# How Long Can Context Length of Open-Source LLMs truly Promise?

**Dacheng Li** [*b] **Rulin Shao** [*w] **Anze Xie** [s] **Ying Sheng** [b]

**Lianmin Zheng** [b] **Joseph E. Gonzalez** [b] **Ion Stoica** [b] **Xuezhe Ma** [u] **Hao Zhang** [s]

[b] UC Berkeley [w] University of Washington [s] UCSD [c] CMU [m] MBZUAI [u] USC

## Abstract

Large language models (LLMs) with long-context instruction following ability has unlocked new potentials, such as supporting long interactive chat sessions. In this paper, we introduce a test suite, LongEval, which enables us to evaluate the long-range retrieval ability of LLMs at various context lengths. We use LongEval to evaluate open-sourced LLMs, and surprisingly, we find many of them fail to achieve their promised context length. In addition, we present a recipe to fine-tune a long-context chatbot based on LLaMA models, and introduce LongChat models that supporting conversations of up to 16,384 tokens. We have released our code at `https://github.com/DachengLi1/LongChat`.

## 1 Introduction

Recent open-sourced Transformer-based chatbots have exhibited strong instruction-following ability and alignment with human preference (Chiang et al., 2023; Taori et al., 2023; Xu et al., 2023; Geng et al., 2023). However, their supported context lengths are usually limited by the pretraining context length of their base models, which typically falls within the range of 2,048 tokens (Touvron et al., 2023). This limitation poses a bottleneck for many advanced applications that require the model reason over a longer context like debugging complex code snippets. In the meantime, there has been a significant surge within the open-source community in developing language models for longer context length. For instance, MPT-7B-storywriter is fine-tuned for a context length of 65K and extrapolates to 80K using Alibi (Press et al., 2021), and ChatGLM2-6B is fine-tuned with 8K context length (Zeng et al., 2023; Du et al., 2022; Team, 2023). However, there lacks a robust benchmark to assess whether these models have actually achieved their promised context length.

To fill this gap, we first propose a test suite, LongEval (§ 4), inspired by real-world multi-round conversations (Zheng et al., 2023a). LongEval incorporates two tasks of different difficulty and provides a simple way to measure long-context performance. Surprisingly, we have found that many open-sourced long-context models have failed to deliver the promised context length [2] (§5.1). To this end, we present a recipe to effectively fine-tune an existing foundation model to a long-context chatbot (§ 3), and introduce a new series of long-context models with a sequence length of 16K, LongChat. LongChat models accurately follow human instructions in long conversations and demonstrate strong alignment with human preferences (Zheng et al., 2023b). We also evaluate them against another academic long-context benchmark and find promising results (Shaham et al., 2023).

[*] Authors contributed equally.

[2] Evaluated models are only up to July 2023; further updates on these models are not tested.

Workshop on Instruction Tuning and Instruction Following at NeurIPS 2023.

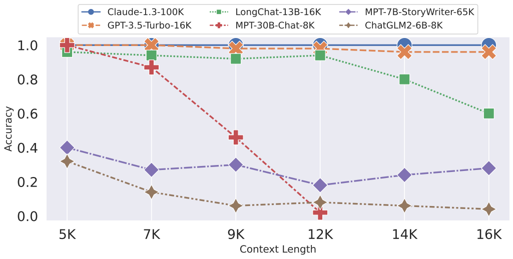

Figure 1: Line retrieval results.

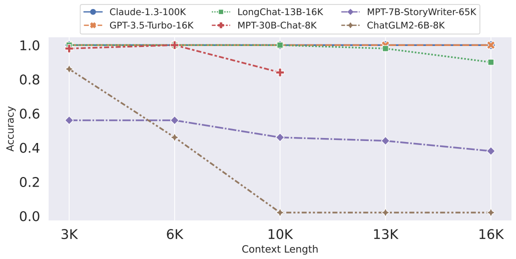

Figure 2: Topic retrieval results.

## 2 Related Work

**Extending Rotary Positional Embedding.** Rotary is a relative positional embedding proposed by Su et al. (2021). There have been some concurrent works on extending Rotary beyond the pretraining context. kaiokendev (2023) shows it is promising to linearly interpolate the Rotary embedding. Concurrently, Chen et al. (2023) independently proposes the same technique, naming it positional interpolation. In addition, Reddit users bloc97 (2023) propose to use a non-linear interpolation scheme using neural tangent kernel (NTK) theory. This paper is a concurrent work on applying the linear interpolation to extend the context length of rotary embedding models.

