CSCI 5942: AI ENGINEERING, LECTURE 2

Architectures

Reading a transformer as a cost model

Required

DeepSeek-V3 Technical Report, architecture sections · arXiv:2412.19437

Ainslie et al., GQA · arXiv:2305.13245

Optional

Shazeer, MQA · arXiv:1911.02150 Fedus et al., Switch · arXiv:2101.03961 Mixtral of Experts · arXiv:2401.04088

Gu and Dao, Mamba · arXiv:2312.00752 Su et al., RoPE · arXiv:2104.09864

Christoffer Heckman

Department of Computer Science, University of Colorado Boulder

August 27, 2026

Quick overview

Unfinished business. Codesign, then the live fit: five Shakespeare models against one law.

Finishing up Tuesday's content

Anatomy. Where eight billion parameters live, counted by hand.

Llama 3 8B from its config file

The cache. Why context length is a memory price.

RoPE, MQA, GQA, MLA as techniques for memory management

Mixture Experts. Models that are made up of 671B parameters and run 37B.

routing, load balance, and what the FLOP count hides

A showdown. Two spec sheets.

Which serves cheaper?

1 CODESIGN

Codesign

What a Budget Buys

Google DeepMind

1,000 H100s × 30 days × 40% MFU?

as measured; imagine why?

MFU is model FLOPs utilization: the fraction of the chips' peak rate the run actually reaches.

$$C \approx 1.0 \times 10^{24}\ \text{FLOPs}$$

$$N^* = \sqrt{C/120} \approx 93\text{B} \qquad D^* \approx 1.9\text{T}$$

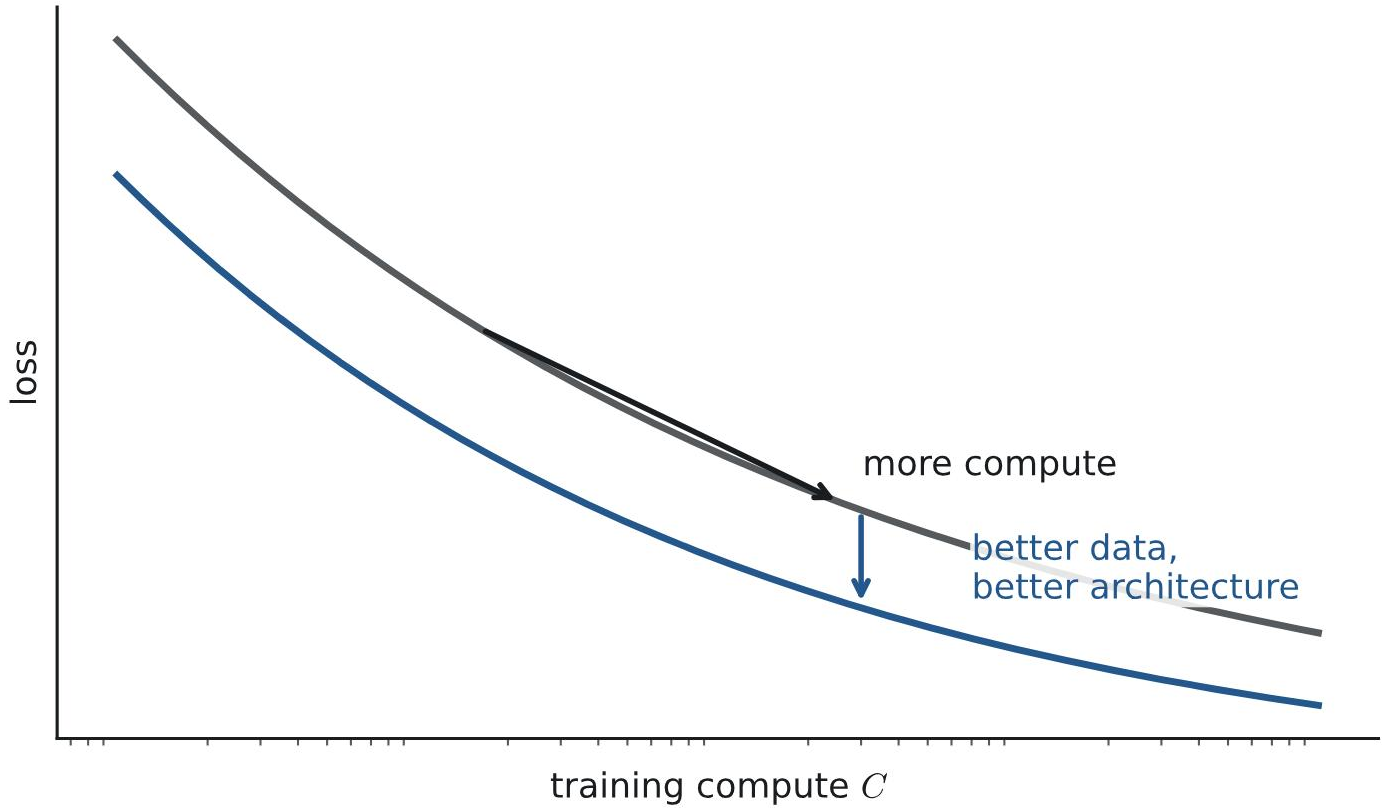
Chinchilla

Overtraining then says: build it smaller, train it longer.

at $2 per GPU-hour, this month of cluster time is about $1.4M, before any inference cost

Sizing rule: Hoffmann et al., "Training Compute-Optimal Large Language Models," arXiv:2203.15556 (2022). Chinchilla photograph: Jaime E. Jimenez, Universidad de Los Lagos, via Wikimedia Commons (attribution license).

One Currency



Measure the shift, but chart its significance in FLOPs.

2 LIVE

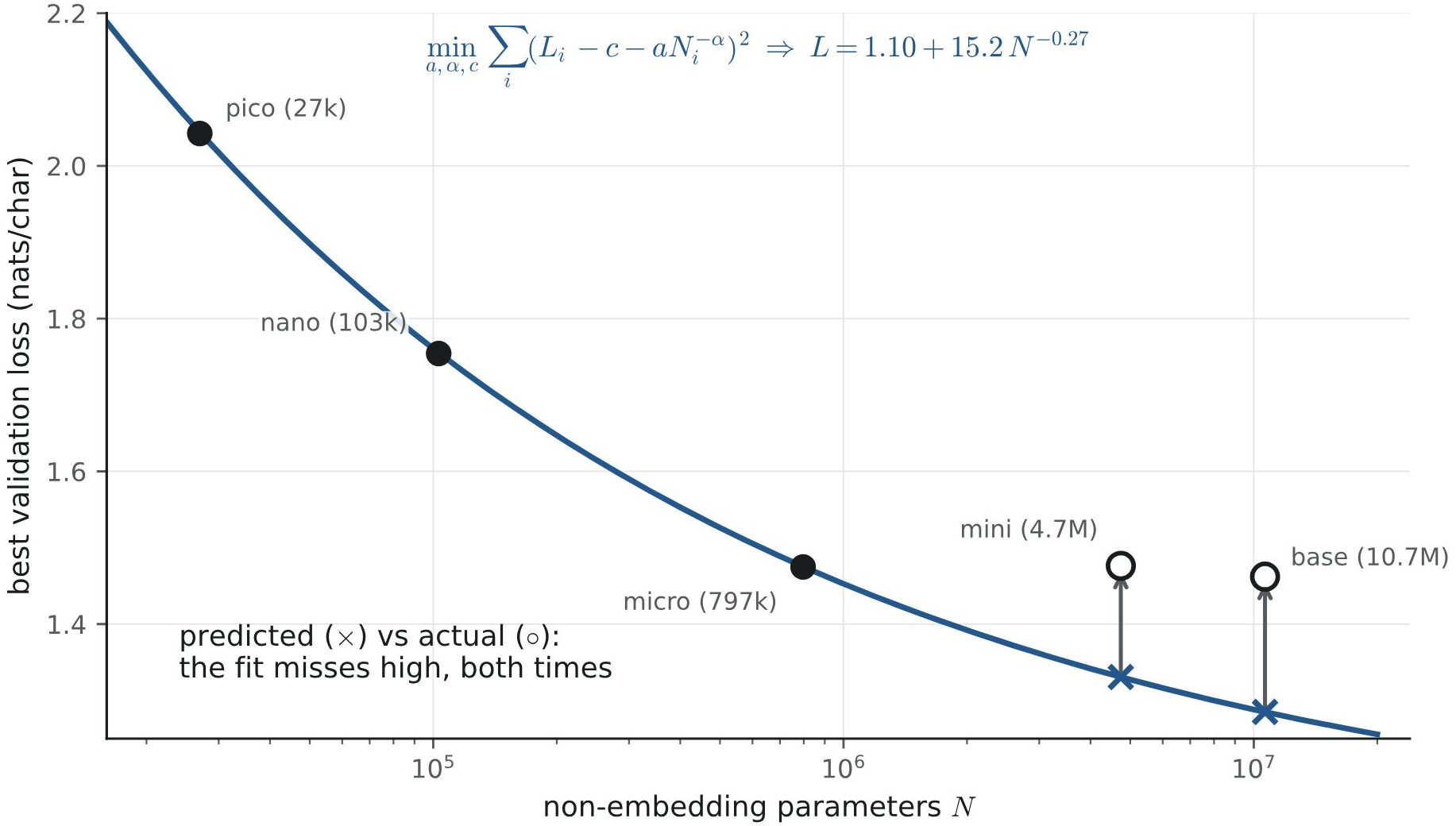
Let’s Play

The Ladder

run	layers	width	N	D
pico	2	32	27k	82M
nano	2	64	103k	82M
micro	4	128	797k	82M
mini	6	256	4.7M	82M
base	6	384	10.6M	82M
your prediction			mini: _____	base: _____

