Project Proposal: Enhancing Scientific Literature Analysis with a RAG System

Jin Huang, Xuteng Luo, Omkar Vodela, Aarya Kulshrestha

Keywords

Scientific Literature Analysis, LLMs

ACM Reference Format:

Jin Huang, Xuteng Luo, Omkar Vodela, Aarya Kulshrestha. 2018. Project Proposal: Enhancing Scientific Literature Analysis with a RAG System. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Understanding scientific literature is critical for advancing research across multiple domains [1, 24]. Large Language Models (LLMs) have improved scientific literature analysis capabilities through natural language processing. Recent work, such as the SciLitLLM pipeline, has demonstrated improvements in LLM performance by combining continuous pre-training (CPT) to inject scientific knowledge with supervised fine-tuning (SFT) to enhance instruction-following abilities. [3, 15, 19]. Despite this, SciLitLLM still requires a large amount of training data, making it expensive and resource intensive to update and limited to the information present in its static training data, making it less adaptable to evolving external knowledge.

One key limitation is illustrated in Figure 1: accurately extracting entities from a biomedicine paper sometimes requires recognizing new terminologies, concepts, clinincal findings and more. While the SciLitLLM pipeline effectively injects domain-specific knowledge and enhances instruction following, some questions require external or "eternal" knowledge not captured in the static training data. To address this limitation, our proposal aims to integrate a Retrieval-Augmented Generation (RAG) module that dynamically supplies up-to-date external context during inference.

1.1 Problem Statement

Current approaches to scientific literature analysis, such as the recent work in SciLitLLM [10] demonstrated a strong ability to analyze scientific literature. Despite its impressive performance on tasks such as entity extraction from biomedicine papers (see Figure 1), this model fails to access and process real-time scientific knowledge without retraining on new datasets. We want to investigate if a RAG system can improve the adaptability and accuracy of scientific literature analysis compared to models trained solely on the domain-specific corpora.

1.2 Motivation

Scientific literature is continuously evolving, and researchers need models that can not only learn from a fixed corpus but also adapt to new knowledge. Traditional pre-trained models are inherently stale, as they cannot adapt to new scientific discovery without costly retraining. This limits researchers and their ability to use AI in making scientific analysis, especially when dealing with questions that require "eternal" or up-to-date information. Enhancing these models with a RAG system is crucial for improving their robustness and applicability in dynamic scientific environments. The impact of this work can extend beyond scientific literature analysis. Any constantly evolving domains such as policy analysis, medical diagnoses, and others can benefit from this system.

1.3 Overview of Proposed Work

Our proposed approach aims to extend the SciLitLLM pipeline by incorporating a RAG module. This module will dynamically retrieve external documents during inference and provide additional contexts. By doing so, we expect to overcome the limitations of static training data and achieve a more flexible and accurate scientific literature analysis system.

2 Related Work

Prior work in adapting LLMs to scientific tasks shows that pre-training on domain-specific corpora, as demonstrated by [2, 17, 22], effectively tailors these models for specialized language tasks. However, while these methods provide strong domain understanding, they are often limited by their static training data and struggle to incorporate evolving scientific information. Several studies highlight that models trained solely on fixed corpora fail to integrate real-time updates or external context, a critical weakness in dynamic fields.

The SciLitLLM method improves upon this by combining continual pre-training and supervised fine-tuning to inject high-quality scientific knowledge and enhance task-specific instruction following [8, 9, 12, 16, 20]. Despite these enhancements, challenges remain in addressing questions that require external or up-to-date information [11, 21].

Recent advances in retrieval-augmented generation provide a promising solution. By dynamically retrieving external documents, methods described in [6, 7, 23] significantly improve performance on tasks that require up-to-date information. These techniques are relevant because they directly tackle the inherent limitations of static pre-training, yet their integration with domain-specific fine-tuning remains to be fully explored.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

3 Proposed Work

Our proposal extends the SciLitLLM approach by incorporating a RAG module, primarily focusing on inference. Before RAG, we will pre-train the SciLitLLM using the CPT and SFT methods. The RAG module be added to the pipeline during inference time, where it will dynamically retrieve relevant information from an external knowledge source and augmenting the model's input with up-to-date context.

The first aspect of our RAG module would be constructing a knowledge source. We plan to find a comprehensive and up-to-date collection of scientific documents, research papers, databases, and other relevant resources. This could include PubMed, arXiv, specialized databases, and even curated web resources. Given our capabilities, we can also try seeing if web crawling of scientific domains would be more effective than searching ourselves. We will then convert the documents into smaller segments and use a vector embedding model to translate the text into vector representations. We will also construct a vector database (e.g., Pinecone, Weaviate, Milvus) to store and efficiently search the document embeddings.

When a query is received, it will be encoded using a sentence transformer model (e.g. SBERT) fine-tuned for scientific text. This encoded query vector will then be used to search our vector database. The top-k most relevant document chunks will be retrieved and formatted into a prompt, which is combined with the original query and fed into the pre-trained SciLitLLM model CPT and SFT.

