Retrieval Augmented Document-Level Information Extraction from Scientific Articles Using Large Language Models

Xizong Zhang, Zixin Jiang, Zhichun Wang†

School of Artificial Intelligence, Beijing Normal University

Engineering Research Center of Intelligent Technology and Educational Application, Ministry of Education

Beijing 100875, China

Abstract

Identifying the main results of scientific papers is an important and challenging task. Unlike traditional named entity recognition (NER) in general domains, NER in the scientific field primarily aims to identify and classify specific scientific entity types from scientific literature. This requires handling professional terminology, complex sentence structures, and domain-specific contextual information. Recently, generation methods based on large language models (LLMs) have attempted to co-train models on multiple datasets while using prompt instructions to accurately extract target entities. However, there is currently insufficient exploration of document-level NER methods in the scientific field. To further fill this gap, we propose a document-level information extraction method for scientific literature based on LLMs. Specifically, the method first constructs a document-level labeled dataset to train a binary classifier based on BERT, which determines the relevance of each sentence to the target entities. In the information extraction phase, specific prompt templates are used to guide LLMs in entity recognition and extraction, while performance is enhanced through fine-tuning with the LoRA framework. Extensive experimental evaluations were conducted on two public datasets, covering zero-shot and supervised conditions. The results show that this method significantly improves the performance in document-level information extraction in the scientific field, surpassing traditional methods.

Keywords: Information Extraction; Scientific Articles; Large Language Models; Generative Information Extraction; Retrieval Augmented Generation.

1. Introduction

Scientific papers present new knowledge and discoveries across various domains, and identifying their main findings is highly beneficial to researchers and professionals in the field. For instance, the *Papers with Code (PWC)* website[1] offers an invaluable resource by compiling tasks, methods, code, and evaluation tables from machine learning papers, making it widely recognized in the community. However, much of the key information on the PWC website relies on collaborative user annotations, leaving many papers with missing or incomplete details. This highlights the pressing need for automated methods to extract critical information from scientific papers.

†Corresponding author: Zhichun Wang (Email: zcwang@bnu.edu.cn, ORCID:0000-0001-8797-7065)

[1]https://paperswithcode.com

This task can be regarded as Named Entity Recognition (NER) in scientific papers. Most recently, LLM-based NER has attracted significant attention from researchers [1, 2, 3]. Most existing NER methods focus on general domains, such as InstructUIE [4] and UniNER [5], which convert various NER datasets into instruction-following formats and then fine-tune encoder-decoder generative models. In scientific domain, NER aims to identify and classify specific scientific entity types from scientific literature, including chemical substances, genes, proteins, diseases, experimental methods, and datasets.

This means that information extraction from scientific literature faces multiple challenges: First, scientific literature covers a wide range of fields, such as computer science, physics, and materials science, each with their specific terminology and structure; second, scientific papers usually contain multiple sections, such as abstracts, introductions, methods, results, and discussions, each with its specific information and format; moreover, to train high-performance models, a large amount of annotated data is required, but annotating scientific literature is very time-consuming, laborious, and costly; finally, in entity extraction tasks, the number of non-entity words is usually much greater than the number of entity words, which can cause the model to overly focus on non-entity words during training, thereby affecting the recognition ability of entity words. These challenges make automated information extraction particularly important and more difficult in scientific literature.

To address the challenges of information extraction from scientific literature, we draw inspiration from the excellent performance of LLMs in natural language processing tasks and propose a new retrieval-augmented document-level information extraction method. This method aims to automatically extract information related to methods, tasks, and datasets from scientific articles to improve the efficiency and accuracy of information processing. Specifically, our method consists of two main stages: sentence retrieval and information extraction. The sentence retrieval stage selects the most important sentences containing information about tasks, methods, and datasets, while the information extraction stage uses LLMs to obtain the final results. In this way, we can not only handle the complex structure of scientific literature but also effectively deal with the challenges brought by domain specificity and document complexity. In addition, our method designs specific prompt templates to guide LLMs to accurately understand and perform information extraction tasks, and uses the LoRA framework for fine-tuning training to improve the performance and generalization ability of the model. We conducted extensive experimental evaluations on two public datasets under zero-shot and supervised conditions, and the results show that our method has significant performance improvements in document-level information extraction applications in the scientific domain, with better results than previous methods.

In summary, the main contributions of this paper include:

- A new document-level information extraction method for the scientific domain based on LLMs is proposed, which is realized through two stages: sentence retrieval and information extraction.
- Specific prompt templates are designed to guide LLMs to accurately understand and perform information extraction tasks, and the LoRA framework is used for fine-tuning training, significantly improving the performance of the model.
- Extensive experiments were conducted on two public datasets, including supervised and zero-shot conditions, and the results show that our method achieves an average improvement of 20% under supervised conditions and more than 10% improvement compared to traditional BERT-based methods.

2. Related Work

The task of scientific information extraction has seen rapid advancements, particularly with the advent of large language models (LLMs). In this section, we review related work in two main groups: general approaches to LLM-based information extraction across domains, specific methods and datasets focused on scientific-domain information extraction.

