RAG-powered Domain Specific/Time Sensitive Q&A Application

Juan Yu **yu.juan@northeastern.edu**

Northeastern University
Khoury College of Computer Sciences

INTRODUCTION

Given their exceptional text generating capability, LLM-backed Q&A systems are gaining significant popularity. However, they are limited by their training data which may not be domain-specific or up-to-date.

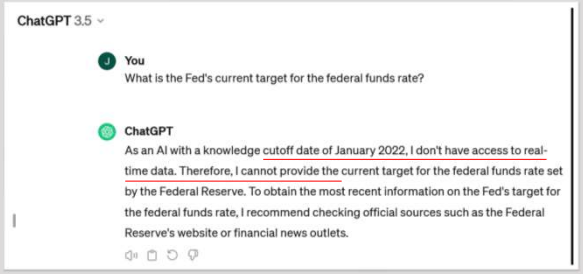

Fig 1. Result from ChatGPT

SOLUTION

Retrieval-Augmented Generation

- ✓ Support input of user-defined knowledge bases
- ✓ Retrieve relevant contexts
- ✓ Generate desirable answers

ARCHITECTURE

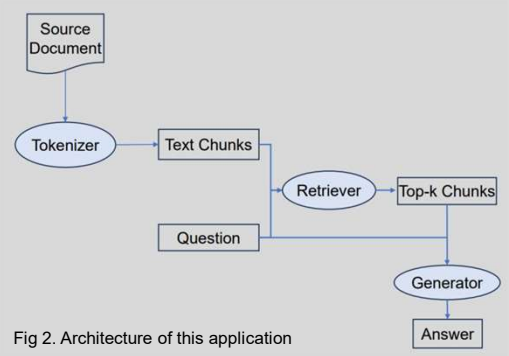

Fig 2. Architecture of this application

- **Tokenizer:** SpaCy
- **Retriever:** Sentence-Transformers + Cosine Similarity
- **Generator:** Llama 2

INTERFACE

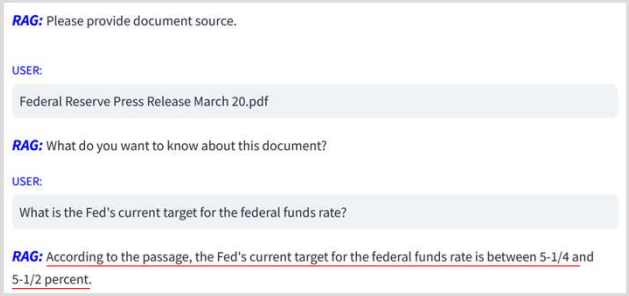

Fig 3. User interface of this application

EVALUATION

- ☐ Generated answers compared to ground truth

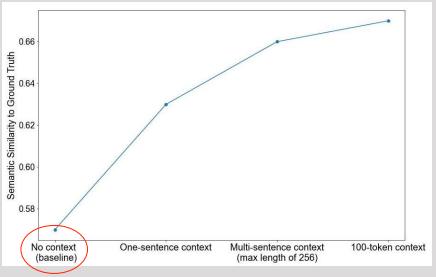

Fig 4. Performance of this application

Note: the evaluation was based on question-context-answer triplets for Apple's Annual Earnings Report 2022.

CONCLUSIONS

- This RAG-framed application can improve the Q&A performance
- Length of context matters a lot
- Future work will be focused on automating the process of analyzing source documents and deciding the right length of context.

REFERENCES

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

[2.] Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. In *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (pp. 6769-6781). (EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference). Association for Computational Linguistics (ACL).

[3.] Wang, Z., Araki, J., Jiang, Z., Parvez, M. R., & Neubig, G. (2023). Learning to filter context for retrieval-augmented generation. *arXiv preprint arXiv:2311.08377*.