

# Efficient Context Scaling with LongCat ZigZag Attention

Chen Zhang, Yang Bai, Jiahuan Li, Anchun Gui, Keheng Wang, Feifan Liu  
 Guanyu Wu, Yuwei Jiang, Defei Bu, Li Wei, Haihang Jing, Hongyin Tang, Xin Chen  
 Xiangzhou Huang, Fengcun Li, Rongxiang Weng\*, Yulei Qian, Yifan Lu, Yerui Sun  
 Jingang Wang\*, Yuchen Xie, Xunliang Cai  
 Meituan, China  
 zhangchen76@meituan.com

## ABSTRACT

We introduce LongCat ZigZag Attention (LoZA), which is a sparse attention scheme designed to transform any existing full-attention models into sparse versions with rather limited compute budget. In long-context scenarios, LoZA can achieve significant speed-ups both for prefill-intensive (e.g., retrieval-augmented generation) and decode-intensive (e.g., tool-integrated reasoning) cases. Specifically, by applying LoZA to LongCat-Flash during mid-training, we serve LongCat-Flash-Exp as a long-context foundation model that can swiftly process up to 1 million tokens, enabling efficient long-term reasoning and long-horizon agentic capabilities.

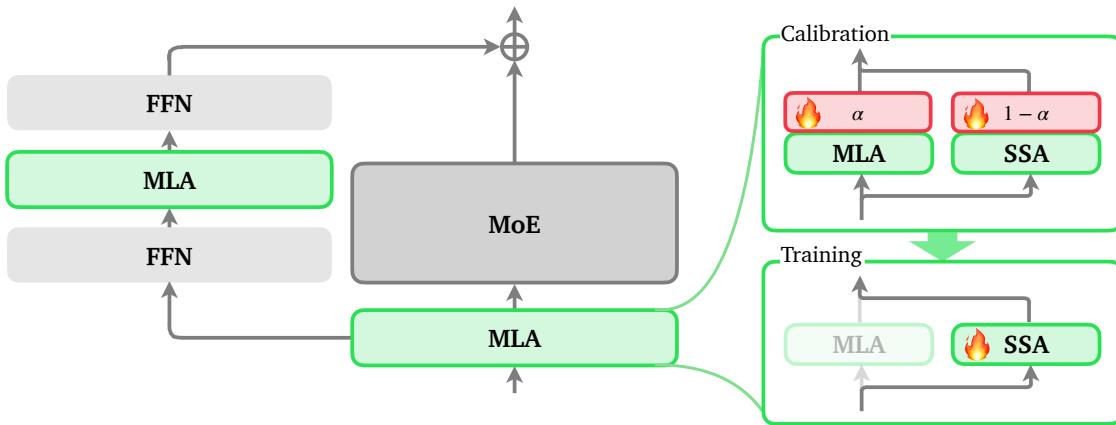


Figure 1 | The illustration of LongCat ZigZag Attention (LoZA), which involves first calibration and then training for realizing the sparsity. The illustration is shown with the exemplar shortcuted-MoE [Cai et al., 2025] in LongCat-Flash, which encompasses two MLA layers [DeepSeek-AI, 2024]. SSA: streaming sparse attention [Xiao et al., 2024].

## 1 Architecture

Built upon any language models (LMs, e.g., LongCat-Flash [LongCat Team et al., 2025]), we can basically alter some of full-attention modules to their sparse-attention alternatives [Tang et al., 2024, Yang et al., 2024, Zhang et al., 2025a, Sun et al., 2025, Zhao et al., 2025, DeepSeek-AI et al., 2025]. The transition from full attention to sparse attention is usually exerted during mid-training or so [Liu et al., 2025a, Zhang et al., 2025b, Tu et al., 2025, Liu et al., 2025b], and the sparse attention outcome is later used in follow-up procedures.

**Full Attention** Attention [Vaswani et al., 2017], or typically full attention, is a key ingredient in modern transformer architectures and thus large language models (LLMs). Usually, the attention takes the softmax

\*Rongxiang and Jingang are the corresponding authors.

form as below:

$$\mathbf{O} = \text{softmax}(\mathbf{Q}\mathbf{K})\mathbf{V}, \quad (1)$$

where other details such like the denominator are omitted.

As noted in the form, the compute essentially scales quadratically with the increase of context, incurring critical burdens for applications such as retrieval-augmented generation [Shi et al., 2024, Yen et al., 2024] or tool-integrated reasoning [Qin et al., 2025, Gou et al., 2024] that runs with extensively long input or output.

**LongCat ZigZag Attention** Pioneers have devoted a lot of efforts in exploring how sparse attention could serve as an alternative as below:

$$\mathbf{O}' = \text{softmax}(\mathbf{Q}\mathbf{K}')\mathbf{V}', \quad (2)$$

where notations that are marked with ' are sparsified ones. For instance,  $\mathbf{K}'$  would only retain 10% elements compared to those indicated in  $\mathbf{K}$ .

In the alternative, the compute minimally remains constant with respect to the context scaling. To echo this, we put forward LongCat ZigZag Attention (short-termed LoZA), as shown in Figure 1. LoZA firstly uncovers the layers that can be sparsified without hurting much performance [Xiao et al., 2025], secondly sparsifies the layers that can be further trained [DeepSeek-AI et al., 2025] to close performance gap. The whole process behaves very much like what has been described in *lottery tickets hypothesis* [Frankle and Carbin, 2019]. In theory, a mid-trained LM is sequentially sparsified, rewound, mid-trained to maximally recover the full performance. In other words, the calibration starts at the end of mid-training while the training starts at the beginning of the mid-training.

