

Optimal Sparsity of Mixture-of-Experts Language Models for Reasoning Tasks



Institute of Science Tokyo
Rio Yokota

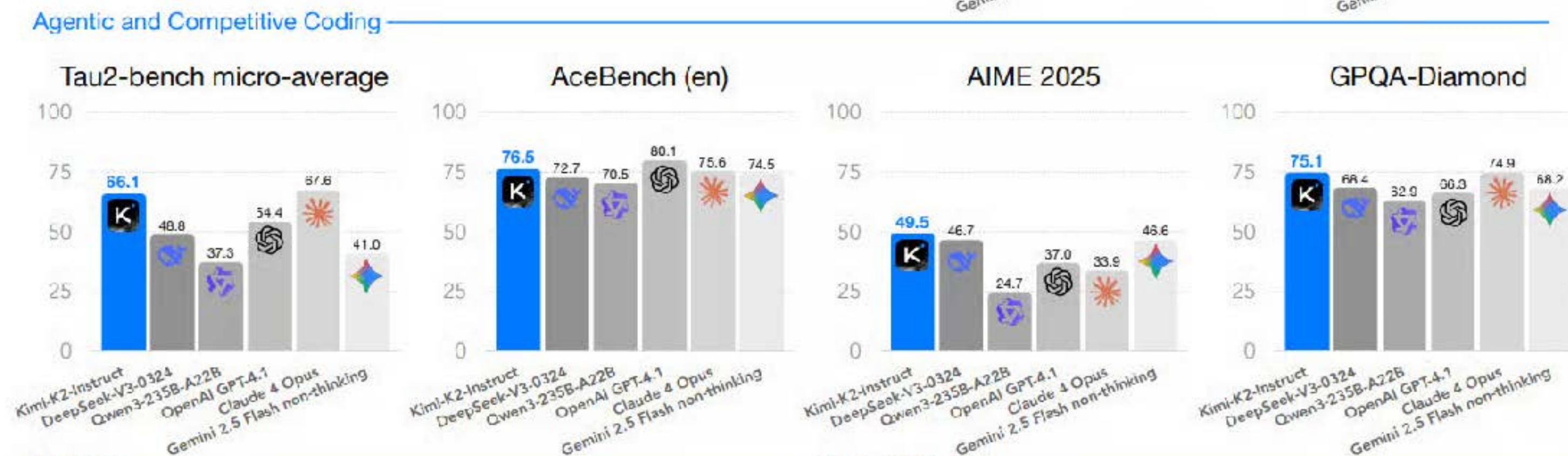
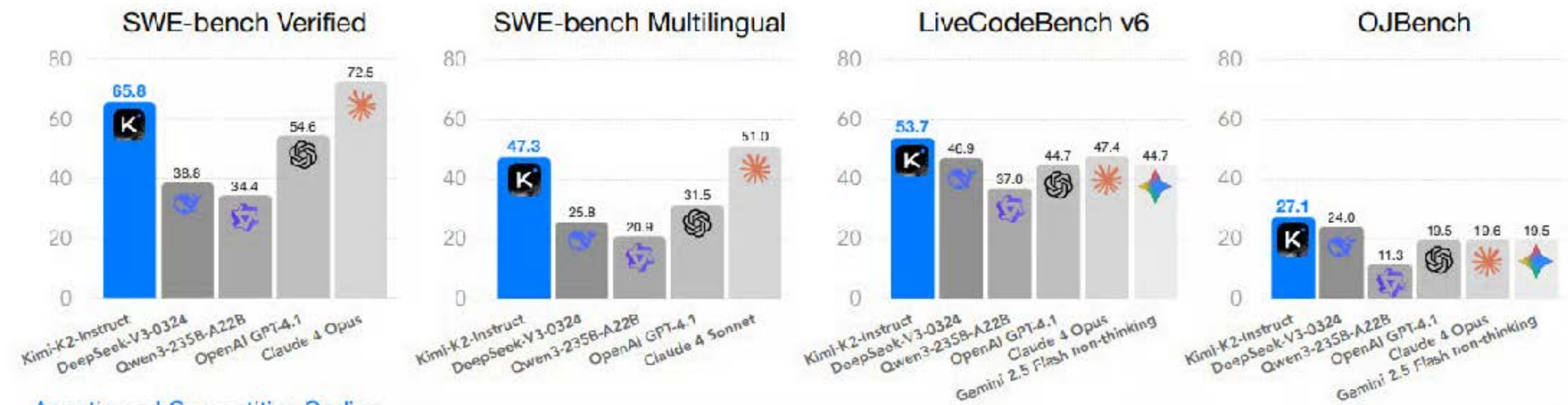
Japanese Symposium on Open Large Language Models
Hitotsubashi Hall
Nov. 25, 2025

Mixture of Experts

Recent open models are:

- Very sparse MoEs
- Trained for test-time/agentic scaling
- Eval'd on math / coding / tool use
- Competitive with frontier models

