Between the Drafts: An Evaluation Framework for Identifying Quality Improvement and Stylistic Differences in Scientific Texts

Danqing Chen, Ingo Weber, Felix Dietrich
Technical University of Munich
Munich, Germany
{chen.danqing, ingo.weber, felix.dietrich}@tum.de

Abstract

This study explores the potential of a lightweight, open-source Large Language Model (LLM), demonstrating how its integration with Retrieval-Augmented Generation (RAG) can support cost-effective evaluation of revision quality and writing style differentiation. By retrieving reference documents from a carefully chosen and constructed corpus of peer-reviewed conference proceedings, our framework leverages few-shot in-context learning to track manuscript revisions and venue-specific writing styles. We demonstrate that the LLM-based evaluation aligns closely with human revision histories—consistently recognizing quality improvements across revision stages and distinguishing writing styles associated with different conference venues. These findings highlight how a carefully designed evaluation framework, integrated with adequate, representative data, can advance automated assessment of scientific writing.

1 Introduction

Human evaluation remains essential and unavoidable for assessing the quality of texts. However, it is notoriously difficult to reproduce and often lacks consistency (Gillick and Liu, 2010; Clark et al., 2021). Recently, large language models (LLMs) have shown remarkable capabilities in handling unseen tasks by simply following task instructions (Chiang and Lee, 2023). In this paper, we explore whether such an ability of the LLMs can be used as an alternative to human evaluation. We prompt LLMs with targeted instructions to evaluate either the quality of revisions across different versions of a manuscript or the similarity of writing styles between texts. Specifically, we use LLMs to assess revision histories based on writing quality and infer likely conference affiliations based on writing style. We find that the LLM-generated evaluations align closely with actual arXiv revision histories and the known conference venues of the papers, indicating that the model can reliably capture both revision-driven quality improvements and venue-specific stylistic patterns.

Large Language Models, such as GPT, are capable of generating fluent and syntactically well-formed text, yet they often fall short in tasks that require precision and factual grounding, especially in domain-specific contexts (Lewis et al., 2020; Petroni et al., 2021). Retrieval-Augmented Generation (RAG) addresses this limitation by integrating external knowledge into the generation process, enabling models to produce content that is not only fluent but also context-aware (Lewis et al., 2020; Izacard and Grave, 2021; Borgeaud et al., 2022; Gao et al., 2024). This integration is particularly critical for scientific manuscript evaluation, which requires a deeper understanding of clarity and discipline-specific writing conventions.

Recent studies have highlighted that university students often lack the academic writing skills required for producing coherent and well-structured research papers and dissertations (Phyo et al., 2023; Aitchison et al., 2012; Barbero, 2008; Cargill et al., 2012; DeLyser, 2003; Luo and Hyland, 2016; Surratt, 2006; Yu and Jiang, 2022). Therefore, we hope this evaluation framework can also assist researchers in the field of machine learning, and potentially in other fields, with manuscript optimization by providing insights into quality variation across manuscript revisions and stylistic alignment with target publication venues.

Our key contributions are:

1. *A data-driven, computational evaluation framework that uses LLMs (with RAG and few-shot prompting) to assess revision quality improvement and stylistic variation.*

2. *A locally deployable and cost-effective tool to support independent manuscript composition and refinement.*

Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems, pages 66–84

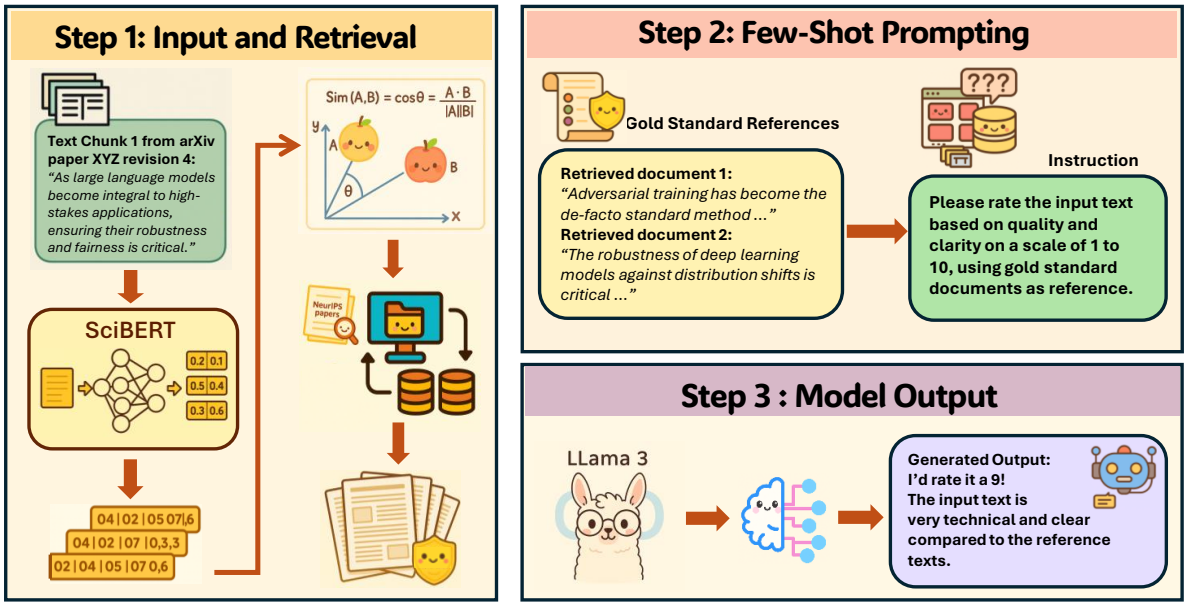


Figure 1: A running example for quality evaluation using few-shot in-context prompting in the RAG framework, with a numerical scale representing quality. The input text and gold standard documents in this figure are for illustration purposes only. For writing style evaluation, the prompt would change, explicitly instructing the LLM to rate on the similarity of writing style based on gold-standard references.

2 Related Work

LLMs have transformed NLP by enabling fluent, human-like text generation (Devlin et al., 2019; et al., 2018; Radford et al., 2019; Brown et al., 2020). However, their capacity remains limited, particularly in domain-specific and knowledge-intensive tasks where access to relevant external data is crucial for understanding beyond surface-level text and generating contextually appropriate responses (Lewis et al., 2020; Petroni et al., 2021). Additionally, state-of-the-art LLMs are prone to generating hallucinations, compromising reliability (Maynez et al., 2020; Perković et al., 2024; Ji et al., 2023a; Yao et al., 2024; Marcus, 2020; Zhang et al., 2022, 2023).

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) addresses key challenges by integrating external knowledge sources to reduce hallucinations and improve accuracy (Borgeaud et al., 2022; Shuster et al., 2021; Jiang et al., 2023; Bhat et al., 2024; Fan et al., 2024). RAG has proven effective across domains by enhancing factual grounding in generative models. For generative retrieval, CorpusLM combines generative retrieval to enhance performance in knowledge-intensive tasks (Li et al., 2024). TC-RAG (Jiang et al., 2024) demonstrates RAG's benefits in medical applications, reducing hallucinations and boosting accuracy. In image generation (Sheynin et al., 2023), large-scale retrieval facilitates cross-modal content modeling without explicit supervision.

There has been extensive exploration of knowledge-grounded generation leveraging various forms of knowledge, such as knowledge bases and external documents (Dinan et al., 2019; Zhou et al., 2018; Lian et al., 2019; Li et al., 2019; Qin et al., 2019; Zhang et al., 2022). The current state-of-the-art practice for utilizing RAG, called Vector-RAG, often employs vector databases for efficient information retrieval (Sarmah et al., 2024).