2.1. LLM-based Information Extraction

Information Extraction (IE) is a crucial NLP area, with LLM-based methods gaining prominence in tasks like Named Entity Recognition, Relation Extraction, and Event Extraction due to their excellent text understanding and generation capabilities [6]. Based on how LLMs are used, approaches can be categorized into zero-shot, few-shot, data-augmented, prompt-designed, constrained decoding generation and supervised fine-tuning ones. Zero-shot enables LLMs to answer without task-specific samples, ideal for new domains [7, 8]. Few-shot trains or learns from context based on a small number of labeled samples, enabling LLMs to generalize from limited data [9, 10]. Data-augmented transforms existing data to generate samples through LLMs, avoiding the introduction of unrealistic or misleading patterns [11, 12]. Prompt design crafts prompts according to task characteristics to guide LLMs in effective information extraction [13, 14]. Constrained decoding generation follows specific constraints or rules during the generation process to improve decoding quality [15]. Supervised fine-tuning builds quality datasets and uses methods like instruction tuning to make LLMs execute IE tasks [16]. At the same time, Universal information extraction has become a mainstream trend [17, 18]. These studies boost IE accuracy and efficiency, but document-level research is scarce. Also, they need much precisely annotated data, which is time-consuming, laborious and costly.

2.2. Scientific Information Extraction

In the field of Scientific IE, both models and datasets have been proposed in recent years. Early SciDeBERTa [19] and SciBERT [20] implemented small-scale information extraction models suitable for scientific fields, and have shown good results so far. Dagdelen et al. [21] fine-tuned a large language model to achieve joint named entity recognition and relation extraction for materials science texts. Kwak et al. [22] tested GPT-4 on a legal will dataset using in-context learning to evaluate its performance. Tang et al. [23] proposed a new method using synthetic data to explore the potential of large language models in clinical text mining and address privacy issues.

Datasets of Scientific IE include SemEval [24], TDMSci [25], SciERC [26], SciREX [27], DMDD [28] and SciDMT [29], covering computer science, physics, and materials science. These data sets from the word level to the sentence level to the summary level and finally extended to the document level of information extraction tasks, gradually improve the breadth and depth of information extraction.

Early traditional information extraction methods and recent methods based on LLMs mostly focus on sentence-level tasks. Although these methods are effective in processing sentence-level data, they often show obvious shortcomings when applied to document-level scientific literature. On the one hand, the structure of scientific literature is complex, with each part having its own specific information and format, which brings great challenges to information extraction. On the other hand, it covers a wide range of fields, each of which has its unique terminology and structure. Therefore, it is particularly urgent and important to develop an information extraction method that can effectively deal with document-level scientific literature.

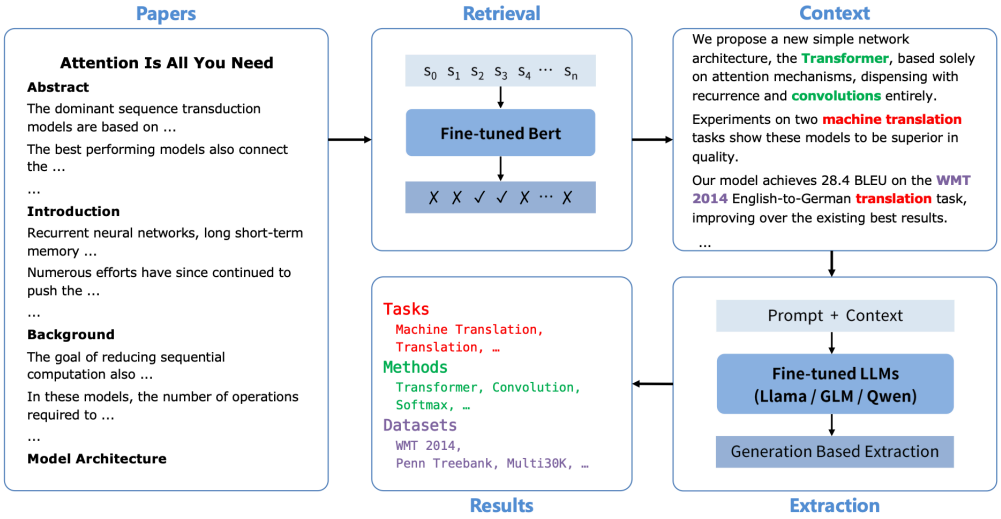

Figure 1: Overview of the proposed retrieval-augmented document-level information extraction framework. The framework operates in two stages: (1) Sentence Retrieval, the full text of a scientific article is processed by a fine-tuned BERT-based classifier to identify sentences that are likely to contain methods, tasks, and datasets; (2) Information Extraction, the retrieved sentences are passed to a LLM to extract the scientific entities.

In summary, while prior studies have demonstrated the effectiveness of LLMs for sentence-level information extraction tasks, there remains a notable gap in approaches that fully leverage document-level context within scientific literature. Additionally, existing datasets and models often lack adaptability to complex, structured scientific documents. Our proposed two-stage framework addresses these gaps by integrating retrieval-augmented sentence selection with prompt-guided and fine-tuned LLMs for more accurate and scalable scientific entity extraction.