Model	Total Parameters	Active Parameters	Experts (Active)	Context Length
Qwen3-Next-80B-A3B	80B	3B	512+1 (10)	262k
Kimi-K2	1T	32B	384+1 (8)	128k
GLM-4.5	355B	32B	256+1 (8)	128k
DeepSeek-V3	671B	37B	256+1 (8)	128k
gpt-oss-120b	117B	5.1B	128 (4)	128k



Math & STEM

Scaling Laws of MoEs

Parameters vs FLOPs: Scaling Laws for Optimal Sparsity for Mixture-of-Experts Language Models

Samira Abnar* Apple
 Harshay Shah* MIT
 Dan Busbridge Apple
 Alaaeldin El-Nouby Apple
 Josh Susskind Apple
 Vimal Thilak* Apple

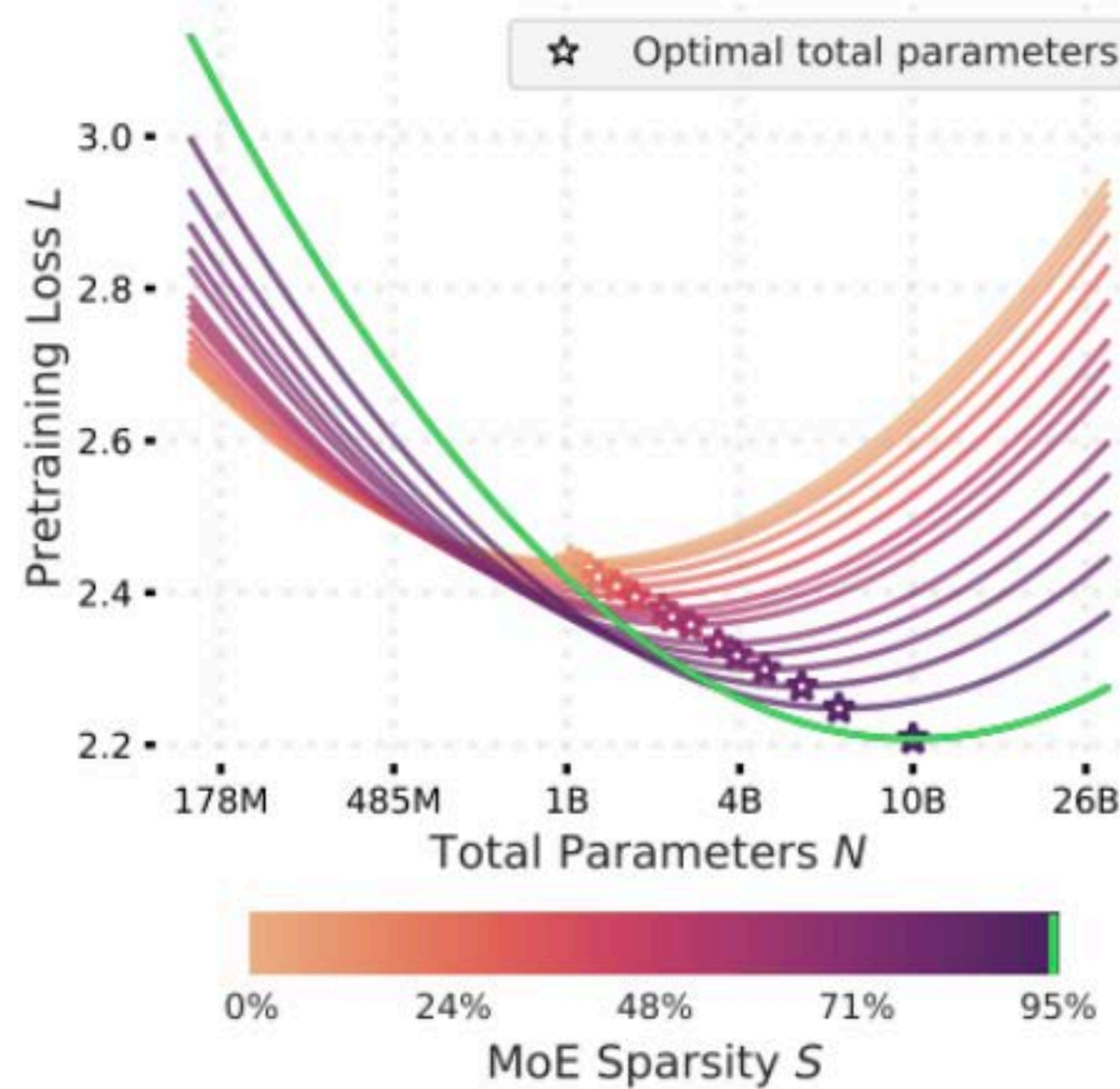
<https://arxiv.org/abs/2501.12370>

Optimal sparsity increases with the total parameters
 → Larger models should be more sparse

