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

Taishi Nakamura[1,2], Satoki Ishikawa[1], Masaki Kawamura[1], Takumi Okamoto[1,2], Daisuke Nohara[1], Jun Suzuki[3,4,2], Rio Yokota[1,2]

[1]Institute of Science Tokyo, [2]NII LLMC, [3]Tohoku University, [4]RIKEN

Cerebras Seminar (January 29)

About Me

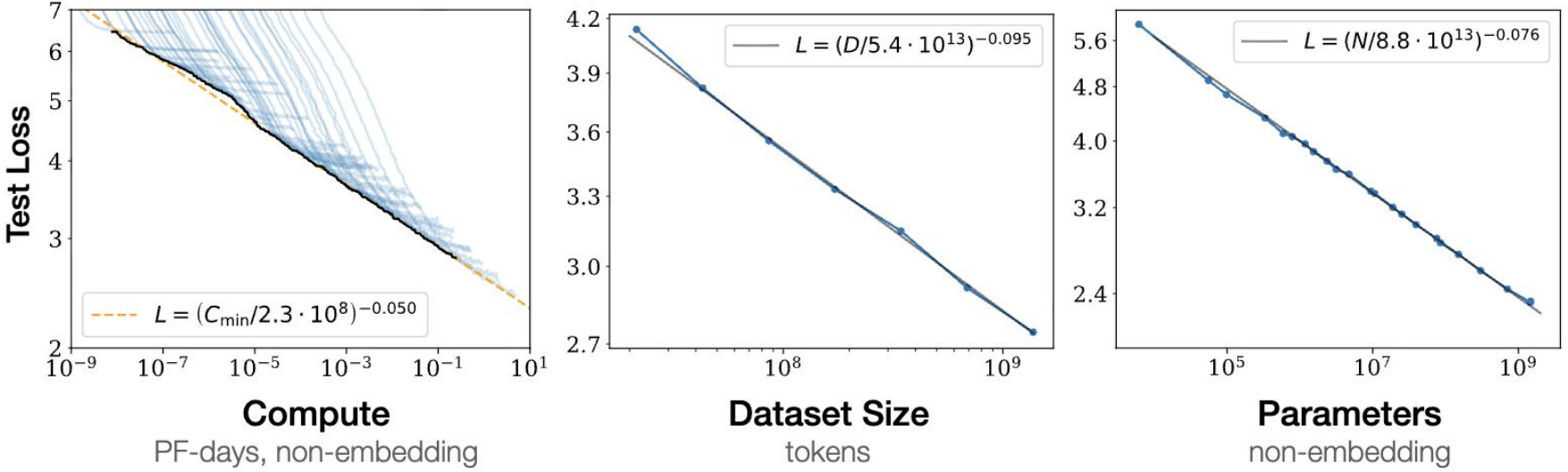
- Taishi Nakamura
- M.S. Student, Institute of Science Tokyo
- Rio Yokota Lab
- Research Assistant at NII LLMC
- Research interests: MoE, reasoning, LLM scaling, efficient training & inference

Outline

- Background & Motivation
- Research Questions
- Experimental Setup
- Main Results
  - Loss vs Reasoning
  - Sparsity & TPP
- Post-training & Test-time Compute
- Conclusion