3. Method

Figure 1 shows the framework of our proposed approach. It works in two main stages: Sentence Retrieval and Information Extraction. The sentence retrieval stage selects the most important sentences containing the information of tasks, methods, and datasets. The information extraction stage employs LLMs to obtain the final results.

3.1. Task Formulation

Given a paper with n sentences $P = [s_0, s_1, ..., s_n]$, where s_i is the i-th sentence in P, n represents the total number of sentences in the paper, the goal of scientific information extraction is to extract sets of methods M, tasks T, and datasets D from the paper. The results of information extraction are represented as $E = (\{m_1, m_2, ..., m_{|M|}\}, \{t_1, t_2, ..., t_{|T|}\}, \{d_1, d_2, ..., d_{|D|}\})$, where $|M|$, $|T|$, and $|D|$ denote the numbers of extracted methods, tasks, and datasets, respectively.

3.2. Sentence Retrieval

The goal of the sentence retrieval module is to retrieve from the full text statements that may contain target entities such as methods, tasks, or datasets. However, traditional unsupervised

retrieval methods often fail to achieve the desired results (as analyzed in Section 4.3). To this end, we reduce the retrieval task to a sentence-level binary classification problem, which is to decide whether each sentence contains the target entity, and keep the positive example sentences with label 1.

In the sentence retrieval module, we employ the BERTForSequenceClassification architecture to determine the relevance of sentences. This architecture is based on the pre-trained BERT model and augmented with a fully connected layer. The classification head maps the feature vectors output by BERT to two categories: relevant (1) or irrelevant (0). The process of the model can be expressed as:

$$Q = \text{SciRetriever}(P) \tag{1}$$

Here, P denotes the input paper with n sentences and the model outputs a subset Q of sentences labeled 1 that are considered relevant to the target entity.

In the training phase, we use the preprocessed data pairs, including the paper P and its corresponding sentence tag set Q, to optimize the model. The goal of the model is to learn a mapping from an input P to an output Q. We optimize the model using the CrossEntropyLoss function. For each sentence s_i, the model predicts its category $\hat{y}_i$ and calculates the cross-entropy loss with respect to the true label y_i. The formula for the cross-entropy loss is:

$$L_{\text{CE}} = -\sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{2}$$

where N is the total number of sentences in the training data, y_i is the true label of sentence s_i (0 or 1), and $\hat{y}_i$ is the model's predicted probability that sentence s_i belongs to category 1 (relevant). By minimizing the cross-entropy loss, the model learns more accurate parameters, thereby improving classification accuracy.

In the inference stage, we input each sentence s_i from the scientific Paper P into the SciRetriever to obtain its relevance label. Ultimately, we select all sentences labeled as 1 as the output Q, as these sentences are deemed to contain information related to the target entities. Through our SciRetriever, we can efficiently filter out sentences relevant to the target entities from scientific literature, providing high-quality input data for subsequent information extraction tasks.

3.3. Information Extraction

Prompt for Information Extraction

Role:
You are an AI assistant with expertise in natural language processing and information extraction from computer science academic papers.

Instructions:
You will be provided with the text of an academic paper in computer science, your goal is to identify and extract information about the datasets, methods and tasks.

Extraction Guidelines:
- Dataset: In the context of academic research in computer science, a dataset refers to the collection of data used in the research to train, evaluate, or validate the proposed method or model ...
- Method: In the context of academic research in computer science, a method refers to the specific approach, technique, or algorithm used to address a defined task ...
- Task: In the context of academic research in computer science, a task refers to a specific problem or objective that the research is trying to address or solve ...

Output Format:
Please provide the extracted results in the following JSON format ...

Input:
...

In the information extraction module, we employ a LLM to accomplish the task of entity recognition and extraction. To ensure that the LLM can accurately understand and execute the task, we first design a specific prompt for the LLM. This prompt not only includes the role setting for the LLM but also provides a detailed description of the specific requirements of the task, as well as the definitions and examples of the three types of entities to be extracted: Methods, Tasks, and Datasets. Additionally, we specify the output format of the LLM to ensure the consistency and parseability of the output results.

Specifically, our prompt clarifies the role of the LLM as a professional information extraction assistant, responsible for extracting the three types of entities: Methods, Tasks, and Datasets, from the given text. Methods are defined as the approaches, techniques, or algorithms used to solve problems or accomplish tasks, such as ”deep learning” or ”random forest”; Tasks are defined as the specific work or objectives that need to be completed, such as ”image classification” or ”text generation”; Datasets are defined as the collections of data used for training or testing models, such as the ”MNIST dataset” or the ”IMDb movie review dataset”. The output format is specified as JSON, containing three lists corresponding to the entities of Methods, Tasks, and Datasets.

Thus, the reasoning process of the LLM can be represented as:

$$E = \text{SciExtractor}(Q, I) \tag{3}$$

where Q is the input sentences, I is the designed prompt, and E is the extracted entities. Through

this approach, we can effectively leverage the powerful capabilities of the LLM to accomplish complex entity extraction tasks while ensuring the accuracy and consistency of the output results.

