Google Cloud

Long-Context LLMs Meet RAG: Overcoming Challenges for Long Inputs in RAG

Bowen Jin[12], Jinsung Yoon[2], Jiawei Han[1], Sercan Arik[2]

[1] University of Illinois at Urbana-Champaign
[2] Google Cloud AI Research

RAG support various applications

- Large language models often struggle with factual inaccuracies and produce hallucinated content when faced with knowledge-intensive questions.
- *Retrieval Augmented Generation (RAG)* incorporates information retrieved from an external knowledge sources into the context to provide up-to-date information and specify obscure facts.

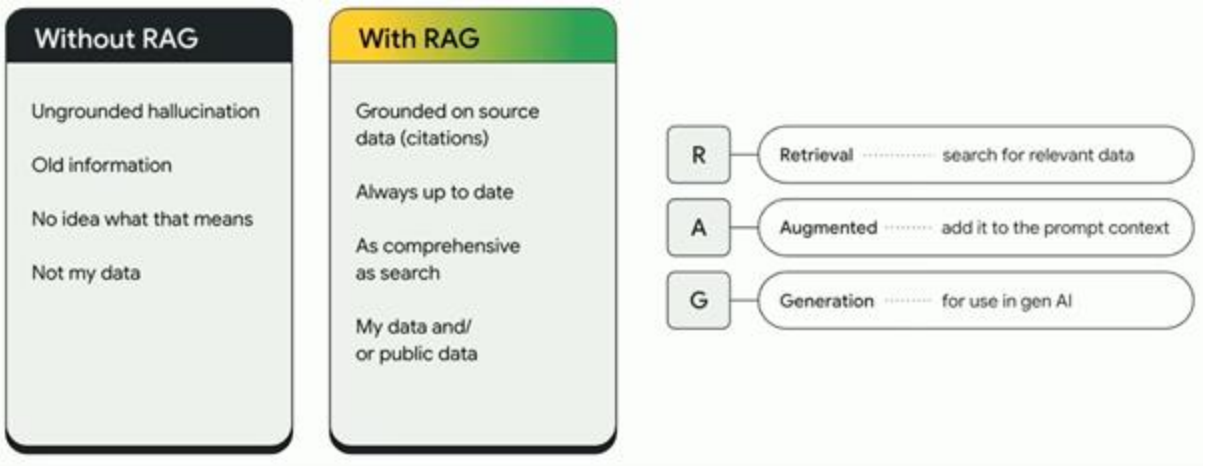
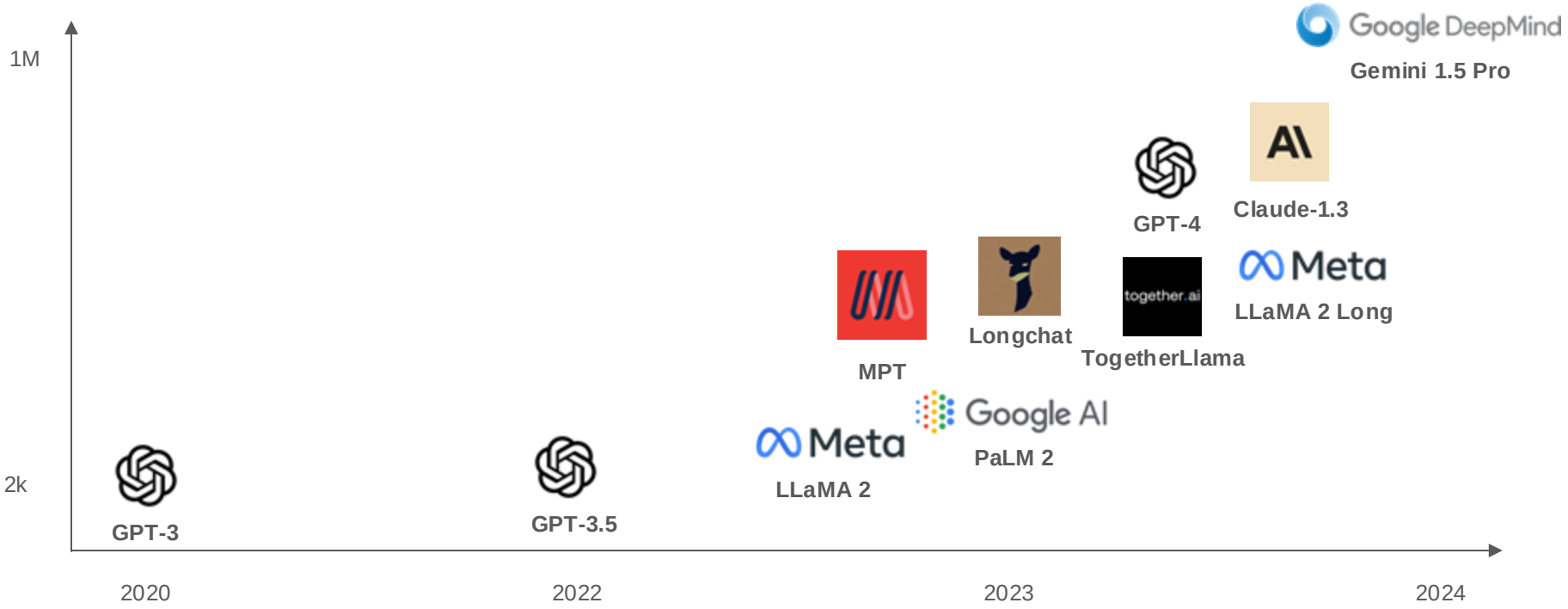


Image from: https://cloud.google.com/use-cases/retrieval-augmented-generation

LLMs are able to support longer context

Advance in compute and research on efficient training have led to increased model context lengths

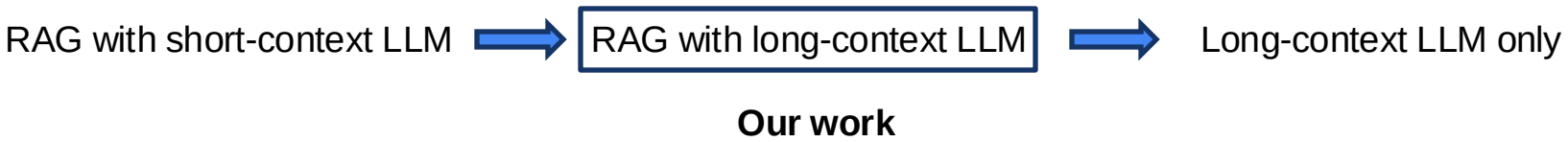


RAG or Long-context LLM?

