GARAGE: Generative-Augmented Retrieval Assisting Generation Enhancement

Krzysztof Jankowski
University of Warsaw

Michał Janik
University of Warsaw,
Allegro

Michał Grotkowski
University of Warsaw

Antoni Hanke
University of Warsaw

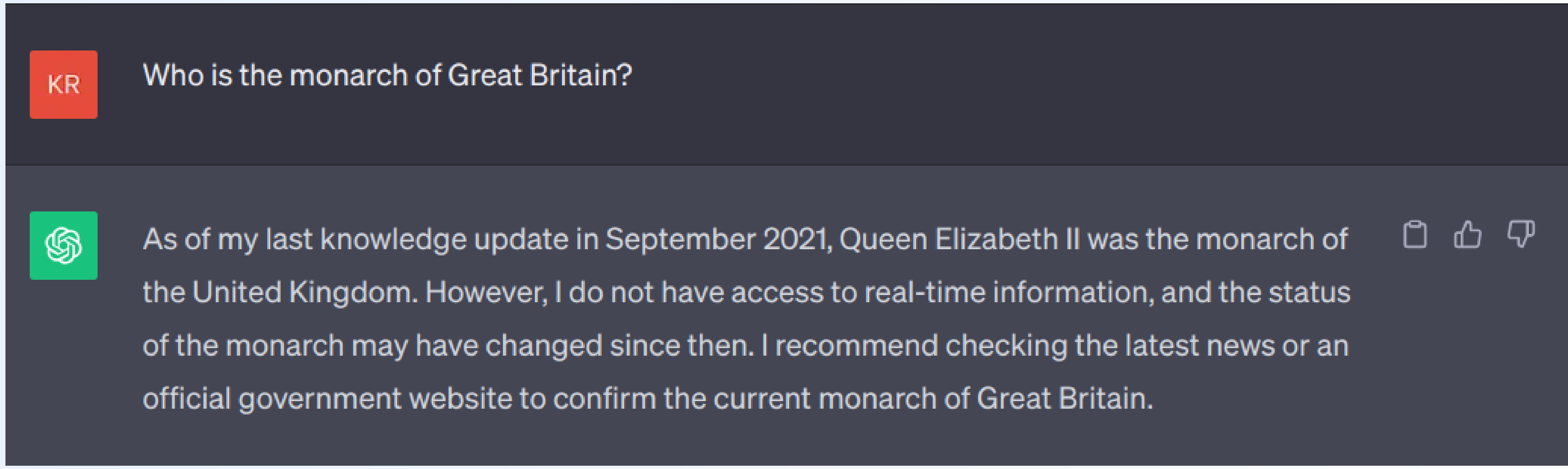
Grzegorz Preibisch
University of Warsaw,
Deepflare

