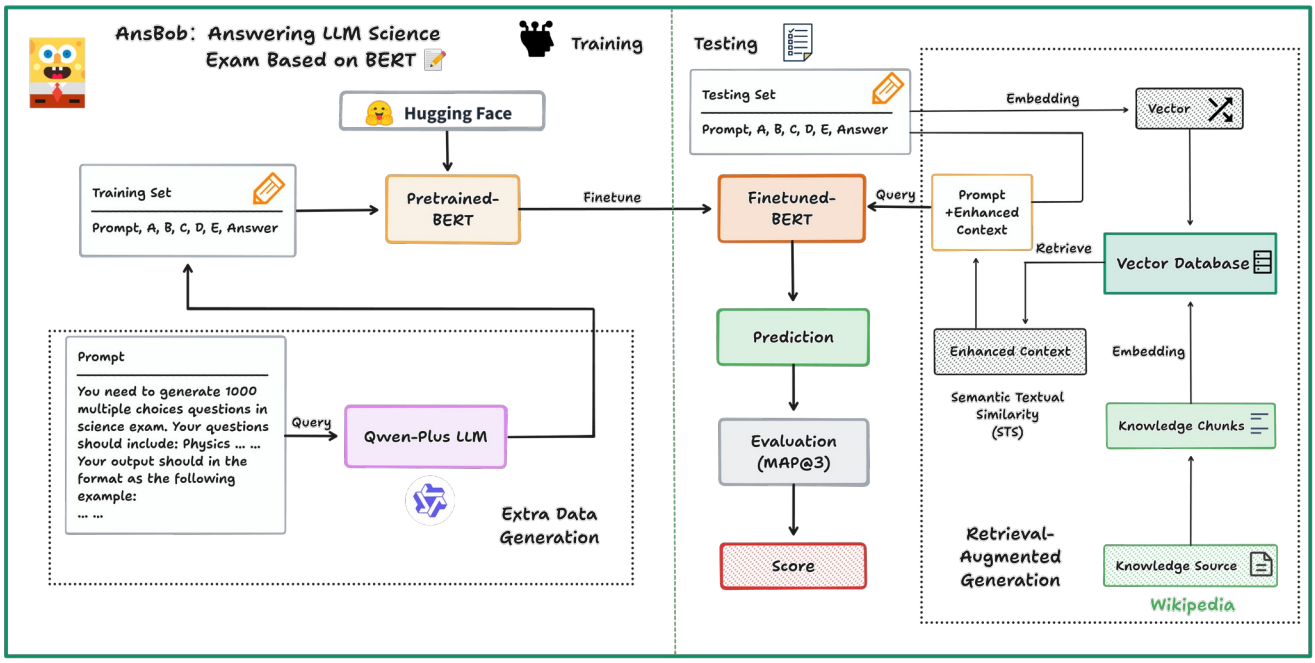
AnsBoB : Answering Science Exam Based on BERT Model



Danbo WANG
Department of Control Science and Engineering, Zhejiang University
Major : *Automation(Control)*

Xinyue YAO
Department of Control Science and Engineering, Zhejiang University
Major : *Robotics Engineering*

Hongze WANG
Department of Control Science and Engineering, Zhejiang University
Major : *Automation(Control)*

Xinyu YU
Department of Computer Science and Engineering, Zhejiang University
Major : *Computer Science and Engineering*

Yutao WANG
Department of Control Science and Engineering, Zhejiang University
Major : *Automation(Control)*

***Abstract*—This report address a problem of answering LLM science exam (generated by ChatGPT) using language model. We propose a novel approach to train and prompt BERT model on a 6k training set and test it on a 500 questions set. Experiment results show that our method has strong performance on the science exam, but still has room for improvement.**

***Keywords*—Natural Language Processing, BERT, Language Models**

I. Introduction

As the scope of large language model capabilities expands, a growing area of research is using LLMs to characterize themselves. Because many preexisting NLP benchmarks have been shown to be trivial for state-of-the-art models, there has also been interesting work showing that LLMs can be used to create more challenging tasks to test ever more powerful models [1]. To answer those tasks, we proposed our model AnsBoB (**Ans**wering science exam **B**ased **o**n **B**ert).

As one of the most popular deep learning based language models, BERT has applications in multiple fields such text classification, linguistic models and chatbot etc. Meanwhile, with the expand of Large Language Model, a novel method to enhance LLM's inference ability: Retrieval Augmented Generation (RAG) has been proposed by et al[2], which matches well with our project. To this end, we have explored and have leveraged these techniques in our AnsBoB model.

In conclusion, our project has three main contributions:

(1) We leveraged Large Language Model to generate and process training data. (2) We fine-tuned BERT model and implement RAG to enhance its capability. (3) We evaluate our result on different training methods.

II. Background

A. LLM Science Exam

The Kaggle LLM Science Exam Competition is an innovative initiative designed to evaluate and advance the capabilities of large language models (LLMs) in the domain of scientific reasoning and knowledge application. Launched on the Kaggle platform, this competition aims to challenge

Equal contribution for each member.

participants to develop LLMs that can effectively tackle highschool-level science-related questions, mirroring the rigor and depth of traditional academic assessments.