Background: Scaling Laws of Dense LLMs

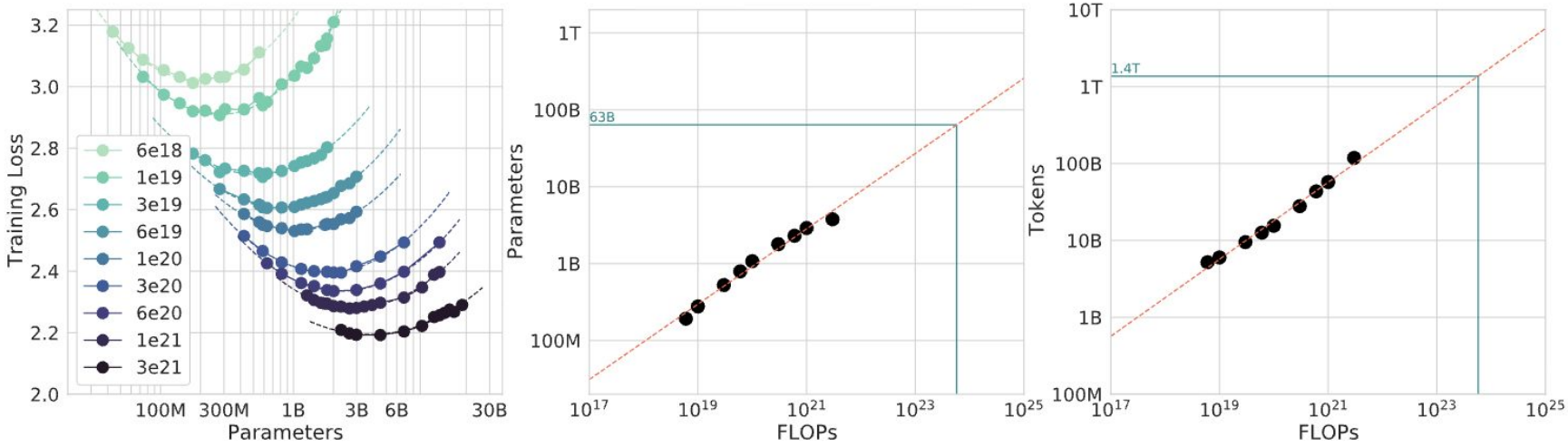
Pretraining loss scales with model size, training tokens, and compute (Kaplan et al., 2020)



Source: Kaplan et al., "Scaling Laws for Neural Language Models" arXiv:2001.08361, Figure 1.

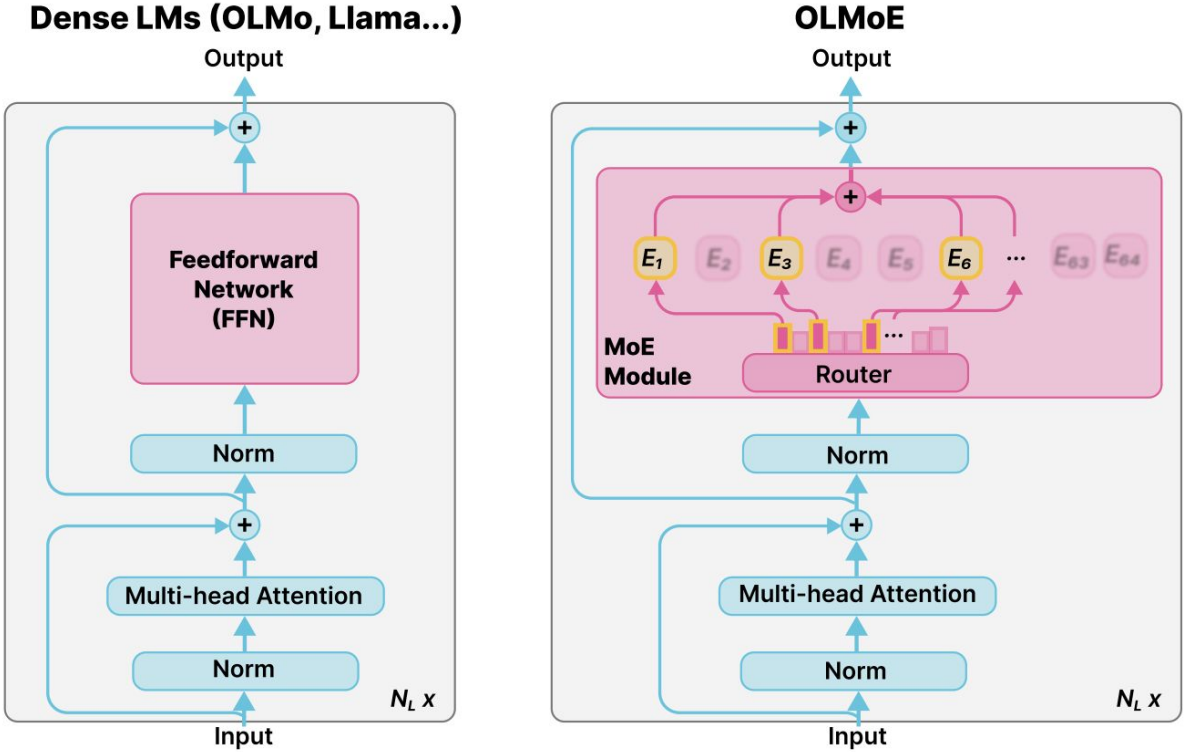
Background: Compute-Optimal Scaling (Dense)

