On the Influence of Context Size and Model Choice in Retrieval-Augmented Generation Systems

Juraj Vladika and Florian Matthes
Technical University of Munich
School of Computation, Information and Technology
Department of Computer Science
Garching, Germany
{juraj.vladika, matthes}@tum.de

Abstract

Retrieval-augmented generation (RAG) has emerged as an approach to augment large language models (LLMs) by reducing their reliance on static knowledge and improving answer factuality. RAG retrieves relevant context snippets and generates an answer based on them. Despite its increasing industrial adoption, systematic exploration of RAG components is lacking, particularly regarding the ideal size of provided context, and the choice of base LLM and retrieval method. To help guide development of robust RAG systems, we evaluate various context sizes, BM25 and semantic search as retrievers, and eight base LLMs. Moving away from the usual RAG evaluation with short answers, we explore the more challenging long-form question answering in two domains, where a good answer has to utilize the entire context. Our findings indicate that final QA performance improves steadily with up to 15 snippets but stagnates or declines beyond that. Finally, we show that different general-purpose LLMs excel in the biomedical domain than the encyclopedic one, and that open-domain evidence retrieval in large corpora is challenging.

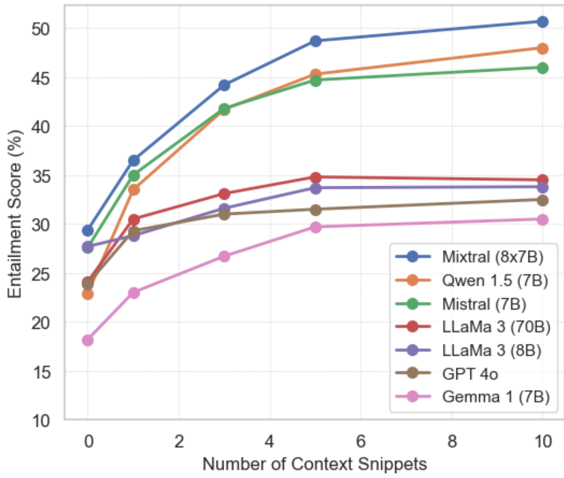

Figure 1: The influence of the number of context snippets passed to the RAG system on the final performance (entailment score) on a biomedical task BioASQ-QA. The performance improves steadily for all models, to a differing extent, and then stagnates after saturation.

1 Introduction

The field of Natural Language Processing (NLP) has been vividly transformed with the advent of large language models (LLMs), massive models that excel on a wide range of complex tasks, including text generation, question answering, and summarization (Zhao et al., 2023). Despite their impressive performance, LLMs have certain limitations. The static nature of the knowledge encoded within their weights can lead to providing outdated content as new information emerges (Zhang et al., 2023). Furthermore, LLMs can generate plausible sounding but factually incorrect responses (*hallucinations*), as they lack a reliable mechanism to verify the accuracy of the information they produce (Ji et al., 2023). Finally, they can lack specialized knowledge related to advanced expert domains.

To address these shortcomings, the concept of Retrieval-Augmented Generation (RAG) has shown great potential (Lewis et al., 2020). RAG systems enhance the capabilities of LLMs by integrating a retrieval component that allows the model to dynamically utilize external knowledge sources during the generation process. By retrieving relevant information from a curated corpus or the web in real-time, RAG models can produce more accurate, up-to-date, and contextually appropriate responses (Fan et al., 2024). RAG systems have also seen wide adoption in various industry branches, where companies leverage them to build tools for accessing their internal documentation via questions posed in human language (Xu et al., 2024).

Despite their increasing popularity and use, there here have been few studies that systematically ex-

Findings of the Association for Computational Linguistics: NAACL 2025, pages 6739–6751

plore different settings of RAG systems, including the size of the provided context, choice of base LLM, and choice of retriever technique (sparse or dense). While recent work has shown that essential information in long context blocks can get "lost in the middle" (Liu et al., 2024) or be affected by noisy context (Cuconasu et al., 2024), most of these studies work with short, factoid question answering (with questions like "*Who won the 2024 Nobel Peace Prize?*") and assume there is one gold context snippet relevant for the answer. There has been less research on how do LLMs use the context for long-form QA, where a holistic final answer has to include multiple or even all context snippets.

To bridge this research gap, our study aims to explore and evaluate various configurations of RAG systems. We systematically investigate how different context sizes, retrieval strategies, and base LLMs impact the performance of RAG systems. We evaluate these parameters through the prism of the generative question-answering task in two different domains: the biomedical BioASQ-QA task (Krithara et al., 2023) and the encyclopedic QuoteSum dataset (Schuster et al., 2024). Both datasets provided the essential resources for our study – inclusion of *gold evidence snippets* and *human-written answers* utilizing these snippets to answer the questions. By conducting a series of experiments and analyses, we aimed to identify best practices for the implementation of RAG systems.

Our contributions include:

- We examine the influence of the number of context snippets in the prompt on the final task performance of the RAG system. We observe the performance to steadily improve from 1 to about 10-15 snippets, but then stagnate or even decline by 20-30 snippets.

- We test the performance of different LLMs of various sizes in their ability to utilize the context snippets for generating accurate answers. The results show Mistral and Qwen to perfrom the best on the biomedical task, while GPT and Llama excel on the encyclopedic task.

- We test the open-domain setting, where gold evidence is not known and has to be retrieved from large knowledge bases. We evaluate two different retrievers and show the impact on final performance. We show that this setting is very challenging and performance is far from the gold setting, with the BM25 optimizing for precision, while semantic search gives a wider coverage of retrieved information.

We make our code available in a public repository on GitHub.[1]

2 Related Work

2.1 Retrieval-Augmented Generation

Early approaches to RAG involved simple retrieval and were developed for the task of question answering (Chen et al., 2017). Recent advancements have seen more sophisticated integration of retrieval and generation processes, thereby significantly enhancing the quality and relevance of the generated text (Lewis et al., 2020). These advancements have been facilitated by improvements in both the retrieval mechanisms, which have become more efficient and effective at finding relevant information, and the generative models, which have become better at integrating and contextualizing the retrieved information (Cai et al., 2022).

