Technology Innovation Institute

Welcome Falcon Mamba:
First Strong Attention-free 7B Language model

Jingwei Zuo

Lead Researcher @**Falcon LLM team, Technology Innovation Institute**

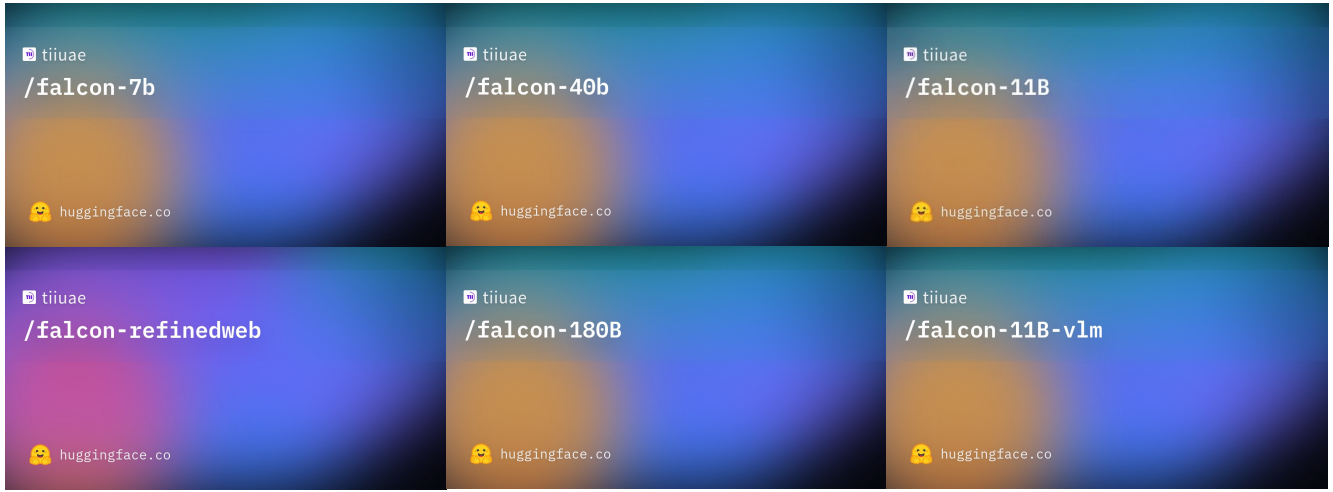
19/12/2024, TotalEnergies Digital Factory

Technology Innovation Institute

Outline

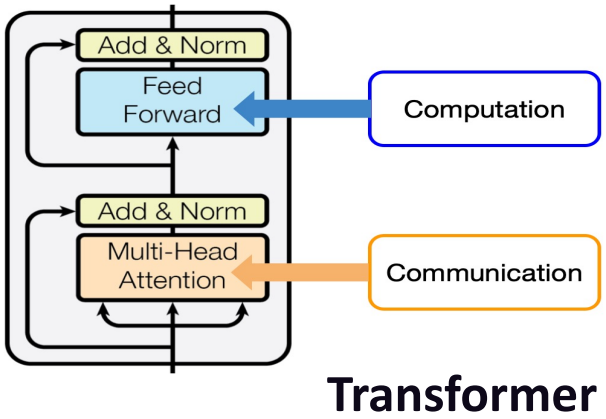
- Context
- Falcon Mamba 7B
- How to use Falcon Mamba 7B?
- More…

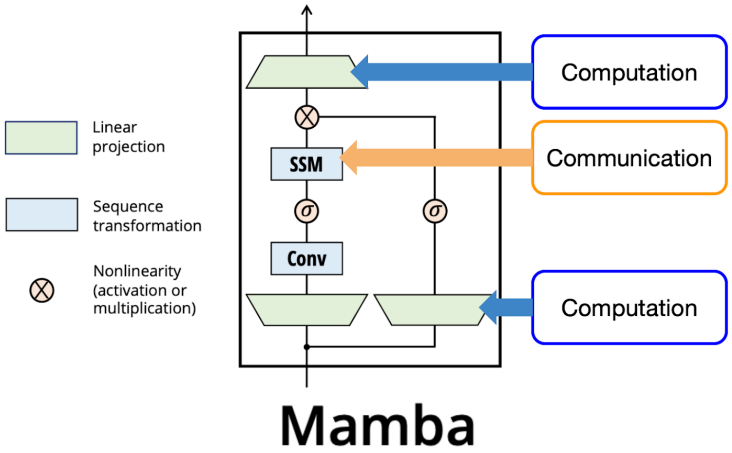
Falcon Model Series



Foundation Model Backbones (e.g., LLMs)*

- **Communication** *between* tokens
- **Computation** *within* a token





*https://www.kolaayonrinde.com/blog/2024/02/11/mamba.html

From State Space Models (SSMs) to Mamba

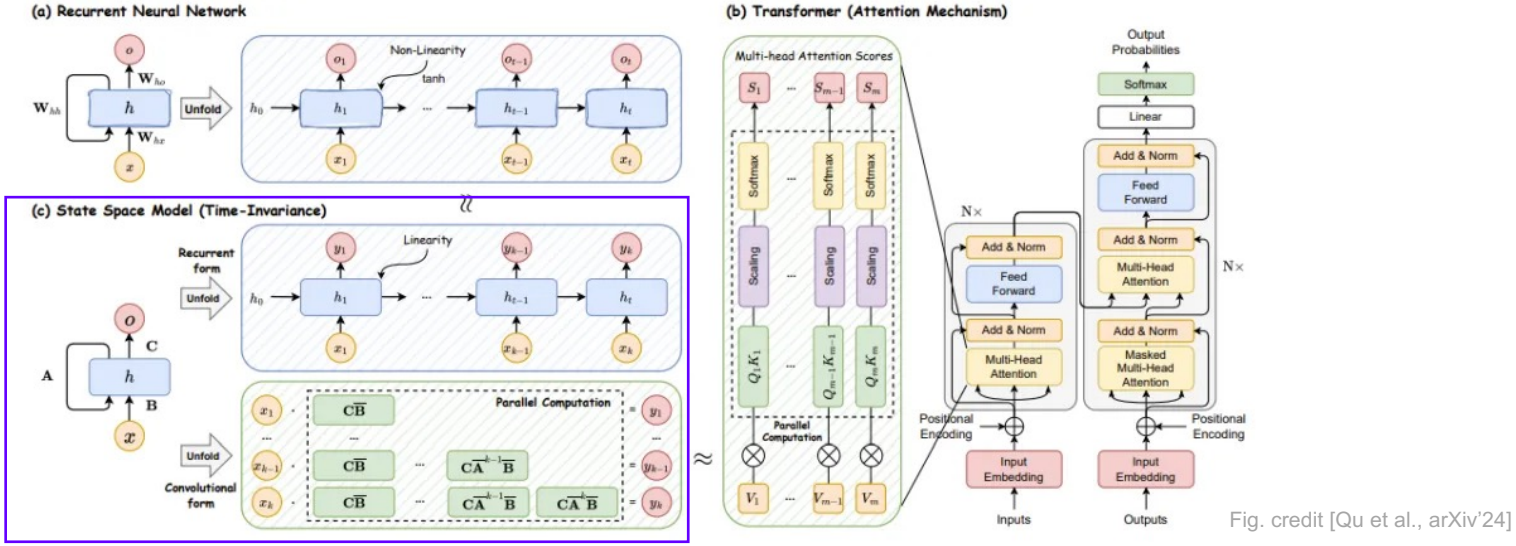
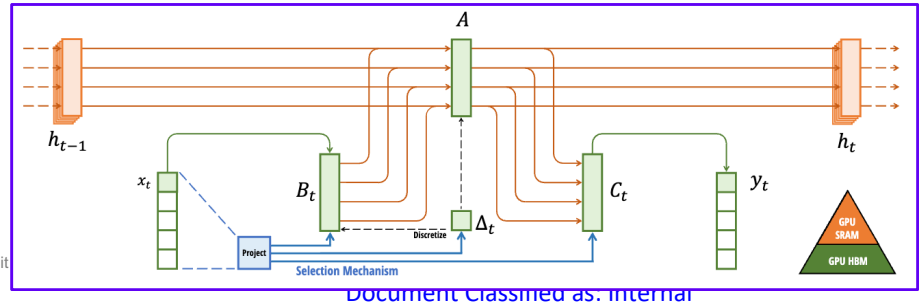