To enhance robustness and address potential challenges, we plan to explore alternative approaches alongside SciLitLLM. While RAG retrieves information based on semantic similarity and excels in context-aware responses, it is prone to hallucinations and may return conceptually related but imprecise results. To mitigate these issues, we will also evaluate lexical retrieval methods such as BM25, a lightweight algorithm that ranks documents using term frequency and inverse document frequency (TF-IDF). BM25 could potentially provide different performance as opposed to semantics it can accurate retrieve documents based on explicit matches.

3.1 Evaluation and Benchmarking

We will conduct a comprehensive evaluation of our enhanced system using SciRIFF benchmarks [19]. Our evaluation strategy will focus on both quantitative performance and qualitative analysis to ensure a robust assessment.

- **Quantitative Benchmarking:** We will compare the RAG-enhanced model against the baseline SciLitLLM system across various metrics, including precision, recall, F1-score, and accuracy, focusing on tasks like entity extraction and context-aware question answering. These metrics are chosen because they provide a balanced view of the system's ability to accurately and comprehensively analyze scientific texts. We will also assess computational efficiency, including inference time and memory usage, to ensure the model's practicality in real-world scenarios.
- **Ablation Studies:** To understand the contribution of each component, we will perform detailed ablation studies. This involves systematically removing or modifying retrieval modules (e.g., disabling RAG or replacing it with BM25) and evaluating the impact on overall performance. Ablation studies are crucial for isolating the effects of individual components, helping to identify the most effective retrieval strategies and guiding future improvements.
- **Robustness Testing:** We will test the system's ability to handle noisy or incomplete data and its adaptability to diverse scientific domains beyond biomedicine. This ensures the system remains reliable and effective in real-world applications where data quality may vary.
- **User-Centric Evaluation:** A qualitative analysis involving domain experts will be conducted to assess the relevance and usefulness of the generated outputs. Expert feedback is invaluable for understanding how well the system meets user needs, focusing on accuracy, clarity, and the ability to provide contextually rich information.

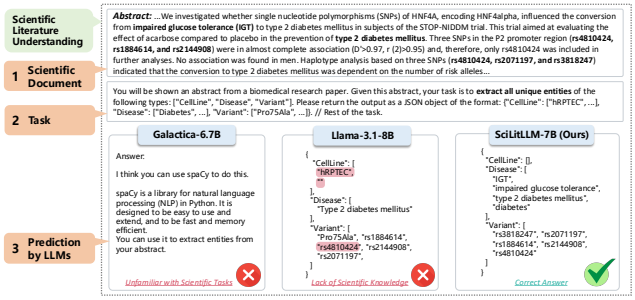

Figure 1: An example of scientific literature understanding in SciRIFF. It involves extracting accurate entities from a biomedicine paper. SciLitLLM-7B demonstrates sufficient scientific knowledge and instruction-following ability to accurately identify and extract these entities.

This multifaceted evaluation approach is designed to provide a thorough understanding of the system's strengths and limitations, ensuring that the enhancements deliver meaningful improvements in scientific literature analysis.

4 Conclusion

In summary, we propose to extend the SciLitLLM pipeline—based on continual pre-training and supervised fine-tuning—with a Retrieval-Augmented Generation (RAG) module. By dynamically retrieving external context, our approach aims to overcome the limitations of static training data and better handle questions requiring external knowledge [1, 24]. Building on established methods [3–5, 8, 13, 14, 16, 18, 19], we expect moderate improvements in the accuracy and relevance of scientific literature understanding.

References

[1] Microsoft Research AI4Science and Microsoft Azure Quantum. 2023. The Impact of Large Language Models on Scientific Discovery: a Preliminary Study using GPT-4. *CoRR* abs/2311.07361 (2023).
[2] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *EMNLP/IJCNLP (1)*. Association for Computational Linguistics, 3613–3618.
[3] Hengxing Cai, Xiaochen Cai, Junhan Chang, Sihang Li, Lin Yao, Changxin Wang, Zhifeng Gao, Hongshuai Wang, Yongge Li, Mujie Lin, Shuwen Yang, Jiankun Wang, Yuqi Yin, Yaqi Li, Linfeng Zhang, and Guolin Ke. 2024. SciAssess:

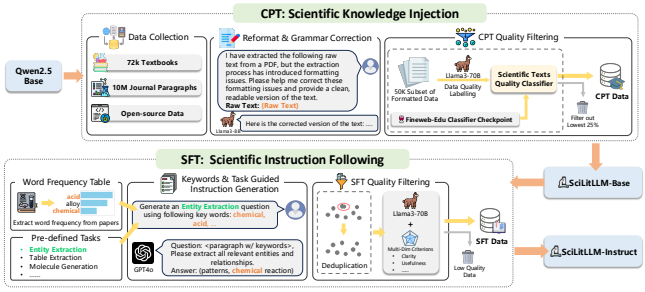

Figure 2: The pipeline of SciLitLLM consists of two key stages: continual pre-training (CPT) for scientific knowledge injection and supervised fine-tuning (SFT) for scientific instruction following.

