Technology Innovation Institute

Falcon-H1: A Family of Hybrid-Head Language Models Redefining Efficiency and Performance

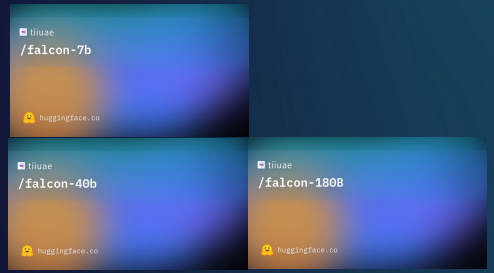
Jingwei Zuo
Lead Researcher @Falcon LLM team
Artificial Intelligence Research Center
Technology Innovation Institute

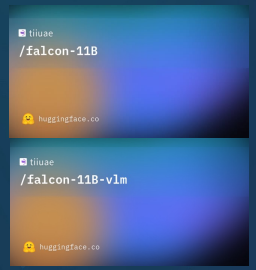
ASAP Seminar

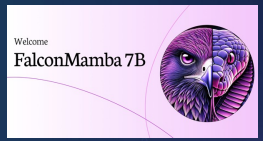
Sep.05, 2025
Online

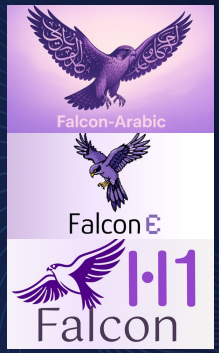
Falcon Model Series

- Open-source commitment
- 2023: Falcon 7B, 40B, 180B
- 2024.05: Falcon2-11B
- 2024.08: Falcon Mamba 7B
- 2024.12: Falcon 3
- 2025.05: Falcon-H1, Falcon-Edge (BitNet), Falcon-Arabic









2023.5 2023.09 2024.05 2024.08 2024.12 2025.05

Outline

1. Context
2. Architecture design and decision
3. Training details
4. Evaluation (highlights and intuitions)
5. General discussions

Context

2024.08

- Mamba1 design
- Untouched model design except the RMSNorm for training stability
- 5.5T token training budget
- Strong base SSM to surpass leading Transformer models such as LLama3.1, Mistral 7B, Gemma 9B.
- 8K context

2024.12

- Mamba1 design
- Untouched model design except the RMSNorm for training stability
- 1.5T token training budget for CPT on Falcon Mamba
- Competitive with hybrid models (Zamba2-7B, Jamba 1.5-mini)
- 32k context, enhanced STEM capabilities

2025.05

- Parallel hybrid attention-Mamba2
- Revisited every aspect of model design, data and training strategy.
- Outperforms and rivals SoTA LLMs (Qwen3, Gemma3, LLama3/4, Mistral3.1) at each scale (non-reasoning).
- Strong open-source community and industry engagement
- Fully production-ready: 0.5B to 34B

From Falcon Mamba, Falcon3 mamba to Falcon-H1

Hybrid Design/Models – Glance on the literature

- 2024.3 ○ Jamba (52B-A12B, 398B-A94B)
- 2024.5 ○ Zamba (7B)
- 2024.6 ○ Samba (3.8B)
- 2024.11 ○ Hymba (1.5B)
- 2024.12 ○ MiniMax-Text-01 (456B-A46B)
- 2025.3 ○ Hunyuan-TurboS
- 2025.4 ○ Nemotron-H (8B, 47B, 56B)
- 2025.5 ○ Falcon-H1 (0.5B, 1.5B, 1.5B-Deep, 3B, 7B, 34B)
- ○ Granite 4.0-tiny-preview (7B-A1B)
- 2025.7 ○ Phi-4-mini-flash-reasoning (3.8B)
- 2025.8 ○ Nemotron-Nano-v2 (9B, 12B), Jet-Nemotron (2B, 4B)

(Non-exhaustive list) Non open weight

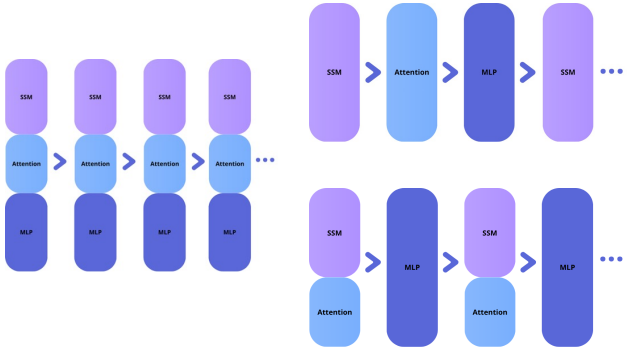
Hybrid Design - Channel Allocation

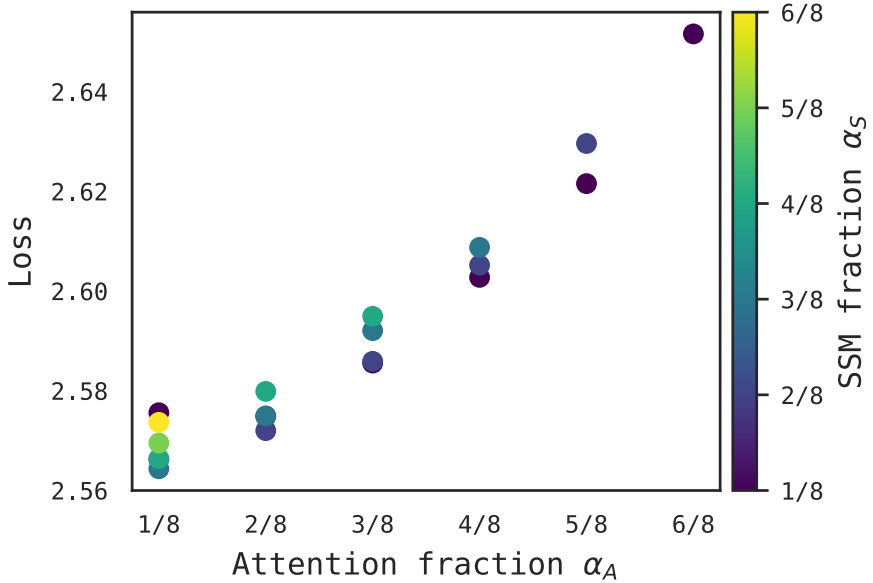
- Fully parallel (SAM), semi-parallel (SA_M), fully sequential (S_A_M)