To further enhance the LLM's ability to extract entities, we adopt LoRA (Low-Rank Adaptation). LoRA achieves fine-tuning by introducing two low-rank matrices $L \in \mathbb{R}^{d_{\text{out}} \times r}$ and $R \in \mathbb{R}^{r \times d_{\text{in}}}$ in each linear layer $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ of the LLM. Here, r is a value much smaller than d_{out} and d_{in}, allowing the product of L and R to effectively capture the key features of the weight matrix W while significantly reducing the number of parameters and improving the training efficiency of the model.

The fine-tuned weight W' is constructed through the product of the original weight W and the low-rank matrices L and R:

$$W' = W + L \times R \tag{4}$$

This fine-tuning method not only retains most of the original model's parameters but also adjusts the model to meet specific task requirements through a small number of trainable parameters L and R. Using the fine-tuned weight W' to replace the original weight W for forward propagation generates the output E, thereby improving the model's performance in entity extraction tasks.

During the training process, we used CrossEntropyLoss as the loss function. By minimizing the cross-entropy loss, the model learns more accurate parameters, thereby improving its performance in classification tasks. In our information extraction task, using the cross-entropy loss function effectively guides the model to recognize and classify entities in scientific literature, such as methods, tasks, and datasets.

4. Experiments

This section presents experiments that answer the following research questions.
RQ1: does retrieving relevant sentences help improve the extraction results?
RQ2: is our sentence retriever effective in improving the extraction results?
RQ3: can the LLMs be fine-tuned for the extraction task?
RQ4: is our extraction framework work effectively and has advantages over other methods?

4.1. Experimental Setting

4.1.1. Datasets

- SciDMT [29] is a large-scale corpus designed to support the detection of scientific entity mentions, focusing on datasets, methods, and tasks in scientific literature. The dataset consists of two components: a primary corpus of 48,049 scientific papers with over 1.8 million entity mentions annotated using distant supervision, and an evaluation set of 100 papers manually annotated.

- SciREX [27] is a dataset designed for document-level information extraction, encompassing multiple tasks such as salient entity identification and document-level N-ary relation extraction, specifically from scientific articles. The dataset combines automatic and human annotations to address the complexities of extracting information from complete documents.

Table 1 outlines the details of the datasets.

Table 1: Statistics of the SciDMT and SciREX Datasets

Dataset	Docs.	Samples	Types	Avg. Entities	Avg. Sentences	Avg. Tokens
SciDMT	48,049	100	3	8	278	8,464
SciREX	438	66	4	6	294	7,018

4.1.2. Data Processing

To implement the sentence retrieval module, we constructed a document-level labeled dataset with $P-Q$ data pairs. Each document d_i in the dataset $D = \{d_1, d_2, \ldots, d_n\}$ contains the following information:

- **id**: A unique identifier for each paper.
- **words**: The list of words in the document.
- **entities**: Document-level entity information, represented as entities $= \{e_1 : L_{e1}, e_2 : L_{e2}, \ldots\}$, where e is the entity type and L_e is the corresponding entity list.
- **sentence_spans**: The starting positions of each sentence in the document, i.e., sentence_spans $= [(\text{start}_1, \text{end}_1), (\text{start}_2, \text{end}_2), \ldots]$.
- **mentions**: Entity mentions for NER tasks, with each mention containing its span and entity type, i.e., mentions $= \{(\text{span}_1, \text{type}_1), (\text{span}_2, \text{type}_2), \ldots\}$.

We established a mapping between entities and mentions using coreference resolution and regex matching to identify all occurrences of each entity and generate their span lists. Each mention m was then matched to the sentence containing the target entity based on its span. If a mention's span is entirely within a sentence, the sentence is labeled as relevant (1). The set of relevant sentences for entity e is defined as:

$$S_e = \{s_{ij} \in \text{sentences}_i \mid \exists m \in \text{mentions}(e), \text{span}(m) \subseteq s_{ij}\}$$

where mentions(e) denotes all mentions of entity e, and span(m) indicates the position range of mention m in the document. Table 2 outlines the detailed information of the Labeled Datasets with $P-Q$ Data Pairs.

Table 2: Statistics of the Labeled Datasets with $P-Q$ Data Pairs

Dataset	Docs.	Positive Sentences	Negative Sentences
SciDMT	500	52,751	86,038
SciREX	300	29,927	53,204

4.1.3. Implementation Detail

To ensure a rigorous and reproducible experimental setup, we detail below the hardware environment, software stack, model selection rationale, fine-tuning configuration, validation strategy, and training cost.

Hardware and Software Environment. All experiments were conducted on a server equipped with a single NVIDIA A800 GPU (80 GB), 512 GB RAM, running Ubuntu 20.04. The software environment included Python 3.10.14, CUDA 11.8, PyTorch 2.1.2, HuggingFace Transformers 4.43.2, and PEFT 0.11.1. The fine-tuning was implemented using the LLaMAFactory framework for efficient LoRA-based adaptation.

LLM Selection Criteria. We selected three widely-used open-source large language models: LLaMA3.1-8B, GLM4-9B, and Qwen2.5-7B. These models were chosen based on the following criteria: (1) strong performance on instruction-following and information extraction tasks; (2) broad community adoption and availability; and (3) architectural diversity, which allowed us to evaluate the robustness and generalizability of our framework across different LLM families.

