Long Context RAG Performance of Large Language Models

Quinn Leng, Jacob Portes, Sam Havens, Matei Zaharia, Michael Carbin

databricks mosaic research

How well do the best open source and commercial models do on long-context RAG tasks?

TLDR:

- Using longer context does not uniformly increase RAG performance.
- LLMs fail at long context RAG in unique ways as a function of context length.

Models

- Commercial: Claude, GPT4, o1
- Open Source: Llama-3, Qwen, etc..

Datasets

- FinanceBench
- Databricks DocsQA
- Natural Questions (NQ)

Observation: Commercial vs OSS

The best commercial models such as o1-mini/preview, GPT-4o, and Claude 3.5 Sonnet steadily, improve performance as a function of context length, while the majority of the open source models first increase and then decrease performance as context length increases

Observation: Distinct failure patterns

- **Claude 3 Sonnet:** refused to answer due to copyright concerns, especially at longer context lengths.
- **Llama 3.1 405B:** maintained consistent failure performance up to 64k tokens,
- **Mixtral-8x7B**: repeated or random content.
- **DBRX:** failed to follow instructions for context lengths above 16k, often summarizing content instead of answering questions directly.

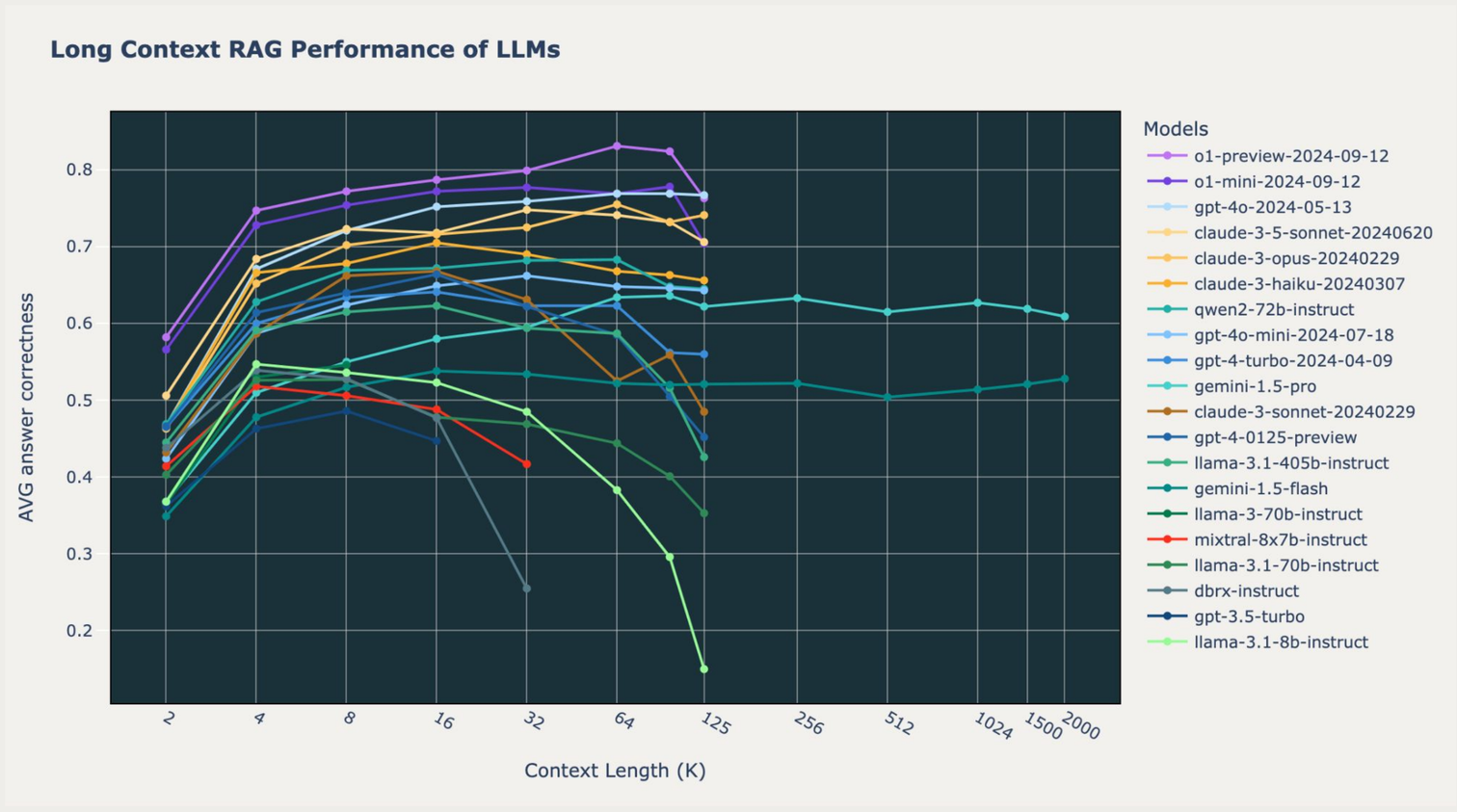
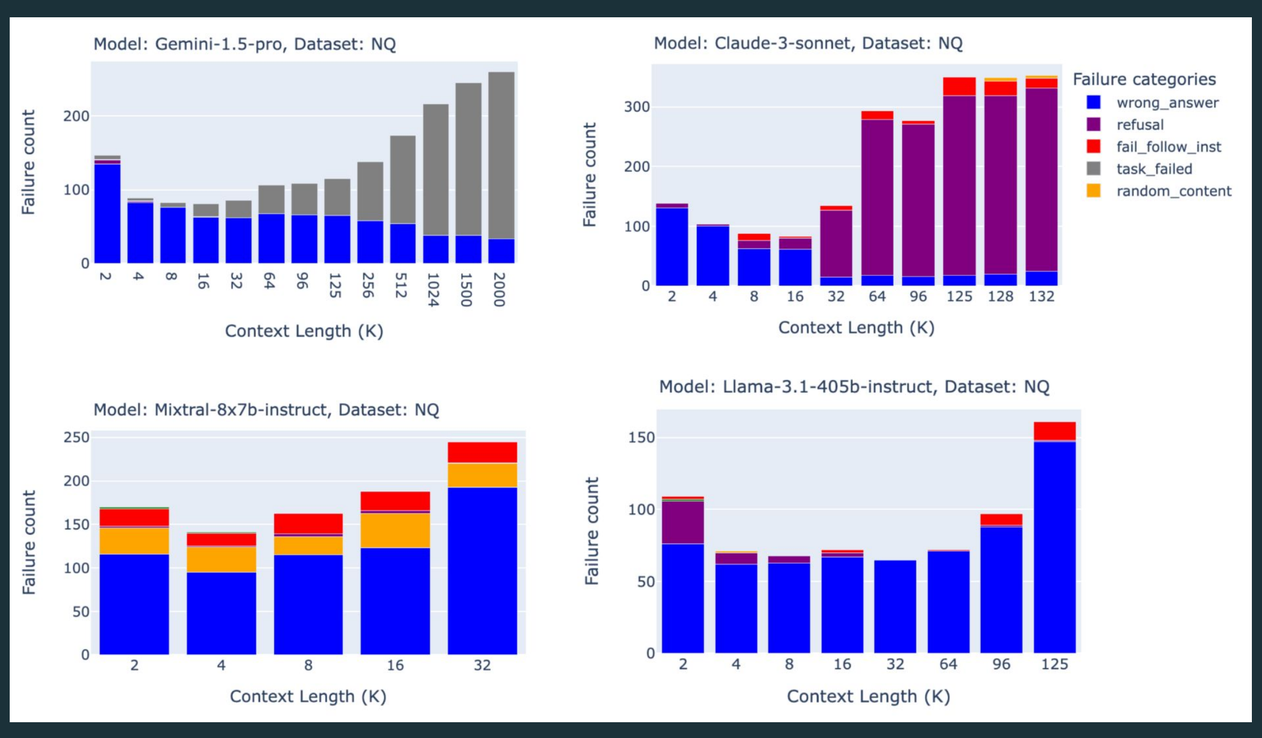


