# Recall with Reasoning: Chain-of-Thought Distillation for Mamba's Long-Context Memory and Extrapolation

**Jun-Yu Ma***, **Tianqing Fang**✉, **Zhisong Zhang**, **Hongming Zhang**, **Haitao Mi**, **Dong Yu**
Tencent AI Lab
*junyu129@outlook.com, ✉tianqfang@tencent.com

## Abstract

Mamba's theoretical infinite-context potential is limited in practice when sequences far exceed training lengths. This work explores unlocking Mamba's long-context memory ability by a simple-yet-effective method, Recall with Reasoning (RwR), by distilling chain-of-thought (CoT) summarization from a teacher model. Specifically, RwR prepends these summarization as CoT prompts during fine-tuning, teaching Mamba to actively recall and reason over long contexts. Experiments on LONGMEMEVAL and HELMET show that RwR outperforms existing long-term memory methods on the Mamba model. Furthermore, under similar pre-training conditions, RwR improves the long-context performance of Mamba relative to comparable Transformer/hybrid baselines while preserving short-context capabilities, all without changing the architecture.

## 1 Introduction

Transformer-based Large Language Models (LLMs) (Vaswani et al., 2017; Touvron et al., 2023a,b) have demonstrated significant capability in various real-world tasks, but suffer from quadratic complexity and poor length extrapolation. In contrast, Mamba (Gu and Dao, 2023) adopts a recurrent inference mode that ensures linear complexity and unlimited input length. However, despite having a theoretical capability of global memorization, empirical studies (Waleffe et al., 2024; Ben-Kish et al., 2024; Ye et al.; Yuan et al., 2024) have shown that Mamba struggles with long-context memory when the length of the processed text exceeds the model's training length.

To address this issue, efforts have focused on compressing unimportant tokens to reduce their negative impact. DeciMamba (Ben-Kish et al., 2024) utilized the selective time-steps of Mamba to filter out unimportant tokens, while ReMamba (Yuan et al., 2024) used the similarity between query and tokens to select important tokens and remove unimportant ones. However, in practice, long-context memory and extrapolation challenges persist when the input length significantly exceeds the training length of Mamba, even after applying such filtering. Moreover, since the original input sentences are compromised, Mamba's language modeling performance may also be negatively impacted in this way (See Table 3).

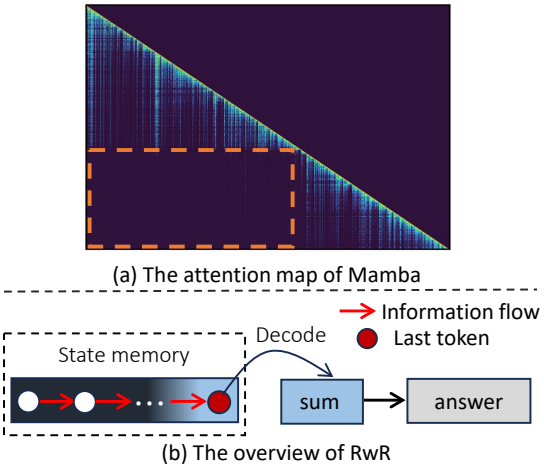


Figure 1: (a) The simulated "attention map" of Mamba[1]. The orange rectangle shows that when Mamba encodes a long text, the representation of the current token is difficult to include the information of the previous token that is far away from it. (b) The state information of the text is gradually gathered from the beginning to the end to the last token. RwR aims to first decode the last token state to obtain a shorter summary, which can be fully accessed by Mamba when answering questions.

Instead of changing the logic of state update, we aims to unlock the long-context memory ability of Mamba in a data-driven chain-of-thought (Wei et al., 2022; DeepSeek-AI, 2025; Jin et al., 2024) paradigm. Conceptually, our approach considers

*Work done during intenship at Tencent AI Lab.

[1]A variant of self-attention in Ben-Kish et al. (2024).



