A Chat Bot for Enrollment of Xi 'an Jiaotong-Liverpool University Based on RAG*

1st Liwei.Xu*
College of Intelligent Engineering
Xi 'an Jiaotong-Liverpool University
China, Suzhou
Liwei.Xu22@student.xjtlu.edu.cn

2nd Jiarui.Liu*
College of Intelligent Engineering
Xi 'an Jiaotong-Liverpool University
China, Suzhou
Jiarui.Liu22@student.xjtlu.edu.cn

2024 8th International Workshop on Control Engineering and Advanced Algorithms (IWCEAA) | 979-8-3315-2888-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/IWCEAA63616.2024.10823979

***Abstract*—Recent advancements in large language models (LLMs) have established pre-training on extensive textual corpora as a foundational methodology. However, in specialised applications such as admissions systems, the focus shifts from general knowledge-based reasoning to ensuring accuracy and relevance in domain-specific responses. This study presents the development of an automated admissions system for Xi'an Jiaotong-Liverpool University, leveraging GLM-4 in conjunction with Retrieval-Augmented Generation (RAG) to handle targeted queries. The implementation of RAG mitigates the occurrence of hallucinations often seen in LLM outputs, thereby enhancing the reliability and alignment of generated responses with real-world data, which is critical for prospective students and their parents. This paper details the construction of the RAG corpus and cue word methodology, and provides an empirical comparison of the efficacy of various major language models with and without RAG integration. The results demonstrate the potential of RAG to significantly improveresponse accuracy in domain-specific tasks, and suggest directions for future research in optimising LLMs for admissions processes.**

Keywords- Chatbots, Large Language Model (LLM), Retriever_Augmented Generation (RAG), Prompt Engineering

I. INTRODUCTION

In the context of the recent advancements in the field of large language models, the pre-training of these models through the utilization of extensive textual data corpora has become a standard methodology [1]. In the context of large models deployed in specialised scenarios, such as question-and-answer processes within the field of jurisprudence, the role of knowledge-based reasoning appears to be less crucial. Instead, the primary objective is to ensure the accuracy of the responses generated based on the available knowledge base, which becomes particularly pivotal in certain domains.

The primary focus of this study is the development of an automated admissions system for Xi'an Jiaotong-Liverpool University, based on GLM-4, and utilizing RAG to address specific admissions queries.

In the case of enrolment systems, it is the students and their parents who constitute the audience. Consequently, it is they who are influenced by the content policy that governs enrolment information. In the context of admissions, it is of paramount importance to ensure the accuracy and completeness of the information provided. As LLM develops, the emergence of hallucinations can result in the generation of content that appears to be plausible but is, in fact, incongruous with the real world [2]. The use of RAG serves to significantly diminish the prevalence of hallucinations and align the generated content more closely with the actual world.

This paper provides a detailed account of the genesis and constitution of the RAG corpus, together with an exposition of the methodology employed in the construction of cue words. The paper demonstrates and contrasts the impact and efficacy of alternative major language models with and without RAG.

This paper will commence with an introduction to the large language model and RAG, after which it will present the methodology employed by our RAG. It will then proceed to evaluate and compare the experiments conducted by the enrolled robots, and to explore the findings and future research directions.

II. RELATED WORK

A. LLM

The field of language modelling has undergone a significant evolution, commencing with the advent of SLM (Statistical Language Modelling), which was developed based on statis- tical learning methods [3]. The fundamental premise is the Markovian hypothesis, which serves as the foundation for developing word prediction models. For example, it predict the subsequent word based on the most recent context. The advent of neural networks has driven the development of language models, with the emergence of neurolinguistic models designed to predict the probability of word sequences. The concept of distributed representations of words has been introduced, and word prediction functions have been constructed under the condition of aggregating contextual features (i.e., distributed word vectors) [4]. Subsequently, the pre-trained language model is based on a highly parallelised Transformer architecture with a self-attention mechanism, which markedly enhances the performance of natural language processing tasks [5]. Large language models are analogous to large pre-trained language models. It has been demonstrated that the greater the number of parameters, the more LLMs can demonstrate remarkable performance. At present, GPT-4, which was developed based on LLM, is capable of performing complex tasks with a high degree of proficiency. Furthermore, its development into a multimodal model has significantly advanced the field of artificial intelligence. While the development of LLMs has resulted in notable advancements in text comprehension and content generation, these systems can also manifest hallucinatory tendencies, leading to generated content

that deviates from the actual world or does not align with user input [6]. This raises concerns about the reliability of these systems. In order to address these issues, this paper employs the use of RAG.