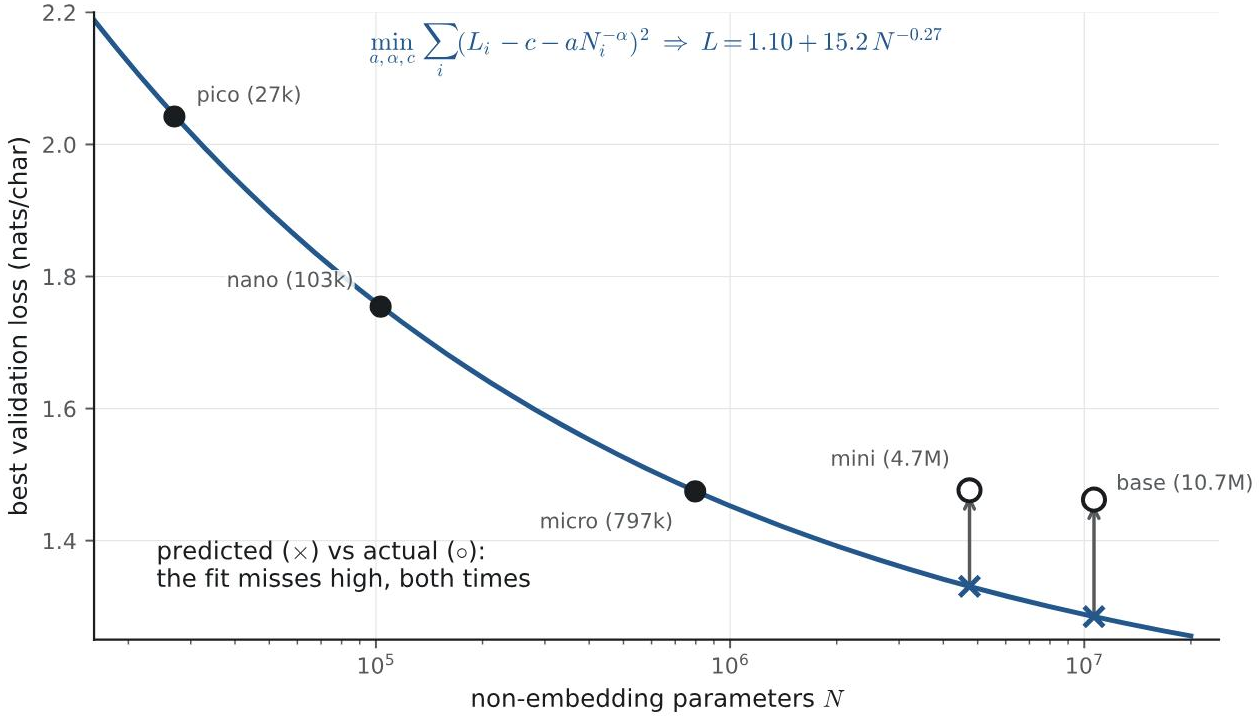
Fit the three smallest. **Predict mini and base.**

The Fit



Assignment 1 codebase, five runs, best validation loss per run. You fit this yourself in Part 3.

What Just Happened?

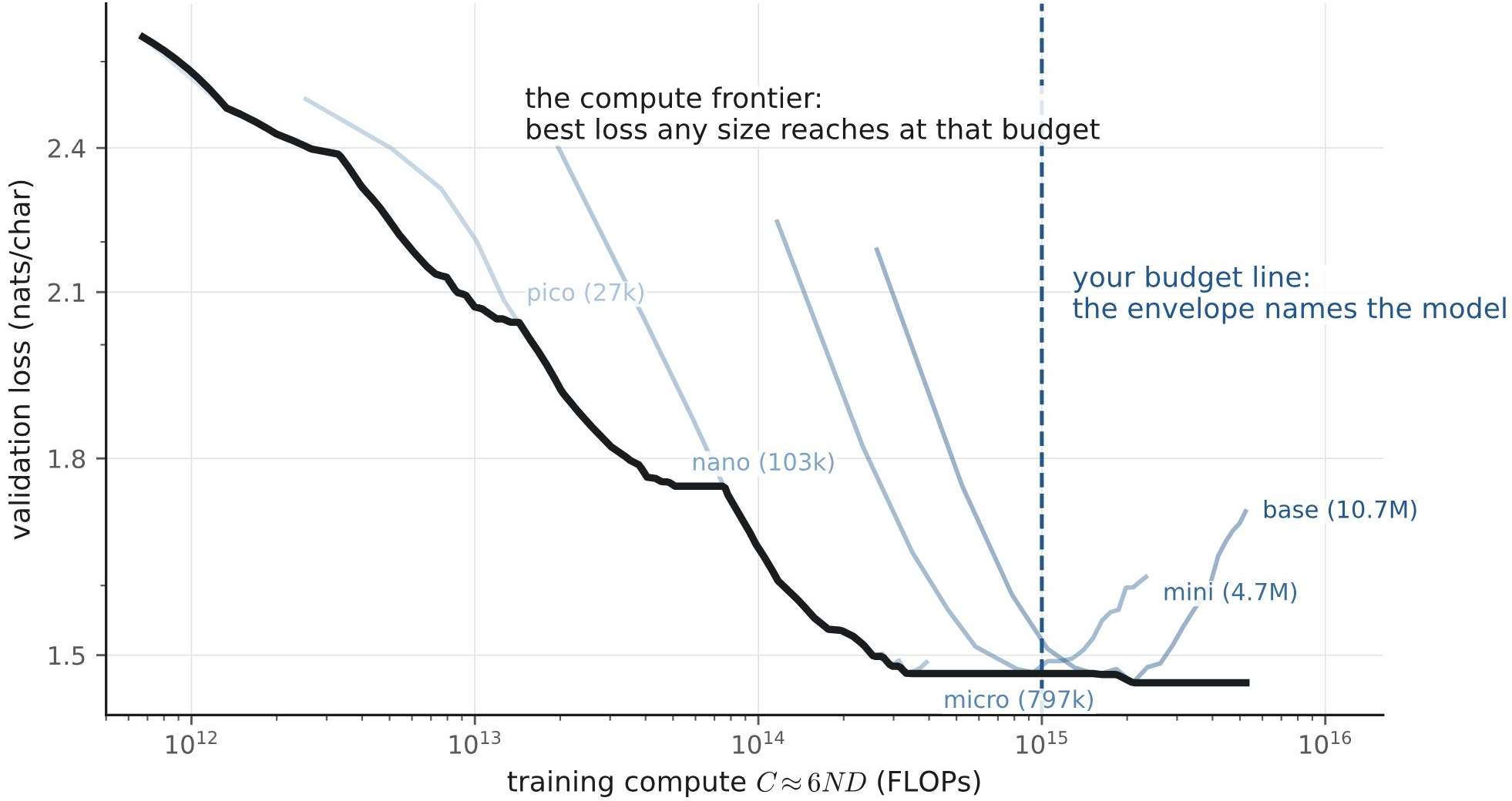


The fit assumed unlimited data.

base sits at 7.7 tokens per parameter, on epoch 80.

Chinchilla's ratio says base wanted 210M fresh tokens. Shakespeare has one million characters.

The Frontier

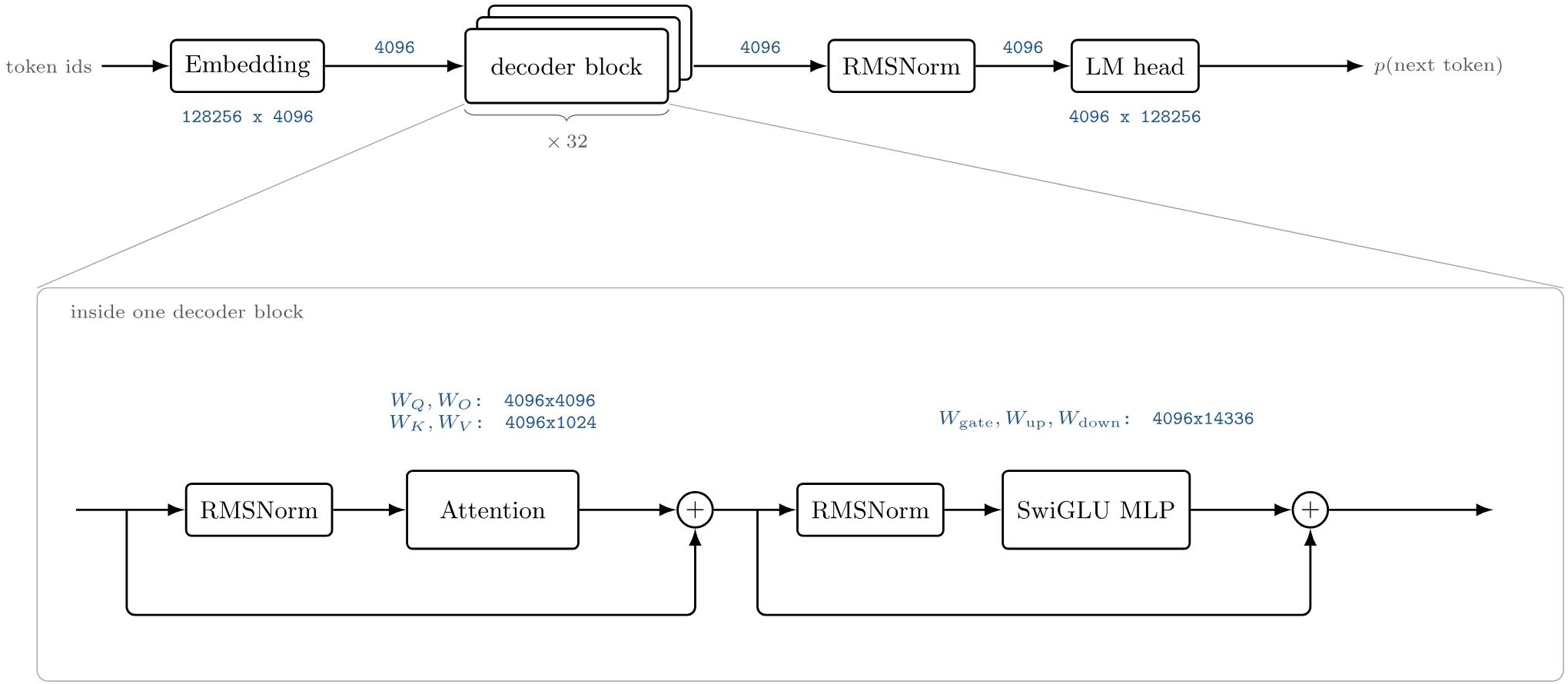


Same five runs, x-axis now training compute C = 6ND accumulated during each run.