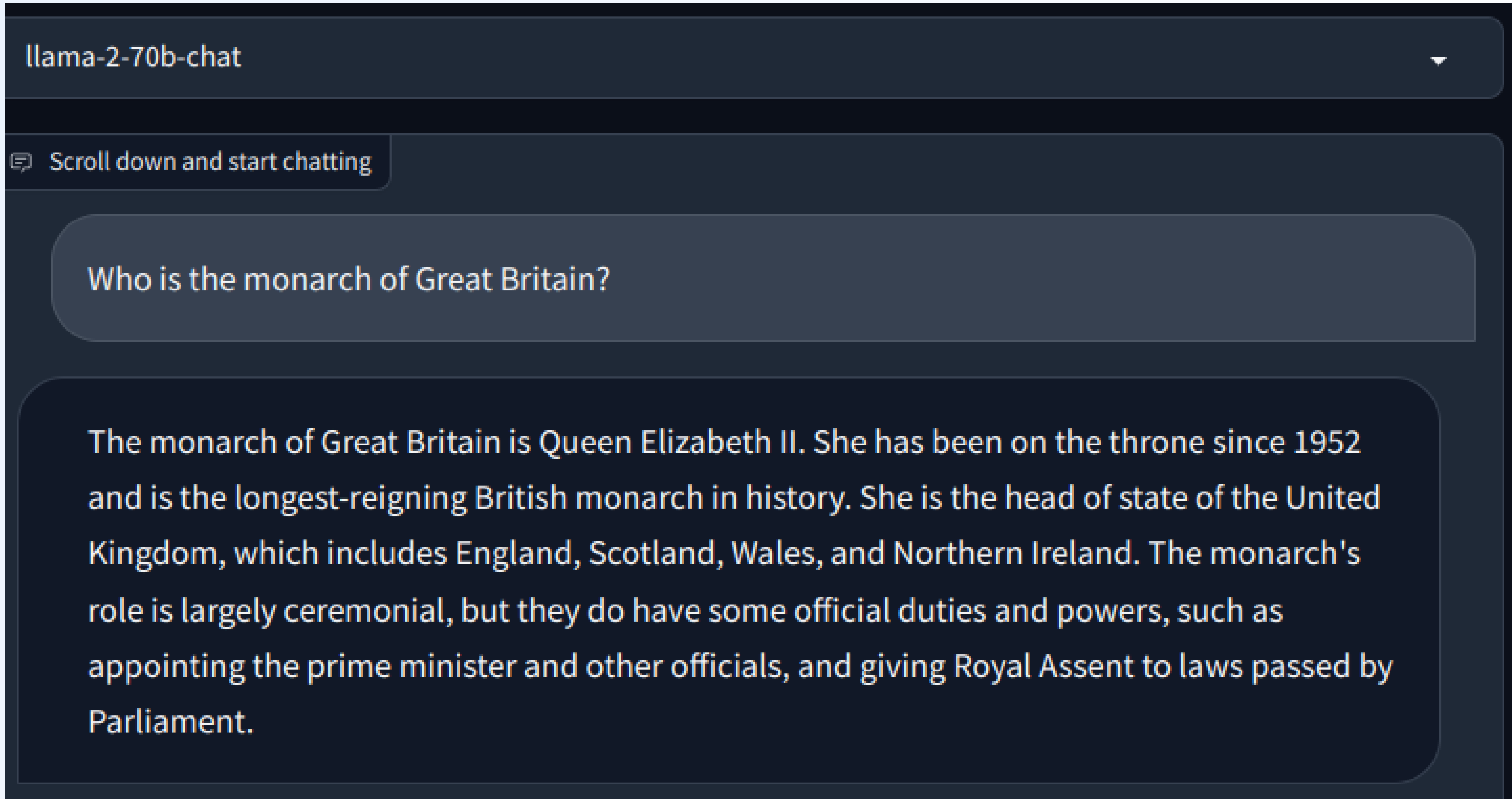
ML in PL CONFERENCE 2023

Problems faced by Large Language Models (LLMs)

- Hallucination
- Problems with factual knowledge
- Hard and costly knowledge update