Fig. credit [Qu et al., arXiv'24]

Selective State Space Models
with hardware-aware state expansions (Mamba)



$$h_t = \overline{A}h_{t-1} + \overline{B}x_t$$
$$y_t = Ch_t$$

- Select to remember/forget
- Read/write from/to the hidden states

Fig. credit [Gu and Dao, arXiv'23]

Mamba Memorization & Forgetting

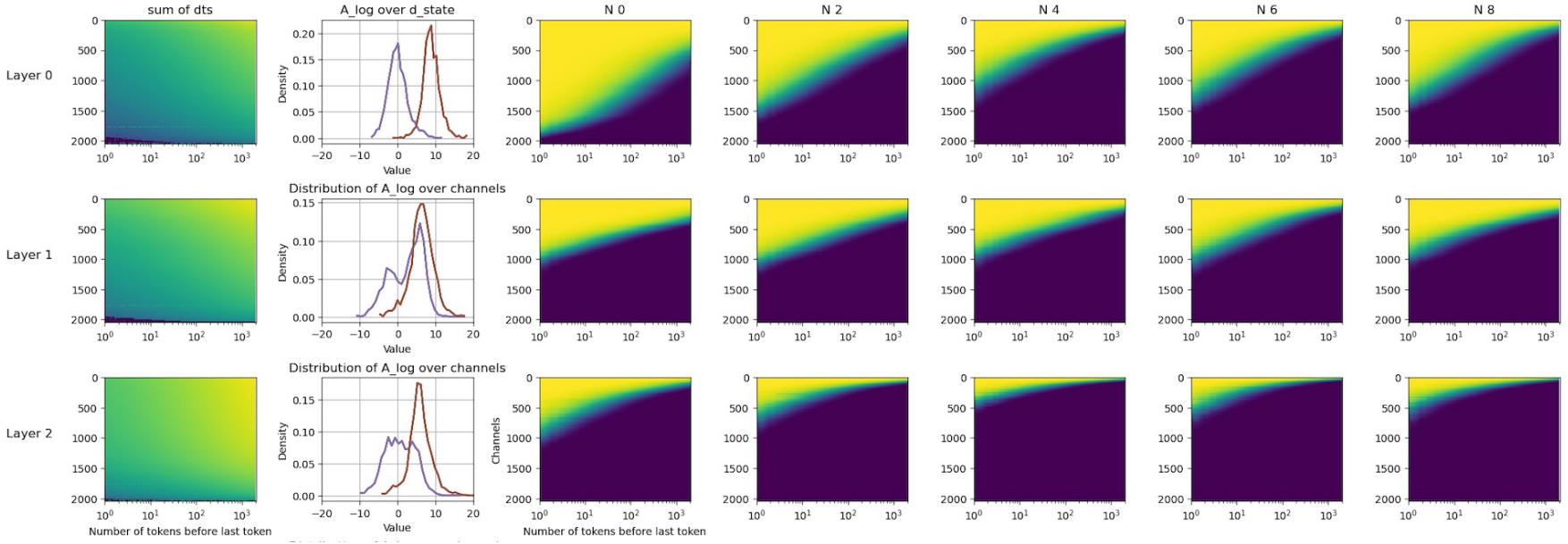


Fig. credit [Gu and Dao, arXiv'23]

Mamba – By-Design Advantages

Transformers

- Quadratic time complexity: $O(n^2)$ regarding context length n
- Memory Constraints: $O(n)$ key-value (KV) cache for each token
- Optimizations: Sparse or local attention (e.g., SWA), CUDA optimizations (e.g., Flash Attention), etc.

Mamba (attention-free)

- Reduced time complexity: $O(n)$ regarding context length n
- Memory Optimization: constant memory cost
- Long-context dependencies: compact state representation

Fig. credit [Gu and Dao, arXiv'23]

Mamba applications in a nutshell

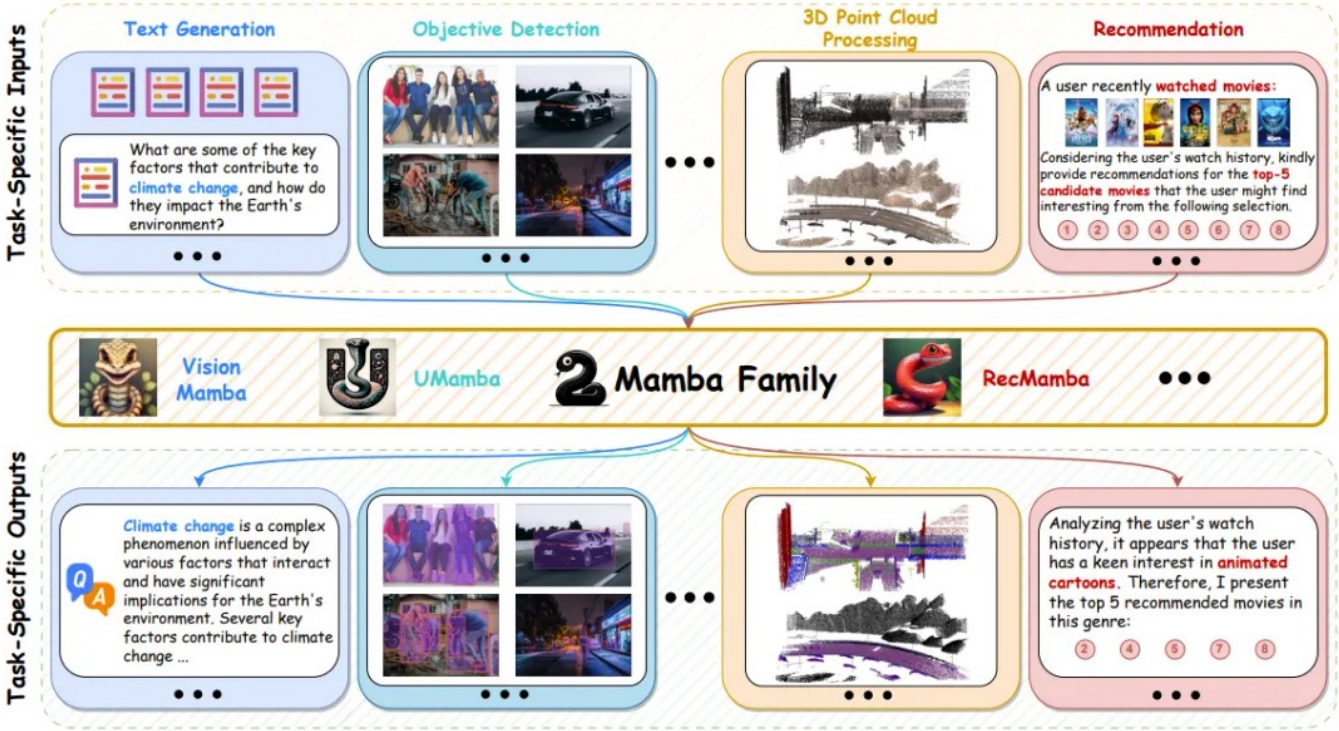


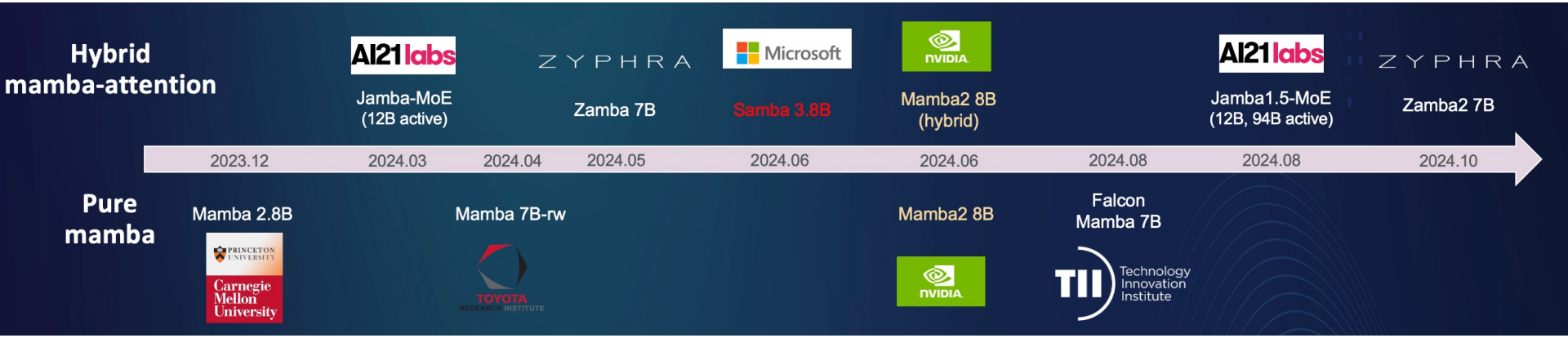
Fig. credit - [Qu et al., arXiv'24]