B. RAG

In some cases, the reasoning employed by LLM is not a significant factor in determining the outcome of a particular scenario. It is, however, the accuracy of the information provided that is of greater consequence. Retrieval-augmented generation is a technique that employs information from private or proprietary data sources to facilitate text gen- eration. It combines a retrieval model, designed for searching large datasets or knowledge bases, with a generative model, such as a large language model, which uses retrieved information to generate readable text responses.

Heydar Soudani, E. Kanoulas and Faegheh Hasibi [7] conducted an analysis of the efficacy of two methods, namely Relational Extraction and Graph Construction (RAG) as well as Fine-Tuning (FT), in enhancing the capabilities of Large Language Models (LLMs) to address low-frequency entity issues. The findings of the study demonstrate that, while the fine-tuning approach yields notable performance improvements in entity recognition across varying levels of popularity, the RAG method surpasses the other comparative approaches in terms of performance. Moreover, with the ongoing development of retrieval and data enhancement techniques, the capacity of the RAG and FT methods to adapt LLM for addressing low-frequency entities has been markedly enhanced.

The construction of a RAG can be approached in a number of ways, with the subsequent sections of this article providing a detailed account of the process.

III. METHODOLOGY

The objective of our research is to harness the power of RAG and other search tools to enhance the model's performance when answering the admission. To achieve this, a systematic methodology (Fig. 1) was drawn up and is detailed below:

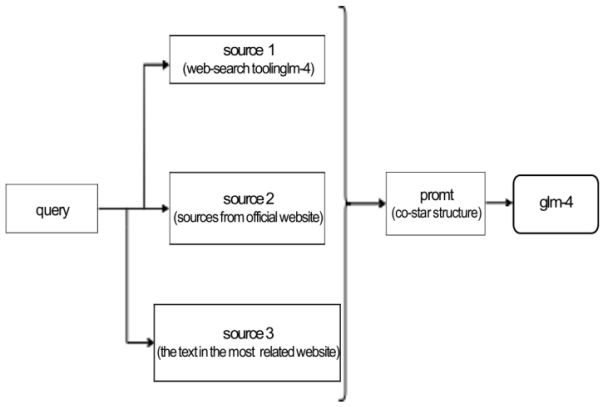

Fig. 1. System architecture of Overall framework

A. Data collected and Preprocessing

To construct the required dataset, the Web crawler technology is utilized to extract all text data from the official website of XJTLU. However, the text form of information is hard for further processing. Thus, the glm-4 was harnessed to summarize the whole text with several short sentences with merely one subject, predicate, and object, whose structure is compact. In contrast, those sentences could reconstruct the text's original meaning without loss. Those sentences were the original form of the features of each text. Furthermore, by using the bge-large-zh-v1.5, a powerful vector embedding tool for Chinese texts [8], each sentence was transformed into a vector, which is an array and stored in a file after adding an index through faiss [9], which is a library for efficient similarity. Meanwhile, a dictionary was also formed where the key stored the index of the text while the value contained the index of the vectors of sentences for the text.

B. Configuring the glm-4 Model

The next phase involves configuring the glm-4 model and integrating it with the RAG and search tools by the prompt constructed in the co-star structure [10]. This process starts with setting up and configuring the environment for glm-4 and setting hyperparameters that the temperature is 0.01 and the t-top is 0.7. Furthermore, the source selected from the dataset by RAG, the other two sources searched from Baidu, the most widely-used Chinese search engine, by different search tools were integrated, and the query into the prompt using the CO-STAR structure. To ensure the control variable, all the models tested in this research harness a similar prompt, which only has differences in the input sources.