RAG or Long-context LLM?

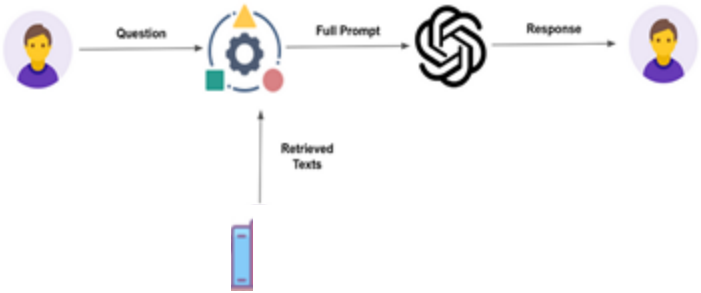
RAG with short-context LLM  Long-context LLM only

RAG or Long-context LLM?

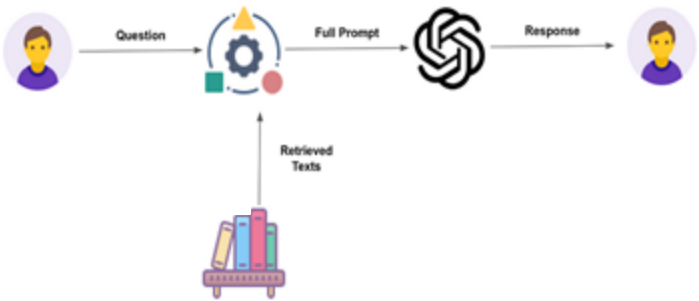


Long context LLM in RAG

- If we adopt a long context LLM in RAG, we are able to increase the number of retrieved document.
  - -> Improve the retrieval recall -> more relevant documents
- Existing works in RAG usually adopt top 1 [1], top 3 [2] or top 5 [3] retrieved documents.



previous



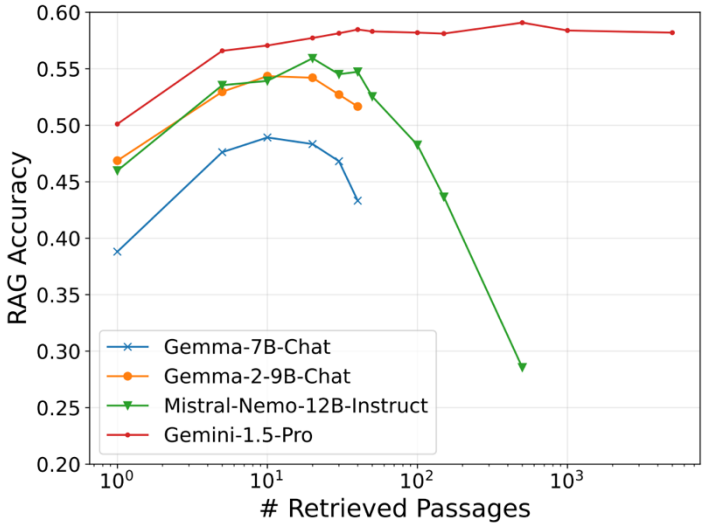
now

[1] Making Retrieval-Augmented Language Models Robust to Irrelevant Context. ICLR 2024.
[2] RA-DIT: Retrieval-Augmented Dual Instruction Tuning. ICLR 2024.
[3] Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. ICLR 2024.

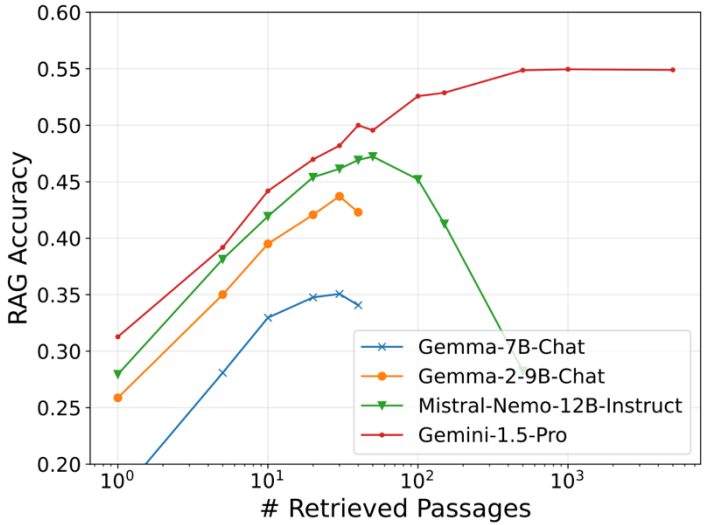
The effect of retrieved context size on RAG performance

Research question: Does a larger volume of retrieved context consistently translate to better performance when using long-context LLMs in RAG?

- Dataset: Natural question
- Retriever: e5 (on the left), bm25 (on the right)
- Generator: Gemma-7B-chat, Gemma-2-9B-chat, Mistral-Nemo-12B-instruct, Gemini-1.5-pro