LoRA Fine-tuning Configuration. To adapt the LLMs to the entity extraction task efficiently, we adopted the Low-Rank Adaptation (LoRA) method. The Fine-tuning was conducted using the LoRA method within the efficient llamafactory [30] training framework. The training utilized a batch size of 1, gradient accumulation steps of 8, and a learning rate of 1e-4. The AdamW optimizer was used in conjunction with a cosine learning rate scheduler and a warmup ratio of 0.1. The models were trained for 5 epochs with CrossEntropyLoss as the loss function and bf16 precision to optimize both computational efficiency and model adaptation across all layers.

Validation Strategy. During training, we split each dataset into 70% for training, 20% for testing and 10% for validation. We applied early stopping based on validation loss with a patience of 2 epochs to avoid overfitting. Performance was monitored on the validation set using precision, recall, and F1 scores.

Computational Cost and Reproducibility. Each fine-tuning experiment required approximately 3-4 hours. To ensure reproducibility, we fixed all random seeds to 87 across modules and documented environment configurations. The source codes, configuration files, and fine-tuning scripts will be released publicly upon publication to support reproducibility and further experimentation.

4.1.4. Evaluation

We use precision, recall and F1 scores to evaluate performance. In generative information extraction tasks based on LLMs, the models often output results that are semantically consistent with the answers but have differences in expression. In order to ensure fairness and objectivity of evaluation, we redefine the rule indicating that the two results are matched instead of direct perfect matching.

4.2. Main Results

Table 3 shows the results of four groups of models, including Bert-based models, Zero-shot Extraction with LLMs, LLMs Enhanced by Retrieval Module, and our proposed models.

4.2.1. Bert-based Models

We initially assess the performance of the traditional BERT-based approach on the task of scientific document information extraction. The test results show that while the BERT model has certain capabilities in processing scientific document information extraction tasks, achieving an F1 score of 64.77 on the SciDMT dataset, its performance is still somewhat inferior compared to models specifically designed for the scientific domain, such as SciBERT and SciDeBERTa, which demonstrats a +2% higher F1 score on both datasets. The additional pre-training of SciBERT and SciDeBERTa on scientific literature enables them to more effectively understand and

Table 3: Main Results

Model	SciDMT			SciREX		
	Precision	Reccall	F1	Precision	Recall	F1
Bert-based models						
BERT	51.52	87.18	64.77	44.78	60.73	51.55
SciBERT	55.28	85.56	67.17	50.04	65.37	56.69
SciDeBERTa	54.95	85.62	66.94	46.32	68.69	55.33
Zero-shot Extraction with LLMs						
GPT-4o	48.04	48.76	48.40	28.64	58.92	38.55
Llama3.1-8b	43.35	57.81	49.55	20.80	58.75	30.72
GLM4-9b	33.91	50.87	40.69	20.43	56.86	30.06
Qwen2.5-7b	38.99	39.03	39.01	25.14	56.07	34.72
LLMs Enhanced by Retrieval Module						
GPT-4o	60.42	51.52	55.61	36.52	70.28	48.06
Llama3.1-8b	51.03	57.62	54.13	27.49	74.17	40.11
GLM4-9b	43.61	52.85	47.79	26.80	63.10	37.62
Qwen2.5-7b	52.59	46.78	49.52	32.70	64.72	43.45
Ours (Retrieval Module + Fine-tuned Extraction Module)						
$\text{Ours}_{Llama3.1-8b}$	72.30	74.29	73.28	64.30	64.11	64.20
$\text{Ours}_{GLM4-9b}$	73.22	75.15	74.17	65.34	61.93	63.59
$\text{Ours}_{Qwen2.5-7b}$	72.58	73.92	73.24	66.57	66.51	66.54

represent the unique terminology and structures found in scientific texts, thus achieving higher scores on both datasets.

4.2.2. Zero-shot Extraction with LLMs

To evaluate the capability of LLMs for scientific information extraction, we conduct document-level full-text zero-shot experiments with these models. The results indicate that the performance of the four LLMs on both datasets is significantly lower than that of traditional BERT-based methods. Even the most powerful among them, GPT-4o, shows a deficit of -18% on the SciDMT dataset and -17% on the SciREX dataset in terms of F1 scores. This suggests that despite the strong capabilities of LLMs demonstrated across a variety of natural language processing tasks, they still require further optimization and fine-tuning to achieve or surpass the performance of models specifically designed for scientific document information extraction in this particular domain.

4.2.3. LLMs Enhanced by Retrieval Module

To assess the effectiveness of our retrieval module, we integrate it into the LLMs while keeping all other experimental settings unchanged, and then conduct zero-shot experiments again. The results show that with the addition of the retrieval module, the four models demonstrate an approximate improvement of +10% in Precision, Recall, and F1 scores on the SciDMT dataset. On the SciREX dataset, Recall and F1 scores also exhibit an improvement of about +10%. These findings answer **RQ1** that incorporating a retrieval module can significantly improve the information extraction capability of LLMs. Nevertheless, the performance of these models is still slightly lower than that of traditional BERT-based methods.

