

# MiniCPM-SALA: Hybridizing Sparse and Linear Attention for Efficient Long-Context Modeling

MiniCPM Team

 <https://huggingface.co/openbmb/MiniCPM-SALA>  
 <https://github.com/OpenBMB/MiniCPM>

## Abstract

The evolution of large language models (LLMs) towards applications with ultra-long contexts faces challenges posed by the high computational and memory costs of the Transformer architecture. While existing sparse and linear attention mechanisms attempt to mitigate these issues, they typically involve a trade-off between memory efficiency and model performance. This paper introduces MiniCPM-SALA<sup>a</sup>, a hybrid architecture that integrates the high-fidelity long-context modeling of sparse attention (InfLLM-V2) with the global efficiency of linear attention (Lightning Attention). By employing a layer selection algorithm to integrate these mechanisms in a 1:3 ratio and utilizing a hybrid positional encoding (HyPE), the model maintains efficiency and performance for long-context tasks. Furthermore, we introduce a cost-effective continual training framework that transforms pre-trained Transformer-based models into hybrid models, which reduces training costs by approximately 75% compared to training from scratch. Extensive experiments show that MiniCPM-SALA maintains general capabilities comparable to full-attention models while offering improved efficiency. On a single NVIDIA A6000D GPU, the model achieves inference speeds up to  $3.5\times$  faster than full attention models at the sequence length of 256K tokens and supports context lengths of up to 1M tokens, a scale where traditional full-attention 8B models fail because of memory constraints.

<sup>a</sup>SALA stands for Sparse Attention and Linear Attention.

## 1 Introduction

As large language models (LLMs) (OpenAI et al., 2024; Comanici et al., 2025; Grattafiori et al., 2024; Yang et al., 2025a; DeepSeek-AI et al., 2025) become increasingly effective, the application scenarios of LLMs are undergoing a profound paradigm shift, transitioning from simple question-answering (Brown et al., 2020) to more advanced applications, such as deep understanding and generation of ultra-long contexts (Bai et al., 2024, 2025; Zhou et al., 2025; Shao et al., 2024), repository-scale code engineering (Guo et al., 2024; Jimenez et al., 2024; Liu et al., 2024), and long-horizon agents for complex tasks (Qian et al., 2024; Mialon et al., 2023; Li et al., 2026). For these advanced applications, models are no longer confined to processing fragmented information. Instead, they must demonstrate the capacity to handle ultra-long contexts, such as grasping entire technical manuals at once, analyzing comprehensive project dependency trees containing tens of thousands of lines of code, and maintaining coherent task states and memory over multi-day human-AI collaborations. This pursuit of holistic contextual information makes the ability to process millions of tokens a critical aspect for advanced LLMs (Kimi Team et al., 2025; NVIDIA et al., 2025b).

However, the Transformer architecture (Vaswani et al., 2017), which is the foundation of modern LLMs, encounters severe computational bottlenecks when handling ultra-long contexts due to its core full-attention mechanism. This bottleneck manifests primarily in two dimensions: (1) the *Compute Bottleneck* of computational complexity: for the standard attention mechanism, the computational cost grows quadratically with the sequence length  $N$ , i.e., its complexity is  $\mathcal{O}(N^2)$ . When the context scales to the level of millions of tokens, the huge overhead causes the inference latency to increase dramatically; (2) the *Memory Bottleneck* of KV-Cache: during the auto-regressive generation process, the model must store the key and value states (KVs) of all historical contextual tokens to avoid redundant computation. For a typical 8B-parameter model, even

when utilizing Grouped Query Attention (GQA) (Ainslie et al., 2023), the KV-Cache required for millions of tokens can reach dozens or even hundreds of gigabytes.

To address the aforementioned challenges, existing solutions have developed two primary paradigms: Sparse Attention (Yuan et al., 2025; DeepSeek-AI et al., 2025; Xiao et al., 2024; Zhao et al., 2025) and Linear Attention (Yang et al., 2024a; Gu & Dao, 2024; Peng et al., 2023; Yang et al., 2024b, 2025b). Both paradigms present distinct advantages and inherent limitations. Sparse attention methods attempt to break the compute bottleneck by computing only the most salient portions of the attention matrix, such as adopting sliding windows or global anchors. However, these methods are hindered by a “sparse computation, dense storage” limitation. While local computation reduces immediate processing overhead, the model must still retain the full KV-Cache to support contextual information retrieval. Linear attention utilizes recurrent formulations to successfully reduce computational complexity to  $\mathcal{O}(N)$ . Nevertheless, this extreme efficiency is achieved by the lossy compression of contextual information and inevitably results in performance degradation.

