LLM-Powered Ensemble Learning for Paper Source Tracing: A GPU-Free Approach

Kunlong Chen
Junjun Wang
Zhaoqun Chen
chenkunlong.ckl@antgroup.com
xingruo.wjj@antgroup.com
zhaoqun.czq@antgroup.com
Ant Group
Hangzhou, Zhejiang, China

Kunjin Chen
kunjin.ckj@taobao.com
Alibaba Group
Hangzhou, Zhejiang, China

Yitian Chen
yitianartsky@gmail.com
Cardinal Operations
Beijing, China

Abstract

We participated in the KDD CUP 2024 paper source tracing competition and achieved the 3rd place. This competition tasked participants with identifying the reference sources (i.e., *ref-sources*, as referred to by the organizers of the competition) of given academic papers. Unlike most teams that addressed this challenge by fine-tuning pre-trained neural language models such as Bert or ChatGLM, our primary approach utilized closed-source large language models (LLMs). With recent advancements in LLM technology, closed-source LLMs have demonstrated the capability to tackle complex reasoning tasks in zero-shot or few-shot scenarios. Consequently, in the absence of GPUs, we employed closed-source LLMs to directly generate predicted reference sources from the provided papers. We further refined these predictions through ensemble learning. Notably, our method was the only one among the award-winning approaches that did not require the use of GPUs for model training. The code can be found in https://github.com/Cklwanfifa/KDDCUP2024-PST

Keywords

Paper source tracing, Large Language Model

1 Introduction

In the rapidly evolving landscape of scientific research, tracing the origins of ideas of academic papers has become increasingly crucial. The ability to accurately identify the source references of a paper not only enhances our understanding of the pathways of scientific progress but also holds significant scientific and societal value [Zhang et al. 2024a]. As the volume of scientific publications continues to grow exponentially [Zha et al. 2019], there is an urgent need for efficient algorithms that can swiftly identify the sources of papers. Such tools are particularly valuable for emerging researchers, enabling them to quickly grasp the developmental trajectory of specific technologies and situate their work within the broader scientific context.

Traditionally, researchers have used methods such as random forests [Valenzuela et al. 2015] or a neural network training [Yin et al. 2021] to identify key citations for each paper. However, these approaches often require substantial time and effort in designing intricate features or fine-tuning pre-trained models to achieve satisfactory performance. The advent of large language model (LLM) technologies has opened new avenues for addressing this challenge. Recent studies have demonstrated that when the parameter count is sufficiently high, LLMs can solve complex problems through zero-shot learning [Kojima et al. 2022].

In this paper, we propose a novel approach that leverages pre-trained LLMs as unsupervised reasoners to identify a paper's source references. Our method utilizes the paper's text and related information to perform direct reasoning, eliminating the need for extensive feature engineering or model fine-tuning. This approach is particularly beneficial for researchers with limited GPU resources, as it offers a balance between computational efficiency and performance accuracy.

The efficacy of our proposed method is validated by its success in the paper source tracing competition of KDD CUP 2024, where it secured third place. Notably, our team was the only award-winning entry that did not rely on GPUs, underscoring the method's ability to achieve competitive results with minimal computational resources. This achievement demonstrates that our approach effectively balances resource consumption and performance, making it a valuable tool for a wide range of researchers and institutions.

2 PRELIMINARIES

2.1 Data Description

The dataset we used in this study is from the Open Academic Graph dataset [Zhang et al. 2024b], which includes training and test sets for multiple tasks under the academic paper mining theme, such as Author Name Disambiguation, Academic question answering, and Paper source tracing. The method we employ is applicable to the data from the Paper source tracing task in the OAG dataset [Zhang et al. 2024a], which is referred to as the PST dataset in the following

text. The PST dataset contains 1576 papers related to computer science and 55014 associated citations.

2.2 Task Description

The primary objective of our study is to identify the most significant reference paper for a given academic work. To formalize this task, we introduce the following notation.