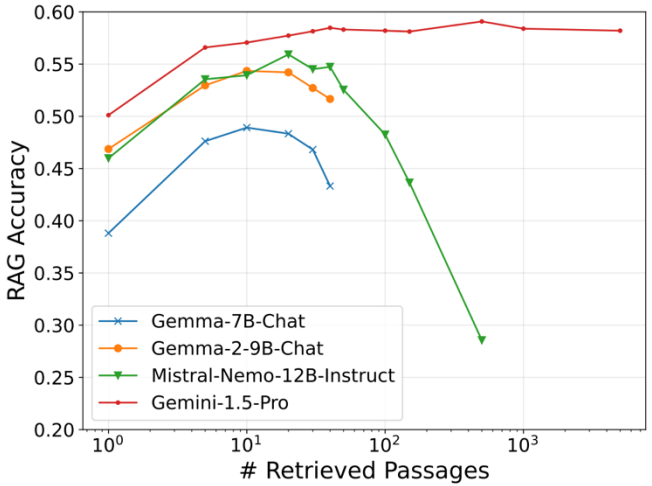


(a) RAG performance with e5 retriever

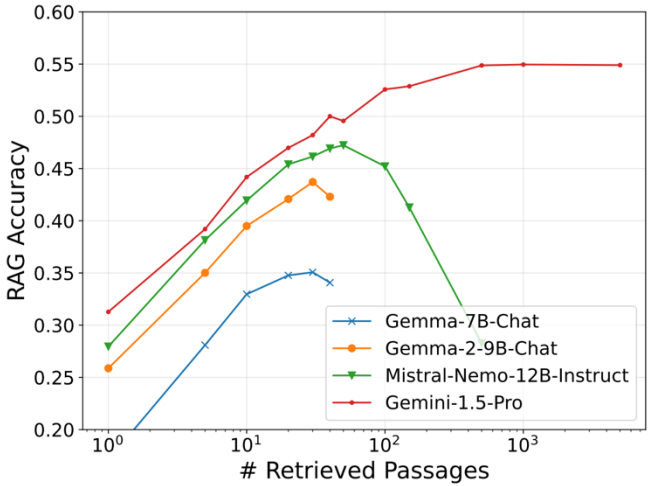


(b) RAG performance with BM25 retriever

The effect of retrieved context size on RAG performance



(a) RAG performance with e5 retriever

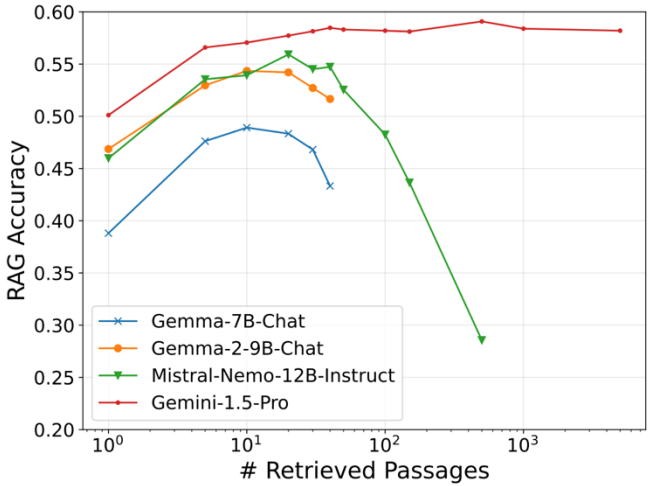


(b) RAG performance with BM25 retriever

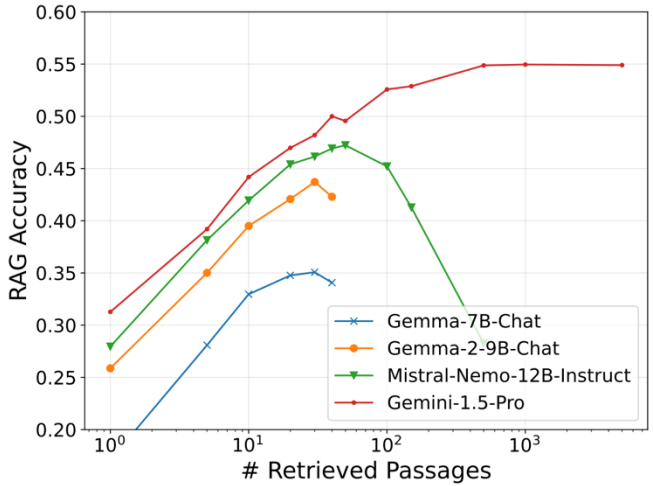
Observations

1) Strong Retriever (e5): Across all LLMs, increasing the number of retrieved passages initially improves performance, but then leads to a sharp decline or plateau.
2) Weak Retriever (BM25): Performance generally exhibits a continuous increase or a slight decrease as the number of retrieved passages increases.

The effect of retrieved context size on RAG performance



(a) RAG performance with e5 retriever



(b) RAG performance with BM25 retriever

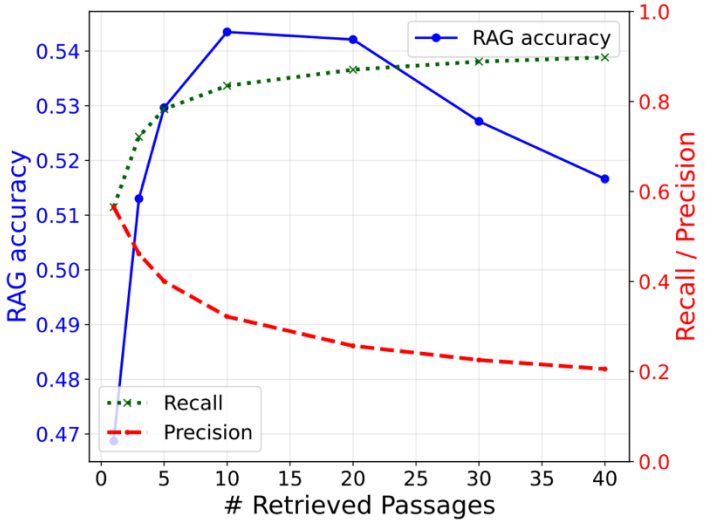
Insights

1) The effectiveness of increasing retrieved context size in RAG depends on the strength of the retriever.
2) With a strong retriever, performance exhibits an “inverted-U pattern”, while a weak retriever shows more consistent, albeit potentially limited, improvement.
3) This suggests that factors beyond simply the amount of retrieved information are at play.