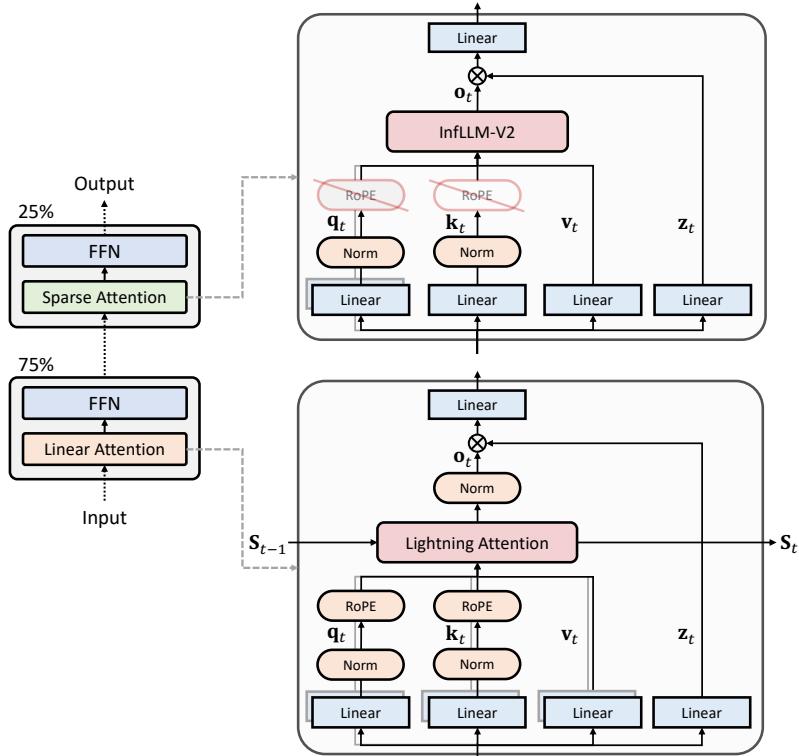
MiniCPM-SALA employs a hybrid architecture of sparse and linear attention (Chen et al., 2026), specifically designed to achieve efficient ultra-long sequence modeling. This architecture combines the high-fidelity long-context modeling capabilities of InfLLM-V2 (Zhao et al., 2025) and the global computational efficiency of Lightning Attention (Qin et al., 2024). Through this integrated approach, the model significantly mitigates inference overhead and memory consumption, while simultaneously addressing the precision bottleneck typical of pure linear architectures in long-range information processing. Consequently, MiniCPM-SALA provides a balanced solution that maintains both efficiency and high performance for long-context tasks. Furthermore, we employ the continual training paradigm to transform a pre-trained Transformer model into our hybrid model. By eschewing training from scratch, this approach significantly reduces the computational costs of model development. While several works have begun exploring the integration of sparse and linear attention (Hu et al., 2025; Hou et al., 2025; He & Garner, 2025), to the best of our knowledge, MiniCPM-SALA is the first to demonstrate through large-scale experimentation that these hybrids can match the performance of full-attention baselines. Furthermore, the model exhibits high efficiency and strong performance in long-context processing.

In summary, the main contributions of this study can be outlined as follows:

- We introduce a Sparse-Linear hybrid attention mechanism integrating 25% InfLLM-V2 and 75% Lightning Attention to strike a balance between throughput and precision. By leveraging the granular focus of sparse attention for local details and the  $\mathcal{O}(N)$  efficiency of linear attention for broad context, the architecture maintains high semantic accuracy as the sequence length scales up.
- We demonstrate that the Transformer-to-hybrid paradigm is a highly effective strategy for building strong hybrid models. This approach circumvents the inefficiencies of cold-start training by performing an architectural transformation on the pre-trained weights, thereby reducing the total training budget to approximately 25% relative to training a comparable model from scratch.
- We adopt HyPE (Hybrid Positional Encoding) (Chen et al., 2026) to effectively harmonize the performance across both short and long contexts. While maintaining general capabilities (e.g., knowledge, mathematics, and coding) comparable to modern full-attention models like Qwen3-8B, MiniCPM-SALA has substantial advantages across multiple long-context benchmarks.
- MiniCPM-SALA demonstrates substantial resource savings and speed advantages in long-context scenarios. On the NVIDIA A6000D GPU, MiniCPM-SALA achieves up to  $3.5\times$  the inference speed of Qwen3-8B at a sequence length of 256K tokens. Furthermore, MiniCPM-SALA supports inference at context lengths of up to 1M tokens on both NVIDIA A6000D and 5090 GPUs, whereas Qwen3-8B fails at this length due to out-of-memory (OOM) errors. These results demonstrate the broad prospects of MiniCPM-SALA in edge-side information-intensive applications.

## 2 Model Development

In this section, we introduce the model architecture and training strategies for MiniCPM-SALA. Specifically, we combine the efficient sparse attention for long-context modeling and linear attention for global efficiency in MiniCPM-SALA. Moreover, we also introduce an efficient training method, which can transform a standard Transformer model into sparse-linear hybrid attention.



**Figure 1:** Architecture of MiniCPM-SALA. The model adopts an efficient hybrid design that combines InfLLM-V2 (Zhao et al., 2025) and Lightning Attention (Qin et al., 2024) modules in a 1:3 ratio. Building on an intermediate MiniCPM-4.0 (MiniCPM-Team et al., 2025) checkpoint, MiniCPM-SALA undergoes a continual training phase to convert a standard Transformer model into a sparse-linear hybrid model.

## 2.1 Model Architecture

The overall architecture of MiniCPM-SALA is illustrated in Figure 1. MiniCPM-SALA adopts a hybrid architecture that interleaves sparse attention layers and linear attention layers. We retain the Feed-Forward Network (FFN) block after each attention block in the Transformer architecture to ensure high-capacity knowledge representation. Inspired by the architectural designs of recent representative studies, such as Qwen3-Next (Qwen Team, 2025) and Kimi-Linear (Kimi Team et al., 2025), as well as our internal small-scale preliminary experiments, we employ a 1:3 mixing ratio: 25% of the layers adopt sparse attention while the remaining 75% employ linear attention.

This hybrid configuration leverages the complementary strengths of both attention mechanisms. Linear attention layers have constant computational and memory complexities with respect to sequence length, facilitating efficient processing of long contexts. On the other hand, sparse attention layers facilitate effective modeling of long-range dependencies. Rather than naively uniformly interleaving the two attention variants, we determine the placement of sparse attention modules using the layer selection mechanism proposed by Chen et al. (2026), which results in superior downstream performance.

**Training Strategy** Existing paradigms for training hybrid models generally fall into two categories: (1) training from scratch (Zuo et al., 2025; Qwen Team, 2025; Kimi Team et al., 2025; NVIDIA et al., 2025b) and (2) converting a pre-trained Transformer model into a hybrid model via cross-architecture distillation (Wang et al., 2024a; Hoshino et al., 2025; Li et al., 2025; Gu et al., 2025). Although training from scratch offers simplicity and maximum architectural flexibility, continual-training conversion is a more resource-efficient alternative that leverages parameter inheritance from established pre-trained models. By recycling pre-trained weights and representations, the continual-training method significantly reduces the immense computational cost typically associated with *de novo* training, achieving competitive performance with a fraction of the budget. Accordingly, MiniCPM-SALA leverages a conversion-based framework that uses continual training to adapt a Transformer into an efficient hybrid version while preserving its core capabilities.

