strong overlap in key points.

6 Discussion

In this section, we discuss the application of dueling bandit algorithms, specifically the Upper Confidence Bound (UCB) Elimination method, to reduce the number of pairwise comparisons in our re-ranking system. We also address the limitations encountered when applying these algorithms in scenarios with large values of K (the number of items to compare).

6.1 Applying Dueling Bandit Algorithm to Reduce Comparison

The UCB Elimination algorithm effectively reduces the number of pairwise comparisons required during the re-ranking process. By focusing on comparisons that are most informative, the algorithm eliminates suboptimal items early, thereby decreasing computational overhead. Our experimental results, as presented in Table 3, demonstrate that for smaller values of K, the performance of UCB Elimination is comparable to exhaustive pairwise comparison, achieving similar MRR scores with significantly fewer comparisons.

6.2 Limitations of Dueling Bandit Algorithms like UCB Elimination in Large K Settings

While UCB Elimination offers efficiency benefits, its performance diminishes as K increases. Unlike exhaustive pairwise comparison, which evaluates all possible pairs, UCB Elimination relies on sampling to estimate upper and lower confidence bounds. This sampling-based approach can lead to less accurate estimations in large K scenarios, as the algorithm may not sample each pair sufficiently to make precise decisions.

A critical factor influencing the accuracy of UCB Elimination is the reliability of the comparison outcomes. In our study, we utilized LLM to judge pairwise comparisons based solely on the titles and abstracts of papers. This limited context can result in less accurate judgments, as the LLM lacks comprehensive information about each paper's content. Our findings, highlighted in Table 3, indicate that augmenting the LLM with in-link citation context

enhances decision accuracy, leading to improved re-ranking performance.

6.3 RAG-powered Academic Paper Writing.

Observations from the Experiment. The quality of generated introductions remains high, with factual consistency scores (SummaCConv: 0.64) and strong semantic similarity (0.87 in representative cases) indicating that the RAG-based approach produces coherent and well-structured academic content. While the generated introductions align closely with author-provided text, minor factual inconsistencies suggest the need for improved integration of nuanced claims and domain-specific details. Additionally, citation handling presents a challenge. Although references are explicitly included, ensuring precise citation alignment remains an issue. The model may overgeneralize certain references or introduce slight misinterpretations, emphasizing the necessity of verification mechanisms for citation fidelity.

Future Improvements. To further enhance the academic writing pipeline, several improvements can be explored. 1) Guided content generation. Guided content generation could involve an intermediate step of generating structured writing guidelines before retrieving relevant sources. 2) Iterative retrieval and writing, using a multi-step process, could ensure a more structured and comprehensive integration of references. 3) Enhanced model training with domain-specific data and fine-tuning LLMs to academic writing patterns could improve coherence and factual accuracy.

6.4 Future work

To address these limitations, future research could explore several avenues:

- Enhanced Prompting Techniques: Developing more effective prompts for the LLM to provide richer context during comparisons. Our current approach utilizes separate citation contexts, which may provide limited information for the LLM during comparisons. To enhance the richness of context, we can adopt methods such as those proposed in ILCiteR (Roy and Han, 2024), which distill citation contexts to obtain grounded evidence for recommendations. By distilling and integrating citation contexts, we can provide the language model with richer, evidence-grounded information, potentially leading to more accurate and contextually appropriate citation recommendations.

- Alternative Dueling Bandit Algorithms: In Active Evaluation (Mohankumar and Khapra, 2022), 13 dueling bandits algorithms are evaluated on NLG tasks. Our research focused solely on the UCB Elimination algorithm, leaving at least 12 other algorithms such as Uncertainty-aware Selection, Random Mixing, and Bayesian Active Learning by Disagreement (BALD) unexplored. Adopting the suitable sampling algorithm may offer improved performance in large K settings.

- Utilizing Advanced Models: Employing more sophisticated models capable of processing extensive contextual information to enhance comparison accuracy.

7 Conclusion

We introduced the Comparison-based LLM RAG Reranker, enhancing academic literature retrieval using pairwise comparisons adjudicated by LLM. To improve efficiency, we integrated the Upper Confidence Bound (UCB) Elimination algorithm, reducing comparisons while maintaining retrieval performance. Incorporating in-link citation contexts further refined reranking accuracy.

Our experiments on a dataset of over 10,000 papers and 490,000 citation contexts showed improved performance over embedding-based retrieval in Mean Reciprocal Rank (MRR). Additionally, we developed a RAG-powered LLM academic writing pipeline, automating literature retrieval and synthesis.

While effective, our approach faces challenges in handling large-scale comparisons. Future work could explore alternative dueling bandit algorithms and improved prompting techniques to refine reranking and citation-based retrieval.