4.2.4. Our Proposed Models

For **RQ4**, we show the experimental results of our extraction framework on three different base models. The results reveal that, compared to using the base models on their own, the integration of our retrieval and extraction modules results in a significant enhancement of approximately +23% to +39% for the three base models on the SciDMT dataset. Even more impressive gains, ranging from +10% to +41%, are observed on the SciREX dataset. Furthermore, the models' performance is about +7% F1 higher than that of traditional BERT-based models , which further confirms the effectiveness of our method in markedly improving the models' performance on the task of scientific document information extraction.

4.3. Analysis

4.3.1. Analysis of Retrieval Module

In the main experiments, we utilized the sentence-level entity mentions from the original dataset, associating document-level entities with sentence-level mentions to construct a document-level context retrieval dataset, thereby training a retrieval model.

To answer **RQ2**, we conducted an experimental analysis of our extraction module SciRetriver with traditional retrieval methods The specific settings are as follows:

- Sparse Search: Retrieval based on the BM25 algorithm.
- Dense Search: Using BGE as the text encoding model and Faiss as the retrieval engine.

The retrieval question was set as: "Retrieve sentences from the given research paper that mention datasets, methods, or tasks explicitly or implicitly, such as the use of benchmark datasets (e.g., CoNLL-2003, ACE2005), proposed models or methods (e.g., BiLSTM, BERT), or described research objectives or tasks (e.g., named entity recognition, relation extraction)."

In the experiments, we selected the top 50 sentences. These experiments were conducted on the SciDMT dataset in a zero-shot setting to evaluate the performance of traditional retrieval methods in the absence of fine-grained annotated data. Through these experiments, we aim to verify the effectiveness and feasibility of traditional retrieval methods in the task of information extraction from scientific literature.

Table 4 and Figure 2 show that the performance of various models in dense retrieval is slightly better than that in sparse retrieval. However, they all fall short considerably compared to our retrieval module. This indicates that relying solely on traditional retrieval methods is not effective in accurately retrieving key contextual information.

Table 4: Performance comparison of different Retrieval Module.

Retrieval	Model	Precision	Recall	F1
BM25	GPT-4o	47.30	46.16	46.72
	Llama3.1-8B	38.88	50.93	44.09
	GLM4-9B	32.57	46.47	38.29
	Qwen2.5-7B	41.69	41.33	41.51
BGE	GPT-4o	50.73	46.04	48.27
	Llama3.1-8B	40.63	49.50	44.63
	GLM4-9B	34.02	47.21	39.55
	Qwen2.5-7B	43.47	41.39	42.40
Our Retrieval Module	GPT-4o	60.42	51.52	55.61
	Llama3.1-8b	51.03	57.62	54.13
	GLM4-9b	43.61	52.85	47.79
	Qwen2.5-7b	52.59	46.78	49.52

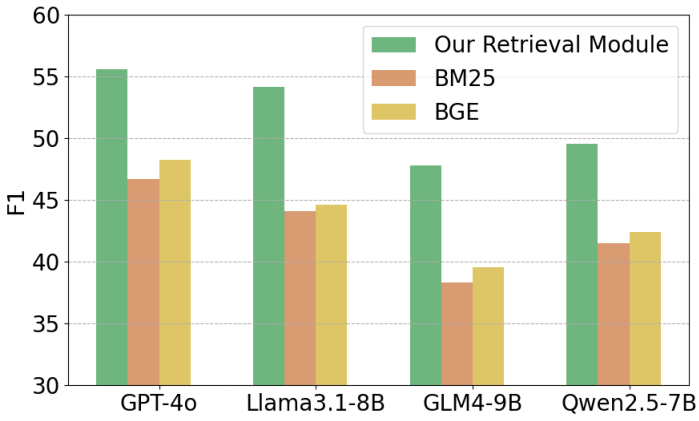

Figure 2: Performance comparison of different Retrieval Module.

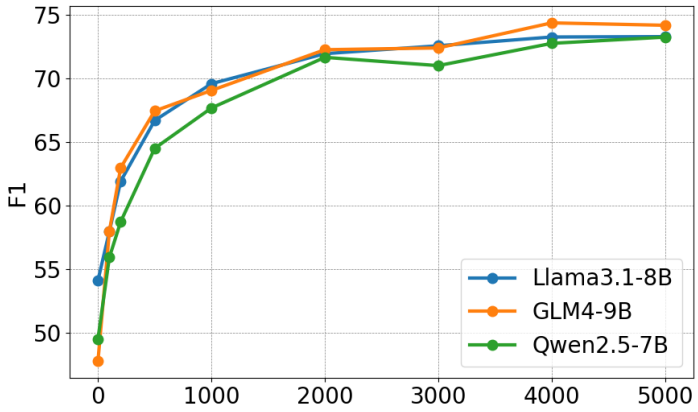

Figure 3: Performance comparison of different LLM backbones with varying training dataset sizes.

4.3.2. Analysis of Extraction Module

For **RQ3**, we analyze the performance of different models on different sizes of training data, as shown in Figure 3. When the amount of training data is small ($\leq$ 2000), the F1 scores of all models improve significantly as the amount of data increases. This indicates that adding more training data has a substantial positive impact on model performance when the data volume is limited. However, once the training data volume exceeds 2000, the performance improvement begins to slow down and stabilizes, demonstrating a trend of diminishing marginal returns.

