SLM-as-a-Judge with Attention Steering for Detailed Topic Extraction from Academic Literature

Takahiro Kawamura[1], Junichiro Mori[1]

[1] Information Technology Center, The University of Tokyo

Abstract

This study introduces domain-specific Small Language Models (SLMs) designed to fact-check the outputs of general-purpose Large Language Models (LLMs). The goal is to accurately and automatically extract detailed technical elements from academic papers to build knowledge graphs. Two SLM types were developed: one through continued pre-training and the other using attention steering. Experiments on information extraction in ‘Fake News Detection’ showed that SLMs improved the sufficiency, accuracy, and stability of extracted information. Future work will involve larger datasets and further knowledge graph construction.

Keywords

Knowledge Graph, LLM, Information Extraction

1. Introduction

A science graph, functioning as a knowledge graph for scientific and technological information, is currently under construction to facilitate the analysis of research trends and enable forecasting [1] through the utilization of metrics such as citation counts and altmetrics. In this process, large language models (LLMs) are employed to extract technical elements corresponding to objectives, methods, and subjects from the full texts of scientific papers, as full-text documents provide significantly more detailed information than abstracts.

However, preliminary tests with general-purpose LLMs (e.g., ChatGPT) showed problems like omitting or hallucinating technical terms, with extraction accuracy varying due to prompt sensitivity. Moreover, LLM outputs may not prioritize the importance of information. To solve these issues, we are developing domain-specific Small Language Models (SLMs) as-a-judge to fact-check and enhance output accuracy, focusing on extracting detailed technical components (e.g., specific model types, datasets, and metrics). For the overall framework, please refer to figure 1. This study proposes two SLM development methods - continuous pre-training and attention steering - and presents initial verification experiments using a small, high-quality dataset in Fake News Detection, a subfield of computer science. The paper also outlines related work, experimental approaches, results, and future directions.

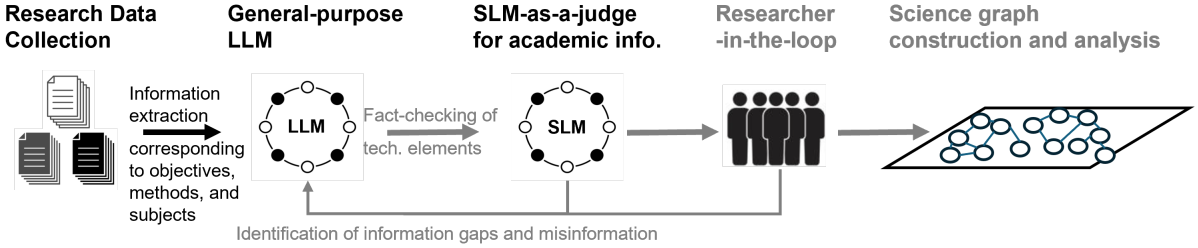
Research Data Collection
General-purpose LLM
SLM-as-a-judge for academic info.
Researcher -in-the-loop
Science graph construction and analysis
Information extraction
corresponding to objectives, methods, and subjects
LLM
Fact-checking of tech. elements
SLM
Identification of information gaps and misinformation

Figure 1: The envisioned overall framework (this paper focused only on SLM development and its comparison with LLM)

ISWC 2025 Companion Volume, November 2–6, 2025, Nara, Japan
Corresponding author.

2. Related Work

Science Knowledge Graph. Recently, there has been growing interest in building knowledge graphs to structure scholarly information. Most existing knowledge graphs focus on bibliographic data such as publications, authors, affiliations, research topics and their relationships (citations, authorship, collaborations, topic similarity). Notable initiatives include OpenAlex[1], as well as projects like Semantic Scholar[2], AMiner[3], and OpenCitations[4]. These graphs typically define entities such as papers, authors, institutions, research fields, and keywords as nodes, with relationships like authorship and citations as edges. Some newer approaches [2] use systems like Babelfy to identify topics and link them to BabelNet entities, constructing graphs from co-occurrence, while others [3] utilize AI-in-the-loop methods (e.g., DeepShovel) to build topic trees and analyze the influence among papers to support idea flow analysis and research forecasting. Recent trends [4, 5, 6, 7] also include the extraction of semantic relationships between technical elements (e.g., research objectives, methods, evaluation metrics) and representing them in graph-based formats allowing for a richer understanding of research content compared to simply listing topics. Techniques such as manual annotation and natural language processing tools are used to extract objectives, methods, results, and their interrelations.

