LLM-Based Iterative Hard Example Mining with Boosting for Academic Question Answering

Yang Zhou
MeiTuan
ShangHai, China
zhouyang96@meituan.com

Haoru Chen
MeiTuan
BeiJing, China
chenhaoru02@meituan.com

Xiaocheng Zhang
MeiTuan
BeiJing, China
zhangxiaocheng@meituan.com

Mengjiao Bao
MeiTuan
BeiJing, China
baomengjiao@meituan.com

Peng Yan[*]
MeiTuan
BeiJing, China
yanpeng04@meituan.com

Abstract

This paper describes the winning solutions of the KDD Cup 2024 Open Academic Graph Challenge (OAG-Challenge) from the BlackPearl Lab team. The challenge was to explore retrieval methods for academic resources, allowing us to answer specialized questions by retrieving relevant papers. This can provide researchers and the general public with high-quality, cutting-edge academic knowledge across various fields.

Our solution includes both recall and ranking processes, with a primary focus on two core ideas: "LLM for Vector" and "Iterative Hard Example Mining with Boosting". In the initial stages of text representation, similarity measures typically relied on autoencoder models, which proved suboptimal for this task. In contrast, vector representations derived from large language models (LLMs) have demonstrated superior performance in recent years, excelling in this specific task as well.

Furthermore, we identified the critical role of negative sample mining, particularly in contexts where "similarity does not necessarily imply correctness". The process of mining hard examples is essential for effective model learning, prompting us to introduce the "Iterative Hard Example Mining with Boosting" strategy. This approach incrementally recalls more challenging negative samples, ultimately integrating them to enhance overall performance.

Our method ranks 1rd place in the final leaderboard,code is publicly available at this link:https://github.com/BlackPearl-Lab/KddCup-2024-OAG-Challenge-1st-Solution.

CCS Concepts

• **Computing methodologies** → Natural language processing.

Keywords

Natural Language Processing, Question Answering Retrieval, Large Language Model, KDDCup 2024

ACM Reference Format:
Yang Zhou, Haoru Chen, Xiaocheng Zhang, Mengjiao Bao, and Peng Yan. 2024. LLM-Based Iterative Hard Example Mining with Boosting for Academic Question Answering. In *Proceedings of KDD 2024 Workshop OAG-Challenge Cup (KDDCup'24).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

1.1 Background

The primary objective of academic data mining is to enhance our comprehension of the dynamics of scientific progress, its intrinsic characteristics, and emerging trends. This field harbors the potential to yield substantial scientific, technological, and educational benefits. For example, in-depth mining of academic data can aid governments in the formulation of science policies, facilitate companies in talent identification, and enable researchers to acquire new knowledge with greater efficiency[1].

The field of academic data mining is rich with entity-centric applications, such as paper retrieval, expert discovery, and conference recommendation. However, the advancement of academic graph mining within the research community has been substantially hindered by the lack of suitable public benchmarks. To address this challenge, the KDD Cup 2024 has introduced the Open Academic Graph Challenge (OAG-Challenge). This initiative encompasses three realistic and challenging datasets, meticulously crafted to drive the latest innovations in academic graph mining.

This paper focuses on Task AQA (Academic Question Answering), where the objective is to retrieve the most relevant papers to answer given specialized questions from a set of candidate papers.

1.2 Dataset Description

The dataset for this competition is sourced from Open Academic Graph(OAG-QA)[6]. OAG-QA retrieves question posts from StackExchange and Zhihu, extracts the URLs of papers mentioned in the answers, and matches them with papers in the OAG. The dataset primarily consists of pairs of questions and related papers. The questions include the question itself and the body content, while

[*]Corresponding author of this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDDCup'24, Aug 25 2024, Barcelona, Spain

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

Table 1: Summary of Data Information

	Number	Avg. Query Length	Avg. Query Body Length	Number of Candidate Papers	Avg. Candidate Paper Text Length
Train	8757	9.25	176.31	395812	160.92
Test A	2919	9.91	148.18	395812	160.92
Test B	3000	11.56	103.19	466387	163.15