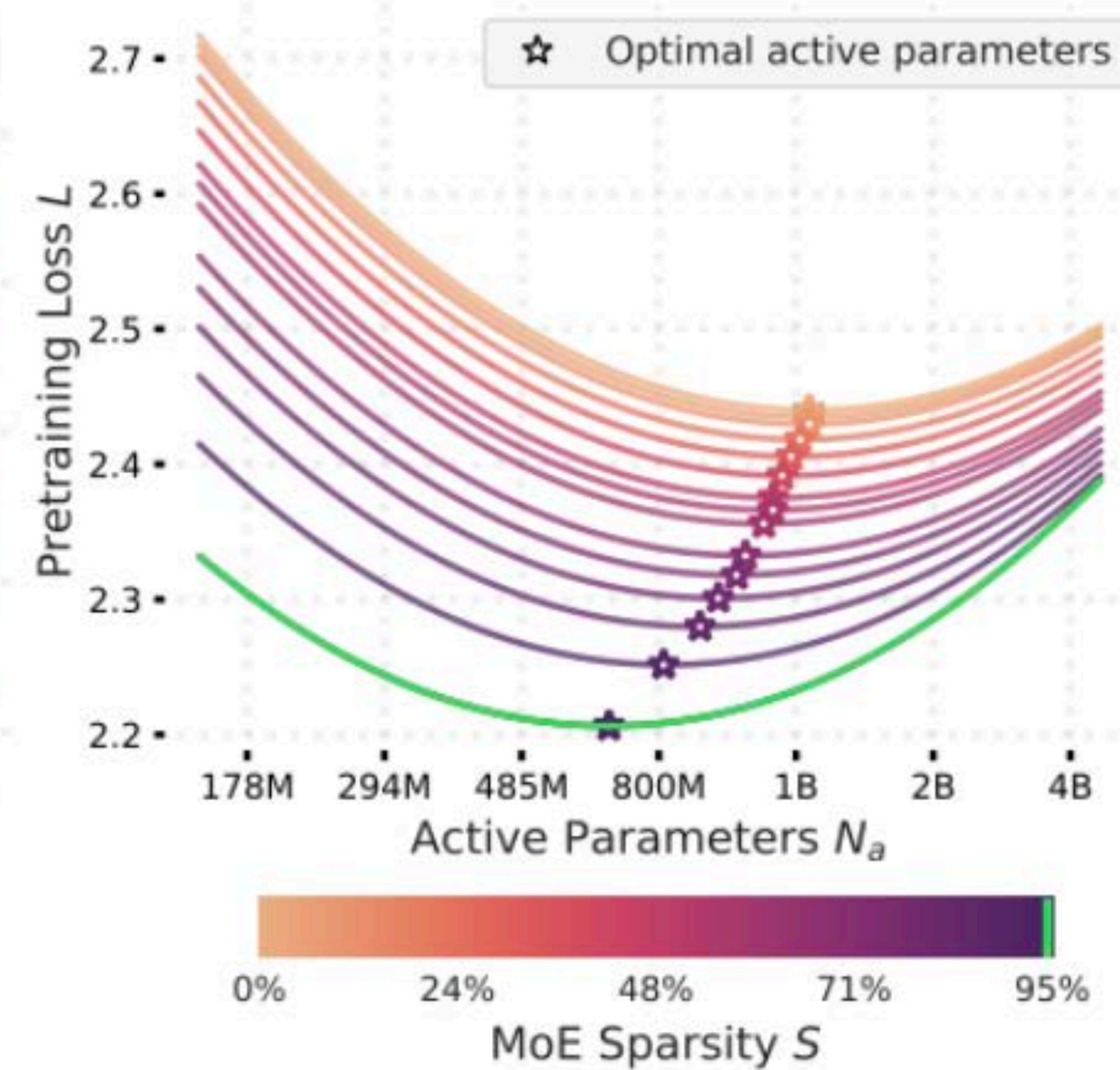
Optimal sparsity mildly decreases with the active parameters

The most sparse model achieves the lowest loss regardless of model size

Fixing Sparsity
 Varying Total Parameters



Fixing Sparsity
 Varying Active Parameters



S: Sparsity

E: Number of experts

K: Number of activated experts

$$S = \frac{E - K}{E}$$

Memorization vs Reasoning

Fixing active parameters and increasing total parameters
→ Fixing the cost and increasing model capacity