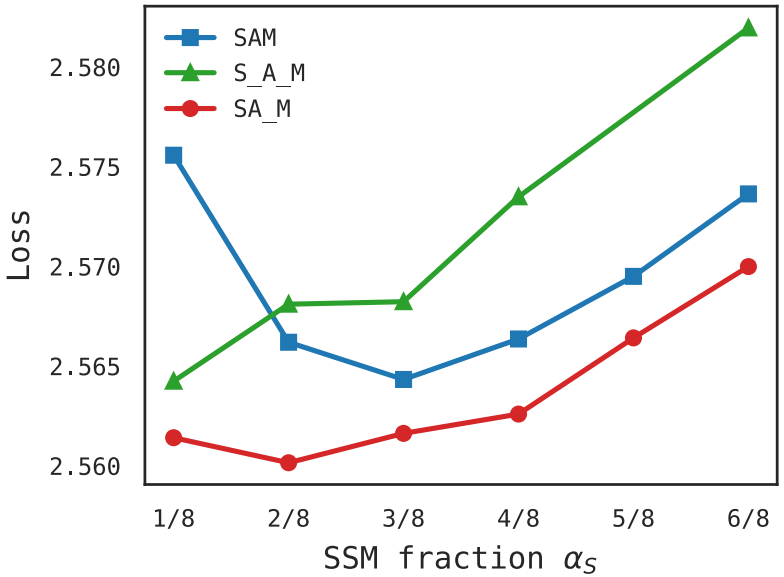
$$\texttt{SAM:}\quad \mathbf{r}_{l+1} = \mathbf{r}_l + \mathcal{F}_l^{\mathrm{MLP}}(\mathcal{N}_l(\mathbf{r}_l)) + \mathcal{F}_l^{\mathrm{attn}}(\mathcal{N}_l(\mathbf{r}_l)) + \mathcal{F}_l^{\mathrm{SSM}}(\mathcal{N}_l(\mathbf{r}_l))$$

$$\texttt{SA\_M:}\quad \mathbf{r}_{l+1} = \mathbf{r}'_l + \mathcal{F}_l^{\mathrm{MLP}}(\mathcal{N}'_l(\mathbf{r}'_l)), \quad \mathbf{r}'_l = \mathbf{r}_l + \mathcal{F}_l^{\mathrm{attn}}(\mathcal{N}_l(\mathbf{r}_l)) + \mathcal{F}_l^{\mathrm{SSM}}(\mathcal{N}_l(\mathbf{r}_l))$$

$$\texttt{S\_A\_M:}\quad \mathbf{r}_{l+1} = \mathbf{r}''_l + \mathcal{F}_l^{\mathrm{MLP}}(\mathcal{N}''_l(\mathbf{r}''_l)), \quad \mathbf{r}''_l = \mathbf{r}'_l + \mathcal{F}_l^{\mathrm{attn}}(\mathcal{N}'_l(\mathbf{r}'_l)), \quad \mathbf{r}'_l = \mathbf{r}_l + \mathcal{F}_l^{\mathrm{SSM}}(\mathcal{N}_l(\mathbf{r}_l))$$







SA_M block, 2:1:5 ratio

Fixed $\alpha_A = \frac{1}{8}$

SSM Module - State Size VS Num Groups

- Revisiting every aspect of the architecture: state size - groups – head dimensions – convolution – chunk size ...
- 2 metrics: Accuracy and efficiency

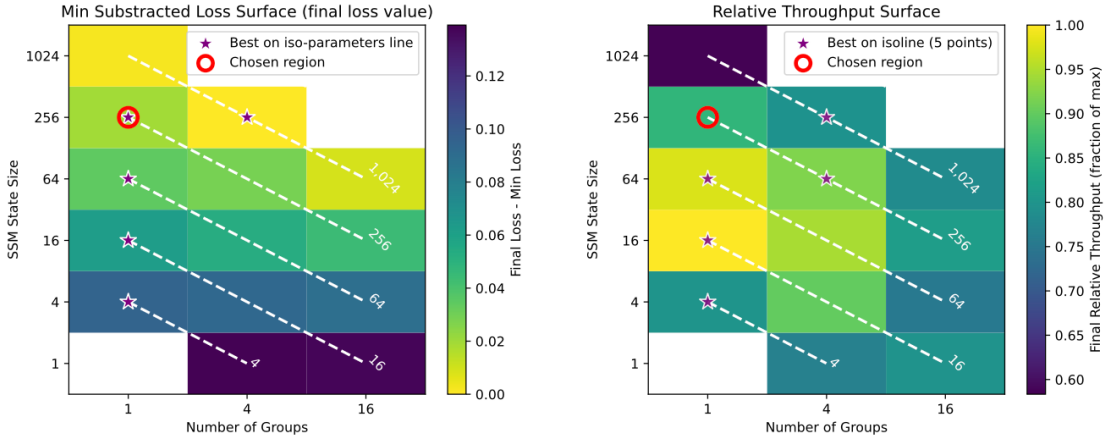
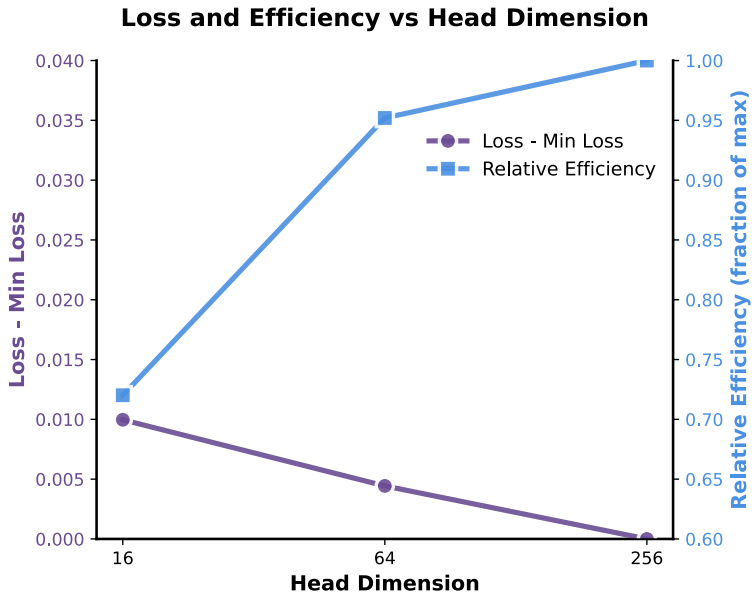
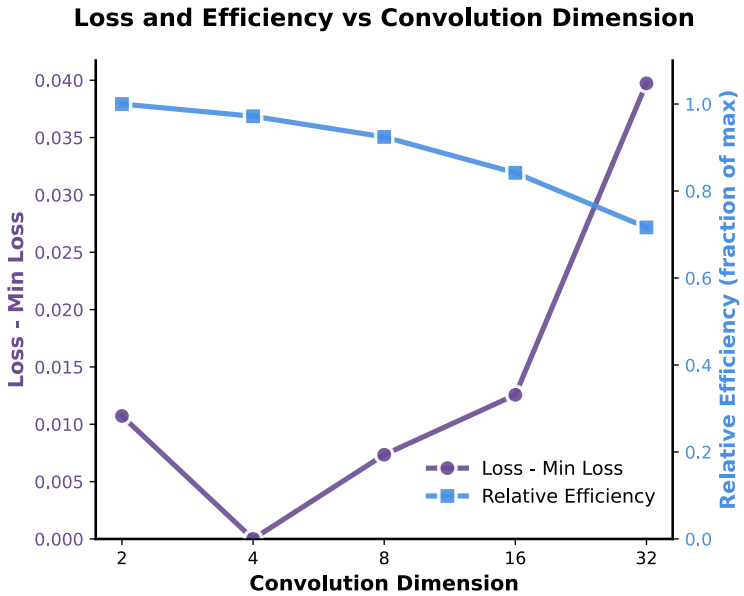