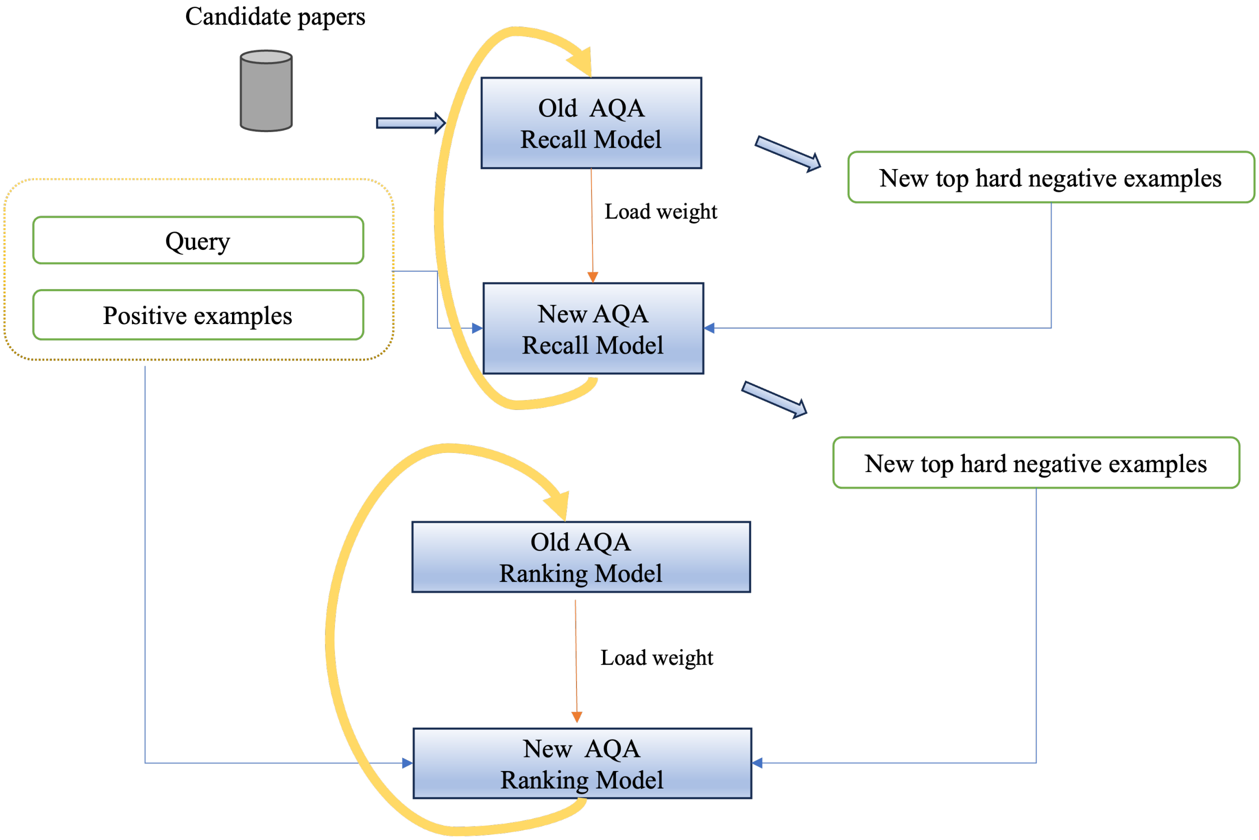

Figure 1: The framework of our proposed solution

the paper information includes the title and abstract. Additionally, a large-scale academic paper repository is provided.The main information of the dataset can be found in Table 1.

1.3 Task Description

The primary objective of this competition is to utilize question-paper pairs to train retrieval models that can effectively address specialized queries by retrieving relevant academic papers. The evaluation metric is the Top-20 Mean Average Precision (MAP@20).