*Calibration.* Regarding the MLA [DeepSeek-AI, 2024] that has been utilized in LMs such as DeepSeek-V3 and LongCat-Flash, LoZA assumes there are totally  $n$  MLAs. LoZA initially attaches a unique parameterized factor  $a_i \in [0, 1]$  per MLA so that the MLA processes as below:

$$\hat{\mathbf{O}}_i = a_i \cdot \mathbf{O}_i + (1 - a_i) \cdot \mathbf{O}'_i, \quad (3)$$

where  $\mathbf{O}_i$  and  $\mathbf{O}'_i$  denote the full attention output and sparse attention output generated by the  $i$ -th MLA separately. Here, sparse attention follows the streaming sparse pattern, where one query token only attends to several sink and local blocks [Xiao et al., 2024].

Then a round of training on the calibration data is carried out by freezing any parameters within the mid-trained LM except all  $a_i$ . After the optimization,  $a_i$  would differ from each other in magnitude, which is used to signify the importance of the corresponding MLA. Notably, by partly sparsifying MLAs with the lowest  $a_i$  values, the performance of the LM is largely preserved.

Based on the observation unearthed in calibration, LoZA later turns 50% MLAs with the lowest  $a_i$  in the mid-trained LM from full attention to streaming sparse attention (SSA) such that:

$$\mathbf{O}^* = \text{softmax}(\mathbf{Q}\mathbf{K}^*)\mathbf{V}^*, \quad (4)$$

where  $\mathbf{K}^*$  and  $\mathbf{V}^*$  are anchored and blocked keys and values (#sink blocks  $s$ , #local blocks  $l$ , block size  $b$ ).

*Training.* Preferably, though the sparsified LM maintains competitive performance, training is further required to close the potential performance gap brought by the sparsification, especially in long-context scenarios.

For budget purpose, we decide to locate the training at mid-training. Since mid-training only consumes hundreds of billions of tokens, it is relatively acceptable for limited compute.

*Pilot Studies.* To elaborate how the sparse pattern and training work, we have preliminarily conducted a study on directly applying an interleaved sparse pattern (i.e., sparsifying one out of two adjacent layers), and then tuning the pattern with only a few tokens.

It shall be observed in Table 1 that 1) hand-crafted sparse pattern yields significant performance drop on long-context data while calibration makes the sparse pattern way better, and 2) training enhances performance on long-context data.

**Desiderata** Our design is to a great extent inspired by Duo-Attention [Xiao et al., 2025]. The reason why we instead bet on layer-level rather than head-level streaming sparse attention as in Duo-Attention is that head-level sparsity in our case would easily lead to both compute imbalance across parallel ranks and schedule recompute across attention layers during inference.

*Kernel.* Head-level sparsity complicates kernel control flow, consuming critical efforts in achieving balanced metadata. In contrast, layer-level sparsity allows the kernel to follow a single uniform schedule, minimizing

Table 1 | The pilot studies on calibration and training. The *interleaved sparse pattern* denotes sparsifying one out of two adjacent layers, and the *calibrated sparse pattern* denotes sparsifying the lowest-valued layers during calibration. The *sparse training* means further training after sparsification. LongEval [Li et al., 2023a] is long-context evaluation and others are short-context evaluation.

Method	BBH	GSM8K	HumanEval+	LongEval
LongCat-Flash	81.0	94.2	66.6	95.7
w/ interleaved sparse pattern	81.2	94.3	65.9	54.1
w/ sparse training	80.2	93.4	70.1	67.4
w/ calibrated sparse pattern	80.9	93.8	67.7	89.6

metadata pressure. Furthermore, it is possible that kernels can process multiple KV groups per thread block to maximize occupancy; consequently, distinct head-level sparse patterns could induce warp divergence.

*Engine.* Head-level sparsity can shard heterogeneous workloads to different ranks (e.g., one device processing all full heads while another processes all sparse ones), creating stragglers that bottleneck global synchronization. Differently, layer-level sparsity guarantees uniform compute across all ranks. Besides, layer-level sparse patterns eliminate the runtime overhead of recomputing schedule metadata across layers.

## 2 Training

The training covers mid-training (specifically only long-context extension phases) and follow-up post-training, and finally yields LongCat-Flash-Exp. The mid-training recipe is roughly the same as those leveraged by LongCat-Flash [LongCat Team et al., 2025].

Contrarily, the post-training recipe is primarily simplified for fast prototyping. And in fact, we could already achieve expected performance with this simple recipe. We leave more fancy post-training to catch even more fascinating performance as future work. It is noteworthy the post-training recipe is mainly utilized for tuning an instruct (or say chat) model.

Nonetheless, to unlock the ability of handling a longer context, we equip these recipes with YaRN [Peng et al., 2024] so that LongCat-Flash-Exp can extrapolate itself to processing up to 1M tokens. In addition to that, we provide a few crucial parameters involved in LoZA. The block size (i.e.,  $b$ ) is 128, the number of sink blocks (i.e.,  $s$ ) is 1, and the number of local blocks (i.e.,  $l$ ) is 7, summing to 1,024 tokens.