Figure 3: Hyperparameter optimization landscapes for SSM number of groups and state dimension size. **(a)** Loss surface showing performance relative to global minimum across number of groups and d_state size. **(b)** Relative throughput surface as fraction of maximum performance. Dashed lines indicate iso-parameter curves (ng × ds = constant), implying constant total parameter count. Red stars mark optimal configurations for each computational budget, revealing distinct trade-offs between model quality and efficiency.

SSM Module - Head Dim, Conv Dim

- Revisiting every aspect of the architecture: state size - groups – head dimensions – convolution – chunk size ...
- 2 metrics: Accuracy and efficiency





SSM Module – Hidden State Resetting

- Long context data (packing): semantic contamination between unrelated contexts
- Attention: cross-document masking (block-diagonal mask)
- SSM (recurrent models)?

$$\mathbf{h}_{t+1} = \overline{A}_t\, \mathbf{h}_t + \mathbf{B}_t dt_t x_t, \qquad y_t = \mathbf{C}_t^{\top}\, \mathbf{h}_t + D x_t.$$

$$\bar{A}_i = \exp\!\left[-e^{A_{\log}}\, \tilde{dt}_i + r_i \cdot (-80)\right] \approx \begin{cases} \mathbf{0}, & r_i = 1, \\ \bar{A}, & r_i = 0. \end{cases}$$

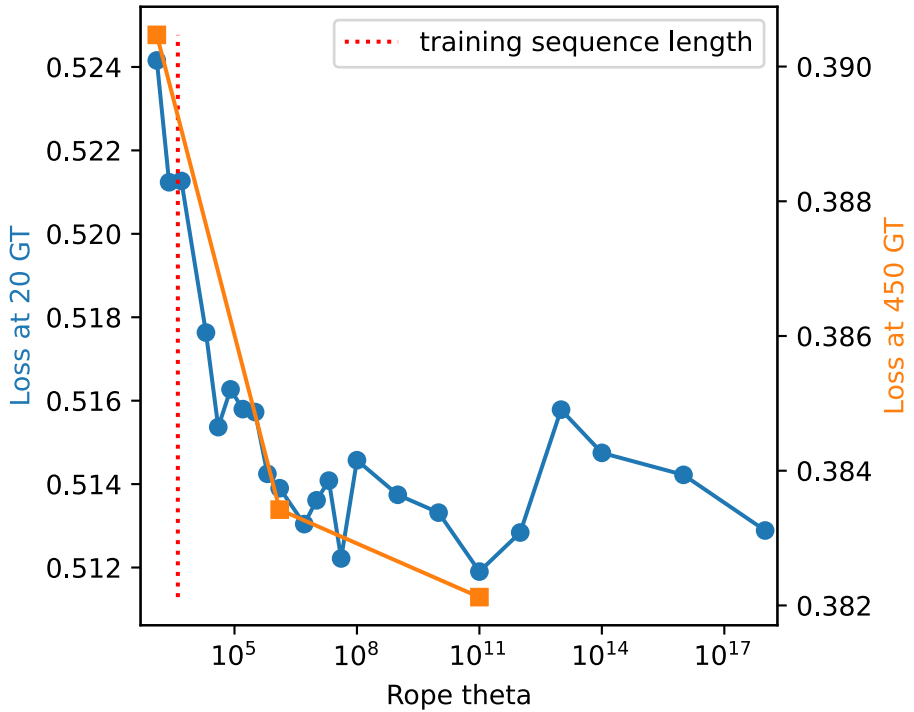
At a boundary $\mathbf{h}_{t+1} = \mathbf{0} \cdot \mathbf{h}_t + \bar{B}\, x_t = \bar{B}\, x_t$

* exp(-80) ≈ 1e-35 ≫ 1e-45
(FP16/BF16 underflow threshold)

Attention Module – RoPE theta

1e11 – optimal value ≫ 1e6 or 1e7 (commonly used)

Open questions: e.g., large RoPE theta – will it work for sequential Hybrid or transformer models?