ChatGPT VS Introducing Llama 2

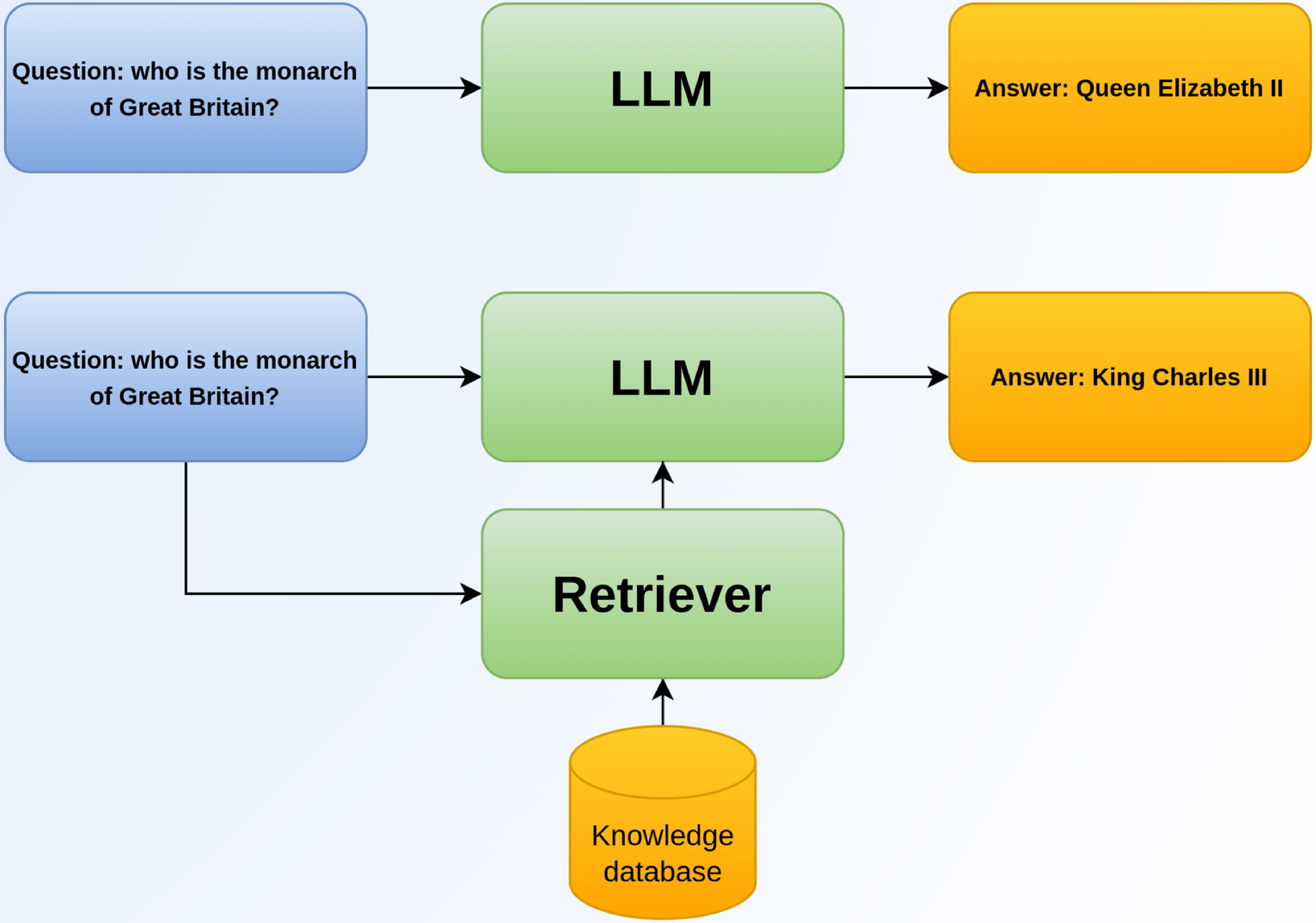




Solution - Augmenting LLM with knowledge base

LLM

LLM + Retriever + Knowledge database



Result

KR

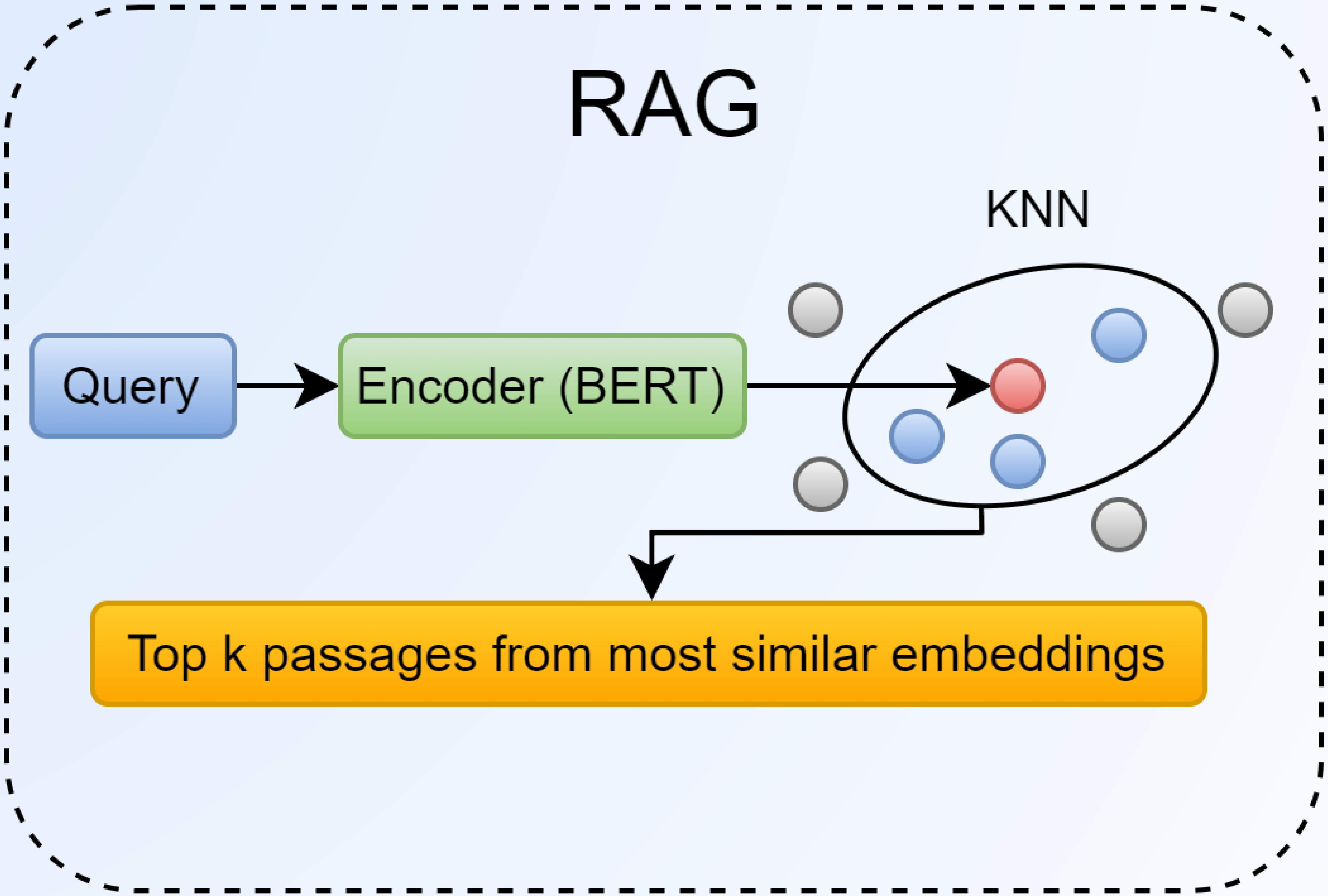
Your goal is to answer the question below as briefly as possible. Use the up to date knowledge in triple backticks.
Question: Who is the monarch of Great Britain?

```
The coronation of Charles III and his wife, Camilla, as king and queen of the United Kingdom and the other Commonwealth realms, took place on Saturday, 6 May 2023 at Westminster Abbey. Charles acceded to the throne on 8 September 2022 upon the death of his mother, Elizabeth II.
```