**Mid-training** For mid-training, we take a data distribution identical to the one used by LongCat-Flash. During mid-training, LongCat-Flash-Exp walks through 32K, 128K, 256K training phases, and is extrapolated to 1M with the power of YaRN. To enhance the long-context ability, we involve 500B tokens during the 32K and 128K stages, followed by 40B tokens during the 256K stage. Regarding long-context data composition, it basically follows below:

- Reasoning-intensive data: enhancing the reasoning potential by expanding reasoning patterns;
- Agentic data: synthesizing large-scale agentic interactions by leveraging a vast array of task-oriented web content and thousands of model context protocol (MCP) servers;
- High-quality long-form data: integrating a diverse collection of long-form data, including curated, open-source, and synthetic books and textbooks;
- Repository-level code: integrating an extensive collection of full-repository codebases to enhance the capacity for solving real-world, cross-file programming challenges.

By strategically diversifying the data mixture, this composition ensures high-fidelity of data.

**Post-training** To quickly validate LongCat-Flash-Exp, we adopt a lightweight post-training pipeline. Specifically, we perform supervised fine-tuning (SFT) using a data distribution identical to that of LongCat-Flash [LongCat Team et al., 2025], but with only 50% of its original volume. To maintain performance, the dataset was carefully curated to span critical domains, including instruction following, mathematics, coding, agentic tasks, and general knowledge. Subsequently, to align with human preferences and optimize model behavior, we employed Direct Preference Optimization (DPO) [Rafailov et al., 2023] alongside Reinforcement Fine-Tuning

Table 2 | The effectiveness of LongCat-Flash-Exp-Base.

Method	MMLU-Pro	GPQA	BBH	GSM8K	HumanEval+	LongEval
LongCat-Flash-Base	70.0	51.2	81.0	94.2	66.6	95.7
LongCat-Flash-Exp-Base	69.9	54.6	81.6	93.8	67.1	99.3

Table 3 | The effectiveness of LongCat-Flash-Exp-Chat. The comparison results are derived on a diverse range of benchmarks covering an array of domains. The hybrid thinking model with  $\dagger$  is evaluated in chat mode for a fair comparison. The best result at each row is **boldfaced**.

	Benchmark <sub>metric</sub>	GLM 4.6 <sup><math>\dagger</math></sup>	DeepSeek V3.2 <sup><math>\dagger</math></sup>	LongCat-Flash Chat	LongCat-Flash Exp-Chat
General	MMLU <sub>Acc</sub>	90.7	<b>91.1</b>	89.7	89.6
	CEval <sub>Acc</sub>	89.6	89.6	<b>90.4</b>	89.9
	CMMLU <sub>Acc</sub>	<b>88.4</b>	87.3	84.3	87.5
	IFEval <sub>Acc</sub>	87.8	88.4	<b>89.7</b>	88.0
	GuideBench <sub>Acc</sub>	83.8	87.0	81.0	<b>90.4</b>
Math	MATH-500 <sub>Acc</sub>	98.6	97.2	96.4	<b>98.8</b>
	AIME-24 <sub>Avg@32</sub>	<b>83.7</b>	73.9	70.4	83.1
	AIME-25 <sub>Avg@32</sub>	<b>80.2</b>	56.5	61.3	74.9
	BeyondAIME <sub>Avg@10</sub>	52.8	42.2	43.0	<b>59.2</b>
Code	Humaneval+ <sub>Pass@1</sub>	<b>92.7</b>	89.0	88.4	87.2
	MBPP+ <sub>Pass@1</sub>	<b>83.6</b>	79.9	79.6	79.1
	LCB <sub>2408-2505, Pass@1</sub>	56.4	<b>59.5</b>	48.0	56.6
	FullStackBench <sub>Pass@1</sub>	61.0	62.5	61.8	<b>64.1</b>
STEM	MMLU-Pro <sub>Acc</sub>	81.5	<b>84.3</b>	82.7	84.0
	GPQA-Diamond <sub>Avg@16</sub>	74.8	75.3	73.2	<b>75.6</b>
Agent	SWE-Bench-Verified <sub>Acc</sub>	68.0	<b>72.1</b>	60.4	63.2
	Terminal-Bench <sub>Acc</sub>	40.5	<b>45.0</b>	39.5	42.5
	$\tau^2$ -Bench <sub>Acc</sub>	69.1	64.0	68.8	<b>69.5</b>
Multilingual	MMMLU <sub>Acc</sub>	<b>87.2</b>	86.7	81.7	85.2
	MGSM <sub>Acc</sub>	91.1	<b>94.9</b>	87.5	94.6
Long Context	LongBenchV2 <sub>Acc</sub>	51.5	<b>54.1</b>	38.2	53.5
	MRCR <sub>Acc</sub>	42.1	37.1	34.4	<b>59.7</b>
	HELMET <sub>Acc</sub>	64.6	59.5	59.1	<b>64.7</b>
	Longform-Writing <sub>Acc</sub>	70.0	<b>73.9</b>	51.3	69.6

(RFT) [OpenAI, 2024]. Compared to large-scale reinforcement learning (RL) approaches, our strategy achieves competitive performance while consuming minimal computational resources.

### 3 Evaluation

**Effectiveness** We first evaluate the base LongCat-Flash-Exp-Base. It is demonstrated in Table 2 that LoZA would not degrade performance. Namely, after mid-training with sparsity, LongCat-Flash-Exp-Base remains comparable to LongCat-Flash-Base.