Mamba LLMs below 2.8B

Model	Token.	Pile ppl ↓	LAMBADA ppl ↓	LAMBADA acc ↑	HellaSwag acc ↑	PIQA acc ↑	Arc-E acc ↑	Arc-C acc ↑	WinoGrande acc ↑	Average acc ↑
Hybrid H3-130M	GPT2	—	89.48	25.77	31.7	64.2	44.4	24.2	50.6	40.1
Pythia-160M	NeoX	29.64	38.10	33.0	30.2	61.4	43.2	24.1	**51.9**	40.6
Mamba-130M	NeoX	**10.56**	**16.07**	**44.3**	**35.3**	**64.5**	**48.0**	**24.3**	**51.9**	**44.7**
Hybrid H3-360M	GPT2	—	12.58	48.0	41.5	68.1	51.4	24.7	54.1	48.0
Pythia-410M	NeoX	9.95	10.84	51.4	40.6	66.9	52.1	24.6	53.8	48.2
Mamba-370M	NeoX	**8.28**	**8.14**	**55.6**	**46.5**	**69.5**	**55.1**	**28.0**	**55.3**	**50.0**
Pythia-1B	NeoX	7.82	7.92	56.1	47.2	70.7	57.0	27.1	53.5	51.9
Mamba-790M	NeoX	**7.33**	**6.02**	**62.7**	**55.1**	**72.1**	**61.2**	**29.5**	**56.1**	**57.1**
GPT-Neo 1.3B	GPT2	—	7.50	57.2	48.9	71.1	56.2	25.9	54.9	52.4
Hybrid H3-1.3B	GPT2	—	11.25	49.6	52.6	71.3	59.2	28.1	56.9	53.0
OPT-1.3B	OPT	—	6.64	58.0	53.7	72.4	56.7	29.6	59.5	55.0
Pythia-1.4B	NeoX	7.51	6.08	61.7	52.1	71.0	60.5	28.5	57.2	55.2
RWKV-1.5B	NeoX	7.70	7.04	56.4	52.5	72.4	60.5	29.4	54.6	54.3
Mamba-1.4B	NeoX	**6.80**	**5.04**	**64.9**	**59.1**	**74.2**	**65.5**	**32.8**	**61.5**	**59.7**
GPT-Neo 2.7B	GPT2	—	5.63	62.2	55.8	72.1	61.1	30.2	57.6	56.5
Hybrid H3-2.7B	GPT2	—	7.92	55.7	59.7	73.3	65.6	32.3	61.4	58.0
OPT-2.7B	OPT	—	5.12	63.6	60.6	74.8	60.8	31.3	61.0	58.7
Pythia-2.8B	NeoX	6.73	5.04	64.7	59.3	74.0	64.1	32.9	59.7	59.1
RWKV-3B	NeoX	7.00	5.24	63.9	59.6	73.7	67.8	33.1	59.6	59.6
Mamba-2.8B	NeoX	**6.22**	**4.23**	**69.2**	**66.1**	**75.2**	**69.7**	**36.3**	**63.5**	**63.3**

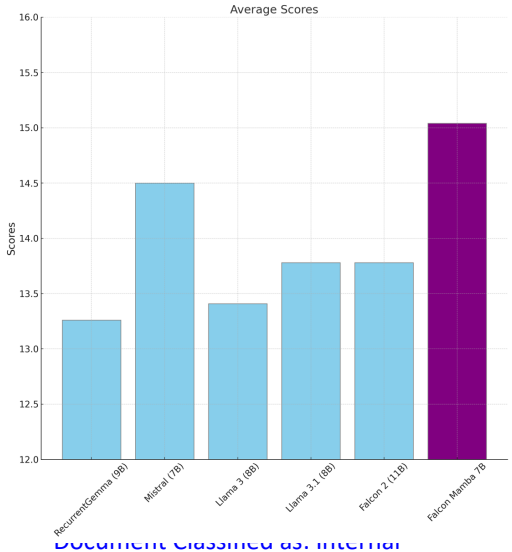
Fig. credit [Gu and Dao, arXiv'23]

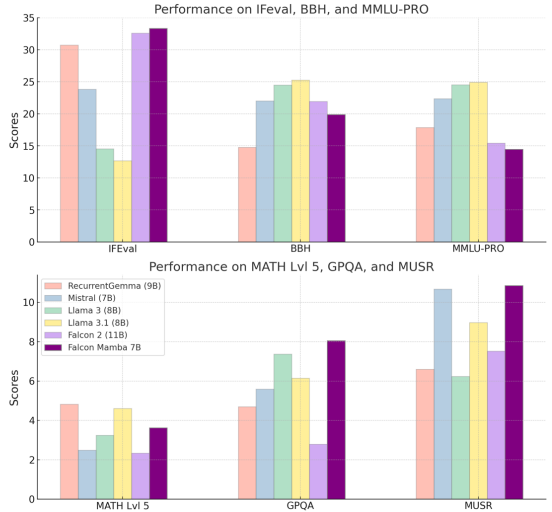
Mamba LLMs beyond 2.8B



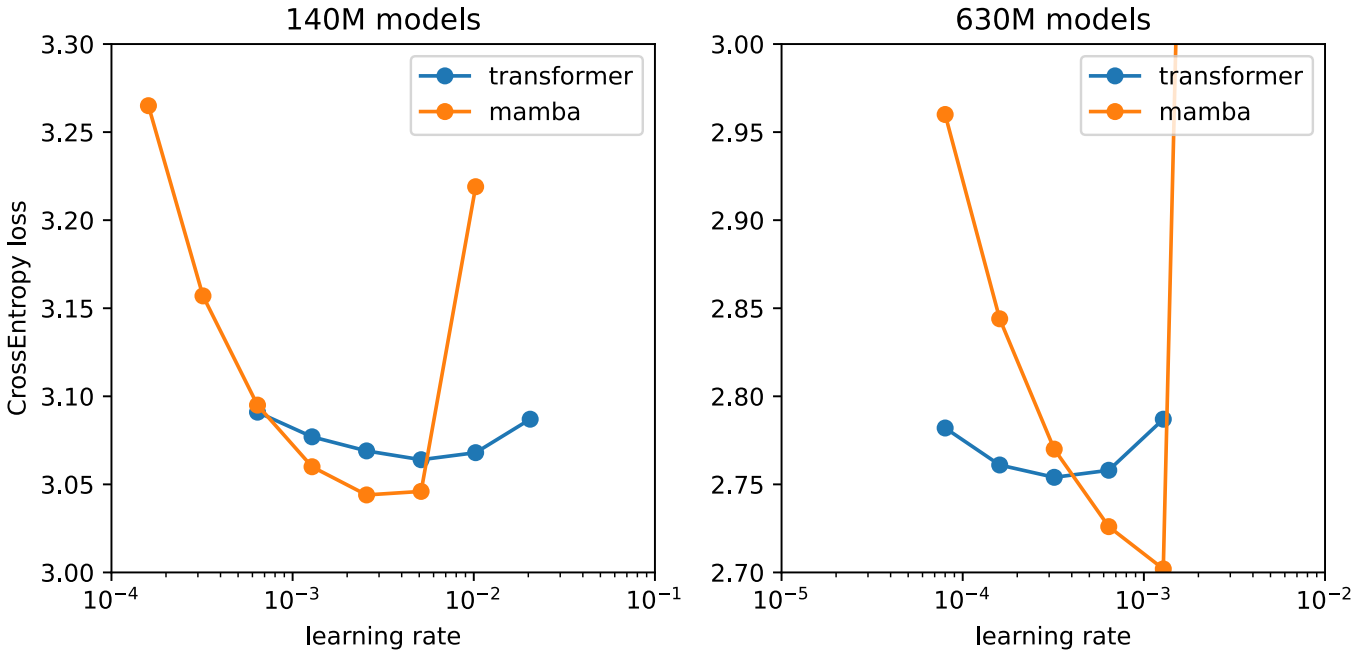
Falcon Mamba 7B

- Best open-source Mamba model at its scale
- Outperforms top Transformer models: Llama 3.1 8B, Mistral 7B, Falcon2 11B
- Supports infinite context length (sequential prefill)