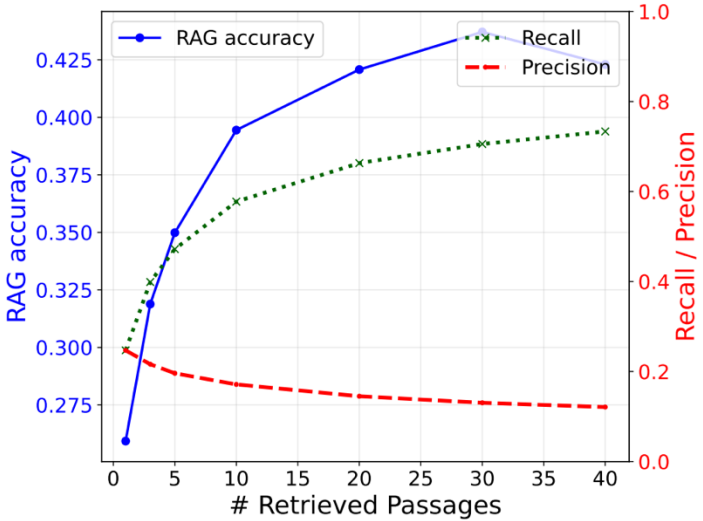
The interplay of retrieval quality and LLM capabilities

Research question: Do the observed performance bottlenecks originate from limitations in the retriever's ability to identify relevant information, or from the long-context LLM's capacity to effectively utilize the retrieved context?

- Retriever: e5 (on the left), bm25 (on the right) ; Generator: Gemma-2-9B-chat.
- Recall@k (Hit@k): if **relevant** document appears in the retrieved documents.
- Precision@k: how much **noise** are there in the context.

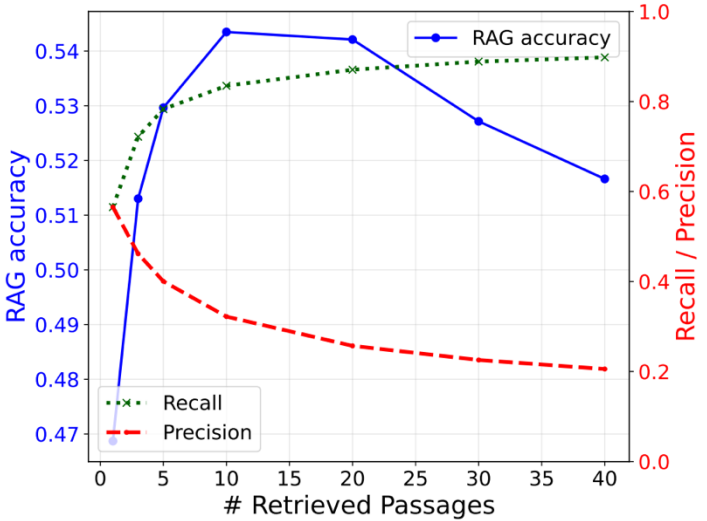


(a) Retrieval with e5 retriever

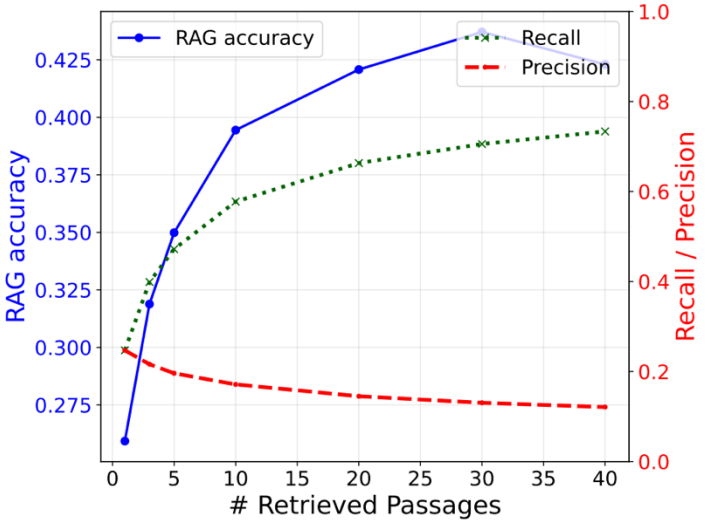


(b) Retrieval with BM25 retriever

The interplay of retrieval quality and LLM capabilities



(a) Retrieval with e5 retriever

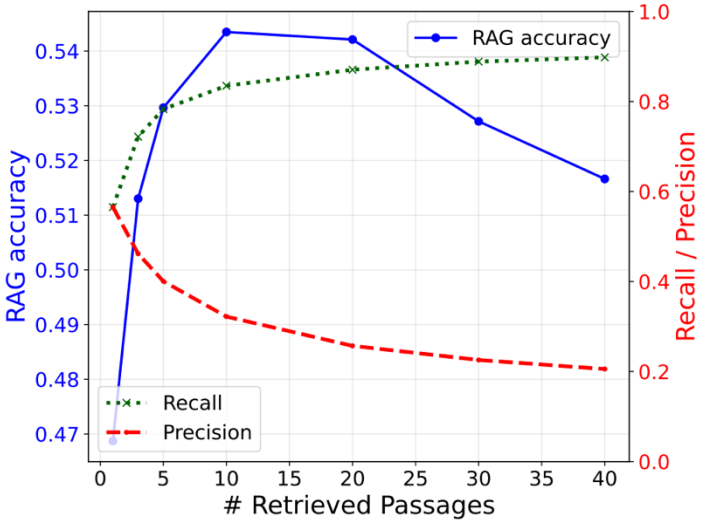


(b) Retrieval with BM25 retriever

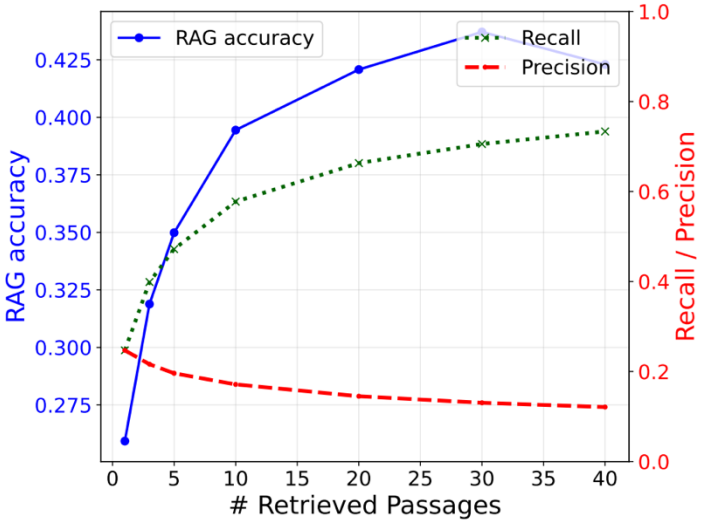
Observations

- Increasing the number of retrieved passages leads to higher recall but lower precision.
- The overall accuracy of the RAG system falls below the recall across all retrieval sizes. -> noise matters
- Despite exhibiting higher precision, the e5 retriever leads to a more pronounced performance degradation as the number of retrieved passages increases compared to BM25.