Limitation and Future Work

Our study was constrained by a limited budget and computational resources, allowing us to conduct experiments on only 20 samples in Reranking and Writing tasks. This small sample size may introduce experimental errors and limit the generalizability of our findings. To mitigate these limitations, future research should aim to increase the sample size to enhance the robustness and reliability of the results.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *Preprint*, arXiv:1903.10676.

Shijie Chen, Bernal Jimenez Gutierrez, and Yu Su. 2025. Attention in large language models yields efficient zero-shot re-rankers. In *The Thirteenth International Conference on Learning Representations*.

Xin Cheng, Xun Wang, Xingxing Zhang, Tao Ge, Si-Qing Chen, Furu Wei, Huishuai Zhang, and Dongyan Zhao. 2024. xrag: Extreme context compression for retrieval-augmented generation with one token. *Preprint*, arXiv:2405.13792.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. *Preprint*, arXiv:2004.07180.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Chanwoo Jeong, Sion Jang, Hyuna Shin, Eunjeong Park, and Sungchul Choi. 2019. A context-aware citation recommendation model with bert and graph convolutional networks. *Preprint*, arXiv:1903.06464.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-shot listwise document reranking with a large language model. *Preprint*, arXiv:2305.02156.

Akash Kumar Mohankumar and Mitesh Khapra. 2022. Active evaluation: Efficient NLG evaluation with few pairwise comparisons. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8761–8781, Dublin, Ireland. Association for Computational Linguistics.

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. *Preprint*, arXiv:2306.17563.

Sayar Ghosh Roy and Jiawei Han. 2024. Ilciter: Evidence-grounded interpretable local citation recommendation. *Preprint*, arXiv:2403.08737.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2022. Scirepeval: A multi-format benchmark for scientific document representations. In *Conference on Empirical Methods in Natural Language Processing*.

Shangyu Wu, Ying Xiong, Yufei Cui, Haolun Wu, Can Chen, Ye Yuan, Lianming Huang, Xue Liu, Tei-Wei Kuo, Nan Guan, and Chun Jason Xue. 2024. Retrieval-augmented generation for natural language processing: A survey. *Preprint*, arXiv:2407.13193.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. 2012. The k-armed dueling bandits problem. *J. Comput. Syst. Sci.*, 78(5):1538–1556.

A Appendix

A.1 UCB Elimination Algorithm

The procedure of UCB Elimination algorithm is shown in Algorithm 1.

Algorithm 1 UCB-Based Response Evaluation

function EVALUATERESPONSES($title, abstract, section, responses, top\_k$)
    Initialize candidates with indices and texts
    Set active candidates, scores, and details
    Define parameters: $\delta = 0.1$, $\epsilon = 0.1$, $max\_rounds = 500$, $min\_matches = 2$
    function UPDATEBOUNDS($candidate$)
        if $candidate.matches = 0$ **then**
            Set $candidate.ucb = 1.0$, $candidate.lcb = 0.0$
        else
            Compute bounds using Wilson score interval
        end if
    end function
    function SELECTPAIR($candidates$)
        if Random() < 0.3 **then**
            Return least-sampled candidate and a random choice
        else
            Return top two candidates by UCB
        end if
    end function
    $round\_count \leftarrow 0$
    while $|active\_candidates| > top\_k$ **and** $round\_count < max\_rounds$ **do**
        $a, b \leftarrow$ SelectPair($active\_candidates$)
        $result \leftarrow$ CompareResponses($title, abstract, section, a.text, b.text$)
        Update scores and matches based on result
        UpdateBounds(a) and UpdateBounds(b)
        Filter out eliminated candidates based on confidence bounds
        Increment $round\_count$
        if $round\_count \geq maximum\_attempts$ **then**
            Break
        end if
    end while
    Sort active candidates by LCB and return top k
    return top responses, scores, details, round count
end function

A.2 LLM Prompt Template

The prompt template used in the experiment is demonstrated in this section. We first introduce the prompt template used in pairwise comparison, as shown in Table 4. {Paper_A} and {Paper_B} will be filled with the paper information. By default, the paper information consists of title and abstract. When augmented with citation context, the paper information consists of title, abstract, and k in-link citation context examples. Each in-link citation context consists of the title of citing paper and the citation context.

Table 4: **Prompt P_t of Pairwise Comparison**.

(System) You are an expert in academic research. You are choosing literature for your own work. Your work's title is {Title} and abstract is {Abstract}. Now you are writing the {Section} section. Given two abstracts from two papers, decide which paper is better to cite when writing the {Section} section. When evaluating the following responses, please ignore their order and base your decision solely on the quality and content of each answer. Your reasoning should focus only on the content and not on whether an answer appears first or second. Please select the response that you would prefer to cite in your work. Please start your response with Ï prefer to cite paperänd provide a short explanation for your choice. (User) Paper A: {Paper_A} Paper B: {Paper_B}