2 Methodology

In the domain of Academic Question Answering (AQA), the challenge of effectively and efficiently retrieving relevant papers from a large-scale collection is considerable. Consequently, the architecture of large-scale question retrieval systems typically adopts a two-stage approach: an initial retrieval phase followed by a ranking phase. A critical challenge in this competition arises from the dataset's derivation from user queries on internet platforms, which inherently introduces a substantial amount of noise. For instance, users may exhibit varying cognitive standards when citing particular papers. Theoretically, a single question can correspond to multiple papers.

Upon employing open-source vector models for text retrieval and conducting a manual review of several recall cases, it was observed that the papers retrieved by these models exhibited a high degree of semantic similarity to the original text. Nevertheless, the correct answer was not invariably the one with the closest semantic match; in certain instances, the correct answer was ranked significantly lower. Furthermore, when fine-tuning an autoencoder-based vector retrieval model using contrastive learning, the model's performance was unexpectedly inferior to its performance without fine-tuning.

To address these challenges, our solution leverages an LLM-based framework for both retrieval and ranking. To further enhance performance, we introduce a hard example mining technique termed Iterative Hard Example Mining with Boosting. Ultimately, we integrate multiple models from various boosting steps to produce the final results. Figure 1 presents the workflow of our proposed solution, while the remainder of this section elaborates on the specifics of our approach in this competition.

2.1 Recall Model

In this competition, the candidate paper set for user queries consists of hundreds of thousands of papers. In real-world business scenarios, the candidate paper set is even larger. Therefore, to ensure efficiency and practicality, splitting the task into a recall and ranking process is a practical and effective approach.

Instruction fine-tuning has been widely applied to train large language models (LLMs) to follow instructions and perform retrieval-augmented generation. Recently, it has also been used to train retrievers and general embedding models (LLM for Vector), which can adjust their output embeddings based on different instructions and task types.

To develop a retriever that performs well in this competition, we further fine-tuned an open-source LLM vector retrieval model using retrieval instruction fine-tuning with the data from this task.

Given a relevant query-document pair, the instructed query follows the instruction template below:

> Given a web search query and a relevant body,retrieve the papers that are pertinent to the query. Query:xx

Let $D = \{\langle q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^- \rangle\}_{i=1}^m$ be the training data consisting of m instances. Each instance contains a query q_i and a relevant (positive) paper description p_i^+, as well as n irrelevant (negative) paper descriptions $p_{i,j}^-$.

We optimize the loss function[5] as the negative log-likelihood of the positive paper:

$$L(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-) = -\log \frac{e^{\mathrm{sim}(q_i,p_i^+)}}{e^{\mathrm{sim}(q_i,p_i^+)} + \sum_{j=1}^{n} e^{\mathrm{sim}(q_i,p_{i,j}^-)}} \quad (1)$$

During training, we adopt the QLoRA[2] training method to fine-tune SFR-Embedding-Mistral[4] with instructions, utilizing in-batch negative samples and hard negative sample mining methods to effectively improve the model's performance. Further improvements can be achieved through heuristic hard negative examples mining, which will be introduced in Section 2.3.

2.2 Ranking Model

Vector retrieval models effectively recall reliable potential positive samples from large-scale document sets by representing queries and documents as vectors, thereby reducing the size of the candidate set. However, since the representations of the query and document are input into the model independently without interaction, the model's comprehension ability is not fully utilized. Therefore, to further enhance performance, it is necessary to develop a ranking model.

We also use LLM as the base model, and the instruction template for ranking is:

> Instruct: Given a query with a relevant body, along with a title and abstract of a paper, determine whether the paper is pertinent to the query by providing a prediction of either 'Yes' or 'No'. Query: xx Paper: xx

During the training phase, the logits corresponding to the 'Yes' token at the final position are utilized as the ranking score. The loss function employed is cross-entropy, with negative samples derived from hard samples retrieved, maintaining a positive to negative ratio of 1:16. We adopted the QLoRA training methodology to fine-tune SOLAR 10.7B[3].

2.3 Iterative Hard Example Mining with Boosting

In the training of recall and ranking models, negative samples are of paramount importance. We have devised a hard negative sample mining methodology to further augment the model's performance.