*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4715–4721
November 4-9, 2025 

two key properties of Mamba: (1) its implicit attention mechanism naturally prioritizes recent tokens during decoding (Ben-Kish et al., 2024) (Figure 1(a)), and (2) the state representation of the last token in the input theoretically encodes the complete history through selective state transitions (Figure 1(b)). Therefore, Recall with Reasoning (RwR) is proposed, which teaches Mamba model to first decode relevant context from Mamba's fixed-size state memory, then performs reasoning based on this distilled historical summary. Specifically, for a pair of long context and query, a more capable Transformer model is employed to generate relevant summary that are needed to answer the query. Such summary is then placed after the context and query as a new entry of training data for further fine-tuning. By augmenting a Supervised Fine-tuning (SFT) dataset with such query-aware summary CoT, Mamba can effectively identify key information from long texts through CoT thinking, thus significantly improving its ability to recall long-context memory. Furthermore, during the inference phase, a simple yet effective strategy of breaking down longer context into smaller pieces is used, which can further improve the performance.

To assess long-context memory and extrapolation ability, two benchmarks are adopted: a chat-form benchmark LONGMEMEVAL (Wu et al., 2025), and an application-centric benchmark HELMET (Yen et al., 2024). Experimental results verify that RwR effectively improves the long-context memory ability of Mamba. In addition, experiments on short-context language modeling tasks, like the RTE (Dagan et al., 2005), GSM8K (Cobbe et al., 2021), Natural Question (Kwiatkowski et al., 2019), and SAMSum (Gliwa et al., 2019), show that our method does not affect the basic language modeling ability of Mamba.

## 2 Related Work

### 2.1 Related Work

**State Space Models** Gu et al. (2022) proposed the S4 model, which is a promising alternative to transformers for capturing long-term dependencies. Building on S4, Gu and Dao (2023) introduced the data-dependent State Space Model (SSM) layer S6, and developed the Mamba (Gu and Dao, 2023) language model backbone. Mamba outperforms transformers of various sizes on large-scale real-world data and scales linearly with sequence length. Additionally, Dao and Gu (2024) provided insights into the performance of recent SSMs compared to transformers in the context of language modeling. Hybrid models like Jamba (Lieber et al., 2024) and Samba (Ren et al., 2024) aim to combine the strengths of attention mechanisms with Mamba's efficient long-range dependency modeling.

**Long-context Memory** Some studies (Peng et al., 2024; Li et al., 2024; Ben-Kish et al., 2024) have highlighted that language models trained on fixed-length contexts tend to suffer performance degradation when applied to longer contexts. Transformer-based models, in particular, face significant computational challenges as the context length increases. For state space models, DeciMamba (Ben-Kish et al., 2024) introduced a non-training method to filter out less important tokens, effectively reducing the input length. Additionally, Yuan et al. (2024) proposed a technique where a network is trained to compress and selectively retain essential information during the initial forward pass. Meanwhile, Ye et al. investigated the fundamental limitations of Mamba in handling long sequences and presented a principled approach to address these issues on various tasks.

**CoT Distillation** The distillation-based transfer of CoT ability to small language models (SLMs) has emerged as a prominent research direction. The predominant methodology leverages CoT prompting to extract rationales from large-scale teacher models (e.g., GPT-4), subsequently transferring these rationales to SLMs via fine-tuning (Ho et al., 2023; Pezeshkpour et al., 2023). Building upon these foundations, Magister et al. (2023) systematically investigated reasoning enhancement across multiple model architectures, empirically establishing the scaling laws governing student model ability and training data volume. In parallel, Wang et al. (2023) addressed the critical challenge of ensuring fidelity and consistency in rationale generation through constrained decoding optimization.

## 3 Method

In this section, we present Recall with Reasoning (RwR). The overview of the framework is presented in Figure 2.

### 3.1 Summary-based CoT Construction

**Valid Summary Extraction** Since Mamba (Gu and Dao, 2023)[2] has not undergone instruction

[2]https://huggingface.co/state-spaces/mamba-2.8b

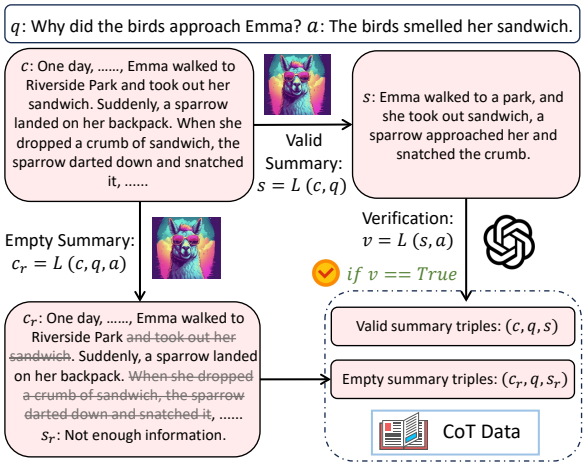


Figure 2: The data generation of RwR. $c$ is the context, $q$ is the query for $c$, and $a$ is the ground-truth answer. $c_r$ is the context in which the correct answer is removed. $v$ refers to whether the valid summary contains correct answer. $L$ refers to the LLM used in this step. "....." refers to a part of the context. The gray part is the deleted text containing the answer to the query.