The interplay of retrieval quality and LLM capabilities



(a) Retrieval with e5 retriever



(b) Retrieval with BM25 retriever

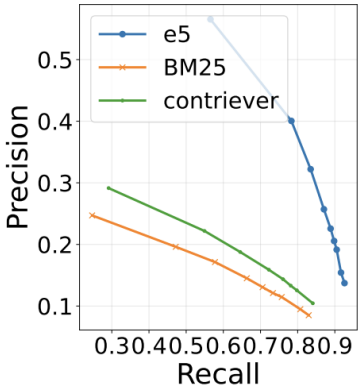
Insights

- **Influence of irrelevant passages**: The discrepancy between retrieval recall and RAG accuracy underscores the detrimental effect of irrelevant retrieved passages ("hard negatives") on the LLMs' performance.
- **Limitations of precision as a metric**: The contrasting performance trends observed with e5 and BM25, despite the former's higher precision, reveal that precision alone is an inadequate measure of retrieval quality in this context, when the end-to-end performance is considered.

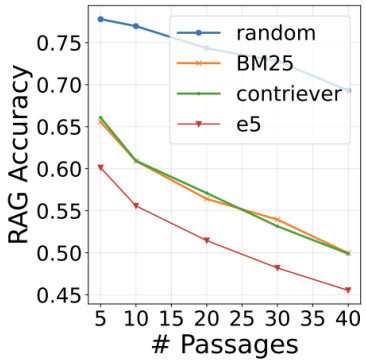
The importance of hard negatives for long-context LLM evaluation

Research question: (1) How robust are current long-context LLMs to these hard negatives? and (2) Does the impact of hard negatives vary with the retriever used?

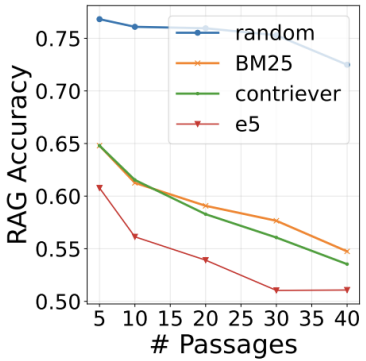
- Setting: one golden document + multiple negatives from different retrievers
- Retriever: random, bm25, contriever, e5
- Generator: Gemma-2-9B-chat, Mistral-Nemo-12B-instruct, Gemini-1.5-pro