For memorization tasks

→ Accuracy increases with the total parameters

For reasoning tasks

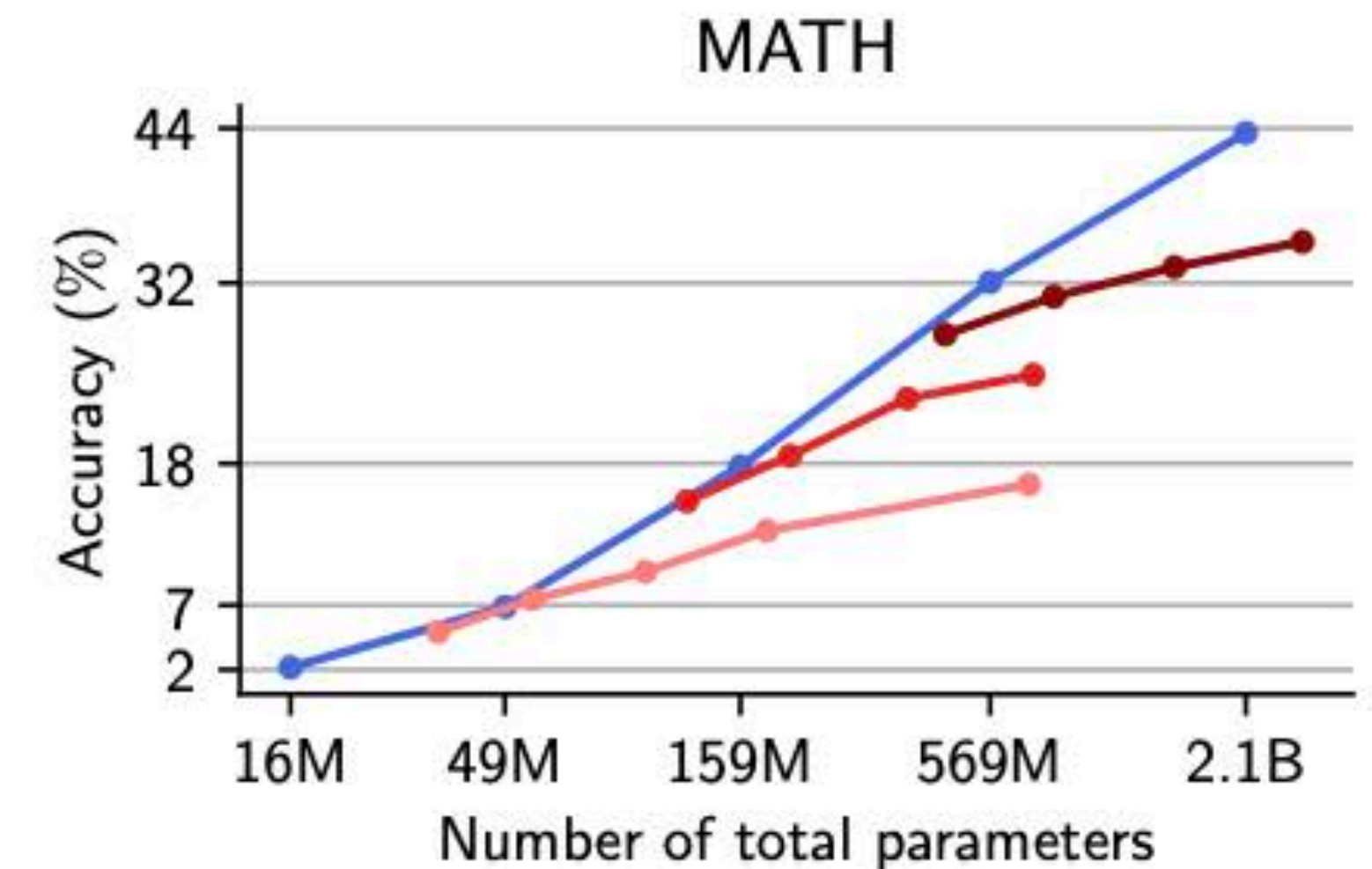