A recent survey by Gao et al. (2024) separates RAG approaches into *naive RAG* and *advanced RAG*. The naive RAG approach follows a traditional process that includes indexing, retrieval, and generation, also called a "Retrieve-then-Read" framework (Zhu et al., 2021). On the other hand, advanced RAG introduces specific improvements to enhance the retrieval quality by employing pre-retrieval and post-retrieval strategies. Pre-retrieval strategies include query rewriting with an LLM (Ma et al., 2023) or query expansion methods like HyDE (Gao et al., 2023). Post-retrieval methods focus on selecting essential information from retrieved documents. This includes reranking the retrieved documents with neural models (Glass et al., 2022) or summarizing the retrieved documents before passing them as context (An et al., 2021).

2.2 Context and Noise in RAG Systems

A lot of recent work has explored how to improve RAG and make it more accurate and robust to imperfect context. This includes fact verification (Li et al., 2024), self-reflection with critique (Asai et al., 2024), learning to re-rank the context (Yu et al., 2024), improved answer attribution (Vladika et al., 2024a), adaptive search strategy (Jeong et al., 2024), and relevance modeling (Wang et al., 2024).

There have also been studies exploring the size of input context and its influence on the performance of RAG systems. Liu et al. (2024) highlight

[1] https://github.com/jvladika/ContextRAG

the effect of information being *lost in the middle*, showing how RAG mostly focuses on the beginning and the ending of the provided context. Similarly, Cuconasu et al. (2024) examine the influence of the position of the most relevant snippet in the context and the influence of noisy snippets on the performance. Both of these studies work with factoid QA dataset where it is assumed one context snippet is the most important for the answer.

Xiong et al. (2024) analyze the effect of number of context snippets on five multiple-choice biomedical QA tasks, while Vladika and Matthes (2024) analyze the impact of the number of snippets as well as context recency and popularity for biomedical QA. Chen et al. (2024a) evaluated the noise robustness and context integration of different LLMs for RAG. Most similar to our work is the study by Hsia et al. (2024), where the influence of different RAG components is tested with eight LLMs and it also includes BioASQ as a benchmark dataset.

While these studies have discovered important principles in context inclusion for RAG systems, they predominantly evaluate it on multiple-choice or short-form QA tasks where there is one clear answer and one most important context snippet. Our work evaluates generative question answering where potentially all snippets could be relevant for inclusion in the answer, which is a more challenging setting. Additionally, we provide a comprehensive evaluation of three main RAG components: the influence of the context size, different retrieval techniques, and choice of base LLMs.

3 Foundations

3.1 RAG System for Question Answering

Typically, a RAG system consists of a *retriever* and a *reader*. Retriever has to search and collect relevant evidence snippets that are passed as *context* inside of a prompt to the reader. Our study investigates the importance of those three aspects (retriever, context, reader) on the final performance of the whole system. We first focus on the influence of context size on the readers' QA capability, followed by the importance of choosing the reader by comparing different base LLMs on the task, and finally, we test the influence of two different retrievers on the final QA performance (BM25 and semantic search). To formally define: Given a question q and context c consisting of context snippets $c_1, c_2, ..., c_n$, the goal is to generate an answer a with a model *reader(q, c) = a*. The context c is provided in the first experiment, but in an open-domain setting, given a document corpus D with documents $d_1, d_{''}, ...d_N$, the idea is for a retriever to *retrieve(q, D) = d1, d2* best matching documents and then from them extract context snippets.

3.2 Datasets

BioASQ-QA (Krithara et al., 2023) is a biomedical question answering (QA) benchmark dataset in English. It has been designed to reflect real information needs of biomedical experts. The questions are written by biomedical experts and the evidence corpus used to answer them is PubMed (White, 2020), the large database of biomedical research papers. The dataset is a part of the ongoing shared challenge, and we use the 2023 version, Task 10b. While the full dataset contains various types of questions (yes/no, factoid, lists), we utilize only the so-called *summary* questions – questions paired with human-selected evidence snippets from PubMed abstracts and human-written "ideal answers", which are essentially natural language summaries of the provided snippets. In total, there are 1130 summary questions.

QuoteSum (Schuster et al., 2024) is a dataset of encyclopedic questions, relevant passages, and human-written semi-extractive answers. The questions are human-written and are paired with up to 8 passages (evidence snippets) from Wikipedia. These passages are used as the main source by annotators to write the answers. QuoteSum contains 805 instances and covers various domains such as geography, history, arts, and technology. An example question is "*Why was Stonehenge built in the first place?*".

These datasets contain the gold evidence snippets and human-written answers based on the snippets, making them a suitable testbed for our study. While BioASQ might be difficult given its language rich with complex biomedical terminology, the main challenge is in successfully utilizing the given context and summarizing it into a concise but informative answer. We intentionally do not benchmark on any biomedical LLM to not give any model a possible advantage.

4 Experiment

4.1 Models

We conduct our experiments using a multitude of different LLMs that serve as readers, i.e., the models reading and comprehending the context and

# docs	GPT 3.5			GPT 4o			LLaMa 3 (70B)			Mixtral (8x7B)		
	R-L	BSc	Ent.%	R-L	BSc	Ent.%	R-L	BSc	Ent.%	R-L	BSc	Ent.%
	BIOASQ											
0	23.2	87.1	22.5	23.8	87.0	23.9	22.9	86.9	24.1	21.8	85,8	29.4
1	28.0	87,9	29.3	28.2	87.9	29.3	28.3	87,8	30.5	29.6	87.9	36.5
3	30.9	88.5	31.4	31.1	88.4	31.0	31.4	88.4	33.1	34.8	88.9	44.2
5	31.9	88.6	32.0	31.9	88.6	31.5	32.0	88.5	**34.8**	36.4	89.0	48.7
10	**32.7**	**88.8**	**32.6**	**32.8**	**88.8**	**32.5**	**32.2**	**88.6**	34.5	**37.7**	**89.2**	**50.7**
	QUOTESUM											
0	27.2	85.0	20.4	26.9	84.9	20.3	26.3	84.4	21.8	22.3	83.6	15.2
1	36.4	87.1	41.7	36.6	87.1	41.7	34.3	86.5	36.5	33.8	86.1	32.0
3	39.0	87.6	42.9	39.1	87.7	43.1	37.4	87.5	40.5	37.2	86.8	35.2
5	**39.9**	87.7	44.0	**39.7**	87.7	**44.2**	38.4	87.5	**41.9**	37.4	86.9	**36.1**
10	39.7	**87.7**	**44.2**	39.6	**87.7**	43.4	**39.4**	**87.5**	41.8	**37.7**	**86.9**	35.9

