GARAGE: Generative-Augmented Retrieval Assisting Generation Enhancement

Krzysztof Jankowski[1]* Michał Janik[1,2]* Michał Grotkowski[1]* Antoni Hanke[1]*
Grzegorz Preibisch[1,3]

*equal contribution [1]University of Warsaw [2]Allegro [3]Deepflare

Contact: kj418274@students.mimuw.edu.pl

Introduction

Large Language Models like ChatGPT have shown to be of great use, but in tasks requiring **factual knowledge** they display two major issues:

- **hallucination**
- **costly model updates to existing knowledge via fine-tuning**

These issues can be addressed by performing **document retrieval** on a pre-existing knowledge database, that provides the LLM with relevant passages, to generate an informed answer.

Our work combines multiple existing **machine learning and classical** techniques to improve document retrieval and answer generation, resulting in a powerful ensemble that **outperforms previous popular models** in domain-specific question answering.

Building Blocks

In the conducted experiment, we focus on combining previous approaches that address different stages of the retrieval-based question-answering pipeline:

- **Generation-Augmented Retrieval (GAR)**[1] (pre-retrieval stage) - a technique that given a query, tries to generate **relevant contexts** that are then used together with the original query in the retrieval stage
- **RAG**[2] (retrieval stage) - a deep learning-based model that retrieves relevant passages based on **similarity between query and passage embeddings**
- **BM25**[3] (retrieval stage) - a **classical retriever** model based on the statistical count of words and inverse document frequency

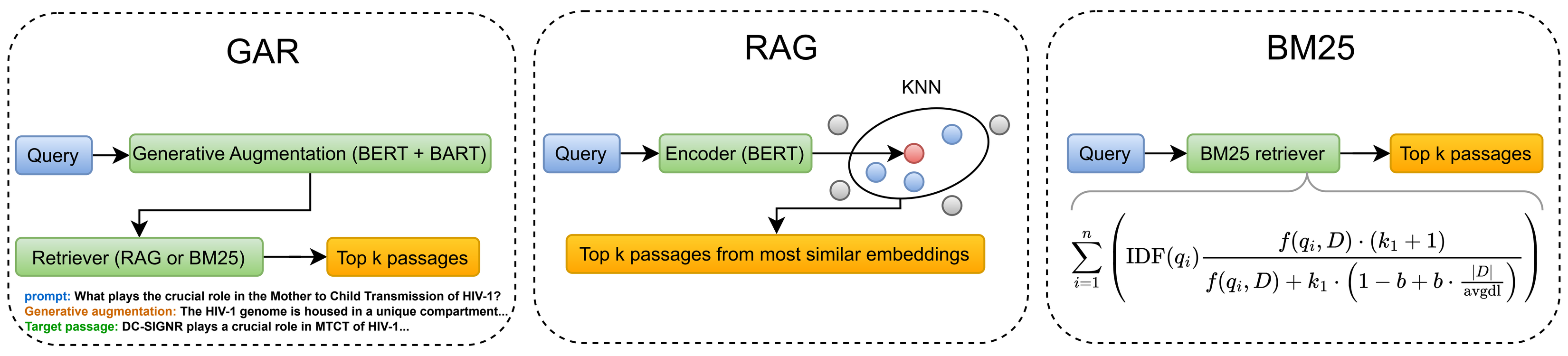

Figure 1: RAG, BM25 and GAR details.

Architecture

Our novel approach, **GARAGE**, improves the RAG setup in two ways: firstly, before the retrieval stage, we **augment** the query via the **GAR** encoder-decoder model (BERT + BART), resulting in more keywords being passed to the retrievers. We then propose combining the passages retrieved by a RAG **neural document retriever** with those retrieved by **BM25**.

These models create a **powerful ensemble of retrievers** and their output is passed to an instruction fine-tuned LLM (ChatGPT) with special prompt engineering.