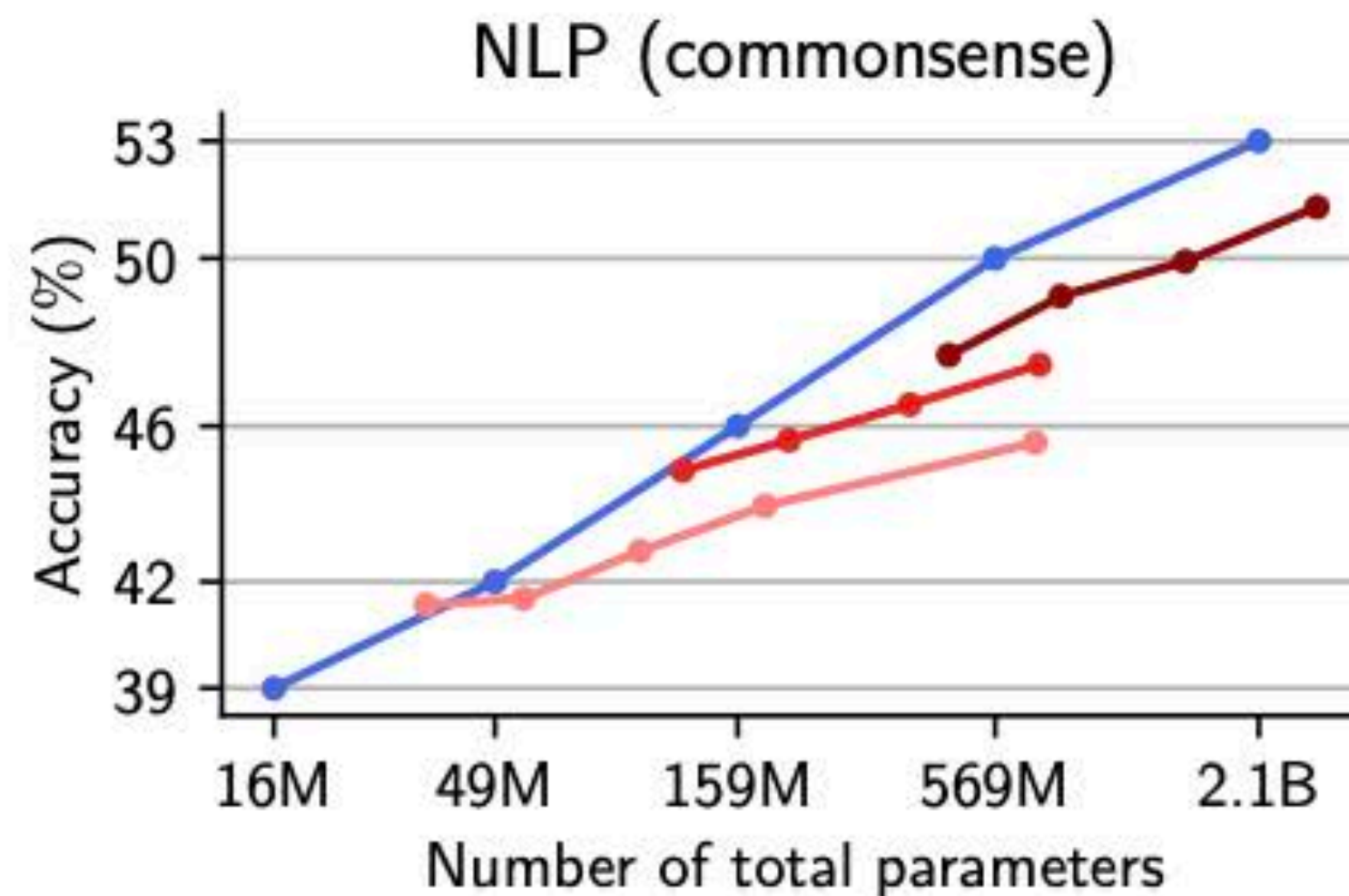
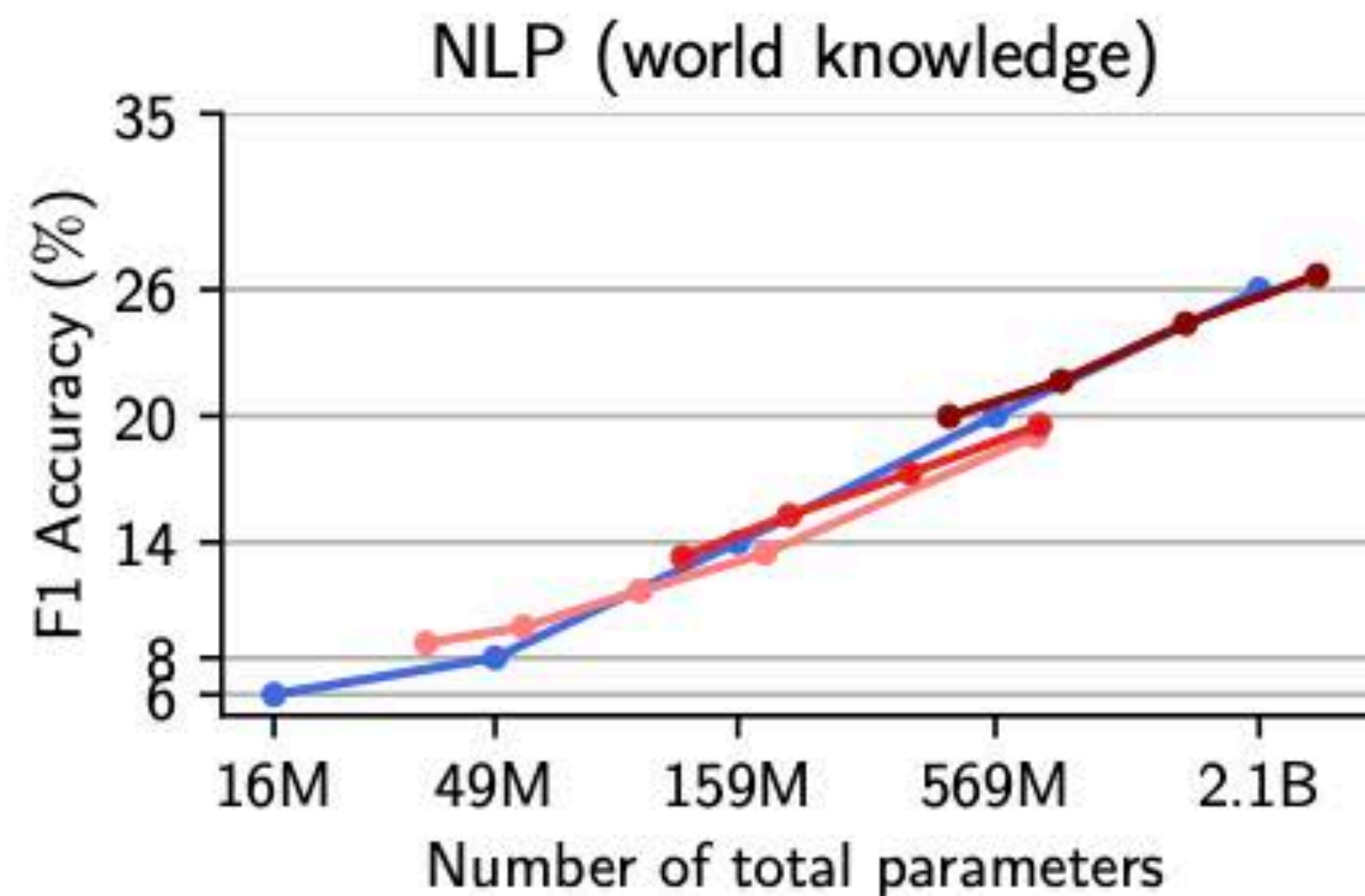
→ Accuracy does not increase with total parameters