3 ANATOMY

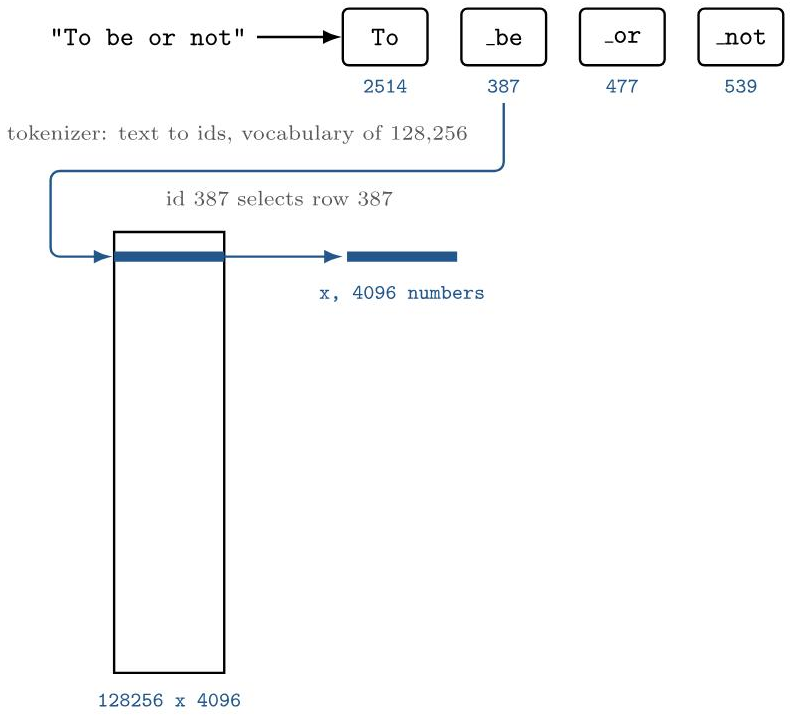
Anatomy

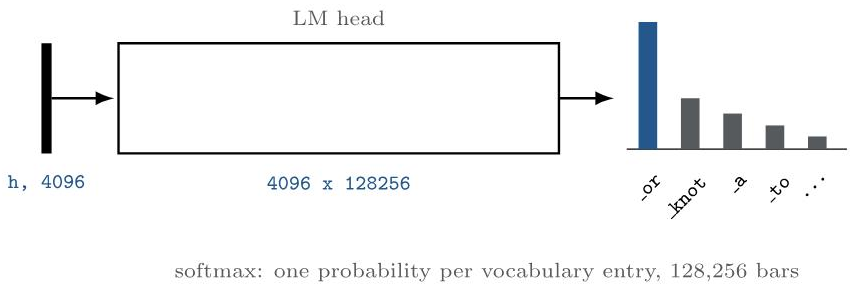
The Decoder Block



Shapes: Grattafiori et al., "The Llama 3 Herd of Models," arXiv:2407.21783 (2024); config.json, meta-llama/Llama-3.1-8B.

Embedding and Head



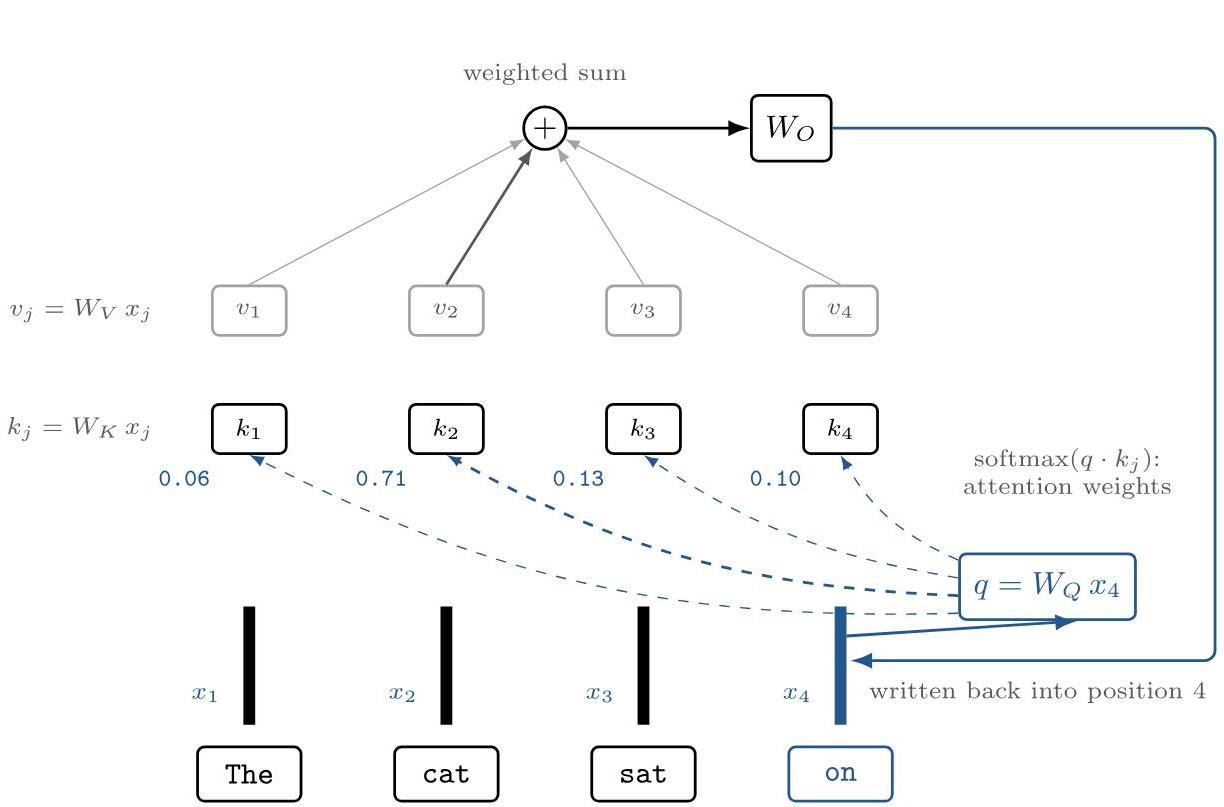


$$\mathrm{RMSNorm}(x) = \frac{x}{\sqrt{\frac{1}{d}\sum_j x_j^2}} \odot g, \qquad g \in \mathbb{R}^{4096}$$

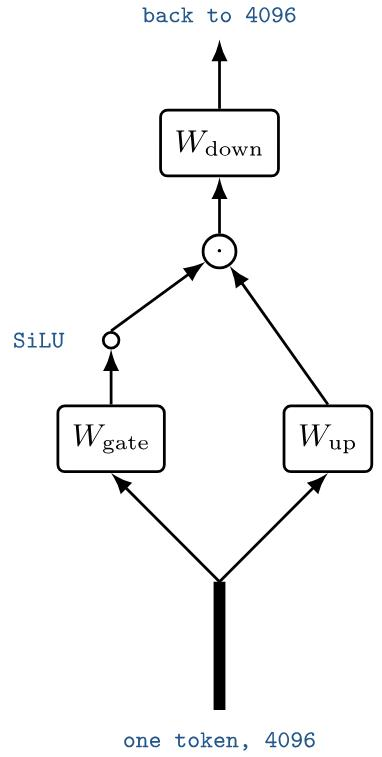
g is learned: one scale per channel, a norm's only parameters

Shapes: config.json, meta-llama/Llama-3.1-8B. Token ids illustrative.

Queries, Keys, Values



attention: q asks, k addresses, v carries, W_O writes back

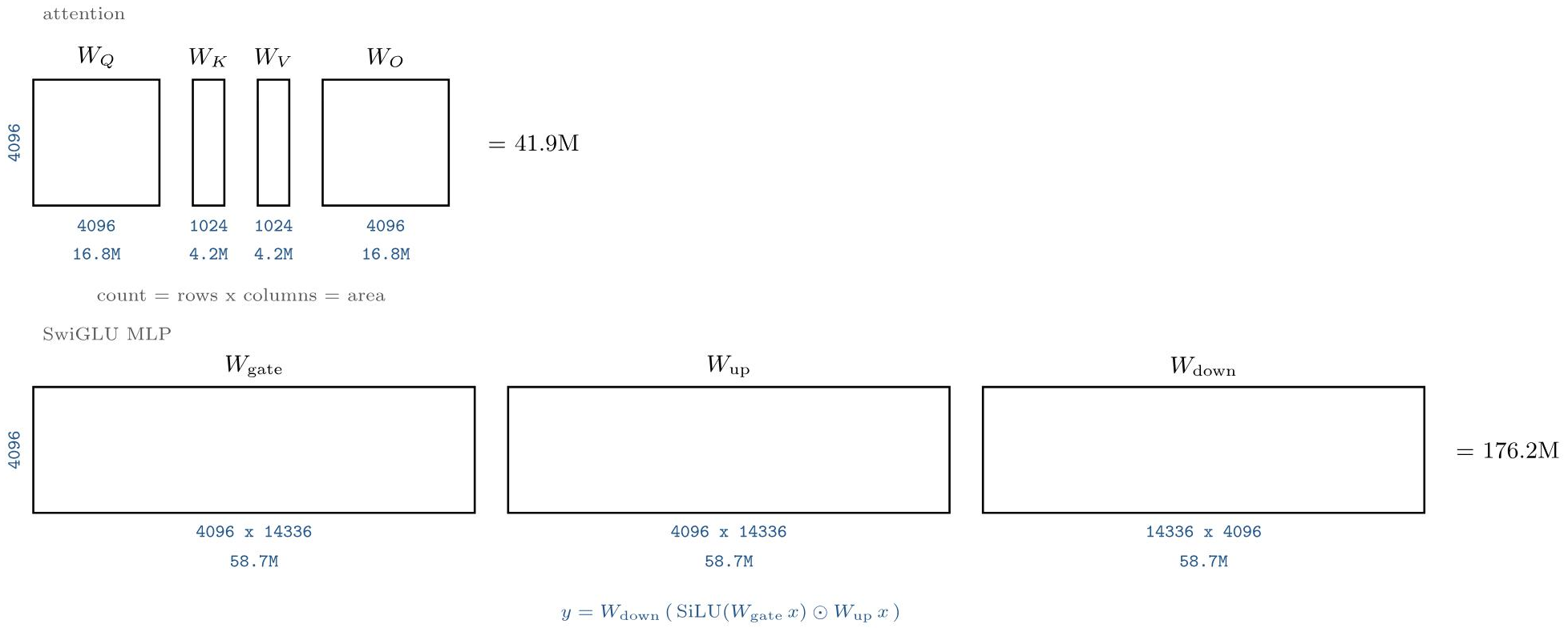


the MLP: each position alone; no other token is read

same kind of object, seven learned matrices; attention's four move information between positions, the MLP's three transform each position

Mechanism: Vaswani et al., "Attention Is All You Need," NeurIPS 2017. Attention weights illustrative; drawn single-head.