C. Retrieval-Augmented Generation (RAG) Implementation

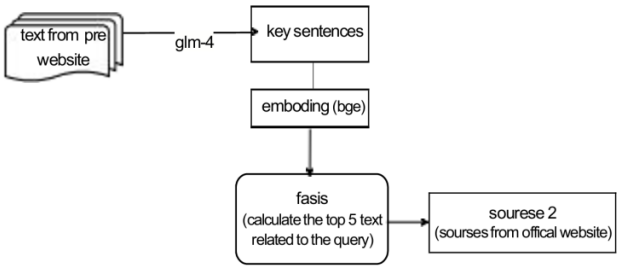

Fig. 2. Architecture of Retrieval-Augmented Generation Implementation

To implement the RAG component effectively, the dataset of vectors and the preconstructed dictionary were utilized in this phase and the setup involved several key steps as the Fig 2 illustrated: 1) Vector Embedding: Transforming the input query from the string into the vector by the bge-large-zh- v1.5, which is the processible form for the faiss. 2) Source searching: Through the faiss, four vectors with the highest cosine similarity to the vector of the query would be selected by the faiss. 3) Source generation: Utilizing the indexes of those four vectors, four indexes of text would be found in the dictionary and then those texts would be added into the prompt as part of the input the glm-4.

D. Implementation of search tools

The source from the search tools would provided in the prompt as the extended materials. Generally, there were two different search tools. One is the original web-searching tool in the model,while the other is an online search tool designed for

the experiments. Both of them would acquire the extended sources from the search engine.

1) web-search tool: The web search tool is contained in the model, the result of which would become part of the context of the model automatically. To exploit its ability, the hyperparameters of the tool were set to search the input query on a specific search engine, which is Baidu. Furthermore, the tool would generate a piece of context, whose size is 1000 tokens.

2) Online search tool: To provide more reliable sources and perform the contrast experiment, the online search tool is designed to select the most related web page link, which is one of the search results of the query on the search engine, based on the rank of cosine similarity of the embedding of the brief introduction on the link and extract the text from the web page opened by the link as the source it provides.

E. Texting queries

A set of queries, which contain 128 different questions related to the possible question for admission, is constructed for the model to answer, while the standard answers of them were formed based on the information of the official website of XJTLU.

F. LLM based judgment

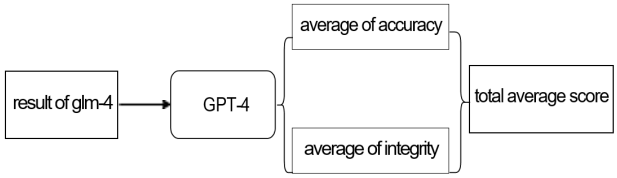

Fig. 3. Framework of evaluation system

The GPT-4 is utilized for judgment through the prompt as Fig 3 displayed, which was constructed in the CO-STAR structure. To evaluate the responses of the model, two criteria, which is accuracy and integrity, were emphasized in the prompt. The criterion of integrity refers to how much current information is provided in the response, while the criterion of accuracy depends on how current the response is and no wrong answer is in the response. Additionally, to avoid the vias of the LLM judgment such as Verbosity bias [11], some recommend ideas against the systemic bias likes Multiple Evidence Calibration(MEC) [12], which refers to require the GPT provided the explain of its score, and other recommend a way to form the prompt, which means to highlight the length and position cannot affect the score.