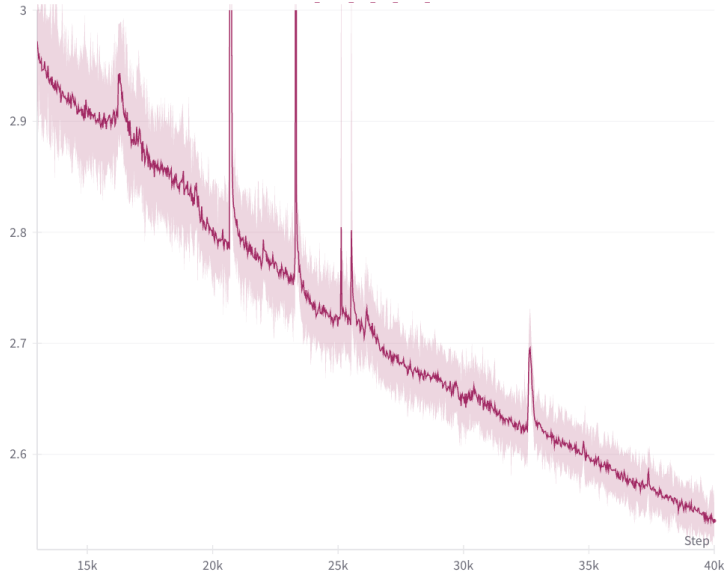


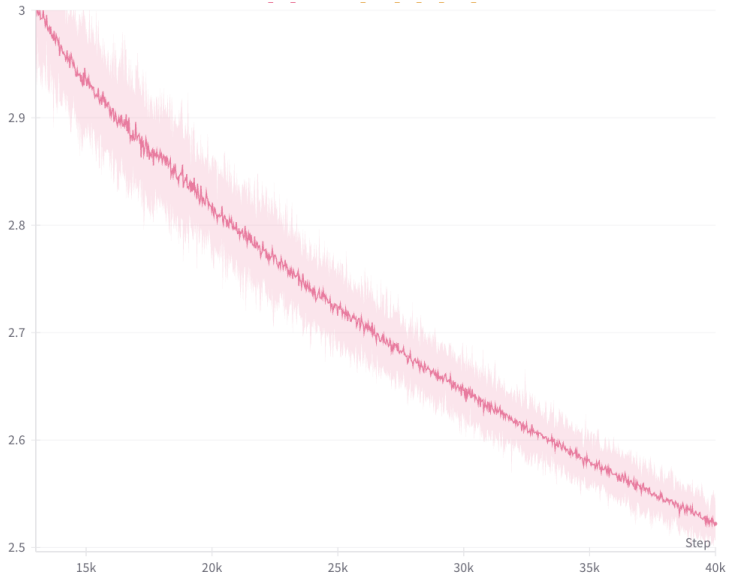


Model's Sensitivity to Learning Rate



Stabilizing Mamba LLM's Pre-Training





Falcon Mamba 7B - Training Details

Training Infrastructure

Gigatron: In-house 5D parallel training framework

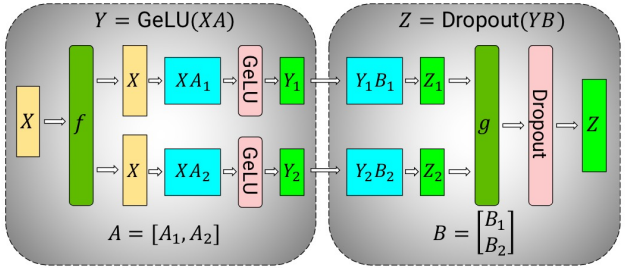
- 5D Parallelism: Data, Tensor, Pipeline, Sequence, Context
- Support both Transformer and Mamba architectures

Trained on 256 H100 80GB GPUs (~ two months)

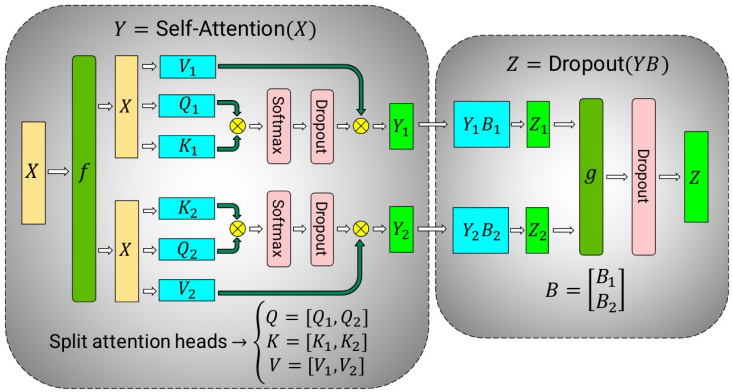
- Only Data Parallelism was applied for Falcon Mamba 7B

Falcon Mamba 7B - Training Details

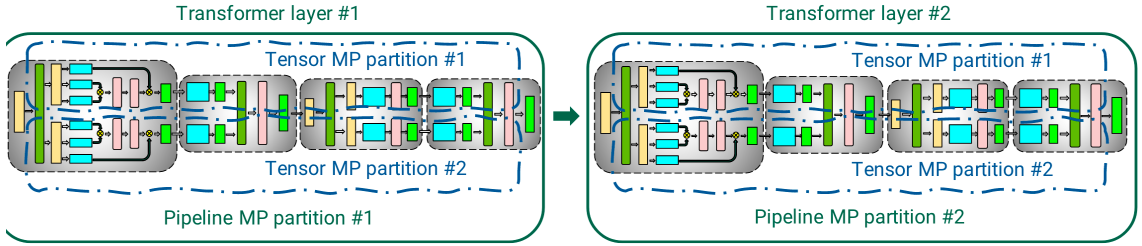
Parallel Training: from Transformer to Mamba

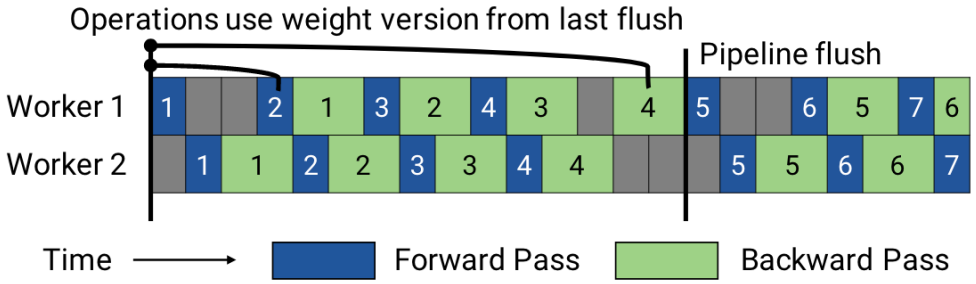


(a) MLP.



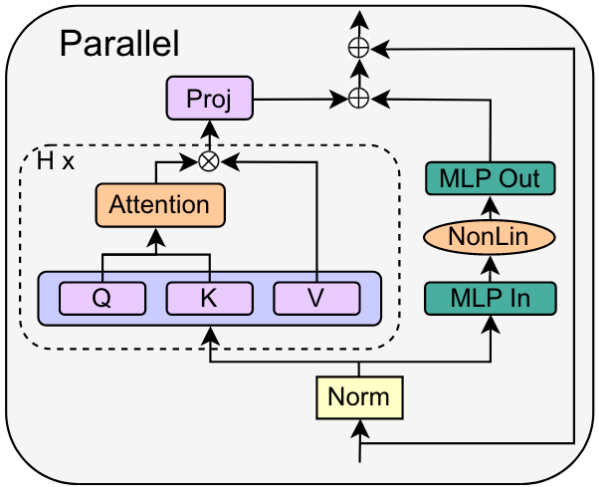
(b) Self-Attention.