Under a fixed compute budget, model size and training data should be scaled proportionally (Hoffmann et al., NeurIPS 2022)



Source: Hoffmann et al., "Training Compute-Optimal Large Language Models" NeurIPS 2022, Figure 3.

Background: Mixture-of-Experts



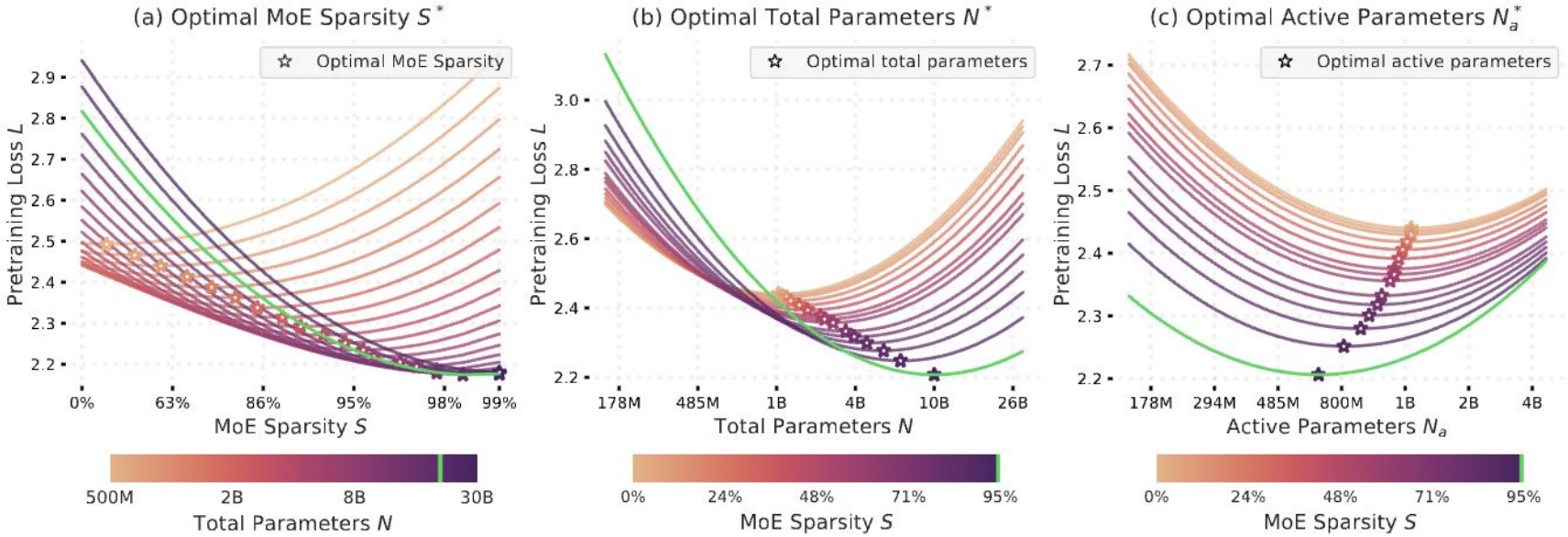
Source: Muennighoff et al., "OLMoE: Open Mixture-of-Experts Language Models" ICLR 2025, Figure C1.