IV. EVALUATION

A. Vertical contrast

The vertical contrasts aim to compare the performance of different models on the queries. A large range of models, which include Deepseek, Ernie-Bot 4.0 from Baidu, Gemini-pro from Google, glm-4 with all sources of text, Spark Max, yi-34b-chat, and Qianwen, are tested and marked by GPT- 4 under the criteria of integrity and accuracy presented in the Table I. With the contribution of rag and other sources of information provided in the input prompt, The model we designed has the highest average score in accuracy, which increases by 0.33 compared with the original model and is larger by around 0.14 scores than the second one, while the average Integrity score is the second highest and is merely lower by 0.01 compared with the first, which has a remarkable increase by more than 0.5 scores. In order to eliminate the possibility of chance, three experiments were conducted, the scores were calculated, and the resulting error was determined. The mean scores and errors are presented in the table. As can be observed, since no fine-tuning of the model or alteration of its parameters was conducted, and the temperature was set to 0.01, which is the smallest one in our experiments. This substantiates the reliability of our experimental conclusions. The same methodology was employed for the subsequent level comparisons. Thus, the glm-4 model with all the sources of text could be considered to have a better performance compared with the other models in the queries. According to this result, utilizing the Rag and other sources of text could significantly enhance the accuracy and integrity of the output result in the question about the university although affected by the faculty of the module the enhancement failed to raise to the highest one in integrity however their score is adjacent indicated that they are in the same level.

TABLE I THE SCORES OF MODULES FOR VERTICAL CONTRAST

model	Integrity	Accuracy	average
deepseek	3.1880 ± 0.0052	3.3756 ± 0.0067	3.2648 ± 0.0200
ERNIE-Bot 4.0	2.9492 ± 0.0054	2.9673 ± 0.0046	2.9324 ± 0.0057
gemini-pro	2.6758 ± 0.0258	2.5868 ± 0.0025	2.8459 ± 0.0152
Qianfen	3.6548 ± 0.0784	3.4765 ± 0.1540	3.4922 ± 0.1201
glm-none	3.0257 ± 0.0254	3.4325 ± 0.1227	3.2567 ± 0.1589
Spark Max	3.3489 ± 0.0556	3.4798 ± 0.1258	3.4102 ± 0.0241
yi-34b-chat-0205	3.4459 ± 0.0230	3.5625 ± 0.1054	3.5058 ± 0.0677
glm-all	3.7267 ± 0.0157	3.7567 ± 0.0030	3.6780 ± 0.0010

B. Horizontal contrast

With the target of exploiting the contribution of differ- ent sources and optimizing the performance, the individual sources, which are the source from RAG, the source from the web-searching tool, and the source from the online searching, and their combinations are tested and evaluated in horizontal contrast. As Table II presented, the sources from Rag and the online searching of queries could enhance the ability of the model in the two criteria, the first of which increased the score of integrity by around 0.3 and more than 0.2 in accuracy, while the latter augment its score of integrity by 0.2 and the same in accuracy. The reason why they could enhance the performance of the model would be the correlated sources they provided. Meanwhile, by adding the source from the web- searching tool the model contains, the model's performance does not improve obviously, which barely enhances the score of integrity slightly

while the score in accuracy dropped by 0.2. However, The combination of sources from Rag and the web-searching tool gained the highest score compared with others and the combination with all the sources is at the same rank though it is slightly lower in the score, the score of which is raised may because more texts that were highly correlated to the query were included as the input. Additionally, maybe because the larger proportion of sources from the Internet were included in the combinations of all the sources it increased the propensity of data pollution and affected its performance. This could be proven by one specific query, the fee of undergraduate level, that all the models that contain the sources from online searching tools responded with a wrong answer while the other models did not.

TABLE II THE ACCURACY OF MODEL WITH FOUR CLUSTERS

source	integrity	accuracy	average
rag,web-searching tool	3.5637± 0.0155	3.7564± 0.0281	3.6032± 0.1340
online searching	3.2354± 0.0256	3.6664± 0.1005	3.4367± 0.1057
rag	3.3357± 0.0015	3.6257± 0.0357	3.5108± 0.0058
web-searching tool	3.0219± 0.0009	3.2934± 0.0047	3.1699± 0.0004
None	3.1527± 0.0017	3.4587± 0.0251	3.2267± 0.0238
All	3.6788± 0.0025	3.8732± 0.0247	3.7982± 0.0014

C. Discussion

Through the vertical and horizontal contrast, The result clearly demonstrates that the benefits of integrating RAG and other online searching facilities with glm-4 for domain-specific questions answering. The enhanced model is proven to provide more accurate and integrate answers to those queries in the field of admission about the university though utilizing the latest dateset excavated from the official website and the assisting online search tools for extended updated sources. These improvements would be critical for generating responses that conform the realistic policies and requirements of the university for the consultants.