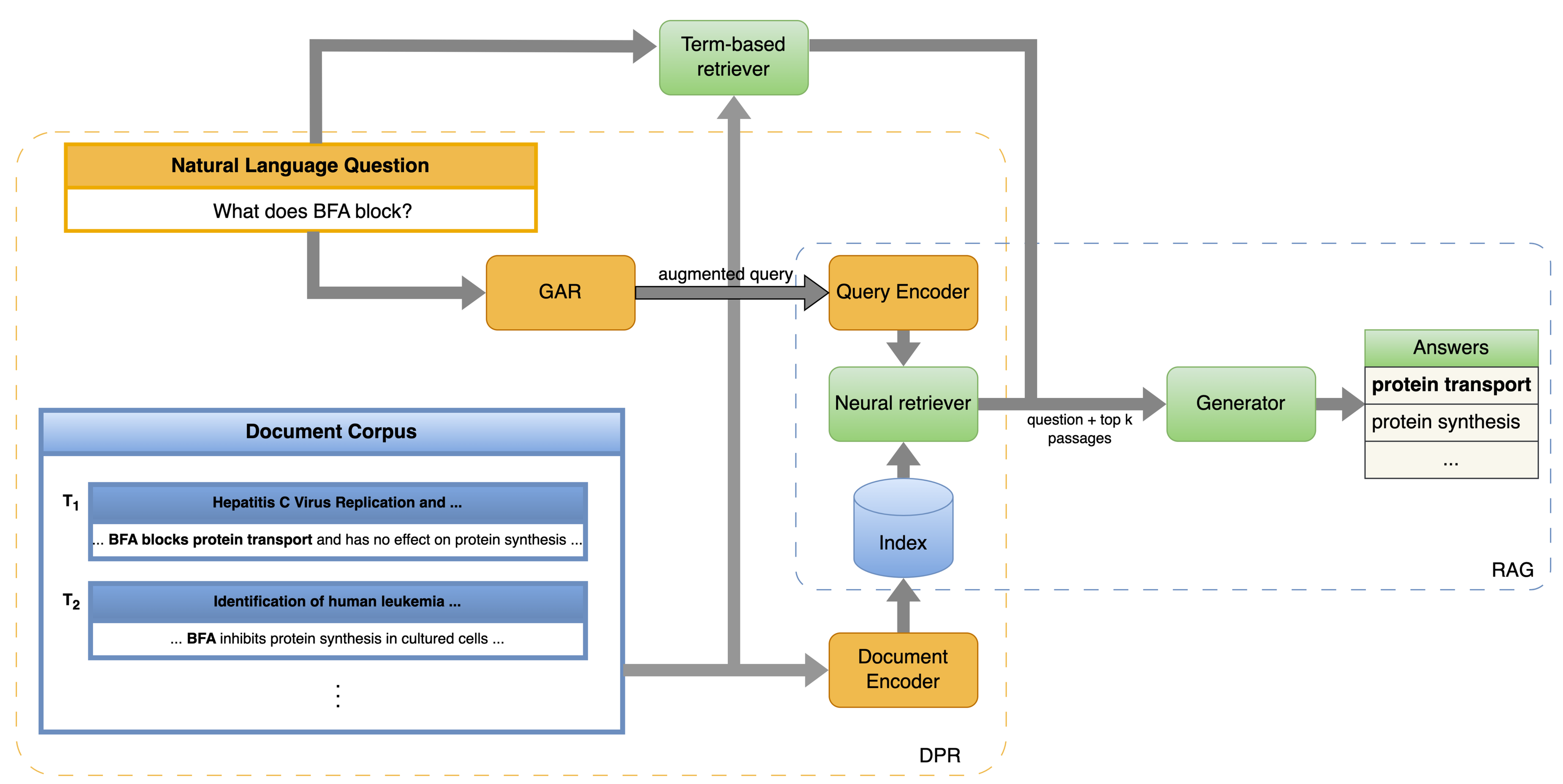

Figure 2: GARAGE architecture.

Hallucination mitigation

Augmenting LLMs with an external non-parametric memory (knowledge base) **significantly reduces hallucination.**

Without passages, ChatGPT often hallucinates answers, but with provided passages, it generates answers based on them **97% of the time**.

If the answer is not in the passages, ChatGPT avoids responding in one-third of the cases. Providing passages shifts ChatGPT from guessing to answering based on sources, addressing safety concerns about hallucinations.

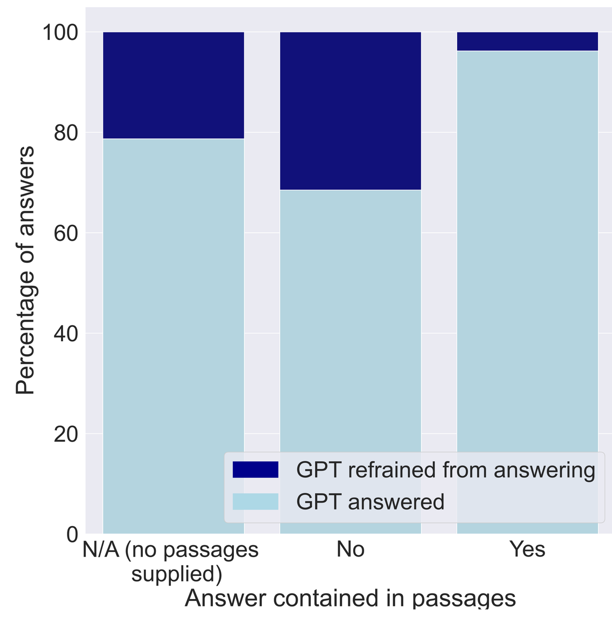

Figure 3: Percentage of unanswered questions by ChatGPT.

