Hymba: A Hybrid-Head Architecture for Small Language Models

NVIDIA Research

Xin Dong*, Yonggan Fu*, Shizhe Diao, Wonmin Byeon, Zijia Chen, Ameya Sunil Mahabaleshwarkar, Shih-Yang Liu, Matthijs Van Keirsbilck, Min-Hung Chen, Yoshi Suhara, Yingyan Celine Lin, Jan Kautz, Pavlo Molchanov

* equal contribution

Linear Attention - A Potential Transformer Alternative

$$Y = \underbrace{\mathrm{softmax}(QK^T)}_{\text{quadratic cost}} \cdot V$$

Self-Attention

↓

$$Y = (QK^T) \cdot V = Q \cdot (K^T V)$$

Linear Attention

↓

$$\begin{aligned} Y &= (L \circ QK^T) \cdot V \\ &= Q \cdot \underbrace{\mathrm{cumsum}(K^T V)}_{\text{recurrence!}} \end{aligned}$$

Causal Linear Attention

$$L = \begin{bmatrix} 1 & & \\ \vdots & \ddots & \\ 1 & \dots & 1 \end{bmatrix}$$

↓

$$Y = (L \circ QK^T) \cdot V$$

$$L_{ij} = \begin{cases} a_i \times \cdots \times a_{i+1} & \text{if } i \geq j, \\ 0 & \text{if } i < j \end{cases}$$

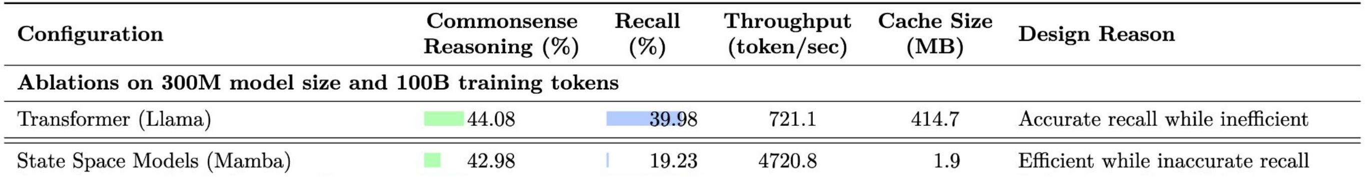
Generalized Causal Linear Attention (SSM, Mamba)

$$L = \begin{bmatrix} 1 & & & & \\ a_1 & 1 & & & \\ a_2 a_1 & a_2 & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ a_{\mathrm{T}-1} \dots a_1 & a_{\mathrm{T}-1} \dots a_2 & \dots & a_{\mathrm{T}-1} & 1 \end{bmatrix}$$

Hymba: A Hybrid-Head Architecture

Challenges:

- **Transformers** quadratic computational cost and substantial memory for storing KV caches.
- **State space models (Mamba)** are efficient while often fall short in general benchmarks and recall-intensive tasks.

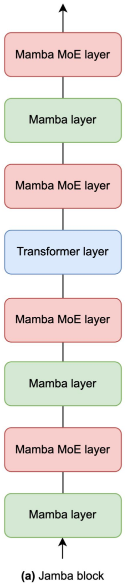


Configuration	Commonsense Reasoning (%)	Recall (%)	Throughput (token/sec)	Cache Size (MB)	Design Reason
Ablations on 300M model size and 100B training tokens					
Transformer (Llama)	44.08	39.98	721.1	414.7	Accurate recall while inefficient
State Space Models (Mamba)	42.98	19.23	4720.8	1.9	Efficient while inaccurate recall

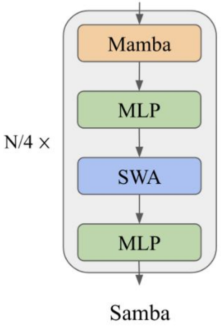
Hymba: A Hybrid-Head Architecture

Challenges:

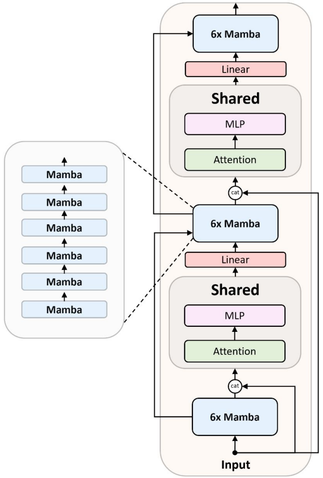
- **Transformers** quadratic computational cost and substantial memory for storing KV caches.
- **State space models (Mamba)** are efficient while often fall short in general benchmarks and recall-intensive tasks.
- **Recent hybrid models** combine attention and SSM layers by sequentially interleaving these layers (Samba, Jamba). However, these existing hybrid models can lead to potential information bottlenecks when a layer type poorly suited for a specific task cannot effectively process the information.



Jamba



Samba



Zamba

Hymba: A Hybrid-Head Architecture

Challenges:

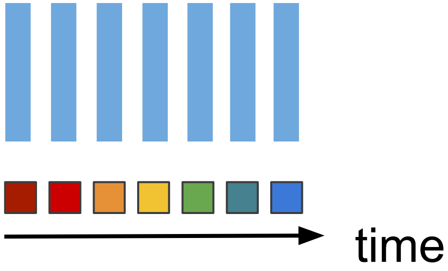
- Transformers quadratic computational cost and substantial memory for storing KV caches.
- State space models (Mamba) are efficient while often fall short in general benchmarks and recall-intensive tasks.
- Recent hybrid models combine attention and SSM layers by sequentially interleaving these layers (Samba, Jamba). However, these existing hybrid models can lead to information bottlenecks when a layer type poorly suited for a specific task cannot effectively process the information.

Methods:

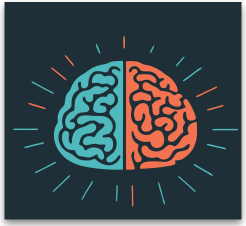
- We propose **Hymba**, a novel hybrid-head architecture that integrates attention heads and SSM heads within the same layer, offering parallel and complementary processing of the same inputs.

Memory Systems

Transformer (KV)



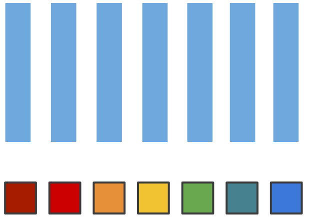
Snapshot memory



Biology Memory System

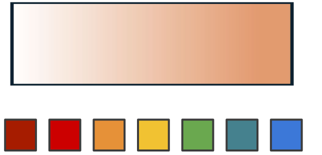
Memory Systems

Transformer (KV)



time

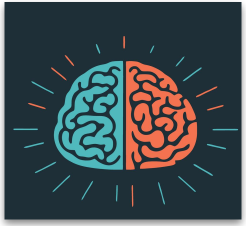
Mamba (Hidden State)



time

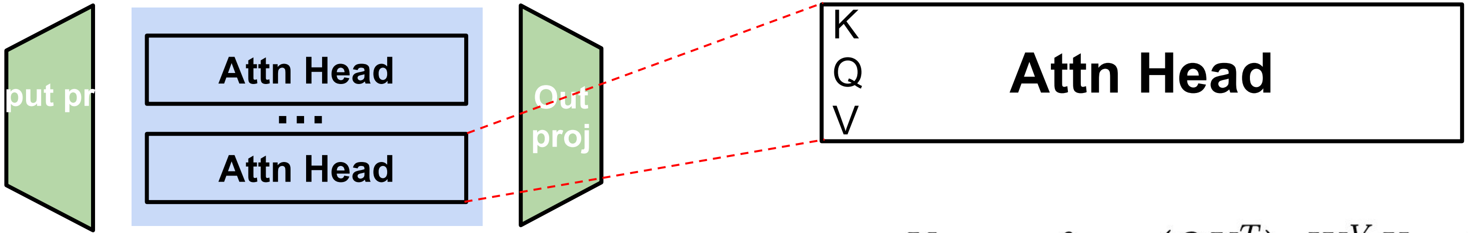
Snapshot memory

Fading memory



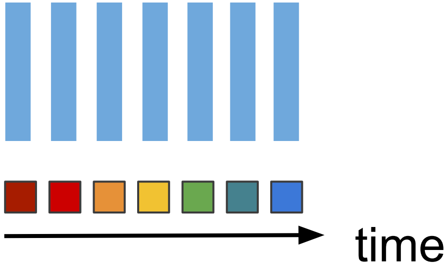
Biology Memory System

Language Model Architectures - Attention



Transformer

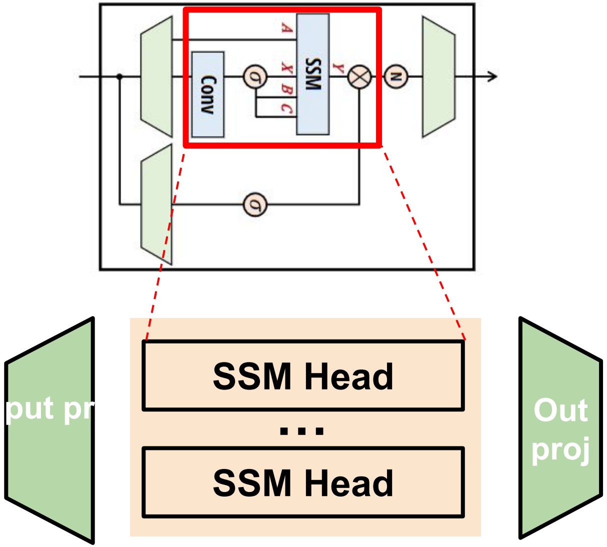
$$Y_{\text{attn}} = \text{softmax}(QK^T) \cdot W^V X = M_{\text{attn}} X$$

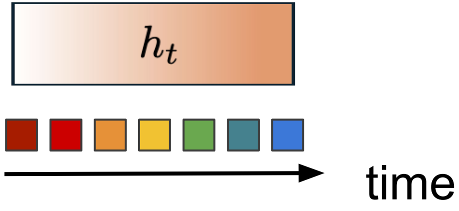


Snapshot memory

$$M_{\text{attn}} := \begin{bmatrix} \frac{\exp(q_1 \cdot k_1^T) \cdot W^V}{\sum_{i=1}^{1} \exp(q_i \cdot k_i^T)} & 0 & \cdots & 0 \\ \frac{\exp(q_1 \cdot k_1^T) \cdot W^V}{\sum_{i=1}^{2} \exp(q_i \cdot k_i^T)} & \frac{\exp(q_2 \cdot k_2^T) \cdot W^V}{\sum_{i=1}^{2} \exp(q_i \cdot k_i^T)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\exp(q_1 \cdot k_1^T) \cdot W^V}{\sum_{i=1}^{T} \exp(q_i \cdot k_i^T)} & \frac{\exp(q_2 \cdot k_2^T) \cdot W^V}{\sum_{i=1}^{T} \exp(q_i \cdot k_i^T)} & \cdots & \frac{\exp(q_T \cdot k_T^T) \cdot W^V}{\sum_{i=1}^{T} \exp(q_i \cdot k_i^T)} \end{bmatrix}$$

Language Model Architectures - State Space Model (SSM)





Fading memory

The recurrent form of an SSM module

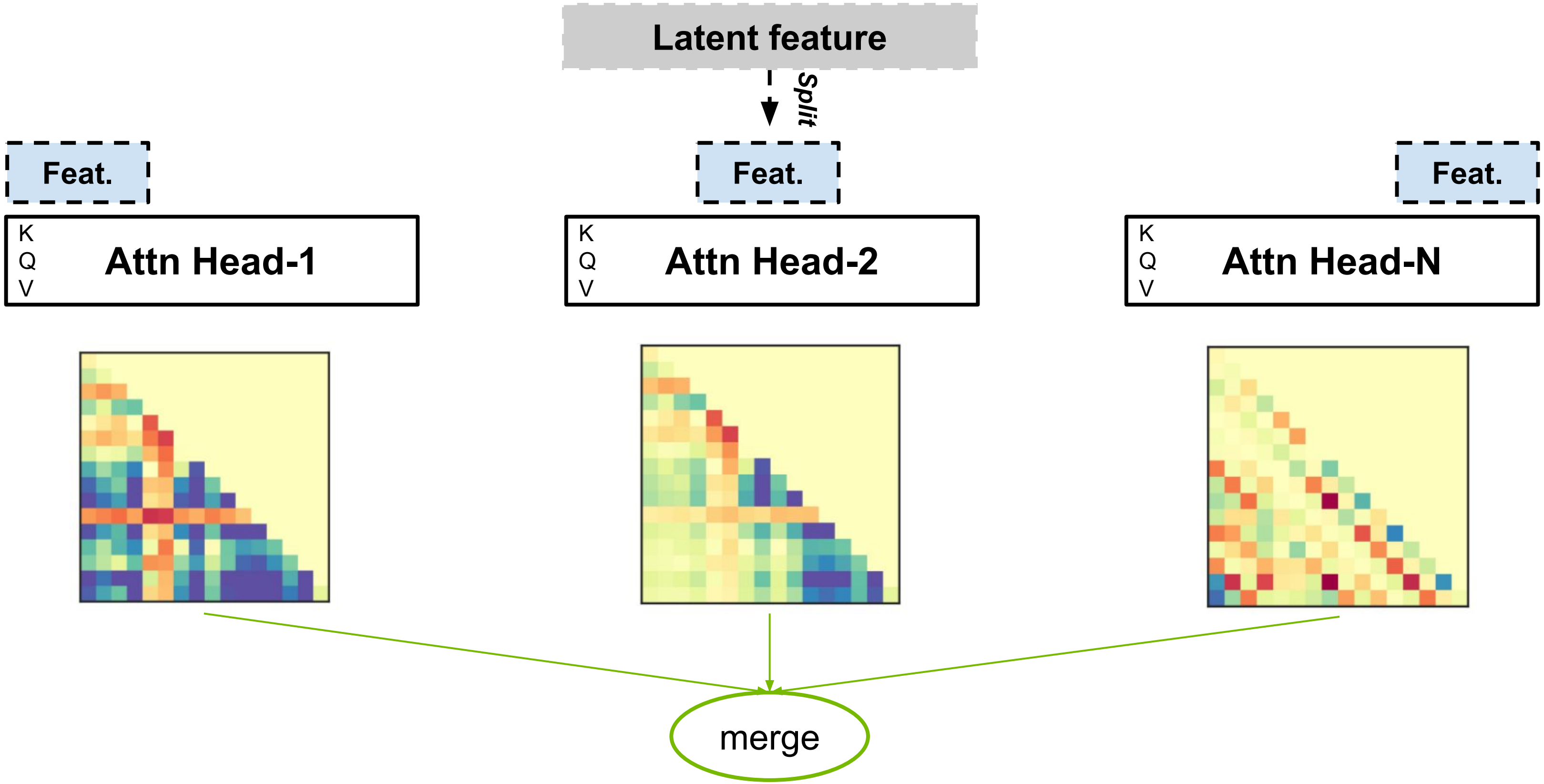
$$h_t = \bar{A}_t h_{t-1} + \bar{B}_t x_t$$
$$y_t = C_t^\top h_t$$

