CSE587 Spring 2025 Final Project
Enhancing Scientific Understanding in LLMs via Research Questions and Approaches

Jiamu Bai Ryo Kamoi Divya Navuluri

GitHub Repository: `https://github.com/ryokamoi/cse587-final`

1 Introduction

While large language models (LLMs) have shown remarkable success in various natural language generation tasks, their performance in structured scientific reasoning remains less explored [2].

This project investigates the ability of LLMs to propose novel and feasible methodological approaches to specific research questions. We take an approach to create a training dataset by extracting research questions and approaches from abstracts of academic papers.

We designed and implemented a two-stage pipeline:

- Dataset Construction: We automatically constructed a high-quality dataset containing (research question, approach) tuples. This was achieved by designing a few-shot prompting template and using Llama 3.3 70B [1] to extract research questions and approaches from over 10,000 paper abstracts drawn from leading NLP conferences (ACL, EMNLP, NAACL, AACL, ARABICNLP, CL, CONLL, EACL, FINDINGS, IWSLT, SEMEVAL, SEM, TACL, WMT and WS). The resulting dataset captures diverse scientific problem statements and solution strategies.
- LLM Fine-tuning and Evaluation: We fine-tuned the base model Llama 3.1 8B [1] using the constructed dataset, with the goal of improving its ability to generate accurate and relevant research approaches when given only the extracted research question. We compared pre-trained and fine-tuned models in terms of output similarity and quality to quantify improvements.

Our experiments show that few-shot prompting tends to produce research approaches with higher novelty, often suggesting creative but complex solutions. In contrast, approaches extracted directly from abstracts demonstrate better feasibility, as they reflect realistic research practices. These findings highlight the trade-off between novelty and feasibility in LLM-generated scientific reasoning.

2 Dataset Collection

We constructed a new dataset of scientific abstracts annotated with their corresponding research questions and approaches. The dataset construction involved the following steps:

- Extract paper abstracts from published papers
- Manually create several few-shot examples of (abstract, research question, approach) triples
- Use Llama 3.3 70B to extract research questions and approaches through few-shot learning

2.1 Data Sources

We collected abstracts from papers published at ACL, EMNLP, NAACL, AACL, ARABICNLP, CL, CONLL, EACL, FINDINGS, IWSLT, SEMEVAL, SEM, TACL, WMT, and WS between 2020 and 2024. We acquired the abstracts from ACL Anthology.[1]

2.2 Dataset size and split

The whole dataset contains the training and test splits in the following way. To avoid data contamination, our test split includes newer papers.

- Training set: 11,728 abstracts, extracted from papers published from 2020 to 2023.
- Test set: 196 abstracts, extracted from papers published in 2024. In our experiments, we only use 20 cases for human annotation.

2.3 Few-shot Prompting for Extracting Research Questions and Approaches from Abstracts

Few-shot examples. We manually created a few-shot prompt with eight examples that include the following information. The binary label, research question, and approach are annotated by a human.

- Abstract
- Does this paper propose a new method? (yes or no)
- Research question
- Approach

The binary label for "Does this paper propose a new method?" is used to filter out papers that do not propose any methods, such as evaluation and survey works. In this project, we focus on research questions about creating new methods.

These examples are used to prompt the `meta-llama/Llama-3.3-70B-Instruct` model [1] to extract structured outputs from each abstract automatically. An example of human annotated example is illustrated in Table 1.

Abstract	*"Text written by humans makes up the vast majority of the data used to pretrain and finetune large language models (LLMs). Many sources of this data – like code, forum posts, personal websites, and books– are easily attributed to one or a few "users". In this paper, we ask if it is possible to infer if any of a user's data was used to train an LLM..."*
New method?	Yes
Research Question	research question: How can we prevent language models from leaking whether a specific user's data was used during training?
Approach	We can reduce the risk of user inference in language models by modifying both the training data and the training procedure to the unknown user specific signals. One main approach is to apply example level privacy preserving techniques such as differential privacy which will limit the model's ability to memorize patterns that are tied to individual users...