Counting by Area



In the 2017 transformer all four attention matrices are square. Llama narrows W_K, W_V by sharing kv heads.

Shapes: config.json, meta-llama/Llama-3.1-8B.

The Shape Table

```
"hidden_size": 4096,
"intermediate_size": 14336,
"num_hidden_layers": 32,
"num_attention_heads": 32,
"num_key_value_heads": 8,
"vocab_size": 128256,
"rope_theta": 500000.0
```

config.json, meta-llama/Llama-3.1-8B

$$W_Q, W_O \in \mathbb{R}^{4096\times4096} \qquad W_K, W_V \in \mathbb{R}^{4096\times1024}$$

$$\text{attention} = \underbrace{2\,(4096^2)}_{\text{two square: } W_Q, W_O} + \underbrace{2\,(4096\cdot1024)}_{\text{two narrow: } W_K, W_V} = 41.9\text{M}$$

$$= 2d^2 + 2\,d\,(n_{kv}\,d_{head}) \text{ with } d = 4096,\ n_{kv} = 8,\ d_{head} = 128$$

$$W_{gate}, W_{up}, W_{down} \in \mathbb{R}^{4096\times14336}$$

$$\text{MLP} = \underbrace{3}_{\text{three matrices}} (4096\cdot14336) = 176.2\text{M}$$

$$= 3\,d\,d_{ff} \text{ with } d_{ff} = 14336$$

one block $= 218.1$M, and the MLP is 81% of it

Eight Billion

$$\underbrace{32}_{L \text{ blocks}} \times \underbrace{218.1\text{M}}_{\text{one block}} = 6.98\text{B}$$

$$\text{embedding} + \text{head} = \underbrace{2}_{\text{two separate tables}} \times (\underbrace{128256}_{V} \cdot \underbrace{4096}_{d}) = 1.05\text{B}$$

$$6.98\text{B} + 1.05\text{B} + \underbrace{0.3\text{M}}_{\text{norms}} = 8.03\text{B}$$

norms: 65 RMSNorm scale vectors (2 per block + 1 final), each length 4096

8.03B, from seven numbers in a config file.

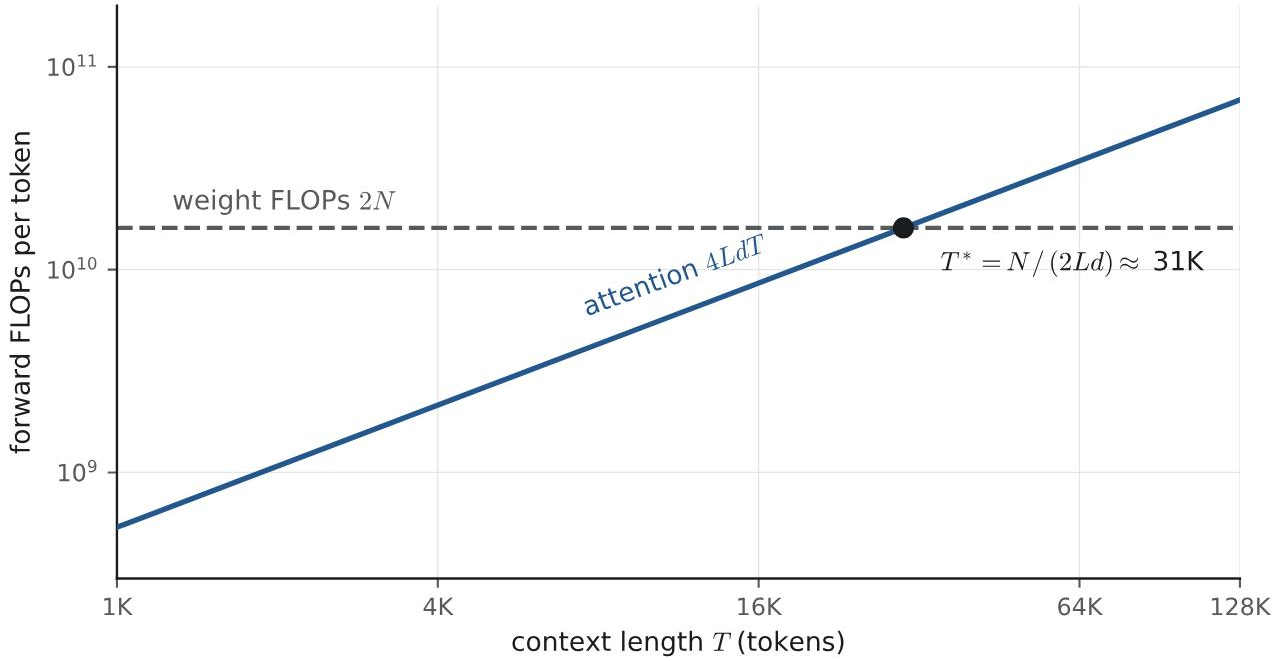
the same recipe on A1 base gives 10.7M; Part 1 asks you to check it

The Attention Term

Tuesday's $C \approx 6ND$ counted only the weights.

$$\text{attention, per layer per token:} \quad \underbrace{2\,T\,d}_{\text{read } T \text{ cached keys}} + \underbrace{2\,T\,d}_{\text{blend } T \text{ values}} = 4\,T\,d$$

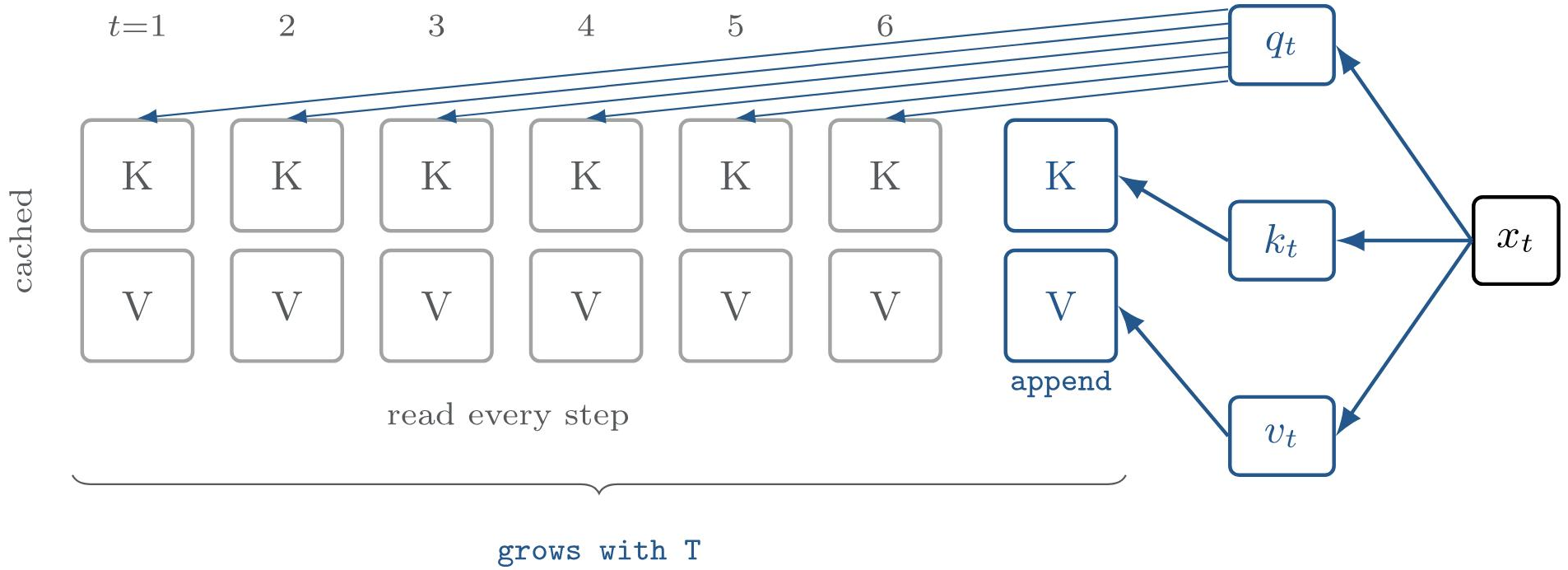
$$\text{forward FLOPs per token} = \underbrace{2N}_{\text{weights}} + \underbrace{4\,L\,d\,T}_{\text{attention}}$$



Counting after Kaplan et al., arXiv:2001.08361, Section 2.1. Shapes: Llama-3.1-8B; decode-step accounting, every new token attending to all T cached keys.