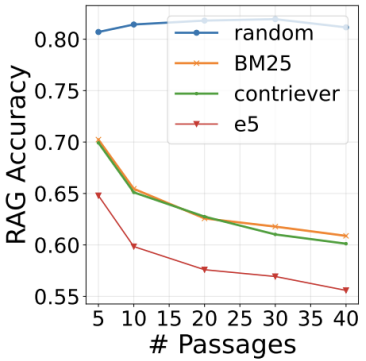
(a) Retrievers



(b) Gemma2-9B-Chat

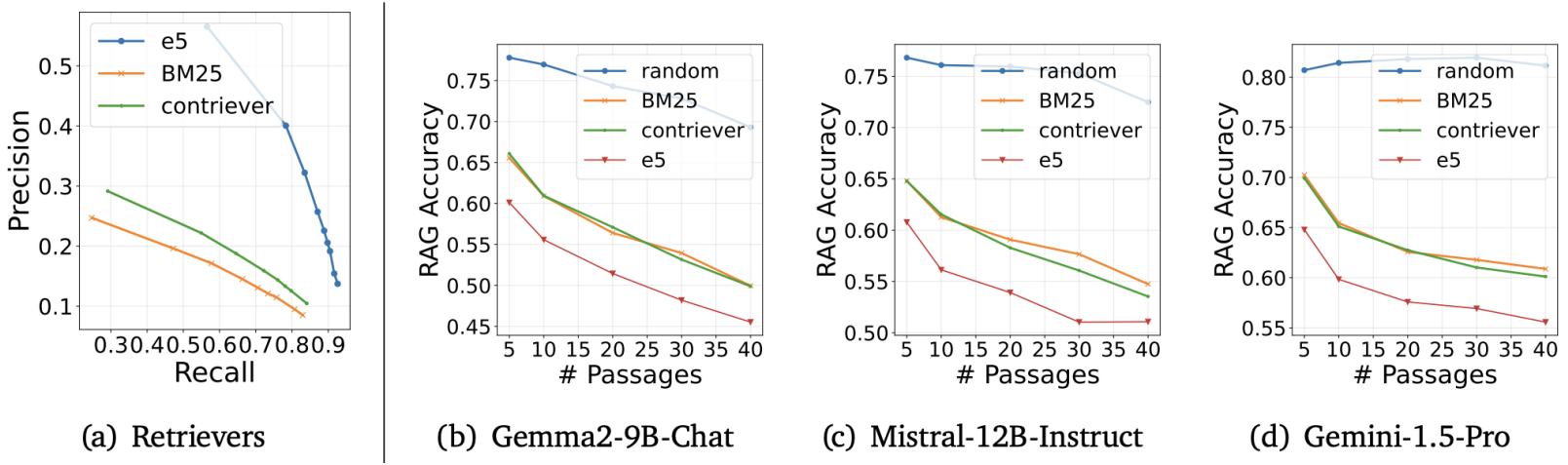


(c) Mistral-12B-Instruct



(d) Gemini-1.5-Pro

The importance of hard negatives for long-context LLM evaluation

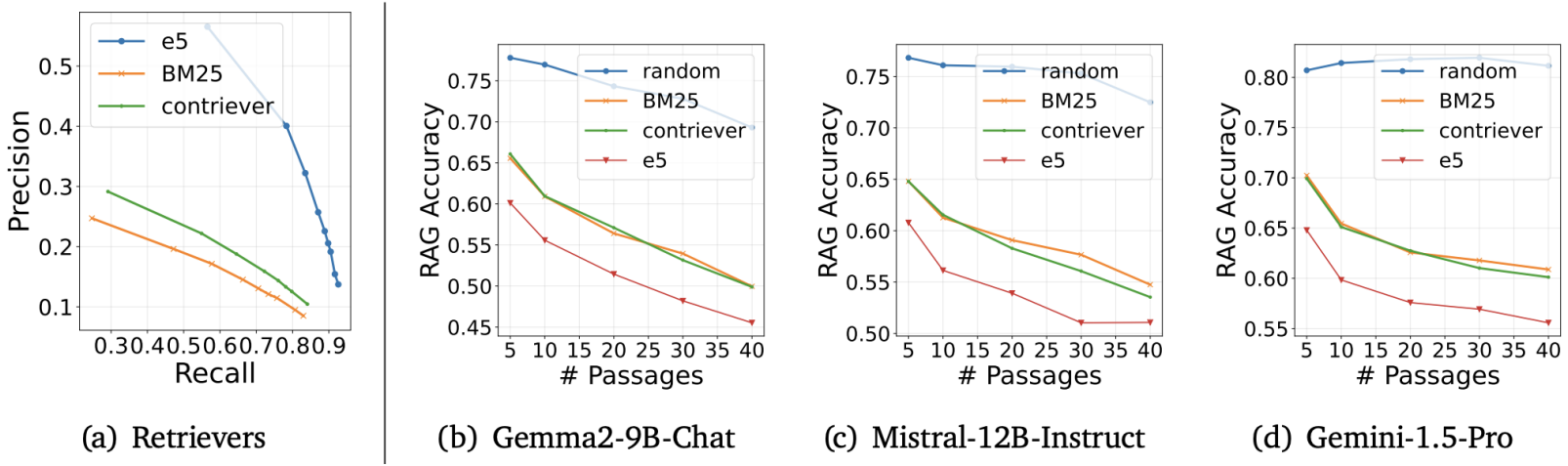


(a) Retrievers (b) Gemma2-9B-Chat (c) Mistral-12B-Instruct (d) Gemini-1.5-Pro

Observations

- **Sensitivity to hard negatives**: Across all LLMs, increasing the number of hard negative passages generally leads to a decline in RAG answer accuracy.
- **Retriever strength and hard negative difficulty**: The strength of the retriever directly correlates with the difficulty of the retrieved hard negatives. LLMs struggle more with hard negatives from stronger retrievers.
- **Distinguishing random and hard negatives**: While Gemini-1.5-Pro demonstrates robustness to random negatives, it remains susceptible to the influence of hard negatives.

The importance of hard negatives for long-context LLM evaluation



(a) Retrievers (b) Gemma2-9B-Chat (c) Mistral-12B-Instruct (d) Gemini-1.5-Pro

Insights

- Existing benchmarks for evaluating long-context LLMs, such as "needle-in-the-haystack" (Kamradt, 2023) and RULER (Hsieh et al., 2024a), predominantly utilize random negatives.
- Our findings demonstrate that such benchmarks may not adequately capture the challenges posed by hard negatives, which are prevalent in real-world RAG applications.

Simple and effective training-free RAG improvement

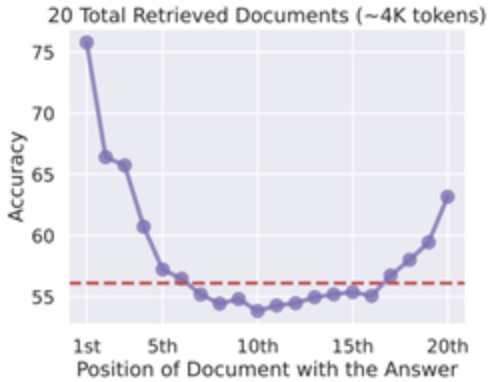
Retrieval reordering

- ”Lost-in-the-middle”: LLMs exhibit a tendency to prioritize information presented at the beginning and end of an input sequence, while paying less attention to the middle.
- Retrieval reordering leverages the inherent "lost-in-the-middle" phenomenon observed in LLMs to mitigate the negative effects of hard negatives.

Given a query q and a set of retrieved passages $d_1, d_2, \ldots, d_k$ with decreasing relevance scores:

$$[I, d_1, d_2, \ldots, d_{k-1}, d_k, q] \longrightarrow [I, d_1, d_3, \ldots, d_4, d_2, q]$$

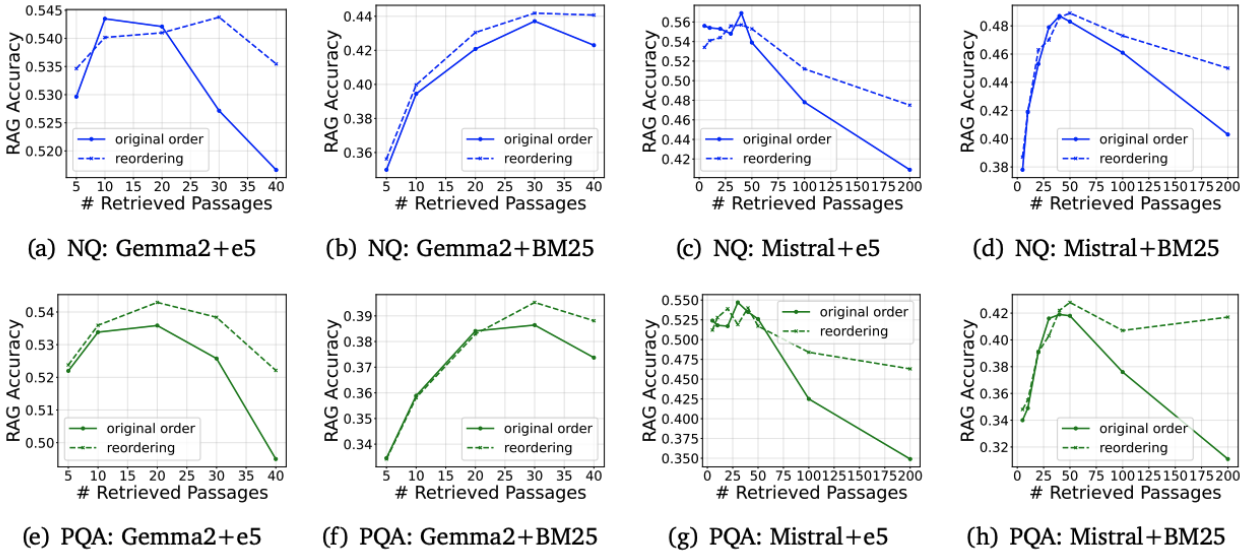
$$Order(d_i) = \begin{cases} \frac{i+1}{2} & \text{if } \mathrm{mod}(i, 2) = 1 \\ (k+1) - \frac{i}{2} & \text{if } \mathrm{mod}(i, 2) = 0 \end{cases}$$