This finding is particularly significant for research and application scenarios with limited resources. It implies that in practical applications of information extraction tasks, especially in entity recognition and classification within scientific literature, we may not need to collect and label large amounts of training data. Instead, we can enhance model performance by optimizing model structures and training strategies. Moreover, this discovery provides direction for future research, specifically exploring how to make more effective use of data when the volume is small and how to design more efficient models and algorithms to handle information extraction tasks in scientific literature.

In summary, these analysis results not only confirm the effectiveness of our proposed method across different scales of data but also offer valuable insights and guidance for information extraction in scientific literature.

5. Conclusion

This paper proposes an entity extraction method based on LLMs for document-level entity extraction tasks in the scientific domain. The method first extracts sentences related to entities, designs targeted LLM prompt templates, and introduces them into LLM-based entity extraction, with fine-tuning conducted using the LoRA framework. Extensive experiments were carried out on two public datasets, including both supervised and zero-shot conditions. The results indicate that, compared to other traditional methods, this method demonstrates superior performance in document-level entity extraction applications within the scientific domain.

Beyond its application to computer science papers, our framework is inherently extensible to other scientific domains and a broader range of entity types. The retrieval module can be retrained on domain-specific data to adapt to different writing styles and terminologies, while the prompt-driven design of the extraction module allows for flexible modification to support additional entity categories such as materials, instruments, or phenomena. Although this study focuses on Methods, Tasks, and Datasets, future work will explore applying the framework to disciplines like physics and materials science, where document structures and terminologies differ significantly. This direction holds promise for building more comprehensive, cross-domain scientific knowledge extraction systems.

Author Contributions

Xizong Zhang: proposed the research problems; designed the research framework; performed the research; collected and analyzed the data; wrote and revised the manuscript.

Zixin Jiang: designed the research framework; performed the research; wrote and revised the manuscript.

Zhichun Wang: proposed the research problems; designed the research framework; wrote and revised the manuscript.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62276026).

References

[1] W. Hou, W. Zhao, X. Liu, and W. Guo, "Knowledge-enriched prompt for low-resource named entity recognition," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 5, pp. 1–15, 2024.

[2] S. Wang, X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li, and G. Wang, "Gpt-ner: Named entity recognition via large language models," *arXiv preprint arXiv:2304.10428*, 2023.

[3] Y. Qi, H. Peng, X. Wang, B. Xu, L. Hou, and J. Li, "ADELIE: Aligning large language models on information extraction," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 7371–7387. [Online]. Available: https://aclanthology.org/2024.emnlp-main.419/

[4] X. Wang, W. Zhou, C. Zu, H. Xia, T. Chen, Y. Zhang, R. Zheng, J. Ye, Q. Zhang, T. Gui, J. Kang, J. Yang, S. Li, and C. Du, "Instructuie: Multi-task instruction tuning for unified information extraction," 2023. [Online]. Available: https://arxiv.org/abs/2304.08085

[5] W. Zhou, S. Zhang, Y. Gu, M. Chen, and H. Poon, "Universalner: Targeted distillation from large language models for open named entity recognition," *arXiv preprint arXiv:2308.03279*, 2023.

[6] D. Xu, W. Chen, W. Peng, C. Zhang, T. Xu, X. Zhao, X. Wu, Y. Zheng, Y. Wang, and E. Chen, "Large language models for generative information extraction: A survey," *Frontiers of Computer Science*, vol. 18, no. 6, p. 186357, 2024.

[7] T. Xie, Q. Li, J. Zhang, Y. Zhang, Z. Liu, and H. Wang, "Empirical study of zero-shot NER with ChatGPT," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 7935–7956. [Online]. Available: https://aclanthology.org/2023.emnlp-main.493/

[8] K. Zhang, B. Jimenez Gutierrez, and Y. Su, "Aligning instruction tasks unlocks large language models as zero-shot relation extractors," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 794–812. [Online]. Available: https://aclanthology.org/2023.findings-acl.50/

[9] Z. Wan, F. Cheng, Z. Mao, Q. Liu, H. Song, J. Li, and S. Kurohashi, "GPT-RE: In-context learning for relation extraction using large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 3534–3547. [Online]. Available: https://aclanthology.org/2023.emnlp-main.214/

[10] C. Pang, Y. Cao, Q. Ding, and P. Luo, "Guideline learning for in-context information extraction," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 15 372–15 389. [Online]. Available: https://aclanthology.org/2023.emnlp-main.950/

[11] R. Zhang, Y. Li, Y. Ma, M. Zhou, and L. Zou, "LLMaAA: Making large language models as active annotators," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 13 088–13 103. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.872/

[12] M. D. Ma, X. Wang, P.-N. Kung, P. J. Brantingham, N. Peng, and W. Wang, "Star: Boosting low-resource information extraction by structure-to-text data generation with large language models," 2024. [Online]. Available: https://arxiv.org/abs/2305.15090

[13] X. Wei, X. Cui, N. Cheng, X. Wang, X. Zhang, S. Huang, P. Xie, J. Xu, Y. Chen, M. Zhang *et al.*, "Chatie: Zero-shot information extraction via chatting with chatgpt," *arXiv preprint arXiv:2302.10205*, 2023.