The Cache

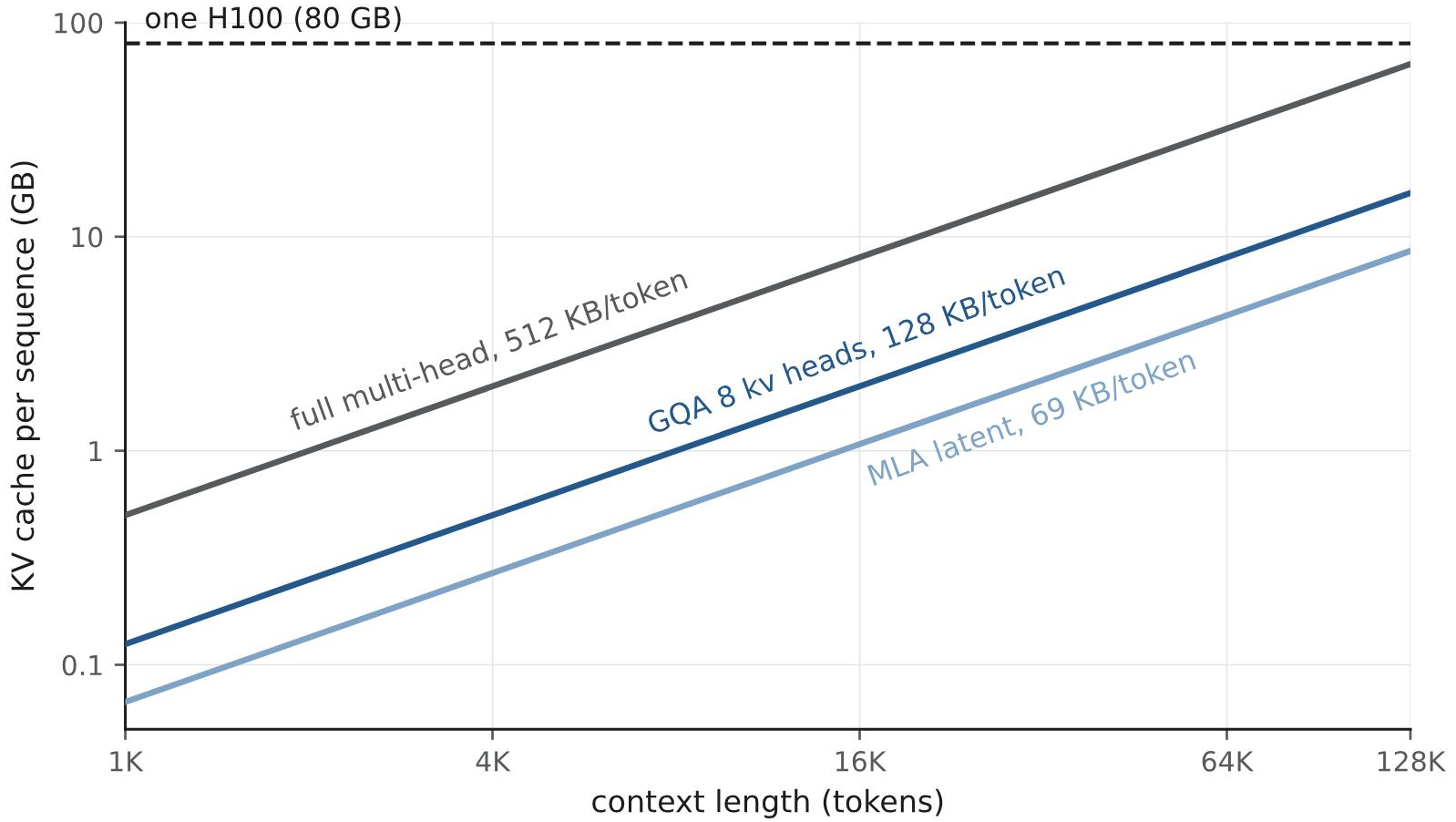
The Decode Step



per new token: compute q_t, k_t, v_t once; read every cached K, V

The Price of Context

$$\text{KV bytes per token} = \underbrace{2}_{K \text{ and } V} \cdot L \cdot n_{kv} \cdot d_{head} \cdot \underbrace{2}_{\text{fp16}}$$

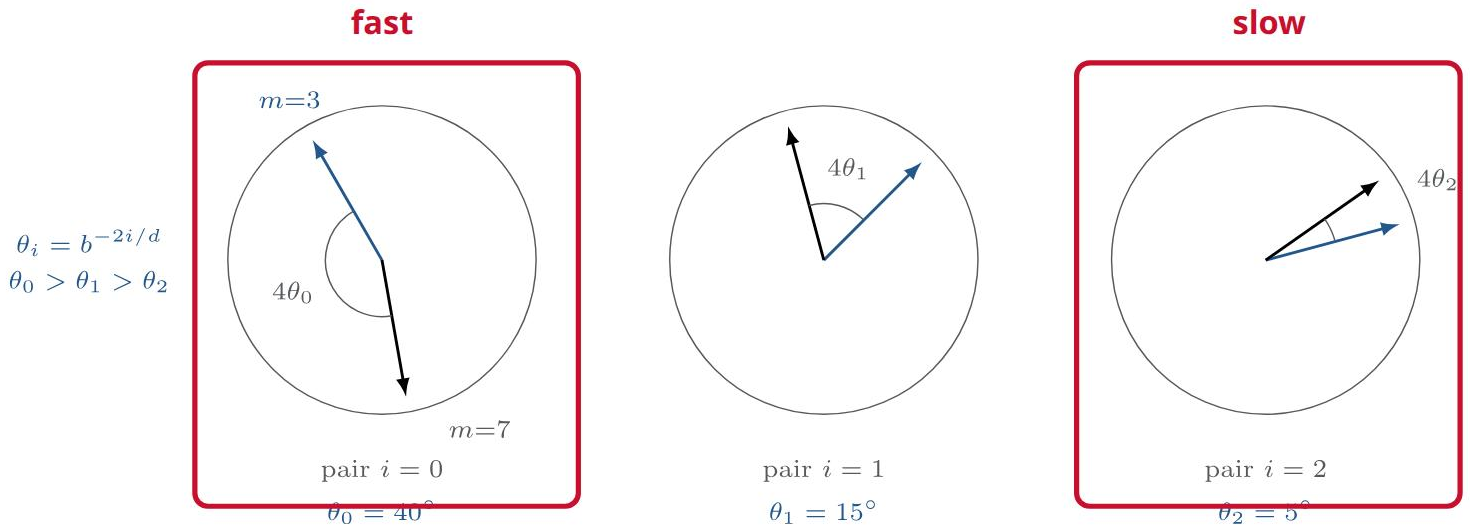


Shapes from each model's config.json; fp16 cache. Llama-3.1 context window 131,072 tokens: Grattafiori et al., arXiv:2407.21783 (2024). DeepSeek-V3 cache dimensions: arXiv:2412.19437.

5 ATTENTION

Attention Variants

RoPE



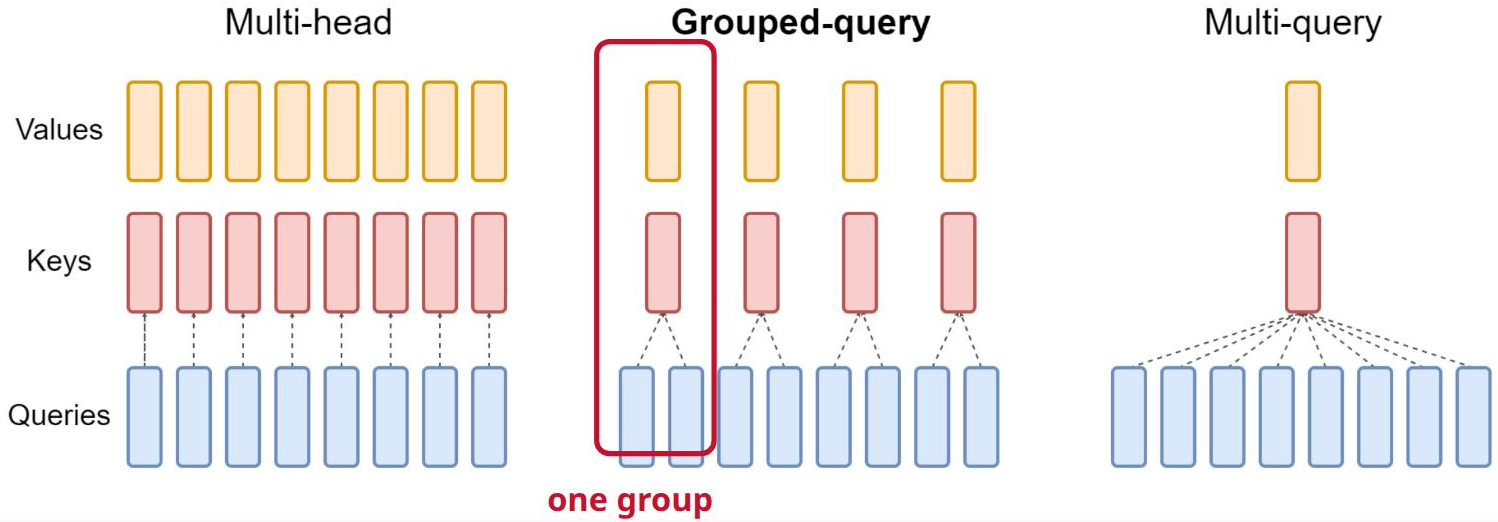
the angle between positions m and n is $(m-n)\,\theta_i$: the same for every absolute shift

$$\begin{pmatrix} q_{2i} \\ q_{2i+1} \end{pmatrix} \mapsto \begin{pmatrix} \cos m\theta_i & -\sin m\theta_i \\ \sin m\theta_i & \cos m\theta_i \end{pmatrix} \begin{pmatrix} q_{2i} \\ q_{2i+1} \end{pmatrix} \qquad \theta_i = b^{-2i/d}$$

$q_m^\top k_n$ depends on the offset $m - n$ only.

Su et al., "RoFormer: Enhanced Transformer with Rotary Position Embedding," arXiv:2104.09864 (2021). Llama 3 sets b = 500,000.