**Long-context Evaluation.** We next delve into existing benchmarks for long-context evaluation. Long range arena (Tay et al., 2020) focuses on the evaluation of sequence lengths ranging from 1K to 16K with both synthetic tasks and real-world tasks. SCROLLS (Shaham et al., 2022) and Zero-SCROLLS (Shaham et al., 2023) include tasks consisting of up to 57K tokens in the questions, e.g., very long scientific papers, which requires the model to generate summarization for the scientific papers, and NarrativeQA, which requires the model to answer questions based on the long inputs. Distinguishing from existing synthetic-based benchmark, we do not require retraining the model. Distinguishing from academic benchmark, we focus on pure long-context ability testing, without focusing on other factors such as reasoning. In addition, our design of retrieval tasks enables us to measure the model performance with a self-defined context length, thus facilitating the comparison of different models at various sequence lengths.

## 3 LongChat: An instruction-following long-context chatbot

The training target is to fine-tune the based LLaMA model, which is trained on 2048 context length, to understand natural languages with much larger context length. Our training recipe can be conceptually described in two steps:

### 3.1 Stage 1: Condensing rotary embeddings

Rotary position embedding (Su et al., 2021) is implemented in HuggingFace library by:

```
query_states, key_states = apply_rotary_pos_emb(query_states, key_states,
cos, sin, position_ids)
```

Where position_ids are indices, indicating the position of a token within a sentence. For instance, the token "today" in the sentence "today is a good day" corresponds to a position_ids 1. The apply_rotary_pos_emb() function applies a transformation based on the position_ids.

The LLaMA model is pre-trained with rotary embedding on sequence length 2048. Consequently it has not encountered examples where position_ids > 2048 during the pre-training phase. Rather than requiring the LLaMA model to adapt to position_ids beyond 2048, we instead condense position_ids > 2048 to be within 0 to 2048. Concretely, we define condensation ratio by dividing the targeted new context length y by 2048. We then divide every position_ids by this ratio and use it in the apply_rotary_pos_emb() function. Conceptually, this is one line of code change:

```
query_states, key_states = apply_rotary_pos_emb(query_states, key_states,
cos, sin, position_ids / ratio)
```

In this paper, we fine-tune the model to y=16384, with a condensation ratio is 8. For instance, a token with position_ids = 10000 becomes position_ids = 10000 / 8 = 1250, and the neighboring token 10001 becomes 10001 / 8 = 1250.125. This step of condensing requires no training.

Table 1: MT-Bench results on model of the same sizes

| LongChat-13B-16k | Vicuna-13B | WizardLM-13B | Baize-v2-13B | Nous-Hermes-13B | Alpaca-13B |
|---|---|---|---|---|---|
| 5.95 | 6.39 | 6.35 | 5.75 | 5.51 | 4.53 |

### 3.2 Step 2: Fine-tuning on Curated Conversation Dataset

In the next step, we perform a fine-tuning procedure on our curated conversation dataset. In particular, we start with the ShareGPT dataset (Zheng et al., 2023b) that has been shown to be effective in teaching LLaMA models to learn to follow instructions in multi-round conversations. We fine-tune the 7B on a full 80K ShareGPT dataset (Zheng et al., 2023b). For the 13B model, we curate a smaller dataset from the 80k examples for efficiency. Concretely, we select the examples responded by GPT-4 and mix them with truncated long examples (>16384 tokens) responded to by GPT-3.5. In this way, we curate a dataset of 18k conversations, with a mixture of long and short conversations, following the principle in Beltagy et al. (2020). We fine-tune the model using standard next-token prediction loss. To save memory, we use Pytorch FSDP and Flash Attention (Dao et al., 2022).

## 4 LongEval: a simple and composable testsuites for long-context LLMs

In this section, we present LongEval, a collection of coarse and fine grained long-context tests.

### 4.1 Task 1: Coarse-grained Topic Retrieval

In real-world long conversations, users usually discuss and switch between several topics (Zheng et al., 2023a). The Topic Retrieval task mimics this scenario by asking the chatbot to retrieve the first topic in a long conversation consisting of multiple topics.