Beyond hybrid design: Width–Depth Trade-offs

Joint sweeps over hidden width and depth, scaling LR inversely with width (µP scaling)

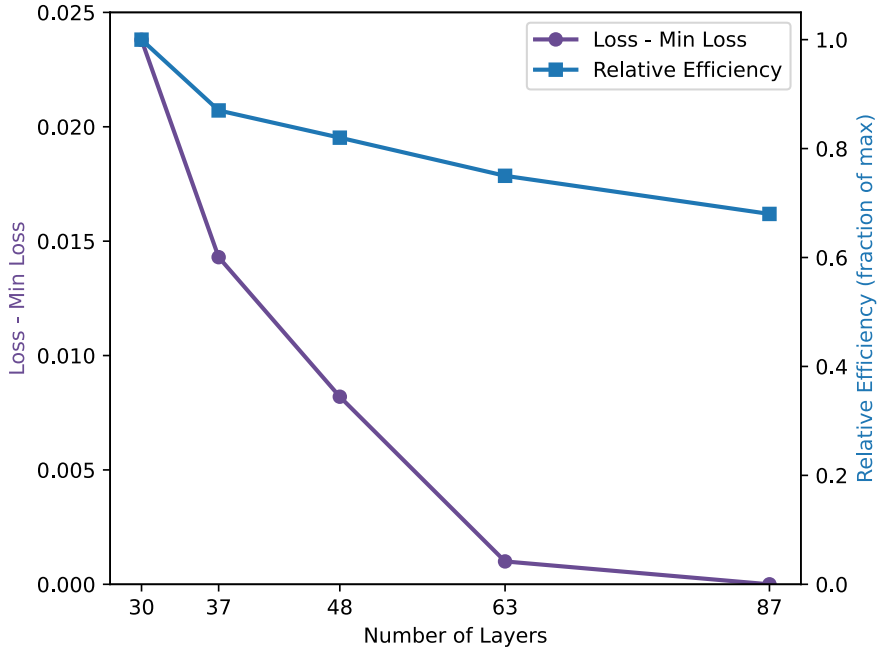
- W1536L87, W1792L63, W2048L48, W2304L37, and W2560L30

The deep version even matched/outperformed 3B and 7B checkpoints

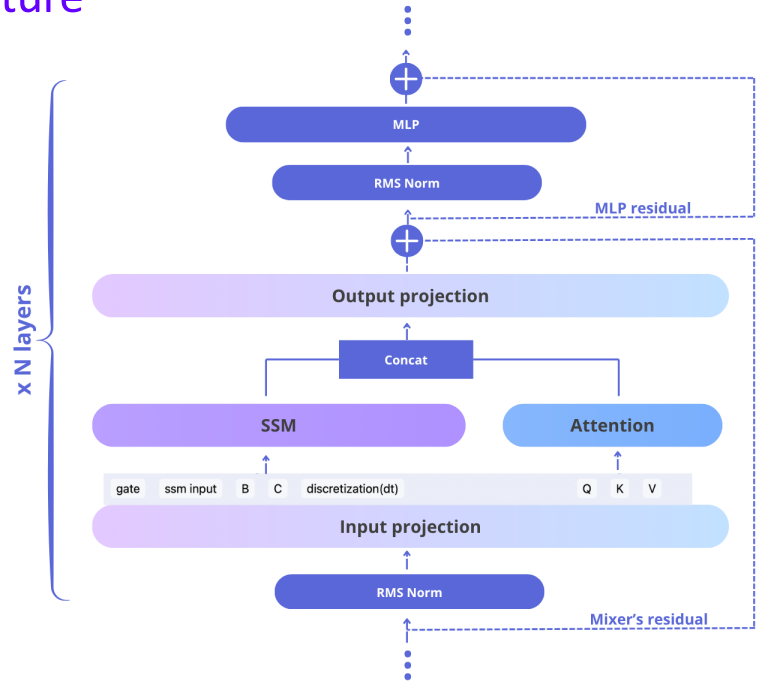
Also seen in GLM4.5: deeper design vs DS-V3 and Kimi K2

Release two 1.5B versions

- Falcon-H1-1.5B (24 layers)
- Falcon-H1-1.5B-Deep (66 layers)



Falcon-H1: Final Architecture



Model	Params (B)	Layers	# Vocab	d_{model}	Heads (Q/KV, SSM)	d_{head} (Attn/SSM)	d_{state}	Context Len.	# Tokens
`Falcon-H1-0.5B`	0.52	36	32,778	1024	8/2, 24	64/64	128	16K	2.5T
`Falcon-H1-1.5B`	1.55	24	65,536	2048	8/2, 48	128/64	256	128K	3T
`Falcon-H1-1.5B-Deep`	1.55	66	65,536	1280	6/2, 24	128/64	256	128K	3T
`Falcon-H1-3B`	3.15	32	65,536	2560	10/2, 32	128/128	256	128K	2.5T
`Falcon-H1-7B`	7.59	44	130,048	3072	12/2, 24	128/128	256	256K	~12T
`Falcon-H1-34B`	33.6	72	261,120	5120	20/4, 32	128/128	256	256K	~18T

Training Dynamics – Spikes removal and µP

High learning rates/Bigger models -> Systematic Spikes

Constraints

- Can not go for high learning rates -> Disables an interesting Hyperparameters area
- Spikey runs tend to have worst loss -> Hard to compare ablation runs

Solution

- Dampening dt works the best: add small µP multiplier to dt-activation

○ µP with tunable multipliers

- 35 multipliers
- 10 stages of iterative optimization
- 500+ runs

Hyperparameters	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
const stage duration	5GT*	25GT	25GT	25GT	65GT	65GT	65GT	65GT	65GT	65GT
decay stage duration	0GT	5GT	5GT	5GT	10GT	10GT	10GT	10GT	10GT	10GT
rampup duration	GT	GT	GT	GT	GT	GT	GT	GT	GT	GT
batch sizes	4M	4M	4M	4M	4M	4M	4M	4M	4M	16M
lr	5.12	5.12	5.12	5.12	2.56	2.56	2.56	2.56	2.56	2.56
wd	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Embedding multipliers										
embedding out	6	6	6	5	4	3	3	3	2.5	2.5
lr	0	0	2	1	1	1	1.5	1.5	2	1.5
wd	0	0	0	-1	-1	-1	-1.5	-1.5	-2	-2.5
Projector multipliers										
projector out	-2	-2	-3	-3	-3.5	-3.5	-4.5	-4.5	-5	-5
lr	0	0	0	-1	-2	-1.5	-0.5	-0.5	0	0
wd	0	0	0	1	1	0.5	-0.5	-0.5	-2	-2
mixer out_proj mlts										
ssm out	0	0	-2	-3	-2	-2	-2.5	-2.5	-2.5	-1.5
attn out	0	0	-2	-2	-2	-1	-1.5	-1.5	-1	-1
lr	0	0	-1	-0.5	-1.5	-2	-2.5	-2.5	-2	-2
wd	0	0	0	0.5	1.5	2	2.5	2.5	2	2
mixer in_proj mlts										
ssm x	0	-2	0	-1	-1	-1.5	-1.5	-1.5	-2	-2
ssm B	-2	-2	-4	-3	-2	-2	-1	-1	-1.5	-1.5
ssm C	0	0	0	0	0	0	0	0	0	-1