Table 1: Results of final QA performance on BioASQ and QuoteSum for different number of **gold snippets**, using the **four big LLM**s as readers: GPT 3.5, GPT 4o, Mixtral (8x7B), LLaMa 3 (70B). The results are measured with ROUGE-L (R-L), BERTScore (BSc), and average entailment prediction of the NLI model (Ent.%).

then generating the answers from it. The experiments were mainly conducted in June 2024 and reflect the up-to-date LLM landscape of that time.

We start with **GPT** as a commercial state-of-the-art LLM in our comparison since it has demonstrated remarkable zero-shot performance on various NLP tasks. Consequently, it is often used as a benchmark for comparing LLMs. We use the **GPT-3.5** (Turbo-0125) as the standard ChatGPT version, and also **GPT-4o** (Turbo-0513), the updated "omni" version of GPT-4 (OpenAI, 2023), which was shown to improve the performance. We then also include two popular open-weights models that achieved impressive performance, namely **Mixtral** (8x7B) (Jiang et al., 2024), based on a sparse mixture-of-experts architecture (Fedus et al., 2022); and **LLaMa 3** (70B) (AI@Meta, 2024), a powerful staple model from Meta. All models are instruction-tuned ("chat") versions.

For the smaller language models, we choose **Mistral-7B** (Instruct-v3) (Jiang et al., 2023), the smaller counterpart to Mixtral; then the **Gemma** (1) (Mesnard et al., 2024), a lightweight open model from Google built from the research and technology used to create Gemini models (Gemini, 2024); and the smaller, 8B version of **LLaMa 3**. We additionally benchmark **Qwen 1.5 (7B)** (Chat), another recently popular and powerful language model (Bai et al., 2023). All of these models are open-source models, and we use the instruction-tuned versions.

4.2 Setup

We use the same prompt and setup for all of the benchmarked models:

> Give a simple answer to the question based on the context.
> QUESTION:
> <the current question>
> CONTEXT:
> [snippet$_1$, snippet$_2$, ..., snippet$_n$]
> ANSWER:

For the internal-knowledge setting with no context, the instruction was changed accordingly to *Give a simple answer to the question based on your best knowledge.* and the *CONTEXT* part removed. While it would have been an interesting experiment to also give the LLMs few-shot examples of QA pairs, we intentionally opt for this zero-shot setting so that the focus of the experiments lies solely on the utilization of provided context for answering and not on potential in-context learning abilities.

GPT models were prompted through the OpenAI API, while all of the open-source models were queried with API calls through the Together AI service[2] platform, which hosts many popular open-source models. We set the token limit to 512 and the temperature parameter to 0, maximizing deterministic generation by favoring high-probability words and thus ensuring reproducibility of the results. One run through the whole dataset with five settings took two computation hours. For embedding models, we used one NVidia V100 GPU card with 16 GB of VRAM.

4.3 Experiment Rounds

Context Size and Reader Performance. The first round of experiments consisted of varying the number of context snippets passed in the prompt and observing how the QA performance changes.

[2] Together AI: https://docs.together.ai/

# docs	Gemma (7B)			LLaMa 3 (8B)			Mistral (7B)			Qwen (7B)		
	R-L	BSc	Ent.%	R-L	BSc	Ent.%	R-L	BSc	Ent.%	R-L	BSc	Ent.%
	BIOASQ											
0	20.7	79.7	18.2	21.1	85.2	27.7	20.6	85.4	27.6	20.3	85.5	22.9
1	25.9	85.7	23.0	28.5	87.8	28.8	28.7	87.9	35.0	28.5	87.7	33.5
3	30.6	86.7	26.7	32.2	88.2	31.6	33.1	88.7	41.8	33.4	88.9	41.7
5	32.2	87.0	29.7	36.7	**88.5**	33.7	34.7	88.9	44.7	35.2	89.0	45.3
10	**33.5**	**87.2**	**30.5**	**37.3**	88.4	**33.8**	**36.4**	**89.2**	**46.0**	**36.8**	**89.1**	**48.0**
	QUOTESUM											
0	8.7	67.4	9.8	24.0	83.6	18.5	25.2	83.9	13.6	25.2	84.3	17.3
1	15.8	54.6	13.7	33.8	86.4	37.2	34.5	86.6	35.4	35.7	86.9	39.6
3	25.2	**77.1**	22.0	38.8	87.0	38.3	37.4	87.2	38.0	39.0	87.4	42.2
5	**24.8**	76.4	22.4	39.4	87.1	37.7	38.6	**87.3**	39.3	39.3	87.5	**43.0**
10	24.7	76.1	**22.4**	**39.7**	**87.1**	**38.0**	**39.0**	87.2	**39.5**	**39.3**	**87.5**	42.9

Table 2: Results of final QA performance on BioASQ and QuoteSum for different number of **gold snippets**, using the **four small LLMs** as readers: Gemma (7B), LLaMa 3 (8B), Mistral v3, and Qwen 1.5 (7B). The results are measured with ROUGE-L (R-L), BERTScore (BSc), and average entailment prediction of the NLI model (Ent.%).

We pass the gold snippets in this experiment since they are known to us and the focus is only on the quantity (context size). While it seems intuitive that adding more snippets will improve the final scores, because an answer based on partial information will be incomplete, we wanted to test: (1) to which extent do different LLMs utilize the provided context, and (2) when do the LLMs get saturated with too much context, leading to stagnation or decline.