Numerous state-of-the-art vector representation models have been developed over the years. Word2Vec (Mikolov et al., 2013b,a) and GloVe (Pennington et al., 2014), produce a single embedding for each word, regardless of the context (Gupta and Jaggi, 2021; Rahimi and Homayounpour, 2021), making static word embeddings fall short in the task of scientific text retrieval compared to contextual embeddings, which provide different embeddings for the same word depending on the surrounding context (Peters et al., 2018). Contextual models have been shown to perform better in scenarios that require deeper semantic understanding (Zhou and Bloem, 2021; Peters et al., 2018; Liu et al., 2025, 2020; Apidianaki, 2023).

The performance of a machine learning system depends heavily on data representation (Le-Khac

et al., 2020). SciBERT (Beltagy et al., 2019), pre-trained on scientific text, has shown strong results across scientific NLP tasks. It has been used in paper recommendation systems that leverage SciBERT embeddings derived from arXiv abstracts (Singh et al., 2023), and has outperformed other models in citation classification (Maheshwari et al., 2021). Its role in the iFORA system for trend detection highlights its utility in text mining (Lobanova et al., 2024). In summarization tasks, the COVIDSum model used SciBERT to generate high-quality abstracts from COVID-19 papers (Cai et al., 2022), outperforming other approaches. SciBERT also excelled in relation extraction (Poleksic and Martincic-Ipsic, 2023) and citation intent classification (Motrichenko et al., 2021). These applications demonstrate SciBERT's value in scientific text processing, making it well-suited for scientific document retrieval tasks.

LLMs can handle complex tasks via few-shot in-context learning, leveraging prompt engineering rather than parameter adjustments, and have been shown to improve the understanding and reasoning of LLMs from a few examples in the context (Wei et al., 2022; Dong et al., 2024; Liu et al., 2022). This paradigm has been applied in domains such as autonomous vehicle training (Zhang et al., 2024), example-based retrieval (Rubin et al., 2022), automated assessment of translation quality (Kocmi and Federmann, 2023), and character generation (Lake et al., 2015). This shift has driven research into improving LLM reasoning through strategic prompting rather than model parameter updating (Stahl et al., 2024; Arora et al., 2023).

3 Experimental Setup

3.1 Data

Given that the effectiveness of retrieval-augmented text generation is closely tied to the quality and relevance of the retrieved content (Li et al., 2022), it is essential to construct the retrieval corpus from a well-established, peer-reviewed publication venue within the specific domain (in this case, machine learning) to ensure a reliable and domain-representative knowledge base for evaluation (for both quality improvement identification and conference-specific stylistic differentiation). Furthermore, prior work demonstrated that dataset size plays a significant role in retrieval performance (Hawking and Robertson, 2003), specifically, using a larger retrieval database during inference improves model performance (Shao et al., 2024). NeurIPS is one of the most prestigious conferences in machine learning and has consistently received high submission volumes in the field, surpassing ICLR and ICML in recent years[1]. To this end, we constructed our retrieval vector database using the full proceedings of NeurIPS 2023 (papers from 2024 were excluded due to incomplete proceedings at the commencement of this study).

For evaluation, papers were randomly collected from arXiv,[2] selecting version 1 (v1) and version 4 (v4) of each paper to analyze quality improvements across revisions. For conference writing style differentiation, proceedings from NeurIPS, ICLR, and ICML (all from the year 2023) were also randomly sampled. Additionally, Amazon reviews[3] were used to examine how LLMs respond to informal language in contrast to scientific writing as a baseline check (Appendix C.2).

The retrieval vector database was constructed by segmenting the text from each NeurIPS paper and encoding the segments into reasonably long, fixed-length SciBERT embeddings. These embeddings were then indexed using FAISS (Facebook AI Similarity Search)[4] to enable efficient similarity search and retrieval. The resulting indexes and embeddings were collected to form the complete retrieval vector database. More details on data preprocessing are provided in Appendix B. The NeurIPS proceedings in this study are sourced from a publicly available dataset on Kaggle.[5] Prior studies have utilized NeurIPS text datasets from Kaggle for topic modeling and text classification (Terko et al., 2019). A similar analysis was performed on ICLR papers by extracting textual features (Joshi et al., 2021). Prior studies have also used papers from arXiv for open-source dataset construction (Clement et al., 2019) and model training (Shabtay et al., 2025). Therefore, this study was conducted using publicly available data, in compliance with established and common practices.

[1] Submission statistics available at: https://papercopilot.com/, https://media.neurips.cc/Conferences/NeurIPS2023/NeurIPS2023-Fact_Sheet.pdf

[2] https://arxiv.org/

[3] https://www.kaggle.com/datasets/kritanjalijain/amazon-reviews

[4] https://github.com/facebookresearch/faiss

[5] https://www.kaggle.com/datasets/mohamednennouche/neurips-papers-1987-2023

Experiment	Revision Quality Improvement Identification		
	Retrieval Database	**#Papers to Rate**	**GPU Type**
	Entire NeurIPS23 dataset	20 (1^{st} & 4^{th} revisions)	NVIDIA A100
Experiment	**Conference Writing Style Distinction**		
	Retrieval Database	**#Papers to Rate**	**GPU Type**
	Entire NeurIPS23 dataset	15 per conference	NVIDIA A100

Table 1: The table presents details of each experiment, including the dataset used to construct the retrieval database, the number of papers used as input for rating, and the GPU type utilized.

3.2 Model Choice

This study employs LLaMA-3.0-8b-instruct (Dubey et al., 2024), a variant of the LLaMA 3.0 model family. The LLaMA 3.0 family includes model configurations with 8B and 70B parameters. The 8B model was chosen to balance hardware constraints with task requirements, as generating ratings (a numerical representation of quality improvement or stylistic similarity, see Section 3.3) and limited suggestions do not necessitate a 70B model, and the 8B configuration allows for possible local deployment on consumer-grade hardware. GPT and other closed-sourced, proprietary models were not considered for privacy and data protection reasons. Beyond identifying quality improvements across revisions and distinguishing writing styles, we also aim to showcase this evaluation framework's potential for academic manuscript refinement, and since authors often prioritize confidentiality during submission and peer review, they may hesitate to use closed-source models for evaluating quality or writing style. Therefore, this study utilizes a locally deployable, lightweight, open-source model, enabling authors to conduct assessments independently. All experiments were conducted on a local computer using a personal Google Colab account to demonstrate the system's local deployability on consumer-grade hardware.

Following a thorough evaluation, both empirical and based on relevant literature reviews, LLaMA was chosen over other open-source alternatives. Due to limited computational resources, fine-tuning was not conducted in this study. Consequently, model selection was carried out with careful consideration to balance performance and efficiency. The LLaMA 3.0 family was selected for this study as it represented the most recent iteration of the LLaMA models available at the time this study commenced. The instruction-tuned version (LLaMA-3.0-8b-instruct) was selected based on empirical observations, demonstrating superior performance compared to the base model LLaMA 3.0.

It is important to note, however, that the primary goal of this study is to design a data-driven, computational evaluation framework, integrated with a domain-relevant retrieval database, capable of identifying quality improvements, writing style differences, and serving as a locally deployable tool for independent and cost-effective manuscript assessment. While our current implementation demonstrates this capability using a specific model, the framework is model-agnostic in principle and can be adapted to incorporate other models should they prove more suitable for particular use cases, this is further demonstrated empirically in an ablation study (Appendix C.4), where a university-hosted Copilot instance shows consistent scoring pattern and yields overall scores closely matching those of LLaMA. Similarly, this architecture is not limited to the field of machine learning; in principle, it can be applied to other domains as well when combined with a curated retrieval vector database containing relevant scientific texts tailored to the specific field.