We then evaluate the instructed LongCat-Flash-Exp-Chat on benchmarks of diverse domains and compare it against LongCat-Flash-Chat to examine the effectiveness of LoZA:

- **General:** MMLU [Hendrycks et al., 2021], CEval [Huang et al., 2023], and CMMLU [Li et al., 2023b] for general knowledge. IFEval [Zhou et al., 2023] and GuideBench [Diao et al., 2025] for instruction following.

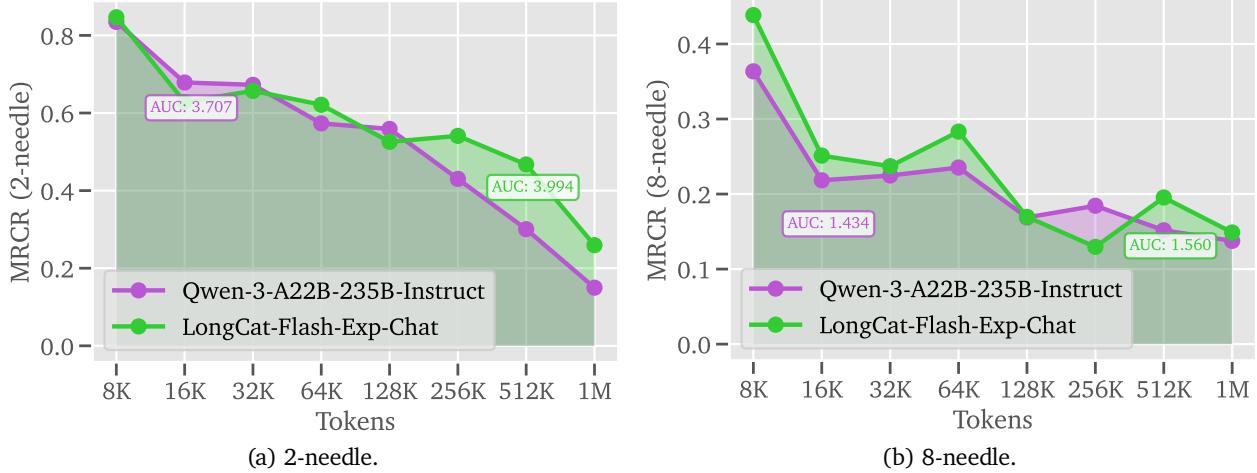


Figure 2 | The effectiveness of LongCat-Flash-Exp-Chat across different context lengths on MRCR. Qwen-3 is considered as a competitive baseline since it also possesses the ability of handling 1M context. AUC: area under curve.

- **Math:** Olympiad-level mathematical benchmarks, including MATH-500 [Lightman et al., 2023], AIME-24 [MAA, 2024] and AIME-25 [MAA, 2025] (American Invitational Mathematics Examinations), and BeyondAIME [ByteDance-Seed, 2025].
- **STEM:** MMLU-Pro [Wang et al., 2024] and GPQA-Diamond [Rein et al., 2024].
- **Code:** Humaneval+ [Liu et al., 2024a], MBPP+ [Liu et al., 2024a], LiveCodeBench (2024.08-2025.05) [Jain et al., 2025], and FullStackBench [Liu et al., 2024b].
- **Agent:** SWE-Bench [Jimenez et al., 2024], sourced from real GitHub issues for evaluating a model’s ability to solve software engineering problems. Terminal-Bench [Terminal-Bench Team, 2025], for evaluating a model’s agentic ability in real terminal environments.  $\tau^2$ -Bench [Barres et al., 2025], a Tool-Augmented Reasoning benchmark.
- **Multilingual:**<sup>1</sup> MMMLU [Hendrycks et al., 2021] and MGSM [Shi et al., 2022].
- **Long Context:** LongBenchV2 [Bai et al., 2025], MRCR [Vodrahalli et al., 2024], and HELMET [Yen et al., 2025] for evaluating long-context understanding, and Longform-Writing [Paech, 2025] for evaluating long text generation.

As shown in Table 3, LoZA would not compromise quality for speed. On the concerned benchmarks, LongCat-Flash-Exp-Chat exhibit competitive performance with LongCat-Flash-Chat. Concretely, LongCat-Flash-Exp excels LongCat-Flash-Chat on long-context benchmarks, largely due to the extended context length. LongCat-Flash-Exp-Chat also stands at the same line with other competitors such like GLM-4.6 concerning chat mode. That is, LongCat-Flash-Exp-Chat obtains similar number of **the best-performing ones** to that of GLM-4.6.

We also provide a micro-benchmarking of LongCat-Flash-Exp-Chat across different context lengths versus Qwen-3, which also possesses the ability of handling 1M context. In Figure 2, we could clearly see that LongCat-Flash-Exp-Chat outweighs Qwen-3 on some context lengths and overall surpasses Qwen-3 in terms of AUC.<sup>2</sup> This implies that LoZA combined with YaRN could efficiently pave the way for context scale of 1M.

**Efficiency** We draw respectively the decode cost of SSA against full attention and the end-to-end timeline of LongCat-Flash-Exp against LongCat-Flash, to showcase how LongCat-Flash-Exp overwhelms LongCat-Flash in real-world serving.