Hyperparameters	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
ssm dt	-1	-2	-3	-2	-1	-1.5	-0.5	-0.5	-1.5	-1.5
ssm gate	0	-2	0	-1	0	-1	-1	-1	-1.5	-1.5
attn key	-2	-2	-3	-2	-2.5	-2.5	-2	-2	-2	-2
lr	0	0	0	-0.5	-1.5	-0.5	-1	-1	-0.5	-0.5
wd	0	0	0	0.5	0.5	0.5	1	1	0.5	0.5
MLP out mlts										
mlp out	0	0	-2	-1	-1	-1	-1	-1	-2	-2
up_proj lr	0	0	-1	-1	-2	-1	0	0	-0.5	-0.5
up_proj wd	0	0	0	1	1	1	0	0	-0.5	-0.5
down_proj lr	0	0	-1	-1	-2	-1	-0.5	-0.5	-0.5	-0.5
down_proj wd	0	0	0	1	1	1	-0.5	-0.5	-0.5	-0.5
MLP gate mlts										
mlp gate out	0	0	-1	-2	-1.5	-1.5	-1	-1	-0.5	-0.5
gate_proj lr	0	0	0	-0.5	-0.5	0	1	1	0.5	0.5
gate_proj wd	0	0	0	0.5	-0.5	-0.5	-0.5	-0.5	0	0
RMS norms lr mlts										
norm (ssm+attn)	0	0	0	0	1	2	2	2	2	2
norm2 (mlp)	0	0	0	0	1	1.5	1.5	1.5	1.5	1.5
norm_gate (mlp gate)	0	0	0	0	1	1	1	1	1	1
norm_f (projector)	0	0	0	0	1	1.5	1.5	1.5	1.5	1.5
SSM bias-like lr mlts										
conv1d.weight	0	0	0	1	2	2	3	3	2.5	2.5
conv1d.bias	0	0	0	0	1	1	2	2	1	1
dt.bias	0	0	0	0	1	1	2	2	1.5	1.5
A_log	0	0	0	0	1	1.5	2	2	1.5	1.5
D	0	0	0	0	1	1	2	2	3	3

Pretraining Infrastructure

- 4096 H100s, in-house training framework with 5D Parallelism.
- Redesigned Context Parallelism (CP) for hybrid attention-mamba2
- Innovative Mixer Parallelism (MP) for parallel attention and SSM heads

Models	Batch Size	Context Len. Stage	DP	TP	PP	CP	MP
Falcon-H1-0.5B	4M	4K, 16K	64	1	1	1	✗
Falcon-H1-1.5B (1.5B-Deep)	4M	16K, 32K	256	1	1	1	✗
		131K	64	1	1	4	✗
Falcon-H1-3B	8M	16K, 32K	256	1	1	1	✗
		131K	64	1	1	4	✗
Falcon-H1-7B	8M	16K, 32K	256	2	1	1	✓
		131K	128	2	1	4	✓
		262K	64	2	1	8	✓
Falcon-H1-34B	26M	16K	448	4	2	1	✓
		32K	192	4	2	2	✓
		131K	48	4	2	8	✓
		262K	24	4	2	16	✓

Pretraining Infrastructure – Mixer Parallelism (MP)

- Parallel-head hybrid models with TP-repartitioning
  - Optimized for training/inference efficiency
  - Balanced distribution of computational and memory overhead
- Improved memory utilization compared to sequential hybrid designs.

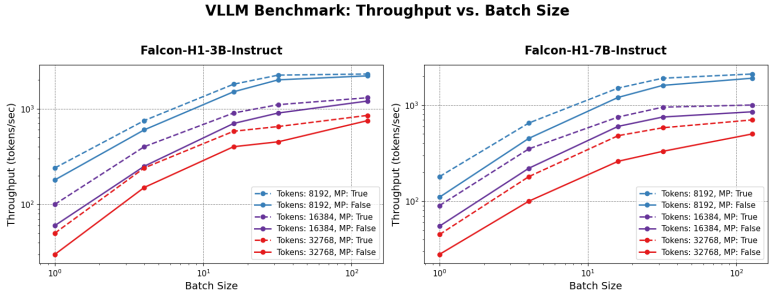
Without Mixer Parallelism

Naive Mixer Parallelism

Interleaved Mixer Parallelism

MP Variant	Throughput (Gtok/hr)	Speedup (ratio)
None (Baseline)	0.2339	1.00
Naive MP	0.2640	1.13
Interleaved MP	**0.3343**	**1.43**

Training Efficiency with MP (2B model)



Inference Efficiency with MP

Falcon-H1: Data & Data strategy

Tokenizer: falcon-world

- Trained on 100+ languages
- English-only version available
- Beating GPT-4o, Qwen2.5, Mistral, Llama3 tokenizers

Tokenizer name	Vocabulary size	Model
`falcon-world-32k`	32768	`tiiuae/Falcon-H1-0.5B*`
`falcon-world-65k`	65536	`tiiuae/Falcon-H1-1.5B*`, `tiiuae/Falcon-H1-3B*`
`falcon-world-131k`	131048	`tiiuae/Falcon-H1-7B*`
`falcon-world-262k`	261120	`tiiuae/Falcon-H1-34B*`

Falcon-H1: Data & Data strategy

Tokenizer: falcon-world

- Trained on 100+ languages
- English-only version available
- Beating GPT-4o, Qwen2.5, Mistral, Llama3 tokenizers

Tokenizer name	Vocabulary size	Model
`falcon-world-32k`	32768	`tiiuae/Falcon-H1-0.5B*`
`falcon-world-65k`	65536	`tiiuae/Falcon-H1-1.5B*`, `tiiuae/Falcon-H1-3B*`
`falcon-world-131k`	131048	`tiiuae/Falcon-H1-7B*`
`falcon-world-262k`	261120	`tiiuae/Falcon-H1-34B*`

Training data (2.5T – 18T)

- Extensive explorations & processing
  - Validating every data source before injecting into training
  - Finding optimal data mixture at various scales
  - Optimal data format, synthetic data, data organization strategies, etc.