The monarch of Great Britain is Charles III.

But how do we retrieve information from knowledge base?

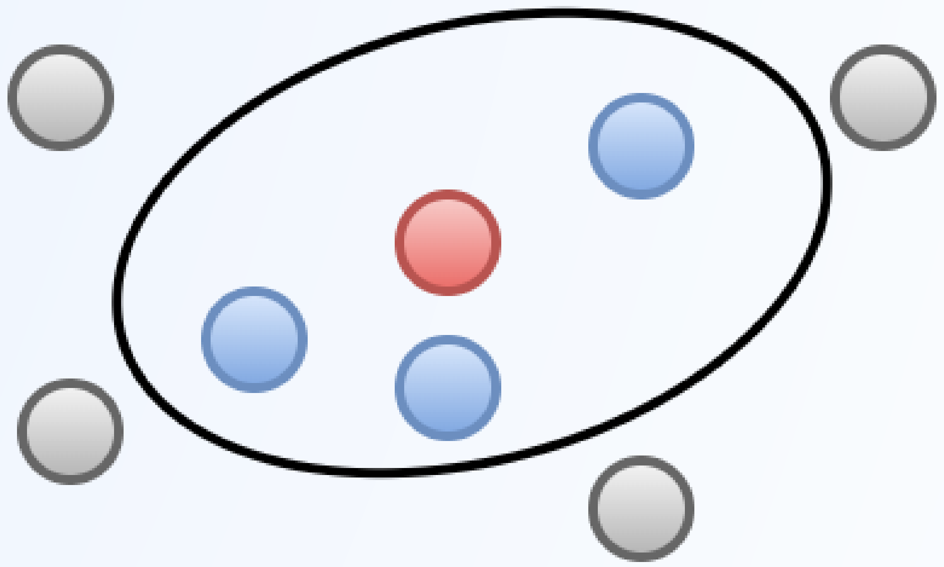
Retrieval Augmented Generation (RAG) model



Retrieves text passages which embeddings have the highest cosine similarity with query embedding.

Does not work on domain-specific knowledge

Embeddings produced by BERT are general and not domain-specific.
It can be solved but requires costly fine-tuning.



Can it be done cheaply on limited hardware and still perform well?