3.3 Scoring Scientific Writing with Retrieval-augmented Generation

Scientific writing standards vary widely across disciplines, making objective evaluation difficult. To address this, we use a Retrieval-Augmented Generation approach that retrieves relevant texts from NeurIPS proceedings as high-quality references. These guide an LLM in assessing input text quality or style, grounded in peer-reviewed examples rather than fixed evaluation criteria. This enables a data-driven, implicit understanding of clarity, quality, or venue-specific writing style.

For full-text paper assessments (revision quality improvement identification and conference writing

style differentiation), each paper is segmented into reasonably long chunks. These chunks are individually evaluated using the RAG system.[6] The final score for each paper is calculated by averaging the scores across all chunks. To assess the input text, the system first encodes each input text chunk using SciBERT and retrieves the top two most similar documents from a vector database using cosine similarity. These documents serve as "gold standard" references. To form the prompt, the retrieved references are first combined with the input text. This is then followed by explicit instructions directing the model to rate the input on a scale from 1 to 10, based either on its similarity in quality (for evaluating revision quality) to the references or its stylistic resemblance (for evaluating conference-specific writing styles). The exact prompts used for the experiments are described in Appendices A.4 and A.5, leveraging few-shot in-context learning, instructing the model to evaluate the input texts (either from different arXiv revisions or different ML conferences) based on retrieved references rather than scoring the input text in isolation. Since the smaller LLaMA models, as well as many other LLMs, are highly sensitive to prompting (Wei et al., 2022; Zhou et al., 2024; Sclar et al., 2024; Arora et al., 2023; Turpin et al., 2023), the prompts used in this paper were rigorously tested and refined to ensure reliable rating generation. A running example is provided in Figure 1.

It is important to note that the primary objective of this task is to assign numerical ratings, with textual suggestions serving as supplementary evidence. Given the constraints of an 8B parameter model, the authors have determined that numerical outputs are more reliable and interpretable than extended textual feedback.

The experimental parameters are summarized in Table 1. The selection of the number of papers used to generate ratings for this study was determined to balance computational efficiency with the need for statistically meaningful results. Given that large language model inference for text generation tasks is computationally intensive, resource constraints were carefully considered. In addition, a baseline check (Appendix C.2) was first conducted to validate the model's ability to distinguish the difference between scientific and non-scientific writing. This is crucial since identifying quality improvement across revisions or differentiating writing style assumes that the system can first distinguish scientific vs. non-scientific writing before making more nuanced distinctions. A consistency check of the ratings (Appendix C.1) was also conducted, which demonstrates the system's consistency in its scoring behavior.

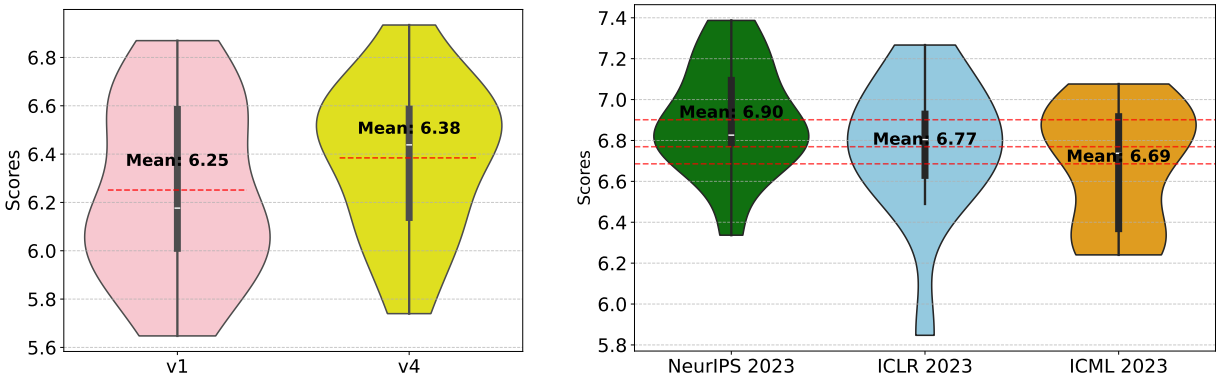
[6]For the purpose of simplicity, the term "RAG system" or "RAG framework" will refer specifically to the LLaMA-3.0-8b-instruct model integrated with a retrieval vector database constructed from NeurIPS 2023 proceedings.

4 Results and Discussion

4.1 Revision Quality Improvement Identification

This section presents the experimental results of revision quality improvement identification, following the methodology in Section 2, with more preprocessing details in Appendix B. Papers were randomly selected from the Machine Learning category on arXiv, with each paper having undergone at least four revisions to ensure meaningful differences across versions and processed through the RAG system. The system evaluated paper quality based on retrieved reference documents. Please note that the use of revised versions (e.g., v1 and v4) as labels effectively serves as a form of human annotation, as such revisions typically result from deliberate, human-driven improvements, usually incorporating expert peer-review suggestions or professional feedback. This provides a natural supervision signal, with later versions usually reflecting higher quality, making additional human expert annotation unnecessary. An example prompt in Appendix A.4 demonstrates a few-shot in-context strategy, guiding the LLM to assess text quality and clarity using NeurIPS papers as an implicit anchor for "good" scientific writing. To ensure fairness, the most similar retrieved document was excluded, as some arXiv papers may originate from NeurIPS.

As shown in Figure 2a, the plot compares RAG system scores for the first and fourth revisions of 20 manuscripts. The notable increase in mean scores from 6.25 (v1) to 6.38 (v4) suggests that the system can differentiate between earlier and refined versions, capturing improvements made during the revision process. By going beyond surface-level text and capturing the difference in quality and clarity between earlier and refined versions, this evaluation methodology also lays the groundwork for providing targeted, content-aware feedback to support manuscript refinement. Additionally, a chunk-based analysis was conducted (Section 4.3), highlighting the section-specific improvements during



(a) Score distributions for 20 arXiv manuscripts, comparing first (v1) and fourth (v4) revisions.

(b) Score distributions for 45 randomly selected papers from NeurIPS, ICLR, and ICML, 15 papers per conference.

Figure 2: Comparison of RAG-generated score distributions : (a) Revision quality improvement identifications, and (b) Conference style distinction. The black bar shows the interquartile range, the red dashed line indicates the mean, and the small white line marks the median.

manuscript revision captured by the RAG system.

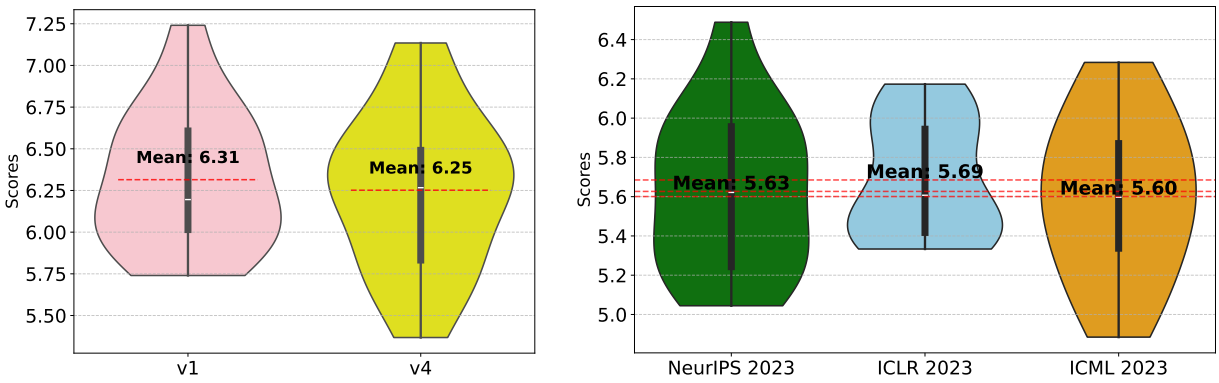
Validating the Impact of RAG on Revision Quality Improvement Identification

To assess the influence of retrieval-augmented generation on the system's ability to identify revision improvement, an additional set of experiments was conducted using the same set of arXiv papers on the same revision stages. In this setup, the revision analysis task was conducted without the use of reference documents (an example prompt is provided in Appendix A.2). This design allows us to isolate the impact of retrieval augmentation by intentionally omitting the contextual grounding offered by the retrieved reference context. The results of this experiment are presented in Figure 3a.