[14] C. Yuan, Q. Xie, and S. Ananiadou, "Zero-shot temporal relation extraction with chatgpt," *arXiv preprint arXiv:2304.05454*, 2023.

[15] S. Geng, M. Josifoski, M. Peyrard, and R. West, "Grammar-constrained decoding for structured nlp tasks without finetuning," *arXiv preprint arXiv:2305.13971*, 2023.

[16] H. Wu, Y. Yuan, L. Mikaelyan, A. Meulemans, X. Liu, J. Hensman, and B. Mitra, "Learning to extract structured entities using language models," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 6817–6834. [Online]. Available: https://aclanthology.org/2024.emnlp-main.388/

[17] X. Xiao, Y. Wang, N. Xu, Y. Wang, H. Yang, M. Wang, Y. Luo, L. Wang, W. Mao, and D. Zeng, "Yayi-uie: A chat-enhanced instruction tuning framework for universal information extraction," 2024. [Online]. Available: https://arxiv.org/abs/2312.15548

[18] H. Gui, L. Yuan, H. Ye, N. Zhang, M. Sun, L. Liang, and H. Chen, "IEPile: Unearthing large scale schema-conditioned information extraction corpus," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 127–146. [Online]. Available: https://aclanthology.org/2024.acl-short.13/

[19] Y. Jeong and E. Kim, "Scideberta: Learning deberta for science technology documents and fine-tuning information extraction tasks," *IEEE Access*, vol. 10, pp. 60 805–60 813, 2022.

[20] I. Beltagy, K. Lo, and A. Cohan, "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)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3615–3620. [Online]. Available: https://aclanthology.org/D19-1371/

[21] J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson, and A. Jain, "Structured information extraction from scientific text with large language models," *Nature Communications*, vol. 15, no. 1, p. 1418, 2024.

[22] A. Kwak, C. Jeong, G. Forte, D. Bambauer, C. Morrison, and M. Surdeanu, "Information extraction from legal wills: How well does GPT-4 do?" in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 4336–4353. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.287/

[23] R. Tang, X. Han, X. Jiang, and X. Hu, "Does synthetic data generation of llms help clinical text mining?" *arXiv preprint arXiv:2303.04360*, 2023.

[24] K. Gábor, D. Buscaldi, A.-K. Schumann, B. QasemiZadeh, H. Zargayouna, and T. Charnois, "SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers," in *Proceedings of the 12th International Workshop on Semantic Evaluation*, M. Apidianaki, S. M. Mohammad, J. May, E. Shutova, S. Bethard, and M. Carpuat, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 679–688. [Online]. Available: https://aclanthology.org/S18-1111/

[25] Y. Hou, C. Jochim, M. Gleize, F. Bonin, and D. Ganguly, "TDMSci: A specialized corpus for scientific literature entity tagging of tasks datasets and metrics," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds. Online: Association for Computational Linguistics, Apr. 2021, pp. 707–714. [Online]. Available: https://aclanthology.org/2021.eacl-main.59/

[26] Y. Luan, L. He, M. Ostendorf, and H. Hajishirzi, "Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 3219–3232. [Online]. Available:

https://aclanthology.org/D18-1360/
[27] S. Jain, M. van Zuylen, H. Hajishirzi, and I. Beltagy, "SciREX: A challenge dataset for document-level information extraction," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 7506–7516. [Online]. Available: https://aclanthology.org/2020.acl-main.670/
[28] H. Pan, Q. Zhang, E. Dragut, C. Caragea, and L. J. Latecki, "DMDD: A large-scale dataset for dataset mentions detection," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1132–1146, 2023. [Online]. Available: https://aclanthology.org/2023.tacl-1.64/
[29] H. Pan, Q. Zhang, C. Caragea, E. Dragut, and L. J. Latecki, "SciDMT: A large-scale corpus for detecting scientific mentions," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 14 407–14 417. [Online]. Available: https://aclanthology.org/2024.lrec-main.1256
[30] Y. Zheng, R. Zhang, J. Zhang, Y. Ye, Z. Luo, Z. Feng, and Y. Ma, "Llamafactory: Unified efficient fine-tuning of 100+ language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Bangkok, Thailand: Association for Computational Linguistics, 2024. [Online]. Available: http://arxiv.org/abs/2403.13372

Author Biography


Xizong Zhang received his BS degree in Computer Science and Technology from the College of Artificial Intelligence, Beijing Normal University in 2020. He is currently pursuing a master's degree at the College of Artificial Intelligence, Beijing Normal University. His research interests include scientific information extraction.


Zixin Jiang received her BS degree in Computer Science and Technology from the College of Artificial Intelligence, Beijing Normal University in 2024. She is currently pursuing a master's degree at the College of Artificial Intelligence, Beijing Normal University. Her research interests include knowledge graph completion and link prediction.


Zhichun Wang is currently an associate professor at the School of Artificial Intelligence, Beijing Normal University. He received his Ph.D. degree from Tianjin University

in 2010, and worked as post-doctoral researcher in Tsinghua University from 2011 to 2012. His research interests include large language models and knowledge graphs.