However, these existing approaches face limitations: many employ NLP techniques that are not necessarily up to date, cover only a small number of papers, are infrequently updated, or are not publicly available.

LLM for academic information. The development of LLMs tailored for scholarly information is active [2], with examples such as Meta's Galactica, domain-specific models like BioMedLM, and language-specific models (e.g., Chinese and English) such as ChatGLM3, all of which have been trained on large scientific datasets. A variety of services, including SciSpace, Elicit, and Consensus, use LLMs to summarize papers, extract key points, and facilitate literature searches, though not all support knowledge graph construction or rely directly on LLMs. However, these LLMs and services were not used in the present research due to differences in purpose, unresolved general LLM issues, and a lack of modifiability.

Instead, this research applies an underexplored method: attention steering, which involves focusing or suppressing attention on specific words during LLM inference to guide output. While similar techniques such as PASTA [8] have shown theoretical and basic task improvements, real-world applications remain rare. Recent studies [9, 10] have explored manipulating attention and activations within LLMs for better content control and long-form processing. This study uniquely applies and evaluates attention steering in developing a specialized SLM for academic information extraction.

3. Proposed Approach

This project aims to build a detailed science graph where research objectives, methods, and subjects are nodes. As an example, extraction should capture the main theme, data (e.g., Slovak language), techniques (deep learning), specific methods (like CNN or LSTM), and even implementation details (such as ReLU or TensorFlow). So, we developed a model with continual pre-training on technical domain data and another using attention steering to focus on technical terms during inference. These outputs were compared to those from a general LLM to check for coverage and accuracy. For attention steering, attention weights in the transformer model were manually biased toward target tokens, intentionally guiding output content. Implemented in frameworks like PyTorch, this involved adjusting attention scores via log-ratio biases during

[1] https://openalex.org/

[2] https://www.semanticscholar.org/

[3] https://www.aminer.cn/

[4] https://search.opencitations.net/

the forward pass as below. Although these methods slow inference, real-time responses are not required for this study.

$$
\begin{aligned}
&= +\log(\text{ratio})\ \mathbf{1}_{=} \\
&= \frac{\exp()}{\exp()}
\end{aligned}
$$

: the position of the specific token to which the bias is to be applied. $\mathbf{1}_{=}$: an indicator function that returns 1 if equals , and 0 otherwise.

3.1. Experimental Setting

For continuous pre-training, we collected 48 survey papers on Fake News Detection (FND), totaling 4.32 million characters, all of which were used for training. The base model was llama-3-8B-Instruct, trained with a learning rate of 5e-6, batch size 2048 tokens, and 1.5 epochs using 8 batches 4 GPUs. Instead of heavy pre-training, we also steered attention toward specific tokens identified by spaCy's NER tool (such as PERSON, LOC, ORG, GPE, PRODUCT, NORP, WORK_OF_ART, LANGUAGE, DATE, TIME, PERCENT, QUANTITY, ORDINAL, and CARDINAL) by adding a bias to their attention scores across all layers. Full-parameter training was used rather than LoRA, since we aimed to apply a moderate bias across all layers. Technical element extraction experiments were done in two ways:

1. Prompting with a paper title and having the model respond based on its internal knowledge about the paper's topics. It has been verified that the paper was included in the training data of the model.
2. Prompting with a 1,000–2,000 character excerpt, selected to contain around 10 topics, and asking the model to extract the topics appearing in the given text.