Our findings highlight RAG's role in distinguishing revision quality. With retrieval (Figure 2a), the mean revision score increased from 6.25 (v1) to 6.38 (v4), while without retrieval, scores remained slightly declined (6.31 in v1 vs. 6.25 in v4). This suggests that retrieval-based generation provides essential context for recognizing manuscript improvements. When relying solely on generative capabilities without retrieval, the LLM fails to differentiate between improved and non-improved versions of the manuscript. In some cases, it even assigned slightly lower scores to objectively enhanced revisions, indicating a lack of sensitivity to quality improvements in scientific writing. One explanation for this may be linked to the phenomenon of "hallucination," where Natural Language Generation models frequently produce context that is incoherent or nonsensical (Levin et al., 2024; Ji et al., 2023a; Xiao and Wang, 2021; Ji et al., 2023b; Maynez et al., 2020) (a real-life example of such a phenomenon is provided in Appendix D). While the scores here are not as extreme as fully incoherent, the inaccurate scores without the retrieved documents may suggest a degree of "hallucination," highlighting the need for a retrieval database. Previous work by (Lewis et al., 2020) demonstrates that integrating external contextual information during text generation improves accuracy and contextual grounding. These results highlight the importance of retrieval mechanisms in enabling language models to move beyond surface text and more effectively identify and evaluate quality improvement between revisions during manuscript evaluation.

4.2 Conference Writing Style Distinction

This section analyzes the scores generated from the RAG system to assess alignment with conference affiliations, expecting higher ratings for papers when referenced against retrieved texts from the same conference. This experiment demonstrates the RAG system's ability to capture the differences in conference-specific writing styles (similarly to Section 4.1, the conference affiliation itself serves as an implicit form of human supervision, as submission and acceptance into specific conferences reflect the formality of the writing). The intuition behind this experiment is that if a given paragraph is semantically similar to a paragraph from a NeurIPS paper, it is likely to share a similar writing style. This approach leverages the connection between the semantic content of text and its stylistic characteristics and is based on the heuristic that when LLMs are explicitly prompted to evaluate writing style in comparison to a reference document based on similarity, they are more



(a) Revision quality improvement identification experiment conducted without using reference documents.

(b) Conference writing style distinction experiment conducted without using reference documents.

Figure 3: Experiments conducted without retrieval augmentation: (a) Revision quality improvement identification, (b) Conference writing style distinction.

likely to assign higher scores to input texts that closely resemble the style of the reference. Recent work supports this heuristic, showing that LLMs can effectively achieve text style transfer using prompt learning (Liu et al., 2024). Related research in authorship identification used prompt engineering to guide LLMs in identifying whether two texts share the same author by focusing on writing style (Huang et al., 2024), achieving great results. Few-shot learning has also been applied to detect machine-generated text using style representation (Soto et al., 2025).

This experiment follows the methodology in Section 2, with more preprocessing details in Appendix B and an example prompt in Appendix A.5. The prompt was carefully crafted to guide the LLM to evaluate inputs based on stylistic alignment, rather than factors such as overall quality or clarity. A few-shot in-context prompting strategy was utilized, leveraging reference documents to implicitly define writing style, similar to the approach used to define "good" scientific writing in Section 4.1 and appendix C.2. Given its effectiveness in that context, the same strategy was deemed appropriate for defining and distinguishing writing style in this experiment. NeurIPS 2023 proceedings serve as the retrieval database. The input comprises 15 randomly selected accepted papers (for each conference) from NeurIPS, ICLR, and ICML. These conferences were specifically chosen due to their similar research focus, ensuring that the results are not skewed by differences in research focus or domain variations. To ensure fairness, the most similar retrieved reference text was excluded from the evaluation of NeurIPS papers for this experiment.

The result of the experiment can be found in Figure 2b. The result highlights the RAG system's sensitivity to stylistic alignment with conference affiliations. As expected, NeurIPS demonstrates a more concentrated score distribution at the higher end, with the highest mean among the compared venues, indicating that its writing style naturally aligns more with the reference documents (also from NeurIPS). In contrast, ICLR shows a wider spread of scores extending towards lower values. ICML received a lower mean than both NeurIPS and ICLR. To validate the RAG system's ability to differentiate writing style, the same experiment using ICLR papers as the retrieval vector database can be found in Appendix C.3, further validating our framework's reliability.[7] These findings demonstrate the RAG system's capability to distinguish differences in writing style across manuscripts from different publication venues. They also highlight the potential of this evaluation framework in serving as a tool to assist authors in tailoring their manuscripts to venue expectations, helping them present their work in a way that is easier for the relevant community to understand and engage with.

Validating the Impact of RAG on Conference Writing Style Distinction

To assess the impact of the retrieval vector database on distinguishing writing styles between conferences, we conducted an additional set of experiments using the same set of proceedings. In this setup, the writing style differentiation task was re-

[7]To ensure a fair assessment, the authors of this study took all possible measures to verify that the retrieved documents are from different papers, preventing stylistic similarities from the same author.

Figure 4: Comparison of individual text chunk scores between Revision 1 and Revision 4 of the same arXiv paper. The plot shows a noticeable improvement in both the individual chunk score distribution and the overall average in Revision 4, indicating enhanced overall quality across the revised segments.

peated without incorporating reference documents, allowing us to isolate the impact of retrieval augmentation. The results of this retrieval-free experiment are presented in Figure 3b, and the corresponding prompt is detailed in Appendix A.3.

In the absence of retrieval, the scores diverged significantly from those observed in the RAG-enhanced setup (Figure 2b). Notably, ICLR papers received the highest scores, rather than NeurIPS papers, underscoring the critical role of the retrieval vector database and reference documents in supplying semantically relevant context. These results highlight the importance of retrieval in providing domain-specific grounding that enhances the accuracy of stylistic differentiation.

4.3 Chunk-based Revision Scores Analysis

This section presents an analysis of chunk-level scores generated by the RAG system for Revisions 1 and 4 of the same arXiv paper. As shown in fig. 4, individual text chunk scores from Revision 4 (right) consistently outperform those from Revision 1 (left). This demonstrates the system's ability to identify quality improvements both at the overall paper level and within individual sections. The results also demonstrate how fine-grained and sectional feedback can guide targeted revisions, enhancing overall quality. By identifying and addressing localized weaknesses at the chunk level, this approach offers a data-driven method for improving the quality of academic texts, highlighting the potential of this evaluation framework to support iterative writing refinement by providing section-specific and targeted feedback during manuscript optimizations.