tuning, following Yuan et al. (2024), we utilized OpenOrca (Mukherjee et al., 2023) as the SFT dataset, which primarily consists of question-answering data. Based on this dataset, context-query-summary triples are constructed to enhance the model's ability to recall key information from long contexts. First, examples that contain background context relevant to the given query are selected from the dataset. Formally, let $c$ denote the context, $q$ the query, and $a$ the ground-truth answer. For simplicity, these selected examples are represented as $D = \{(c, q, a)\}$. Next, as shown in Figure 2, for each example $e \in D$, a Transformer model[3] is employed to extract a summary of the context that is relevant to the query. Since mamba is not trained with SFT, it cannot be used directly to extract summary here. The prompt used for this task is: "<$c$> <$q$> Please extract a note relevant to the query:" The extracted summary is denoted as $s$. To ensure the quality of the generated summary $s$, GPT-4o is used to verify their consistency with the correct answers $a$. Samples with inconsistencies are filtered out. Finally, the filtered examples are denoted as $D_f = \{(c, q, s)\}$.

**Empty Summary Construction** In real-world scenarios, not all queries can be answered based on the context provided. Training the model only on the above dataset may lead to overconfidence. For example, when context information is insufficient to answer the query, it will fail to answer: "There is no enough information here", but it will forcibly generate wrong answers. To mitigate this problem, we construct examples where the given context does not contain the answer to the query to enhance the model's ability to distinguish between relevant and irrelevant paragraphs. As shown in Figure 2, $e \in D$ is selected, and then the Llama-3.1-8B-Instruct is used to locate the paragraphs in context $c$ that contain the correct answer. These paragraphs are removed to obtain the modified context $c_r$. This process results in a set of empty-summary data, denoted as $D_e = \{(c_r, q, empty)\}$. Subsequently, the summary CoT data is the combination of $D_f$ and $D_e$, which is denoted as $D_s = [D_f, D_e]$.

Finally, Mamba is trained using the OpenOrca dataset along with the constructed dataset $D_s$. In this way, the model can unlock the ability of recalling from long input context via CoT.

### 3.2 Segmented Summarization for Answering

For scenarios with very long input lengths, a simple yet effective strategy is to segment longer context into smaller pieces. First, the long context is divided into multiple parts. Since the trained model has the ability to extract summaries, a summary is generated for each part. Finally, these summaries are fed together into the model to answer the query. This approach ensures that each processing step remains within a manageable length, which benefits the model's memory. Moreover, Mamba has linear computational complexity as the input length increases, so this strategy does not increase the demand for computing resources.

## 4 Experiments

### 4.1 Experiment Settings

**Evaluation Datasets & Metrics** LONGMEMEVAL was a challenging benchmark designed to evaluate the long-context memory ability of chat assistants across six memory tasks, including single-session-user (SU), single-session-assistant (SA), single-session-preference (SP), multi-session reasoning (MR), knowledge update (KU), temporal reasoning (TR). It also contains two datasets of different length. HELMET was a comprehensive benchmark for long-context language models covering seven diverse categories of tasks. This paper selected several tasks in HELMET and set the evaluation length to 16k.

Following previous work, the LONG-

[3]In this paper, the Llama-3.1-8B-Instruct was used.