Background: Scaling with Mixture-of-Experts

- MoE enables scaling model capacity
- MoE often achieves better loss scaling efficiency than dense models
  (e.g., Clark et al., ICML 2022; Krajewski et al., ICML 2024)

Background: Sparsity in MoE

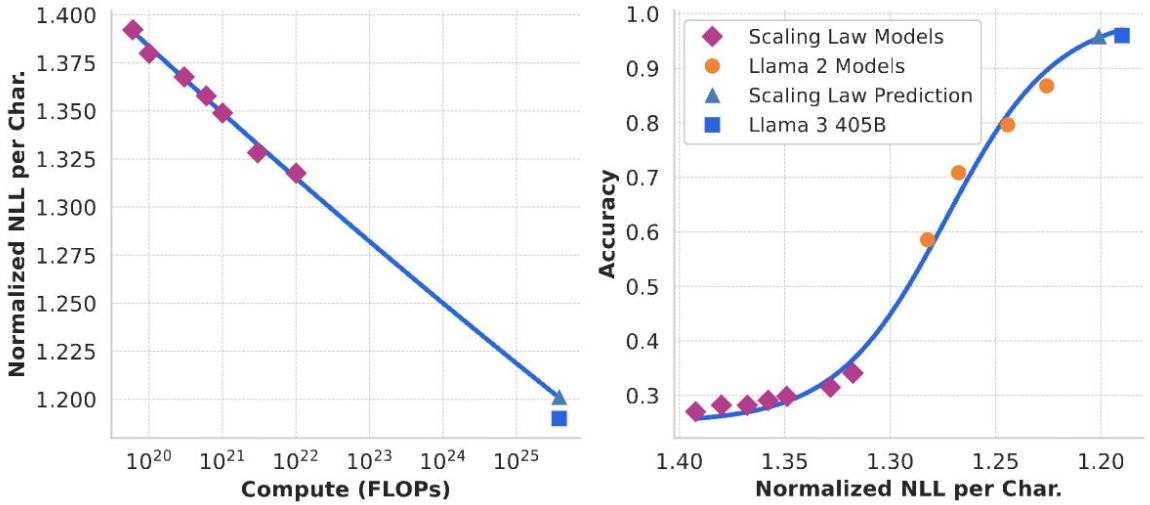
Under fixed compute (IsoFLOP), optimal sparsity increases with model size (Abnar et al., 2025).



Source: Abnar et al., "Parameters vs FLOPs: Scaling Laws for Optimal Sparsity for Mixture-of-Experts Language Models" ICML 2025, Figure 2.

Background: Loss-Based Performance Prediction

- Scaling laws to predict downstream performance from loss
- However, this relationship is task-dependent and unreliable for reasoning (Lourie et al., 2025; Jelassi et al., ICLR 2025)