	34B		**7B**		**3B**	**1.5B**	**0.5B**
Data Source	Start	End	Start	End	Mix	Mix	Mix
Raw data	99.47	43.45	81.07	42.26	39.70	23.20	11.50
Web	40.00	14.60	25.00	12.35	11.60	10.20	6.50
Curated	25.00	15.93	26.00	16.47	11.68	4.75	0.00
Code	20.00	10.05	20.00	10.74	14.00	8.00	5.00
Math	14.47	2.87	10.07	2.70	2.42	0.25	0.00
Rewritten data	0.23	52.05	10.56	53.04	56.80	69.80	75.50
Web & Curated	0.00	20.36	0.00	18.12	22.08	23.75	20.50
Code & Math	0.23	31.69	10.56	34.92	34.72	46.05	55.00
Synthetic data*	0.30	4.50	8.37	4.70	3.50	7.00	13.00
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00

* Fully synthetic samples, not derived from rewriting existing raw data.

Falcon-H1: Data & Data strategy

Tokenizer: falcon-world

- Trained on 100+ languages
- English-only version available
- Beating GPT-4o, Qwen2.5, Mistral, Llama3 tokenizers

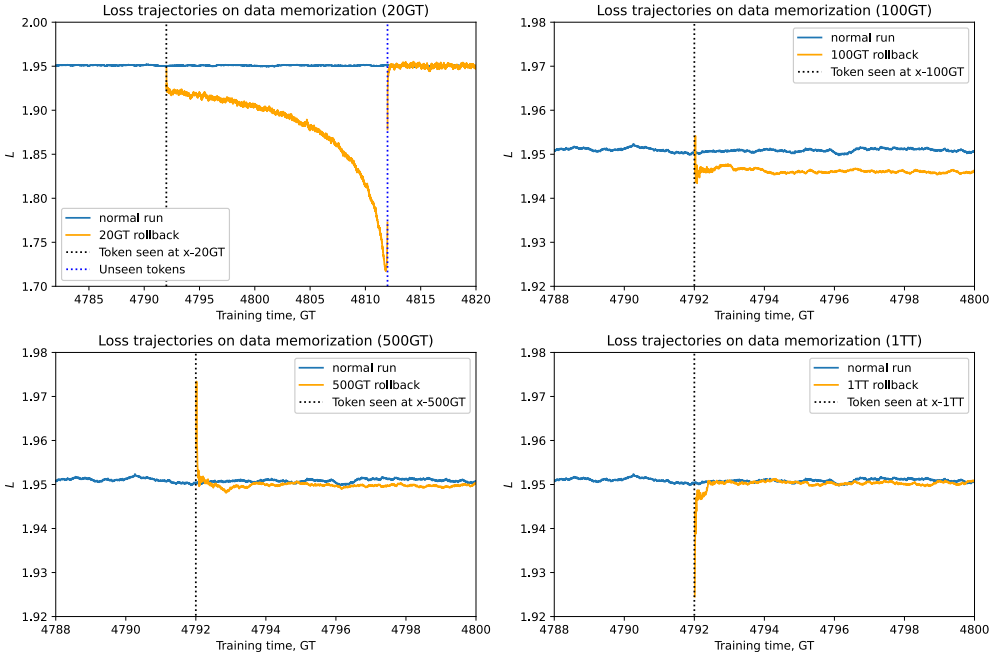
Tokenizer name	Vocabulary size	Model
`falcon-world-32k`	32768	`tiiuae/Falcon-H1-0.5B*`
`falcon-world-65k`	65536	`tiiuae/Falcon-H1-1.5B*`, `tiiuae/Falcon-H1-3B*`
`falcon-world-131k`	131048	`tiiuae/Falcon-H1-7B*`
`falcon-world-262k`	261120	`tiiuae/Falcon-H1-34B*`

Training data (2.5T – 18T)

- Extensive explorations & processing
  - Validating every data source before injecting into training
  - Finding optimal data mixture at various scales
  - Optimal data format, synthetic data, data organization strategies, etc.

Data strategies

- Multi-epoch training, model memorization and forgetting window, dynamic mixture on-the-fly, etc.
- Anti-curriculum learning in pre-training
- 16K context pretraining, extension to up-to 256K …

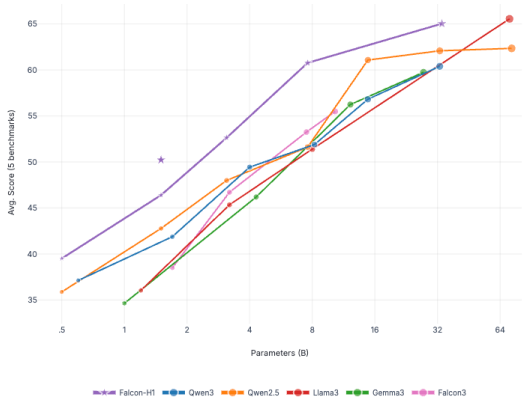


Evaluation – During Training

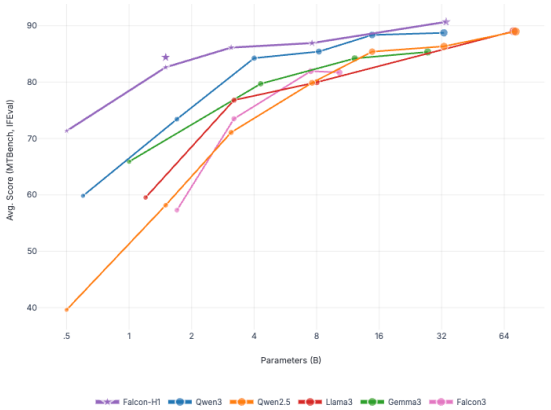
- Frequent benchmark evaluations to capture better signals
- Frequent vibe checks on intermediate checkpoints, to avoid unintended specialization or domain biases

Evaluation – Final Results (Instruct)

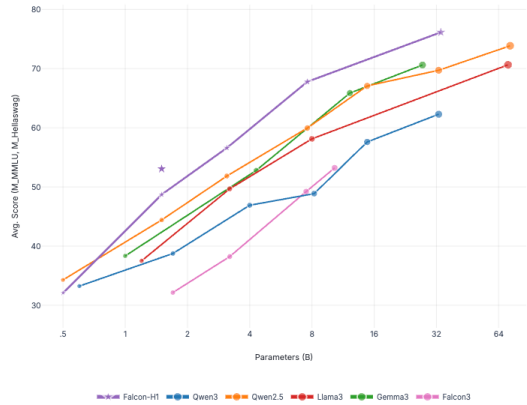
General tasks
(LiveBench, BBH, ARC-C, Hellaswag, TruthfulQA)