| Tasks | SU | SA | SP | MR | KU | TR | Avg |
|---|---|---|---|---|---|---|---|
| Orcale(10k length) | | | | | | | |
| Mamba (Pre) | 0 | 1.8 | 0 | 0.75 | 0 | 0 | 0.4 |
| Mamba (SFT) | 31.4 | 39.2 | 6.7 | 8.3 | 21.8 | 14.3 | 18.6 |
| DeciMamba | 47.0 | 41.1 | 6.7 | 3.0 | 19.2 | 14.3 | 19.3 |
| ReMamba | 45.7 | 33.9 | **13.3** | 10.5 | 28.2 | 23.2 | 24.4 |
| RwR | **48.6** | **44.6** | 10.0 | **13.5** | **33.3** | **24.1** | **27.6** |
| S(100k length) | | | | | | | |
| Mamba (Pre) | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Mamba (SFT) | 11.4 | 5.7 | 0 | 4.5 | 10.3 | 11.2 | 8.0 |
| +SSA | 7.1 | 7.1 | 0 | 5.3 | 5.1 | 9.8 | 6.6 |
| DeciMamba | 0 | 0 | 0 | 0 | 0 | 0.75 | 0.2 |
| ReMamba | 2.9 | 1.8 | 0 | 3.8 | 7.7 | 4.5 | 4.0 |
| RwR | 10.0 | 7.1 | 0 | 6.3 | 11.5 | **16.3** | 9.8 |
| +SSA | **12.9** | **8.9** | 0 | **7.5** | **16.7** | 15.0 | **11.4** |

Table 1: The evaluation results (%) on LONGMEMEVAL. We reported the average results of three rounds. ORCALE dataset contains only context related to the question, while the S dataset also contains some irrelevant context. "SSA" refers to the proposed strategy in Section 3.2. Since the number of examples is different in different tasks, Avg is the weighted average.

| Tasks | RAG | ICL | SR | Avg |
|---|---|---|---|---|
| Mamba (Pre) | 0 | 0 | 0 | 0 |
| Mamba (SFT) | 44.6 | 39.0 | 0.5 | 28.0 |
| DeciMamba | 39.8 | 20.0 | **3.5** | 21.1 |
| ReMamba | 41.8 | 52.0 | 1.7 | 31.8 |
| RwR | **47.3** | **54.0** | 1.6 | **34.3** |

Table 2: The evaluation results (%) on HELMET. The length of examples is set to 16k.

MEMEVAL (Wu et al., 2025), was evaluated by GPT-4o. For the HELMET benchmark, the Retrieval-augmented generation (RAG) and Synthetic recall (SR) tasks was evaluated by SubEM, and the Many-shot in-context learning (ICL) task was evaluated by Accuracy. For short context tasks, the metrics were as follows:

**Reasoning** on the GSM8K (Cobbe et al., 2021), and the results were measured by solve rate.

**Summarization** on the SAMSum (Gliwa et al., 2019), and the results were measured by the average of ROUGE-1, ROUGE-2 and ROUGE-L.

**Open-domain QA** on the Natural Question (Kwiatkowski et al., 2019), and the results were measured by exact match (EM) with the reference answer after minor normalization as in Chen et al. (2017) and Lee et al. (2019).

**Natural language inference (NLI)** on the RTE (Dagan et al., 2005), and the results were measured by accuracy of two-way classification.

**Training Details** Mamba-2.8b (Gu and Dao, 2023) was used as the backbone. OpenOrca was used as SFT data. To accommodate device memory constraints, the training examples were truncated to a maximum length of 6,000. Finally, 100,000 OpenOrca data and 10,000 constructed summary data were used for the proposed RwR.

**Baselines** We included the untuned Mamba, Mamba fine-tuned on the OpenOrca dataset, Decimamba (Ben-Kish et al., 2024) and Remamba (Yuan et al., 2024) as baselines. Decimamba and Remamba were compression methods designed for the long-context memory of Mamba.

### 4.2 Results

**Long-context Memory Tasks** Table 1 and 2 reported the long-context memory evaluation results on two benchmarks. We observed that while previous long-context memory and extrapolation methods (e.g., DeciMamba and ReMamba) improve performance in the 10k length setting, their effectiveness decreased significantly in the 100k length setting and even underperformed directly fine-tuned Mamba (SFT). This suggests that existing methods have notable limitations in extending the Mamba's context length. In nearly all tasks, RwR enhanced the performance of Mamba across all context lengths. This demonstrates that CoT can effectively extend the model's processing length and improve its long-context memory ability. Furthermore, for the 100k length settings, our SSA strategy further improved the performance for RwR. However, for the Mamba (SFT) that was trained without our constructed data, using this strategy results in a decrease in performance. This shows that our method can effectively improve the model's ability to extract summaries, thereby indirectly improving the model's long-context memory ability.

