CS6207 Project Proposal

Zuo Xinyue (A0194561Y), Zhang Yifan (A0289527L), Liu Chenyan (A0232487Y)

1 INTRODUCTION

Although Large Language Models (LLMs) have garnered considerable attention, they have been criticised for hallucination and static knowledge[2], both of which will compromise the credibility of the model. To overcome this problem, integrating external knowledge into the generation has been proven effective towards generating high-fidelity and up-to-date answers.

2 PROJECT DESCRIPTION

2.1 Project Description & Target

The primary objective of this project is to integrate an external text knowledge database into a question-answering LLM. Instead of solely generating responses in an auto-regressive manner when presented with questions, the LLM is designed to dynamically retrieve up-to-date and real-world factual knowledge from an external text database and incorporate it into the final output.

2.2 Input & Output

The system consists of a LLM and an external database. The input consists of common-sense questions from the real world in text format. The output should be answers in text format, combining the effort of the LLM and knowledge from the external database.

3 APPROACH

3.1 Approach

In our approach, we first focus on generating effective embedding for query inputs and information stored in the external database. Next, LLM will perform query processing, which includes understanding the query, analysing the context, and reformulating the query for better information retrieval if necessary. After that, the system will perform database querying and information retrieval. In this step, LLM will retrieve the information with the highest matching score. The next step is to integrate the information into the generation of response by LLM. Lastly, the generated response will be post-processed and optimized.

3.2 Novelty

We base our project on the work from Facebook [4]. We will focus on improving retrieval efficiency by adopting clustering algorithms to reduce redundancy within the external database and promote retrieval efficiency.

3.3 Dataset

We use the publicly available dataset - Stanford Question Answering Dataset (SQuAD) in this project. SQuAD, a reading comprehension dataset, consists of questions posed by crowd workers on Wikipedia articles, where the answers are segments of text or spans. For the external knowledge database, due to the enormous size of the Wikipedia database and constrained computing resources, we construct a subset of Wikipedia on our own.

4 RALATED WORK

Knowledge-intensive tasks for LLMs have been a very popular research topic. There's a surge in research focused on fine-tuning LLMs for specific domains such as legal [1], medical [5] and financial fields [6]. Several common practices include training models on specialized datasets and integrating structured knowledge into question-answering systems [3].

There are different forms of external databases, for example, [7] focus on integrating LLMs with relational databases, and [4] focus on text databases, which are more related to our project and will serve as the basis for our project code.

5 TASK ASSIGNMENT

Member	Task
Liu Chenyan	Crawl database, Report, Clustering Algo Implementation
Zhang Yifan	RAG (dataset, LLM+DB), Research/Implement Optimization, Report
Zuo Xinyue	RAG (dataset, LLM+DB), Research/Implement Optimization, Report

Table 1: Task assignment

6 SCHEDULE

Date	Tasks
2.5 - 2.26	Database Preparation, Baseline Understanding, Implementation
2.26 - 3.4	Research Optimization Techniques (Implement)
3.4 - 3.25	Implement Optimization (Clustering)
3.25 - 4.8	Report & Presentation

Table 2: Schedule

REFERENCES

[1] Fei, Z., Shen, X., Zhu, D., Zhou, F., Han, Z., Zhang, S., Chen, K., Shen, Z., and Ge, J. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289* (2023).

[2] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232* (2023).

[3] Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906* (2020).

[4] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems 33* (2020), 9459–9474.

[5] Pal, S., Bhattacharya, M., Lee, S.-S., and Chakraborty, C. A domain-specific next-generation large language model (llm) or chatgpt is required for biomedical engineering and research. *Annals of Biomedical Engineering* (2023), 1–4.

[6] Zhang, L., Cai, W., Liu, Z., Yang, Z., Dai, W., Liao, Y., Qin, Q., Li, Y., Liu, X., Liu, Z., et al. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. *arXiv preprint arXiv:2308.09975* (2023).

[7] Zhou, X., Sun, Z., and Li, G. Db-gpt: Large language model meets database. *Data Science and Engineering* (2024), 1–10.