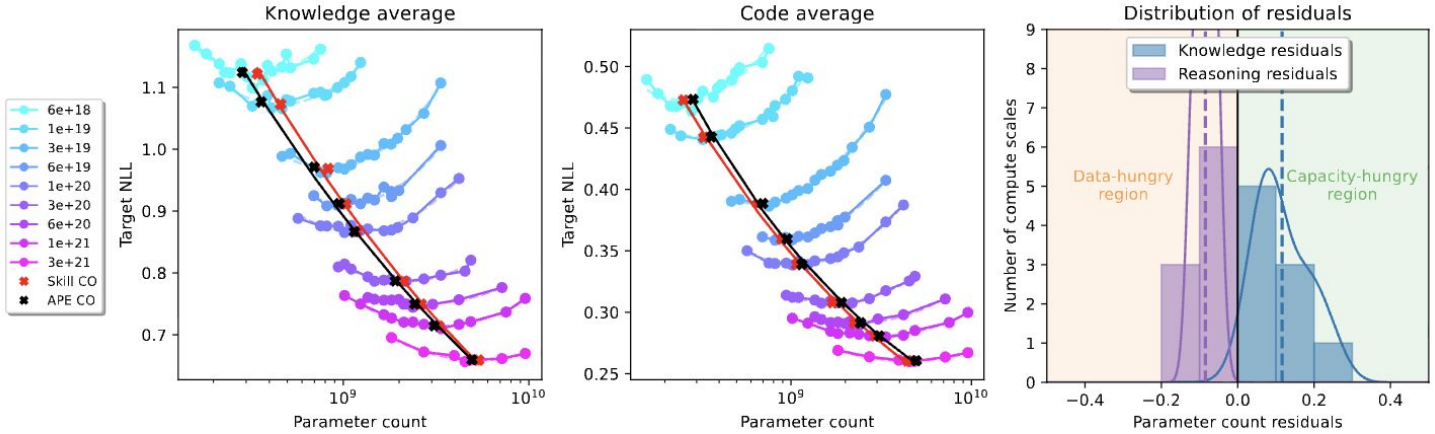


Source: Grattafiori et al., "The Llama 3 Herd of Models" arXiv:2407.21783, Figure 4.

Background: Skill-Dependent Scaling

Scaling behaviour is not uniform across tasks (Roberts et al., 2025)

Reasoning and knowledge exhibit different compute-optimal trends



(a) Knowledge QA. (b) Code. (c) Data- vs capacity-hunger.

Source: Roberts et al., "Compute Optimal Scaling of Skills: Knowledge vs Reasoning" arXiv:2503.10061 Figure 1.

Research Questions

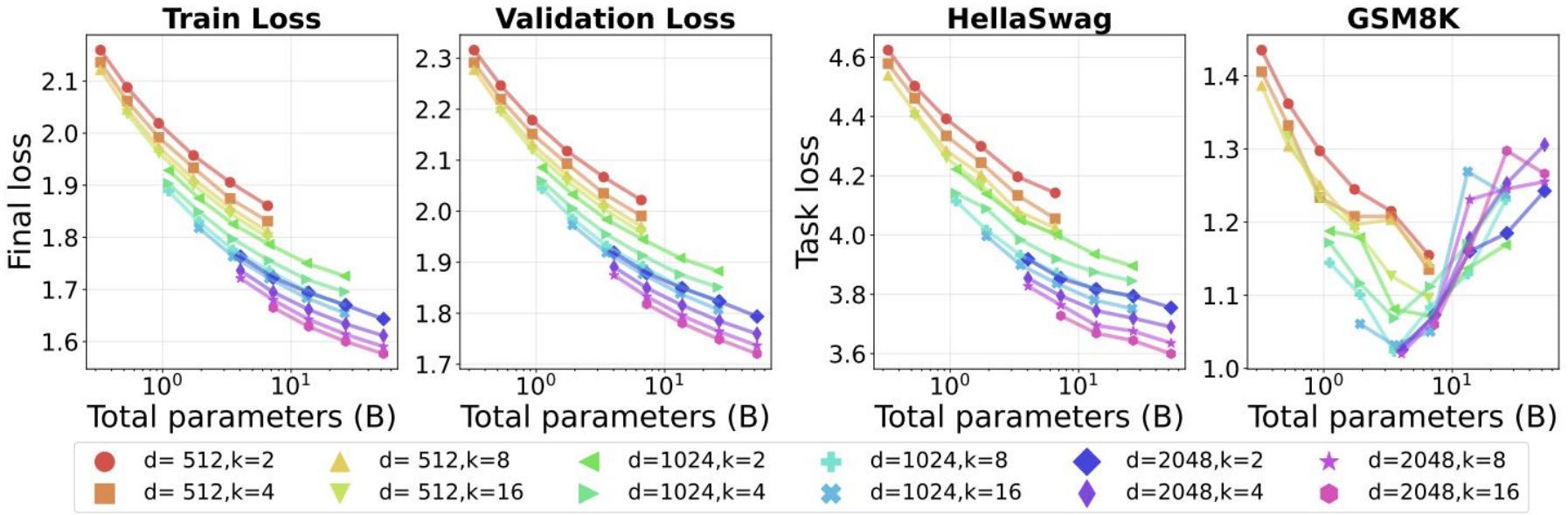
- What is the compute-optimal sparsity for MoE models on reasoning tasks, and how do total and active parameter counts relate to reasoning performance?
- Can common techniques improve overly sparse MoE models?
  - Test-time compute (Self-Consistency)
  - RL post-training (GRPO)