**Table 1:** Overview of the whole training process to build MiniCPM-SALA.

Stage	Trainable Parameters	Sparse Attention	Sequence Length	# Tokens
Architecture Conversion (HALO)	Linear Attention	Disabled	0.5K	1.3B
Continual Stable-Training	All Parameters	Disabled	4K	314.6B
Short-Decay Training	All Parameters	Disabled	4K	1006.6B
Long-Decay Training	All Parameters	Enabled	32K 160K 520K	102.2B 62.9B 50.6B
Supervised Fine-Tuning	All Parameters	Enabled	64K 140K	204.5B 213.3B

**Sparse Attention and Linear Attention** For the sparse attention layers, we incorporate InfLLM-V2 (Zhao et al., 2025), which offers the distinct advantage of introducing no additional parameters to the architecture. Its inherent flexibility and ability to switch seamlessly between dense and sparse modes are highly compatible with our conversion process. This compatibility facilitates a stable training initialization by allowing sparse modules to inherit dense weights without architectural discrepancies, ensuring that the conversion to a hybrid structure does not compromise the model capacity. For the linear attention layers, we utilize Lightning Attention (Qin et al., 2024). Given our Transformer-to-hybrid conversion paradigm, Lightning Attention is selected for its functional proximity to the standard softmax attention. This structural alignment is intended to mitigate the complexities of parameter adaptation, thereby preserving pre-trained knowledge and ensuring robust downstream performance. Lightning Attention also provides better length generalization capabilities according to Chen et al. (2026), which may improve data efficiency during long-context continual-training.

**Other Architectural Improvements** Following HypeNet (Chen et al., 2026), we also introduce several architectural modifications to enhance the expressivity and training stability of MiniCPM-SALA. These include QK-Normalization (Henry et al., 2020), HyPE (Chen et al., 2026), and the integration of output gates.

- **QK-Normalization:** This is applied to all attention layers (both sparse and linear layers) to prevent the activation spikes that often occur in long-context training and further improve and boost the expressivity of linear attention modules.
- **HyPE (Hybrid Positional Encoding):** To balance rich positional awareness and long-range information retention, we employ a hybrid approach to positional encoding. We apply Rotary Positional Embedding (RoPE) (Su et al., 2023) to the linear attention layers to facilitate position-sensitive memory, allowing the model to preserve the relative order of tokens within the global context. On the other hand, we remove RoPE in the sparse attention layers. This strategic omission prevents the decay of long-distance information often associated with RoPE, thereby enabling more precise recall over extended contexts.
- **Output gates:** Furthermore, we incorporate an output gate after each attention block (both sparse and linear). This architectural choice aligns with recent advances in the gated attention mechanism (Qiu et al., 2025), in which the output gate has been shown to effectively mitigate issues such as attention sink. By regulating the information flow, the output gate prevents excessive focus on specific tokens and ensures a more flexible distribution of attention weights. Empirically, we observe that integrating output gates into both linear and sparse attention significantly improves model stability and performance.

## 2.2 Model Training

The training of MiniCPM-SALA is conducted through a multi-stage process that starts from an intermediate checkpoint of MiniCPM-4.0 (MiniCPM-Team et al., 2025), which has already been trained on 7T tokens. This methodology represents an extended implementation of Hybrid Attention via Layer Optimization (HALO) (Chen et al., 2026). In the initial phase, we use the HALO framework to convert softmax attention to linear attention. This conversion serves as the starting point for subsequent pipeline stages, including continual pre-training and post-training. By leveraging this approach, the model can transition from a dense architecture to a hybrid structure while preserving the general capabilities acquired during the backbone’s earlier training phases. The entire conversion process, consisting of five stages, is shown in Table 1. It is worth noting that the

Transformer-to-hybrid training of MiniCPM-SALA consumes approximately 2T tokens. This corresponds to roughly 25% of the data volume required to train MiniCPM-4.0 from scratch (8T tokens).

**Architecture Conversion (HALO)** The first stage uses HALO to convert the Transformer model from a full attention architecture to a hybrid architecture. During this phase, the training configuration of MiniCPM-SALA differs from the standard HALO approach in two aspects. First, regarding layer selection, we keep the first and last layers unconverted to improve training stability. For the remaining layers, we utilize the HALO selection algorithm to determine which layers are preserved as softmax attention layers. These preserved softmax attention layers are subsequently trained as sparse attention in later stages. The second difference from standard HALO is that we do not perform the final fine-tuning step of the original HALO process. Instead, we conduct more extensive continual pre-training and post-training, which comprise the subsequent stages of our methodology. The training process at this stage is highly efficient, using only 1.3B tokens with a sequence length of 512 tokens. Furthermore, only the converted linear-attention layers are trainable during this stage, while all other parameters remain frozen.

**Continual Stable-Training** The second stage is continual stable-training. We use the checkpoint from the previous stage as the starting point for further training on the MiniCPM-4.0 pre-training dataset. The primary objective of this phase is to facilitate better coordination between the converted linear attention layers and other model components, including full attention layers, FFN layers, and embeddings. The sequence length for this process is set to 4K tokens, with a total training volume of 314.6B tokens. Since the sequence length remains relatively short, the sparse attention is disabled at this stage to maintain computational efficiency. For the hyperparameter configuration, the learning rate (LR) is set to  $7.5 \times 10^{-3}$  and held constant after a 2,000-step LR warmup period. Accounting for the sequence length and the number of GPUs, the global batch size is set to 7.8M tokens.

**Short-Decay Training** The third stage is short-decay training, during which the LR undergoes exponential decay from  $7.5 \times 10^{-3}$  to  $3.75 \times 10^{-4}$ . This process utilizes a sequence length of 4K tokens and a global batch size of 7.8M tokens. This stage involves training on 1T tokens, representing the most extensive data volume in the entire development pipeline. Building on the MiniCPM-4.0 decay strategy, we significantly increase the weight of L2 high-quality selection data (Wang et al., 2026) and introduce a large volume of PDF corpora and L3 synthetic data. This approach aims to enhance general capabilities and logical reasoning using high-information-density training data, achieving the efficient compression and internalization of massive amounts of knowledge.