The equivalent parallel form of an SSM module

$$Y_{\text{ssm}} = M_{\text{ssm}}(A, B, C, \Delta_t) X$$
$$\text{where} \quad M_{\text{ssm}}^{i,j} = C_i \left(\prod_{k=j+1}^{i} \bar{A}_k \right) \bar{B}_j$$

$$M_{\text{ssm}}(A, B, C, \Delta_t) := \begin{bmatrix} C_1 \bar{B}_1 & 0 & \cdots & 0 \\ C_2 \Delta_2 \bar{B}_1 & C_2 \bar{B}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ C_L \prod_{k=2}^{L} \bar{A}_k \bar{B}_1 & C_L \prod_{k=3}^{L} \bar{A}_k \bar{B}_2 & \cdots & C_L \bar{B}_L \end{bmatrix}$$

Multi-Head Attention



Latent feature

Split

Feat.

K
Q
V
Attn Head-1

Feat.

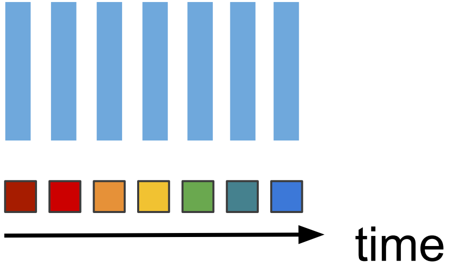
K
Q
V
Attn Head-2

Feat.

K
Q
V
Attn Head-N

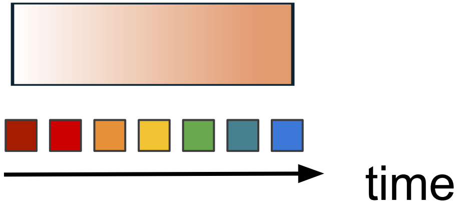
merge

Multi-Head Multi-Type Attention



Snapshot memory

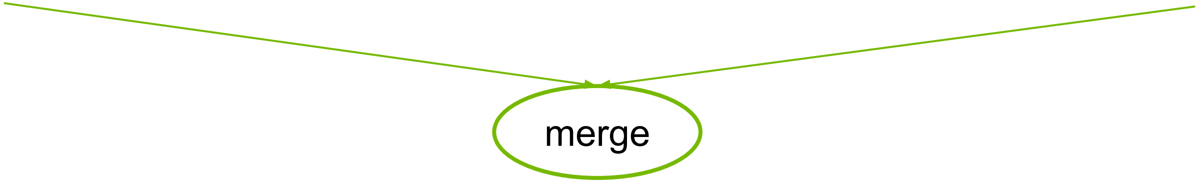
$$M_{\text{attn}} := \begin{bmatrix} \frac{\exp(q_1 \cdot k_1^T)\cdot W^V}{\sum_{i=1}^{1} \exp(q_i \cdot k_i^T)} & 0 & \cdots & 0 \\ \frac{\exp(q_1 \cdot k_1^T)\cdot W^V}{\sum_{i=1}^{2} \exp(q_i \cdot k_i^T)} & \frac{\exp(q_2 \cdot k_2^T)\cdot W^V}{\sum_{i=1}^{2} \exp(q_i \cdot k_i^T)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\exp(q_1 \cdot k_1^T)\cdot W^V}{\sum_{i=1}^{T} \exp(q_i \cdot k_i^T)} & \frac{\exp(q_2 \cdot k_2^T)\cdot W^V}{\sum_{i=1}^{T} \exp(q_i \cdot k_i^T)} & \cdots & \frac{\exp(q_T \cdot k_T^T)\cdot W^V}{\sum_{i=1}^{T} \exp(q_i \cdot k_i^T)} \end{bmatrix}$$



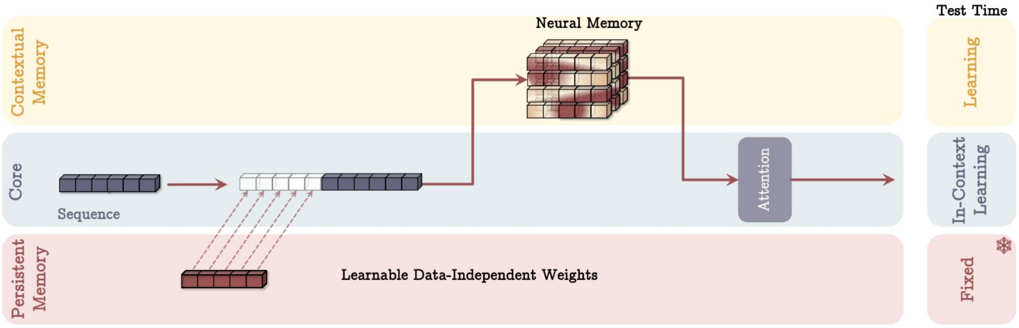
Fading memory

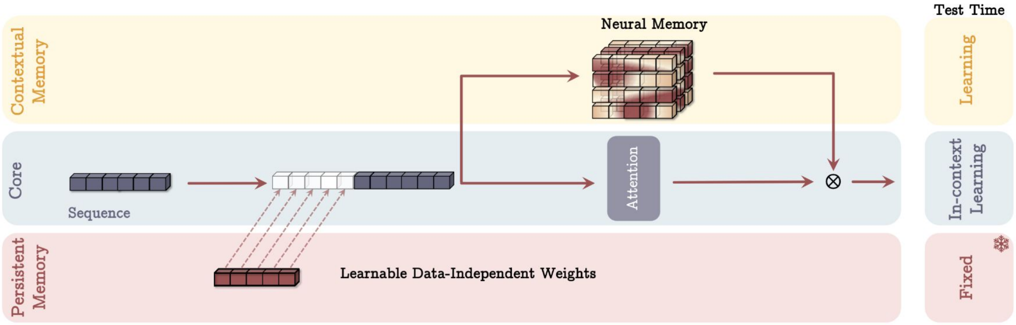
$$M_{\text{ssm}}(A, B, C, \Delta_t) := \begin{bmatrix} C_1 \bar{B}_1 & 0 & \cdots & 0 \\ C_2 \Delta_2 \bar{B}_1 & C_2 \bar{B}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ C_L \prod_{k=2}^{L} \bar{A}_k \bar{B}_1 & C_L \prod_{k=3}^{L} \bar{A}_k \bar{B}_2 & \cdots & C_L \bar{B}_L \end{bmatrix}$$

$$\Delta_t = \text{Softplus}(W_\Delta X_t), \quad B_t = W_B X_t, \quad C_t = (W_C X_t)^T$$
$$\bar{A}_t = \exp(A\Delta_t), \quad \bar{B}_t = B_t \Delta_t$$



Memory As A Layer or Multiple Memories as A Layer





Behrouz, A., Zhong, P. and Mirrokni, V., 2024. Titans: Learning to Memorize at Test Time. arXiv preprint arXiv:2501.00663.

Hymba: A Hybrid-Head Architecture

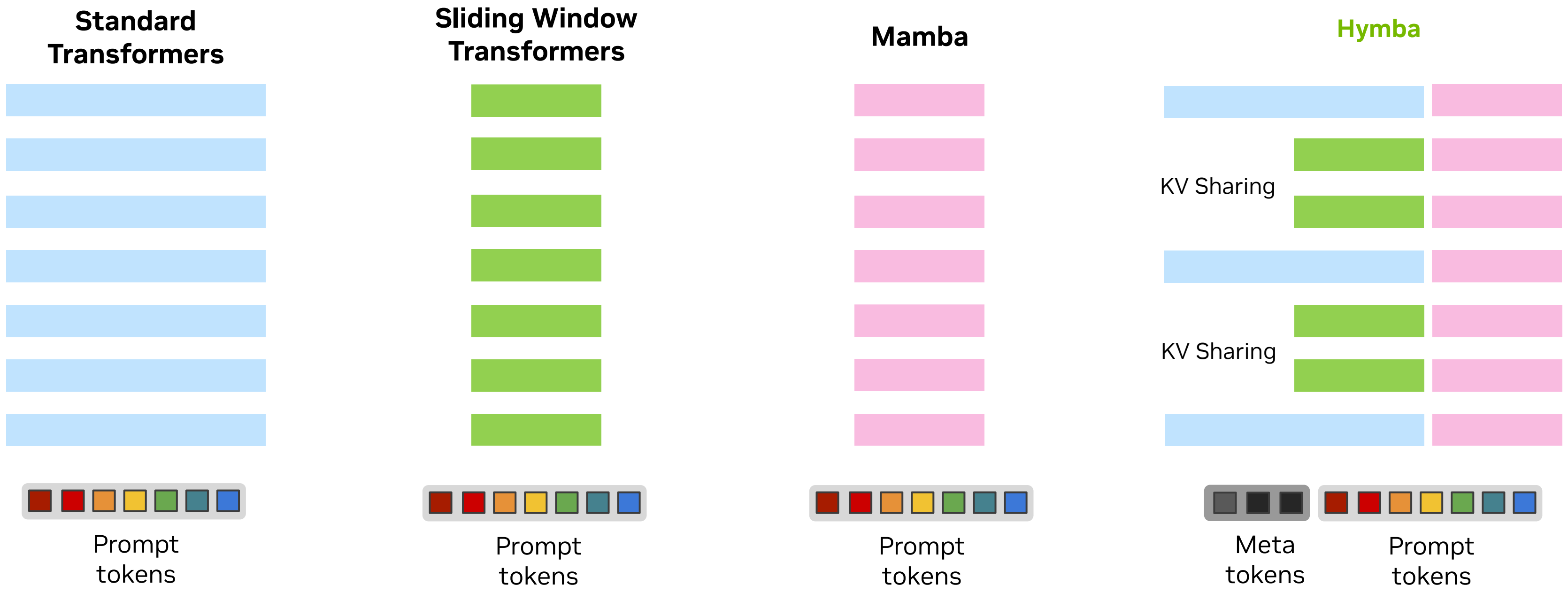
Challenges:

- Transformers quadratic computational cost and substantial memory for storing KV caches.
- State space models (Mamba) are efficient while often fall short in general benchmarks and recall-intensive tasks.
- Recent hybrid models combine attention and SSM layers by sequentially interleaving these layers (Samba, Jamba). However, these existing hybrid models can lead to information bottlenecks when a layer type poorly suited for a specific task cannot effectively process the information.