- Let $\mathcal{A}$ denote the set of all references that have been cited at least once across the entire dataset.
- For each paper P_i, we define $R_i = [r_{i,1}, r_{i,2}, \cdots, r_{i,J_i}]$ as the set of its references, where each $r_{i,j} \in \mathcal{A}$.
- We denote $S_i \subset R_i$ as the subset of references that serve as source references for paper P_i.

Our task is to assign a probability score $p_{i,j} \in [0, 1]$ to each reference $r_{i,j}$ in R_i, where higher values indicate a greater likelihood of the reference being a key source for the paper. To evaluate the performance of our model, we employ the following metrics:

- For each paper P_i, we calculate the Average Precision (AP), denoted as AP_i.
- The overall model performance is assessed using the Mean Average Precision (MAP), computed as $\text{MAP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i$, where N is the total number of papers in the dataset.

3 METHOD

Given the constraints of limited access to high-performance GPUs, we developed an innovative approach that diverges from traditional methods such as fine-tuning large pre-trained models or employing graph machine learning techniques. Our methodology primarily leverages closed-source LLMs in conjunction with feature engineering and ensemble learning to address the paper source tracing problem effectively.

3.1 Feature Engineering

While our method relies heavily on LLMs for text comprehension and problem-solving, we found that incorporating carefully engineered features from the papers enhances performance. Our feature extraction process focuses on the following key aspects:

(1) Paper Metadata: We extract information such as the publication venue (journal/conference name), citation count, and author details (nationality and affiliation) for each paper.
(2) Citation Statistics: We compute the total occurrences of each cited reference within the paper, as well as its distribution across different sections or chapters.
(3) Reference Metadata: Similar to the main paper, we extract publication venue and author information for each cited reference.
(4) Contextual Keywords: We count the occurrences of specific keywords (e.g., *motivated by*, *inspired by*) in the vicinity of each citation to gauge its importance.

These features provide a rich, structured representation of the papers and their citations, complementing the semantic understanding capabilities of the LLMs.

3.2 LLM-generated Answers

To generate diverse and high-quality answers, we employ four state-of-the-art LLMs: GPT-4 Turbo, GPT-4o, Gemini 1.5 Pro, and Claude 3 Opus. Our selection criterion was based on empirical evidence suggesting that LLMs with an MMLU (Massive Multitask Language Understanding) score above 85 demonstrate superior logical reasoning capabilities in academic contexts.

3.2.1 Prompt Engineering. We designed a set of prompts to elicit nuanced responses from the LLMs:

(1) Base prompt: A carefully crafted instruction set (illustrated in Figure 1) that guides the LLM in identifying source citations.
(2) Inspiration-focused prompt: This variant asks the LLM to categorize citations as "direct inspiration," "indirect inspiration," or "other inspiration," emphasizing the keyword "inspiration."
(3) Title-enriched prompt: An extension of the base prompt that includes the titles of cited articles for additional context.
(4) Meta-optimized prompt: We utilize GPT-4 to refine the base prompt itself, leveraging the LLM's meta-learning capabilities.
(5) Notes-based prompt: For papers with available "notes" fields (potentially containing annotators' insights), we instruct the LLM to identify key citations based on these descriptions.

3.2.2 Answer Generation. Let N_{LLM} denote the number of LLM types and N_{prompt} the number of prompt variants. Theoretically, for each academic paper, we could generate up to $M_p = N_{LLM} \cdot N_{prompt} \cdot M$ answers, where M is the maximum number of answers per LLM-prompt combination. However, due to practical constraints, we limit this to M_p answers per paper.

For each cited reference $r_{i,j}$ in the citation set R_i of paper P_i, we generate a probability list $P_{i,j}^{LLM}$ of length M_p, where each element represents the LLM-estimated probability of $r_{i,j}$ being a source citation. Formally:

$$P_{i,j}^{LLM} = [p_1, p_2, \ldots, p_{M_p}], \quad p_k \in [0, 1] \tag{1}$$

3.3 Base Models

To complement the LLM-generated probabilities, we employ two gradient boosting frameworks: LightGBM [Ke et al. 2017] and Cat-Boost [Prokhorenkova et al. 2018]. These models operate on the engineered features described in Section 3.1.