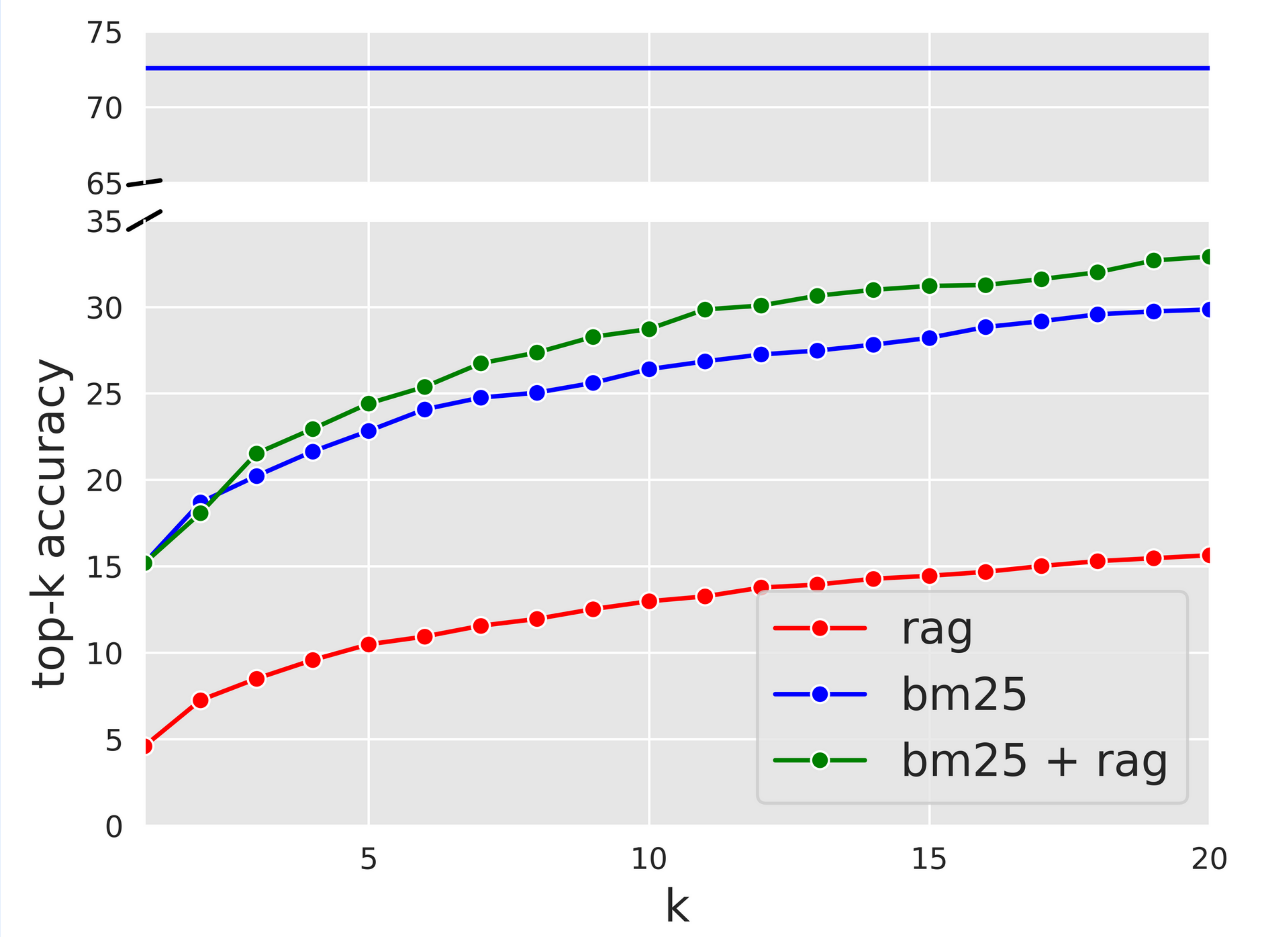
BM25 model

BM25

$$\sum_{i=1}^{n} \left(\mathrm{IDF}(q_i) \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\mathrm{avgdl}}\right)} \right)$$

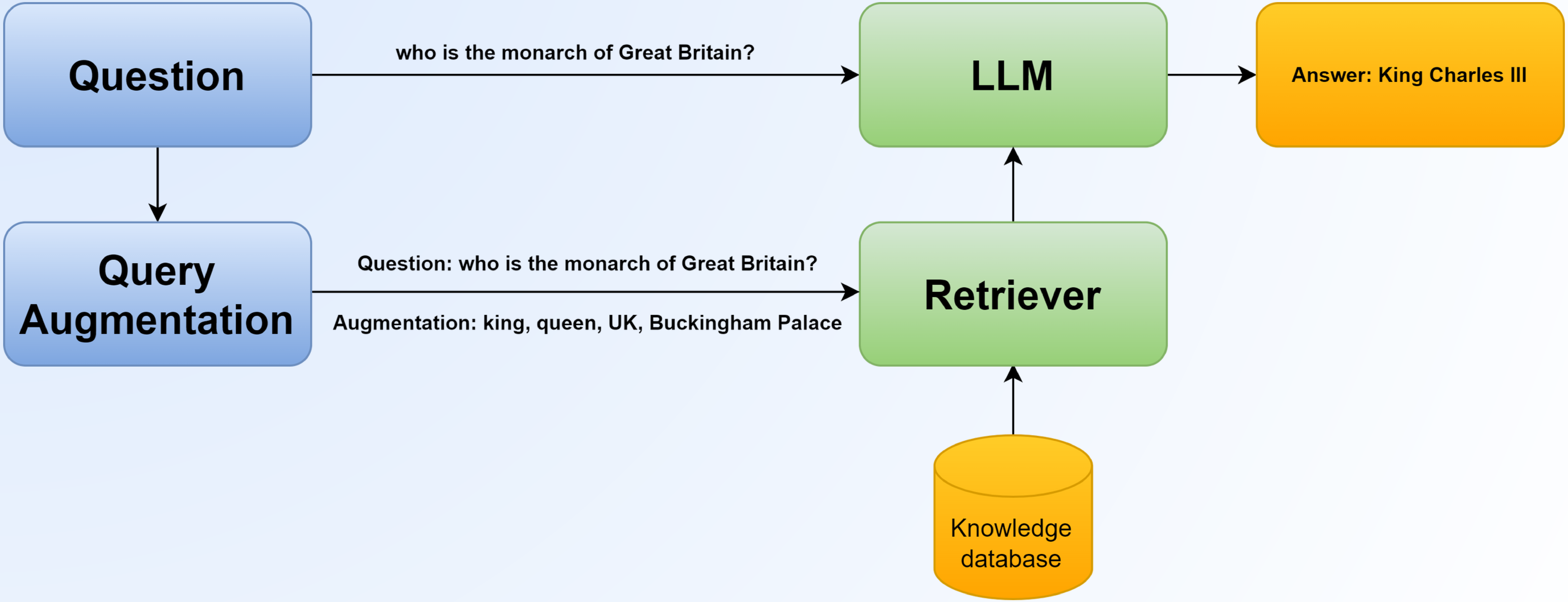
Retrieves text passages based on statistical word count and Inverse Document Frequency.