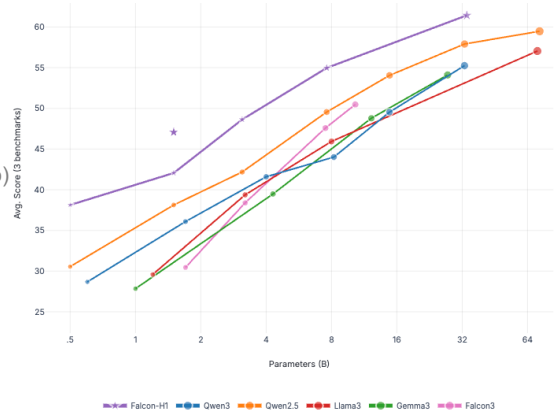
Instruct following
(IFEval, MTBench)



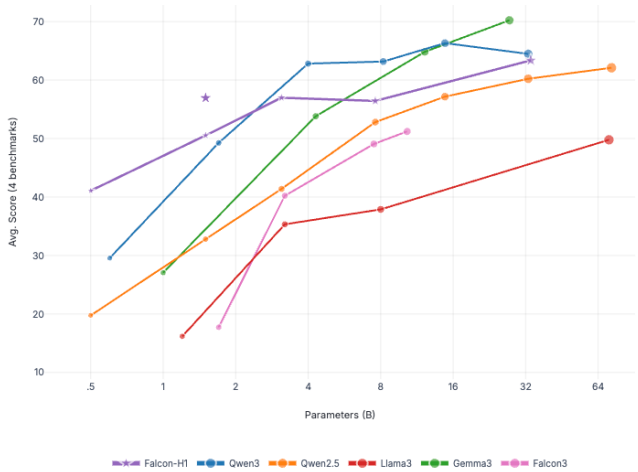
Multilingual
(M_MMLU, M_Hellaswag)



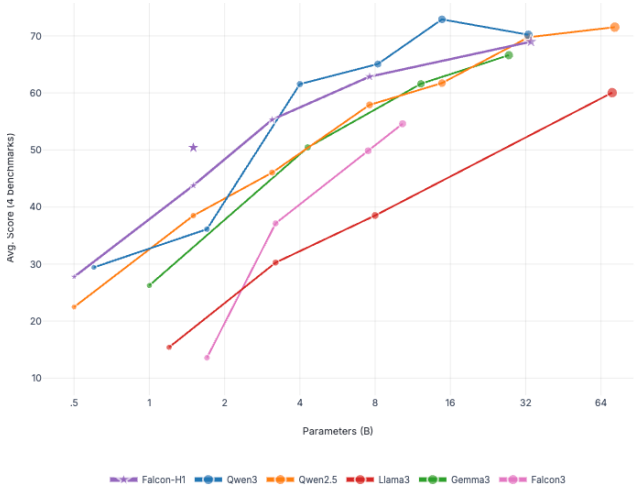
Science et al.
(GPQA, MMLU, MMLU-Pro)



Evaluation – Final Results (Instruct)



Math
(GSM8K, MATH-500, AMC23, AIME2025)



Coding
(HumanEval+, MBPP+, LiveCodeBench, CRUXEval)

- Recent models have shifted their strengths toward math- and reasoning-intensive tasks, while sacrificing performance on general and knowledge-intensive tasks
- Model comparisons beyond architecture – Data mixture matters – business use cases, model positioning, and real-world deployment scenarios, etc.

Evaluation – Long Context Results

Falcon-H1-34B-Instruct

- Shines on RAG tasks
- On Recall, longQA tasks, outperforms Qwen2.5-72B ,but lag behind Qwen3-32B and Llama-3.3-70B (beyond 32k)

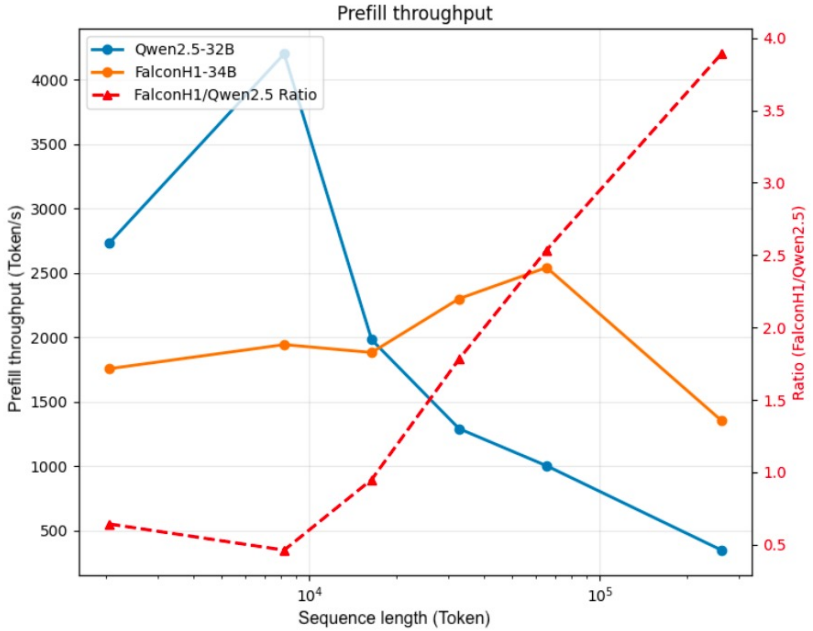
Seq. Length	Falcon-H1-34B-Instruct	Qwen2.5-72B-Instruct	Qwen3-32B	Llama-3.3-70B-Instruct
HELMET-RAG				
8k	72.17	72.21	69.25	**74.29**
16k	81.46	80.42	77.92	**82.33**
32k	67.96	70.08	64.83	**70.21**
65k	67.08	63.25	61.96	**69.08**
131k	**62.21**	42.33	57.08	55.38
HELMET-Recall				
8k	100.00	100.00	100.00	100.00
16k	100.00	100.00	100.00	100.00
32k	97.50	98.38	**100.00**	99.63
65k	80.69	71.75	96.50	**98.81**
131k	56.63	38.81	**86.13**	82.19
HELMET-longQA				
8k	32.87	**35.20**	31.63	33.67
16k	34.64	39.13	35.68	**39.75**
32k	35.09	39.22	41.15	**47.53**
65k	32.45	36.71	47.47	**48.57**
131k	33.81	32.94	**53.52**	46.06