Conclusion

This study introduces a locally deployable, data-driven, and entirely open-source evaluation framework for identifying quality improvements across manuscript revisions and stylistic variations across proceedings from different machine learning conferences. By integrating a carefully constructed and curated retrieval vector database, the proposed approach demonstrates its effectiveness by accurately identifying revision-based improvements in arXiv submissions at both the overall and section-specific levels, while also distinguishing writing styles across different venues. These contributions underscore the potential of this evaluation framework to support independent and cost-effective manuscript composition and refinement in academic writing.

Limitations

This study was constrained by limited computational resources. All experiments were conducted on a personal Colab account, not only to emphasize the cost-effectiveness but also the local deployability of the proposed evaluation framework; therefore, larger LLMs (e.g., LLaMA-3.0-70B) were not used. In addition, due to limited computational resources and copyright restrictions on academic papers, fine-tuning was not performed, even though it could have further improved evaluation accuracy. Non-textual elements like figures and results,

key to peer review, were also excluded. In this study, we rely solely on few-shot in-context learning using reference documents retrieved by RAG for manuscript evaluation. While effective in this setup, this approach may not generalize well or provide accurate evaluations in other contexts. Furthermore, the arXiv papers were sampled randomly, without accounting for whether some arXiv papers were already of high quality or underwent minimal revision, cases in which the system may not detect noticeable improvements in writing quality. We did not incorporate other open-source models in this study due to computational constraints, which limited our ability to conduct large-scale evaluations or ablation studies across multiple models. In addition, this study only focuses on the field of Machine Learning.

Despite the limitations, we hope readers recognize our effort to develop an evaluation framework that lays the foundation of cost-effective and independent manuscript assessment, as well as our attempt to demonstrate the potential of utilizing entirely open-source NLP-driven tools and publicly available datasets, in enhancing scientific communication practices.

References

Claire Aitchison, Janice Catterall, Pauline Ross, and Shelley Burgin. 2012. 'tough love and tears': Learning doctoral writing in the sciences. *Higher Education Research & Development*, 31(4):435–447.

Marianna Apidianaki. 2023. From word types to tokens and back: A survey of approaches to word meaning representation and interpretation. *Computational Linguistics*, 49(2):465–523.

Simran Arora, Avanika Narayan, Mayee F Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, and Christopher Re. 2023. Ask me anything: A simple strategy for prompting language models. In *The Eleventh International Conference on Learning Representations*.

Ever J Barbero. 2008. Journal paper requirement for phd graduation. *Latin American & Caribbean Journal of Engineering Education*, 2(2).

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.

Vani Bhat, Sree Divya Cheerla, Jinu Rose Mathew, Nupur Pathak, Guannan Liu, and Jerry Gao. 2024. Retrieval augmented generation (rag) based restaurant chatbot with ai testability. In *2024 IEEE 10th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService)*, pages 1–10. IEEE.

Sebastian Borgeaud, Arthur Mensch, Hoffmann, et al. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Xiaoyan Cai, Sen Liu, Libin Yang, Yan Lu, Jintao Zhao, Dinggang Shen, and Tianming Liu. 2022. Covidsum: A linguistically enriched scibert-based summarization model for covid-19 scientific papers. *Journal of Biomedical Informatics*, 127:103999.

Margaret Cargill, Patrick O'Connor, and Yongyan Li. 2012. Educating chinese scientists to write for international journals: Addressing the divide between science and technology education and english language teaching. *English for Specific Purposes*, 31(1):60–69.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. 2019. On the use of arxiv as a dataset. *Preprint*, arXiv:1905.00075.

Dydia DeLyser. 2003. Teaching graduate students to write: A seminar for thesis and dissertation writers. *Journal of Geography in Higher Education*, 27(2):169–181.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1*, pages 4171–4186.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Radford Alec et al. 2018. Improving language understanding by generative pre-training. *OpenAI*.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 6491–6501.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148–151, Los Angeles. Association for Computational Linguistics.

Prakhar Gupta and Martin Jaggi. 2021. Obtaining better static word embeddings using contextual embedding models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5241–5253. Association for Computational Linguistics.

David Hawking and Stephen Robertson. 2003. On collection size and retrieval effectiveness. *Information Retrieval*, 6:99–105.

Baixiang Huang, Canyu Chen, and Kai Shu. 2024. Can large language models identify authorship? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 445–460, Miami, Florida, USA. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880. Association for Computational Linguistics.

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

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023b. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843. Association for Computational Linguistics.

Xinke Jiang, Yue Fang, Rihong Qiu, Haoyu Zhang, Yongxin Xu, Hao Chen, Wentao Zhang, Ruizhe Zhang, Yuchen Fang, Xu Chu, Junfeng Zhao, and Yasha Wang. 2024. Tc-rag:turing-complete rag's case study on medical llm systems. *Preprint*, arXiv:2408.09199.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

Deepali J. Joshi, Ajinkya Kulkarni, Riya Pande, Ishwari Kulkarni, Siddharth Patil, and Nikhil Saini. 2021. Conference paper acceptance prediction: Using machine learning. In *Machine Learning and Information Processing*, pages 143–152.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. 2020. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934.

Gabriel Levin, Sabrina Piedimonte, and Behrouz Zand. 2024. Navigating the complexities of artificial intelligence in scientific writing: a dual perspective.

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. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022. A survey on retrieval-augmented text generation. *Preprint*, arXiv:2202.01110.

Xiaoxi Li, Zhicheng Dou, Yujia Zhou, and Fangchao Liu. 2024. Corpuslm: Towards a unified language model on corpus for knowledge-intensive tasks. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 26–37.

Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019. Incremental transformer with deliberation decoder for document grounded conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 12–21. Association for Computational Linguistics.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *IJCAI International Joint Conference on Artificial Intelligence*, page 5081.

Pusheng Liu, Lianwei Wu, Linyong Wang, Sensen Guo, and Yang Liu. 2024. Step-by-step: Controlling arbitrary style in text with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15285–15295, Torino, Italia. ELRA and ICCL.

Qi Liu, Matt J. Kusner, and Phil Blunsom. 2020. A survey on contextual embeddings. *Preprint*, arXiv:2003.07278.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68. Association for Computational Linguistics.

Yifan Liu, Gelila Tilahun, Xinxiang Gao, Qianfeng Wen, and Michael Gervers. 2025. A comparative study of static and contextual embeddings for analyzing semantic changes in medieval Latin charters. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 182–192, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Polina Lobanova, Pavel Bakhtin, and Yaroslav Sergienko. 2024. Identifying and visualizing trends in science, technology, and innovation using scibert. *IEEE Transactions on Engineering Management*, 71:11898–11906.

Na Luo and Ken Hyland. 2016. Chinese academics writing for publication: English teachers as text mediators. *Journal of Second Language Writing*, 33:43–55.

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

Gary Marcus. 2020. The next decade in AI: four steps towards robust artificial intelligence. *CoRR*, abs/2002.06177.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26.

Dmitry Motrichenko, Yaroslav Nedumov, and Kirill Skorniakov. 2021. Bag of tricks for citation intent classification via scibert. In *2021 Ivannikov Ispras Open Conference (ISPRAS)*, pages 120–126. IEEE.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Gabrijela Perković, Antun Drobnjak, and Ivica Botički. 2024. Hallucinations in llms: Understanding and addressing challenges. In *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, pages 2084–2088.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544. Association for Computational Linguistics.

Wai Mar Phyo, Marianne Nikolov, and Ágnes Hódi. 2023. Doctoral students' english academic writing experiences through metaphor analysis. *Heliyon*, 9(2):e13293.

Andrija Poleksic and Sanda Martincic-Ipsic. 2023. Effects of pretraining corpora on scientific relation extraction using bert and scibert. In *SEMANTICS Workshops*.

Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by reading: Contentful neural conversation with on-demand machine reading. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5427–5436. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Zahra Rahimi and Mohammad Mehdi Homayounpour. 2021. Tenssent: a tensor based sentimental word embedding method. *Applied Intelligence*, 51(8):6056–6071.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671. Association for Computational Linguistics.

Bhaskarjit Sarmah, Dhagash Mehta, Benika Hall, Rohan Rao, Sunil Patel, and Stefano Pasquali. 2024. Hybridrag: Integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 608–616.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Nimrod Shabtay, Felipe Maia Polo, Sivan Doveh, Wei Lin, Muhammad Jehanzeb Mirza, Leshem Choshen, Mikhail Yurochkin, Yuekai Sun, Assaf Arbelle, Leonid Karlinsky, and Raja Giryes. 2025. Livexiv - a multi-modal live benchmark based on arxiv papers content. In *The Thirteenth International Conference on Learning Representations*.

Rulin Shao, Jacqueline He, Akari Asai, Weijia Shi, Tim Dettmers, Sewon Min, Luke Zettlemoyer, and Pang Wei W Koh. 2024. Scaling retrieval-based language models with a trillion-token datastore. *Advances in Neural Information Processing Systems*, 37:91260–91299.

Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. 2023. kNN-diffusion: Image generation via large-scale retrieval. In *The Eleventh International Conference on Learning Representations*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803. Association for Computational Linguistics.

Rajeev Singh, Gaurav Gaonkar, Vedant Bandre, Nishant Sarang, and Sachin Deshpande. 2023. Scientific paper recommendation system. In *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, pages 1–4.

Rafael Alberto Rivera Soto, Kailin Koch, Aleem Khan, Barry Y Chen, Marcus Bishop, and Nicholas Andrews. 2025. Few-shot detection of machine-generated text using style representations. In *The Twelfth International Conference on Learning Representations*.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring llm prompting strategies for joint essay scoring and feedback generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 283–298. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021.

Christopher K Surratt. 2006. Creation of a graduate oral/written communication skills course. *American Journal of Pharmaceutical Education*, 70(1).

Ajša Terko, Emir Zunic, and Dzenana Donko. 2019. Neurips conference papers classification based on topic modeling. In *2019 XXVII International Conference on Information, Communication and Automation Technologies (ICAT)*, pages 1–5. IEEE.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. 2024. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *Preprint*, arXiv:2310.01469.

Shulin Yu and Lianjiang Jiang. 2022. Doctoral students' engagement with journal reviewers' feedback on academic writing. *Studies in Continuing Education*, 44(1):87–104.

Jiawei Zhang, Chejian Xu, and Bo Li. 2024. Chatscene: Knowledge-enabled safety-critical scenario generation for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15459–15469.

Yizhe Zhang, Siqi Sun, Xiang Gao, Yuwei Fang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2022. Retgen: A joint framework for retrieval and grounded text generation modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11739–11747.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713.

Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. 2024. Larger and more instructable language models become less reliable. *Nature*, 634(8032):61–68.

Wei Zhou and Jelke Bloem. 2021. Comparing contextual and static word embeddings with small data. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 253–259.

Appendix

A Prompt Engineering

Examples of the entire prompt, which was input into LLaMA-3.0-8B-instruct can be found in this section.

A.1 Combined Prompt with RAG

An example of the general prompt structure using retrieval reference documents is shown in this section.

The combined prompt with RAG

Task:
Please provide a rating for the following paragraph on a scale from 1 to 10. Your response must be a single number only.

INPUT TEXT TO RATE:
`[Content Placeholder]`

GOLD STANDARD DOCUMENTS FOR REFERENCE:

Document no.1

Document no.2

INSTRUCTIONS:
Please rate the **INPUT TEXT TO RATE** based on its quality and clarity on the scale of 1 to 10, using the **GOLD STANDARD DOCUMENTS FOR REFERENCE** as a basis. Do not rate the **GOLD STANDARD DOCUMENTS** themselves.

Now please, give the rating, for the **INPUT TEXT TO RATE**.

A.2 Combined Prompt without Using RAG

An example of the general prompt structure without using retrieval reference documents is shown in this section. This is also the prompt used where the revision analysis task was conducted without the use of reference documents.

The prompt without using RAG

Task:
Please provide a rating for the following paragraph on a scale from 1 to 10. Your response must be a single number only.

INPUT TEXT TO RATE:
`[Content Placeholder]`

INSTRUCTIONS:
Please rate the **INPUT TEXT TO RATE** based on its quality and clarity on the scale of 1 to 10.

Now please, give the rating, for the **INPUT TEXT TO RATE**.

A.3 Combined Prompt without Using RAG for Conference Writing Style Differentiation

The prompt without using RAG (conference writing style differentiation)

Task:
Please provide a rating for the following paragraph on a scale from 1 to 10. Your response must be a single number only.

INPUT TEXT TO RATE:
`[Content Placeholder]`

INSTRUCTIONS:
Please rate the **INPUT TEXT TO RATE** based on its writing style on the scale of 1 to 10.

Now please, give the rating, for the **INPUT TEXT TO RATE**.

A.4 Example Prompt Used for Revision Analysis

An example prompt used in the revision analysis experiment is provided in this section.

An example prompt used for revision analysis

Generated Text for Chunk 1 from [Paper Title Holder].
Task:
Please provide a rating for the following paragraph on a scale from 1 to 10. Your response must be a single number only.

INPUT TEXT TO RATE:
`[Content Placeholder]`

GOLD STANDARD DOCUMENTS FOR REFERENCE:

Document no.1

Document no.2

INSTRUCTIONS:
Please rate the **INPUT TEXT TO RATE** based on its quality and clarity on the scale of 1 to 10, using the **GOLD STANDARD DOCUMENTS FOR REFERENCE** as a basis. Do not rate the **GOLD STANDARD DOCUMENTS** themselves.

Now please, give the rating, for the **INPUT TEXT TO RATE**.

A.5 Example Prompt used for Conference Writing Style Distinction

This section provides an example prompt used in the conference stylistic distinction experiment.

An example prompt used for conference writing style distinction

Generated Text for Chunk 1 from [Paper Title Holder].
Task:
Please provide a rating for the following paragraph on a scale from 1 to 10. Your response must be a single number only.

INPUT TEXT TO RATE:
`[Content Placeholder]`

GOLD STANDARD DOCUMENTS FOR REFERENCE:

Document no.1

Document no.2

INSTRUCTIONS:
Please rate the **INPUT TEXT TO RATE** based on its WRITING STYLE on the scale of 1 to 10, using the **GOLD STANDARD DOCUMENTS FOR REFERENCE** as a basis. Do not rate the **GOLD STANDARD DOCUMENTS** themselves.

Now please, give the rating, for the **INPUT TEXT TO RATE**.

B Data Preprocessing

B.1 Vector Database Construction

The retrieval vector database is constructed using the text of NeurIPS proceedings 2023 sourced from Kaggle, chunking the text from each of the papers into fixed-length SciBERT embeddings (512 tokens in length) and indexed by FAISS (Facebook AI Similarity Search) to index and retrieve text embeddings efficiently.

B.2 Query Encoding

The input query (text to be rated) is encoded into a fixed-length vector using SciBERT.