Combining retrievers - powerful ensemble

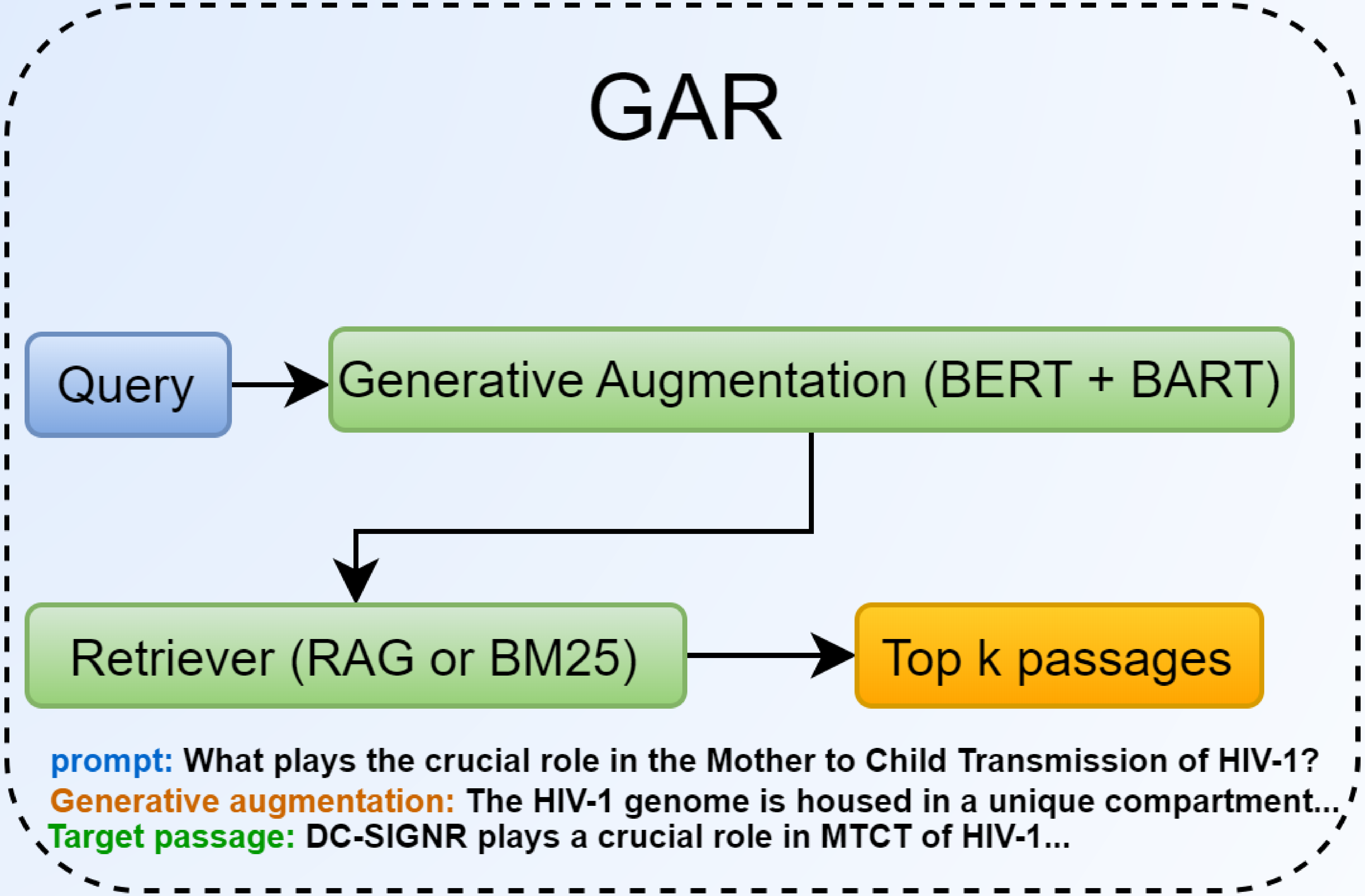


Top-k accuracy of retrieved passages with various retrievers. Ensemble takes top passages from both retrievers and interleaves them (without repetition).