**Long-Decay Training** The fourth stage, long-decay, progressively extends the context length from 4K to 32K, 160K, and finally 520K tokens. These processes use data volumes of 102.2B tokens, 62.9B tokens, and 50.6B tokens, respectively. To accommodate the increased sequence lengths, the global batch size is adjusted to 7.8M, 9.8M, and 10.1M tokens, while the LR is systematically decays from  $3 \times 10^{-4}$  to  $2 \times 10^{-4}$  at 32K, then to  $1 \times 10^{-4}$  at 160K, and finally to  $3.75 \times 10^{-5}$  at 520K to conclude the process. At this stage, we up-sample the proportion of long-context data to better align the model with long-sequence distributions. Given the growing computational advantages of sparse attention at longer sequences, we enable the sparse attention mechanism at this stage and maintain full-parameter training, thereby allowing the model to effectively learn the synergy between sparse attention and linear attention.

**Supervised Fine-Tuning** The SFT corpus for this stage is composed of high-quality reasoning-intensive data, encompassing code, mathematics, knowledge, function calls, and general dialogue. This selection is designed to fully catalyze the reasoning and task-execution capabilities under complex logic. Furthermore, we specifically synthesize long-context data to enhance the precision of information retrieval and cross-document comprehension within extended sequences. During the SFT stage, the context length is set to 64K and increased to 140K afterwards, utilizing 204.5B and 213.3B tokens, respectively. Sparse attention remains enabled throughout this entire process. By bridging shorter and longer contexts, this strategy allows the model to better balance general capabilities with long-context proficiency. For both phases, the LR follows a schedule with a 1,000-step warmup to a peak of  $1 \times 10^{-3}$  before decaying to  $1 \times 10^{-4}$ , while the global batch sizes are set to 15.7M for the 64K phase and 17.8M for the 140K phase.

### 3 Experiments

#### 3.1 Model Performance

**Benchmarks** To thoroughly assess the general capabilities of the model, we conducted evaluations across a diverse array of benchmarks. These include knowledge-intensive tasks (CMLU (Li et al., 2023), MMLU-

**Table 2:** Standard evaluation results of MiniCPM-SALA and other open-source LLMs.

Models	Qwen3	Nemotron-Nano-v2	MiniCPM-4.1	Minstral-3-R	Falcon-H1R	MiniCPM-SALA
# Param.	8B	9B	8B	8B	7B	9B
Knowledge						
CMMLU	81.68	61.59	84.72	71.74	63.55	81.55
MMLU-Pro	73.26	71.79	72.70	68.75	70.98	67.04
Code						
HumanEval	93.90	93.90	91.46	96.95	96.34	95.12
LCB-v5	56.89	68.26	56.89	65.87	67.66	60.48
LCB-v6	48.57	60.00	51.43	53.71	57.71	52.00
MBPP	81.32	93.39	91.05	94.16	91.05	89.11
Math						
AIME24	73.33	71.67	80.83	81.46	86.67	83.75
AIME25	66.67	56.67	72.08	75.00	81.04	78.33
Other						
BBH	74.17	74.28	82.68	64.39	63.17	81.55
IFEval	84.66	86.69	77.45	70.06	86.32	76.34
Average	73.45	73.82	76.13	74.21	76.45	<b>76.53</b>

**Table 3:** Long-context evaluation results of MiniCPM-SALA and other open-source LLMs.

Models	Qwen3	Nemotron-Nano-v2	Minstral-3-R	Falcon-H1R	MiniCPM-SALA	
# Param.	8B	9B	8B	7B	9B	
RULER	64K	80.53	88.77	70.66	56.50	92.65
	128K	71.74	68.01	45.09	36.33	89.37
MRCR	64K-2N	29.20	20.91	44.02	13.18	29.77
	64K-4N	21.56	13.69	35.80	9.06	20.57
	64K-8N	17.82	13.24	17.23	6.93	16.56
	128K-2N	26.50	14.61	50.30	9.17	28.62
	128K-4N	14.75	12.20	22.66	8.22	19.62
	128K-8N	12.15	7.55	14.47	7.54	10.12
	32K	43.40	19.69	3.78	14.89	54.54
NoLiMa	64K	23.35	11.82	2.48	9.87	42.95
	128K	11.25	5.80	3.48	4.73	23.86
Average	32.02	25.12	28.18	16.04	<b>38.97</b>	

Pro (Wang et al., 2024b)), coding benchmarks (HumanEval (Chen et al., 2021), LCB-v5/v6 Jain et al. (2025), MBPP (Austin et al., 2021)), and mathematical reasoning sets (AIME24/25 (AIME, 2025)), alongside other representative benchmarks such as BBH (Suzgun et al., 2022) and IFEval (Zhou et al., 2023). We further evaluated long-context capabilities using RULER (Hsieh et al., 2024), MRCR<sup>1</sup>, and NoliMa (Modarressi et al., 2025). We utilized the OpenCompass framework (Contributors, 2023) to conduct the evaluations.

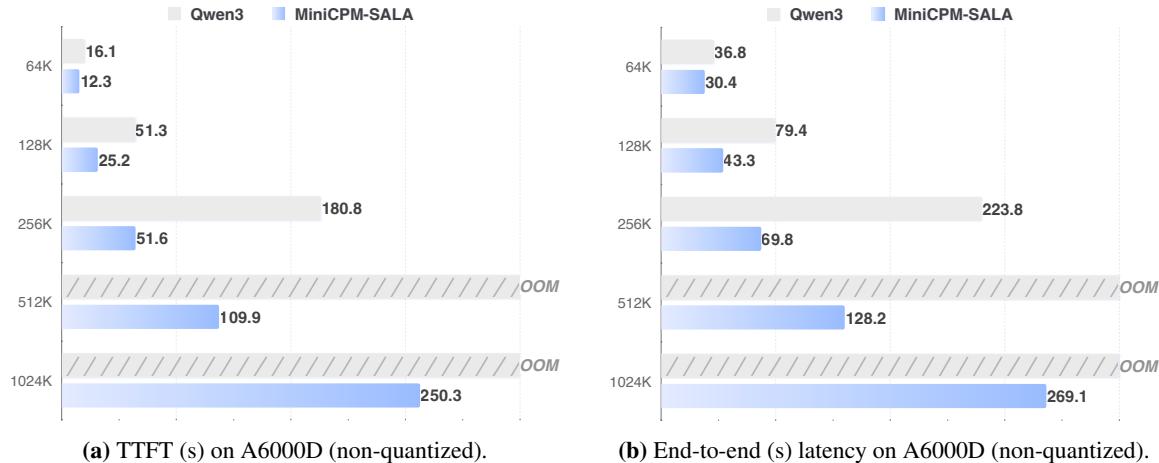
**Baseline Models** Given that MiniCPM-SALAI is a 9B-parameter model, we selected a series of modern baselines of comparable size, encompassing both hybrid and full-attention architectures. Specifically, the baselines include Qwen3-8B (Yang et al., 2025a), Nemotron-Nano-v2-9B (NVIDIA et al., 2025a), MiniCPM-4.1-8B (MiniCPM-Team et al., 2025), Minstral-3-Reasoning-8B (Liu et al., 2026), and Falcon-H1R-7B (Team et al., 2026). We exclude MiniCPM-4.1-8B from the evaluation of long contexts because of its limitation to a context length of 64K.