Experimental setup

We benchmark our model on CovidQA - the same subset of CORD-19 dataset[4] consisting of **5,000 medical articles** used in baseline models: RAG and fine-tuned RAG.[5]

We use top-k accuracy metrics for retrieved passages, and exact match and F1 score for answer generation. All the experiments are conducted on a **single 16GB GPU**.

Results

We benchmark GARAGE against the original RAG and its improved RAG-end2end-QA variant, in passage retrieval and answer generation. Additionally, we compare it to ChatGPT in answer generation.

Our model **outperforms other models** in almost all metrics. We conduct experiments for **various combinations** of retrievers and their proportions of contributed passages to the final top-k passages.

Retriever	Top-5	Top-20
BM25 + RAG	22.83	32.92
GAR (RAG)	8.33	11.05
80%(BM25+RAG) + 20%GAR(BM25)	**24.48**	**35.98**
BM25	22.83	29.86
RAG	10.48	15.64
RAG-end2end-QA	19.85	26.91

Table 1: Top-k accuracy for document retrieval on CovidQA.

Method	EM	F1
BM25 + BART	5.78	13.56
GAR (RAG)	1.87	5.59
40%BM25 + 60%RAG + ChatGPT	2.21	**18.74**
ChatGPT zero-shot	0.74	12.32
RAG	1.87	6.17
RAG-end2end-QA	**8.08**	18.38

Table 2: Exact Match and F1 score with top 5 retrieved passages (except ChatGPT zero-shot).

Efficiency of Hybrid Passage Retrieval

By integrating top-ranked passages from **both BM25 and RAG retrievers**, we observed a notable enhancement in retrieval accuracy. This improvement is especially pronounced in unfamiliar and challenging domains. Our hybrid approach consistently surpasses other methods across various passage numbers.

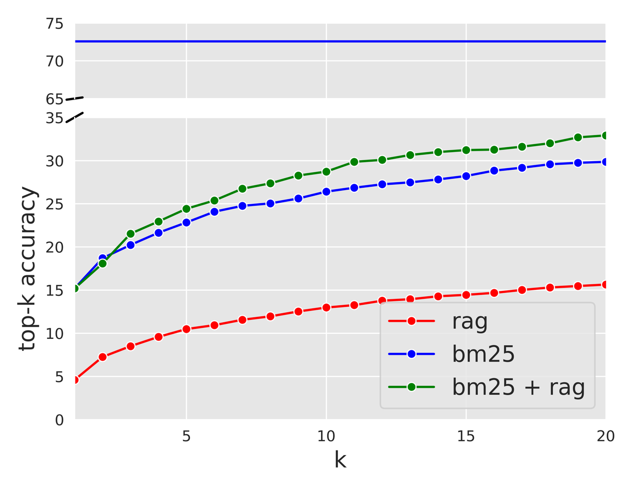

Figure 4: Top-k accuracy of retrievers.

Summary

We show that by **combining classical and neural** retrieval approaches in domain-specific question answering we can **outperform** fine-tuned models with a significantly **smaller compute budget**. Thanks to this approach popular LLMs can hallucinate less, be more specific in domain question answering, and give users more control of the model's knowledge base.

References

[1] Y. Mao, Y. Mao, P. He, X. Liu, X. Liu, Y. Shen, J. Gao, J. Gao, J. Han, and W. Chen, "Generation-augmented retrieval for open-domain question answering," *arXiv: Computation and Language*, 2020.

[2] P. S. H. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *CoRR*, vol. abs/2005.11401, 2020.

[3] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford, "Okapi at trec-3.," pp. 0–, 01 1994.

[4] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. Kinney, Y. Li, Z. Liu, W. Merrill, P. Mooney, D. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. Wade, K. Wang, N. X. R. Wang, C. Wilhelm, B. Xie, D. Raymond, D. S. Weld, O. Etzioni, and S. Kohlmeier, "Cord-19: The covid-19 open research dataset," 2020.

[5] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering," *Transactions of the Association for Computational Linguistics*, 2022.