Main Experimental Setup

- Model
  - 16-layer Mixtral-style Transformer
- Swept Dimensions
  - Model width $d \in \{512, 1024, 2048\}$
  - Number of experts $E \in \{8, 16, 32, 64, 128, 256\}$ ($E \leq 128$ for $d = 2048$)
  - Top-k $\in \{2, 4, 8, 16\}$
- Training
  - AdamW, cosine LR
  - 125B tokens
    - Web (~43B), Math (~32B), STEM + Wikipedia (~49B)
- Evaluation Tasks
  - Memorization: TriviaQA, HellaSwag
  - Reasoning: GSM8K, GSM-Plus

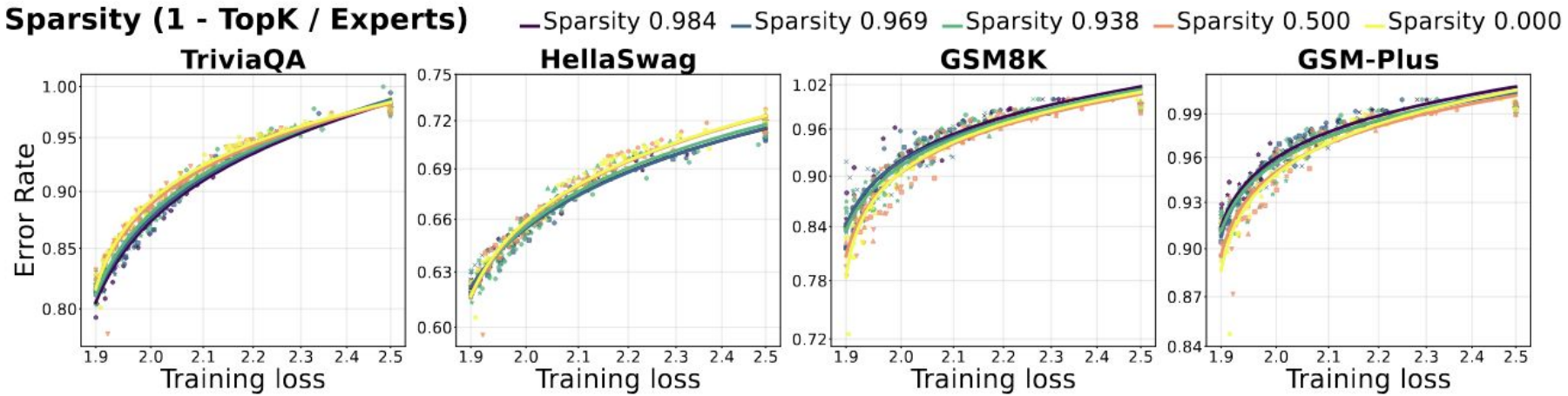
Loss Does Not Predict Task Loss

- Cross-entropy loss decreases monotonically with model size
- But task loss for reasoning is non-monotonic