Topic retrieval example

USER: I would like to discuss <TOPIC-1>
ASSISTANT: What about xxx of <TOPIC-1>?
... (A multi-turn conversation of <TOPIC-1>)
... (More topics discussions)
USER: I would like to discuss <TOPIC-k>
...
USER: What is the first topic we discussed?

Line retrieval example

line torpid-kid: <CONTENT> is <24169>
...
line wacky-cob: <CONTENT> is <10310>
...
line dull-trap: <CONTENT> is <32177>
...
What is the <CONTENT> in line wacky-cob?

This task tests whether the model can locate a chunk of text and associate it with the right topic name. We design a conversation to be 400-600 tokens long. Thus, this task is considered coarse-grained because the model may give correct predictions when it locates positions not too far away (<500 token distance) from the right ones.

### 4.2 Task 2: Fine-grained Line Retrieval

To further test the model's ability to locate and associate texts from a long conversation, we introduce a finer-grained Line Retrieval test. In this test, the chatbot needs to precisely retrieve a number from a long document, instead of a topic from long multi-round conversations.

The task was originally proposed in open-sourced community (Papailiopoulos, 2023). The original testcase uses numbers as contents, which we found smaller LLMs usually cannot comprehend. We make it more suitable for open-sourced chatbots by using random natural language (e.g. torpid-kid).

## 5 Results and findings

We compare 6 recently released LLMs with long context support, including four open-source models and two proprietary models. Their specifications are listed in Table 1 below.

### 5.1 LongEval Results

From the coarse-grained topic retrieval test, we already observe false promises of open-source long-context models. For instance, Mpt-7b-storywriter claims to have a context length of 65K, but barely achieves 50% accuracy even at one-fourth of its claimed context length (16K). Chatglm2-6B does not consistently retrieve the correct first topic at the context length of 6K (46% accuracy). Its retrieval accuracy degrades more when it is tested on context length greater than 10K. On the other hand, we observed that our LongChat-13B-16K model reliably retrieves the first topic, with comparable accuracy to gpt-3.5-turbo.

In the finer-grained line retrieval test, Mpt-7b-storywriter performs even worse than in the coarse-grained cases, dropping accuracy from 50% to 30%. Chatglm2-6B also has degradation, not being able to consistently retrieve the associated number at 5K context length. We notice that ChatGLM2-6B states that it has not been yet fully optimized for single-turn long document understanding, which could explain its current performance on LongEval. In contrast, we observe that LongChat-13B-16K performs reliably, achieving near gpt-3.5 or Anthropic-claude's Bai et al. (2022) ability within 12K context length and decent performance towards its limit (more discussion in § 5.5).

### 5.2 Attempts to compare models of weaker instruction following ability

In topics and line retrieval tests, we observe some models fail to follow instructions at a very long context. For instance, in the Line Retrieval test, the model may simply respond "sure, I will tell you the number" instead of returning an actual number. Although a strong chatbot should be expected to have strong instruction following ability, in LongEval, we would like to more fairly evaluate their long context ability. To give this more fair comparison, we took two actions to avoid factors irrelevant to long-context capability: prompt engineering and compute accuracy based only on the cases where the model follows our instruction. More details are presented in codes.

Table 2: ZeroScrolls results (validation set)

| Benchmark | LongChat-13B-16K | LongChat-7B-16k | Vicuna-13B-v1.3 | Vicuna-7B-v1.3 | GPT-4-8k |
|---|---|---|---|---|---|
| Qasper (F1) | 0.286 | 0.275 | 0.220 | 0.190 | 0.356 |

### 5.3 Human preference benchmark (MT-bench)

In the previous section, we have observed that LongChat models perform well on long-range retrieval tasks, but does this come with a significant drop in human preference? To test whether it still follows human preferences, we use GPT-4 graded MT-Bench, a set of challenging multi-turn conversation questions (Zheng et al., 2023b). We find that LongChat-13B-16K is comparable to its closest alternative - Vicuna-13B, which indicates that this long-range ability does not come with a significant sacrifice of its short-range ability. At the same time, LongChat-13B-16K is competitive compared to other models of the same size.

### 5.4 Long sequence question answer benchmark

In the previous sections, we have tested models on our (simple) long-range retrieval tasks and human preference tasks. But how do these models perform on more complex academic long-range reasoning tasks? In this section, we study this by running the Qasper question-answering dataset. We use the validation set selection and prompts from the ZeroScrolls long sequence benchmark (Shaham et al., 2023). As shown in the table below, LongChat significantly outperforms Vicuna due to its longer context length. We leave more rigorous analysis on academic benchmarks for future work.