As a starting point, we first pose the question to LLMs with no context (0 snippets), thus testing their *internal knowledge recall*. While an interesting research caveat on its own, we use it here just as a baseline. Afterward, we vary the numbers of context snippets in the array of 1, 3, 5, 10; to give the idea of a general trend. In case a question has fewer snippets than the given k, then all of the snippets for that question were used. In BioASQ, more than 80% of all questions have at least 3 snippets, around 60% at least 5 snippets, and around one third at least 10 snippets – using more than 10 wouldn't make a lot of sense given that only 18% of questions have more than 10 snippets. Similarly, QuoteSum has around 75% questions with at least 3 snippets, 50% with at least 5, and 30% have the maximum 8 snippets (labeled "10" in tables for consistency).

Closed Retrieval Apart from the easier setting where gold snippets are provided to the model, we also explore the more challenging setup with evidence retrieval "in the wild" (Chen et al., 2024b). In this case, the RAG system first has to retrieve the evidence from a knowledge base before producing an answer based on it. For the closed retrieval setting in BioASQ, we only use the PubMed documents required to answer its 1130 questions. This results in around 8000 documents as a knowledge base. This mimics the common RAG use case in the industry where one would be working with a limited knowledge corpus containing internal company documents. The abstracts are saved in a vector database and embedded with a sentence embedding model (we use the biomedical *S-PubMedBERT-MS-MARCO*[3] from Deka et al., 2022). Afterward, the top 10 most similar documents to the question are retrieved (using cosine similarity and the embedding model), and the most similar sentence from each document is selected as an evidence snippet. The amount of selected sentences/documents is also varied with amounts 1, 3, 5, 10; to align with the first round of experiments. For QuoteSum, we omit this experiment as the subset of Wikipedia articles is not provided.

Open Retrieval. Finally, we test the QA performance of both datasets in the most challenging open setting – using a large knowledge base where the retriever first has to sift through millions of documents to discover the most relevant ones. For BioASQ, we test two different retrievers – semantic search with the same sentence embedding model as in last round (*S-PubMedBERT-MS-MARCO*) and the sparse retrieval technique BM25, which has a long-established track record of good performance for information retrieval (IR) tasks.

For BioASQ, we use MEDLINE, a snapshot of currently available abstracts in PubMed that is updated once a year. We used the 2022 version found on the official website.[4] We filter it to a 10-year

[3] https://huggingface.co/pritamdeka/S-PubMedBert-MS-MARCO

[4] https://www.nlm.nih.gov/databases/download/pubmed_medline.html

# docs	GPT 3.5			GPT 4o			LLaMa 3 (70B)			Mixtral (8x7B)		
	R-L	BSc	Ent.%	R-L	BSc	Ent.%	R-L	BSc	Ent.%	R-L	BSc	Ent.%
0	23.2	87.1	22.5	23.8	87.0	23.9	22.9	86.9	24.1	21.8	85.8	29.4
1	24.2	87.8	25.2	24.5	87.2	25.4	24.0	86.9	25.8	25.7	87.0	30.1
3	32.7	88.0	27.4	34.5	87.9	27.2	28.1	87.7	27.7	30.3	87.8	36.1
5	30.5	88.1	28.6	30.0	88.2	28.8	29.3	88.0	30.0	31.6	87.9	39.9
10	**32.0**	**88.7**	**31.4**	**31.8**	**88.6**	**31.0**	**31.0**	**88.3**	**32.1**	**32.9**	**88.2**	**44.4**

Table 3: Results of final QA performance on BioASQ for different number of **retrieved context** snippets in the **closed retrieval** setting (with a corpus of 8 thousand PubMed documents), using the four big LLMs. The results are measured with ROUGE-L (R-L), BERTScore (BSc), and average entailment prediction of the NLI model (Ent.%).

span from 2012 to 2022, following BioASQ's time range – this results with 10.6 million abstracts in total. For QuoteSum, we use Wikipedia since the dataset is based on it. We query the Wikipedia search API directly through a link.[5] Wikipedia search is based on BM25.[6] While a popular way to benchmark retrievers is using common IR metrics like *recall@k*; we focus only on benchmarking the final QA performance, as this both keeps it consistent with the previous experiments and also highlights the fact that the final answer is the most important artefact of a QA system.

Unlike gold snippets where we often only had up to 10 provided in the original dataset, the open setting allows to keep increasing the number of snippets indefinitely. Therefore, we additionally evaluate with 15, 20, 30 snippets, to test the effect of context saturation.

4.4 Evaluation

To evaluate the quality of the generated answers, we use three main metrics. Given that the dataset contains ideal answers, we can use reference-based metrics. Evaluating LLMs for long-form QA is a challenging, ongoing research problem, and no metric is ideal (Xu et al., 2023). Still, we cover a variety of metrics, to gain an overview.

The first metric is **ROUGE** (Lin, 2004),[7] which looks at the recall between the reference answer and the generated answer. Specifically, we use the ROUGE-L, which looks at the longest overlapping sequence between the reference and generated answer. Since this metric focuses solely on lexical overlaps, we use two additional semantic metrics. We also apply the **BERTScore** metric, which captures semantic similarity by using the BERT model's embeddings (Zhang et al., 2020).

The third metric utilizes the concept of natural language inference (**NLI**), by using the reference answer as the hypothesis and the generated answer as the premise. The intuition behind this approach is that a good answer should logically entail the reference answer. Using NLI this way has been done to evaluate the quality of summaries and text generation (Laban et al., 2022). We use the model DeBERTa-v3 (He et al., 2023), which was shown to work well with NLI and reasoning tasks We use the version *Tasksource* that was fine-tuned on a wide array of NLI datasets and other classification datasets (Sileo, 2023).[8] This model predicts three scores (entailment, neutral, contradiction) and we report on the average **entailment** score as **Ent%**.