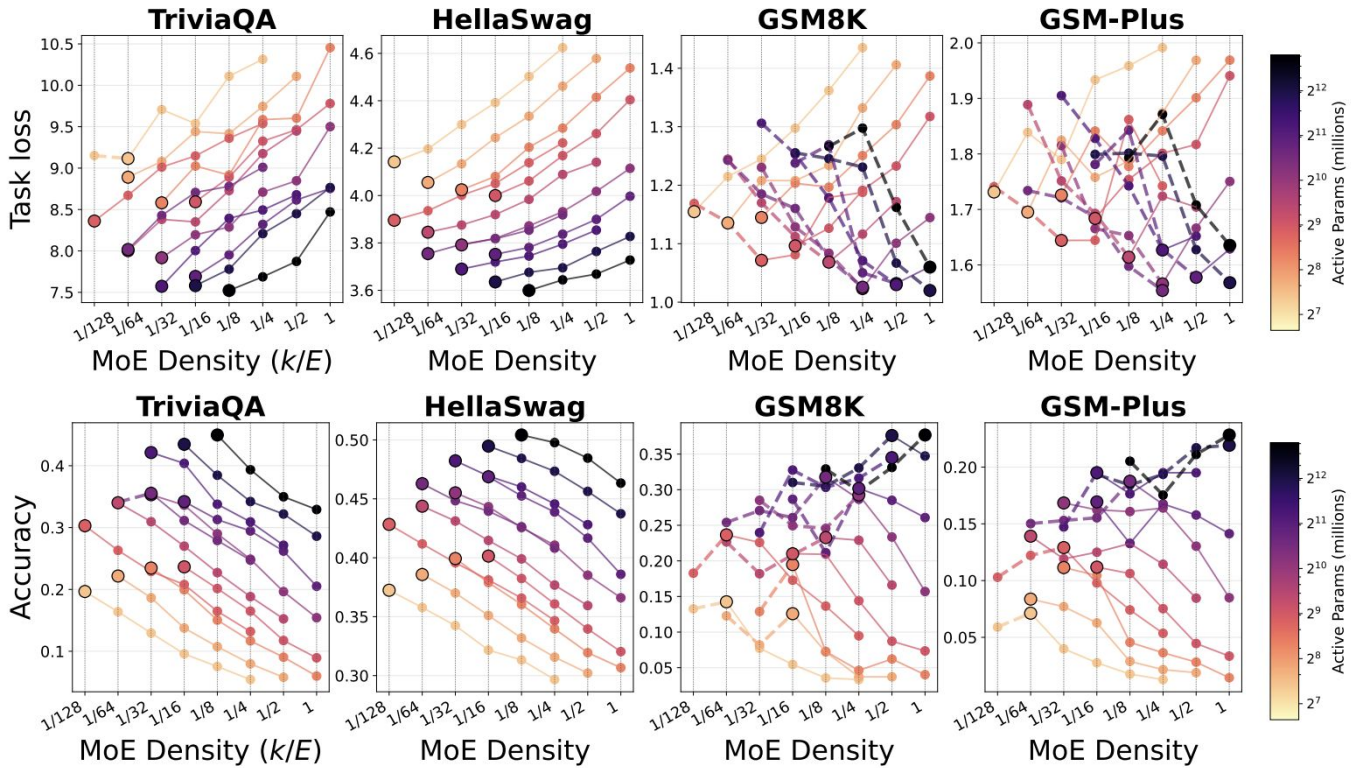
Same Training Loss, Different Reasoning Performance

- For the same cross-entropy loss, higher sparsity leads to worse reasoning performance
- Even at the same loss, dense-leaning MoE is better for reasoning.