Simple and effective training-free RAG improvement

Retrieval reordering

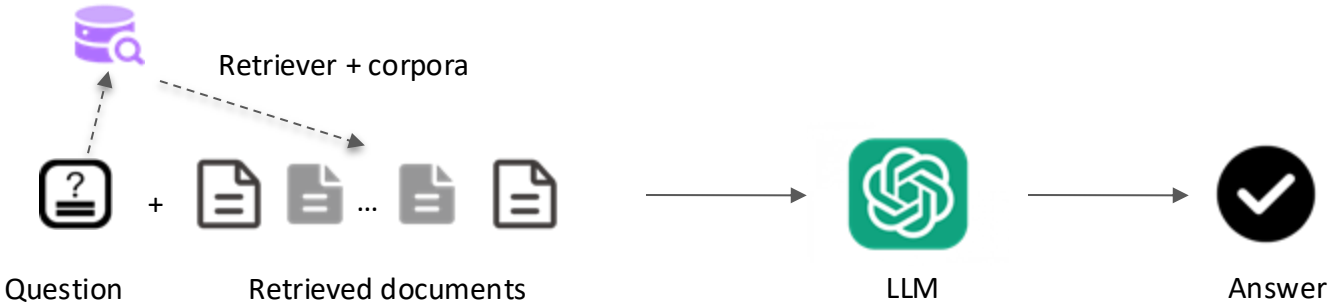
- ”Lost-in-the-middle”: LLMs exhibit a tendency to prioritize information presented at the beginning and end of an input sequence, while paying less attention to the middle.
- Retrieval reordering leverages the inherent "lost-in-the-middle" phenomenon observed in LLMs to mitigate the negative effects of hard negatives.



Improving Robustness for RAG via Data-Augmented Fine-Tuning

Implicitly improving LLM robustness through fine-tuning

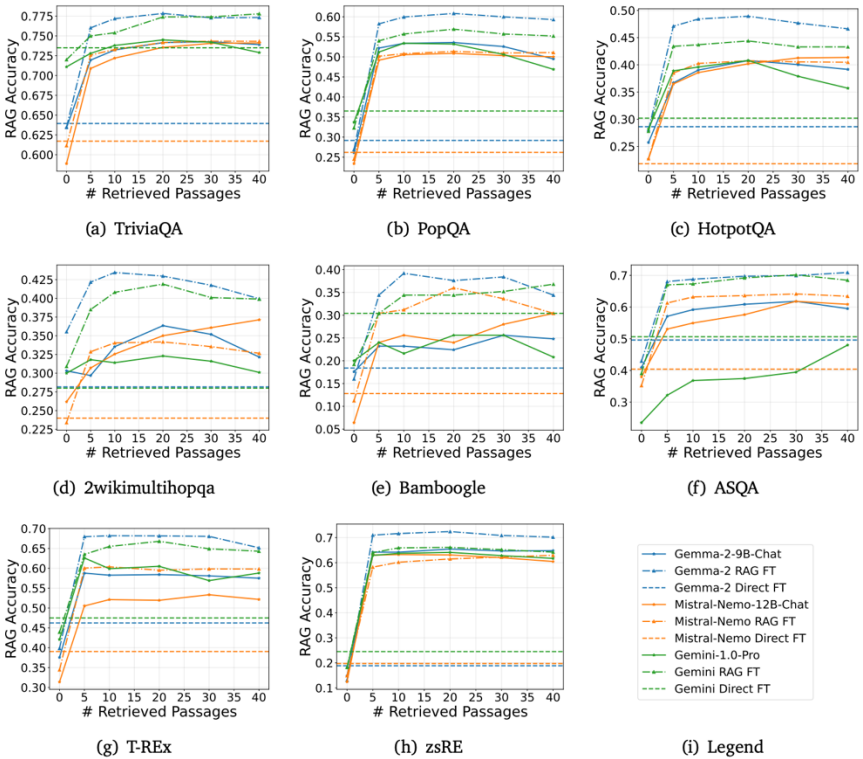
Finetune the LLM to obtain knowledge



Finetune the long context LLM to be robust to hard negatives

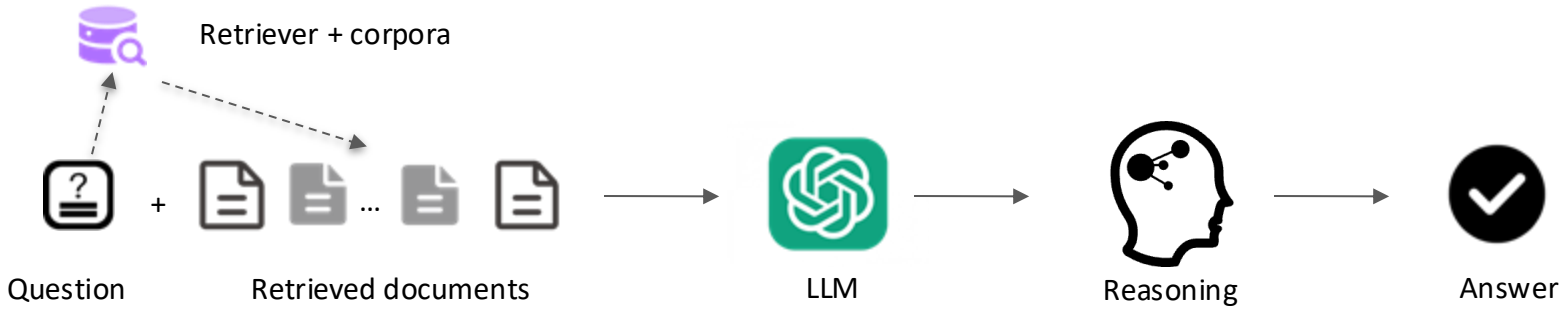
Improving Robustness for RAG via Data-Augmented Fine-Tuning