**Results of Standard Evaluation** Table 2 presents the performance of MiniCPM-SALA across a variety of standard benchmarks. The model achieves an average score of 76.53, which represents a competitive level

<sup>1</sup><https://huggingface.co/datasets/openai/mrcr>

**Table 4:** Ultra-long context evaluation results of MiniCPM-SALA and other open-source LLMs. \* denotes results cited from the official Qwen3-Next documentation.

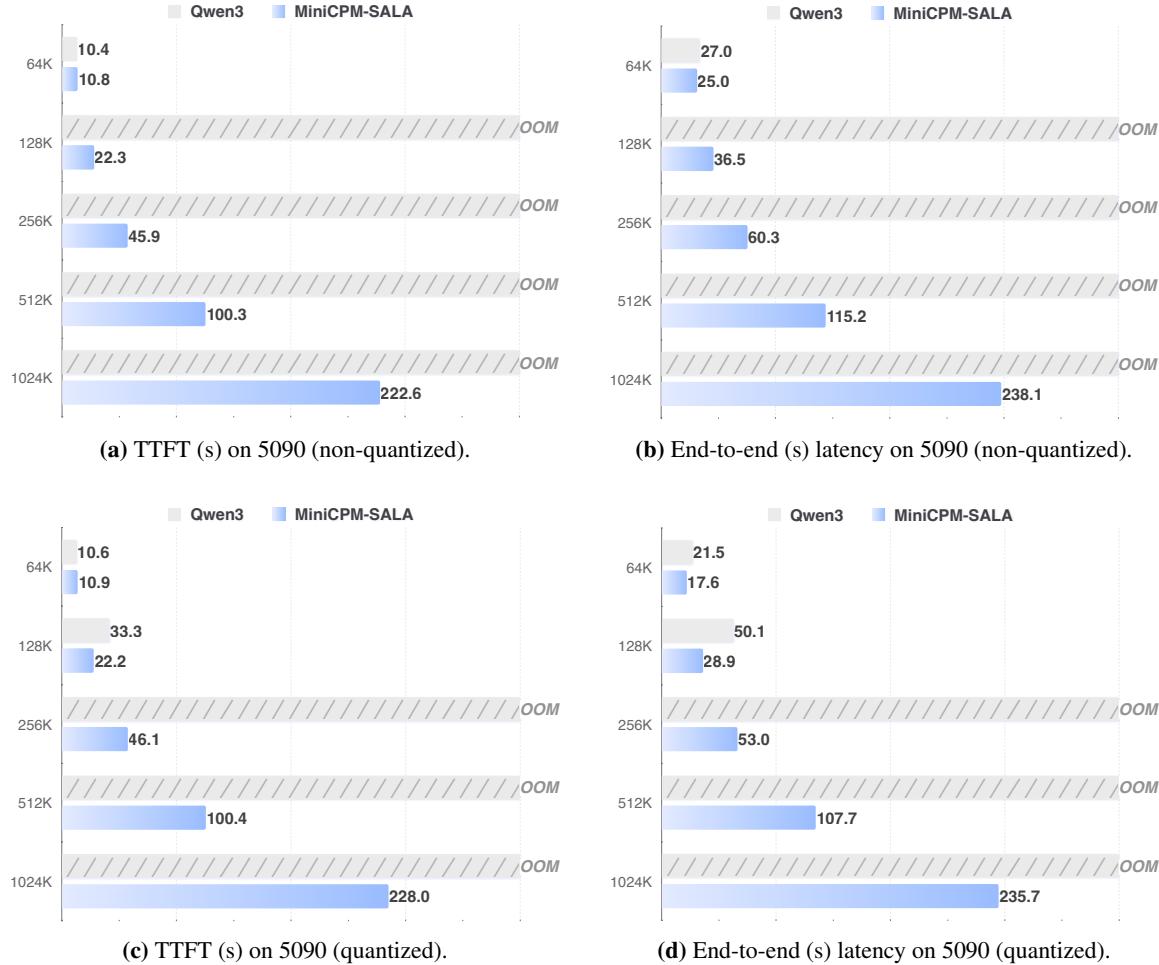
RULER	128K	512K	1000K	2048K
Qwen3-30B-A3B-Instruct-2507*	89.1	78.4	72.8	-
Qwen3-235B-A22B-Instruct-2507*	93.9	90.9	84.5	-
Qwen3-Next-80B-A3B-Instruct*	96.0	86.9	80.3	-
MiniCPM-SALA (9B)	89.4	87.1	86.3	81.6



**Figure 2:** Inference speed comparison between Qwen3-8B and MiniCPM-SALA. For each tested sequence length, the models process a specified input (prefilling) and generate 1K tokens (decoding). “TTFT” denotes Time To First Token, representing the prefilling latency, while “End-to-end” measures the total latency including both prefilling and decoding phases.

among open-source models of a similar scale. In coding tasks, the model demonstrates high proficiency with scores of 95.12 on HumanEval and 89.11 on MBPP. Mathematical reasoning capabilities also remain robust, as evidenced by the scores of 83.75 on AIME24 and 78.33 on AIME25. These results indicate that the integration of long-context mechanisms does not result in a significant degradation of general capabilities or short-context performance. The model maintains a performance profile that is comparable to, and in some cases exceeds, the performance of models such as Qwen3-8B and Falcon-H1R-7B in standard evaluation settings.