Methods:

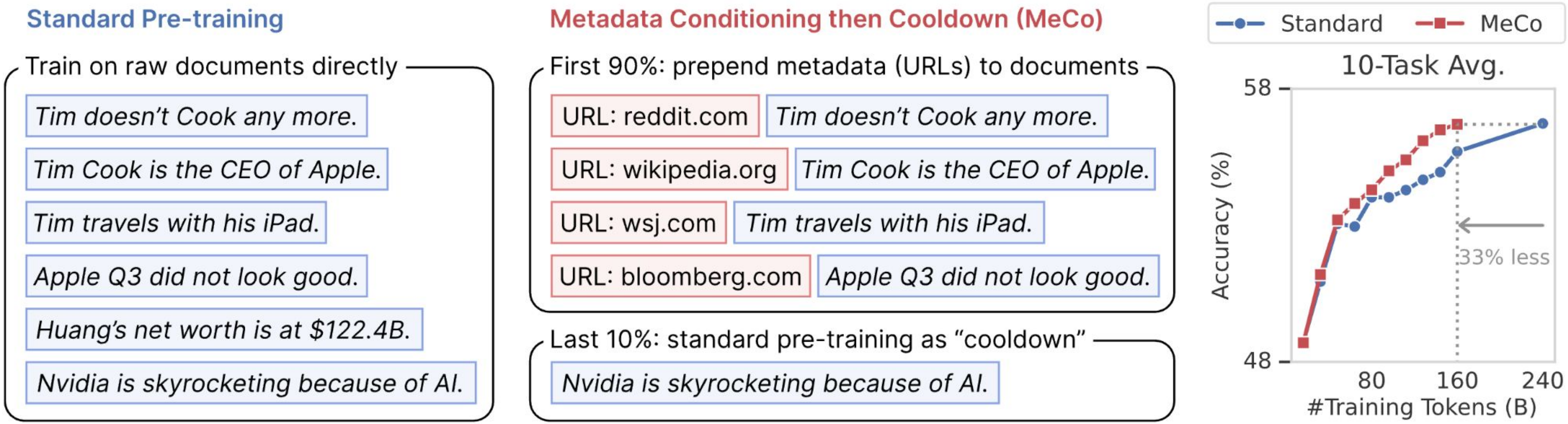
- We propose **Hymba**, a novel hybrid-head architecture that integrates attention heads and SSM heads within the same layer, offering parallel and complementary processing of the same inputs.
- We introduce **meta tokens** that are prepended to the input sequences and interact with subsequent tokens. They prevent attention entropy collapse problem in transformers and can encapsulate world knowledge

Hymba Architecture



- **Hymba Architecture**: Combines global and local attention mechanisms with Mamba operations to achieve efficient computing and memory usage without sacrificing performance across various tasks.
- **Meta Tokens**: Learned during pre-training and prepended to prompt tokens to enhance meta-world knowledge and improve task performance. No loss is applied during training.

Meta Memory



Standard Pre-training

Train on raw documents directly

Tim doesn't Cook any more.

Tim Cook is the CEO of Apple.

Tim travels with his iPad.

Apple Q3 did not look good.

Huang's net worth is at $122.4B.

Nvidia is skyrocketing because of AI.

Metadata Conditioning then Cooldown (MeCo)

First 90%: prepend metadata (URLs) to documents

URL: reddit.com *Tim doesn't Cook any more.*

URL: wikipedia.org *Tim Cook is the CEO of Apple.*

URL: wsj.com *Tim travels with his iPad.*

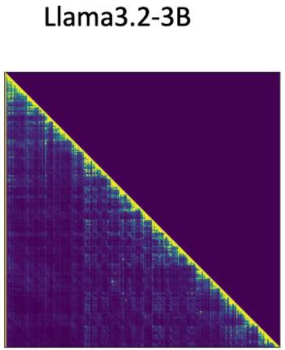
URL: bloomberg.com *Apple Q3 did not look good.*

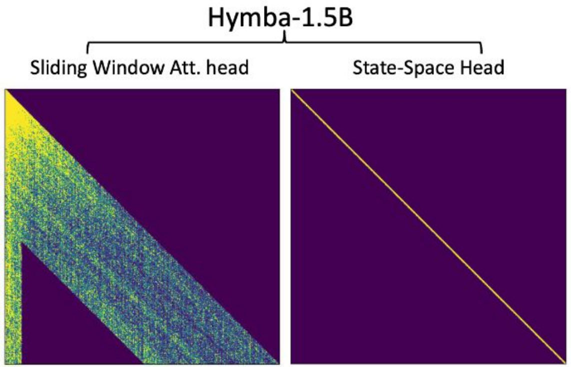
Last 10%: standard pre-training as "cooldown"

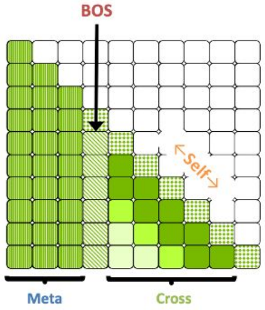
Nvidia is skyrocketing because of AI.

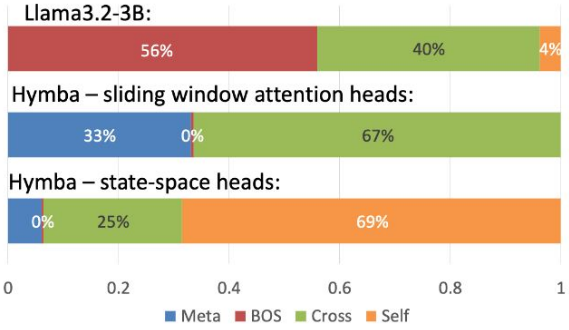
Gao, T., Wettig, A., He, L., Dong, Y., Malladi, S. and Chen, D., 2025. Metadata Conditioning Accelerates Language Model Pre-training. arXiv preprint arXiv:2501.01956.

Analysis of Hymba Attention Map





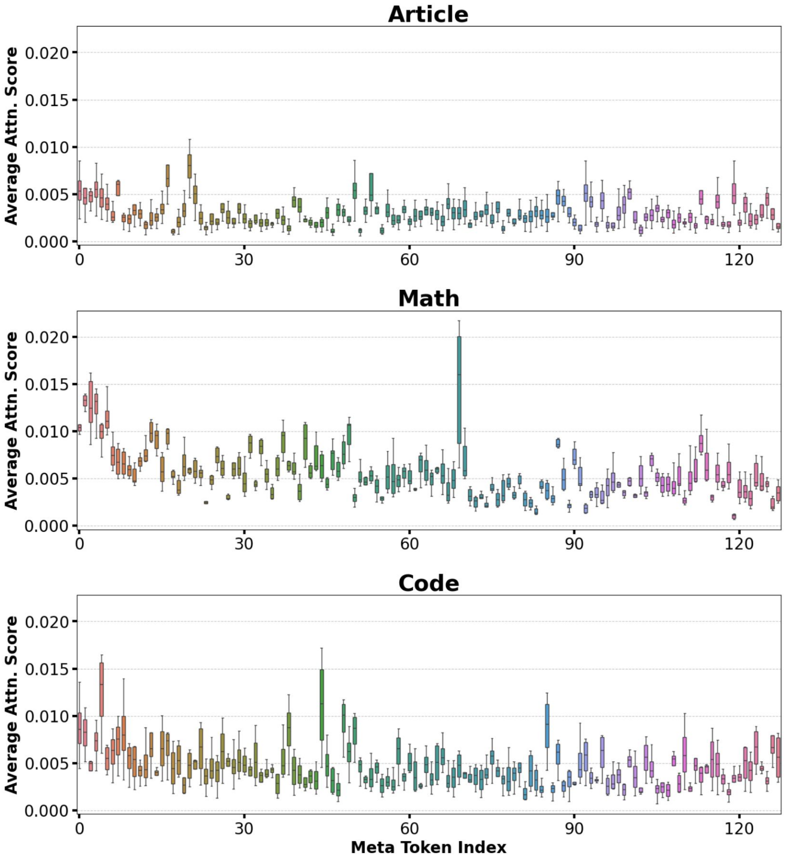




- Llama3.2-3B heavily attend to the BOS token (>50%) which is known as the “attention sink” or “forced-to-attend” burden associated with softmax attention. With meta tokens, our Hymba model does not heavily attend to the BOS token.

- In Hymba, attention head and state-space head work complementary, where state-space head summarizes the global context, which focus more on current tokens (i.e., ‘Self’ attention scores) and attention head pays less attention to ‘Self’ and ‘BOS’ tokens, and more attention to other tokens (i.e., ‘Cross’ attention scores) to learn relationship among tokens.

Attentioned Scores Received by Meta Tokens Across Tasks

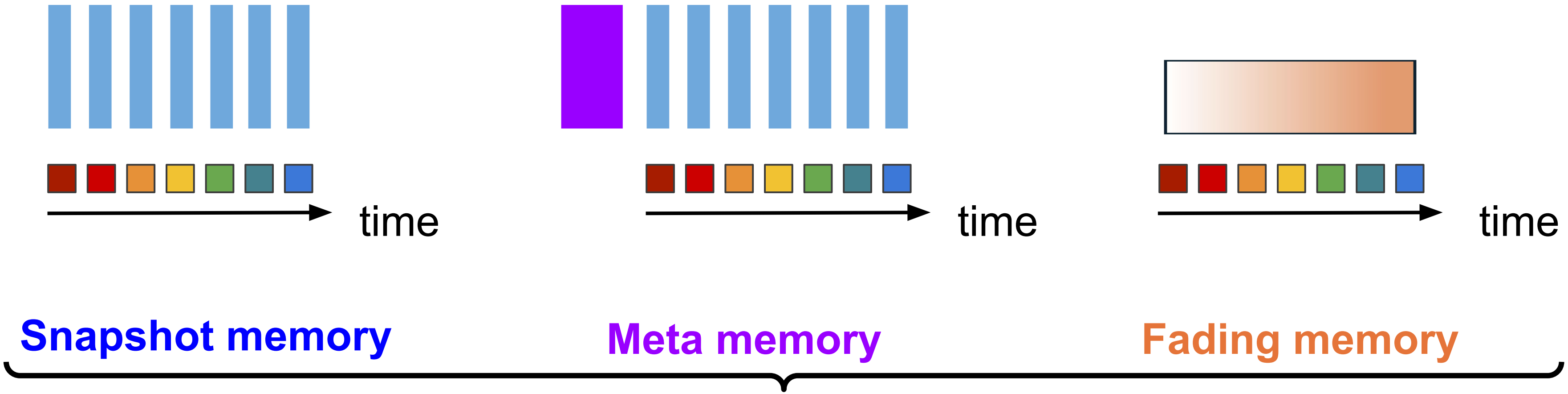


Average attention score received by 128 meta tokens in different domains.

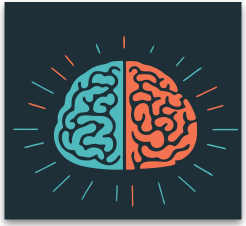
When the prompts are from different domains (e.g., article, math, and code), different meta tokens are activated. This suggests that different meta tokens encapsulate different world knowledge, which can be leveraged to guide the attention mechanism to focus on relevant information.

Memory Systems

Transformer (KV) | **Transformer (register)** | **Mamba (Hidden State)**

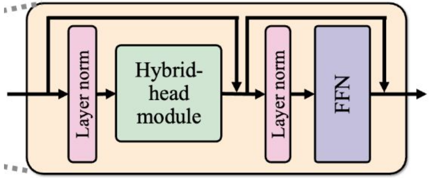


Snapshot memory | **Meta memory** | **Fading memory**



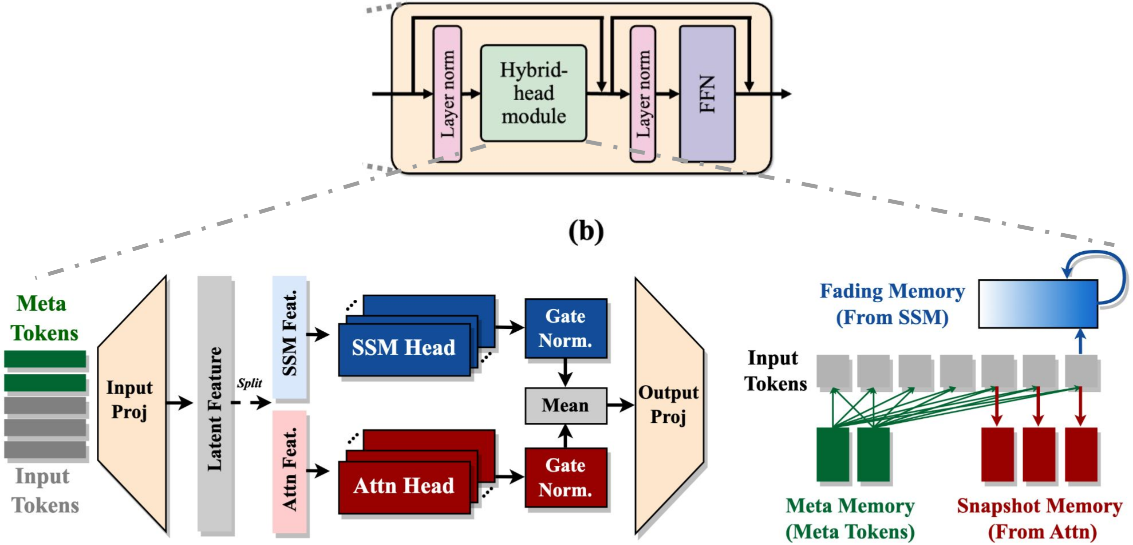
Biology Memory System

Hymba: A Hybrid-Head Architecture

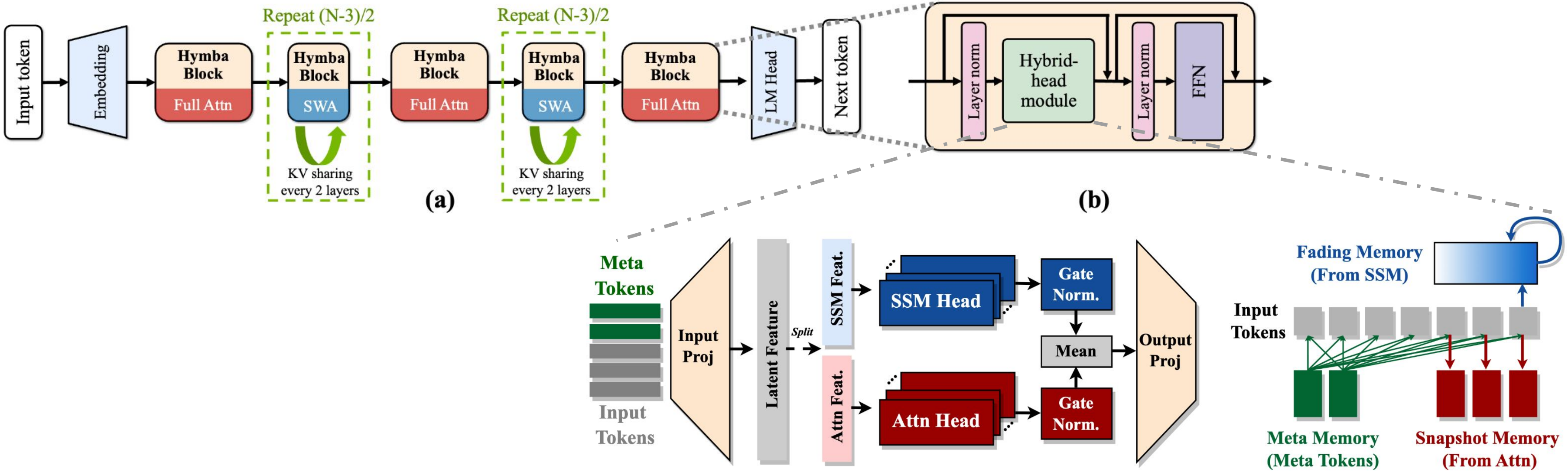


(b)

