ChatQA 2: Bridging the Gap to Proprietary LLMs in Long Context and RAG Capabilities

Peng Xu, Wei Ping, Xianchao Wu, Chejian Xu, Zihan Liu, Mohammad Shoeybi, Bryan Catanzaro

1. Motivation

1. The open LLM community has made significant progress in advancing the **long context capabilities** of open-access LLMs.
2. **Performance gaps** to frontier proprietary models, e.g., GPT-4-Turbo, still **exist**.
3. The absence of **training data and reproduction recipe** makes it challenging
4. Evaluation on **synthetic** tasks does not represent **real-world downstream task**

2. Contributions

1. We propose a **two-step** approach to establish the long context capability of Llama3-70B
2. Llama3-ChatQA-2-70B with 128K context window achieves **better accuracy** than most **existing state-of-the-art models**, including GPT-4-Turbo-2024-04-09, Qwen2-72B-Instruct, and Llama3.1-70B-Instruct, on both long-context tasks exceeding 100K tokens and the RAG benchmark for conversational QA tasks within a 4K window.
3. We found RAG performance of our Llama3-ChatQA-2-70B is **highly robust to variations in chunk size** when using a long-context retriever. Even more promising, accuracy **consistently improves** as the total number of retrieved tokens for input increases.

5. Ablation of Retrievers

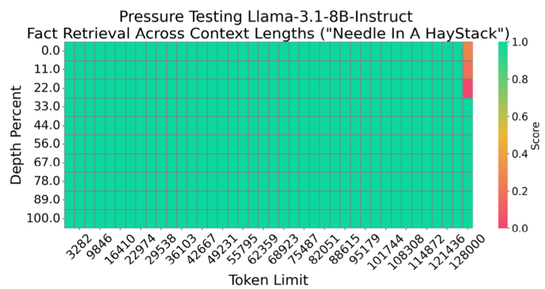
The choice of retrievers do not make much difference.

Retriever	Long (32K)	En.MC	En.QA
E5-Mistral	47.59	73.80	40.66
NV-emb-v2	46.43	74.24	41.19

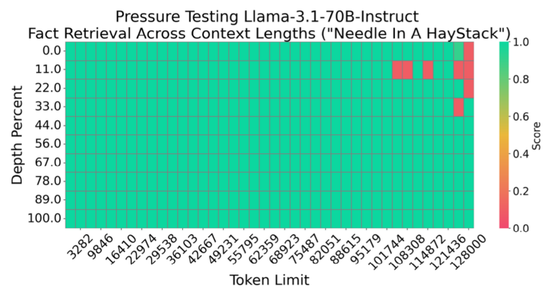
3. Main Results

Model Type	Long Context LLM (128K)	Ultra-long (>100K) 4 tasks	Long (32K) 6 tasks	Short (4K) 9 tasks (RAG)
Proprietary	GPT-4-Turbo-2024-04-09	33.16	**51.93**	54.72
OPEN ACCESS	Qwen2-72B-Instruct	39.77	49.94	54.06
	Llama3.1-70B-Instruct	39.81	49.92	52.12
Open Source	`Llama3-ChatQA-2-70B`	**41.04**	48.15	**56.30**

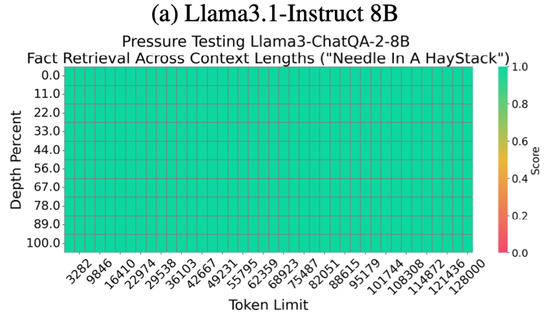
4. Needle in a Haystack Test



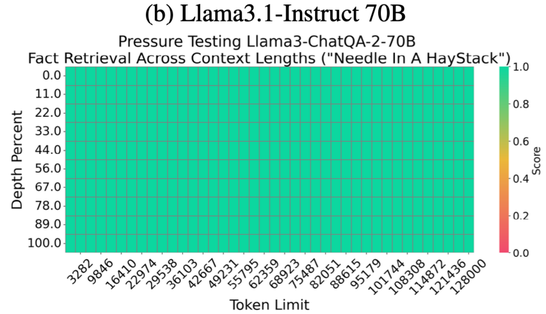
(a) Llama3.1-Instruct 8B



(b) Llama3.1-Instruct 70B

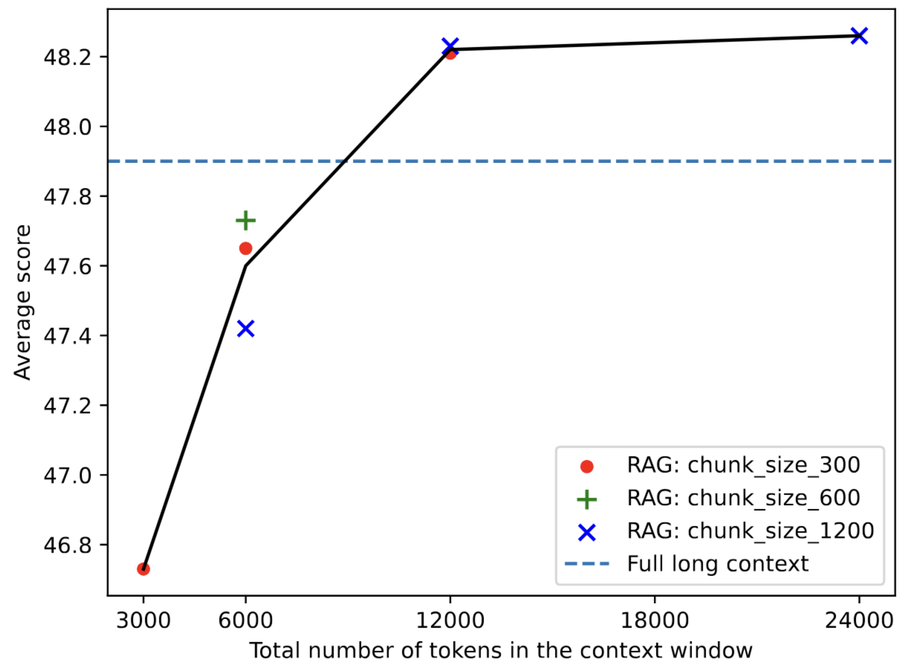


(c) Llama3-ChatQA-2-8B



(d) Llama3-ChatQA-2-70B

6. Long Context vs RAG



7. Comparing to Llama3.1

	`Llama3-ChatQA-2-8B` (new)	Llama3.1-8B-Instruct
Short (ChatRAG)	**53.79**	48.79
Long (32K)	42.05	**42.42**
Ultra long (>100K)	**35.18**	33.17
HumanEval	66.46	**70.73**
MMLU	65.73	**67.59**
MT-bench	8.09	**8.42**
GSM8K (0-shot)	**87.41**	83.70