**Results of Long-Context Evaluation** The evaluation of long-context capabilities is summarized in Table 3, covering benchmarks such as RULER, MRCR, and NoLiMa. MiniCPM-SALA shows a notable proficiency in managing extended input sequences. On the RULER benchmark at a 128K context length, the model maintains



**Figure 3:** Inference speed comparison between Qwen3-8B and MiniCPM-SALA. For each tested sequence length, the models process a specified input (prefilling) and generate 1K tokens (decoding).

a score of 89.37, while many other baselines exhibit a more pronounced decrease in accuracy at the same scale. The advantage of the model is particularly visible in the NoLiMa benchmark, where it achieves a score of 23.86 at the 128K level. This performance is substantially higher than the scores recorded for other models in the comparison. With an overall average long-context score of 38.97, the model demonstrates improved stability and effective information retrieval across large context windows.

**Results of Ultra-Long Context** As demonstrated in Table 4, MiniCPM-SALA exhibits surprising length extrapolation capabilities. The results for the Qwen3 models are sourced from the official Qwen3-Next documentation<sup>2</sup>. Despite being restricted to a 520K training length, the model successfully extrapolates to 2048K tokens without a significant degradation in performance, maintaining a score of 81.6. It is worth noting that this extrapolation requires no auxiliary techniques (e.g., YaRN (Peng et al., 2024)). This result highlights the efficacy of our approach in handling context windows far beyond the training stage. Additionally, MiniCPM-SALA shows remarkable parameter efficiency, surpassing the performance of the Qwen3-Next-80B-A3B-Instruct model at the 1000K context length (86.3 vs. 80.3), proving that effective long-context processing does not necessarily require massive parameter counts. The length extrapolation capabilities of MiniCPM-SALA can be attributed to the NoPE configuration within the sparse attention layers. In this design, the stored KV-Cache does not require combination with positional information, which can otherwise hinder the capture of long-range dependencies.

<sup>2</sup><https://huggingface.co/Qwen/Qwen3-Next-80B-A3B-Instruct>

### 3.2 Inference Speed

We assessed the inference speed of MiniCPM-SALA and Qwen3-8B across different hardware and sequence lengths. To verify the long-text processing capabilities of the model in edge computing scenarios, we conducted experiments not only on cloud-grade inference chips, such as the NVIDIA A6000D, but also on consumer-grade edge GPUs, such as the NVIDIA 5090. For each sequence length, we measured both the Time To First Token (TTFT) and the end-to-end latency. The former serves as an indicator of the prefilling speed, while the latter reflects the combined performance of the prefilling and decoding phases. To align the evaluation with practical deployment scenarios, we assessed the inference latency for both non-quantized models and models compressed via GPTQ (Frantar et al., 2023) INT4 quantization.

Figure 2 presents a comprehensive comparison of inference latency between Qwen3-8B and MiniCPM-SALA on an NVIDIA A6000D GPU (96GB VRAM). We evaluated performance across sequence lengths ranging from 64K to 1024K tokens. As illustrated, MiniCPM-SALA demonstrates a significant performance advantage over the baseline across all tested configurations. In non-quantized settings, MiniCPM-SALA consistently achieves lower latency. Notably, at a sequence length of 256K, MiniCPM-SALA reduces the TTFT from 180.8s (Qwen3) to just 51.6s, representing a speedup of approximately  $3.5\times$ .

Crucially, the results highlight a distinct advantage in memory efficiency. While Qwen3-8B encounters OOM failures at sequence lengths of 512K and 1024K, MiniCPM-SALA successfully processes these extended contexts. For example, at 1024K tokens, MiniCPM-SALA maintains a TTFT of 250.3s (non-quantized) and 256.9s (quantized), whereas the baseline fails to complete the inference. This trend persists in the end-to-end latency metrics, proving that MiniCPM-SALA is robust enough for ultra-long context generation tasks where full-attention models fail.

Figure 3 demonstrates the critical advantage of MiniCPM-SALA on memory-constrained hardware. On the RTX 5090 (32GB VRAM), the baseline Qwen3-8B hits a “memory wall” significantly earlier than on the A6000D, triggering OOM errors at just 128K tokens in non-quantized settings and 256K in quantized settings. In stark contrast, MiniCPM-SALA successfully scales to 1024K context lengths without memory failure. This suggests that MiniCPM-SALA effectively democratizes long-context inference, enabling 1M-token processing on consumer-level GPUs where full-attention architectures are unusable.

## 4 Conclusion

In this paper, we presented MiniCPM-SALA, a hybrid architecture that combines sparse and linear attention to overcome the computational and memory bottlenecks of ultra-long context modeling. By utilizing a cost-effective Transformer-to-hybrid training paradigm, we successfully retained the general capabilities of full-attention models while reducing training costs by approximately 75%. Experimental results confirm that MiniCPM-SALA achieves a  $2.5\times$  inference speedup and enables 1M-token context processing on single GPUs (e.g., NVIDIA A6000D), surpassing the limitations of standard 8B models. These findings establish MiniCPM-SALA as a scalable and accessible solution for next-generation, information-intensive applications.

## 5 Contributions and Acknowledgments

MiniCPM-SALA is the result of the collective efforts of all members of our team. Please refer to Chen et al. (2026) and Zhao et al. (2025) for model architecture details.

