More Documents, Same Length:
Isolating the Challenge of Multiple Documents in RAG

Anonymous ACL submission

Abstract

Retrieval-augmented generation (RAG) provides LLMs with relevant documents. Although previous studies noted that retrieving many documents can degrade performance, they did not isolate how the quantity of documents affects performance while controlling for context length. We evaluate various language models on custom datasets derived from a multi-hop QA task. We keep the context length and position of relevant information constant while varying the number of documents, and find that increasing the document count in RAG settings poses significant challenges for LLMs. Additionally, our results indicate that processing multiple documents is a separate challenge from handling long contexts. We will publicly release the datasets and code upon publication to facilitate further research in multi-document retrieval.

1 Introduction

The RAG approach enriches prompts with relevant documents, retrieved according to an input query (Karpukhin et al., 2020). For example, given a question about a certain historical period, RAG techniques can retrieve documents related to the time from a large historical corpus.

Recent work has noted a drop in RAG performance when retrieving many documents. For example, in multi-hop QA, LLMs struggle when the number of retrieved documents grows, even when presented with all the needed information (Press et al., 2022; Liu et al., 2023; Levy et al., 2024; Wang et al., 2024). Such deficiencies were observed without controlling for the *number of tokens* in which the information is conveyed, i.e., when the number of documents grew, so did the number of overall tokens, thus conflating between the challenge of long context and multi document.
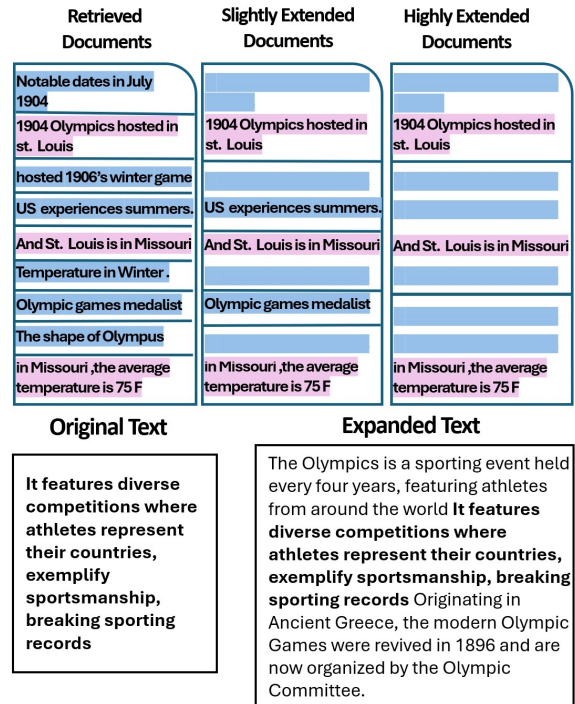
---
*Equal contribution.

Figure 1: **More Documents, Same Length.** We create various sets containing the same questions but differing in the number of distractor documents. Each set includes a multi-hop question, all of the supporting documents that contain the information to answer the question (pink), and varying distractor documents (blue). We begin with a 20-document version (left) and then reduce the number of documents while maintaining a fixed context size. When fewer documents are used, the remaining documents are extended (blue without text) so that concatenating them yields the same total length.