**Short-context Tasks** In order to verify whether the proposed method has negative effects while improving long-context memory and extrapolation ability, several short-context tasks were selected for evaluation, and the results were shown in Table 3. As shown in the table, compared with Mamba (SFT), the short-context language modeling ability of our method RwR has been slightly improved. However, the short-context abilities of DeciMamba and ReMamba were significantly reduced, which indicates that the previous compression methods affect the language modeling ability of Mamba and bring challenges in practical application.

| Tasks | Dialogue | NLI | Reasoning | Open-QA |
|---|---|---|---|---|
| Mamba (Pre) | 25.2 | 45.7 | 71.5 | 21.5 |
| Mamba (SFT) | 23.3 | 40.6 | 88.5 | 23.4 |
| DeciMamba | 24.7 | 36.5 | 92.8 | 9.9 |
| ReMamba | 5.1 | **50.5** | 61.5 | 21.4 |
| RwR | **28.1** | 46.9 | **93.0** | **23.7** |

Table 3: The results (%) on short-context tasks.

### 4.3 Other Architectures Study

**Extrapolation study** In order to verify the superiority of Mamba in extrapolation ability compared with other model architectures, length extrapolation experiments were conducted on the Transformer model and the hybrid SSM-Transformer model. Specifically, we selected Transformer model Phi-2 (Javaheripi et al., 2023) and hybrid model Hymba (Dong et al., 2024), as they closely match Mamba in both the pre-training context length and model size. Then these models were fine-tuned using the same data as those used by Mamba in this study. The performance of the fine-tuned Phi-2 and Hymba models was evaluated on the LONGMEMEVAL benchmark. As shown in Table 4, at 10k length, the average performance of both Phi-2 and Hymba was slightly worse than that of Mamba. However, their performance was almost 0 at the 100k length setting, significantly lower than that of Mamba, which indicates that their length extrapolation ability is very limited. In addition, the Phi-2 model only achieved certain performance on the single-session-user (SU) and single-session-assistant (SA) tasks. These suggest that the Phi-2 model retains only some simple ability during extrapolating the length and is not generalizable.

**Efficiency study** The rightmost column of Table 4 presents the average time taken by different models to process samples of varying dataset lengths. For a data length of 10k, the time differences between models were minimal, with the Transformer model requiring less time than the hybrid model. This is likely due to the more complex structure of the Hymba model, which requires additional processing steps. However, when processing data of length 100k setting, the Transformer model took significantly longer than the other models, which highlights the efficiency of the SSM model in processing long texts.

## 5 Conclusion

This paper focuses on the long-context memory and extrapolation of Mamba. A method called RwR is proposed to guide the CoT of Mamba to focus on summarizing and identifying key information in the previous context, thereby enhancing memory ability. Experiments on the LONGMEMEVAL and HELMET datasets demonstrate that the proposed method effectively enhances the model's long-context memory abilities, and in the meantime retaining the basic language modeling ability on other short-context tasks. Further analysis shows that Mamba has better length extrapolation ability than the Transformer and hybrid models.

| Tasks | SU | SA | SP | MR | KU | TR | Avg | Time |
|---|---|---|---|---|---|---|---|---|
| Orcale(10k length) | | | | | | | | |
| RwR | 48.6 | 44.6 | 10.0 | **13.5** | **33.3** | **24.1** | **27.6** | 1.7s |
| Phi-2 | **61.4** | **67.8** | 0 | 1.5 | 5.1 | 4.5 | 18.5 | 2.5s |
| Hymba | 40.0 | 46.4 | 13.3 | 6.0 | 35.9 | 22.6 | 24.8 | 4.3s |
| S(100k length) | | | | | | | | |
| RwR | **10.0** | **7.1** | 0 | **6.3** | **11.5** | **16.3** | **9.8** | 10.8s |
| Phi-2 | 0 | 0 | 0 | 0.75 | 1.3 | 0 | 0.4 | 30.6s |
| Hymba | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 17.5s |

Table 4: The evaluation results (%) of SSM model RwR, Transformer model Phi-2 and hybrid model Hymba on the LONGMEMEVAL. "Time" refers to the average time consumed by each sample calculation.

## Limitations

There are several limitations for this paper. First, this paper only conducts experiments on Mamba-2.8b, but whether it is effective on other SSM models such as Mamba2 (Dao and Gu, 2024) or Falcon mamba (Zuo et al., 2024) is still unknown and needs to be explored in future work. Second, the longest test length in this paper is about 100k, but longer lengths, such as 200k, are not explored due to computational costs. Third, since the pre-training length of Mamba is limited to 2k, which is much shorter than more advanced Transformer models (such as Llama-3.3, which has a pre-training sequence length of up to 128k), the current Mamba model is not comparable to the state-of-the-art Transformer models.