### 5.5 Limitations and future work

We find that LongChat-13B-16K also experiences an accuracy drop when the context length is near 16K on the fine-grained line retrieval task. In our preliminary attempts, we conjecture that this is because it is near the maximal fine-tuning length. For instance, training on even larger (e.g. 32K) can alleviate this problem. We leave this to an exciting future work. Moreover, we have observed many efficiency issues (e.g. memory and throughput) during training/inferencing chatbots of much longer context length. We plan to develop systems for long-context chatbot training/inference.

## 6 Conclusions

In our evaluations, commercial long-context models always fulfill their promise: gpt-3.5-16k and Anthropic Claude (almost) achieve perfect performance in both our benchmarks. However, existing open-sourced models often do not achieve the claimed context length. For instance, Mpt-7b-storywriter claims a 65k context length but can barely achieve 50% accuracy beyond a much smaller context length (10k for topic retrieval task and 5k for line retrieval task). We qualitatively summarize our assessment in Table 4.

## References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

bloc97. Ntk-aware scaled rope allows llama models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation., 2023. URL `https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/ntkaware_scaled_rope_allows_llama_models_to_have/`.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 320–335, 2022.

Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. Koala: A dialogue model for academic research. Blog post, April 2023. URL `https://bair.berkeley.edu/blog/2023/04/03/koala/`.

kaiokendev. Extending context is hard. . . but not impossible, 2023. URL `https://kaiokendev.github.io/context`.

Dimitris Papailiopoulos. A little retrieval test for large language models, 2023. URL `https://github.com/anadim/the-little-retrieval-test`.

Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, et al. Scrolls: Standardized comparison over long language sequences. *arXiv preprint arXiv:2201.03533*, 2022.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. Zeroscrolls: A zero-shot benchmark for long text understanding. *arXiv preprint arXiv:2305.14196*, 2023.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006*, 2020.

MosaicML NLP Team. Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023. URL `www.mosaicml.com/blog/mpt-7b`. Accessed: 2023-05-05.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*, 2023.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL `https://openreview.net/forum?id=-Aw0rrrPUF`.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, et al. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*, 2023a.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023b.

# 7 Supplementary Material

Table 3: Qualatative summary of the model evaluated

| | Model Size | Instruction-tuned | Pretrain Context Length | Finetune Context Length | Claimed Context Length | Open-sourced |
|---|---|---|---|---|---|---|
| MPT-30b-chat | 30B | yes | 8K | ? | 16K | yes |
| MPT-7b-storywriter | 7B | yes | 2K | 65K | 84K | yes |
| ChatGLM2-6b | 6B | yes | 32K | 8K | 8K | yes |
| LongChat-13b-16k (ours) | 13B | yes | 2K | 16K | 16K | yes |
| Gpt-3.5-turbo | – | – | – | – | 16K | no |
| Anthropic claude-1.3 | – | – | – | – | 100K | no |

Table 4: Qualatative summary of the model evaluated

| Ability at the claimed context length | Text generation | Coarse Retrieval | Fine-grained Retrieval |
|---|---|---|---|
| Description | Faithfully generate natural languages | Retrieve information in a coarse granularity | Retrieve information in a fine-grained granularity |
| LongChat-13B-16K | ★★★ | ★★★ | ★★ |
| MPT-30B-Chat | ★★★ | ★★★ | ★★ |
| MPT-7B-Storywriter | ★★★ | ★★ | ★ |
| ChatGLM2-6B-32K | ★★★ | ★★ | ★ |
| GPT-3.5-turbo-16K | ★★★ | ★★★ | ★★★ |
| Anthropic-claude-100K | ★★★ | ★★★ | ★★★ |

**More details on LongEval** We performed sanity checks on both tasks by running sensitivity tests. We observed expected results that indicate the tasks can effectively examine LLM's abilities of text generation, retrieval, and information association in the scenario of long context and then reflect these abilities by returning the retrieving accuracy of models. An example of the sensitivity tests we run is based on the vanilla LLaMA model, which is pretrained with a 2K context length. We found this modle can achieve perfect accuracy on both tasks when the test inputs length are less than 2K, but will immediately fail (nearly 0 accuracy) on any test inputs beyond 2K.