For each paper-citation pair $(P_i, r_{i,j})$ in the training set, we construct a feature vector $X_{i,j}$ and a corresponding binary label $Y_{i,j} \in \{0, 1\}$, where $Y_{i,j} = 1$ indicates that $r_{i,j}$ is a reference source of P_i. We train two classifiers:

$$f_{lgb}(X_{i,j}) \to [0, 1] \quad \text{(LightGBM)} \tag{2}$$

$$f_{cb}(X_{i,j}) \to [0, 1] \quad \text{(CatBoost)} \tag{3}$$

These classifiers provide probability estimates for each citation being a source citation based on the engineered features.

3.4 Ensemble Approach

Our final prediction model combines the strengths of LLM-generated answers and traditional machine learning classifiers. For each paper-citation pair $(P_i, r_{i,j})$, our goal is to find a function that combines the prediction results of different methods as:

$$P_{i,j} = f_{ensemble}((P_{i,j}^{LLM}), f_{lgb}(X_{i,j}), f_{cb}(X_{i,j})) \quad (4)$$

This ensemble approach allows us to leverage the semantic understanding of LLMs while benefiting from the structured feature representations captured by traditional models.

3.5 Our framework

In this section, we present our complete approach. The algorithm takes as input the main text of paper P_i and its list of cited references R_i, and outputs the probability $p_{i,j}$ for each reference $r_{i,j}$ being a source citation of paper P_i.

(1) Using f_{lgb} and f_{cb} trained on the training set, we score each reference $r_{i,j}$, where the score represents the probability of it being a source citation. This score is represented as $p_{i,j}^{base} = p_{i,j}^{lgb} \times \omega^{lgb} + p_{i,j}^{cb} \times \omega^{cb}$.

(2) We modify $p_{i,j}^{base}$ using the results from LLMs. This process involves two main steps:

- First, we group the LLM outputs into N_{LLM}^g groups based on prompt type and base model type. Within each group, we ask each LLM to provide a confidence score for its predicted key citations, as represented by the prob2score function (Equation 5). Additionally, we assign a weight ω_i^{llm} to each group based on expert knowledge, representing the importance of that LLM group. Finally, we aggregate the results of all LLM groups according to their weights to obtain $score_{i,j}^{bonus}$, which serves as a bonus to be added to $p_{i,j}^{base}$.

- In the second step, for each candidate reference, we aggregate the results from all LLMs. If the score given by the LLM at the p_{neg}-th percentile is less than $p_{threshold}^{neg}$, we reduce the bonus magnitude from the previous step. Our approach is to divide $score_{i,j}^{bonus}$ by a constant C_{neg}.

$$\text{prob2score}(x) = \begin{cases} 3, & \text{if } x \geq 0.9 \\ 2, & \text{if } 0.5 \leq x < 0.9 \\ 1, & \text{if } 0.4 < x < 0.5 \\ 0, & \text{if } x \leq 0.4 \end{cases} \quad (5)$$

(3) Our final prediction can be expressed as:

$$p_{i,j}^{final} = p_{i,j}^{base} + \omega_f \cdot score_{i,j}^{bonus} \quad (6)$$

where ω_f is the weight used to adjust the base score using the LLM prediction results.

4 Result

We present the performance of our proposed method on the validation set in Table 1. As can be observed, using LightGBM or LLM individually does not yield high scores. When the two approaches are combined, however, a significant improvement in the score

Table 1: MAP scores on validation set

Method	Score
SciBert (baseline)	0.29
Our method (only LightGBM)	0.42
Our method (only LLM)	0.45
Our method (combined)	0.50

Table 2: The configuration of parameters

Parameters	Value
ω^{lgb}	0.4
ω^{cb}	0.6
ω_f	0.035
$p_{threshold}^{neg}$	0.2
C_{neg}	4
p_{neg}	0.4

can be achieved. This demonstrates the effectiveness of ensemble learning techniques [Zhou 2021]. In addition, we present the configuration of parameters in Table 1.