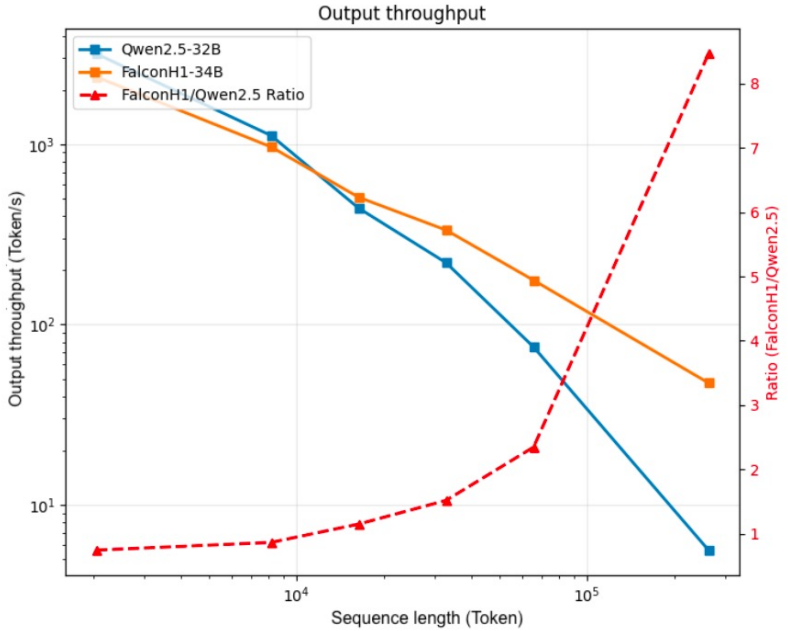
Long context benchmarking on Helmet, check full results in Falcon-H1's tech report

Evaluation – Model Efficiency

Falcon-H1-34B VS Qwen2.5 32B

- Prefill test: Input seq_len (2k to 262k), output 2k (batch_size 32) -> up to 4x speedup
- Generation test: Input 4k (batch_size 32), output seq_len (2k to 262k) -> up to 8x speedup





Evaluation Results – Model Depth Wins

Falcon-H1-1.5B

- SoTA 1B-scale model to date

Falcon-H1-1.5B-Deep

- Matches or surpasses Qwen2.5-7B across most benchmarks (check tech report)

Tasks	Falcon-H1-1.5B-Deep	Falcon-H1-1.5B	Qwen3-1.7B	Qwen2.5-1.5B	Gemma3-1B	Llama3.2-1B	Falcon3-1.6B
General							
BBH	**54.43**	46.47	35.18	42.41	35.86	33.21	34.47
ARC-C	**43.86**	42.06	34.81	40.53	34.13	34.64	43.09
TruthfulQA	**50.48**	49.39	45.98	47.05	42.17	42.08	42.31
HellaSwag	**65.54**	63.33	49.27	62.23	42.24	55.30	58.53
MMLU	**66.11**	62.03	57.04	59.76	40.87	45.93	46.10
Math							
GSM8k	**82.34**	74.98	69.83	57.47	42.38	44.28	44.05
MATH-500	**77.80**	74.00	73.00	48.40	45.40	13.20	19.80
AMC-23	**56.56**	46.09	43.59	24.06	19.22	7.19	6.87
AIME-24	**14.37**	12.50	11.25	2.29	0.42	1.46	0.41
AIME-25	**11.04**	9.58	8.12	1.25	1.25	0.00	0.21
Science							
GPQA	**33.22**	26.34	27.68	26.26	28.19	26.59	26.76
GPQA_Diamond	**40.57**	35.19	33.33	25.59	21.55	25.08	31.31
MMLU-Pro	**41.89**	37.80	23.54	28.35	14.46	16.20	18.49
MMLU-stem	**67.30**	64.13	54.30	54.04	35.39	39.16	39.64
Code							
HumanEval	**73.78**	68.29	67.68	56.10	40.85	34.15	22.56
HumanEval+	**68.90**	61.59	60.96	50.61	37.20	29.88	20.73
MBPP	**68.25**	64.81	58.73	64.81	57.67	33.60	20.63
MBPP+	**56.61**	56.35	49.74	56.08	50.00	29.37	17.20
LiveCodeBench	**23.87**	17.61	14.87	12.52	5.09	2.35	0.78
CRUXEval	**52.32**	39.57	18.88	34.76	12.70	0.06	15.58
Instruction Following							
IFEval	**83.50**	80.66	70.77	45.33	61.48	55.34	54.26
Alpaca-Eval	27.12	**28.18**	21.89	9.54	17.87	9.38	6.98
MTBench	**8.53**	8.46	7.61	7.10	7.03	6.37	6.03
LiveBench	36.83	34.13	**40.73**	21.65	18.79	14.97	14.10
Multilingual							
Multi-Hellaswag	**53.14**	49.38	37.89	42.93	41.77	39.78	32.04
Multi-MMLU	**53.00**	48.06	39.60	45.90	34.91	35.24	32.25
MGSM	**60.00**	58.00	52.40	45.20	-	29.73	15.33

Performance of the 1B+ Instruct models

Falcon-H1: embracing the open-source ecosystem

Supported in most popular framworks

- General usage, fine-tuning, local/cloud deployment, quantization, etc
- More are on the way…

General Discussions

Model Architecture

- Sets the Lower Bound of model performance
- Defines key traits: efficiency, memory cost, etc.

Data Mixture

- Sets the Upper bound of model performance
- Much stronger impact than architecture once lower bound is guaranteed

Anti-intuition Observations

- Common practices, or research conclusions often valid only under narrow conditions
- Surprises in scaling, long context, broad validation
- Model performance shaped by interdependent factors → requires multi-dimensional optimization guided by experiences and exhaustive testing

General Discussions

Future work

- Extend context beyond **256K**
- **Inference optimization** with vLLM, llama.cpp, MLX, SGLang, etc.
- Advance **reasoning, coding, and agentic** capabilities
- Improve **token efficiency** in training
- Enhance data: broader, more diverse, high-quality datasets
- Etc.

Thank you & Take aways

- Technical report https://arxiv.org/abs/2507.22448
- Github https://github.com/tiiuae/falcon-h1
- Blogpost https://falcon-lm.github.io/blog/falcon-h1/
- HuggingFace collection https://huggingface.co/collections/tiiuae/falcon-h1-6819f2795bc406da60fab8df