V. Conclusion

The research trial explores the integration of Retrieval-Augmented Generation (RAG) and other search tools with the glm-4 model to support consultants for the information of a specific university. With the combination of those sources to enhance the performance of the model, we aim to construct a model whose responses could be conformed to reality and trialed to reduce the propensity to “hallucinate” . Our methodology included data collection from the official website of Xian Jiaotong-Liverpool University, preprocessing of the collected data, and implementation of RAG and two different search tools providing the sources from Baidu, the search engine, to enhance the capability of question answering of the model. The experimental results, evaluated and scored by the GPT-4 under the criteria of integrity and accuracy, demonstrated that the enhanced model, which utilized the sources retrieved by the RAG and original web searching tool, could provide more contextually relevant and accurate responses compared to other models and glm-4 with different combination of sources. However, there are several areas for improvement.

One primary limitation is the reliance on the updated knowl- edge base. that limited the performance of the model when the query is about the latest updated pieces of information. The updated data set for RAG would be one of the reliable strategies for this limitation. Future work could force on automatic update and maintenance system of the collected dataset to ensure the accuracy and integrity of the responses on real-updated data.

Another remarkable concern is the issue of data pollu- tion when utilizing those search tools, which gain correlated sources from the Internet. These aspects would also be im- portant to ensure the accuracy and improve the performance. Further works of constructing a filter for those search tools are expected.

Moreover, the bias of artificial intelligence may be another affection when evaluating, while certain recommended techniques [12]have been utilized to weaken the influence of this bias. Meanwhile, integrating with more comprehensive human evaluation is expected to provide deeply insights into those criteria.

In conclusion, while there are still limitations, this research demonstrates the potential of glm-4 enhanced by the RAG and other search tools that could support the consultants in different domains. Through constructing the current framework and discovering those limitations, our work would contribute to more robust and generally used solutions to increase the accuracy and integrity and further against the propensity of LLMs to “hallucinate in a large range of specific segmentation areas.

References

[1] T. Zhang, S. G. Patil, N. Jain et al, “RAFT: Adapting Language Model to Domain Specific RAG,” *arXiv e-prints*, p. arXiv:2403.10131, Mar. 2024.

[2] L. Huang, W. Yu, W. Ma et al, “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions,” *arXiv e-prints*, p. arXiv:2311.05232, Nov. 2023.

[3] A. Stolcke, “Srilm - an extensible language model- ing toolkit,” in *Interspeech*, 2002. [Online]. Available: https://api.semanticscholar.org/CorpusID:1988103

[4] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, “A neural probabilistic language model,” *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, 2003. [Online]. Available: https://api.semanticscholar.org/CorpusID:221275765

[5] M. E. Peters, M. Neumann, M. Iyyer et al, “Deep contextualized word representations,” ArXiv, vol. abs/1802.05365, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:3626819

[6] L. Huang, W. Yu, W. Ma et al, “A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions,” ArXiv, vol. abs/2311.05232, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265067168

[7] H. Soudani, E. Kanoulas, and F. Hasibi, “Fine tuning vs. retrieval augmented generation for less popular knowl- edge,” ArXiv, vol. abs/2403.01432, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268248396

[8] S. Xiao, Z. Liu, P. Zhang et al, “C-pack: Packed resources for general chinese embeddings,” in Proceedings of the 47th International ACM

SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 641–649. [Online]. Available: https://doi.org/10.1145/3626772.3657878

[9] J. Johnson, M. Douze, and H. J$\acute{e}$gou, "Billion-scale similarity search with GPUs," IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535– 547, 2019.

[10] DAIR.AI, Prompt engineering playbook. GovTech Data Science & AI Division, 2023, ch. 4, pp. 26–41.

[11] L. Zheng, W.-L. Chiang, Y. Sheng et al, "Judging llm-as-a-judge with mt-bench and chatbot arena," in Advances in Neural Information Processing Systems, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 46 595–46 623.

[12] P. Wang, L. Li, L. Chen et al, "Large language models are not fair evaluators,"2023. [Online]. Available: https://arxiv.org/abs/2305.17926