Table 1: Illustration of human annotated few-shot examples

Dataset. Each record in our dataset includes:

- Research Question (LLM-generated)
- Research Approach (LLM-generated)
- Abstract (metadata, not used in training or evaluation)

[1] `https://github.com/acl-org/acl-anthology`

3 Experiment Setup

We fine-tuned a large language model using our custom dataset to improve its capability in generating research approaches from scientific abstracts.

3.1 Models

We evaluate the performance of three models on the research approach generation. As a baseline, we compare the fine-tuned model against two models with few-shot prompting.

- Llama 3.1 8B fine-tuned on our training data (ours)
- Llama 3.1 8B with few-shot prompting (baseline)

3.2 Training Strategy

The model was fine-tuned on the research question and approach pairs. The target of the training is to help LLM generate reasonable research approaches given questions. We use LlamaFactory [3] for training.

Hyperparameters We train the model for one epoch with a learning rate of 1e-5 with the AdamW optimizer. We apply a linear annealing of the learning rate with the warm-up ratio of 0.5. We use a batch size of 16.

3.3 Evaluation

We evaluated the impact of fine-tuning on approach generation using human annotation. We select a small subset from the model output with the number of 60 annotations and manually review whether the generated research approaches from finetuned model are improved compared with those generated from few-shot learning.

Specifically, we compare the following pairs of generated approaches.

- Fine-tuned Llama 3.1 8B (ours) vs. Llama 3.1 8B with few-shot prompting
- Fine-tuned Llama 3.1 8B (ours) vs. approaches extracted from abstracts.
- Llama 3.1 8B with few-shot prompting vs. approaches extracted from abstracts.

4 Results and Analysis

Methods		Better Novelty (%)			Better Feasibility (%)		
Method 1	Method 2	Method 1	Tie	Method 2	Method 1	Tie	Method 2
Fine-tuned	Few-shot	5	5	90	30	25	45
Extracted	Few-shot	5	20	75	55	45	0
Fine-tuned	Extracted	10	65	25	15	50	35

Table 2: Pairwise comparison of generated research approaches based on novelty and feasibility

Table 2 presents a pairwise comparison of the three methods: 1. few-shot prompting with Llama 3.1 8B, 2. a fine-tuned version of Llama 3.1 8B trained on our custom dataset, and 3. approaches extracted directly from paper abstracts using few-shot prompting with Llama 3.3 70B. Each row reflects the percentage of times one method was judged better than another in terms of novelty or feasibility, based on human annotations.

From Table 2, we observe the following:

- The few-shot prompting result shows higher novelty than both fine-tuned results and extracted results.
- The extracted result generally has better feasibility.

4.1 Analysis

Few-shot prompting model shows highest novelty in general. Few-shot prompting often results in overly complex approaches involving ideas like pretraining or using external knowledge graphs. While these methods boost perceived novelty, they are often impractical and difficult to implement, reducing their feasibility. In contrast, the fine-tuned model tends to produce more grounded but less innovative responses.

The extracted result generally has better feasibility. The extracted results generally have better feasibility because they are originally written by paper authors and reflect realistic, well-scoped research plans. These approaches are usually grounded in practical methods that are more likely to be implemented successfully. In contrast, LLM-generated outputs often aim for novelty but may overlook feasibility, suggesting ideas that sound impressive but are hard to execute.

5 Conclusion

In this project, we explored the ability of large language models to generate structured scientific insights by constructing a dataset of (research question, approach) pairs extracted from paper abstracts. We fine-tuned a Llama 3.1 8B model on this dataset and evaluated its ability to generate research approaches given only the research question. Our results show that while fine-tuning improves feasibility over few-shot prompting, it still lags behind the quality of approaches extracted directly from abstracts. This highlights both the potential and limitations of current LLMs in scientific reasoning tasks, suggesting that integrating structured extraction with targeted fine-tuning may be a promising direction for future work.

References

[1] Llama Team, AI @ Meta. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[2] Renze Lou, Hanzi Xu, Sijia Wang, Jiangshu Du, Ryo Kamoi, Xiaoxin Lu, Jian Xie, Yuxuan Sun, Yusen Zhang, Jihyun Janice Ahn, et al. Aaar-1.0: Assessing ai's potential to assist research. *arXiv preprint arXiv:2410.22394*, 2024.

[3] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang 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, 2024. Association for Computational Linguistics.