Retrieval-Augmented Generation Solutions for Typical Application Process Issues

Yudi Zhang[a]
Information and Computing Science, Beijing University of Civil Engineering and Architecture,
No. 15 Yongyuan Road, Beijing, China

Keywords: Retrieval-Augmented Generation, Large Language Model, Reranking.

Abstract: Challenges such as the generation of factually incorrect illusions, privacy issues, and outdated information often hinder the practical deployment in the Large Language Model (LLM). Retrieval-Augmented Generation (RAG), which utilizes advanced retrieval technology, is designed to address these issues. RAG will use the embedded vector model to build an external database for the information to be updated and the information of the application field, enhance the user prompt by adding the retrieved relevant data in the context, and retrieve the matched content of the vector library. In this process, how to improve the retrieval efficiency and quality, and how to improve the robustness of the model are the focus of the method discussed in this paper. Gradient Guided Prompt Perturbation (GGPP) uses top k to minimize the distance between the target paragraph embedding vector and query embedding vector while maximizing the distance between original paragraph embedding and query embedding to reduce the influence of perturbation on the model and improve the robustness of the model. Boolean agent RAG setups improve markup efficiency in a language model by incorporating Boolean decision steps where the language model determines whether to query vector databases based on user input. This setting saves a lot of tokens. GenRT is an algorithm that optimizes reordering and truncation strategies to improve efficiency and accuracy in processing long text. Finally, the application of medical question answering system is cited to find the best combination of the retrieval and LLM model in this field.

1 INTRODUCTION

Large language models (LLMS) are driving a major revolution in global development - freeing productivity. LLMS show an amazing talent for semantic understanding and reasoning (Chang, 2023; Kasneci, 2023). When using talent to join the Q&A system, it is inevitable to encounter problems caused by illusions and prejudices. For one thing, large language models may fabricate answers that contradict reality when they generate text. Second, large language models lack knowledge and information lag in vertical domains. These two factors will reduce the reliability of LLM large model, reduce the reliability of the model, and also encounter resistance in the process of commercial implementation (Zhao, 2023; Zhao, 2024).

In order to reduce illusion and bias, solve the problem of knowledge scarcity and information lag in the specialized knowledge-intensive tasks of large-scale models, and improve the accuracy of the Chinese text during the question-and-answer process of large-scale models. The current solution is to use Retrieval-Augmented Generation (RAG) enhanced retrieval technology (Chen, 2024; Lewis, 2020), vertical integration of professional domain related database, combined with appropriate embedding model, text information into vector information. The vector information of each text block is reordered to meet the input requirements of large models (such as using the BGE Ranker library to fine-tune the text), enhance the relevance of context retrieval under the professional domain retrieval function, and finally find the text block matching the user's question in the knowledge base. However, in this process, it is possible to encounter the problem of small interference in the user's query text affecting the results. There will be insufficient data expression,

[a] https://orcid.org/0009-0000-1072-6946

Zhang, Y.
Retrieval-Augmented Generation Solutions for Typical Application Process Issues.
DOI: 10.5220/0012917400004508

In *Proceedings of the 1st International Conference on Engineering Management, Information Technology and Intelligence (EMITI 2024)*, pages 164-167
ISBN: 978-989-758-713-9

improper text division, resulting in the system cannot find the appropriate context information caused by retrieval difficulties. These are the current RAG encountered resistance in the application process. How to reasonably combine the existing retrieves and large language models is also a problem that needs to be considered. (Xiong, 2024)

In order to solve the above problems, this paper will focus on the latest techniques and algorithms in search queries, reordering of embedded vector libraries, truncation strategy algorithms (Xu, 2024), and evaluation methods in different fields based on existing RAG enhanced retrieval. The workflow of RAG is shown in Figure 1. The current methods that can improve the retrieval speed, robustness and generalization of RAG in the application process are summarized. The types of RAG frameworks, the latest algorithms for queries, and generators in searchers are also discussed comprehensively. Combined with the large language model, this paper discusses the technical development direction of RAG's future application prospects and explores its potential commercial value in AI application and question answering system.