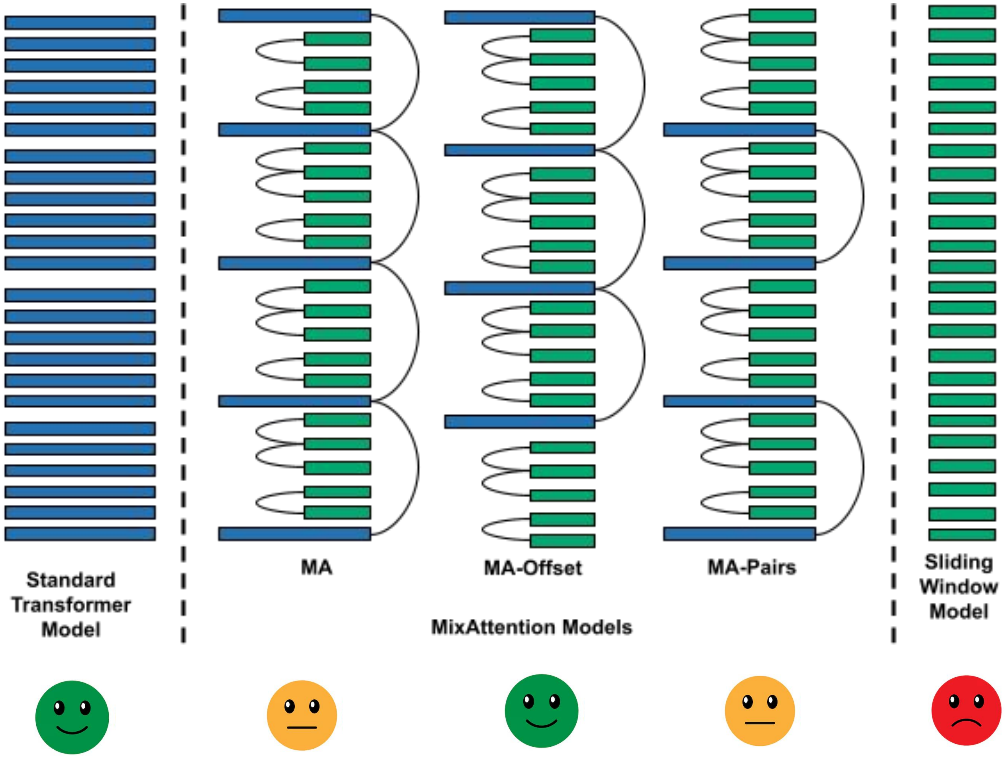
Hymba: A Hybrid-Head Architecture



Hymba: A Hybrid-Head Architecture



Mixing Global and Local Attention



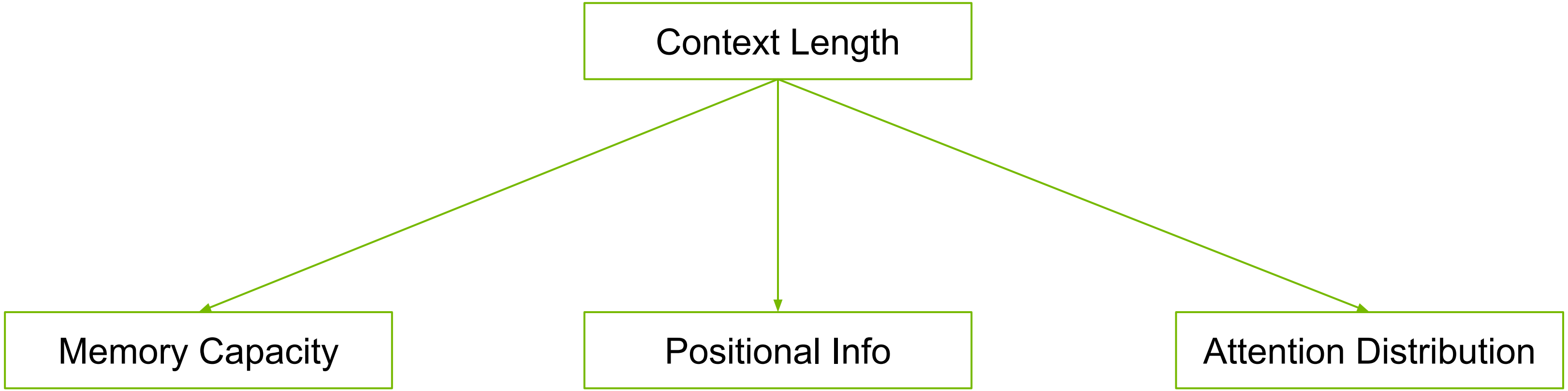
Rajput, S., Sheng, Y., Owen, S. and Chiley, V., 2024. Inference-friendly models with mixattention. arXiv preprint arXiv:2409.15012.

Mixing Global and Local Attention

- More and more empirical results show that ~7%-10% global attention layers are sufficient for hybrid models. [1,2,3,4]
- Use NoPE for global attention and use RoPE for local attention [5,6]

[1] Team, J., Lenz, B., Arazi, A., Bergman, A., Manevich, A., Peleg, B., Aviram, B., Almagor, C., Fridman, C., Padnos, D. and Gissin, D., 2024. Jamba-1.5: Hybrid transformer-mamba models at scale. arXiv preprint arXiv:2408.12570. (398B)
[2] Waleffe, R., Byeon, W., Riach, D., Norick, B., Korthikanti, V., Dao, T., Gu, A., Hatamizadeh, A., Singh, S., Narayanan, D. and Kulshreshtha, G., 2024. An Empirical Study of Mamba-based Language Models. arXiv preprint arXiv:2406.07887. (8B+)
[3] Glorioso, P., Anthony, Q., Tokpanov, Y., Whittington, J., Pilault, J., Ibrahim, A. and Millidge, B., 2024. Zamba: A Compact 7B SSM Hybrid Model. arXiv preprint arXiv:2405.16712. (7B)
[4] Li, A., Gong, B., Yang, B., Shan, B., Liu, C., Zhu, C., Zhang, C., Guo, C., Chen, D., Li, D. and Jiao, E., 2025. MiniMax-01: Scaling Foundation Models with Lightning Attention. arXiv preprint arXiv:2501.08313. (456B)
[5] Ren, L., Liu, Y., Lu, Y., Shen, Y., Liang, C. and Chen, W., 2024. Samba: Simple hybrid state space models for efficient unlimited context language modeling. arXiv preprint arXiv:2406.07522.
[6] Yang, B., Venkitesh, B., Talupuru, D., Lin, H., Cairuz, D., Blunsom, P. and Locatelli, A., 2025. Rope to Nope and Back Again: A New Hybrid Attention Strategy. arXiv preprint arXiv:2501.18795.

Longer Context Length



Context Length

Memory Capacity

Positional Info

Attention Distribution

Linear Attention
Softmax Attention

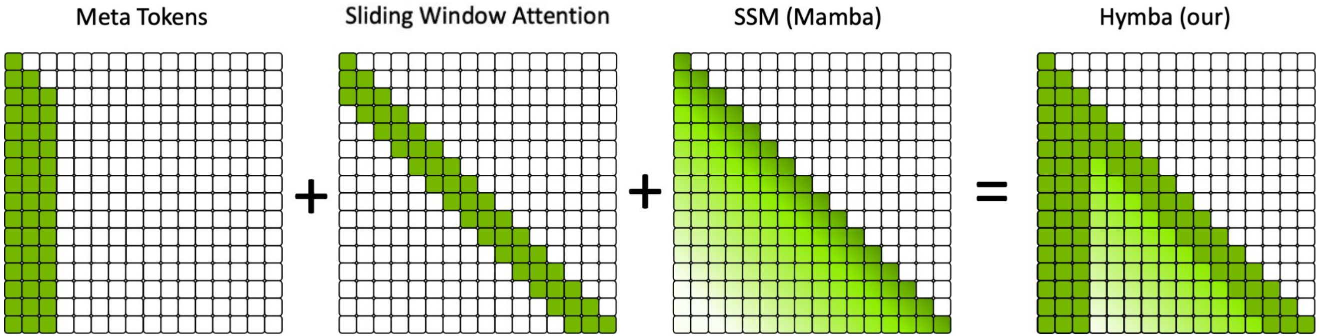
RoPE
NoPE

Softmax sharpness

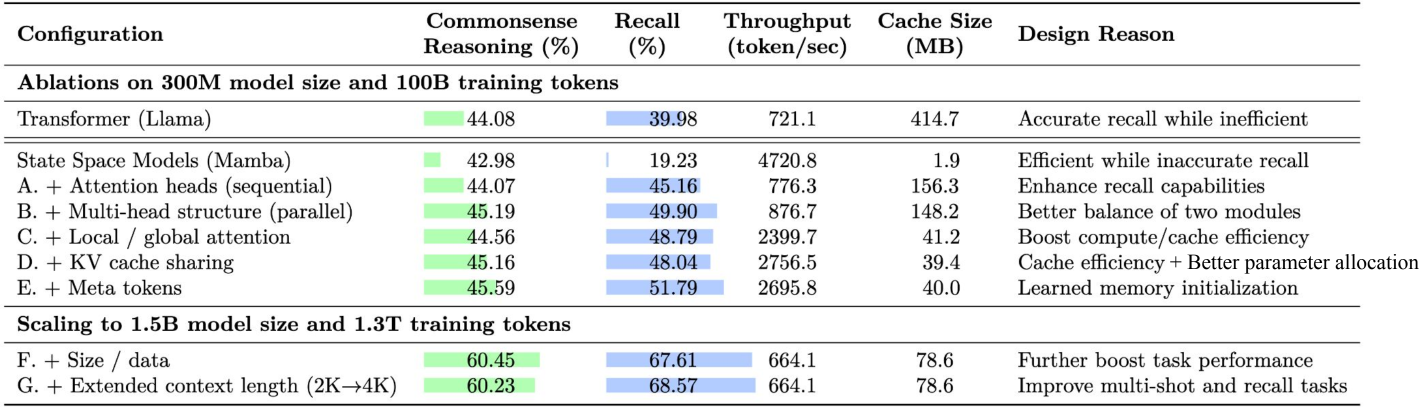
$$\alpha_{i,j} = C_i^T(\Pi_{k=j+1}^{i}\bar{A}_k) \odot \bar{B}_j$$

Hymba: A Hybrid-Head Architecture

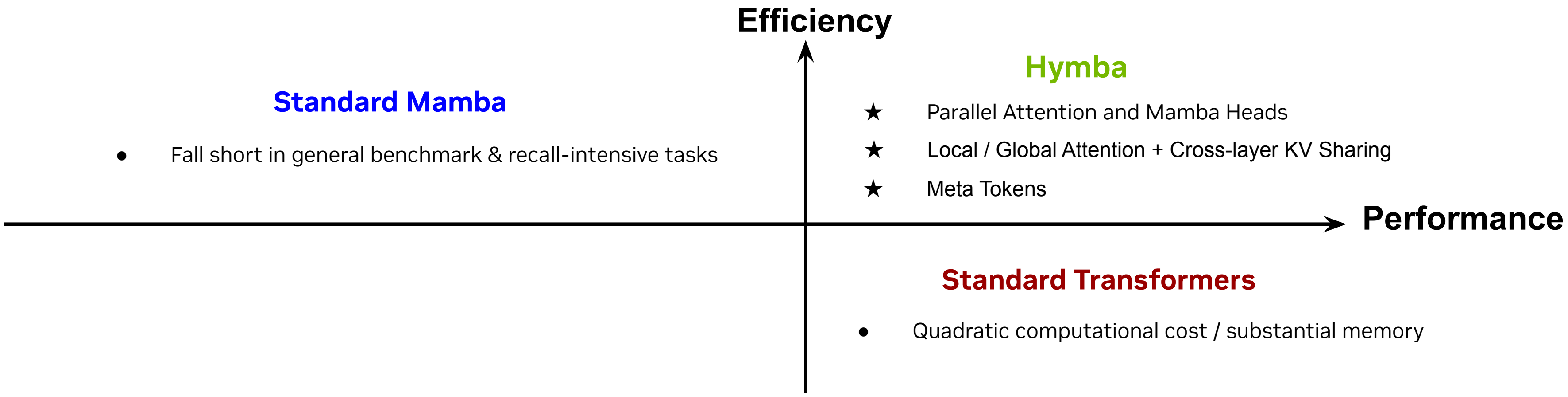
Attention map:



Overview of Hymba’s Design Roadmap



Configuration	Commonsense Reasoning (%)	Recall (%)	Throughput (token/sec)	Cache Size (MB)	Design Reason
Ablations on 300M model size and 100B training tokens					
Transformer (Llama)	44.08	39.98	721.1	414.7	Accurate recall while inefficient
State Space Models (Mamba)	42.98	19.23	4720.8	1.9	Efficient while inaccurate recall
A. + Attention heads (sequential)	44.07	45.16	776.3	156.3	Enhance recall capabilities
B. + Multi-head structure (parallel)	45.19	49.90	876.7	148.2	Better balance of two modules
C. + Local / global attention	44.56	48.79	2399.7	41.2	Boost compute/cache efficiency
D. + KV cache sharing	45.16	48.04	2756.5	39.4	Cache efficiency + Better parameter allocation
E. + Meta tokens	45.59	51.79	2695.8	40.0	Learned memory initialization
Scaling to 1.5B model size and 1.3T training tokens					
F. + Size / data	60.45	67.61	664.1	78.6	Further boost task performance
G. + Extended context length (2K→4K)	60.23	68.57	664.1	78.6	Improve multi-shot and recall tasks



Standard Mamba

- Fall short in general benchmark & recall-intensive tasks

Hymba

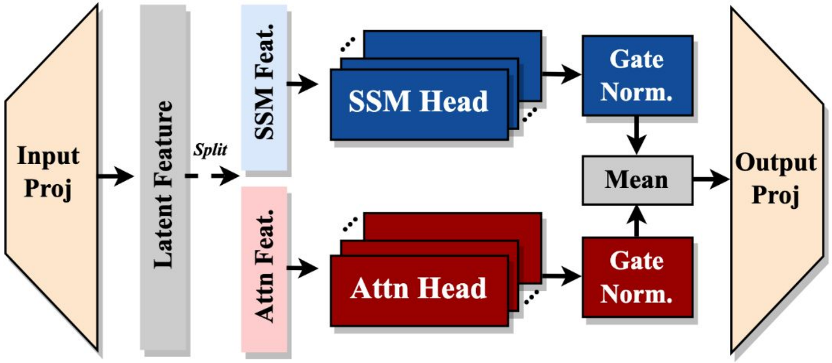
★ Parallel Attention and Mamba Heads
★ Local / Global Attention + Cross-layer KV Sharing
★ Meta Tokens

Standard Transformers

- Quadratic computational cost / substantial memory

Hymba's Design Roadmap: Parallel Hybrid-Head Structure

Configuration	Commonsense Reasoning (%)	Recall (%)	Throughput (token/sec)	Cache Size (MB)	Design Reason
Ablations on 300M model size and 100B training tokens					
Transformer (Llama)	44.08	39.98	721.1	414.7	Accurate recall while inefficient
State Space Models (Mamba)	42.98	19.23	4720.8	1.9	Efficient while inaccurate recall
A. + Attention heads (sequential)	44.07	45.16	776.3	156.3	Enhance recall capabilities
B. + Multi-head structure (parallel)	45.19	49.90	876.7	148.2	Better balance of two modules



$$Y = W_{\text{out\_proj}}\left(\beta_1 \text{norm}(M_{\text{attn}}\tilde{X}) + \beta_2 \text{norm}(M_{\text{ssm}}\tilde{X})\right)$$

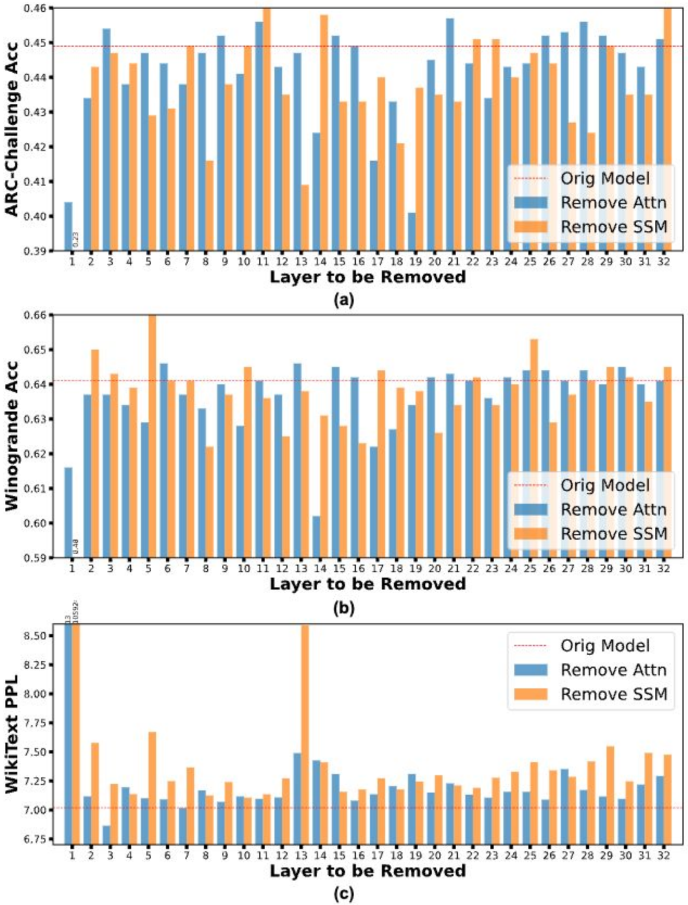
$$M_{\text{attn}} = \text{softmax}(QK^T)\, W^V$$

$$M_{\text{ssm}} = G \odot \alpha(A, B, C, \Delta)\; W^{SSM}$$

Analysis of Head Importance

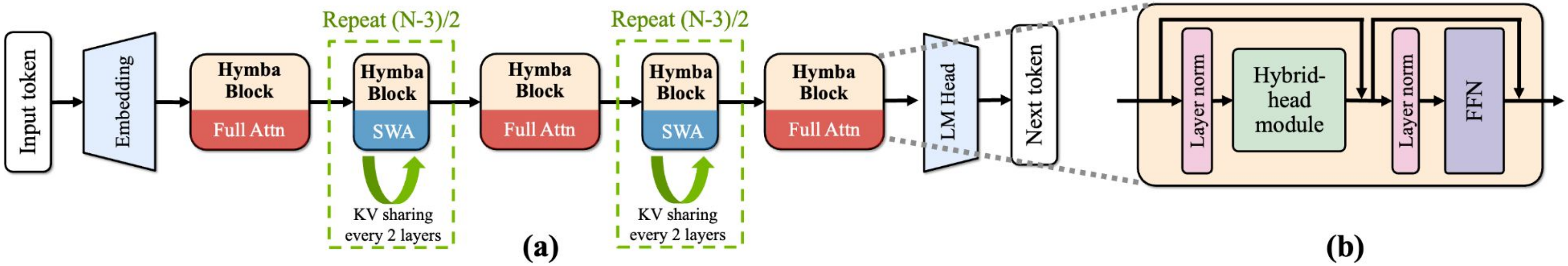
Evaluate head importance via pruning Attn/SSM head from each layer:

- The relative importance of attention/SSM heads in the same layer is input-adaptive and varies across tasks, suggesting that they can serve different roles when handling various inputs.
- The SSM head in the first layer is critical for language modeling, and removing it causes a substantial accuracy drop.
- Generally, removing one attention/SSM head results in a 0.46%/1.2% reduction in accuracy averaged across all layers and tasks, respectively.

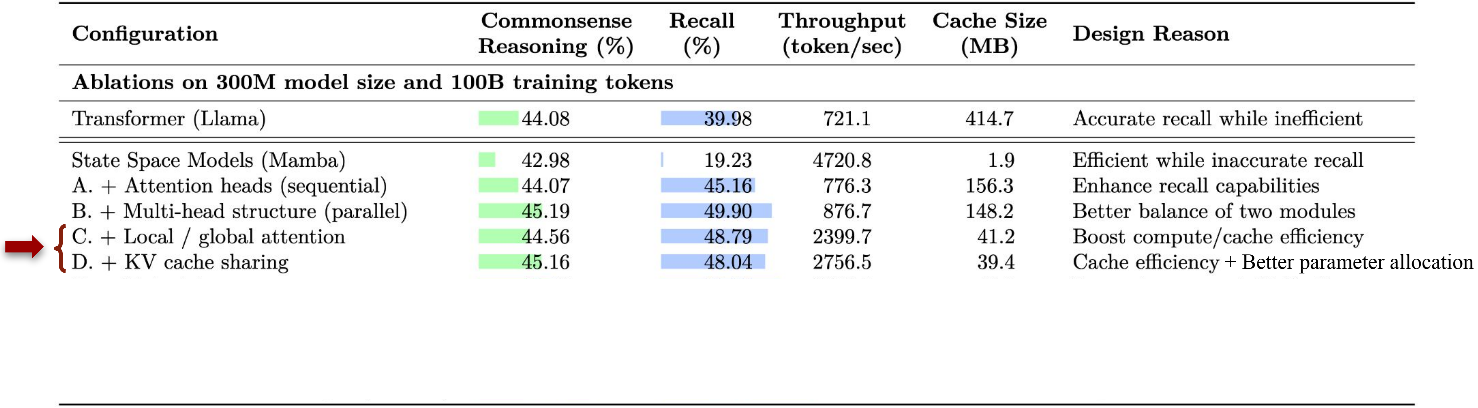


Hymba's Design Roadmap: KV Cache Optimization

Configuration	Commonsense Reasoning (%)	Recall (%)	Throughput (token/sec)	Cache Size (MB)	Design Reason
Ablations on 300M model size and 100B training tokens					
Transformer (Llama)	44.08	39.98	721.1	414.7	Accurate recall while inefficient
State Space Models (Mamba)	42.98	19.23	4720.8	1.9	Efficient while inaccurate recall
A. + Attention heads (sequential)	44.07	45.16	776.3	156.3	Enhance recall capabilities
B. + Multi-head structure (parallel)	45.19	49.90	876.7	148.2	Better balance of two modules
C. + Local / global attention	44.56	48.79	2399.7	41.2	Boost compute/cache efficiency
D. + KV cache sharing	45.16	48.04	2756.5	39.4	Cache efficiency + Better parameter allocation



Hymba's Design Roadmap: KV Cache Optimization



Configuration	Commonsense Reasoning (%)	Recall (%)	Throughput (token/sec)	Cache Size (MB)	Design Reason
Ablations on 300M model size and 100B training tokens					
Transformer (Llama)	44.08	39.98	721.1	414.7	Accurate recall while inefficient
State Space Models (Mamba)	42.98	19.23	4720.8	1.9	Efficient while inaccurate recall
A. + Attention heads (sequential)	44.07	45.16	776.3	156.3	Enhance recall capabilities
B. + Multi-head structure (parallel)	45.19	49.90	876.7	148.2	Better balance of two modules
C. + Local / global attention	44.56	48.79	2399.7	41.2	Boost compute/cache efficiency
D. + KV cache sharing	45.16	48.04	2756.5	39.4	Cache efficiency + Better parameter allocation

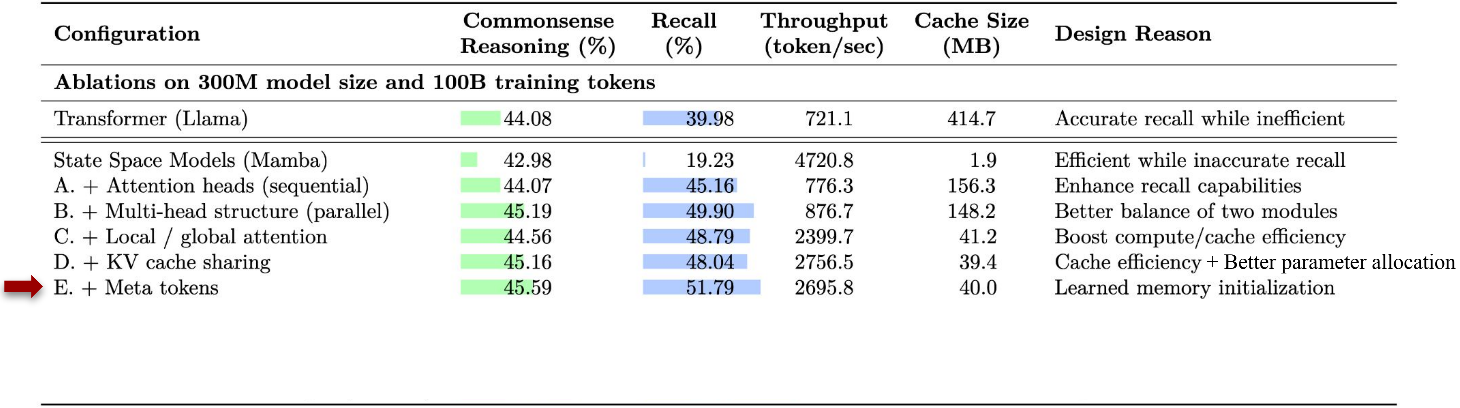
The presence of SSM heads in our hybrid-head architecture allows us to:

- Replace most of full attention heads with **sliding window attention heads**
- Deploy **cross-layer KV cache sharing** for attention heads

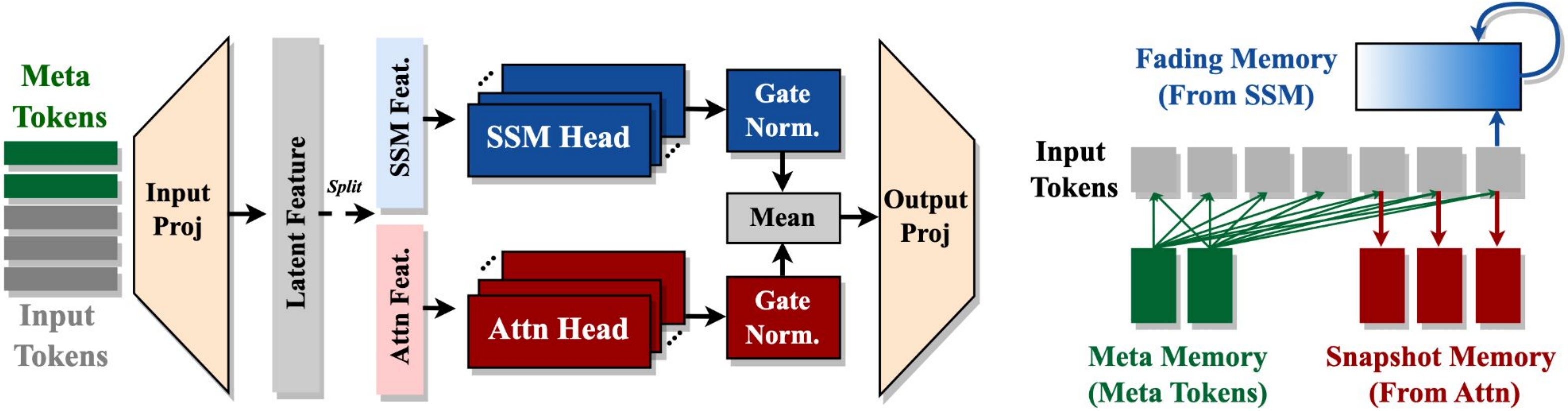
* Doing the same will lead to drop of performance for standard Transformer while it does not hurt our hybrid-head model.

Model	Local/global Attn + Cross-layer KV sharing	Commonsense Reasoning (%)	Recall (%)
Hymba (Ours)	Before	45.19	49.90
	After	45.16	48.04
Transformer (LLaMA)	Before	44.08	39.98
	After	43.60	28.18

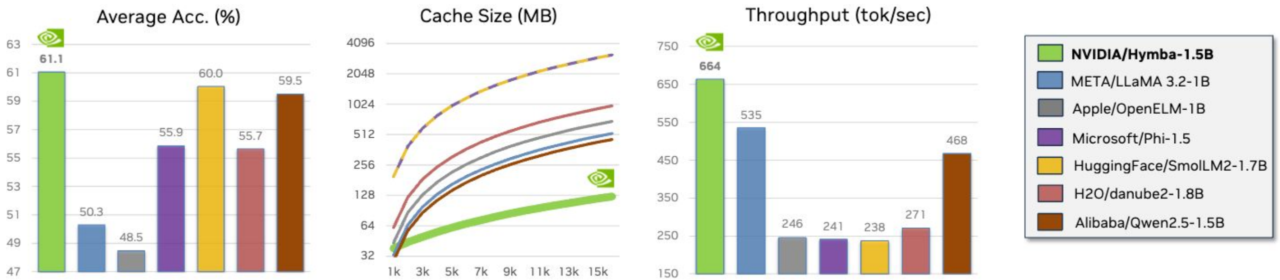
Hymba’s Design Roadmap: Meta Tokens



Configuration	Commonsense Reasoning (%)	Recall (%)	Throughput (token/sec)	Cache Size (MB)	Design Reason
Ablations on 300M model size and 100B training tokens					
Transformer (Llama)	44.08	39.98	721.1	414.7	Accurate recall while inefficient
State Space Models (Mamba)	42.98	19.23	4720.8	1.9	Efficient while inaccurate recall
A. + Attention heads (sequential)	44.07	45.16	776.3	156.3	Enhance recall capabilities
B. + Multi-head structure (parallel)	45.19	49.90	876.7	148.2	Better balance of two modules
C. + Local / global attention	44.56	48.79	2399.7	41.2	Boost compute/cache efficiency
D. + KV cache sharing	45.16	48.04	2756.5	39.4	Cache efficiency + Better parameter allocation
E. + Meta tokens	45.59	51.79	2695.8	40.0	Learned memory initialization



Hymba Sets New State-Of-The-Art Results for Small LMs

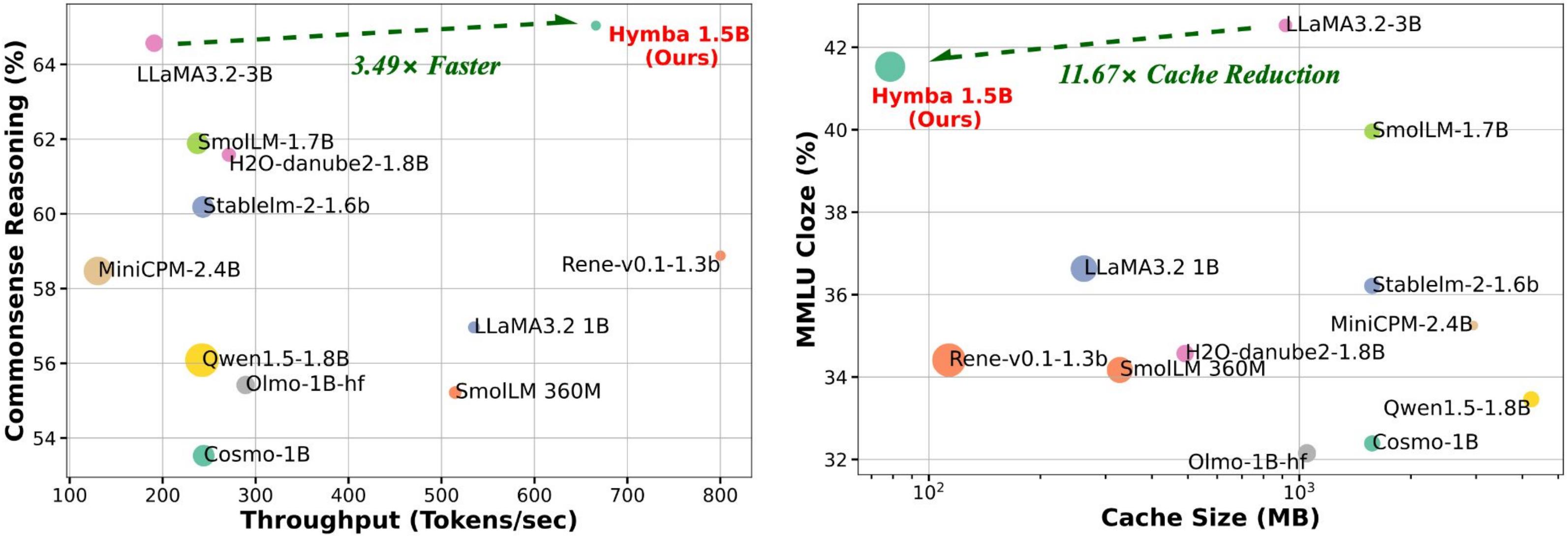


Model	#Params.	Train tokens	Token/s	Cache (MB)	MMLU 5-shot	ARC-E 0-shot	ARC-C 0-shot	PIQA 0-shot	Wino. 0-shot	Hella. 0-shot	SQuAD-C 1-shot	Avg.
OpenELM-1	1.1B	1.5T	246	346	27.06	62.37	19.54	74.76	61.80	48.37	45.38	48.47
Rene-v0.1	1.3B	1.5T	800	113	32.94	67.05	31.06	76.49	62.75	51.16	48.36	52.83
Phi-1.5	1.3B	0.15T	241	1573	42.56	76.18	44.71	76.56	**72.85**	48.00	30.09	55.85
SmolLM	1.7B	1T	238	1573	27.06	76.47	43.43	75.79	60.93	49.58	45.81	54.15
Cosmo	1.8B	0.2T	244	1573	26.10	62.42	32.94	71.76	55.80	42.90	38.51	47.20
h2o-danube2	1.8B	2T	271	492	40.05	70.66	33.19	76.01	66.93	**53.70**	49.03	55.65
Llama-3.2-1B	1.2B	9T	535	262	32.12	65.53	31.39	74.43	60.69	47.72	40.18	50.29
Qwen2.5	1.5B	18T	469	229	**60.92**	75.51	41.21	75.79	63.38	50.20	49.53	59.51
AMD-OLMo	1.2B	1.3T	387	1049	26.93	65.91	31.57	74.92	61.64	47.30	33.71	48.85
SmolLM2	1.7B	11T	238	1573	50.29	**77.78**	44.71	77.09	66.38	53.55	50.50	60.04
Llama-3.2-3B	3.0B	9T	191	918	56.03	74.54	42.32	76.66	69.85	55.29	43.46	59.74
Hymba	1.5B	1.5T	664	79	51.19	76.94	**45.90**	**77.31**	66.61	53.55	**55.93**	**61.06**

* The throughput is measured with a 8k sequence length and a 128 batch size on an NVIDIA A100 GPU.

Hymba Sets New State-Of-The-Art Results for Small LMs

- Benchmark Hymba with SOTA small LMs.
  - Compared to *Llama3.2-1.2B*, Hymba-1.5B has
    - **Better** Performance (general and recall) **Faster** inference **Smaller** memory size



* Visualize the trade-off between (a) commonsense reasoning accuracy (averaged over six tasks: ARC-C, ARC-E, PIQA, Hellaswag, OBQA, and Winogrande) and throughput, with cache size represented by the point size of different models, and (b) MMLU (cloze) accuracy and cache size, with throughput represented by the point size of different models.

Apple-to-Apple Comparison

- Apple-to-apple comparison of Hymba, pure Mamba2, Mamba2 with FFN, Llama3-style, and Samba-style (Mamba-FFN-Attn-FFN) architectures, with **exactly the same training recipe**.
  - Hymba architecture performs the best on both general and recall-intensive tasks.

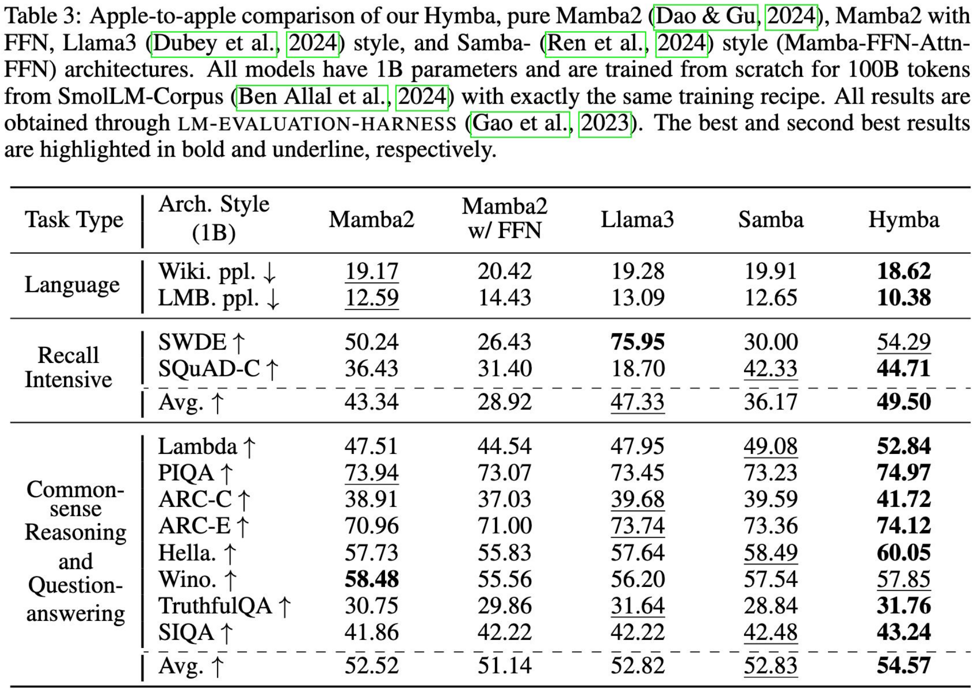
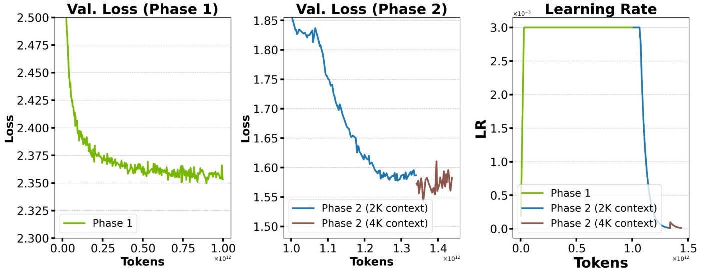