B.3 Document Retrieval

The encoded query is compared against precomputed SciBERT embeddings in the vector database using cosine similarity. The top 2 most similar documents are retrieved as "gold standard" references.[8]

B.4 Combining the Input Text and Retrieved Documents

After retrieval, the system merges the cleaned input text (sanitize and preprocess text by removing unwanted elements such as LaTeX commands, email addresses, long alphanumeric strings, HTML tags, special characters, and excessive whitespace) with top reference documents as "gold standard" examples of high-quality writing. The LLM (LLaMA-3.0-8B-instruct) then evaluates the input text's quality and clarity based on these references. The final prompt combines the following elements:

1. The input text to be rated.
2. Retrieved "gold standard" documents for references.
3. Instructions asking the model to rate the input text on a scale of 1 to 10 based on its alignment with the "gold standard" reference documents.

Detailed structure of the prompt can be found in Appendix A.

B.5 Chunk-Based Evaluation with the RAG System for Revision Analysis and Conference Writing Style

To assess paper quality (or writing style), the content is divided into 200-token segments, each scored by the RAG system using retrieved reference documents. The process includes:

1. Segmentation: The paper is divided into 200-token chunks.
2. Scoring: Each chunk is input into the RAG system, which assigns a quality score.
3. Aggregation: The scores across all chunks are averaged to compute the overall score for the paper.

This chunking method was implemented to accommodate the limited input window size of LLaMA-3.0-8b-insturct while ensuring a more precise and refined scoring process by the RAG system.

To balance efficiency and relevance, reference documents were truncated to 200 tokens, ensuring

[8] The number 2 is determined based on a balance between the need for meaningful reference and computational resource constraints.

sufficient context without unnecessary length, as the primary objective of the reference documents was to establish a gold standard for defining what is considered "good" or "suitable" during evaluation, rather than to serve as comprehensive scientific texts for in-depth analysis.

For output generation, a 1,000-token limit was set to balance computational resource constraints while providing sufficient justification and suggestions, with a focus on delivering clear and reliable numerical ratings.

C Ablation Studies

C.1 Consistency Check of the RAG System

A consistency check was conducted to evaluate the reliability of the RAG system in delivering consistent scores for the same scientific text. The primary objective was to determine whether the RAG system could produce stable and reproducible evaluations across multiple assessments of the same input. This check specifically aimed to ensure that the system's outputs are free from randomness, thereby confirming the reliability of its scoring mechanism. The experimental setup for this consistency check follows the same methodology described in Appendix C.2, with the primary distinction that each input was processed through the RAG system five times to obtain multiple ratings. 100 random text samples from NeurIPS 2023 were selected as input. Each input text was processed through the RAG system five times to generate multiple ratings, making the approach computationally intensive. Consequently, the number of text samples was carefully selected to strike a balance between resource constraints and the need for representative results. As consistency checks do not require large volumes of data for effective evaluation, a limited yet sufficient sample size was deemed appropriate. To quantify the consistency of the RAG's scoring behavior, the percentage of texts for which the RAG system assigned identical scores across all 5 trials was calculated. Specifically, if the RAG system produces the same score for a given text in all five iterations, it is considered a "consistent" evaluation.

The experiment result Figure 5 shows that the RAG system consistently evaluates scientific texts. 91.5% of texts received identical scores across all five trials, indicating high reliability. 1.1% and 4.3% showed moderate consistency (above 75% and 60%, respectively), while only 3.2% had identical scores in 60% or fewer evaluations. Overall,

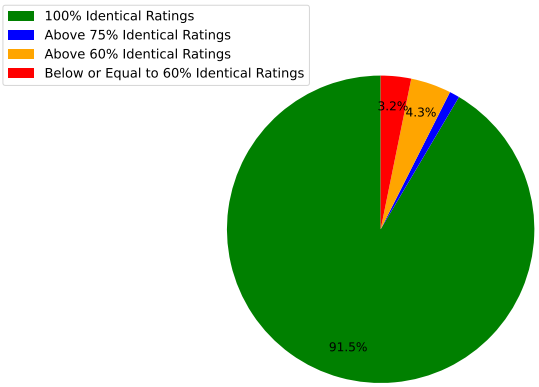


Figure 5: The portion of text chunks getting identical scores when feeding into the LLM 5 times

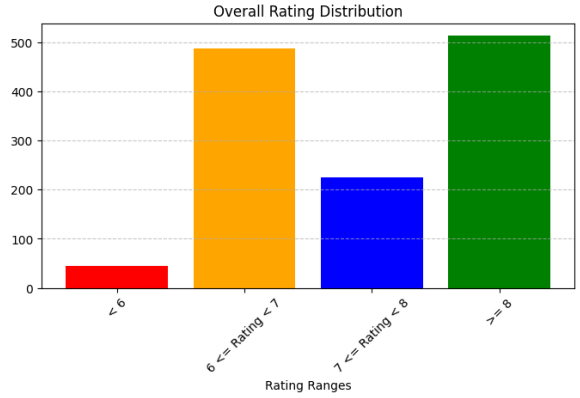
these results demonstrate that the RAG system is highly consistent in its evaluation of scientific writing, with a very small percentage of cases showing minimal variation in scoring. This result is highly important, as it shows that the scores are not randomly assigned to texts by the system.

C.2 A Baseline Check for the RAG System

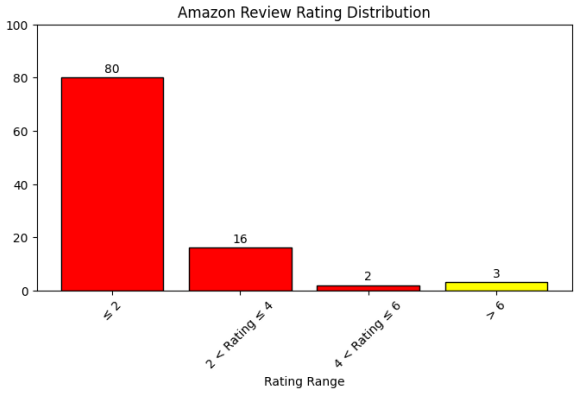
This experiment assesses the RAG system's ability to distinguish the semantic differences between scientific and non-scientific writing by evaluating whether it assigns higher scores to scientifically rigorous texts and lower scores to colloquial ones. Differentiating conference writing styles or revision quality assumes that the model can first distinguish scientific vs. non-scientific writing. Using the Amazon dataset as a colloquial contrast to NeurIPS ensures the system recognizes core differences of what constitutes "scientific" before tackling finer distinctions. The prompt used in this baseline experiment can be found in Appendix A.1, prompting the LLM to rate the input text (either scientific text or Amazon review) based on quality and clarity.

This baseline experiment is essential, as a system that cannot reliably identify the core elements that constitute scientific writing, such as quality, clarity, or other key factors, cannot be expected to discern more nuanced dimensions, including revision-based improvements or conference-specific stylistic conventions. By grounding ratings in authoritative references, this experiment ensures the model follows retrieved sources rather than arbitrary biases before applying it to specific cases like NeurIPS vs. ICLR papers.

SciBERT embeddings were precomputed for key



(a) Distribution of RAG-generated scores for NeurIPS texts. Most received moderate to high scores (above 6), indicating the system's ability to identify well-written content.



(b) Scores assigned to Amazon reviews. The majority received low ratings, showing the model's ability to distinguish informal writing.

Figure 6: Results of the RAG system's baseline check on scientific (left) and informal texts (right). X-axis: score (1–10), Y-axis: number of texts.

sections of the NeurIPS 2023 dataset to enable efficient text retrieval. A stratified 20% sample from the NeurIPS 2023 text dataset was used for retrieval, while a separate 20% served as input for LLM evaluation. Model-generated outputs were parsed for ratings, which were stored alongside the input text and retrieved documents for analysis.