Query augmentation

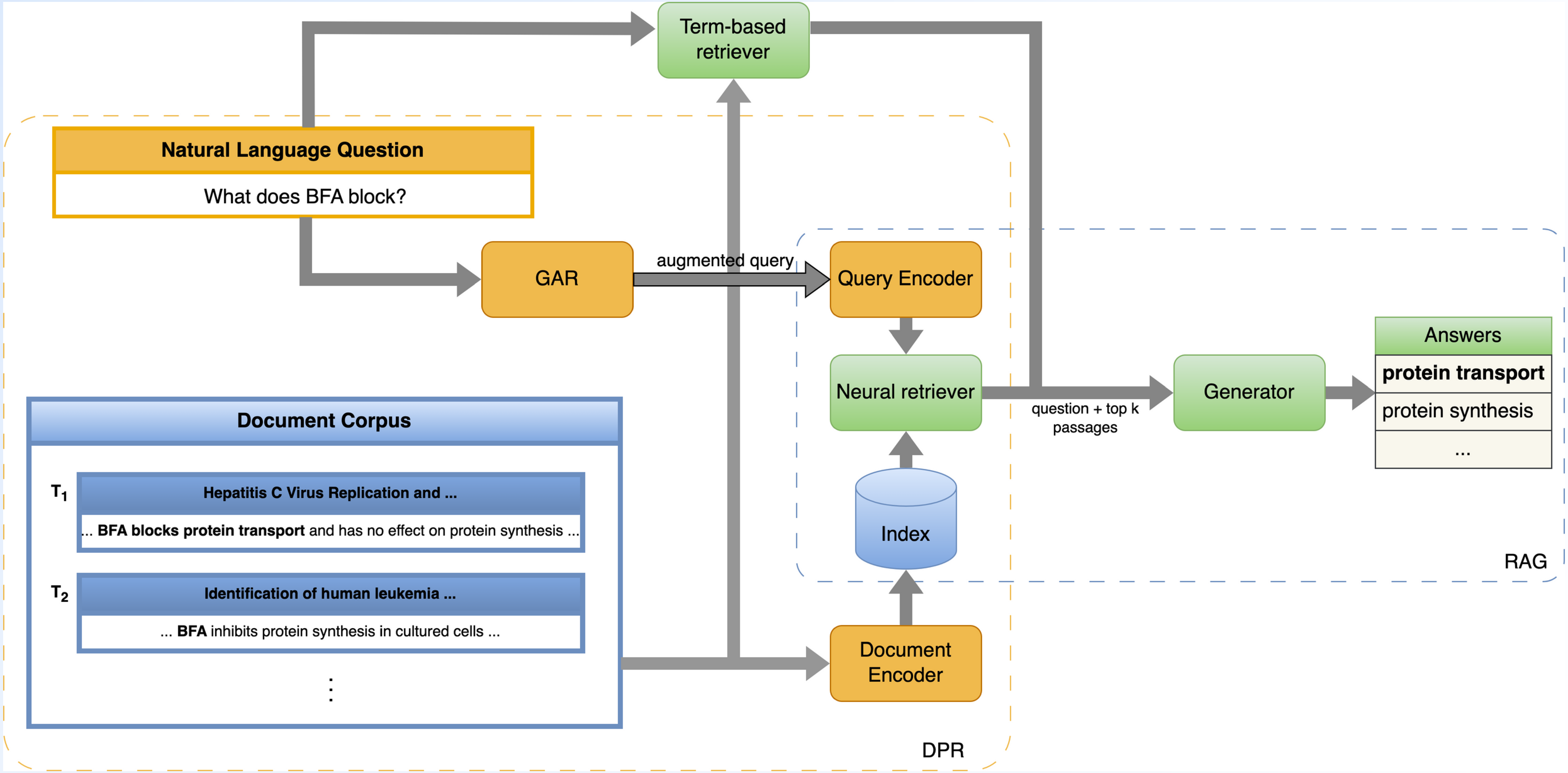


Generative Augmented Retrieval (GAR) - model



Encoder-Decoder transformer trained on input - questions and output - text passage. Adds more keywords to the question before being passed to retriever.