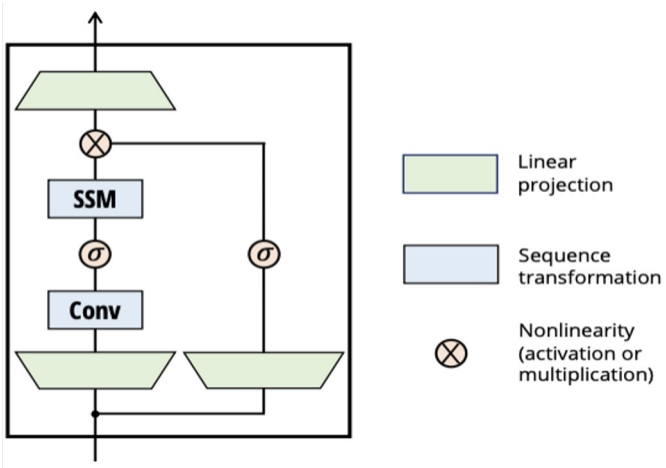




Falcon Mamba 7B - Training Details

Parallel Training: from Transformer to Mamba





Falcon Mamba 7B - Training Details

Training Recipes & Procedure

- WSD (warmup-stable-decay) learning rate schedule
  - Flexible pretraining & continual training
- Batch size ramp-up, batch scaling (during ramp-up), etc.
- RMSNorms within Mamba blocks

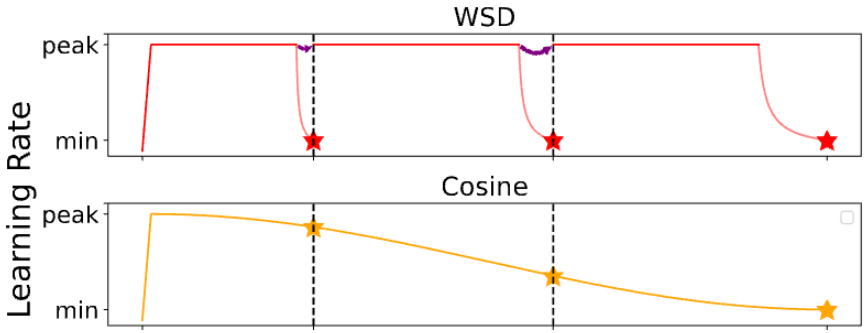


Fig. credit - [Wen et al., arXiv'24]

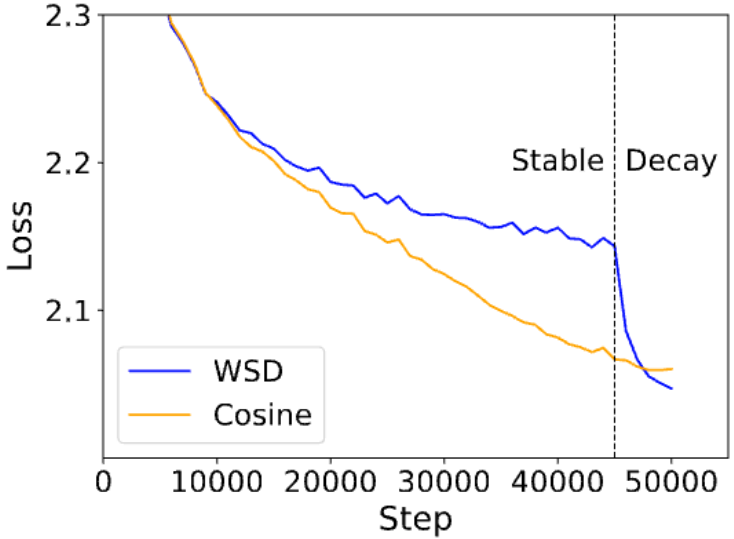
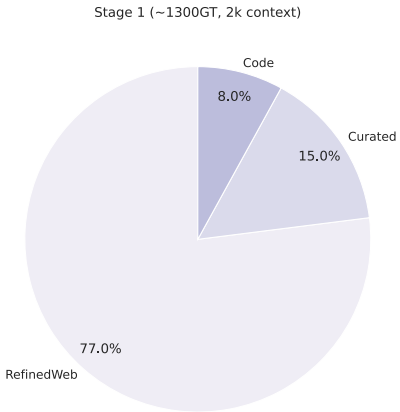


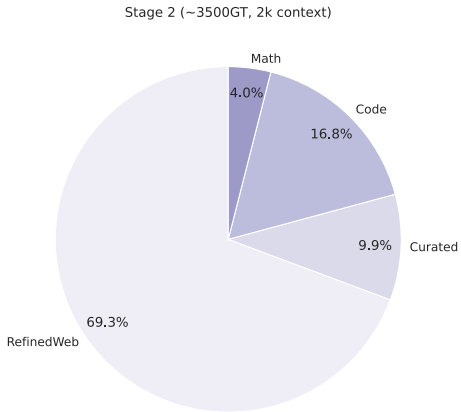
Fig. credit - [Wen et al., arXiv'24]

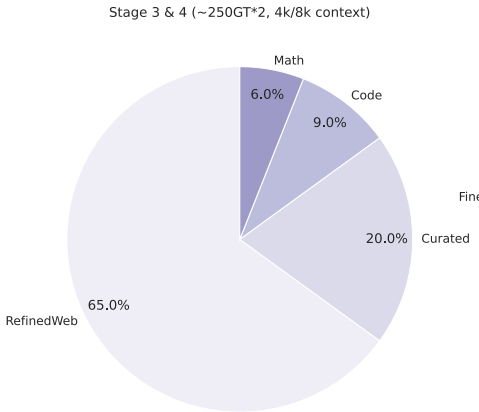
Falcon Mamba 7B - Training Details

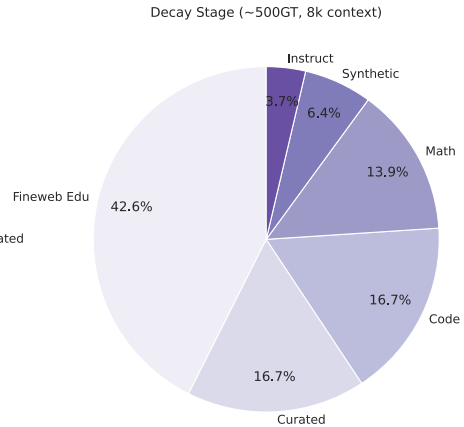
Pretraining Data

- Mostly from Falcon 2 11B
- ~**5,800GT** from multiple resources, e.g., RefinedWeb, code, math, etc.









Falcon Mamba 7B VS SoTA LLMs

Table 2: Model Performance on HF Leaderboard v1 tasks: **bold** (best), underline (second best)