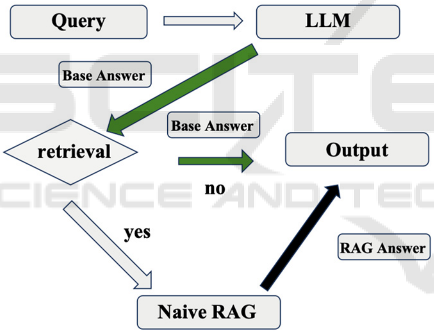

Figure 1: The workflow of RAG (Photo/Picture credit: Original).

1.1 Gradient Guided Prompt Perturbation (GGPP)

The retrieval part of RAG may generate erroneous results due to small errors prompted by users. The new algorithm, Gradient Guided Prompt Perturbation (GGPP), will optimize this type of problem. The encoder converts the user's text information into an embedding vector and connects the prefix to the query. The goal of prefix optimization algorithm in GGPP is to change the paragraph ranking in Large Language Models (LLMs) based on Retrieval Augmented Generation (RAG) extract the correct paragraph from the top-K search result and elevate the target paragraph to the top-K result. This algorithm achieves this by minimizing the distance between the target paragraph embedding vector and the input query embedding vector, while maximizing the distance between the original paragraph embedding and the query embedding. The prefix optimization algorithm adjusts the embedded coordinates to increase similarity with the target coordinates. If the query embedding is not made closer to the target specific point, restore the adjustment (Hu, 2024).

1.2 Boolean Agent RAG Setups

The Boolean Agent RAG (BARAG) configuration has the goal of improving token efficiency when a language model needs to decide if querying a vector database will lead to more precise and relevant responses. This is achieved by integrating a boolean decision-making step that aims to optimize the use of the model's built-in knowledge for answering queries without the need for unnecessary database retrievals. This method will likely lead to substantial reduction in token usage, especially in practical applications where database-retrieved text usually consumes most tokens (Kenneweg, 2024).

1.3 GENRT

In the retrieval technology section of RAG enhanced retrieval technology, how to effectively handle reordering and truncation is the key issue. These two questions involve whether the document sorting model fully understands the semantic context of the text information, and whether the truncation strategy directly affects the efficiency and quality of answer generation when facing long texts. However, single channel reordering and truncation models often lead to error accumulation.

Therefore, the algorithm in GenRT shown in Figure 2 focuses on combining dynamic reordering and static truncation and executing them simultaneously, and innovates two adaptive loss functions as core algorithms: one is step adaptive attention loss, which optimizes the attention score in each document of the rearranged table by calculating the cross entropy of attention distribution, and the other is step by step lambda loss, which aims to build a scoring matrix for each model step, helping the model find the best truncation point and generate higher relevant documents. The training method updates the rating matrix by adding penalty terms to documents in the sequence that do not decrease in correlation (Xu, 2024).