We additionally use for the first experiment **METEOR** (Lavie and Agarwal, 2007) (in *nltk*), a metric that looks at word overlaps like ROUGE but relaxes the matching criteria – it takes into account word stems and synonyms. Finally, we also report on the average **cosine similarity** (Cos) of text embeddings between generated and reference answers, a metric that emphasizes the semantic similarity of these two strings. For that we use the sentence transformer *all-mpnet-base-v2*.[9] The results are in Tables 7 and 8 in Appendix.

5 Results

5.1 Gold Snippets

The results of four large LLMs with gold snippets are present in Table 1. All models observe a similar pattern: after starting with a rather low zero-shot performance, already utilizing just one context snippet leads to a big jump in performance. After that, most models slowly and steadily improve their answers as measured by all three metrics. Looking at different models, for BioASQ, GPT 4o and LLaMa

[5]https://en.wikipedia.org/w/api.php?action=query&list=search&srsearch=\{text\}&format=json

[6]Source: https://wikimedia-research.github.io/Discovery-Search-Test-BM25/

[7]https://pypi.org/project/rouge/

[8]https://huggingface.co/sileod/deberta-v3-base-tasksource-nli

[9]https://huggingface.co/sentence-transformers/all-mpnet-base-v2

# docs	GPT 4o (semantic)			Mixtral (semantic)			GPT 4o (BM25)			Mixtral (BM25)		
	R-L	BSc	Ent.%	R-L	BSc	Ent.%	R-L	BSc	Ent.%	R-L	BSc	Ent.%
0	23.8	87.0	23.9	21.8	85.8	29.4	23.8	87.0	23.9	21.8	85.8	29.4
1	20.9	86.5	18.9	22.1	86.0	23.4	20.7	86.5	18.9	22.2	86.0	23.5
3	23.8	86.6	19.0	23.1	86.1	23.9	22.4	86.7	20.1	22.8	86.0	24.1
5	22.9	86.7	19.5	23.3	86.0	25.8	23.2	86.9	20.9	23.0	86.1	26.1
10	23.0	86.8	19.9	23.2	86.0	27.6	23.4	86.9	21.5	23.3	86.0	28.9
15	25.1	87.1	26.9	24.7	86.2	31.4	24.9	87.2	26.5	25.0	86.4	31.1
20	25.3	**87.3**	**27.6**	24.6	86.2	**31.9**	**25.5**	**87.4**	**27.9**	**25.2**	**86.5**	**32.0**
30	**25.7**	87.2	27.5	**24.9**	**86.3**	31.5	25.4	87.2	27.4	25.1	86.3	31.6

Table 4: Results of final QA performance on BioASQ for different number of **retrieved context** snippets in the **open retrieval** setting, using PubMed (10 million doc.) with two big LLMs. Semantic refers to semantic search using dense vector embeddings and BM25 is a sparse retrieval technique, which showed better performance here. The results are measured with ROUGE-L (R-L), BERTScore (BSc), and average entailment prediction of the NLI model (Ent.%).

#	GPT 4o			Mixtral		
	R-L	BSc	E.%	R-L	BSc	E.%
1	23.0	83.7	12.9	25.4	84.7	16.7
3	24.3	84.0	13.6	26.2	84.9	16.4
5	24.8	84.1	14.4	26.4	84.9	17.2
10	25.4	84.2	15.4	27.2	85.0	17.3
20	25.8	**84.4**	**16.3**	27.8	**85.2**	**18.9**
30	**26.1**	84.2	16.2	**28.0**	85.1	18.8

Table 5: Results of final QA performance on QuoteSum for different number of **retrieved context** snippets from **retrieved documents** from Wikipedia, using Wikipedia's built-in BM25-based search.

3 (70B) had a similar performance, with LLaMa slightly outperforming GPT. Mixtral showed for BioASQ by far the strongest performance among the models across all three metrics. The biggest jump is observed in the entailment metric, showing how the answers generated by Mixtral had a higher entailment score – meaning a higher logical alignment with the reference answer. On the other hand, for QuoteSum, the situation is the other way around. GPT models performed the best, followed by LLaMa, and Mixtral came in last place. The zero-context performance was a lot lower than any context-based setting, showing how questions from this dataset are highly dependent on context.

The difference in performance for BioASQ could be explained by the different levels of biomedical knowledge that some models encode compared to others. In related studies, Mixtral and Mistral were found to encode more recent and higher quality biomedical knowledge than GPT 4 (Vladika et al., 2024b), while Mistral was found to perform better on biomedical QA tasks than the domain-specific biomedical variation BioMistral (Dada et al., 2024).

The results of four smaller LLMs with gold snippets are shown in Table 2. A similar pattern is observed – the top-1 snippet helps improve the performance significantly, while further additions lead to a more linear improvement. This holds true for LLaMa 3 (8B) and Gemma (7B). Mistral, just like its larger counterpart Mixtral, led to excellent performance as measured by all three metrics. The best performance for top-5 and top-10 was done by Qwen (7B). For BioASQ, even compared to the way bigger models LLaMa 3 (70B) and GPT 4o in the previous table, Mistral and Qwen demonstrated a lot stronger performance and context utilization, showing that the model size is not always the most important factor in RAG systems and comprehensive answering. For QuoteSum, Qwen outperfromed LLaMa 3 (70B), while Mistral outperformed its larger counterpart Mixtral.

Table 9 shows an example of a question from BioASQ, its top 3 gold snippets, and the ideal answer from our dataset. It then shows the responses generated by different LLMs for the given question and given three snippets, demonstrating the differences in addressing the question based on the given context. Mixtral seems to have picked up the most details while keeping it concise, Qwen gave the longest answer, while GPT 4o and LLaMa 3 (70B) left some crucial details out. Interestingly, LLaMa 3 (8B) gave a more informative answer than its bigger counterpart.