Since LoZA enables 50% sparsity in LongCat-Flash-Exp, the compute brought by attention should be ideally reduced by a factor of 2. For long-context circumstances where attention dominates the compute, the efficiency

<sup>1</sup>We report the average performance of eight widely-used languages, i.e., Arabic, French, German, Spanish, Portuguese, Indonesian, Japanese, and Korean.

<sup>2</sup>This metric for long-context evaluation is originally proposed in <https://contextarena.ai/>.

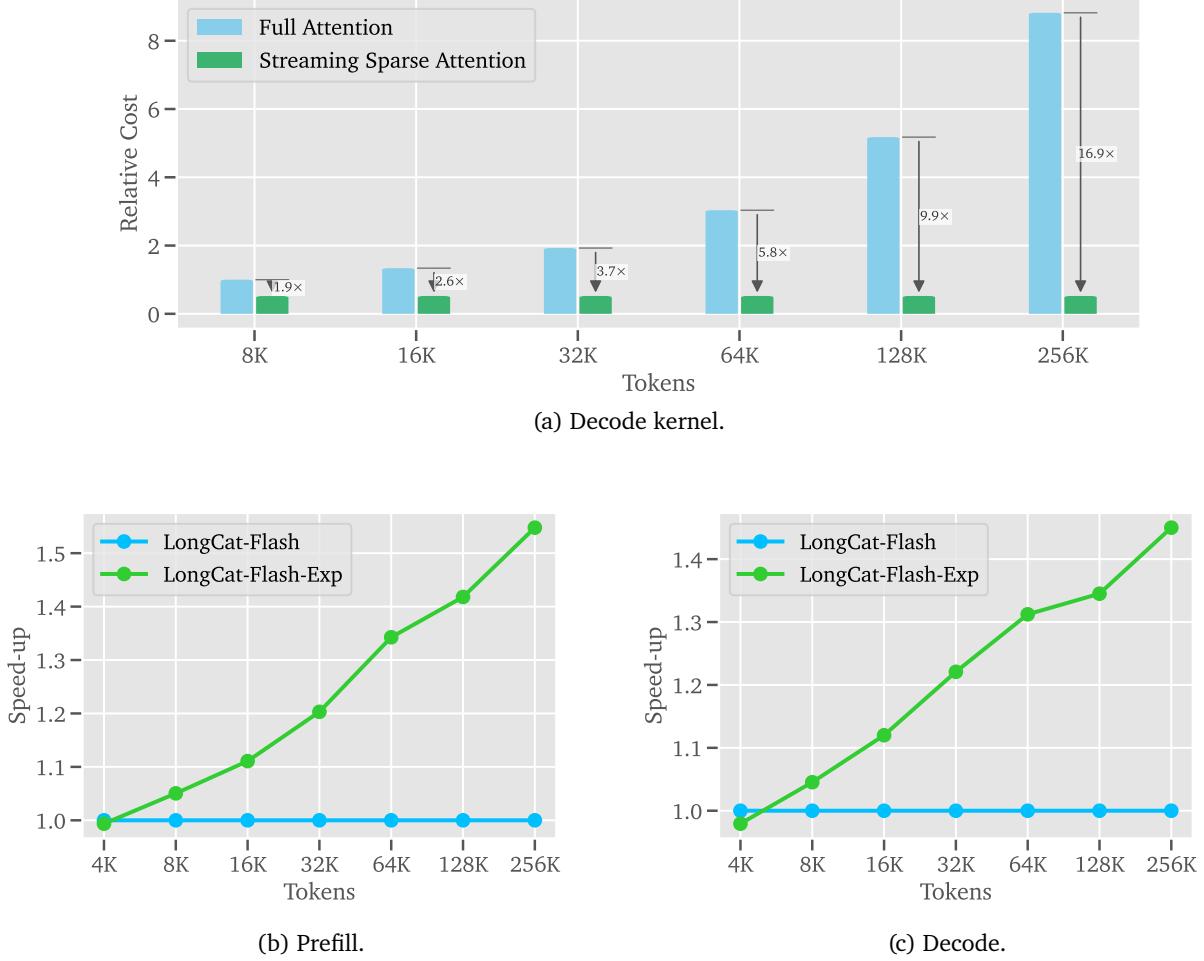


Figure 3 | The efficiency of LoZA. The relative cost and speed-up are practically measured on inference clusters.

could be maximally lifted to 2 times of the original. Promoted by our efforts in kernel and engine customizations, in Figure 3, streaming sparse attention kernel could minimally use 90% less cost in decode compared to full attention kernel (i.e., FlashMLA [Dege et al., 2025]) for a context of 128K tokens. Meanwhile, in end-to-end benchmarking, LongCat-Flash-Exp realizes more than 50% speed-up in prefill and saves over 30% cost in decode for a context of 256K tokens.

## 4 Conclusion

We present LoZA, a sparse attention algorithm built upon MLA, that is universally applicable to full-attention LMs and based off LongCat-Flash results in LongCat-Flash-Exp. During mid-training, LoZA transforms LongCat-Flash via calibration, sparsification, and training. It is worthy of mentioning that the process essentially follows the pace of *lottery tickets hypothesis*, providing adequate theoretical guarantee for LoZA. With specialized design efforts, LoZA realizes principal speed-ups in both prefill and decode phases. This enables efficient long-term reasoning and long-horizon agentic capabilities, thereby making context-native (i.e., context as memory) applications viable. LoZA might also broadly impact related work that is aims at improving attention, especially for that intends to transform the MLA to sparse one. We would extremely like to invite the community to embed LoZA into any other open-source LMs that use MLA, and perhaps large multi-modal models [Li et al., 2025].