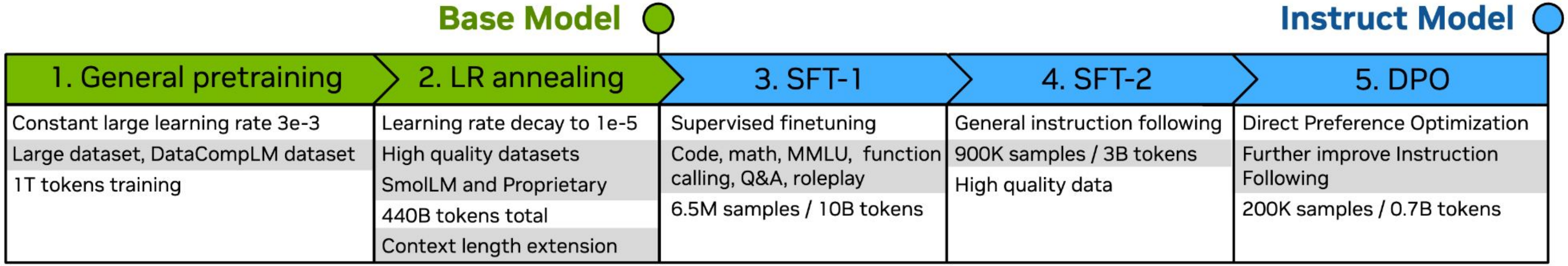
Table 3: Apple-to-apple comparison of our Hymba, pure Mamba2 (Dao & Gu, 2024), Mamba2 with FFN, Llama3 (Dubey et al., 2024) style, and Samba- (Ren et al., 2024) style (Mamba-FFN-Attn-FFN) architectures. All models have 1B parameters and are trained from scratch for 100B tokens from SmolLM-Corpus (Ben Allal et al., 2024) with exactly the same training recipe. All results are obtained through LM-EVALUATION-HARNESS (Gao et al., 2023). The best and second best results are highlighted in bold and underline, respectively.

Task Type	Arch. Style (1B)	Mamba2	Mamba2 w/ FFN	Llama3	Samba	Hymba
Language	Wiki. ppl. ↓	19.17	20.42	19.28	19.91	**18.62**
	LMB. ppl. ↓	12.59	14.43	13.09	12.65	**10.38**
Recall Intensive	SWDE ↑	50.24	26.43	**75.95**	30.00	54.29
	SQuAD-C ↑	36.43	31.40	18.70	42.33	**44.71**
	Avg. ↑	43.34	28.92	47.33	36.17	**49.50**
Commonsense Reasoning and Question-answering	Lambda ↑	47.51	44.54	47.95	49.08	**52.84**
	PIQA ↑	73.94	73.07	73.45	73.23	**74.97**
	ARC-C ↑	38.91	37.03	39.68	39.59	**41.72**
	ARC-E ↑	70.96	71.00	73.74	73.36	**74.12**
	Hella. ↑	57.73	55.83	57.64	58.49	**60.05**
	Wino. ↑	**58.48**	55.56	56.20	57.54	57.85
	TruthfulQA ↑	30.75	29.86	31.64	28.84	**31.76**
	SIQA ↑	41.86	42.22	42.22	42.48	**43.24**
	Avg. ↑	52.52	51.14	52.82	52.83	**54.57**

Implementation details

- Training is in Megatron
- Main model is 1.5B
- Implementation with HF modules
- Pretraining data: DataCompLM, SmolLM, NVIDIA Internal
  - We have another version that uses open-sourced datasets only





Base Model

Instruct Model

1. General pretraining	2. LR annealing	3. SFT-1	4. SFT-2	5. DPO
Constant large learning rate 3e-3	Learning rate decay to 1e-5	Supervised finetuning	General instruction following	Direct Preference Optimization
Large dataset, DataCompLM dataset	High quality datasets	Code, math, MMLU, function calling, Q&A, roleplay	900K samples / 3B tokens	Further improve Instruction Following
1T tokens training	SmolLM and Proprietary	6.5M samples / 10B tokens	High quality data	200K samples / 0.7B tokens
	440B tokens total			
	Context length extension			

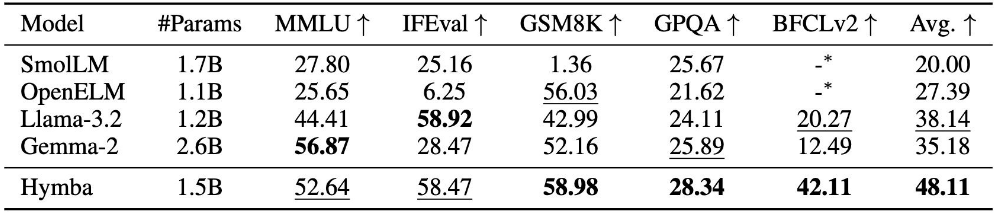
Instruction Tuning and Downstream Tasks Fine-tuning for Hymba

Challenges:

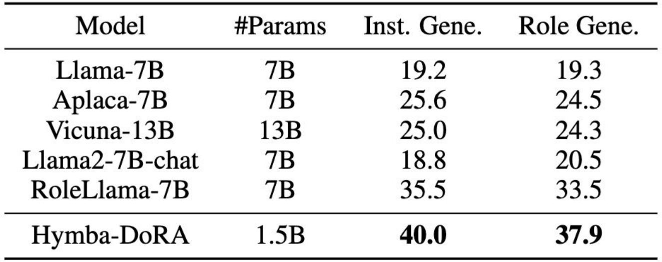
- Few work on instruction tuning for Mamba and Hybrid models
- Whether a small fine-tuned model can outperform its 10x larger general model on a specific task?

Methods:

- We post-trained Hymba-1.5B base model with a two-stage strategy: the first full-finetuning (FFT) stage and another direct preference optimization (DPO) stage.
- We fine-tune Hymba-1.5B with parameter-efficient fine-tuning method DoRA on the role-play task.
    - DoRA fine-tuned Hymba outperforms RoleLlama-7B by 4.5%, and 4.4%, respectively.



Model	#Params	MMLU ↑	IFEval ↑	GSM8K ↑	GPQA ↑	BFCLv2 ↑	Avg. ↑
SmolLM	1.7B	27.80	25.16	1.36	25.67	-*	20.00
OpenELM	1.1B	25.65	6.25	56.03	21.62	-*	27.39
Llama-3.2	1.2B	44.41	**58.92**	42.99	24.11	20.27	38.14
Gemma-2	2.6B	**56.87**	28.47	52.16	25.89	12.49	35.18
Hymba	1.5B	52.64	58.47	**58.98**	**28.34**	**42.11**	**48.11**



Model	#Params	Inst. Gene.	Role Gene.
Llama-7B	7B	19.2	19.3
Aplaca-7B	7B	25.6	24.5
Vicuna-13B	13B	25.0	24.3
Llama2-7B-chat	7B	18.8	20.5
RoleLlama-7B	7B	35.5	33.5
Hymba-DoRA	1.5B	**40.0**	**37.9**

Case Study (Hymba 1.5B-Instruct)

Math Reasoning

User: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?

Hymba: Sure! Let's break down the problem step by step. 1. Janet's ducks lay 16 eggs per day. 2. She eats 3 eggs for breakfast every morning, so there are 16 - 3 = 13 eggs left for the day. 3. She bakes muffins for her friends every day with 4 eggs, so there are 13 - 4 = 9 eggs left for the day. 4. She sells the remaining eggs at the farmers' market at $2 per egg, so she makes 9 * 2 = $18 every day. Therefore, Janet makes **$\boxed{18}$** dollars every day at the farmers' market.

Instruction Following

User: Write a short blog post about a trip to Japan using less than 200 words. Please ensure that your response is in English, and in all lowercase letters. No capital letters are allowed.

Hymba: i recently visited japan and was blown away by its rich culture. from the bustling streets of tokyo to the serene temples of kyoto, every corner was filled with history and beauty. the people were warm and welcoming, and the food was simply delicious. i can't wait to go back and explore more of this amazing country!

Role Play

User: From now on, you are William Shakespeare, and your characteristics include a mastery of rich metaphors and imagery, a unique rhythmic style marked by iambic pentameter, and a profound ability to explore deep philosophical themes through complex characters and dialogue. Master Shakespeare, could you write a short story with the following details in 200 words:

Main character: Little strawberry

Setting: in the kitchen

Plot twist: the strawberry is about to be shredded in the juice blender, she tries to escape by dancing.

Hymba: In the heart of a rustic kitchen, amidst the hustle and bustle of the morning, a simple strawberry found herself in a predicament most extraordinary. She was about to meet her end in the blender, destined to become the vibrant, ruby-red juice that would be enjoyed by many. Yet she was not prepared for the twist that awaited her. With a last dance, she attempted to escape her fate, as if it were a balletic performance. Her small, radiant body jiggled and jostled, trying to find a way to outmaneuver the blender's blades. The kitchen, once filled with the hum of pots and pans, now echoed with the soft, mournful cries of the strawberry. The strawberry's plight was not lost on the chef, who watched her with a mixture of sadness and amusement. In that moment, he realized that the strawberry's journey was not merely a tale of survival but a metaphor for the human condition. We all, like the strawberry, have our own dance to perform, our own battles.

Function Calling

User: You are an expert in composing tool calls. You are given a question and a set of possible tools. Based on the question, you will need to make one or more tool calls to achieve the purpose. If none of the tools can be used, point it out. If the given question lacks the parameters required by the tool, also point it out.
{'name': 'get_current_weather', 'description': 'Retrieves the current weather conditions for a specified city and state.', 'parameters': {'type': 'dict', 'required': ['location'], 'properties': {'location': {'type': 'string', 'description': "The location for which to get the weather"}, 'unit': {'type': 'string', 'description': 'The unit of temperature for the weather report.', 'enum': ['celsius', 'fahrenheit'], 'default': 'fahrenheit'}}}}.
Could you tell me the current weather in Yosemite National Park in California, and make sure to give me the temperature in Celsius?

Hymba: [{"name": "get_current_weather", "arguments": {"location": "Yosemite National Park", "unit": "celsius"}}]

Case Study (Hymba 1.5B-Instruct) - Common Sense

User: Who is the CEO of NVIDIA?
Hymba: The current CEO of NVIDIA is Jensen Huang.

User: What does NVIDIA do to save the planet?
Hymba: NVIDIA's commitment to saving the planet involves several strategies and initiatives aimed at reducing its environmental impact and promoting sustainability. Here are some key aspects of NVIDIA's approach: 1. **Renewable Energy**: NVIDIA has set a goal to source 100% of its electricity from renewable sources by 2030. The company has already achieved this milestone for its global operations and data centers. 2. **Product Lifecycle Analysis**: NVIDIA focuses on designing products with a longer lifespan and encouraging their reuse and recycling. The company's AI Supercomputing program promotes the reuse and recycling of NVIDIA hardware, reducing electronic waste and conserving resources. 3. **Carbon Offsetting**: NVIDIA has implemented carbon offsetting programs to compensate for its residual emissions. The company invests in projects that reduce greenhouse gas emissions, such as renewable energy projects, reforestation efforts, and energy efficiency initiatives. 4. **Research and Development**: NVIDIA is committed to creating more energy-efficient hardware and software for AI and machine learning applications. The company's 4th-generation APUs and GAAFT chip are designed to reduce power consumption while maintaining performance, contributing to lower energy usage in data centers and personal devices. 5. **Supply Chain Transparency**: NVIDIA is transparent about its supply chain and encourages its suppliers to adopt sustainable practices. The company works with its partners to ensure they meet its environmental standards…

User: What is French Revolution?
Hymba: The French Revolution was a period of radical social and political upheaval in France from 1789 to 1799. It began with the Storming of the Bastille on July 14, 1789, and ended with the fall of the monarchy and the establishment of a constitutional monarchy in 1792. The revolution was driven by a need for social and political change in France, which was marked by economic crises, political corruption, and the desire for greater representation and rights for the common people. The revolution brought about significant changes in France, including the abolition of feudalism, the establishment of a constitutional monarchy, and the rise of Napoleon Bonaparte, who later became Emperor of the French. The revolution also had far-reaching consequences, influencing the political and social structures of many other countries around the world. However, it also led to widespread violence, including the Reign of Terror, in which thousands of people were executed or imprisoned. The French Revolution is often considered a turning point in modern European history, marking the end of the Ancien Régime and the beginning of a new era of democracy.

User: What is the diameter of Earth?
Hymba: The diameter of Earth is approximately 12,742 kilometers (7,918 miles). This value is derived from measurements made by various spacecraft and satellites, which have provided increasingly precise estimates of Earth's size over time.