In this work, we address the following question: *Assuming a fixed input length, how is LLM performance affected by the number of retrieved documents?* This disentangles the challenge of long context from the challenge in processing collections of related documents – which often contain redundancies, conflicting information, and implicit
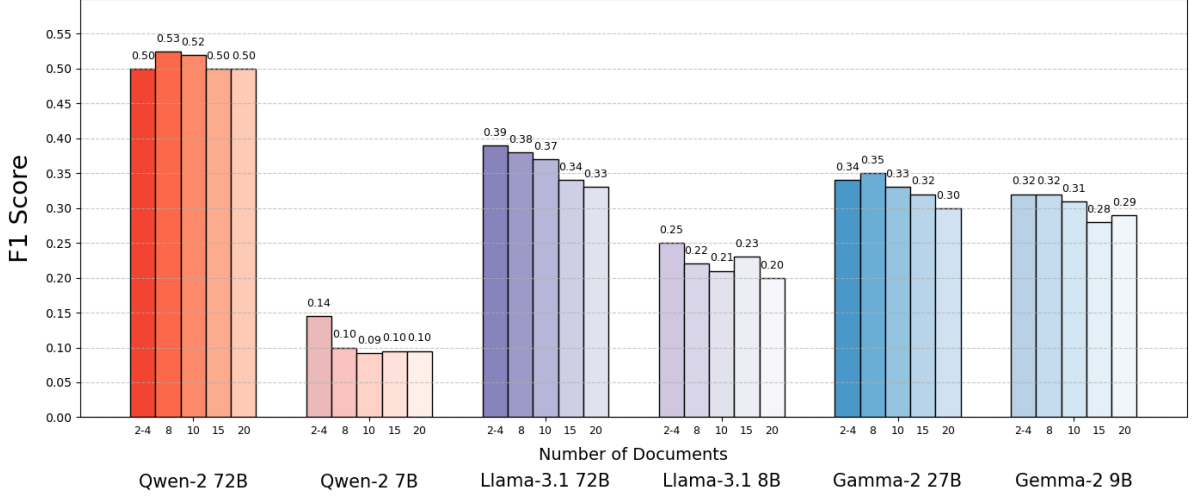
Figure 2: **Increasing the number of retrieved document can hurt performance.** In retrieval setups with fixed context windows, adding more documents could reduce performance by up to 10 percent. Two models (Llama- 3.1 and Gemma-2) showed worse performance, while Qwen-2 remained unaffected. The smaller versions of the LLMs (7–9B) show a similar trend as their larger counterparts but the effect is weaker. The hues of the bars represent the amount of retrieved documents.

inter-document relations (Hirsch et al., 2023; Lior et al., 2024). From a practical perspective, answering this question can help understand a breadth versus depth tradeoff — i.e., whether to strive to retrieve shorter context out of many documents or whether to aim to retrieve longer context out of fewer documents.

An ideal experimental setup would have the exact information conveyed *in the same number of tokens* across varying number of documents, from a long and self-contained single document to a large, multi-document corpus. We find that the custom sets we constructed from MuSiQue (Trivedi et al., 2022), a multi-hop QA dataset, serve as a convenient approximation, allowing us to explore the relationship between long-context and multi-document comprehension in a controlled environment with real-world texts.

Each instance in MuSiQue consists of a question and a set of 20 documents, where each document is an excerpt from a Wikipedia article retrieved according to the input question. MuSiQue is constructed such that the question can be answered based on only a few of the input documents (2-4), while the other documents serve as realistic distractors in retrieval settings, as they revolve around the question's topic but do not contain information required to answer the question. We vary the *number of documents* in the input by gradually removing the distractor documents. When removing a distractor document, we respectively extend each of the remaining documents with distracting content from their corresponding Wikipedia article. Importantly, the process preserves the position of the relevant information within the cotext. This process is illustrated in Fig. 1.

If the context length is the sole challenge, we should expect the performance to remain similar regardless of the number of input documents. Conversely, if processing multiple related documents presents an additional challenge, we would expect an inverse correlation between performance and the number of input documents.

The results of evaluating several state-of-the-art models (Llama-3.1, Qwen2, and Gemma2) which are presented in Fig. 2, indicate that in most cases, reducing the number of documents improves performance by up to 10%. An exception is Qwen2, which may indicate that it better handles multi-document collections.

Our work has several major implications and avenues for future work. First, from a practical perspective, RAG systems should take the number of retrieved documents into consideration, as the introduction of additional documents into the prompt may hurt performance. Second, future work should explore novel approaches for multi-document processing, which according to our findings presents a separate challenge from mere long context. Such work can make use of our paradigm and data for training and evaluation.

2

2 Multi-Document Evaluation with Controlled Token Count

Our goal is to understand how the number of retrieved documents affects LLM performance when controlling the input length. To this end, we evaluate several models on multi-document multi-hop question answering, which requires models to find relevant information within a given context to answer a specific question. In particular, we make controlled adjustments to the number of documents in the input, while preserving the position of the key information needed to answer the questions, and keeping the context length consistent.