## Acknowledgement

We sincerely thank the infrastructure team and evaluation team of LongCat for their constructive feedback and promptly support.

## References

- Weilin Cai, Juyong Jiang, Le Qin, Junwei Cui, Sunghun Kim, and Jiayi Huang. Shortcut-connected expert parallelism for accelerating mixture of experts. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net, 2025. URL <https://openreview.net/forum?id=GKly3FkxN4>.
- DeepSeek-AI. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *CoRR*, abs/2405.04434, 2024. doi:[10.48550/ARXIV.2405.04434](https://doi.org/10.48550/ARXIV.2405.04434). URL <https://doi.org/10.48550/arXiv.2405.04434>.
- Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=NG7sS51zVF>.
- Meituan LongCat Team, Bayan, Bei Li, Bingye Lei, Bo Wang, Bolin Rong, Chao Wang, Chao Zhang, Chen Gao, Chen Zhang, Cheng Sun, Chengcheng Han, Chenguang Xi, Chi Zhang, Chong Peng, Chuan Qin, Chuyu Zhang, Cong Chen, Congkui Wang, Dan Ma, Daoru Pan, Defei Bu, Dengchang Zhao, Deyang Kong, Dishan Liu, Feiye Huo, Fengcun Li, Fubao Zhang, Gan Dong, Gang Liu, Gang Xu, Ge Li, Guoqiang Tan, Guoyuan Lin, Haihang Jing, Haomin Fu, Haonan Yan, Haoxing Wen, Haozhe Zhao, Hong Liu, Hongmei Shi, Hongyan Hao, Hongyin Tang, Huantian Lv, Hui Su, Jiacheng Li, Jiahao Liu, Jiahuan Li, Jiajun Yang, Jiaming Wang, Jian Yang, Jianchao Tan, Jiaqi Sun, Jiaqi Zhang, Jiawei Fu, Jiawei Yang, Jiaxi Hu, Jiayu Qin, Jingang Wang, Jiyuan He, Jun Kuang, Junhui Mei, Kai Liang, Ke He, Kefeng Zhang, Keheng Wang, Keqing He, Liang Gao, Liang Shi, Lianhui Ma, Lin Qiu, Lingbin Kong, Lingtong Si, Linkun Lyu, Linsen Guo, Liqi Yang, Lizhi Yan, Mai Xia, Man Gao, Manyuan Zhang, Meng Zhou, Mengxia Shen, Mingxiang Tuo, Mingyang Zhu, Peiguang Li, Peng Pei, Peng Zhao, Pengcheng Jia, Pingwei Sun, Qi Gu, Qianyun Li, Qingyuan Li, Qiong Huang, Qiyuan Duan, Ran Meng, Rongxiang Weng, Ruichen Shao, Rumei Li, Shizhe Wu, and Shuai Liang. Longcat-flash technical report. *CoRR*, abs/2509.01322, 2025. doi:[10.48550/ARXIV.2509.01322](https://doi.org/10.48550/ARXIV.2509.01322). URL <https://doi.org/10.48550/arXiv.2509.01322>.
- Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. QUEST: query-aware sparsity for efficient long-context LLM inference. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=KzACYwOMTV>.
- Shuo Yang, Ying Sheng, Joseph E. Gonzalez, Ion Stoica, and Lianmin Zheng. Post-training sparse attention with double sparsity. *CoRR*, abs/2408.07092, 2024. doi:[10.48550/ARXIV.2408.07092](https://doi.org/10.48550/ARXIV.2408.07092). URL <https://doi.org/10.48550/arXiv.2408.07092>.
- Jintao Zhang, Chendong Xiang, Haofeng Huang, Jia Wei, Haocheng Xi, Jun Zhu, and Jianfei Chen. Spargeattention: Accurate and training-free sparse attention accelerating any model inference. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net, 2025a. URL <https://openreview.net/forum?id=74c3Wwk8Tc>.
- Yutao Sun, Tianzhu Ye, Li Dong, Yuqing Xia, Jian Chen, Yizhao Gao, Shijie Cao, Jianyong Wang, and Furui Wei. Rectified sparse attention. *CoRR*, abs/2506.04108, 2025. doi:[10.48550/ARXIV.2506.04108](https://doi.org/10.48550/ARXIV.2506.04108). URL <https://doi.org/10.48550/arXiv.2506.04108>.
- Weilin Zhao, Zihan Zhou, Zhou Su, Chaojun Xiao, Yuxuan Li, Yanghao Li, Yudi Zhang, Weilun Zhao, Zhen Li, Yuxiang Huang, Ao Sun, Xu Han, and Zhiyuan Liu. Infllm-v2: Dense-sparse switchable attention for seamless short-to-long adaptation. *CoRR*, abs/2509.24663, 2025. doi:[10.48550/ARXIV.2509.24663](https://doi.org/10.48550/ARXIV.2509.24663). URL <https://doi.org/10.48550/arXiv.2509.24663>.
- DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Erhang Li, Fangqi Zhou, Fangyun Lin, Fucong Dai, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Li, Haofen Liang, Haoran Wei, Haowei Zhang, Haowen Luo, Haozhe Ji, Honghui Ding, Hongxuan Tang, Huanqi Cao, Huazuo Gao, Hui Qu, Hui Zeng, Jialiang Huang, Jiashi Li, Jiaxin Xu, Jiewen Hu, Jingchang Chen, Jingting Xiang, Jingyang Yuan, Jingyuan Cheng, Jinhua Zhu, Jun Ran, Junguang Jiang, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kaige Gao,