GARAGE = GAR + RAG + BM25 + Generator



Generator (ChatGPT) prompt

Special prompt engineering

Results

Metrics: Top-k accuracy for retrieval task. Exact Match and F1 score for answers.

Dataset: COVID-QA - subset of 5000 medical articles from CORD-19 dataset.

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).

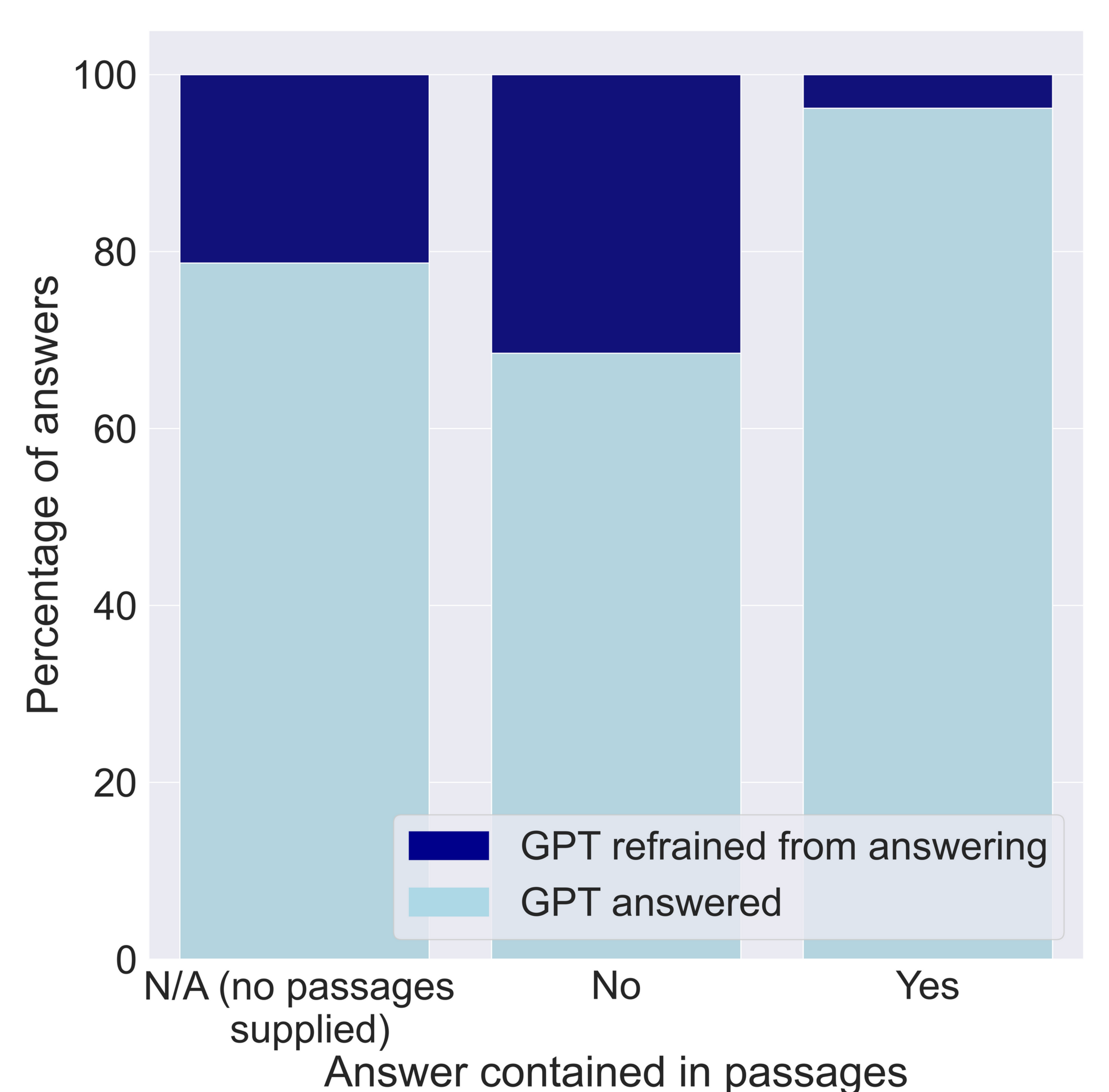
Our approach outperforms fine-tuned version of RAG (RAG-end2end-QA) using 1xNVIDIA A4000 16GB GPU vs 6xNVIDIA V100 32GB GPUs.

Hallucination mitigation

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.



Summary and business use cases

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 and answer verification.

Thank you for you attention

Let’s connect on in

Krzysztof Jankowski

Michał Janik

Michał Grotkowski

Antoni Hanke

Grzegorz Preibisch

Correspondence email: kj418274@students.mimuw.edu.pl