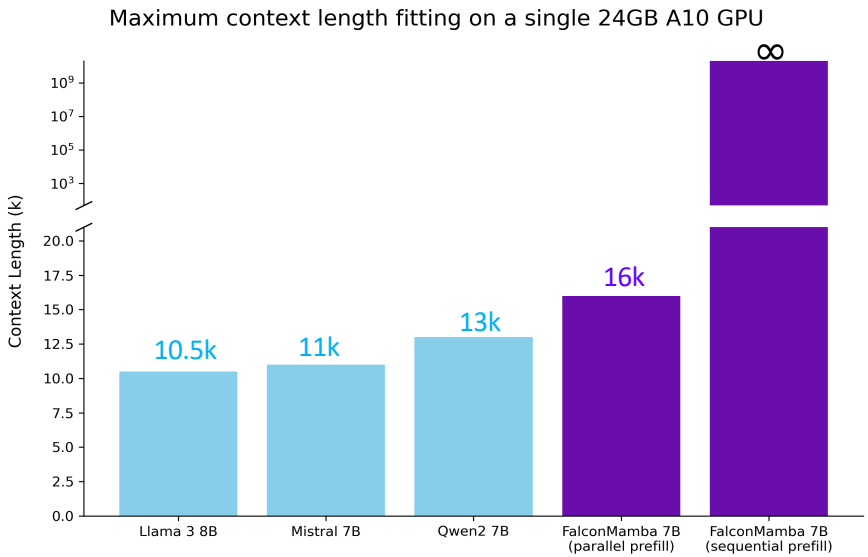
Model Name	ARC-25	HellaSwag-10	MMLU-5	Winogrande-5	TruthfulQA-0	GSM8K-5	Average
RWKV models							
RWKV-v6-Finch-7B*	43.86	75.19	41.69	68.27	42.19	19.64	48.47
RWKV-v6-Finch-14B*	47.44	78.86	52.33	71.27	45.45	38.06	55.57
Transformer models							
Falcon2-11B	59.73	82.91	58.37	78.30	52.56	53.83	**64.28**
Meta-llama-3-8B	60.24	82.23	**66.70**	78.45	42.93	45.19	62.62
Meta-llama-3.1-8B	58.53	82.13	66.43	74.35	44.29	47.92	62.28
Mistral-7B-v0.1	59.98	**83.31**	64.16	78.37	42.15	37.83	60.97
Mistral-Nemo-Base-2407 (12B)	57.94	82.82	64.43	73.72	49.14	**55.27**	63.89
Gemma-7B	61.09	82.20	64.56	79.01	44.79	50.87	63.75
Hybrid SSM-attention models							
RecurrentGemma-9b**	52.00	80.40	60.50	73.60	38.60	42.60	57.95
Zyphra/Zamba-7B-v1*	56.14	82.23	58.11	**79.87**	52.88	30.78	60.00
Pure SSM models							
TRI-ML/mamba-7b-rw*	51.25	80.85	33.41	71.11	32.08	4.70	45.52
FalconMamba-7B (pre-decay)*	49.23	80.25	57.27	70.88	37.28	21.83	52.79
FalconMamba-7B*	**62.03**	80.82	62.11	73.64	**53.42**	52.54	64.09

Table 3: Model Performance on HF Leaderboard v2: **bold** (best), underline (second best)

Model Name	IFEval-0	BBH-3	Math-Lvl5-4	GPQA-0	MuSR-0	MMLU-PRO-5	Average
RWKV models							
RWKV-v6-Finch-7B	27.65	9.04	1.11	2.81	2.25	5.85	8.12
RWKV-v6-Finch-14B	29.81	12.89	1.13	5.01	3.16	11.3	10.55
Transformer models							
Falcon2-11B	32.61	21.94	2.34	2.80	7.53	15.44	13.78
Meta-llama-3-8B	14.55	24.50	3.25	7.38	6.24	24.55	13.41
Meta-llama-3.1-8B	12.70	25.29	4.61	6.15	8.98	24.95	13.78
Mistral-7B-v0.1	23.86	22.02	2.49	5.59	10.68	22.36	14.50
Mistral-Nemo-Base-2407 (12B)	16.83	**29.37**	4.98	5.82	6.52	**27.46**	15.08
Gemma-7B	26.59	21.12	**6.42**	4.92	**10.98**	21.64	**15.28**
Hybrid SSM-attention models							
RecurrentGemma-9b	30.76	14.80	4.83	4.70	6.60	17.88	13.20
Zyphra/Zamba-7B-v1*	24.06	21.12	3.32	3.03	7.74	16.02	12.55
Pure SSM models							
TRI-ML/mamba-7b-rw*	22.46	6.71	0.45	1.12	5.51	1.69	6.25
FalconMamba-7B (pre-decay)*	24.05	11.01	1.71	3.05	8.68	8.59	9.52
FalconMamba-7B	**33.36**	19.88	3.63	**8.05**	10.86	14.47	15.04

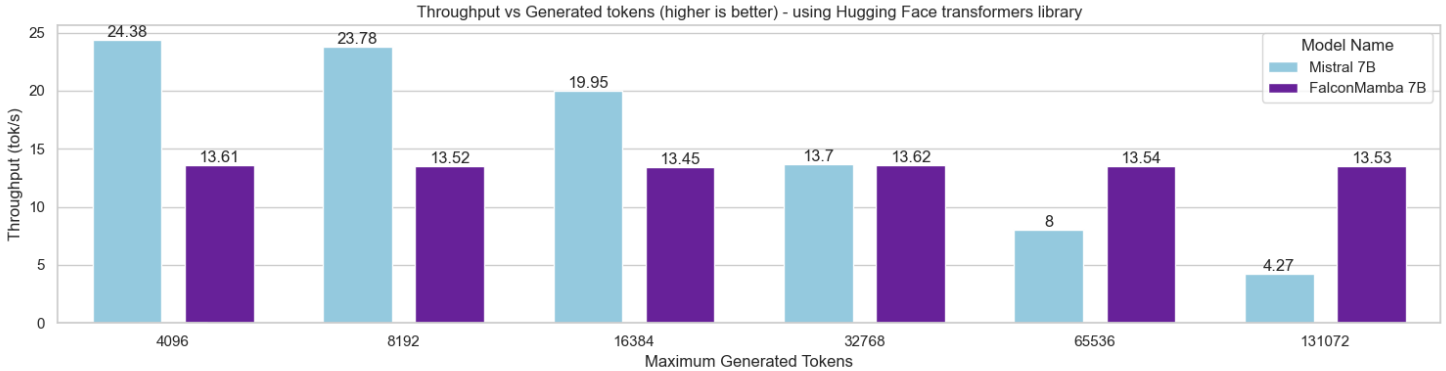
Falcon Mamba 7B VS SoTA LLMs

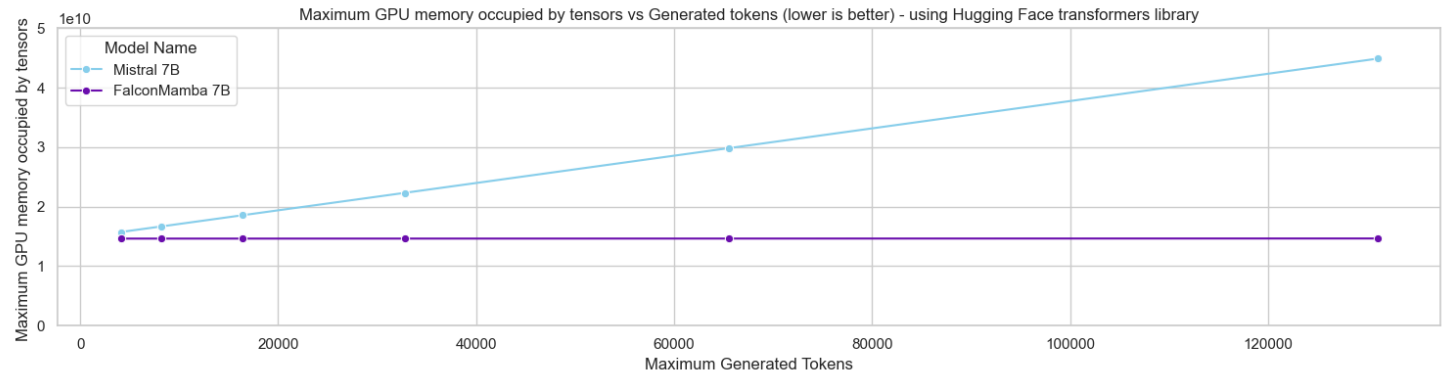
- Infinite context length (sequential prefill)
- ~16K context length (parallel prefill), outperforming Transformer models



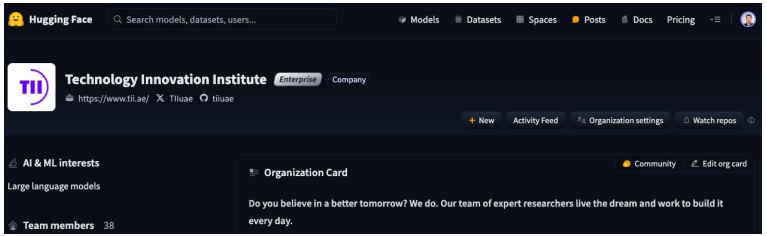
Falcon Mamba 7B VS SoTA LLMs

- Constant throughput regardless the generation length
- Constant memory usage and lower memory cost than Mistral 7B





How to use Falcon Mamba 7B



HuggingFace Ecosystem

Direct use with HF Transformers

- Multiple precisions (quantized & non quantized)
- Multiple platforms (CPU & GPUs)

Fine tuning

- SFTrainer, PEFT, etc.

```python
# pip install accelerate
from transformers import AutoTokenizer, AutoModelForCausalLM

tokenizer = AutoTokenizer.from_pretrained("tiiuae/falcon-mamba-7b")
model = AutoModelForCausalLM.from_pretrained("tiiuae/falcon-mamba-7b", device_map

input_text = "Question: How many hours in one day? Answer: "
input_ids = tokenizer(input_text, return_tensors="pt").input_ids.to("cuda")

outputs = model.generate(input_ids)
print(tokenizer.decode(outputs[0]))
```

Falcon Mamba 7B Chat model (Instruct)

Online, full precision

- HuggingFace Playground

Offline, multiple precisions

- Llama.cpp, LM Studio, etc.