5.2 Closed Retrieval

This setting used the small knowledge base of eight thousand PubMed articles that were used as gold evidence in BioASQ. The results of the experiments are shown in Table 3. In this setting, it is visible that the performance dropped when compared to Table 1. Even in Mixtral, which was the best performing model, the performance dropped on average. Still, the performance kept improving

Question	Gold answer	Mixtral (w/ top 3 snippets)	Mixtral (internal knowledge)
For what is Protein A from Staphylococcus aureus used in biochemistry?	Protein A from the bacterium Staphylococcus aureus (SpA) is used as an **affinity ligand for purification of immunoglobulin G (IgG)**.	Protein A from Staphylococcus aureus is used in biochemistry as a powerful diagnostic reagent, and it also has applications in human disease therapy. It is isolated from the bacterial pathogen and is known to be produced exclusively as an extracellular protein in certain strains.	Protein A from Staphylococcus aureus is often used in biochemistry and molecular biology for the purification of antibodies and the detection of antibody-antigen complexes. It has a high affinity for the Fc region of IgG antibodies, making it a useful tool in various laboratory techniques such as immunoprecipitation, immunoblotting, and ELISA.
What is the function of Taraxasterol in rheumatoid arthritis?	Taraxasterol suppresses **inflammation** in **rheumatoid arthritis**.	Based on the context provided, there is no direct information about the function of Taraxasterol in rheumatoid arthritis.	Taraxasterol is a plant-derived compound that has been studied for its potential anti-inflammatory effects, which are beneficial in managing symptoms of rheumatoid arthritis.

Table 6: Two examples where internal knowledge of LLMs gave a better answer than the RAG answer based on top 3 retrieved snippets

with each increase of *top k* snippets selected, once again demonstrating that with more context, the performance was better. This was especially apparent in the top-10 setting, since the more evidence snippets selected, the higher the chances of selecting some of the gold evidence snippets used for the generation of the ideal answer.

5.3 Open Retrieval

The final setting used around 10 million PubMed articles as its knowledge base for retrieval. The idea of this experiment is to see (1) how much the performance in the open setting differs from the closed setting with gold evidence and (2) what the influence of different retrievers is on this performance. Results for BioASQ are shown in Table 4, while results for QuoteSum are shown in Table 5. Since we stored an offline copy of PubMed documents, we could use both BM25 and semantic search (with local vector embeddings), while for Wikipedia, we used its search API based on BM25.

When compared to previous tables, it is evident that open retrieval is the most challenging setting, with lowest average scores overall. It is also interesting to observe that retrieving the documents with BM25 led to a slightly better final performance compared to semantic search.

6 Discussion

6.1 Retrieval Techniques

Looking at Table 3, BM25 led to a better performance overall. Given that it works with keyword matches, this retrieval technique optimizes for precision in search results rather than recall, thus ensuring that more documents will actually be discussing the same concepts (words) mentioned in the question itself. This shows that optimizing for precision and matching the keywords of the query to the knowledge contained in the knowledge base can lead to improved performance. Especially in critical applications like the biomedical domain of question answering, optimizing for precision and robust answers can be more important than the recall provided by semantic search.

6.2 Internal vs. External Knowledge Conflict

An interesting remark from open retrieval in Table 4 is that both GPT and Mixtral have better scores for their zero-shot answers (with 0 context snippets) than the answers where up to 10 context snippets were provided. After we analyzed many outputs, a potential explanation of this phenomenon is that, while snippets discovered in the corpus can be semantically similar to the question, they do not always provide all the important information. On the other hand, when using just the vanilla prompt, LLM answers based on its "internal" knowledge — these answers reflect the collected knowledge of LLMs gained from the large pre-training corpora. Therefore, the internal LLM answers can be more informative than the RAG setting where an LLM is instructed to answer only using the provided short snippets. As more snippets are added, the informativeness of RAG answers starts surpassing LLM's internal knowledge. Recent studies have also observed that for biomedical tasks, it can sometimes be more beneficial to generate internal answers than retrieve external context (Frisoni et al., 2024).

Consider the first example in Table 6 – the answer from Mixtral's internal knowledge mentions purification and IgG, the same as in the gold answer, while the answer based on the top 3 snippets produced an incomplete answer. In general, the bottleneck is often tied to incorrect retrieval — sometimes, the retrieved snippets did not address the question at all, especially for complex biomedical terms found in BioASQ. On the other hand, LLMs in the vanilla setting will always provide the answer based on their best knowledge, thus outperforming cases of bad retrieval. This is apparent in the second example in Table 6. This demonstrates the well-known challenges of knowledge conflict between the internal knowledge of LLMs and the knowledge passed to them in the context (Marjanovic et al., 2024) and is an interesting future research direction following from our study.

6.3 Context Saturation

Another insight of the study visible in Table 4 is that there is a certain upper limit to the performance improvements. As we kept on adding more and more context, increasing to 20, the performance stalled and then slightly dropped for 30 retrieved context snippets. As the saturation point is reached, adding more context to the prompt just leads to noise and confusion in answering. This confirms the previous findings from literature that context can get "lost in the middle" of long prompts and ignored by the reader LLM when answering the questions (Liu et al., 2024; Hsieh et al., 2024).

7 Conclusion

In this study, we explored the effectiveness of Retrieval-Augmented Generation (RAG) systems for long-form question answering using two datasets. We systematically evaluated the impact of various settings of retrieval strategies, context sizes, and base LLMs on RAG performance. Our findings indicate that increasing context snippets enhances performance up to around 15 snippets. For biomedical QA, models like Mixtral and Qwen performed the best, while they were outperformed by GPT-4o and LLaMa 3 for encyclopedic QA. In open retrieval setting, BM25 yielded better results for biomedical QA, with open challenges remaining for exploring knowledge conflict between internal LLM knowledge and external context. We envision future work to explore the effects of query expansion methods and evidence re-ranking. We hope our work provides valuable insights for optimizing applied RAG systems in practice.

Limitations

Our study is limited to two datasets, thus making it possible that some findings would not universally generalize to different domains and tasks. Additionally, we only evaluate the models in a zero-shot setting, whereas a few-shot setting with some examples of questions and answers would have led to a more uniform performance across models.