→ Sparsity is not good for reasoning?

MIXTURE OF PARROTS 🦜🦜🦜: EXPERTS IMPROVE MEMORIZATION MORE THAN REASONING

Samy Jelassi* Harvard University	Clara Mohri Harvard University	David Brandfonbrener Harvard University Kempner Institute	Alex Gu MIT
Nikhil Vyas Harvard University	Nikhil Anand Harvard University Kempner Institute	David Alvarez-Melis Harvard University Kempner Institute	Yuanzhi Li Microsoft Research
Sham M. Kakade Harvard University Kempner Institute		Eran Malach Harvard University Kempner Institute	

<https://arxiv.org/abs/2410.19034>



■ Dense transformer ■ MoE (18M active parameters) ■ MoE (58M active parameters) ■ MoE (200M active parameters)

Experimental Setup

Optimal Sparsity of Mixture-of-Experts Language Models for Reasoning Tasks

Taishi Nakamura^{1,2} Satoki Ishikawa¹ Masaki Kawamura¹ Takumi Okamoto^{1,2} Daisuke Nohara¹
Jun Suzuki^{3,2,4} Rio Yokota^{1,2}

Model architecture

Base model: Mixtral [Jiang et al., 2024]

Normalization: RMSNorm [Zhang & Sennrich, 2019]

Activation: SwiGLU [Shazeer, 2020]

Positional embedding: RoPE [Su et al., 2024]

Router: Drop-less token-choice top-k routing [Gale et al., 2023]

Architectural hyperparameters

Model width: $d = \{512, 1024, 2048\}$

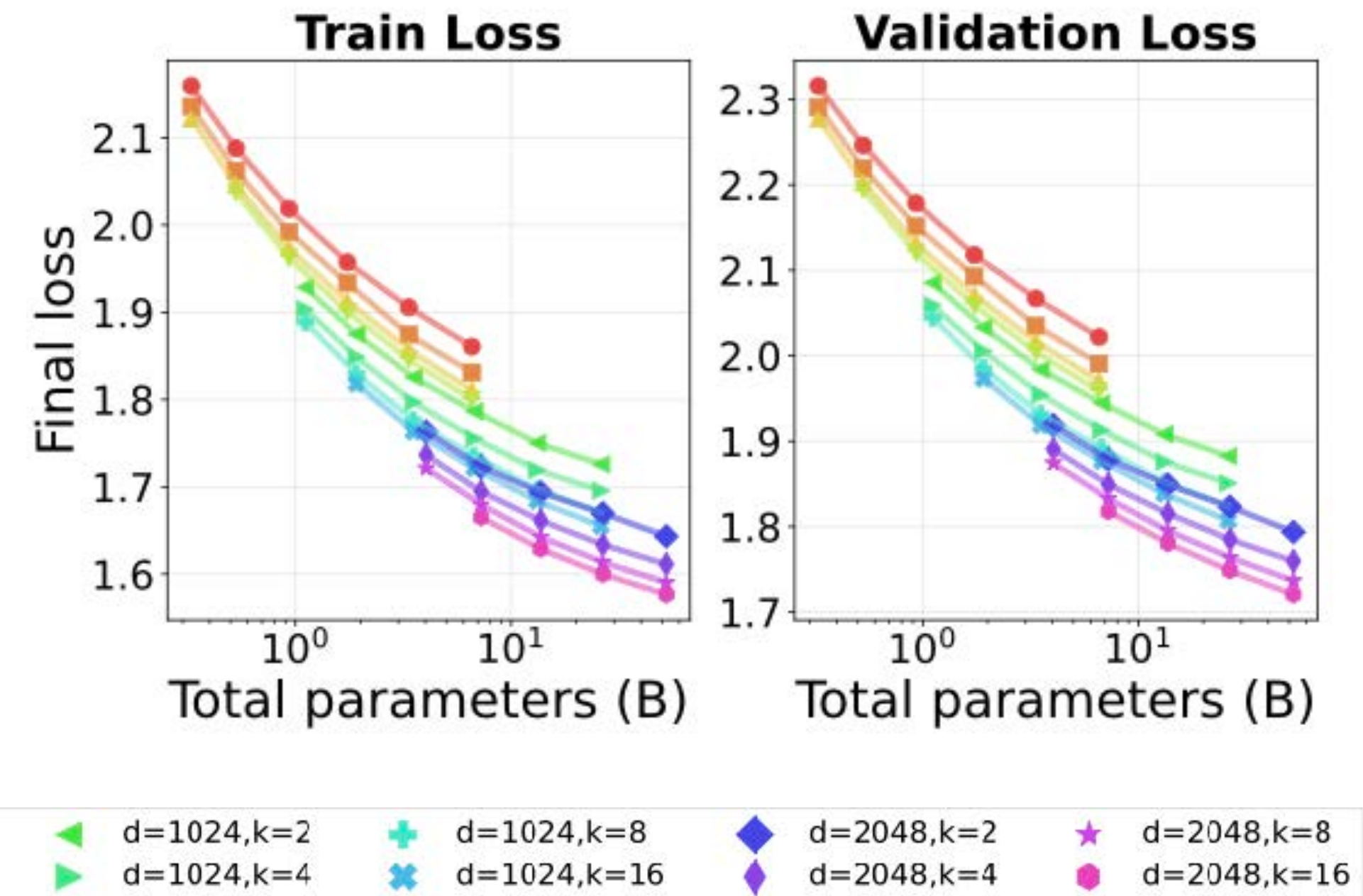
Number of experts: $E = \{8, 16, 32, 64, 128, 256\}$