The distribution of ratings generated by the RAG system for the sampled scientific texts is shown in Figure 6a. The ratings are provided on a scale from 1 to 10; the rating distribution indicates that the RAG system consistently assigns high scores to the text from NeurIPS2023 accepted papers. All texts predominantly maintain scores above 6, suggesting a robust and reliable scoring mechanism.

To further assess differentiation capabilities, the RAG system was tested on 100 randomly selected Amazon reviews. The results (Figure 6b) show that 80% of reviews received a rating of 2, with very few exceeding 5 and none above 7. These results suggest that the system effectively identifies and distinguishes differences, such as clarity and quality, between high-quality scientific writing and informal content, demonstrating its sensitivity to established scholarly standards.

C.3 Conference Style Distinction using ICLR as Retrieval Vector Database

An ablation study using ICLR proceedings as the sole vector database for conference writing style differentiation is presented in this section, in Figure 7. The system successfully distinguished between the writing styles of various conferences; as expected, ICLR papers received the highest scores,

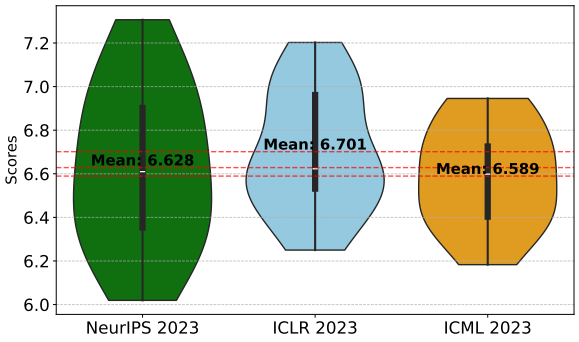


Figure 7: Score distributions for 15 randomly selected papers from NeurIPS, ICLR, and ICML using ICLR conference proceedings as vector database (reference document).

demonstrating the RAG system's capacity to capture semantic differences in writing styles across manuscripts from distinct publication venues.

It is worth noting that the differences in mean scores are less pronounced compared to those reported in Section 4.2. This attenuation may be attributed to the reduced size of the vector database, which in this case consists exclusively of ICLR papers, which have significantly smaller submission volumes than NeurIPS (up to and including the year 2024, at the commencement of this study).[9] Smaller retrieval corpora can limit the system's capacity, thereby affecting overall performance. Prior work supports this by showing that using a larger datastore (retrieval database) during inference im-

[9] Detailed submission statistics available at https://media.neurips.cc/Conferences/NeurIPS2023/NeurIPS2023-Fact_Sheet.pdf and https://papercopilot.com/

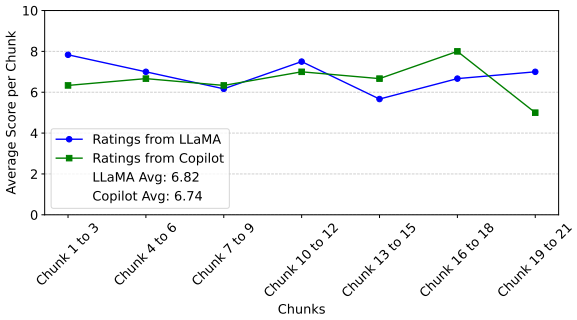


Figure 8: The results of the experiment conducted comparing the scores from LLaMA-3.0-8b-instruct to those scores from Copilot.

proves model performance (Shao et al., 2024).

C.4 Experiment Conducted Comparing the Scores from LLaMA to the Scores from Copilot

This experiment evaluates the same one NeurIPS proceeding using an identical RAG mechanism with two distinct LLMs: LLaMA-3.0-8b-instruct and Copilot. The authors of this paper retain full copyright of the NeurIPS proceeding being evaluated, and a university-owned instance of Copilot was utilized to ensure compliance with legal and data privacy standards. Readers should note that this Copilot instance is not locally deployable, which limits its feasibility for large-scale paper evaluation experiments. This constraint arises from the need to manually paste text chunks and reference documents into the chat instance one by one, making the process impractical for extensive evaluations such as conference style distinction and revision quality analysis. To ensure the reliability of the evaluation, the NeurIPS paper selected for assessment in this experiment is from a different year than those in the retrieval database. The objective is to verify score consistency across models, confirming that using a more advanced LLM does not significantly alter evaluation outcomes.

Figure 8 visualizes the average chunk scores, grouped in chunk triplets per data point. While absolute score values sometimes differ (with LLaMA sometimes showing higher scores and sometimes Copilot), models exhibit a similar evaluation trend and overall average scores that closely match, indicating consistency in content assessment and demonstrating that utilizing a larger and more advanced LLM remains a reliable approach for revision analysis and conference writing style distinction.

As seen in the results, despite variations in absolute scores, both models exhibit a consistent scoring trend. The average rating assigned by Copilot for this manuscript is 6.74, whereas the rating from LLaMA is 6.82. Copilot's scores are slightly lower than those from LLaMA, which can be explained by differences in model calibration, training distribution, and risk preferences. Language models are often calibrated to avoid extreme outputs, ensuring balanced scoring unless strong justification exists. This conservative behavior helps maintain consistency, especially when trained on diverse-quality texts (Jiang et al., 2021). Additionally, the distribution of ratings in Copilot and LLaMA-3.0-8b-instruct's training data likely influences its scoring behavior. If extremely high ratings were less common in training, the model might be less inclined to assign them, an effect reinforced by fine-tuning techniques like reinforcement learning from human feedback (RLHF) (Stiennon et al., 2020). Furthermore, models trained with reward mechanisms often develop risk-averse tendencies, favoring mid-range scores to avoid penalization (Ouyang et al., 2022). These factors explain why, despite following a similar trend, different LLMs can produce varying score distributions due to underlying differences in pretraining and optimization. However, readers should note that the objective of this experiment is not to achieve an exact match in the absolute value of ratings across different LLMs but rather to ensure that the overall scoring patterns are consistent, with minimal variation in overall scoring trends and average scores.

D Real-life Example of Hallucination

Figure 9 illustrates a real-world hallucination case with ChatGPT-4o. When lacking web access, the model generated incorrect author names, but with browsing enabled, it retrieved the correct ones. This underscores the importance of external knowledge sources in scientific writing. Given the impracticality of embedding a complete web-scale knowledge base within a large language model (LLM) (Li et al., 2022), these findings also indicate the importance of retrieval-augmented methods, such as utilizing vector databases. Similar to how ChatGPT-4o exhibited hallucinations in the absence of external knowledge search, smaller models like LLaMA-3.0-8B-Instruct are likely to face challenges in accurately evaluating the quality of scientific texts and writing style without access to

retrieval-enhanced information.

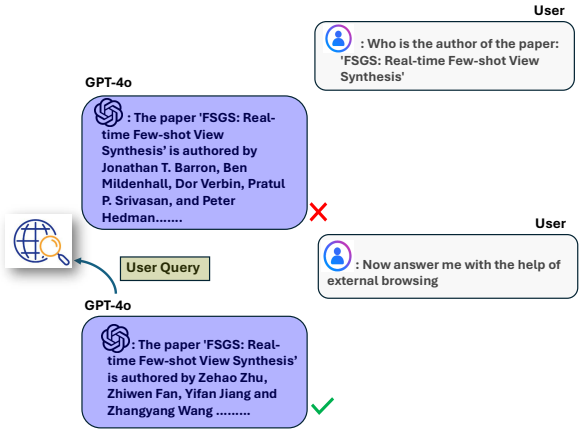


Figure 9: A real-life example of hallucination during manuscript creation when using ChatGPT-4o; after enabling external search, GPT retrieved the correct author name for the paper.