Sharing Keys

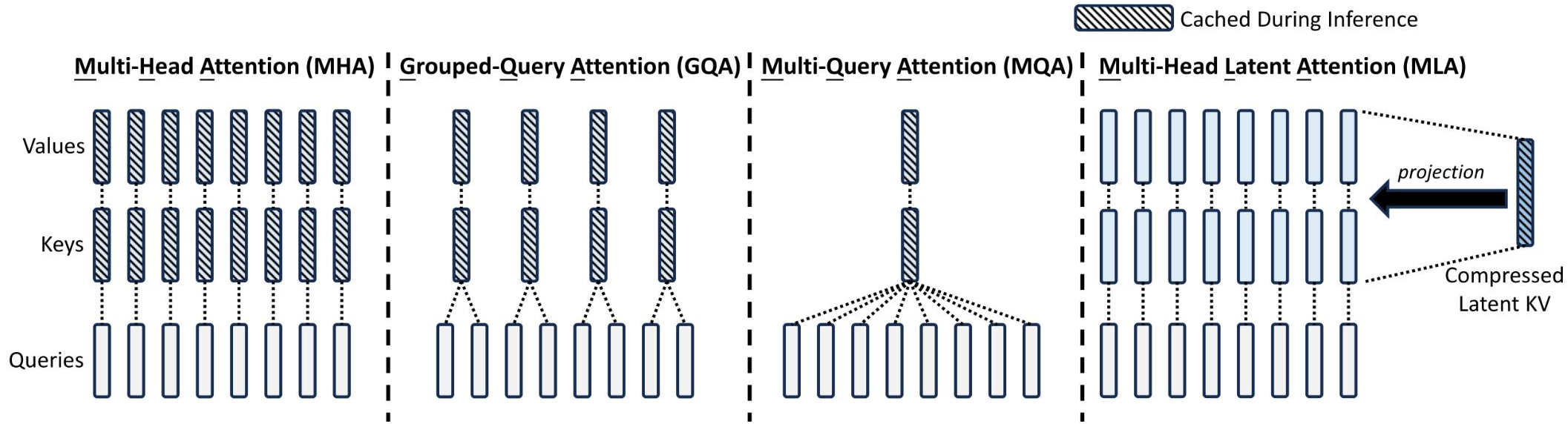


	kv heads	cache per token
multi-head	32	512 KB
grouped-query	8	128 KB
multi-query	1	16 KB

Converting a trained MHA model costs about 5% of its pretraining compute.

Ainslie et al., "GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints," EMNLP 2023, arXiv:2305.13245, Figure 2. Shazeer, "Fast Transformer Decoding," arXiv:1911.02150 (2019). Cache column at Llama-3-8B shapes, fp16.

The Latent Cache



$$\text{cache per token per layer} = \underbrace{512}_{\text{latent } c_{KV}} + \underbrace{64}_{\text{RoPE key}} \quad \text{values}$$

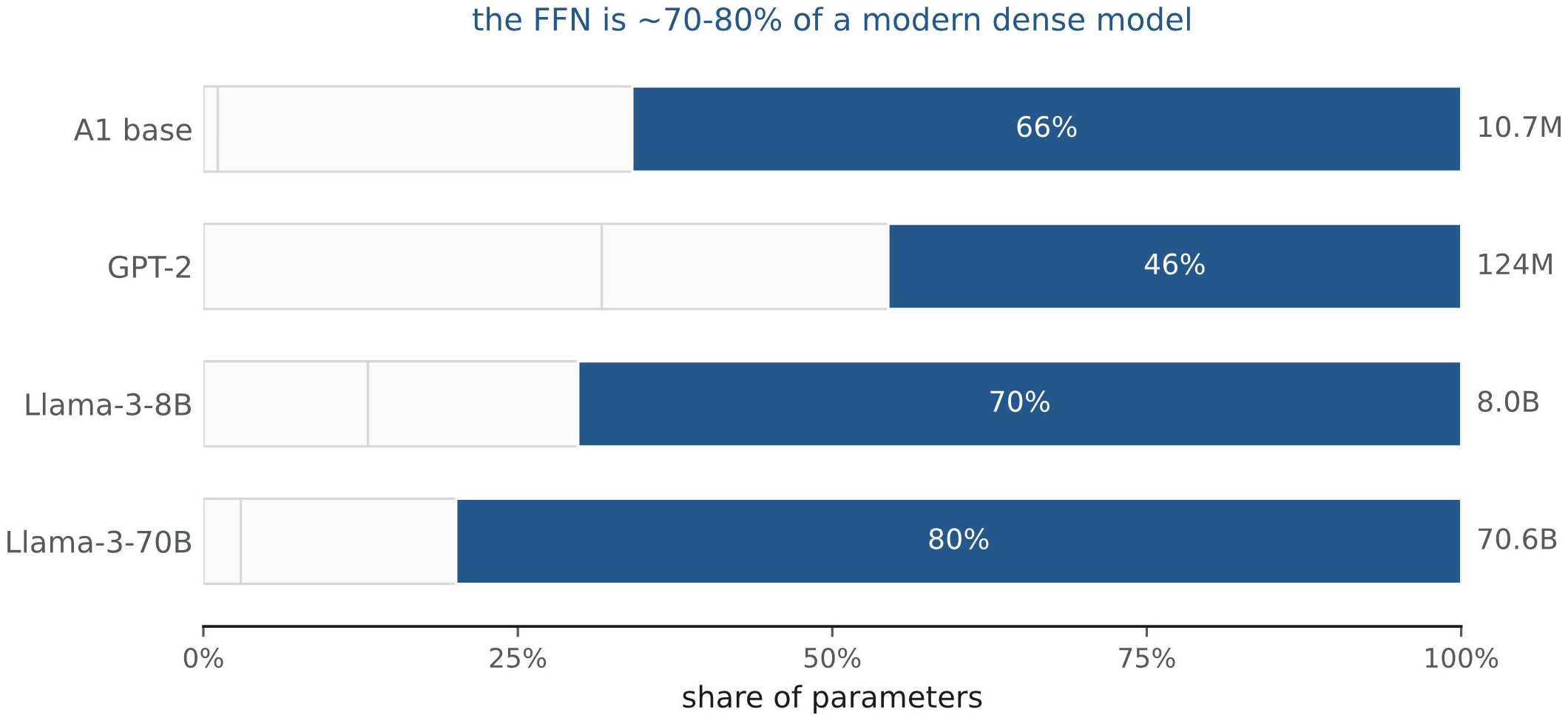
61 layers $\Rightarrow$ 69 KB per token, on a 671B model.

DeepSeek-AI, "DeepSeek-V2," arXiv:2405.04434 (2024), Figure 3. Cache dimensions: DeepSeek-V3 Technical Report, arXiv:2412.19437.

6 EXPERTS

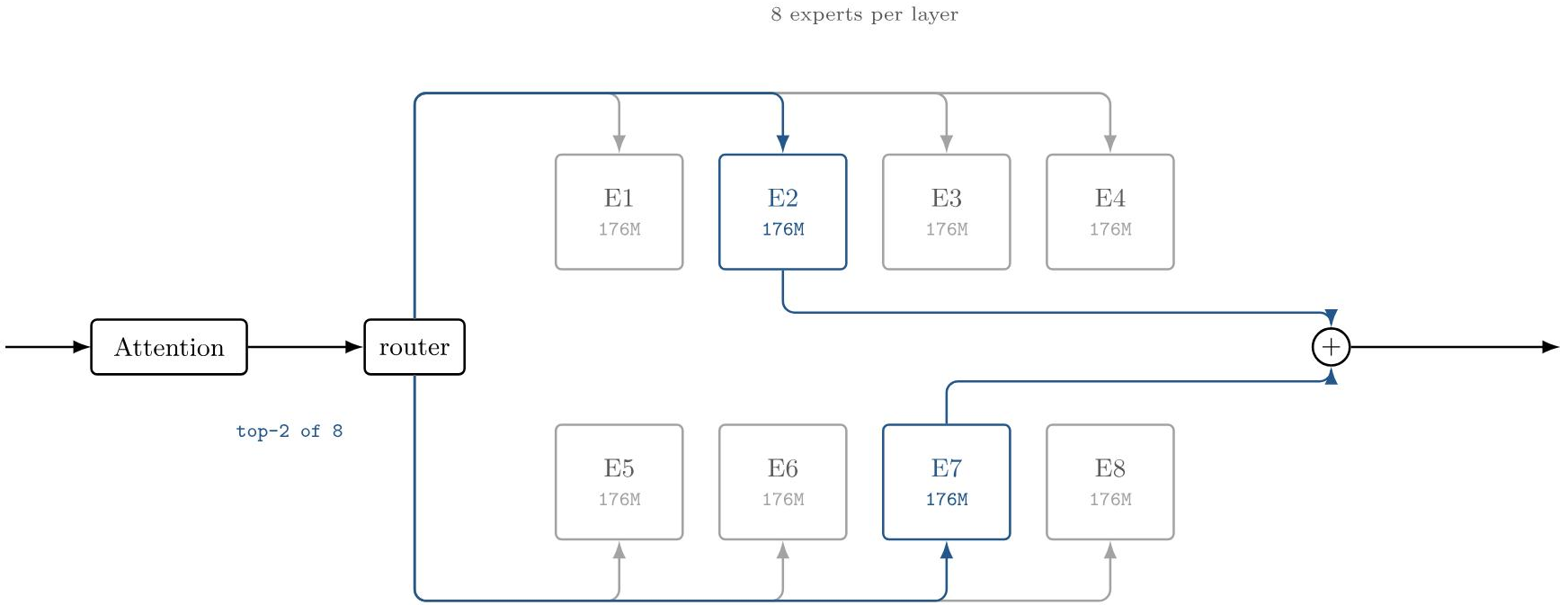
Experts

Where Parameters Live



Computed from each model's config.json: A1 base (Assignment 1 codebase), GPT-2 124M (Radford et al. 2019), Llama-3.1-8B and 70B (Grattafiori et al., arXiv:2407.21783).