5 Conclusion

In this paper, we have briefly introduced our method that secured third place in the KDD CUP 2024 paper source tracing competition. Our approach leverages the zero-shot capabilities of LLMs, achieving impressive performance without the need for GPU-intensive model training. We demonstrate that ensemble learning techniques can significantly enhance the zero-shot performance of LLMs.

We believe that our work highlights the potential of combining multiple LLMs to harness their collective intelligence for complex reasoning tasks. This approach is particularly valuable in resource-constrained environments, as it enables sophisticated inference without the need for extensive computational resources.

Looking ahead, we encourage the machine learning community to focus on developing methods that effectively integrate the strengths of diverse LLMs for inference tasks. Such research directions hold promise for expanding the applicability of advanced AI systems across a wide range of domains and computational settings.

References

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 30 (2017), 3149–3157.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. (2022).

Liudmila Ostroumova Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost - unbiased boosting with categorical features. 31 (2018), 6639–6649.

Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015. Identifying Meaningful Citations.

Da Yin, Weng Lam Tam, Ming Ding, and Jie Tang. 2021. MRT: Tracing the Evolution of Scientific Publications. 35 (2021), 711–724.

Hanwen Zha, Wenhu Chen, Keqian Li, and Xifeng Yan. 2019. Mining Algorithm Roadmap in Scientific Publications. (2019), 1083–1092.

Fanjin Zhang, Kun Cao, Yukuo Cen, Jifan Yu, Da Yin, and Jie Tang. 2024a. PST-Bench: Tracing and Benchmarking the Source of Publications. *arXiv preprint arXiv:2402.16009* (2024).

Fanjin Zhang, Shijie Shi, Yifan Zhu, Bo Chen, Yukuo Cen, Jifan Yu, Yelin Chen, Lulu Wang, Qingfei Zhao, Yuqing Cheng, et al. 2024b. OAG-Bench: A Human-Curated Benchmark for Academic Graph Mining. *arXiv preprint arXiv:2402.15810* (2024).

Zhi-Hua Zhou. 2021. Ensemble Learning. (2021), 181–210.

```
def get_prompt_gpt(text):
    return f'''**** I have a task to identify the source papers of a given paper, which
        author indicates that they inspire them most based on its text.
                **** I will now give you a text of academic papers, to find the most
                    pertinent source papers:
                Firstly, Determine the primary challenges outlined in the paper, and
                    understand the algorithm proposed by the author.
                Then, look for key phrases such as "inspired by", "motivated by", "inspired
                    us", "motivated us", "take inspiration", "the pioneering/previous work", "
                    following.. we adopt ... to solve the challenge/problem", "we use... based
                     on to achieve..." or other language that indicates a strong reliance on
                    previous research for developing the paper's core contributions.
                If key phrases exist, locate the key phrases in the text and find the sources
                     papers which are indicated by these key phrases.
                If key phrases do not exist or for other reasons, identify the novel methods
                    and approaches the paper introduces to tackle these challenges and locate
                    references that are directly linked to these main challenges and methods.
                Verify that your answer do not include the ref papers appearing at the
                    begining part of the text which describe the historical findings  like "
                    someone et al. proposed...", normally they are not direct related to the
                    paper's topic.
                Verify that the source papers are directly relevant to the paper's novel
                    contributions very directly.
                Specifically highlight any references that are preceded by phrases indicating
                     direct inspiration or motivation, such as 'Inspired by [reference]', and
                    make these references a priority in the list
                Please provide a concise list of source papers based on the aforementioned
                    criteria, ideally limiting the selection to the most central references
                    that heavily influenced the main contributions of the paper.
                **** Normally you should return less than 8 source papers. ****
                **** Please re-evaluate your result by the following metric:
                * Is the main idea of paper p inspired by the r e f e r e n c e
                * Is the core method of paper p derived from the r e f e r e n c e
                * Is the reference essential for paper p? Without the work of this reference,
                    paper p cannot be completed.
                Then, you should return your result in the json format, with the key is
                     reference   n u m b e r  and the value is   confidence   s c o r e  between 0
                    and 1.
                **** The text of the paper is:{text}'''
```

Figure 1: The basic prompt for generating the answer.