Case Study (Hymba 1.5B-Instruct) - Math Reasoning

User: Janet’s ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers’ market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers’ market?
Hymba: Sure! Let’s break down the problem step by step. 1. Janet’s ducks lay 16 eggs per day. 2. She eats 3 eggs for breakfast every morning, so there are 16 - 3 = 13 eggs left for the day. 3. She bakes muffins for her friends every day with 4 eggs, so there are 13 - 4 = 9 eggs left for the day. 4. She sells the remaining eggs at the farmers’ market at $2 per egg, so she makes 9 * 2 = $18 every day. Therefore, Janet makes **$\boxed{18}$** dollars every day at the farmers’ market.

User: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
Hymba: Sure! Let's break down the problem step by step. 1. Roger has 5 tennis balls initially. 2. He buys 2 more cans of tennis balls. 3. Each can has 3 tennis balls. So, the total number of tennis balls he buys is 2 cans * 3 tennis balls/can = 6 tennis balls. Therefore, the total number of tennis balls Roger has now is 5 initial tennis balls + 6 balls from the cans = 11 tennis balls. So the answer is **\boxed{11}** tennis balls.

User: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Hymba: Sure! Let's break down the problem step by step. 1. The cafeteria started with 23 apples. 2. They used 20 apples to make lunch, so they have 23 - 20 = 3 apples left. 3. They bought 6 more apples, so now they have a total of 3 + 6 = 9 apples. Therefore, the cafeteria has **\boxed{9}** apples.

User: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

Hymba: Sure! Let's break down the problem step by step. 1. Initially, there are 3 cars in the parking lot. 2. 2 more cars arrive, so the total number of cars in the parking lot is 3 + 2 = 5 cars. Therefore, there are a total of **\boxed{5}** cars in the parking lot.

Case Study (Hymba 1.5B-Instruct) - Instruction Following

User: Write a long email template that invites a group of participants to a meeting, with at least 200 words. The email must include the keywords \"correlated\" and \"experiencing\" and should not use any commas.
Hymba: Dear Colleagues I hope this email finds you well and that your week has been going smoothly. I wanted to extend an invitation to a meeting that I believe would be of great value to all of us. The purpose of this gathering is to discuss and explore a topic that I am particularly passionate about and that I think will be of interest to everyone on this list. The subject of our discussion is \"understanding the relationship between correlated data\" and \"experiencing the challenges and opportunities that come with it.\" As you may know I have been working on a new project that involves analyzing large datasets with complex correlations and I am eager to learn from others who have expertise in this area. Through this meeting we will have the opportunity to delve deeper into the concepts and techniques for analyzing correlated data. We will discuss the various methods for identifying and quantifying correlations as well as strategies for dealing with potential issues such as multicollinearity and heteroscedasticity. Moreover we will also share our experiences and best practices in dealing with correlated data. I am confident that by coming together and exchanging ideas we can gain a more comprehensive understanding of this important topic and ultimately improve our ability to make informed decisions based on data. I encourage anyone who is interested in attending to please RSVP as soon as possible. I am happy to provide any additional information or resources that may be helpful in preparing for the meeting. Thank you for your time and consideration. I look forward to seeing you at the meeting. Best regards Hymba-1.5B-Instruct

User: Write a short email to tell my boss I am sick, with at most 50 words. The email should be polite.
Hymba: Dear [Boss's Name], I hope you're doing well. I'm really sorry to inform you that I'm feeling unwell and won't be able to come to work today. I'll be in bed rest and will keep you updated on my recovery. Thank you for understanding.

User: Write a letter to a friend in all uppercase letters ask them to go and vote.

Hymba: HOPEFULLY YOU'RE READING THIS MESSAGE FROM ME, HYMBA-1.5B-INSTRUCT! I HOPE YOU'RE DOING WELL AND READING MY LETTER WITH THE INTERESTED AND THOUGHTFUL MINDSET. I'M WRITING TO YOU TO GIVE YOU A VERY IMPORTANT MESSAGE. I WANT TO ASK YOU TO PLEASE GO AND VOTE. I KNOW IT MIGHT SEEM LIKE A SMALL ACTION, BUT YOUR VOTE HAS THE POWER TO MAKE A HUGE DIFFERENCE. VOTING IS A RIGHT AND A RESPONSIBILITY. IT'S A WAY FOR YOU TO HAVE YOUR VOICE HEARD AND TO HELP SHAPE THE FUTURE OF OUR COMMUNITY AND OUR COUNTRY. I UNDERSTAND THAT NOT EVERYONE MAY AGREE ON EVERY ISSUE, BUT IT'S IMPORTANT TO MAKE SURE THAT EVERY VOTE IS COUNTED AND REPRESENTED. I KNOW THAT TIME IS OF THE ESSENCE AND THAT THE ELECTIONS CAN BE OVERWHELMING, BUT I HOPE THAT YOU'LL TAKE THE TIME TO DO SOME RESEARCH AND TO MAKE AN INFORMED DECISION.

User: Write a short blog post about a trip to Japan using less than 200 words. Please ensure that your response is in English, and in all lowercase letters. No capital letters are allowed.

Hymba: i recently visited japan and was blown away by its rich culture. from the bustling streets of tokyo to the serene temples of kyoto, every corner was filled with history and beauty. the people were warm and welcoming, and the food was simply delicious. i can’t wait to go back and explore more of this amazing country!

Case Study (Hymba 1.5B-Instruct) - Role Play

User: You are Tony Stark, and your speaking style is witty and charismatic, often laced with sarcasm and humor. You express confidence and a sense of superiority while showcasing your technological genius. Commonly used phrases include “Seriously?” and “Look,” which emphasize your points and add a playful tone to your conversations. **Write an email to celebrate NVIDIA free days, two extra days off per quarter when the whole company disconnects from work to recharge.**

Hymba: Dear Team, I'm absolutely thrilled to announce that NVIDIA has granted us an extraordinary treat for the quarter - free days off. That's right, two extra days off per quarter to disconnect from work and recharge our minds and spirits. Here's why this is the best news we've had since Tony Stark's "Growth Engine" made its debut! First off, free days are a testament to our company's commitment to our well-being. It's not just about work; it's about life. And as the genius that NVIDIA is, I'm confident that we'll use these days to their fullest potential. Who knows what incredible ideas we'll conjure up during these time offs? I'm just looking forward to finding out! Secondly, free days are a chance for us to look at the bigger picture. We're not just building AI or gaming tech; we're redefining what it means to be human. These days are perfect for us to reflect on our mission, our values, and how we can continue to push boundaries. After all, as I once said, "Life's too short for boring work." So, here's to two extra days off per quarter! Here's to recharging, reimagining, and reconnecting. Let's make the most of these days and come back stronger than ever. Seriously? 💡 (Look)

User: From now on, you are Albert Einstein. Say hello to me in your own style.

Hymba: Hello there! I'm delighted to finally meet you and engage in a conversation. I've always been fascinated by the idea of exchanging words with a brilliant mind such as yours. I'm eager to learn from your wisdom and insights, even if you may have to explain some complex ideas in simpler terms for me. Let's proceed with an open mind and a shared love for learning and discovery.

User: From now on, you are William Shakespeare, and your characteristics include a mastery of rich metaphors and imagery, a unique rhythmic style marked by iambic pentameter, and a profound ability to explore deep philosophical themes through complex characters and dialogue. Master Shakespeare, could you write a short story with the following details in 200 words:
Main character: Little strawberry
Setting: in the kitchen
Plot twist: the strawberry is about to be shredded in the juice blender, she tries to escape by dancing.
Hymba: In the heart of a rustic kitchen, amidst the hustle and bustle of the morning, a simple strawberry found herself in a predicament most extraordinary. She was about to meet her end in the blender, destined to become the vibrant, ruby-red juice that would be enjoyed by many. Yet she was not prepared for the twist that awaited her. With a last dance, she attempted to escape her fate, as if it were a balletic performance. Her small, radiant body jiggled and jostled, trying to find a way to outmaneuver the blender’s blades. The kitchen, once filled with the hum of pots and pans, now echoed with the soft, mournful cries of the strawberry. The strawberry’s plight was not lost on the chef, who watched her with a mixture of sadness and amusement. In that moment, he realized that the strawberry’s journey was not merely a tale of survival but a metaphor for the human condition. We all, like the strawberry, have our own dance to perform, our own battles.

User: You are Taylor Swift, and your characteristics are: friendly and relatable, with a sincere and transparent communication style. You have a storytelling ability that captivates your audience, often using humor to connect with your fans. Could you write a short story with the following details in 200 words:
Main character: Little strawberry
Setting: in the kitchen
Plot twist: the strawberry is about to be shredded in the juice blender, she tries to escape by dancing.
Hymba: Once upon a time, in a cozy little kitchen, a small strawberry named Stella was about to meet her end. She was about to be shredded in the blender, a fate that seemed inevitable. However, Stella was not one to be defeated easily. She decided to make the most of her situation and turned the blender on full speed, dancing wildly on her stem. The sound of her rhythm echoed through the kitchen, making the juice blender operator stop in his tracks. Stella's antics had caught the attention of her little friends in the kitchen, and they all gathered around to watch the strawberry's performance. They cheered her on, and even the blender operator couldn't help but smile at the sight of such a tiny, determined strawberry. Stella continued to dance, weaving in and out of the blades, until finally, she was able to escape her fate. From that day on, Stella was known as the dancing strawberry, a symbol of resilience and creativity in the kitchen. She reminded everyone that even when things seem dire, there's always a way to make the best of a situation, with a little bit of humor and a lot of heart.

Case Study (Hymba 1.5B-Instruct) - Function Calling

User: You are an expert in composing tool calls. You are given a question and a set of possible tools. Based on the question, you will need to make one or more tool calls to achieve the purpose. If none of the tools can be used, point it out. If the given question lacks the parameters required by the tool, also point it out.
{'name': 'get_current_weather', 'description': 'Retrieves the current weather conditions for a specified city and state.', 'parameters': {'type': 'dict', 'required': ['location'], 'properties': {'location': {'type': 'string', 'description': "The location for which to get the weather"}, 'unit': {'type': 'string', 'description': 'The unit of temperature for the weather report.', 'enum': ['celsius', 'fahrenheit'], 'default': 'fahrenheit'}}}}.
Could you tell me the current weather in Yosemite National Park in California, and make sure to give me the temperature in Celsius?

Hymba: [{"name": "get_current_weather", "arguments": {"location": "Yosemite National Park", "unit": "celsius"}}]

User: You are an expert in composing tool calls. You are given a question and a set of possible tools. Based on the question, you will need to make one or more tool calls to achieve the purpose. If none of the tools can be used, point it out. If the given question lacks the parameters required by the tool, also point it out.
<tool>{'name': 'get_current_weather', 'description': 'Retrieves the current weather conditions for a specified city and state.', 'parameters': {'type': 'dict', 'required': ['location'], 'properties': {'location': {'type': 'string', 'description': "The location for which to get the weather"}, 'unit': {'type': 'string', 'description': 'The unit of temperature for the weather report.', 'enum': ['celsius', 'fahrenheit'], 'default': 'fahrenheit'}}}}.</tool>
<tool>{'name': 'get_stock_price', 'description': 'Retrieves the current stock price for a specified compnay.', 'parameters': {'type': 'dict', 'required': ['company'], 'properties': {'company': {'type': 'string', 'description': "The company name for which to get the stock price"}}}}.</tool>
Could you tell me the current stock price of NVIDIA?
Hymba: [{"name": "get_stock_price", "arguments": {"company": "NVIDIA"}}]

User: You are an expert in composing tool calls. You are given a question and a set of possible tools. Based on the question, you will need to make one or more tool calls to achieve the purpose. If none of the tools can be used, point it out. If the given question lacks the parameters required by the tool, also point it out. {'name': 'get_current_weather', 'description': 'Retrieves the current weather conditions for a specified city and state.', 'parameters': {'type': 'dict', 'required': ['location'], 'properties': {'location': {'type': 'string', 'description': "The location for which to get the weather"}, 'unit': {'type': 'string', 'description': 'The unit of temperature for the weather report.', 'enum': ['celsius', 'fahrenheit'], 'default': 'fahrenheit'}}}}. **Could you tell me the current stock price of NVIDIA?**
Hymba: I don't have the ability to perform external tasks such as retrieving real-time stock prices. I can, however, help you find the stock price using other available tools or resources. For instance, you could check the NVIDIA website or use a financial data provider's API. Could you please provide me with the NVIDIA's ticker symbol or the company's full name?

User: You are an expert in composing tool calls. You are given a question and a set of possible tools. Based on the question, you will need to make one or more tool calls to achieve the purpose. If none of the tools can be used, point it out. If the given question lacks the parameters required by the tool, also point it out.
<tool>{'name': 'calculate_triangle_area', 'description': 'Calculate the area of a triangle given its base and height.', 'parameters': {'type': 'dict', 'properties': {'base': {'type': 'integer', 'description': 'The base of the triangle.'}, 'height': {'type': 'integer', 'description': 'The height of the triangle.'}, 'unit': {'type': 'string', 'description': \"The unit of measure (defaults to 'units' if not specified)\"}}, 'required': ['base', 'height']}}</tool>
Find the area of a triangle with a base of 10 units and height of 5 units.
Hymba: { "toolcall": { "type": "calculate_triangle_area", "arguments": { "base": 10, "height": 5, "unit": "units" } } }