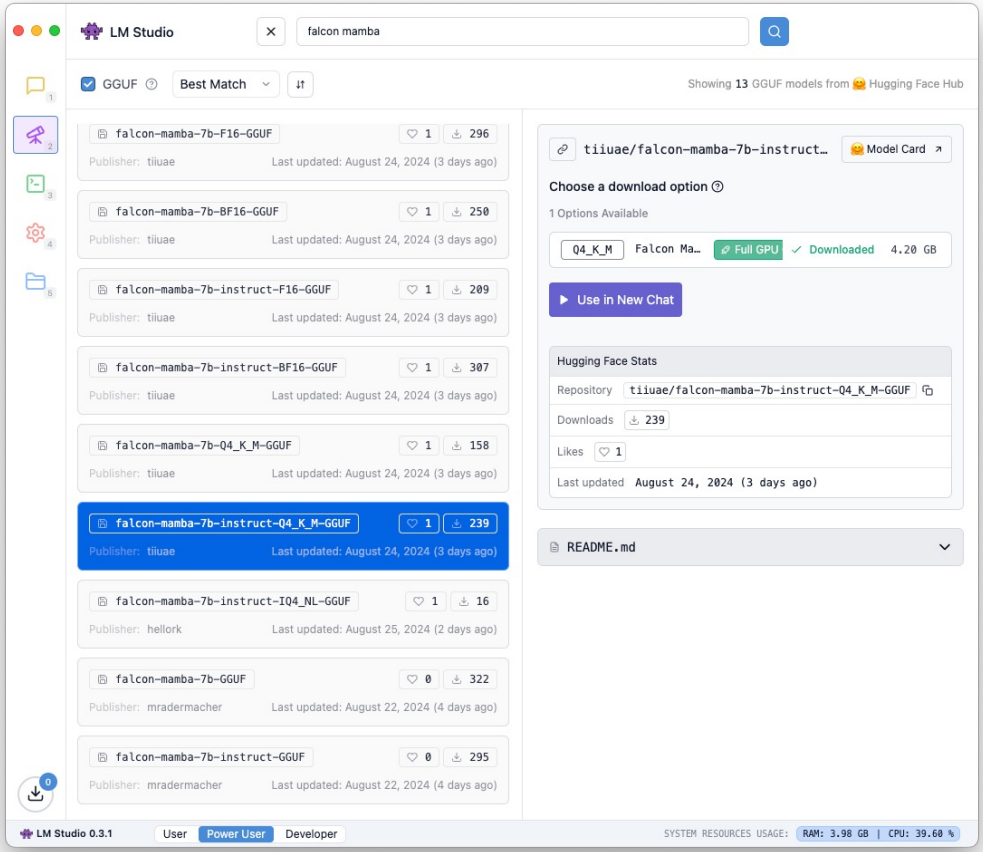
Hugging Face

LLaMA C++

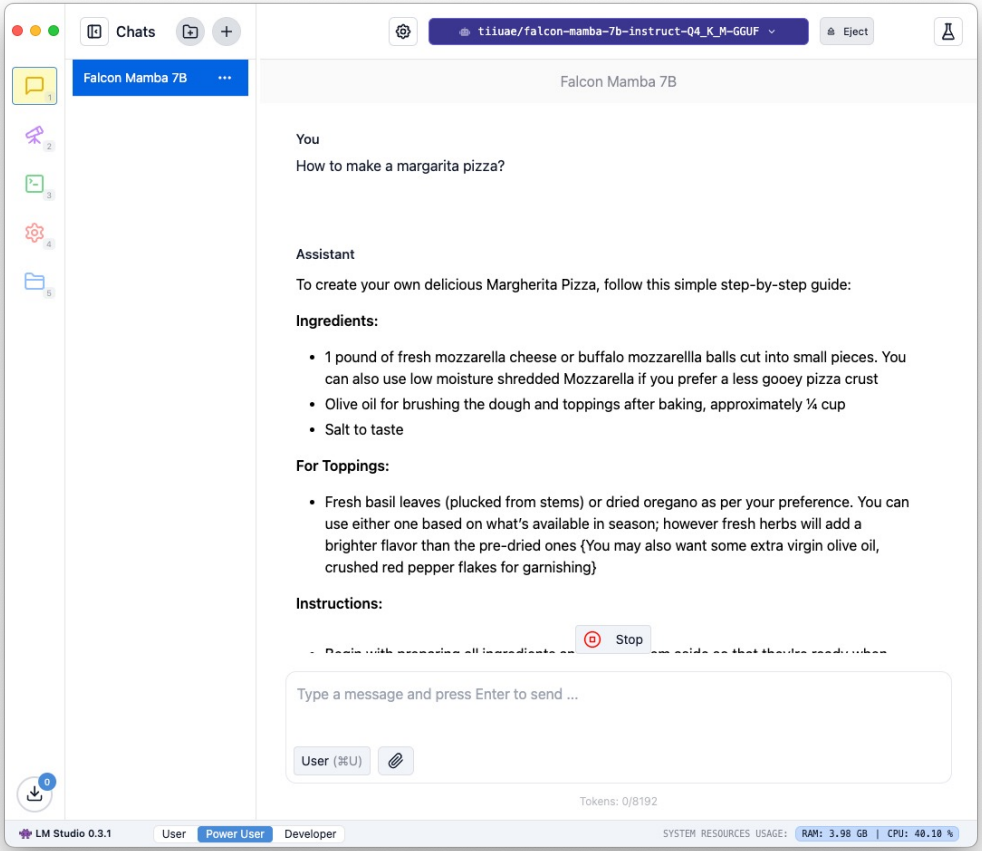
LM Studio

MLX

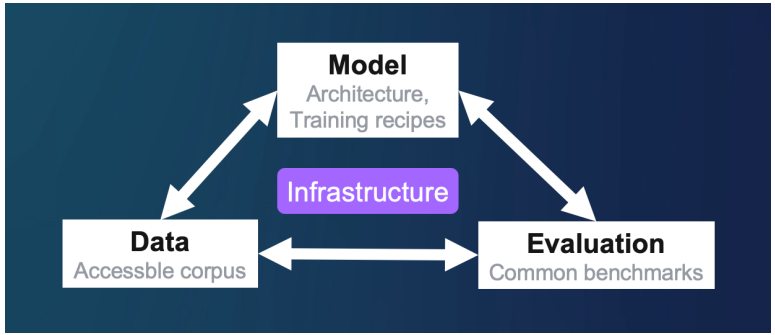
How to use Falcon Mamba 7B - LM Studio



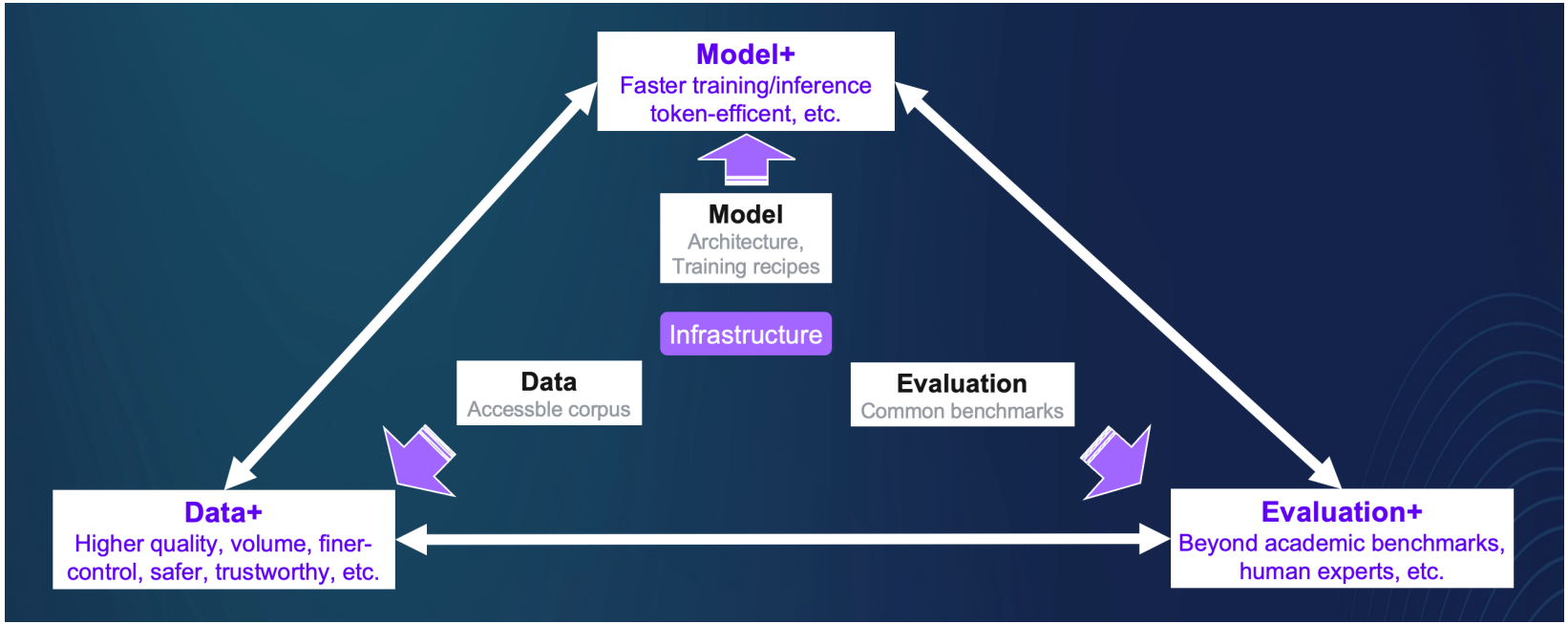
How to use Falcon Mamba 7B - LM Studio



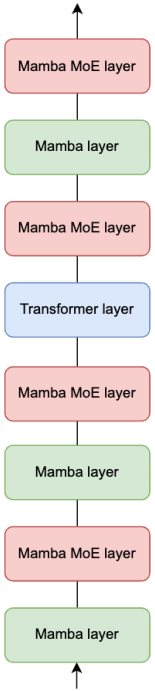
LLMs - Model & Data & Evaluation



LLMs - Model & Data & Evaluation

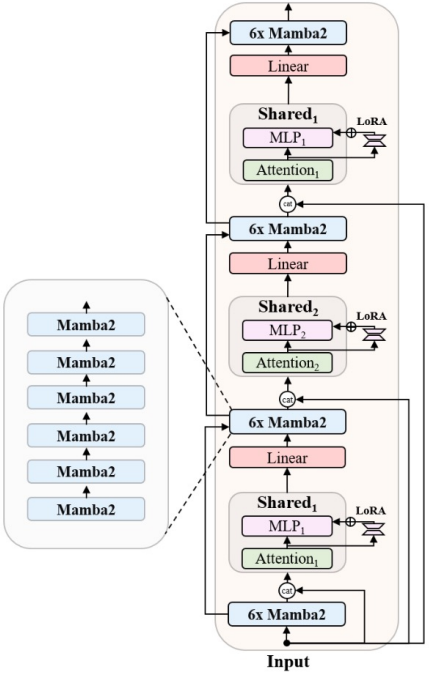


Hybrid Model - The Future Design of LLMs?



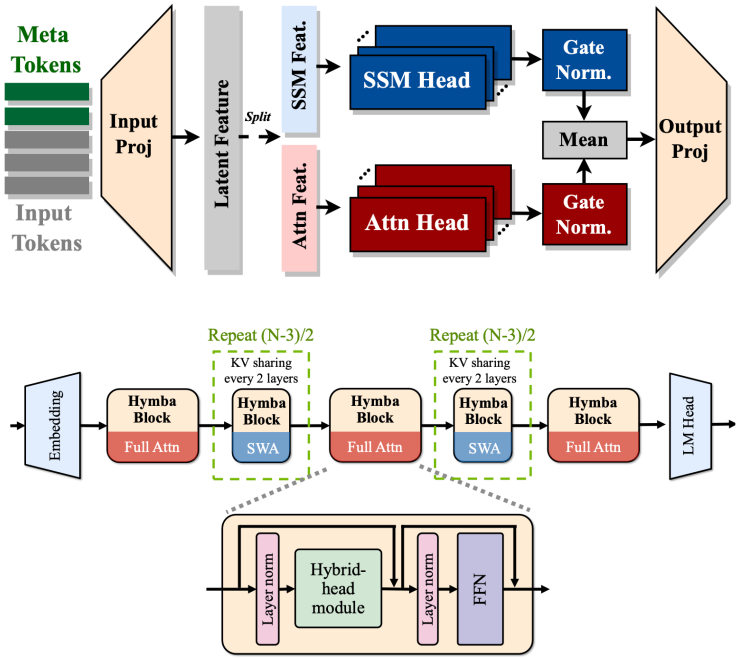
Jamba-MoE 52B (12B)

Fig. credit - [Lieber et al., arXiv'24]



Zamba 2 7B

Fig. credit - Zyphra



Hymba 1.5B

Fig. credit - [Dong et al., arXiv'24]

Falcon 3 Family of Open Models

Welcome to the Falcon 3 Family of Open Models!

Upvoted 84 +72

Published December 17, 2024

Update on GitHub

Abdalgader Abdalgader Abubaker tiiuae
abdulmuneertii Abdul Muneer tiiuae
AdamWma Adam tiiuae
AbdulazizAlshamsi Abdulaziz Alshamsi tiiuae
alice-tii Alice Pagnoux tiiuae
Almansoorialikhalifa Ali Khalifa Almansoori tiiuae
amztheory Ahmed Alzubaidi tiiuae
billelmokeddem Billel Mokeddem tiiuae
Dhia-GB Dhia Garbaya tiiuae
DhiyaEddine Rhaiem tiiuae