Our dataset is based on the validation set of MuSiQue (Trivedi et al., 2022), a multi-hop QA dataset that consists of 2,417 answerable questions. Each question is associated with 20 paragraphs sampled from individual documents, retrieved from Wikipedia according to the question. Of these paragraphs, 2–4 contain the supporting information necessary to answer the question, while the remaining paragraphs serve as realistic distractors in a RAG setup, as they are retrieved from related topics but do not contain relevant information to answer the question. Fig. 1 shows an example query, and a list of retrieved documents, where three are relevant to the question (marked in pink), and the rest are distractors (marked in blue).

Leveraging MuSiQue's structure, we constructed several data partitions to investigate the impact of the number of retrieved documents in a controlled manner. The process involved the following steps:

1. **Select the total number of documents:** We reduced the number of documents from the original 20 to 15, then 10, 8, and finally down to the 2–4 documents consisting of the relevant information to answer the question.

2. **Choose the supporting and non-supporting documents:** We always keep the documents that support the answer to ensure that the question remains answerable, and randomly select the remaining ones from the non-supporting set. Non-supporting documents remain consistent across different document counts, i.e., each set includes all documents from the smaller sets. Fig. 1 shows such document selection in the two right columns, note that relevant documents (blue) are always kept.

3. **Expand the selected documents:** Since the original documents are Wikipedia paragraphs, we located their source Wikipedia pages and added text preceding and following the paragraphs to match the original token count. This replaces distracting content from different documents with distracting content from the same document. In Fig. 1, we show that each of the remaining documents is expanded to keep the original token count, while ensuring that information from the supporting documents appeared in similar positions across all sets.

3 Evaluation

3.1 Experimental Setup

We evaluated six instruction-tunes LLMs from three model families: Llama-3.1 8B/70B (AI@Meta, 2024), Qwen2 7B/72B (Yang et al., 2024), and Gemma2 9B/27B (Team et al., 2024). We used Together.ai[1] platform to run the large versions of the models, while the smaller version of the models were run using A6000 GPU. We used decoding temperature of 0.8 for all models, as recommended in previous LLM evaluations (Chen et al., 2021).

For evaluation, we measured the overlap F1 score between the gold and the predicted outputs, as suggested in MuSiQue (Trivedi et al., 2022). The prompt, formats, and evaluation code were implemented using the SEAM benchmark (Lior et al., 2024).

3.2 Results: Adding documents can hurt RAG by up to 10%

Our key findings (Fig. 2) reveal that in a retrieval setup, LLMs suffer when presented with more documents, even when the total context length is the same. This may be due to the unique challenges in multi-document processing, which involves processing information that is spread across multiple sources, which can introduce conflicting or overlapping details. Almost all models perform better when presented with fewer documents, with scores improving by 5% to 10% on average. We find that the smaller versions of all LLMs exhibit a similar pattern, albeit to a lesser degree.

An exception is Qwen2, which may indicate that it better handles multi-document collections. Its
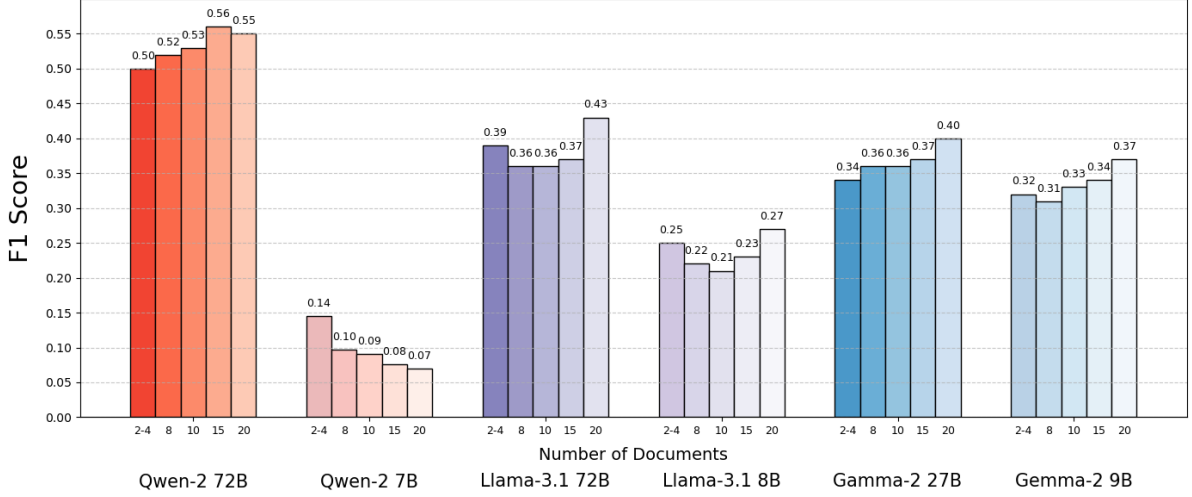
---