Implicitly improving LLM robustness through fine-tuning

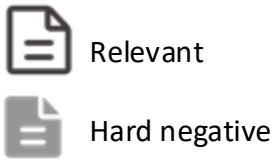


Improving Robustness for RAG via Data-Augmented Fine-Tuning

Enhancing relevance identification through reasoning augmentation

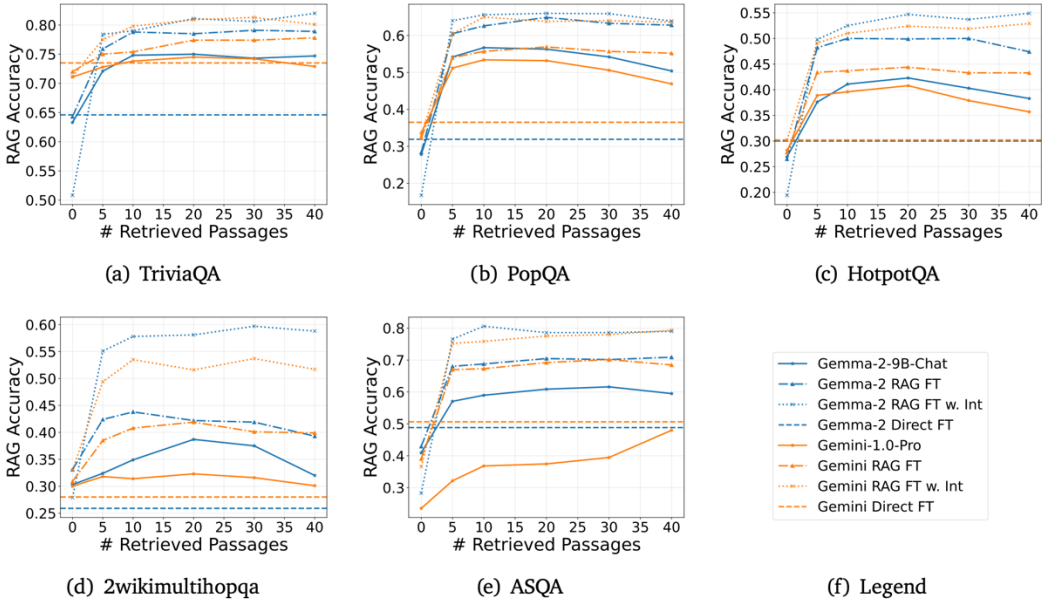


Finetune the long context LLM to reason to identify hard negatives



Improving Robustness for RAG via Data-Augmented Fine-Tuning

Enhancing relevance identification through reasoning augmentation



Improving Robustness for RAG via Data-Augmented Fine-Tuning

Case study

Question	Which English chemist discovered the most elements?
Ground Truth	Humphry Davy
Retrieved Passages	**Doc 1** (Title: Chemistry) J.J. Thomson of Cambridge University discovered the electron and soon after the French scientist Becquerel as well as the couple Pierre and Marie Curie investigated the phenomenon of radioactivity. In a series of pioneering scattering experiments Ernest Rutherford at the University of Manchester discovered the internal structure of the atom and the existence of the proton, classified and explained the different types of radioactivity and successfully transmuted the first element by bombarding nitrogen with alpha particles. His work on atomic structure was improved on by his students, the Danish physicist Niels Bohr and Henry Moseley. The electronic theory of chemical ... **Doc 2** (Title: Chemistry) weights. The development of the electrochemical theory of chemical combinations occurred in the early 19th century as the result of the work of two scientists in particular, J.J. Berzelius and Humphry Davy, made possible by the prior invention of the voltaic pile by Alessandro Volta. Davy discovered nine new elements including the alkali metals by extracting them from their oxides with electric current. British William Prout first proposed ordering all the elements by their atomic weight as all atoms had a weight that was an exact multiple of the atomic weight of hydrogen. J.A.R. Newlands devised an early table of ... ... **Doc 40** (Title: "History of chemistry") many had already applied the concept of atomism to chemistry. A major example was the ion theory of Svante Arrhenius which anticipated ideas about atomic substructure that did not fully develop until the 20th century. Michael Faraday was another early worker, whose major contribution to chemistry was electrochemistry, in which (among other things) a certain quantity of electricity during electrolysis or electrodeposition of metals was shown to be associated with certain quantities of chemical elements, and fixed quantities of the elements therefore with each other, in specific ratios. These findings, like those of Dalton's combining ratios, were early clues ...

Model Predictions	
Gemma-2-9B-Chat	William Hyde Wollaston
Implicit RAG FT	William Ramsay
RAG FT with Reasoning	The best document to answer this question is **Doc 2(Chemistry)**. This document states that **Humphry Davy** "discovered nine new elements including the alkali metals by extracting them from their oxides with electric current." Therefore, the answer is *Humphry Davy*. While other documents mention chemists and their discoveries, none of them contradict or provide a higher number of discovered elements than Davy's nine. This makes Doc 2 the most useful in identifying the English chemist who discovered the most elements. It's important to note that this answer is based on Doc 2 alone. If other documents mentioned an English chemist who discovered more than nine elements, they would supersede the information in Doc 2. However, without that information, we can confidently say that based on these documents, Humphry Davy discovered the most elements among the English chemists mentioned. **Conclusion:** Based on the given documents, *Humphry Davy* is the English chemist who discovered the most elements.

Take aways

Analysis

- Increasing the number of retrieved documents in RAG can result in an inversed-U performance curve.
- This is potentially because of the hard negatives retrieved by the retriever.
- Retrieval precision is not an accurate metric to measure the amount of noise.
- The stronger the retriever is, the harder its retrieved negatives are.

Methodology

- Simply reorder the retrieved documents work well with a large number of retrieved document in RAG.
  - Efficient, effective, plug-and-play, retriever-agnostic, LLM-agnostic
- Implicit RAG-specific LLM tuning
  - Improved performance, generalizable to unseen dataset
- RAG-specific LLM tuning with intermediate reasoning
  - Best performance, expensive to get the reasoning labels

Thank you!