## References

Assaf Ben-Kish, Itamar Zimerman, Shady Abu-Hussein, Nadav Cohen, Amir Globerson, Lior Wolf, and Raja Giryes. 2024. Decimamba: Exploring the length extrapolation potential of mamba. *CoRR*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual*

*Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1870–1879. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer.

Tri Dao and Albert Gu. 2024. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Xin Dong, Yonggan Fu, Shizhe Diao, Wonmin Byeon, Zijia Chen, Ameya Sunil Mahabaleshwarkar, Shih-Yang Liu, Matthijs Van Keirsbilck, Min-Hung Chen, Yoshi Suhara, et al. 2024. Hymba: A hybrid-head architecture for small language models. *arXiv preprint arXiv:2411.13676*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *CoRR*, abs/2312.00752.

Albert Gu, Karan Goel, and Christopher Ré. 2022. Efficiently modeling long sequences with structured state spaces. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14852–14882. Association for Computational Linguistics.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. 2023. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 1(3):3.

Mingyu Jin, Weidi Luo, Sitao Cheng, Xinyi Wang, Wenyue Hua, Ruixiang Tang, William Yang Wang, and Yongdong Zhang. 2024. Disentangling memory and reasoning ability in large language models. *CoRR*, abs/2411.13504.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6086–6096. Association for Computational Linguistics.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024. Long-context llms struggle with long in-context learning. *CoRR*, abs/2404.02060.

Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, Omri Abend, Raz Alon, Tomer Asida, Amir Bergman, Roman Glozman, Michael Gokhman, Avashalom Manevich, Nir Ratner, Noam Rozen, Erez Shwartz, Mor Zusman, and Yoav Shoham. 2024. Jamba: A hybrid transformer-mamba language model. *CoRR*, abs/2403.19887.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adámek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1773–1781. Association for Computational Linguistics.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of GPT-4. *CoRR*, abs/2306.02707.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024. 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.

Pouya Pezeshkpour, Hayate Iso, Thom Lake, Nikita Bhutani, and Estevam Hruschka. 2023. Distilling large language models using skill-occupation graph context for hr-related tasks. *CoRR*, abs/2311.06383.

Liliang Ren, Yang Liu, Yadong Lu, Yelong Shen, Chen Liang, and Weizhu Chen. 2024. Samba: Simple hybrid state space models for efficient unlimited context language modeling. *CoRR*, abs/2406.07522.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *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.

Roger Waleffe, Wonmin Byeon, Duncan Riach, Brandon Norick, Vijay Korthikanti, Tri Dao, Albert Gu, Ali Hatamizadeh, Sudhakar Singh, Deepak Narayanan, Garvit Kulshreshtha, Vartika Singh, Jared Casper, Jan Kautz, Mohammad Shoeybi, and Bryan Catanzaro. 2024. An empirical study of mamba-based language models. *CoRR*, abs/2406.07887.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023. SCOTT: self-consistent chain-of-thought distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5546–5558. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*. OpenReview.net.

Zhifan Ye, Kejing Xia, Yonggan Fu, Xin Dong, Jihoon Hong, Xiangchi Yuan, Shizhe Diao, Jan Kautz, Pavlo Molchanov, and Yingyan Celine Lin. Longmamba: Enhancing mamba's long-context capabilities via training-free receptive field enlargement. In *The Thirteenth International Conference on Learning Representations*.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. 2024. HELMET: how to evaluate long-context language models effectively and thoroughly. *CoRR*, abs/2410.02694.

Danlong Yuan, Jiahao Liu, Bei Li, Huishuai Zhang, Jingang Wang, Xunliang Cai, and Dongyan Zhao. 2024. Remamba: Equip mamba with effective long-sequence modeling. *CoRR*, abs/2408.15496.

Jingwei Zuo, Maksim Velikanov, Dhia Eddine Rhaiem, Ilyas Chahed, Younes Belkada, Guillaume Kunsch, and Hakim Hacid. 2024. Falcon mamba: The first competitive attention-free 7b language model. *Preprint*, arXiv:2410.05355.