[1] https://www.together.ai

Figure 3: **The effects of adding non-related documents.** When adding irrelevant documents, LLMs' performance improves.

Model	Supporting documents only	No documents
Qwen-2 72B	0.61	0.01
Qwen-2 7B	0.25	0.04
Llama-3.1 70B	0.44	0.02
Llama-3.1 8B	0.19	0.02
Gemma-2 27B	0.52	0.02
Gemma-2 9B	0.50	0.05

Table 1: F1 scores for the large and small versions of each model in two scenarios. In the first scenario, only the supporting documents are provided (without expanding the context). In the second scenario, only the question is provided (without any supporting documents).

scores were similar across all tested settings, and slightly higher for 8 and 10 documents.

3.3 Analysis

To contextualize our results, we created three additional versions of our data, discussed below along with the respective findings.

Additional context hurts performance. We tested the performance when models are given only the supporting documents, thus providing a much shorter context and eliminating any distracting content. The performance of the LLMs on this set was significantly higher compared to the experimental sets that contained external information. Full results are shown in Table 1.

Contamination does not seem to interfere with our results. To evaluate whether the models' knowledge is already encoded in their parameters, we run the models only on the questions, with no additional retrieved context. The results showed a consistent poor performance of approximately 0.02 F1 score across all models, mitigating the concern of data contamination. The complete set of results can be found in Table 1.

Random distractors mitigate confusion. Finally, we evaluated all the models against a version of the data where we use randomly selected Wikipedia paragraphs, instead of using the retrieved distractors. As shown in Fig. 3, for the large versions of the LLMs, the performance of the models actually improved as more documents appeared within the input with random distractors. This suggests that similar but unrelated documents, which are often retrieved in RAG, may confuse the model and decrease performance.

4 Conclusions

We assessed the challenges of multi-document retrieval tasks when varying the number of documents. Our results indicate that input that includes more documents complicates the task in an environment of retrieval settings, highlighting the need for retrieval systems to balance relevance and diversity to minimize conflicts. Future models could benefit from mechanisms to identify and discard conflicting information while leveraging document variety.

5   Limitations

This study does not address prompt variations or the effects of data order within inputs. Future work should explore alternative datasets to ensure more robust evaluations. While our experiments focused on extreme scenarios (highly distracting or random contexts) and document counts between 2–20, future research should investigate more nuanced setups and larger document sets to better reflect real-world conditions. All datasets from this study will be publicly available upon publication for further research in multi-document processing.

References

AI@Meta. 2024. Llama 3 model card.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Eran Hirsch, Valentina Pyatkin, Ruben Wolhandler, Avi Caciularu, Asi Shefer, and Ido Dagan. 2023. Revisiting sentence union generation as a testbed for text consolidation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7038–7058, Toronto, Canada. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models.

Gili Lior, Avi Caciularu, Arie Cattan, Shahar Levy, Ori Shapira, and Gabriel Stanovsky. 2024. Seam: A stochastic benchmark for multi-document tasks.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multihop questions via single-hop question composition.

Minzheng Wang, Longze Chen, Cheng Fu, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang, and Yongbin Li. 2024. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.