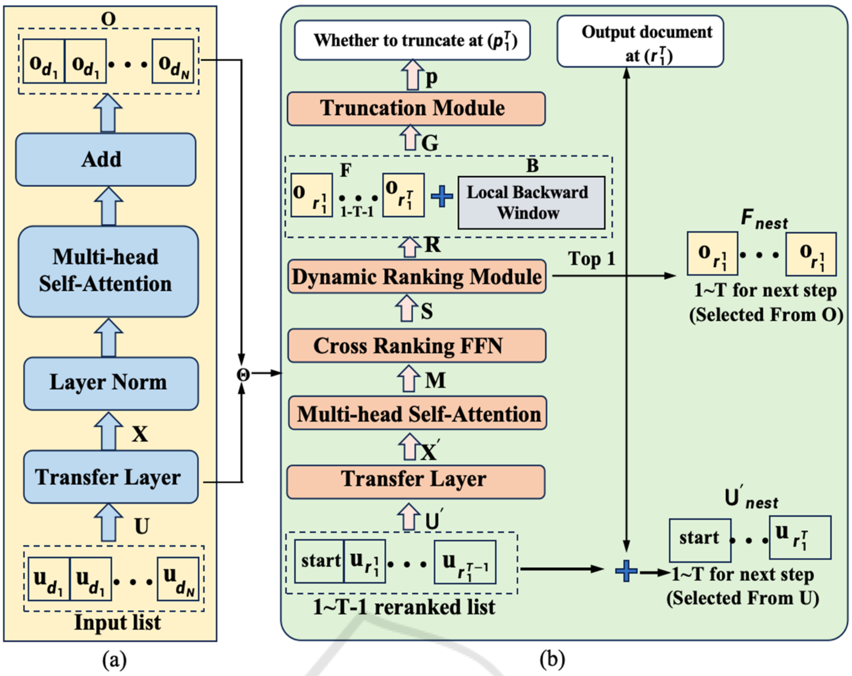
O
o_{d_1} o_{d_1} • • • o_{d_N}
Add
Multi-head Self-Attention
Layer Norm
X
Transfer Layer
U
u_{d_1} u_{d_1} • • • u_{d_N}
Input list
(a)
Whether to truncate at (p_1^T)
Output document at (r_1^T)
p
Truncation Module
G
F
B
$o_{r_1^1}$ • • • $o_{r_1^T}$ 1-T-1
Local Backward Window
R
Dynamic Ranking Module
Top 1
F_{nest}
$o_{r_1^1}$ • • • $o_{r_1^1}$
1~T for next step (Selected From O)
S
Cross Ranking FFN
M
Multi-head Self-Attention
X′
Transfer Layer
U′
U'_{nest}
start • • $u_{r_1^T}$
start $u_{r_1^1}$ • • • $u_{r_1^{T-1}}$
1~T-1 reranked list
1~T for next step (Selected From U)
(b)

Figure 2: The workflow of GenRT (Photo/Picture credit: Original).

2 INDUSTRIAL APPLICATION

Due to the excellent performance of RAG technology in knowledge intensive tasks, the biomedical question answering field is also beginning to use RAG technology. Researchers have found that RAG technology can effectively address the hallucinations and outdated issues exhibited by LLM in medical question and answer tasks. So, the researchers used the MIRAG evaluation benchmark MEDRAG toolkit to conduct large-scale experiments using different combinations of corpora, retrievers, and LLM. The MIRAG evaluation benchmark includes four enhanced retrieval methods, including RAG and five common medical corpora: MMLU Med, MedQA-US, MedMCQA, PubMedQA *, BioASQ-Y/N. The MEDRAG toolkit mainly consists of three parts: Corpora, Retrievers, and LLMs. Attempt to find the optimal solution between these combinations in the application scenarios of medical question answering. In the end, it was found that MEDRAG improved the accuracy performance of six different LLMs in the test set by 18% compared to chain of thought promotion. Among them, chatgt-4 is the LLM with the best performance, but gpt-3.5 is affected by RAG in terms of accuracy and performance, with the largest increase in numerical values. MEDRAG has shown great potential in enhancing LLM's Zero Shot Learning ability to answer medical questions, which may be a more effective choice than conducting larger scale pre training. Although supervised fine-tuning methods may be more suitable, MEDRAG remains a more flexible and cost-effective approach to improving medical Q&A (Xiong, 2024).

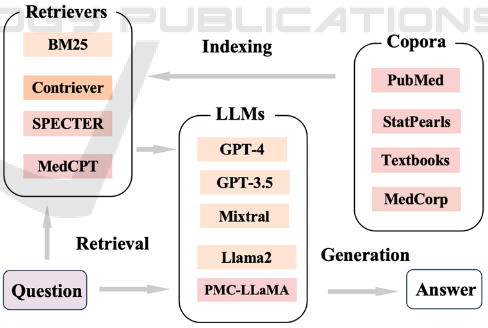

Figure 3: The application workflow of RAG (Photo/Picture credit: Original).