We employ an iterative approach for hard negative sample mining, as illustrated in Figure 1. During the training of the recall model, in the initial iteration, we utilize the open-source SFR-Embedding-Mistral[4] model to retrieve the top n hard negative samples. These samples are then used for fine-tuning to obtain an enhanced model. In subsequent iterations, the improved model from the previous round is employed to retrieve the top n hard negative samples, followed by further fine-tuning. For the ranking model, we continue to utilize the hard negative samples retrieved by the recall model for training. This iterative process ensures that each round of training enhances the performance of the model from the previous iteration.

3 Experiment Results

In this section, we present our main results.

Table 3 presents the results of our model using the Iterative Hard Negative Example Mining with Boosting strategy. The base model weights are derived from SFR-Embedding-Mistral and SOLAR-10.7B-v1.0. Table 2 shows the main training hyperparameters used in each iteration.

Table 2: Hyperparameters used in each iteration

Method	LoRA parameters	Learning Rate
Recall Model	QLoRA, r=32, rank=64	1e-4
Ranking Model	QLoRA, r=32, rank=64	1e-4

Without fine-tuning, even advanced LLM-based models exhibit suboptimal performance. However, through the implementation of our fine-tuning strategy, we observed an initial improvement of 0.7 in the first iteration. As illustrated in the accompanying table, the MAP@20 score incrementally improves with each successive iteration. Upon reaching the sixth iteration, the rate of improvement began to decelerate. Ultimately, we conducted a total of eight iterations and applied rank averaging (Rank avg) to the final scores, culminating in a MAP@20 of 0.301.

Table 3: Our Main Result

Iteration	Recall Model MAP@20	Rank Model MAP@20
0	0.154	0.158
1	0.221	0.252
2	0.234	0.260
3	0.243	0.266
4	0.251	0.273
5	0.256	0.281
6	0.260	0.283
7	0.263	0.284
8	0.273	0.285
Rank avg ensemble	-	0.301

4 CONCLUSION

In this paper, we propose an LLM-Based Iterative Hard Example Mining with Boosting method for Academic Question Answering (AQA). This approach iteratively mines hard negative examples to boost the performance of the LLM-based model. Evaluation results demonstrate that this method effectively utilizes hard examples and continuously improves the model's performance through iterative processes. Our method secured the first place in Task 3 of the KDD CUP 2024 (we also achieved first place in all other tracks of the OAG-Challenge).

References

[1] Bo Chen, Jing Zhang, Fanjin Zhang, Tianyi Han, Yuqing Cheng, Xiaoyan Li, Yuxiao Dong, and Jie Tang. 2023. Web-scale academic name disambiguation: the WhoIsWho benchmark, leaderboard, and toolkit. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3817–3828.

[2] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. arXiv:2305.14314 [cs.LG] https://arxiv.org/abs/2305.14314

[3] Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2024. SOLAR 10.7B: Scaling Large Language Models with Simple yet Effective Depth Up-Scaling. arXiv:2312.15166 [cs.CL] https://arxiv.org/abs/2312.15166

[4] Shafiq Rayhan Joty Caiming Xiong Yingbo Zhou Semih Yavuz Rui Meng, Ye Liu. 2024. SFR-Embedding-Mistral:Enhance Text Retrieval with Transfer Learning. Salesforce AI Research Blog. https://blog.salesforceairesearch.com/sfr-embedded-mistral/

[5] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving Text Embeddings with Large Language Models. arXiv:2401.00368 [cs.CL] https://arxiv.org/abs/2401.00368

[6] Fanjin Zhang, Shijie Shi, Yifan Zhu, Bo Chen, Yukuo Cen, Jifan Yu, Yelin Chen, Lulu Wang, Qingfei Zhao, Yuqing Cheng, Tianyi Han, Yuwei An, Dan Zhang, Weng Lam Tam, Kun Cao, Yunhe Pang, Xinyu Guan, Huihui Yuan, Jian Song, Xiaoyan Li, Yuxiao Dong, and Jie Tang. 2024. OAG-Bench: A Human-Curated Benchmark for Academic Graph Mining. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.