Benchmarking LLM Proficiency in Scientific Literature Analysis. *CoRR* (2024). https://doi.org/10.48550/arXiv.2403.01976

[4] Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. 2024. Instruction Pre-Training: Language Models are Supervised Multitask Learners. *arXiv preprint arXiv:2406.14491* (2024).

[5] Daixuan Cheng, Shaohan Huang, and Furu Wei. 2023. Adapting Large Language Models via Reading Comprehension. *CoRR* abs/2309.09530 (2023).

[6] David 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), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 15607–15631. https://doi.org/10.18653/v1/2023.acl-long.870

[7] Ronen Eldan and Yuanzhi Li. 2023. TinyStories: How Small Can Language Models Be and Still Speak Coherent English? *CoRR* abs/2305.07759 (2023). https://doi.org/10.48550/arXiv.2305.07759

[8] Xisen Jin, Dejiao Zhang, Henghui Zhu, Wei Xiao, Shang-Wen Li, Xiaokai Wei, Andrew O. Arnold, and Xiang Ren. 2022. Lifelong Pretraining: Continually Adapting Language Models to Emerging Corpora. In *NAACL-HLT*. Association for Computational Linguistics, 4764–4780.

[9] Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2023. Continual Pre-training of Language Models. In *ICLR*. OpenReview.net.

[10] Sihang Li, Jin Huang, Jiaxi Zhuang, Yaorui Shi, Xiaochen Cai, Mingjun Xu, Xiang Wang, Linfeng Zhang, Guolin Ke, and Hengxing Cai. 2024. Scilitllm: How to adapt llms for scientific literature understanding. *arXiv preprint arXiv:2408.15545* (2024).

[11] Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Carl Yang, and Liang Zhao. 2023. Beyond One-Model-Fits-All: A Survey of Domain Specialization for Large Language Models. *CoRR* abs/2305.18703 (2023). doi:10.48550/ARXIV.2305.18703 arXiv:2305.18703

[12] Sanket Vaibhav Mehta, Darshan Patil, Sarath Chandar, and Emma Strubell. 2023. An Empirical Investigation of the Role of Pre-training in Lifelong Learning. *J. Mach. Learn. Res.* 24 (2023), 214:1–214:50.

[13] OpenAI. 2023. GPT-4 Technical Report. *CoRR* abs/2303.08774 (2023). https://doi.org/10.48550/arXiv.2303.08774

[14] Team Qwen. 2024. Qwen2.5: A Party of Foundation Models. https://qwenlm.github.io/blog/qwen2.5/

[15] Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. SciRepEval: A Multi-Format Benchmark for Scientific Document Representations. In *EMNLP*. 5548–5566. https://doi.org/10.18653/v1/2023.emnlp-main.338

[16] Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding. In *AAAI*. AAAI Press, 8968–8975.

[17] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A Large Language Model for Science. *CoRR* abs/2211.09085 (2022). https://doi.org/10.48550/arXiv.2211.09085

[18] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR* abs/2302.13971 (2023).

[19] David Wadden, Kejian Shi, Jacob Morrison, Aakanksha Naik, Shruti Singh, Nitzan Barzilay, Kyle Lo, Tom Hope, Luca Soldaini, Shannon Zejiang Shen, Doug Downey, Hannaneh Hajishirzi, and Arman Cohan. 2024. SciRIFF: A Resource to Enhance Language Model Instruction-Following over Scientific Literature. *CoRR* abs/2406.07835 (2024). https://doi.org/10.48550/arXiv.2406.07835

[20] Tongtong Wu, Massimo Caccia, Zhuang Li, Yuan-Fang Li, Guilin Qi, and Gholamreza Haffari. 2022. Pretrained Language Model in Continual Learning: A Comparative Study. In *ICLR*. OpenReview.net.

[21] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. FinGPT: Open-Source Financial Large Language Models. *CoRR* abs/2306.06031 (2023). https://doi.org/10.48550/arXiv.2306.06031

[22] Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2022. A Deep-learning System Bridging Molecule Structure and Biomedical Text with Comprehension Comparable to Human Professionals. *Nature communications* 13, 862 (2022).

[23] Xingjian Zhang, Yutong Xie, Jin Huang, Jinge Ma, Zhaoying Pan, Qijia Liu, Ziyang Xiong, Tolga Ergen, Dongsub Shim, Honglak Lee, and Qiaozhu Mei. 2024. MASSW: A New Dataset and Benchmark Tasks for AI-Assisted Scientific Workflows. *CoRR* abs/2406.06357 (2024). https://doi.org/10.48550/arXiv.2406.06357

[24] Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh T. N. Nguyen, Lauren T. May, Geoffrey I. Webb, and Shirui Pan. 2023. Large Language Models for Scientific Synthesis, Inference and Explanation. *CoRR* abs/2310.07984 (2023).

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009