Kang Guan, Kexin Huang, Kexing Zhou, Kezhao Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Wang, Liang Zhao, Liangsheng Yin, Lihua Guo, Lingxiao Luo, Linwang Ma, Litong Wang, Liyue Zhang, M. S. Di, M. Y Xu, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Panpan Huang, Peixin Cong, Peiyi Wang, Qiancheng Wang, Qihao Zhu, Qingyang Li, Qinyu Chen, Qiushi Du, Ruiling Xu, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runqiu Yin, Runxin Xu, Ruomeng Shen, Ruoyu Zhang, S. H. Liu, Shanghao Lu, Shangyan Zhou, Shanhua Chen, Shaofei Cai, Shaoyuan Chen, Shengding Hu, Shengyu Liu, Shiqiang Hu, Shirong Ma, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, Songyang Zhou, Tao Ni, Tao Yun, Tian Pei, Tian Ye, Tianyuan Yue, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjie Pang, Wenjing Luo, Wenjun Gao, Wentao Zhang, Xi Gao, Xiangwen Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaokang Zhang, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xingyou Li, Xinyu Yang, Xinyuan Li, Xu Chen, Xuecheng Su, Xuehai Pan, Xuheng Lin, Xuwei Fu, Y. Q. Wang, Yang Zhang, Yanhong Xu, Yanru Ma, Yao Li, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Qian, Yi Yu, Yichao Zhang, Yifan Ding, Yifan Shi, Yiliang Xiong, Ying He, Ying Zhou, Yinmin Zhong, Yishi Piao, Yisong Wang, Yixiao Chen, Yixuan Tan, Yixuan Wei, Yiyang Ma, Yiyuan Liu, Yonglun Yang, Yongqiang Guo, Yongtong Wu, Yu Wu, Yuan Cheng, Yuan Ou, Yuanfan Xu, Yuduan Wang, Yue Gong, Yuhan Wu, Yuheng Zou, Yukun Li, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehua Zhao, Zehui Ren, Zhangli Sha, Zhe Fu, Zhehan Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhixian Huang, Zhiyu Wu, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihui Gu, Zijia Zhu, Zilin Li, Zipeng Zhang, Ziwei Xie, Ziyi Gao, Zizheng Pan, Zongqing Yao, Bei Feng, Hui Li, J. L. Cai, Jiaqi Ni, Lei Xu, Meng Li, Ning Tian, R. J. Chen, R. L. Jin, S. S. Li, Shuang Zhou, Tianyu Sun, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xinnan Song, Xinyi Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, Dongjie Ji, Jian Liang, Jianzhong Guo, Jin Chen, Leyi Xia, Miaojun Wang, Mingming Li, Peng Zhang, Ruyi Chen, Shangmian Sun, Shaoqing Wu, Shengfeng Ye, T. Wang, W. L. Xiao, Wei An, Xianzu Wang, Xiaowen Sun, Xiaoxiang Wang, Ying Tang, Yukun Zha, Zekai Zhang, Zhe Ju, Zhen Zhang, and Zihua Qu. Deepseek-v3.2: Pushing the frontier of open large language models, 2025. URL <https://arxiv.org/abs/2512.02556>.

Emmy Liu, Graham Neubig, and Chenyan Xiong. Midtraining bridges pretraining and posttraining distributions. *CoRR*, abs/2510.14865, 2025a. doi:[10.48550/ARXIV.2510.14865](https://doi.org/10.48550/arXiv.2510.14865). URL <https://doi.org/10.48550/arXiv.2510.14865>.

Charlie Zhang, Graham Neubig, and Xiang Yue. On the interplay of pre-training, mid-training, and rl on reasoning language models, 2025b. URL <https://arxiv.org/abs/2512.07783>.

Chengying Tu, Xuemiao Zhang, Rongxiang Weng, Rumei Li, Chen Zhang, Yang Bai, Hongfei Yan, Jingang Wang, and Xunliang Cai. A survey on LLM mid-training. *CoRR*, abs/2510.23081, 2025. doi:[10.48550/ARXIV.2510.23081](https://doi.org/10.48550/ARXIV.2510.23081). URL <https://doi.org/10.48550/arXiv.2510.23081>.

Zhuorui Liu, Chen Zhang, and Dawei Song. Zigzagattention: Efficient long-context inference with exclusive retrieval and streaming heads. *CoRR*, abs/2508.12407, 2025b. doi:[10.48550/ARXIV.2508.12407](https://doi.org/10.48550/ARXIV.2508.12407). URL <https://doi.org/10.48550/arXiv.2508.12407>.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fdb053c1c4a845aa-Abstract.html>.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. REPLUG: retrieval-augmented black-box language models. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, NAACL 2024, Mexico City, Mexico, June 16-21, 2024, pages 8371–8384. Association for Computational Linguistics, 2024. doi:[10.18653/V1/2024.NAACL-LONG.463](https://doi.org/10.18653/V1/2024.NAACL-LONG.463). URL <https://doi.org/10.18653/v1/2024.nacl-long.463>.