Figure 1: Long context RAG performance of o1, GPT-4, Claude 3/3.5, Gemini 1.5, Llama 3/3.1, Qwen 2, Mistral and DBRX models on 3 curated RAG datasets

Dataset \ Details	Category	Corpus #docs	# queries	AVG doc length (tokens)	Description
DocsQA (v2)	**corporate question-answering**	7563	139	2856	DocsQA is a question-answering dataset using information from Databricks public documentation and real user questions and labeled answers. Each of the document in the retrieving corpus is a webpage.
FinanceBench (150 tasks)	**finance question-answering**	53399	150	811	FinanceBench is an academic RAG dataset that includes pages from 360 SEC 10k filings from public companies and the corresponding questions and ground truth answers based on the SEC10k documents. More details can be found in the paper https://arxiv.org/abs/2311.11944. We use a proprietary (closed source) version of the full dataset from Patronus. Each of the document in our corpus correspond to a page from the SEC 10k PDF files.
NQ-doc-dev	**knowledge (wikipedia) question-answering**	7369	534	11354	Natural question is an academic question-answering dataset from Google, discussed in their 2019 paper. The queries are Google search queries. Each query is answered using content from wikipedia pages in the search result.

Table 1: We benchmarked all LLMs on 3 curated RAG datasets that were formatted for both
retrieval and generation.

LLMs Fail at Long Context RAG in Different Ways



We extracted the answers for each model at different context lengths and defined the following broad failure categories:

- **repeated_content**: when the LLM answer is completely (nonsensical) repeated words or characters.
- **random_content**: when the model produces an answer that is completely random, irrelevant to the content, or doesn't make logical or grammatical sense.
- **fail_to_follow_instruction**: when the model doesn't understand the intent of the instruction or fails to follow the instruction specified in the question. For example, when the instruction is about answering a question based on the given context while the model is trying to summarize the context.
- **wrong_answer**: when the model attempts to follow the instruction but the provided answer is wrong.
- **others**: the failure doesn't fall under any of the categories listed above