**Contributors** (Ordered by the last name) Wenhao An, Yingfa Chen, Yewei Fang, Jiayi Li, Xin Li, Yaohui Li, Yishan Li, Yuxuan Li, Biyuan Lin, Chuan Liu\*, Hezi Liu, Siyuan Liu, Hongya Lyu, Yinxu Pan, Shixin Ren, Xingyu Shen, Zhou Su, Haojun Sun, Yangang Sun, Zhen Leng Thai, Xin Tian, Rui Wang\*, Xiaorong Wang, Yudong Wang, Bo Wu, Xiaoyue Xu, Dong Xu, Shuaikang Xue, Jiawei Yang, Bowen Zhang, Jinqian Zhang, Letian Zhang, Shengnan Zhang, Xinyu Zhang, Xinyuan Zhang\*, Zhu Zhang, Hengyu Zhao, Jiacheng Zhao\*, Jie Zhou, Zihan Zhou

**Project Design and Coordination** Shuo Wang, Chaojun Xiao, Xu Han, Zhiyuan Liu, Maosong Sun

**Affiliations** Contributors marked with \* are affiliated with XCORE SIGMA, while the remaining contributors are affiliated with OpenBMB.

## References

- AIME. AIME problems and solutions, 2025. URL [https://artofproblemsolving.com/wiki/index.php/AIME\\_Problems\\_and\\_Solutions](https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions).
- Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL <https://aclanthology.org/2023.emnlp-main.298/>.
- Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.
- Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. LongAlign: A recipe for long context alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. URL <https://aclanthology.org/2024.findings-emnlp.74/>.
- Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. Longwriter: Unleashing 10,000+ word generation from long context LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=kQ5s9Yh0WI>.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf).
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harry Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- Yingfa Chen, Zhen Leng Thai, Zihan Zhou, Zhu Zhang, Xingyu Shen, Shuo Wang, Chaojun Xiao, Xu Han, and Zhiyuan Liu. Hybrid linear attention done right: Efficient distillation and effective architectures for extremely long contexts, 2026. URL <https://arxiv.org/abs/2601.22156>.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL <https://arxiv.org/abs/2507.06261>.
- OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. <https://github.com/open-compass/opencompass>, 2023.
- DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3.2: Pushing the frontier of open large language models, 2025. URL <https://arxiv.org/abs/2512.02556>.
- Elias Frantar, Saleh Ashkboos, Torsten Hoefer, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers, 2023. URL <https://arxiv.org/abs/2210.17323>.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=tEYskw1VY2>.
- Yuxian Gu, Qinghao Hu, Shang Yang, Haocheng Xi, Junyu Chen, Song Han, and Han Cai. Jet-nemotron: Efficient language model with post neural architecture search, 2025. URL <https://arxiv.org/abs/2508.15884>.

- Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y. Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. Deepseek-coder: When the large language model meets programming – the rise of code intelligence, 2024. URL <https://arxiv.org/abs/2401.14196>.
- Mutian He and Philip N. Garner. Alleviating forgetfulness of linear attention by hybrid sparse attention and contextualized learnable token eviction, 2025. URL <https://arxiv.org/abs/2510.20787>.
- Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4246–4253, 2020.
- Yuichiro Hoshino, Hideyuki Tachibana, Muneyoshi Inahara, and Hiroto Takegawa. Rad: Redundancy-aware distillation for hybrid models via self-speculative decoding, 2025. URL <https://arxiv.org/abs/2505.22135>.
- Haowen Hou, Zhiyi Huang, Kaifeng Tan, Rongchang Lu, and Fei Richard Yu. Rwkv-x: A linear complexity hybrid language model, 2025. URL <https://arxiv.org/abs/2504.21463>.
- Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. RULER: What’s the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=kIoBbc76Sy>.
- Xiang Hu, Jiaqi Leng, Jun Zhao, Kewei Tu, and Wei Wu. Hardware-aligned hierarchical sparse attention for efficient long-term memory access, 2025. URL <https://arxiv.org/abs/2504.16795>.
- 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. URL <https://openreview.net/forum?id=chfJJYC3iL>.
- 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. URL <https://openreview.net/forum?id=VTF8yNQM66>.
- Kimi Team, Yu Zhang, Zongyu Lin, Xingcheng Yao, Jiaxi Hu, Fanqing Meng, Chengyin Liu, Xin Men, Songlin Yang, Zhiyuan Li, Wentao Li, et al. Kimi linear: An expressive, efficient attention architecture, 2025. URL <https://arxiv.org/abs/2510.26692>.
- 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, 2023.
- Keyu Li, Junhao Shi, Yang Xiao, Mohan Jiang, Jie Sun, Yunze Wu, Shijie Xia, Xiaojie Cai, Tianze Xu, Weiye Si, Wenjie Li, Dequan Wang, and Pengfei Liu. Agencybench: Benchmarking the frontiers of autonomous agents in 1m-token real-world contexts, 2026. URL <https://arxiv.org/abs/2601.11044>.
- Yanhong Li, Songlin Yang, Shawn Tan, Mayank Mishra, Rameswar Panda, Jiawei Zhou, and Yoon Kim. Distilling to hybrid attention models via kl-guided layer selection, 2025. URL <https://arxiv.org/abs/2512.20569>.
- Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, et al. Minstral 3, 2026. URL <https://arxiv.org/abs/2601.08584>.
- Tianyang Liu, Canwen Xu, and Julian McAuley. Repobench: Benchmarking repository-level code auto-completion systems. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=pPjZIOuQuF>.
- Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants, 2023. URL <https://arxiv.org/abs/2311.12983>.
- MiniCPM-Team, Chaojun Xiao, Yuxuan Li, Xu Han, Yuzhuo Bai, Jie Cai, Haotian Chen, Wentong Chen, Xin Cong, Ganqu Cui, Ning Ding, et al. Minicpm4: Ultra-efficient llms on end devices, 2025. URL <https://arxiv.org/abs/2506.07900>.

- Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schuetze. Nolima: Long-context evaluation beyond literal matching. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=0OshX1hiSa>.
- NVIDIA, Aarti Basant, Abhijit Khairnar, Abhijit Paithankar, Abhinav Khattar, Adithya Renduchintala, Aditya Malte, Akhiad Bercovich, Akshay Hazare, Alejandra Rico, et al. Nvidia nemotron nano 2: An accurate and efficient hybrid mamba-transformer reasoning model, 2025a. URL <https://arxiv.org/abs/2508.14444>.
- NVIDIA, Aaron Blakeman, Aaron Grattafiori, Aarti Basant, Abhibha Gupta, Abhinav Khattar, Adi Renduchintala, Aditya Vavre, Akanksha Shukla, Akhiad Bercovich, Aleksander Ficek, et al. Nemotron 3 nano: Open, efficient mixture-of-experts hybrid mamba-transformer model for agentic reasoning, 2025b. URL <https://arxiv.org/abs/2512.20848>.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report, 2024. URL <https://arxiv.org/abs/2303.08774>.
- Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Gv, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartomiej Koptyra, Hayden Lau, Jiaju Lin, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Johan Wind, Stanisław Woźniak, Zhenyuan Zhang, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. RWKV: Reinventing RNNs for the transformer era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023. URL <https://aclanthology.org/2023.findings-emnlp.936/>.
- 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*, 2024. URL <https://openreview.net/forum?id=wHBfxhZu1u>.
- Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024. URL <https://aclanthology.org/2024.acl-long.810/>.
- Zhen Qin, Weigao Sun, Dong Li, Xuyang Shen, Weixuan Sun, and Yiran Zhong. Various lengths, constant speed: Efficient language modeling with lightning attention. In *Forty-first International Conference on Machine Learning*, 2024. URL <https://openreview.net/forum?id=Lwm6TiUP4X>.
- Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=1b7wh04SfY>.
- Qwen Team. Qwen3-Next: Towards Ultimate Training & Inference Efficiency, 2025. URL <https://qwen.ai/blog?id=4074cca80393150c248e508aa62983f9cb7d27cd>.
- Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. Assisting in writing Wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024. URL <https://aclanthology.org/2024.nacl-long.347>.
- Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL <https://arxiv.org/abs/2104.09864>.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.
- Falcon LLM Team, Iheb Chaabane, Puneesh Khanna, Suhail Mohmad, Slim Frikha, Shi Hu, Abdalgader Abubaker, Reda Alami, Mikhail Lubinets, Mohamed El Amine Seddik, and Hakim Hacid. Falcon-h1r: Pushing the reasoning frontiers with a hybrid model for efficient test-time scaling, 2026. URL <https://arxiv.org/abs/2601.02346>.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/3f5ee243547dee91fdb053c1c4a845aa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fdb053c1c4a845aa-Paper.pdf).

Junxiong Wang, Daniele Paliotta, Avner May, Alexander M. Rush, and Tri Dao. The mamba in the llama: Distilling and accelerating hybrid models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 62432–62457. Curran Associates, Inc., 2024a. doi: 10.52202/079017-1996. URL [https://proceedings.neurips.cc/paper\\_files/paper/2024/file/723933067ad315269b620bc0d2c05cba-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/723933067ad315269b620bc0d2c05cba-Paper-Conference.pdf).

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. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 95266–95290. Curran Associates, Inc., 2024b. doi: 10.52202/079017-3018. URL [https://proceedings.neurips.cc/paper\\_files/paper/2024/file/ad236edc564f3e3156e1b2feafb99a24-Paper-Datasets\\_and\\_Benchmarks\\_Track.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/ad236edc564f3e3156e1b2feafb99a24-Paper-Datasets_and_Benchmarks_Track.pdf).

Yudong Wang, Zixuan Fu, Hengyu Zhao, Chen Zhao, Chuyue Zhou, Xinle Lin, Hongya Lyu, Shuaikang Xue, Yi Yi, Yingjiao Wang, Zhi Zheng, Yuzhou Zhang, Jie Zhou, Chaojun Xiao, Xu Han, Zhiyuan Liu, and Maosong Sun. Data science and technology towards agi part i: Tiered data management, 2026. URL <https://arxiv.org/abs/2602.09003>.

Chaojun Xiao, Pingle Zhang, Xu Han, Guangxuan Xiao, Yankai Lin, Zhengyan Zhang, Zhiyuan Liu, and Maosong Sun. Inflm: Training-free long-context extrapolation for llms with an efficient context memory. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 119638–119661. Curran Associates, Inc., 2024. doi: 10.52202/079017-3801. URL [https://proceedings.neurips.cc/paper\\_files/paper/2024/file/d842425e4bf79ba039352da0f658a906-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/d842425e4bf79ba039352da0f658a906-Paper-Conference.pdf).

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report, 2025a. URL <https://arxiv.org/abs/2505.09388>.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In *Forty-first International Conference on Machine Learning*, 2024a. URL <https://openreview.net/forum?id=ia5XvxFUJT>.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL <https://openreview.net/forum?id=y8Rm4VNRPH>.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL <https://openreview.net/forum?id=r8H7xhYPwz>.

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, Yuxing Wei, Lean Wang, Zhiping Xiao, Yuqing Wang, Chong Ruan, Ming Zhang, Wenfeng Liang, and Wangding Zeng. Native sparse attention: Hardware-aligned and natively trainable sparse attention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025. URL <https://aclanthology.org/2025.acl-long.1126/>.

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. Inflm-v2: Dense-sparse switchable attention for seamless short-to-long adaptation, 2025. URL <https://arxiv.org/abs/2509.24663>.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL <https://arxiv.org/abs/2311.07911>.

Zihan Zhou, Chong Li, Xinyi Chen, Shuo Wang, Yu Chao, Zhili Li, Haoyu Wang, Qi Shi, Zhixing Tan, Xu Han, Xiaodong Shi, Zhiyuan Liu, and Maosong Sun. LLM $\times$ MapReduce: Simplified long-sequence processing using large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025. URL <https://aclanthology.org/2025.acl-long.1341/>.

Jingwei Zuo, Maksim Velikanov, Ilyas Chahed, Younes Belkada, Dhia Eddine Rhayem, Guillaume Kunsch, Hakim Hacid, Hamza Yous, Brahim Farhat, Ibrahim Khadraoui, Mugariya Farooq, Giulia Campesan, Ruxandra Cojocaru, Yasser Djilali, Shi Hu, Iheb Chaabane, Puneesh Khanna, Mohamed El Amine Seddik, Ngoc Dung Huynh, Phuc Le Khac, Leen AlQadi, Billel Mokeddem, Mohamed Chami, Abdalgader Abubaker, Mikhail Lubinets, Kacper Piskorski, and Slim Frikha. Falcon-h1: A family of hybrid-head language models redefining efficiency and performance, 2025. URL <https://arxiv.org/abs/2507.22448>.