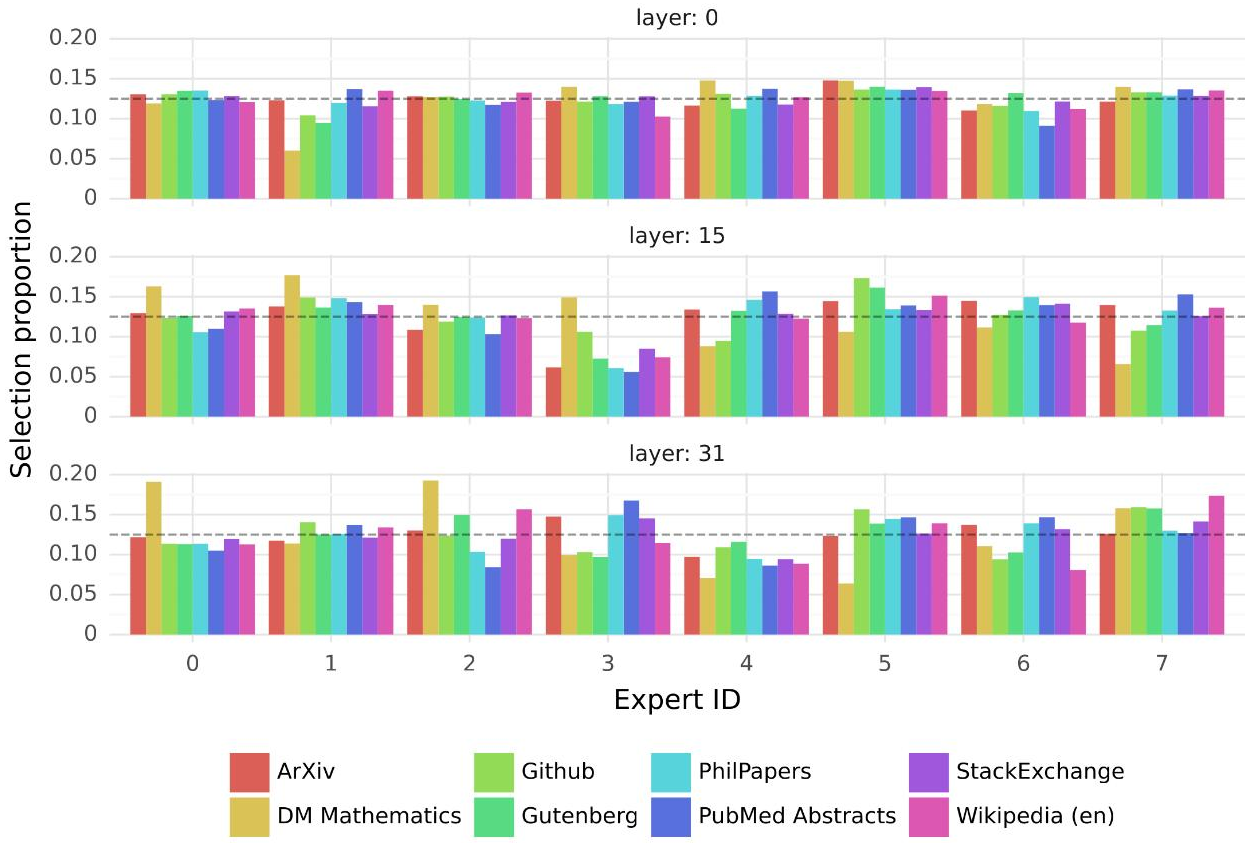
Routing



$$y = \sum_{i \in \text{top-2}} g_i(x)\, E_i(x)$$

Shazeer et al., "Outrageously Large Neural Networks," arXiv:1701.06538 (2017). Fedus et al., "Switch Transformers," arXiv:2101.03961 (2021). Jiang et al., "Mixtral of Experts," arXiv:2401.04088 (2024). Shapes: Mixtral 8x7B.

Keeping Experts Busy



Busy experts get more gradient and get busier. Left alone, routing collapses.

$$\mathcal{L}_{\text{bal}} = \alpha \, E \sum_{i=1}^{E} f_i \, P_i$$

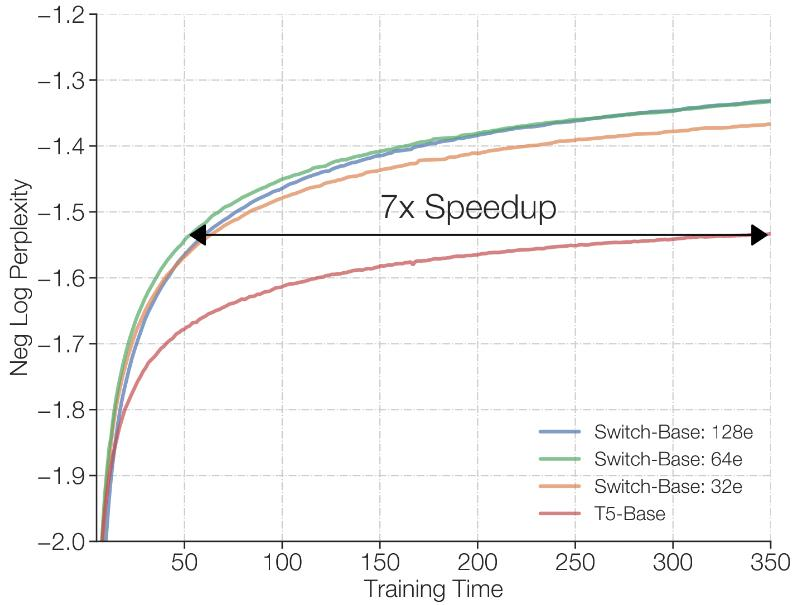
The figure is Mixtral trained with this loss on: selection hugs the dashed 1/8 line.

DeepSeek-V3 drops the loss: a per-expert bias nudges routing instead.

Jiang et al., "Mixtral of Experts," arXiv:2401.04088 (2024). Loss form: Fedus et al., arXiv:2101.03961, Eq. 4. Wang et al., "Auxiliary-Loss-Free Load Balancing," arXiv:2408.15664; DeepSeek-V3, arXiv:2412.19437.

Three Generations

	year	routing	total	active
Switch-C	2021	top-1 of 2048	1.57T	1.4B
Mixtral 8x7B	2023	top-2 of 8	46.7B	12.9B
DeepSeek-V3	2024	top-8 of 256, +1 shared	671B	37B

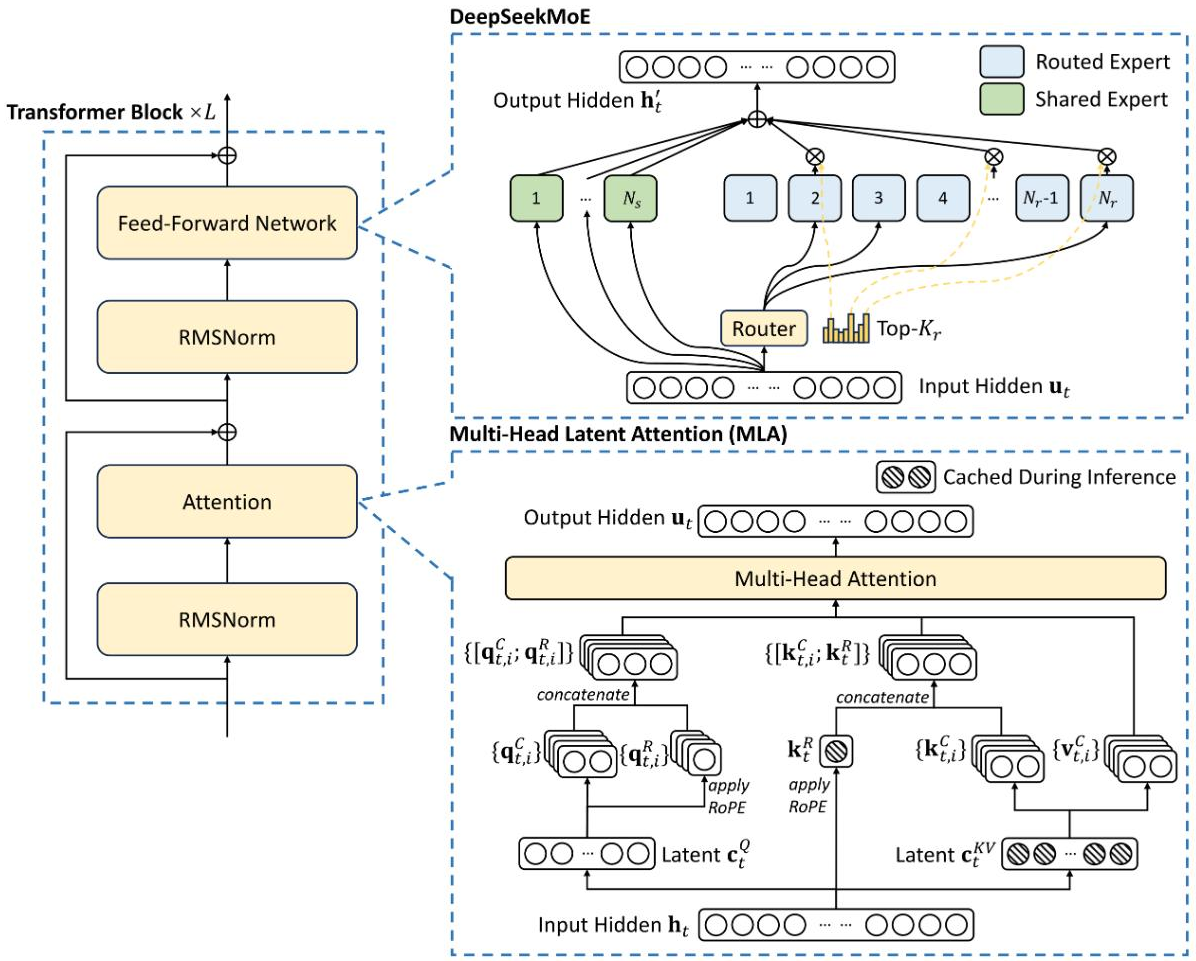


Total parameters set the memory. Active parameters set the speed.

active = attention + embeddings + the experts one token actually visits

Fedus et al., "Switch Transformers," arXiv:2101.03961 (2021), Figure 5 and Table 9; Switch-C active params computed from its Table 9 shapes. Jiang et al., arXiv:2401.04088. DeepSeek-AI, arXiv:2412.19437.