The use of automated metrics for natural language generation tasks is not ideal, and they have certain drawbacks. ROUGE score focuses too much on word overlaps with no semantic matching, BERTScore often gives scores in a very tight range, and NLI models can struggle with long text as input. Ideally, a human evaluation would bring a more rigorous result assessment, but hiring human annotators, especially domain experts for the medical text, was outside of our scope and resources.

Finally, the LLMs, embedding models, and retriever models tested in this study represent only a subset of the quickly evolving landscape of NLP models and technologies. We selected some of the most popular and trending ones, but there are certainly other models that warrant discussion and would have led to an improved performance. Since most experiments were conducted in June 2024, the choice of benchmarked models reflects that. In the meantime, GPT 4o-mini has superseded GPT 3.5, LLaMa 3.3 is a continuation of LLaMa 3, Gemma 2 was released, as well as Qwen 2 after Qwen 1.5.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable suggestions. This research has been supported by the German Federal Ministry of Education and Research (BMBF) grant 01IS17049 Software Campus 2.0 (TU München).

References

AI@Meta. 2024. Llama 3 model card.

Chenxin An, Ming Zhong, Zhichao Geng, Jianqiang Yang, and Xipeng Qiu. 2021. Retrievalsum: A retrieval enhanced framework for abstractive summarization. *Preprint*, arXiv:2109.07943.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Deng Cai, Yan Wang, Lemao Liu, and Shuming Shi. 2022. Recent advances in retrieval-augmented text generation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 3417–3419.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024a. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024b. Complex claim verification with evidence retrieved in the wild. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3569–3587, Mexico City, Mexico. Association for Computational Linguistics.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. *Preprint*, arXiv:2401.14887.

Amin Dada, Marie Bauer, Amanda Butler Contreras, Osman Alperen Koraş, Constantin Marc Seibold, Kaleb E Smith, and Jens Kleesiek. 2024. Does biomedical training lead to better medical performance? *Preprint*, arXiv:2404.04067.

Pritam Deka, Anna Jurek-Loughrey, and P Deepak. 2022. Improved methods to aid unsupervised evidence-based fact checking for online health news. *Journal of Data Intelligence*, 3(4):474–504.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

William Fedus, Jeff Dean, and Barret Zoph. 2022. A review of sparse expert models in deep learning. *Preprint*, arXiv:2209.01667.

Giacomo Frisoni, Alessio Cocchieri, Alex Presepi, Gianluca Moro, and Zaiqiao Meng. 2024. To generate or to retrieve? on the effectiveness of artificial contexts for medical open-domain question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9878–9919, Bangkok, Thailand. Association for Computational Linguistics.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Gemini. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2G: Retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715, Seattle, United States. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Jennifer Hsia, Afreen Shaikh, Zhiruo Wang, and Graham Neubig. 2024. RAGGED: Towards informed design of retrieval augmented generation systems. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. 2024. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14982–14995, Bangkok, Thailand. Association for Computational Linguistics.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7036–7050, Mexico City, Mexico. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1).

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Xiaonan Li, Changtai Zhu, Linyang Li, Zhangyue Yin, Tianxiang Sun, and Xipeng Qiu. 2024. LLatrieval: LLM-verified retrieval for verifiable generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5453–5471, Mexico City, Mexico. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315, Singapore. Association for Computational Linguistics.

Sara Vera Marjanovic, Haeun Yu, Pepa Atanasova, Maria Maistro, Christina Lioma, and Isabelle Augenstein. 2024. DYNAMICQA: Tracing internal knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14346–14360, Miami, Florida, USA. Association for Computational Linguistics.

Gemma Team Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, L. Sifre, Morgane Riviere, Mihir Kale, J Christopher Love, Pouya Dehghani Tafti, L'eonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Am'elie H'eliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Cl'ement Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikula, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Pier Giuseppe Sessa, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vladimir Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Brian Warkentin, Ludovic Peran, Minh Giang, Cl'ement Farabet, Oriol Vinyals, Jeffrey Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *ArXiv*, abs/2403.08295.

OpenAI. 2023. Gpt-4 technical report.

Tal Schuster, Adam Lelkes, Haitian Sun, Jai Gupta, Jonathan Berant, William Cohen, and Donald Metzler. 2024. SEMQA: Semi-extractive multi-source question answering. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1363–1381, Mexico City, Mexico. Association for Computational Linguistics.

Damien Sileo. 2023. tasksource: Structured dataset preprocessing annotations for frictionless extreme multi-task learning and evaluation. *arXiv preprint arXiv:2301.05948*.

Juraj Vladika and Florian Matthes. 2024. Improving health question answering with reliable and time-aware evidence retrieval. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4752–4763, Mexico City, Mexico. Association for Computational Linguistics.

Juraj Vladika, Luca Mülln, and Florian Matthes. 2024a. Enhancing answer attribution for faithful text generation with large language models. In *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KDIR*, pages 147–158. INSTICC, SciTePress.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024b. MedREQAL: Examining medical knowledge recall of large language models via question answering. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14459–14469, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yuhao Wang, Ruiyang Ren, Junyi Li, Xin Zhao, Jing Liu, and Ji-Rong Wen. 2024. REAR: A relevance-aware retrieval-augmented framework for open-domain question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5613–5626, Miami, Florida, USA. Association for Computational Linguistics.

Jacob White. 2020. Pubmed 2.0. *Medical reference services quarterly*, 39(4):382–387.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

Anbang Xu, Tan Yu, Min Du, Pritam Gundecha, Yufan Guo, Xinliang Zhu, May Wang, Ping Li, and Xinyun Chen. 2024. Generative ai and retrieval-augmented generation (rag) systems for enterprise. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5599–5602.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245, Toronto, Canada. Association for Computational Linguistics.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. RankRAG: Unifying context ranking with retrieval-augmented generation in LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zihan Zhang, Meng Fang, Lingxi Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023. How do large language models capture the ever-changing world knowledge? a review of recent advances. In *Conference on Empirical Methods in Natural Language Processing*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *Preprint*, arXiv:2303.18223.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *Preprint*, arXiv:2101.00774.