Four models were compared: continuous pre-training (CP), CP plus steered attention (SA), original llama-3-8B-Instruct, and GPT4.1. Evaluation for (1) knowledge-based extraction used precision, recall, and F1-score, while (2) prompt-based extraction used Normalized Discounted Cumulative Gain (NDCG) to check the order of the extracted topics. The ground-truth data was manually created by project members specializing in computer science, who visually extracted information from 10 papers on FND. The average number of topics to be extracted per paper was 41.3. Temperature was set to 1.0.

4. Experimental Results

4.1. Extraction from Internal Knowledge

In Table 1, both CP and SA achieved higher F1 scores compared to GPT and the base llama model, primarily because they extracted a greater number of technical elements, thus increasing recall. Notably, the SA model also attained the highest precision, demonstrating that the intended effect of attention steering was realized. Although GPT had high precision, its recall was low likely due to more conservative outputs, resulting in the lowest F1 score. The number of elements to be extracted was not specified, since the appropriate number of technical elements differs for each paper and is unknown in practical use. While CP had the highest F1 score overall, its lower precision and larger standard deviation suggested instability. In direct comparison, CP and SA each outperformed the other in five out of ten cases, but SA's marginally lower F1 score was offset by its higher precision and greater stability, making it the more desirable model. However, both CP and SA often generated structurally awkward sentences and substantially underperformed GPT in language fluency. As a result, it is recommended that SA be used in tandem with GPT, relying on GPT for language generation tasks.

Table 1
Extraction results

LLM/SLMs	Metrics	Ave.	SD	LLM/SLMs	Metrics	Ave.	SD
GPT-4.1	P	0.88	0.15	llama+CP	P	0.76	**0.22**
	R	0.11	0.05		R	**0.23**	**0.14**
	F1	0.19	0.08		F1	**0.33**	**0.16**
	NDGC	0.88	0.15		NDGC	0.86	0.14
llama-3-8B-Instruct	P	0.83	0.20	llama+CP+SA	P	**0.92**	0.09
	R	0.14	0.11		R	0.20	0.11
	F1	0.23	0.13		F1	0.32	0.14
	NDGC	0.86	**0.19**		NDGC	**0.96**	0.07

4.2. Extraction from Prompts

Since most LLMs cannot process the full text of a paper in a single prompt (with the exception of GPT-4.1), the experiment extracted technical elements using only 1,000–2,000 character excerpts as prompts. This approach matches the current 'Deep Research' trend, where papers are segmented and summarized before being fed to LLMs, and also reflects practices used in Retrieval-Augmented Generation (RAG) systems. Extraction accuracy for technical elements from such excerpt lengths was nearly 100% across all models, leading to minimal performance difference among them. Consequently, NDCG@5 was adopted to evaluate the ranking order of the extracted elements, with relevance scores manually assigned to the ground truths. As shown in Table 1, the SA model produced the best results, with low variance across runs. In contrast, the CP model sometimes hallucinated by outputting information not present in the excerpt. For the SA model, the attention bias was set to a very small value ($\log(1 + 1\ \ 16)$), unlike the larger bias ($\log(1.5)$) used in prior experiments, because a larger bias degrades linguistic quality when input tokens are short.

5. Discussion and Future Work

In this paper, the authors proposed a fact-checking Small Language Model (SLM) designed specifically for extracting domain-specific technical elements from academic literature, with the ultimate aim of constructing a detailed science graph. Consequently, GPT rarely makes outright errors in responses to single queries, but its answers are often overly simple and lack sufficient detail. Repeating questions can draw out more information, but this eventually leads to hallucinations (speculative or inaccurate content) which makes it hard to distinguish between factual and fabricated information. Continuous pre-training enables models to produce more detailed responses, particularly regarding numerical data (historically difficult for LLMs), yet it also introduces greater variance between experimental runs and sometimes produces seriously flawed outputs. Such models also tend to have reduced comprehension of prompts and decreased fluency in generated language. This is consistent with the well-known phenomenon that acquiring highly specialized knowledge can diminish a model's general language abilities and performance on unrelated tasks. Attention steering, when used appropriately, may enhance the detail and stability of the information generated. However, it is important to note that these findings are based on experiments conducted with a relatively small dataset, and the subjectivity of the ground-truth data remains a concern.