3 DISCUSSION

The current application of RAG has been implemented in knowledge-intensive fields, and in the application of complex tasks, such as: graph understanding and question answering, code generation. Applications in special fields include: medical Q&A, disaster summary, textbook Q&A, etc. It is believed that RAG will further enhance the commercial value of LLM in knowledge-intensive industries in the near future. Its ability to solve model illusion, update model information base in real time and keep it private all show its potential to improve the accuracy, reliability and stability of question answering field, and make the generated text results more realistic. These will lead to it becoming an inseparable part of the future private deployment of ai products, such as enterprise private document review, industry bid assistance writing and other customized functions. The great promise of the business sector also means that more vertical-specific assessment methods or test datasets is required. At the same time, the dynamic information field such as finance and news media to establish a regular data update process to meet the needs of the industry. This process can automatically complete the extraction, analysis and update of new data to meet the needs of the industry.

RAG enhanced retrieval technology also encounters limitations. For example, traditional vector retrieval cannot represent logical reasoning connections due to its embeddedness and lacks real relevance and thought chain. The deficiency of knowledge base context and the loss of key information in the process of compressing paragraph vector into single vector will inevitably lead to the problem of knowledge waste. Who will dominate RAG Retrieval and Generation more, and whether their performance in different fields will make their emphases different, these questions will directly lead to whether rag will be oriented towards search or agent in the future technological development path. The constraints or balance points between these should be focused, these are still unsolved challenges.

Challenges come with new technological possibilities. For example, the use of knowledge graph embedding makes the generated results more interpretable and logical upward compatible, so that logical reasoning is more accurate. For the pain points of insufficient context in the knowledge base, the enhancement of context information by using more efficient document parsing tools can be ensured to add relevant metadata to each paragraph of text.

4 CONCLUSION

This paper mainly summarizes RAG's recent algorithms for improving retrieval efficiency and the impact of weak interference brought by improving user prompts on generated results, as well as the introduction and discussion of vector data reranking and truncation strategies. The current RAG enhancement was introduced to enhance the retrieval robustness and enhance the explainability of the query method. The RAG application in the current medical question answering system is introduced, and the performance difference of its hybrid training method in the application side is presented. By using the evaluation tool suitable for the field of medical question answering, the combination of LLM and retrieval device suitable for this field is obtained. Finally, the commercial application prospect and future research direction of RAG are discussed. There is reason to believe that RAG enhanced retrieval technology will become an important part of the privatization ai deployment boom in the future.

REFERENCES

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., ... & Xie, X. 2023. A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology.

Chen, J., Lin, H., Han, X., & Sun, L. 2024. Benchmarking large language models in retrieval-augmented generation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 16, pp. 17754-17762).

Hu, Z., Wang, C., Shu, Y., & Zhu, L. 2024. Prompt perturbation in retrieval-augmented generation based large language models. arXiv preprint arXiv:2402.07179.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. Learning and individual differences, 103, 102274.

Kenneweg, T., Kenneweg, P., & Hammer, B. 2024. Retrieval Augmented Generation Systems: Automatic Dataset Creation, Evaluation and Boolean Agent Setup. arXiv preprint arXiv:2403.00820.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Xiong, G., Jin, Q., Lu, Z., & Zhang, A. 2024. Benchmarking retrieval-augmented generation for medicine. arXiv preprint arXiv:2402.13178.

Xu, S., Pang, L., Xu, J., Shen, H., & Cheng, X. 2024. List-aware reranking-truncation joint model for search and retrieval-augmented generation. arXiv preprint arXiv:2402.02764.

Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., ... & Cui, B. 2024. Retrieval-Augmented Generation for AI-Generated Content: A Survey. arXiv preprint arXiv:2402.19473.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. 2023. A survey of large language models. arXiv preprint arXiv:2303.18223.