Activated experts: $k = \{2, 4, 8, 16\}$

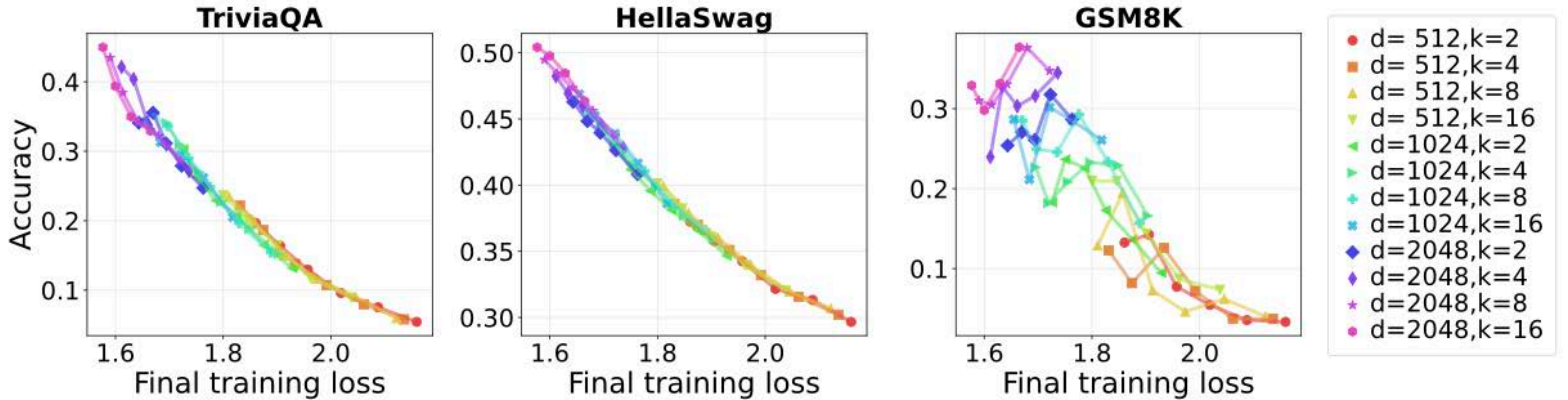
Optimizer: AdamW

Learning rate: peak $4e-4$, 2k linear warmup, cosine decay

Weight decay: 0.1



Downstream Accuracy vs Training Loss

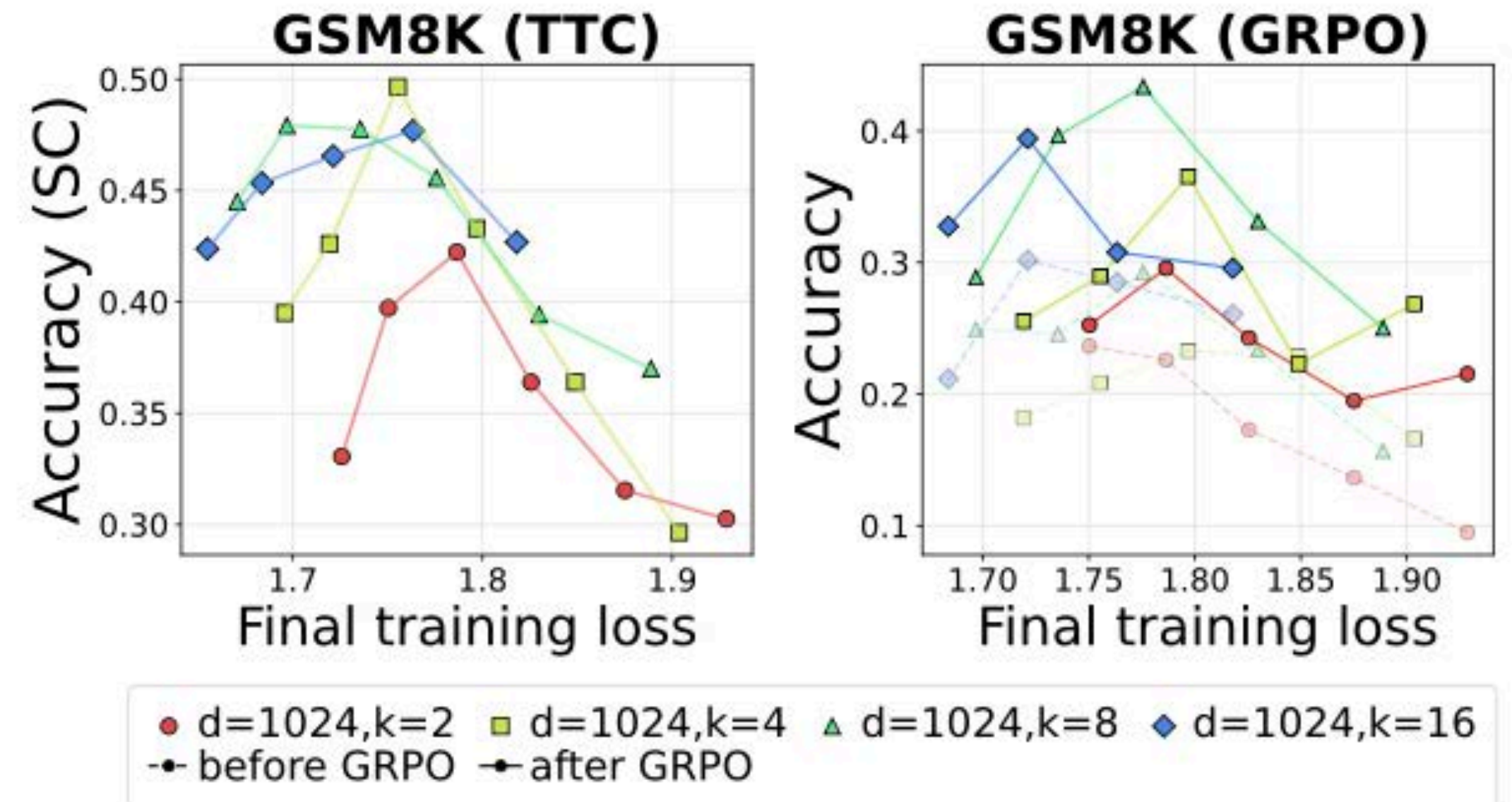


We want the downstream accuracy to inversely scale with the training loss

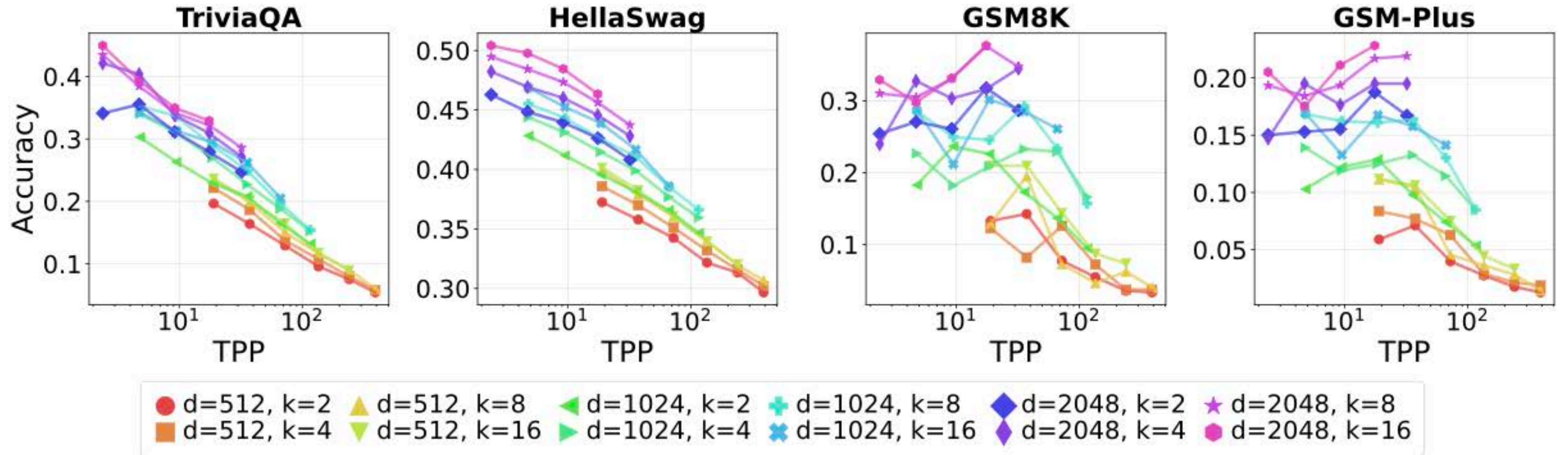
TriviaQA: World knowledge 🙌

Hellaswag: Commonsense reasoning 🙌

GSM8K, GSM-Plus: Mathematical reasoning 🤔



Downstream Accuracy vs Tokens per Parameter



Chinchilla scaling law [Hoffman et al., 2022]: 20 tokens per parameter (TPP)

→ Optimal TPP is task dependent [Roberts et al., 2025]

- Memorization benefits from lower TPP
- Reasoning benefits from higher TPP

By plotting against the TPP, we can see a more uniform trend for math reasoning tasks

→ Math reasoning tasks have an optimal TPP regardless of the model size