For future work, the authors' top priority is to expand their dataset, given that no gold-standard dataset fully satisfies their requirements. They plan to validate their approach on a larger scale using general-purpose technical term extraction datasets. It is also essential to define the scope and size of the target domain — for instance, while current experiments focused only on FND within Computer Science, expanding the domain is expected to make the model's behavior more similar to general-purpose LLMs. This makes it important to carefully define the domain in line with the intended application. From an algorithmic perspective, unresolved challenges include developing techniques for introducing bias into the attention mechanism, quantitatively assessing how varying the degree of bias affects output, and evaluating potential thresholds as well as the risk of model collapse from excessive bias. The authors plan to use the extracted information to build science graphs and proceed with assessments in domains where analysis is demonstrably needed. Toward the goal of science and technology analysis, we are working with other major Japanese research institutions on science graph sharing and joint analysis, and also plan to share their methodologies and data internationally. We hope to promote the development of new, internationally recognized science and technology assessment methods by collaborating with reputable overseas institutions.

6. Declaration on Generative AI

During the preparation of this work, the authors used **ChatGPT** in order to: **Text Translation**. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] X. Gu, M. Krenn, Forecasting high-impact research topics via machine learning on evolving knowledge graphs, Machine Learning: Science and Technology 6 (2025). doi:10.1088/2632-2153/add6ef.

[2] M. D. L. Tosi, J. C. dos Reis, Scikgraph: A knowledge graph approach to structure a scientific field, Journal of Informetrics 15 (2021).

[3] X. Wang, L. Fu, X. Gan, Y. Wen, G. Zheng, J. Ding, L. Xiang, N. Ye, M. Jin, S. Liang, B. Lu, H. Wang, Y. Xu, C. Deng, S. Zhang, H. Kang, X. Wang, Q. Li, Z. Guo, J. Qi, P. Liu, Y. Ren, L. Wu, J. Yang, J. Zhou, C. Zhou, Acemap: Knowledge discovery through academic graph, 2024. doi:10.48550/arXiv.2403.02576.

[4] S. Fathalla, S. Vahdati, S. Auer, C. Lange, Towards a knowledge graph representing research findings by semantifying survey articles„ in: International Conference on Theory and Practice of Digital Libraries, 2017.

[5] D. Dessì, F. Osborne, D. R. Recupero, D. Buscaldi, E. Motta, Generating knowledge graphs by employing natural language processing and machine learning techniques within the scholarly domain, Future Generation Computer Systems 116 (2021).

[6] S. Auer, V. Ilangovan, M. Stocker, S. Tiwari, L. Vogt, H. Hussein, Open Research Knowledge Graph, 2024.

[7] D. Dessí, F. Osborne, D. Buscaldi, D. R. Recupero, E. Motta, Cs-kg 2.0: A large-scale knowledge graph of computer science, Sci Data 12 (2025). doi:10.1038/s41597-025-05200-8.

[8] Q. Zhang, C. Singh, L. Liu, X. Liu, B. Yu, J. Gao, T. Zhao, Tell your model where to attend: Post-hoc attention steering for llms, in: 12th International Conference on Learning Representations, 2024.

[9] W. Wang, J. Yang, W. Peng, Semantics-adaptive activation intervention for llms via dynamic steering vectors, in: 13th International Conference on Learning Representations (poster), 2025. doi:10.48550/arXiv.2410.12299.

[10] Z. Gu, J. Yao, K. Du, Llmsteer: Improving long-context llm inference by steering attention on reused contexts, in: Machine Learning for Systems Workshop at the 38th Annual Conference on Neural Information ProcessingSystems, 2024. doi:10.48550/arXiv.2411.13009.