dunghuynh ngoc dung huynh guest
fedyanin Kirill tiiuae
gcamp Giulia Campesan tiiuae
Gkunsch Guillaume Kunsch tiiuae
gokulkarthik Gokul Karthik tiiuae
griffintaur Ankit Singh tiiuae
HakimHacid Hakim Hacid tiiuae
HamzaYous89 Hamza Yous tiiuae
hangzou Hang Zou tiiuae
ibrahim-khadraoui-TII ibrahim khadraoui tiiuae
ifarhat1993 BrahimFarhat tiiuae
Iheb-Chaabane Iheb Chaabane tiiuae
jaadariF Fedi Jaadari tiiuae
JingweiZuo Jingwei Zuo tiiuae
karnakar Karna tiiuae
kasper-piskorski Kasper Piskorski tiiuae
IChahed Ilyas Chahed tiiuae
lkhphuc Phúc Lê Khắc guest
LeenAlQadi Leen AlQadi guest
Ludovick Lepauloux tiiuae
melaseddik Mohamed El Amine Seddik tiiuae
mersinvald Mike Lubinets tiiuae
Mughaira Mughaira tiiuae
puneeshkhanna Puneesh Khanna tiiuae
qiyang-zhao Qiyang Zhao tiiuae
rcojocaru Ruxandra Cojocaru tiiuae
RedaAlami Reda alami tiiuae
rishabh-saraf Rishabh Saraf tiiuae
SanathNarayan Sanath Narayan tiiuae
shihux Shi Hu tiiuae
slimfrikha-tii Slim Frikha tiiuae
wamreyaz wamiq para tiiuae
wdevazelhes William de Vazelhes tiiuae
yellowvm Maksim Velikanov tiiuae
ybelkada Younes Belkada tiiuae
yasserTII Yasser Dahou tiiuae

Technology Innovation Institute

Thank you!