DeepSeek-V3



61 layers · $d = 7168$

MLA: 576 cached values per layer

256 routed + 1 shared expert, top-8

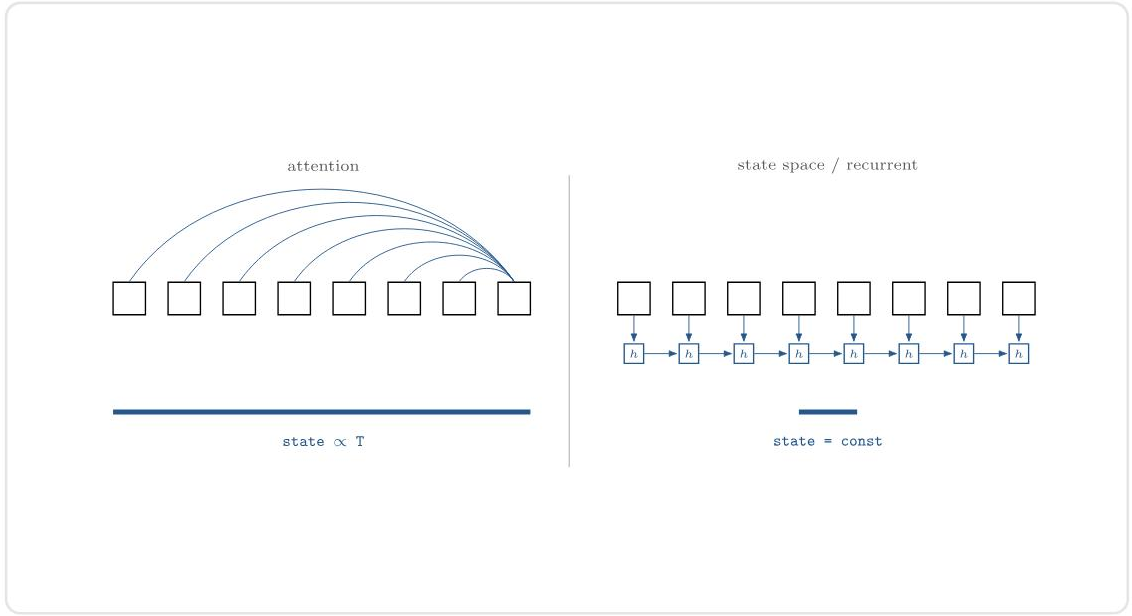
FP8 training

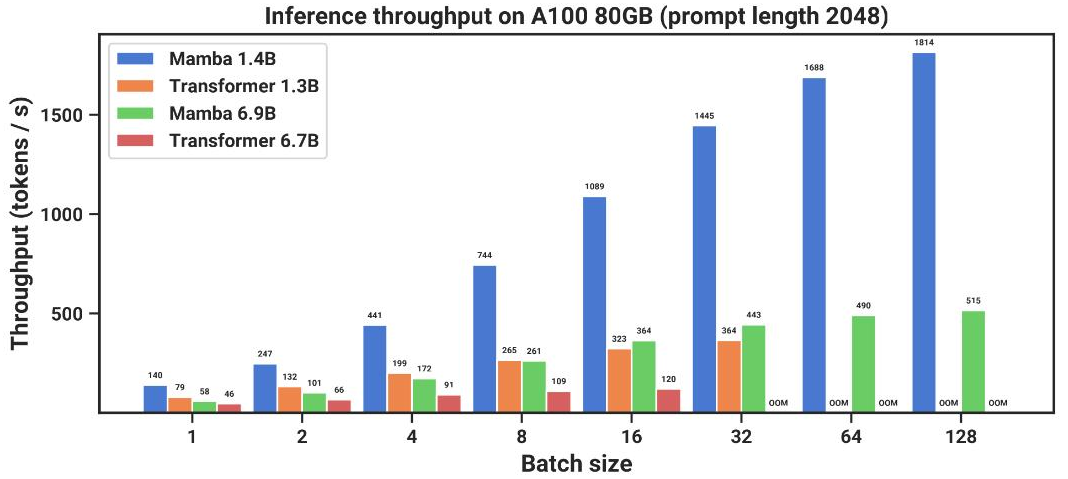
14.8T tokens

2.79M H800-hours for the full run

DeepSeek-AI, "DeepSeek-V3 Technical Report," arXiv:2412.19437 (2024), Figure 2 and Tables 1-2. Shared + fine-grained experts: Dai et al., "DeepSeekMoE," arXiv:2401.06066 (2024). Cost figure covers the final training run at rental prices; ablations and research runs excluded.

Constant State





AI21 Jamba interleaves one attention layer per seven Mamba layers.

Gu and Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," arXiv:2312.00752 (2023). Lieber et al., "Jamba: A Hybrid Transformer-Mamba Language Model," arXiv:2403.19887 (2024).

7 SHOWDOWN

Showdown

Spec Sheets

Card A · Qwen2.5-14B

dense, 14.7B

48 layers, GQA 8 kv heads

active per token: 14.7B

fp16 weights: 29 GB

Card B · Mixtral 8x7B

MoE, 46.7B total

32 layers, top-2 of 8, GQA 8 kv heads

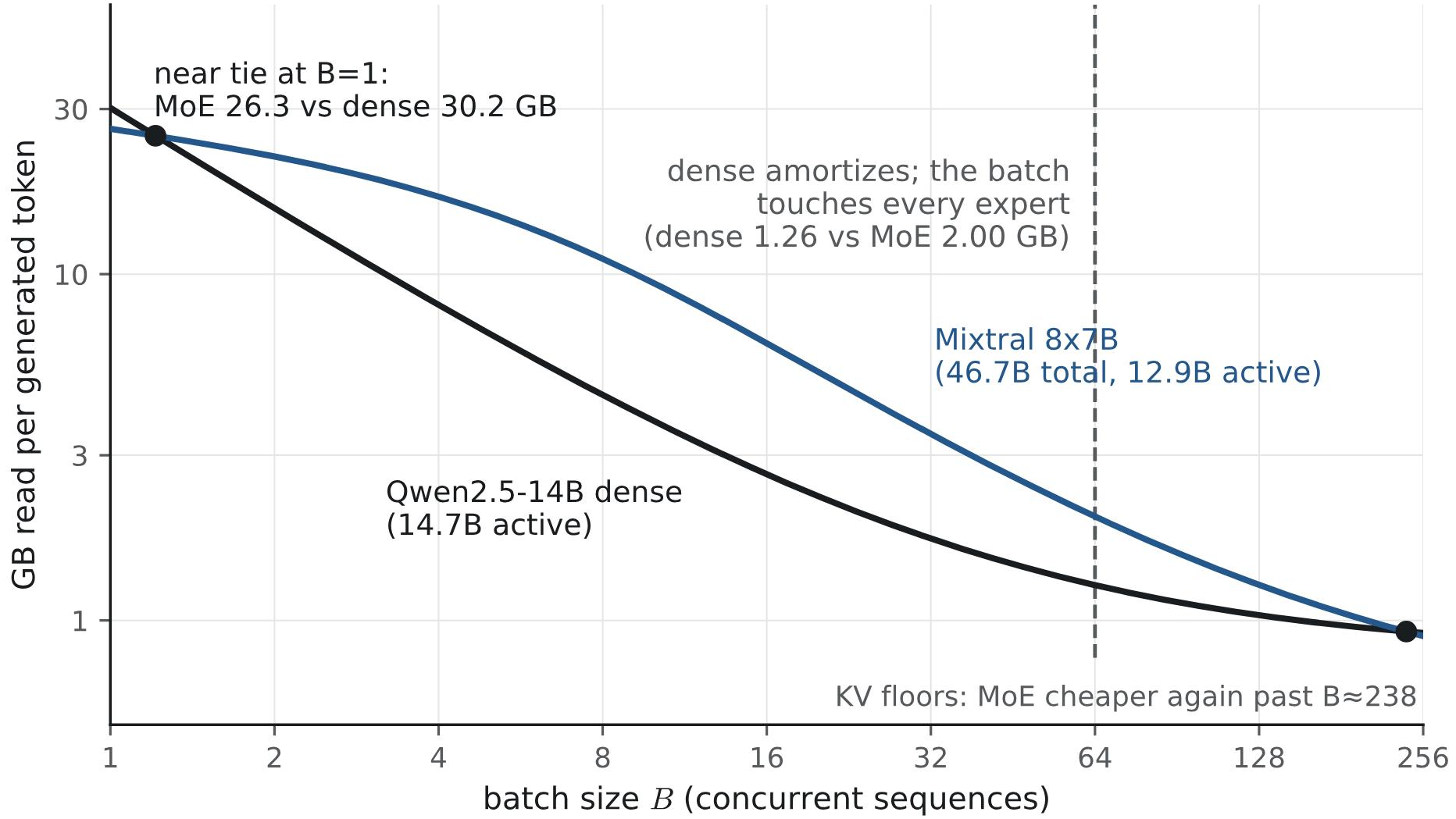
active per token: 12.9B

fp16 weights: 93 GB

Same FLOPs per token. Which decodes cheaper?

	batch 1	batch 64
your prediction	A / B / tie	A / B / tie

The Reveal



Arithmetic from each model's config.json shapes: weights read once per step amortized over the batch, plus each sequence's own KV cache; context 4,096, fp16, uniform top-2 routing. Serving engines change the constants, not the shape.

Reading

Required. DeepSeek-V3 Technical Report, architecture sections. Ainslie et al., GQA.

Optional. Shazeer 2019 · Switch 2021 · RoPE 2021 · Mamba 2023 · Mixtral 2024

Assignment 1 is out: build the block we counted today.

the causality test catches the classic mask bugs; Part 3 is the live fit you saw

Tuesday: Mark. Thursday: Evaluation, and Assignment 2.

bring one benchmark you trust and one reason to distrust it