Howard Yen, Tianyu Gao, and Danqi Chen. Long-context language modeling with parallel context encoding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 2588–2610. Association for Computational Linguistics, 2024. doi:[10.18653/V1/2024.ACL-LONG.142](https://doi.org/10.18653/V1/2024.ACL-LONG.142). URL <https://doi.org/10.18653/v1/2024.acl-long.142>.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Xuanhe Zhou, Yufei Huang, Chaojun Xiao, Chi Han, Yi R. Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian,

Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi, Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Guoliang Li, Zhiyuan Liu, and Maosong Sun. Tool learning with foundation models. *ACM Comput. Surv.*, 57(4):101:1–101:40, 2025. doi:10.1145/3704435. URL <https://doi.org/10.1145/3704435>.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. Tora: A tool-integrated reasoning agent for mathematical problem solving. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=Ep0TtjVoap>.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. Duoattention: Efficient long-context LLM inference with retrieval and streaming heads. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL <https://openreview.net/forum?id=cFu7ze7xUm>.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL <https://openreview.net/forum?id=rJ1-b3RcF7>.

Dacheng Li, Rulin Shao, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang Xuezhe Ma. How long can open-source llms truly promise on context length?, June 2023a. URL <https://lmsys.org/blog/2023-06-29-longchat>.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=wHBfxhZulu>.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL [http://papers.nips.cc/paper\\_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html).

OpenAI. Reinforcement fine-tuning. <https://platform.openai.com/docs/guides/reinforcement-fine-tuning>, 2024. Accessed: 2025-12-24.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2021.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. C-Eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*, 2023.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. CMMLU: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*, 2023b.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Lingxiao Diao, Xinyue Xu, Wanxuan Sun, Cheng Yang, and Zhuosheng Zhang. GuideBench: Benchmarking domain-oriented guideline following for LLM agents. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11361–11399, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi:10.18653/v1/2025.acl-long.557. URL <https://aclanthology.org/2025.acl-long.557/>.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.

MAA. Aime 2024, 2024. URL <https://maa.org/math-competitions/american-invitational-mathematics-examination-aime>.

MAA. Aime 2025, 2025. URL <https://artofproblemsolving.com/wiki/index.php/AIMEProblemsandSolutions>.

- ByteDance-Seed. Beyondaime: Advancing math reasoning evaluation beyond high school olympiads, 2025.  
URL <https://huggingface.co/datasets/ByteDance-Seed/BeyondaIME>.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.
- Jiawei Liu, Songrun Xie, Junhao Wang, Yuxiang Wei, Yifeng Ding, and Lingming Zhang. Evaluating language models for efficient code generation. *arXiv preprint arXiv:2408.06450*, 2024a.
- Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Siyao Liu, He Zhu, Jerry Liu, Shulin Xin, Aoyan Li, Rui Long, Li Chen, Jack Yang, Jinxiang Xia, Z. Y. Peng, Shukai Liu, Zhaoxiang Zhang, Ge Zhang, Wenhao Huang, Kai Shen, and Liang Xiang. Fullstack bench: Evaluating llms as full stack coders, 2024b. URL <https://arxiv.org/abs/2412.00535>.
- Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*, 2024.
- The Terminal-Bench Team. Terminal-bench: A benchmark for ai agents in terminal environments, Apr 2025.  
URL <https://github.com/laude-institute/terminal-bench>.
- Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan.  $\tau^2$ -bench: Evaluating conversational agents in a dual-control environment. *arXiv preprint arXiv:2506.07982*, 2025.
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners, 2022. URL <https://arxiv.org/abs/2210.03057>.
- Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3639–3664, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi:10.18653/v1/2025.acl-long.183. URL <https://aclanthology.org/2025.acl-long.183>.
- Kiran Vodrahalli, Santiago Ontanon, Nilesh Tripuraneni, Kelvin Xu, Sanil Jain, Rakesh Shivanna, Jeffrey Hui, Nishanth Dikkala, Mehran Kazemi, Bahare Fatemi, Rohan Anil, Ethan Dyer, Siamak Shakeri, Roopal Vij, Harsh Mehta, Vinay Ramasesh, Quoc Le, Ed Chi, Yifeng Lu, Orhan Firat, Angeliki Lazaridou, Jean-Baptiste Lespiau, Nithya Attaluri, and Kate Olszewska. Michelangelo: Long context evaluations beyond haystacks via latent structure queries, 2024. URL <https://arxiv.org/abs/2409.12640>.
- Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. Helmet: How to evaluate long-context language models effectively and thoroughly, 2025. URL <https://arxiv.org/abs/2410.02694>.
- S.J. Paech. Longform creative writing benchmark, 2025. URL <https://github.com/EQ-bench/longform-writing-bench>. GitHub repository.
- Pengcuo Dege, Qiuming Luo, Rui Mao, and Chang Kong. Flashmla-etap: Efficient transpose attention pipeline for accelerating MLA inference on NVIDIA H20 gpus. *CoRR*, abs/2506.01969, 2025. doi:10.48550/ARXIV.2506.01969. URL <https://doi.org/10.48550/arXiv.2506.01969>.
- Han Li, Xinyu Peng, Yaoming Wang, Zelin Peng, Xin Chen, Rongxiang Weng, Jingang Wang, Xunliang Cai, Wenrui Dai, and Hongkai Xiong. Onecat: Decoder-only auto-regressive model for unified understanding and generation. *arXiv preprint arXiv:2509.03498*, 2025.