A Appendix

A.1 Additional Metrics

The results of the experiments with gold snippets were additionally evaluated using the METEOR and cosine similarity metrics. The results are shown in Tables 7 and 8. For BioASQ, the results with these two metrics mostly follow the patterns observed with the original metrics from the main part, with a big jump in performance for the first snippet and then continued to increase. It is a similar case for QuoteSum, but in this dataset, the two metrics seem to peak at the top 5 snippets and then slightly drop and deteriorate when including all top 10 snippets.

A.2 Example outputs

Example outputs of 6 models for a question from BioASQ, together with top 3 gold snippets and ideal answer, are shown in Table 9.

#	GPT 3.5		GPT 4o		LLaMa 70B		Mixtral		Gemma 7B		LLaMa 8B		Mistral		Qwen 7B	
	MET	Cos	MET	Cos	MET	Cos	MET	Cos	MET	Cos	MET	Cos	MET	Cos	MET	Cos
0	19.6	74.5	21.3	75.1	21.9	75.0	25.8	75.3	20.6	66.3	25.3	72.9	24.6	73.0	22.9	70.7
1	21.2	76.1	21.3	76.4	22.9	76.4	30.7	80.2	23.7	70.5	25.8	76.6	29.6	79.6	26.7	79.0
3	24.5	79.3	24.6	79.2	26.7	79.4	37.9	82.9	28.5	74.0	27.2	78.8	35.2	82.1	32.5	81.8
5	25.7	80.1	25.5	80.0	27.8	79.8	40.4	84.0	30.6	75.2	30.0	79.9	37.7	83.1	35.1	83.0
10	26.5	80.6	26.6	80.6	28.7	80.3	42.8	84.8	32.3	76.1	31.1	80.3	39.8	84.1	37.8	84.2

Table 7: Results of final QA performance on BioASQ for different number of **gold context snippets** using the four big LLMs and four small LLMs. The results are measured with Meteor (MET) and average cosine similarity of text embeddings using *all-mpnet-base-v2*.

#	GPT 3.5		GPT 4o		LLaMa 70B		Mixtral		Gemma 7B		LLaMa 8B		Mistral		Qwen 7B	
	MET	Cos	MET	Cos	MET	Cos	MET	Cos	MET	Cos	MET	Cos	MET	Cos	MET	Cos
0	23.6	73.0	23.4	73.0	23.2	72.3	26.0	72.6	6.3	29.3	21.8	70.1	25.4	72.2	24.6	72.9
1	25.8	76.1	25.9	76.2	24.1	72.5	31.1	76.6	9.8	36.2	23.5	71.1	28.0	76.6	27.4	76.7
3	30.4	78.1	30.6	78.2	33.0	77.6	36.3	78.7	17.1	56.9	29.1	74.9	34.6	79.0	32.1	78.6
5	31.9	78.5	31.5	78.5	34.6	77.5	36.8	78.8	17.3	56.1	30.3	75.3	34.8	78.9	32.7	79.0
10	31.5	78.3	31.5	78.4	34.5	77.5	36.8	79.0	17.2	55.9	30.5	75.5	34.8	78.9	32.8	79.0

Table 8: Results of final QA performance on QuoteSum for different number of **gold context snippets** using the four big LLMs and four small LLMs. The results are measured with Meteor (MET) and average cosine similarity of text embeddings using *all-mpnet-base-v2*.

Question	What is the role of nimotuzumab in treatment of pontine glioma?
Top 3 snippets	(1) We found that nimotuzumab with or without radiotherapy, chemotherapy in newly diagnosed or recurrent HGG, such as glioblastoma multiforme (GBM), anaplastic astrocytomas (AA), and diffuse intrinsic pontine glioma (DIPG), might improve the response rate or the survival time. (2) In conclusion, nimotuzumab is a very well-tolerated drug with acceptable toxicity, and it may have promising value in the combination treatment. (3) Combining craniospinal irradiation (CSI) with concurrent temozolomide and nimotuzumab therapy may slightly improve tumor control and overall survival.
Ideal Answer	Nimotuzumab (an anti-EGFR monoclonal antibody) is being used for treatment of pontine gliomas. Nimotuzumab is a very well-tolerated drug with acceptable toxicity, and it may have promising value in the combination treatment. Clinical trials evaluating efficacy of nimotuzumab are ongoing.
GPT-4	The role of nimotuzumab in the treatment of pontine glioma is to potentially improve the response rate or survival time when used in combination with radiotherapy and chemotherapy.
Mixtral	The context suggests that nimotuzumab, a well-tolerated drug with acceptable toxicity, could potentially improve the response rate or survival time when used in combination with radiotherapy and/or chemotherapy for treating pontine glioma (DIPG) and other high-grade gliomas like glioblastoma multiforme (GBM) and anaplastic astrocytomas (AA).
LLaMa 3 (70B)	Nimotuzumab may improve the response rate or survival time in the treatment of pontine glioma.
LLaMa 3 (8B)	According to the context, the role of nimotuzumab in the treatment of pontine glioma is to potentially improve the response rate or survival time, either as a standalone treatment or in combination with radiotherapy and chemotherapy.
Mistral	Nimotuzumab may potentially improve the response rate or survival time when used in combination with radiotherapy, chemotherapy for newly diagnosed or recurrent High-Grade Gliomas (HGG), including Diffuse Intrinsic Pontine Glioma (DIPG).
Qwen (7B)	The role of nimotuzumab in the treatment of pontine glioma, as mentioned in the context, is that it may improve the response rate or survival time when used in combination with radiotherapy and chemotherapy for newly diagnosed or recurrent glioblastoma multiforme (GBM), anaplastic astrocytomas (AA), and diffuse intrinsic pontine glioma (DIPG). It is described as a well-tolerated drug with acceptable toxicity, suggesting it may have potential as a part of a combination therapy.

Table 9: Example of a question from BioASQ with its top 3 gold snippets and ideal answer. Below are responses generated by different LLMs, demonstrating the differences in addressing the question based on the given context.