B. BERT

Bidirectional Encoder Representations from Transformers (BERT) is a groundbreaking model introduced by Devlin et al. [3] in 2018 that has significantly advanced the field of natural language processing (NLP). BERT is built on the Transformer architecture, which relies on self-attention mechanisms to process text efficiently. Unlike previous models that utilized unidirectional context (processing text in a left-to-right or right-to-left manner), BERT employs a bidirectional approach, allowing it to capture the full context of a word based on its surrounding words.

BERT is pre-trained on a massive corpus of text using two primary tasks: masked language modeling (MLM) and next sentence prediction (NSP). In MLM, random words in a sentence are masked and the model learns to predict them based on context, while NSP involves predicting whether one sentence logically follows another. This pre-training enables BERT to generate rich contextual embeddings that can be fine-tuned for various downstream tasks, such as question answering, sentiment analysis, and named entity recognition.

III. METHOD

As we have mentioned in section I, AnsBoB has three main modules. (i) a module that generate training data (ii) a training module that fine-tune BERT (iii) RAG module. In this section, we will introduce this three modules in detail.

A. Data Generation from LLM

In this module, we employed the Qwen-Plus API to generate multiple-choice questions (MCQs) tailored for scientific content. The process was designed to ensure that the questions reflect a comprehensive understanding of key concepts across various scientific disciplines while maintaining a rigorous standard for assessment.

The prompt we designed icould be found at Appendix.

B. Fine-tune BERT model

We fine-tuned our BERT model with following steps:

- Preprocessing training data, which includes tokenization using BERT's specialized tokenizer
- Obtain pre-trained model from huggingface. Here we use bert-base-cased model.
- We set our batch size as 4, learning rate as 0.00005, weight decay as 0.01, and training the model with 5 epochs.

C. Retrieval Augmented Generation

Retrieval-Augmented Generation (RAG) is an innovative approach that synergizes retrieval mechanisms with generative models to enhance the performance of natural language processing tasks. The concept of knowledge library fits our need well and thus we implement it in our model.

We leveraged knowledge from wikipedia and made Semantic Textual Similarity(STS) matches between the embeddings in vector database and each choice. Combined with the original predictions score from fine-tuned BERT model, we can receive 3 top possible answer from AnsBoB.

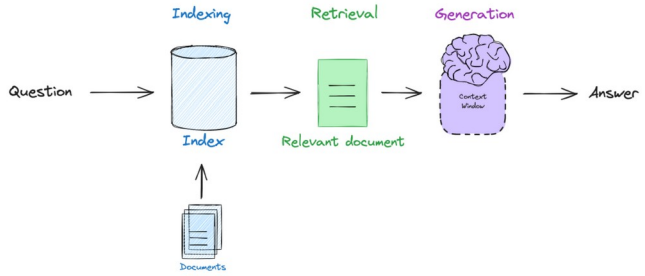


Figure 1. shows how a RAG module works to enhance LLM.

IV. EXPERIMENT

We train the model on a PC with a single 1080ti GPU. We evaluate our result with Mean Average Point(MAP) method and do a ablation experiment to verify our work.

Referring to the competition requirement, for each question, our model need to predict top 3 possible answer and value it. To this end, the MAP should be adapted.

$$MAP@3=\frac{1}{U}\sum_{u=1}^{U}\sum_{k=1}^{min(n,3)}P(k)\times rel(k)$$

where U is the number of questions in the test set, P(k) is the precision at cutoff k, n is the number predictions per question, and rel(k) is an indicator function equaling 1 if the item at rank k is a relevant (correct) label, zero otherwise.

A. Training on different data sets with RAG

TABLE I. RESULT WITHOUT RAG

Scores in different training set with RAG		
6000	***500***	***0***
0.4925..	0.5408	0.3543..

Each score is a mean value from 3 independent training

B. Training on different data sets without RAG

TABLE II. RESULT WITHOUT RAG

Scores in different training set without RAG		
6000	***500***	***0***
0.4925	0.5408..	0.3543

Each score is a mean value from 3 independent training

C. Ablation experiment

TABLE III. RESULT OF ABLATION

RAG+6k training	***6k training***	***Pre-trained model***
0.4925	0.4453	0.3543

Each score is a mean value from 3 independent training

V. CONCLUSION

In conclusion, our project has explore the possibility of Large Language Model in answering science exam. Firstly, we effectively utilized LLMs to generate and process high-quality training data, enriching the learning experience of our model. Secondly, we fine-tuned the BERT architecture and integrated the Retrieval-Augmented Generation (RAG) methodology to enhance the model's inference capabilities, enabling it to tackle complex scientific examination tasks with greater efficacy. Finally, we conducted comprehensive evaluations across various training methods, providing insights into the effectiveness of our AnsBoB model. Collectively, these contributions not only advance our understanding of LLM capabilities but also pave the way for future research aimed at developing more robust and challenging projects in the field of NLP.

ACKNOWLEDGMENT

Many thanks to our 2 prof. Wang, TA and our group members!

REFERENCES

[1] *Will Lifferth, Walter Reade, and Addison Howard. Kaggle - LLM Science Exam. https://kaggle.com/competitions/kaggle-llm-science-exam, 2023. Kaggle.*

[2] Patrick Lewis, Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, NeurIPS 2020

[3] Jacob Delvin, Ming-Wei Chang and Kenton Lee, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, https://doi.org/10.48550/arXiv.1810.04805

[4] https://www.kaggle.com/code/andersonhklein/ucsc-nlp-final-project/input

[5] https://www.kaggle.com/datasets/radek1/additional-train-data-for-llm-science-exam