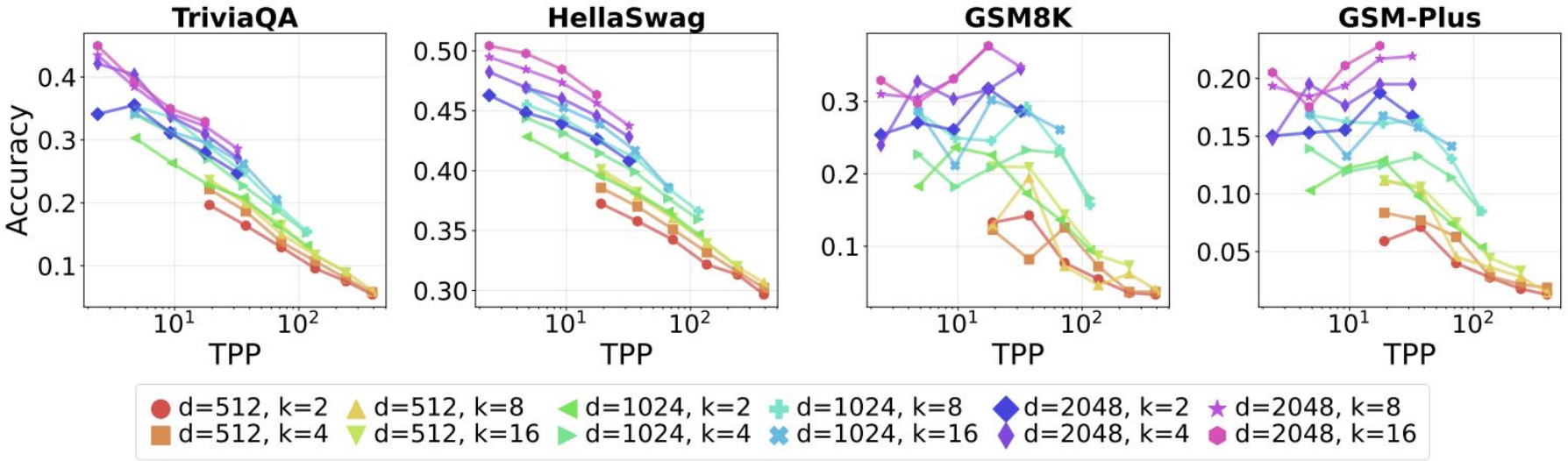
Sparsity Matters for Reasoning in MoE

For reasoning tasks (GSM8K, GSM-Plus), sparsity (1-k/E) is crucial



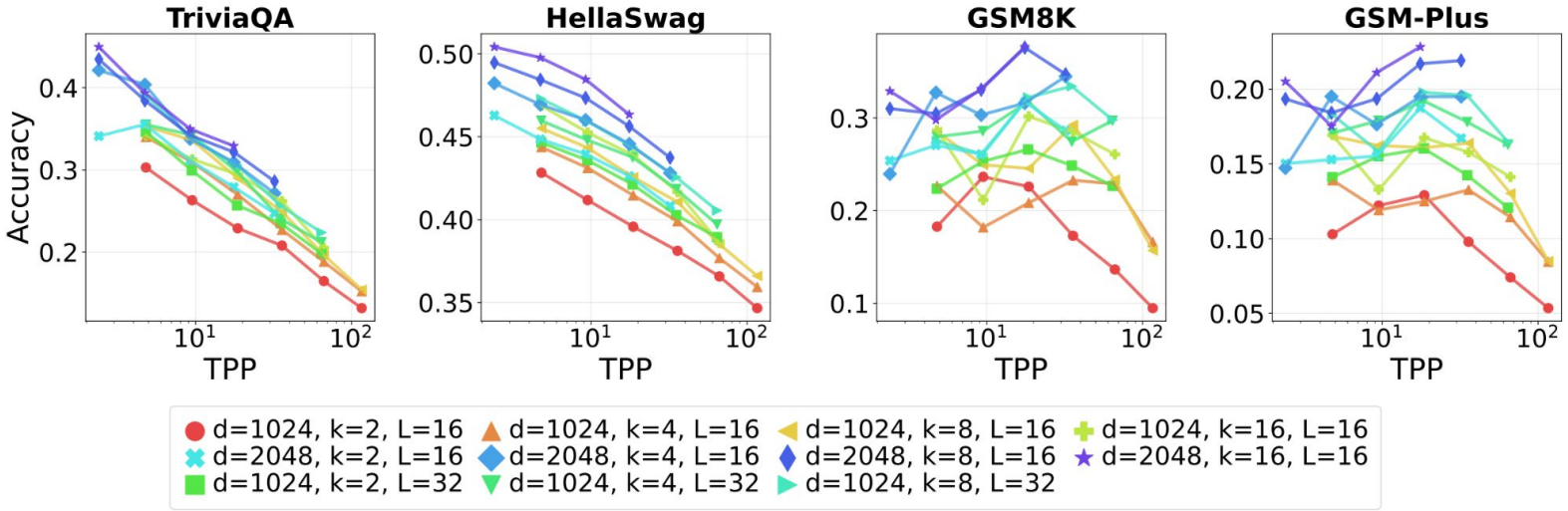
Tokens Per Parameter Matter for Reasoning in MoE

- For reasoning tasks (GSM8K, GSM-Plus), total tokens per parameter (TPP) is crucial
- Top-k also plays an important role



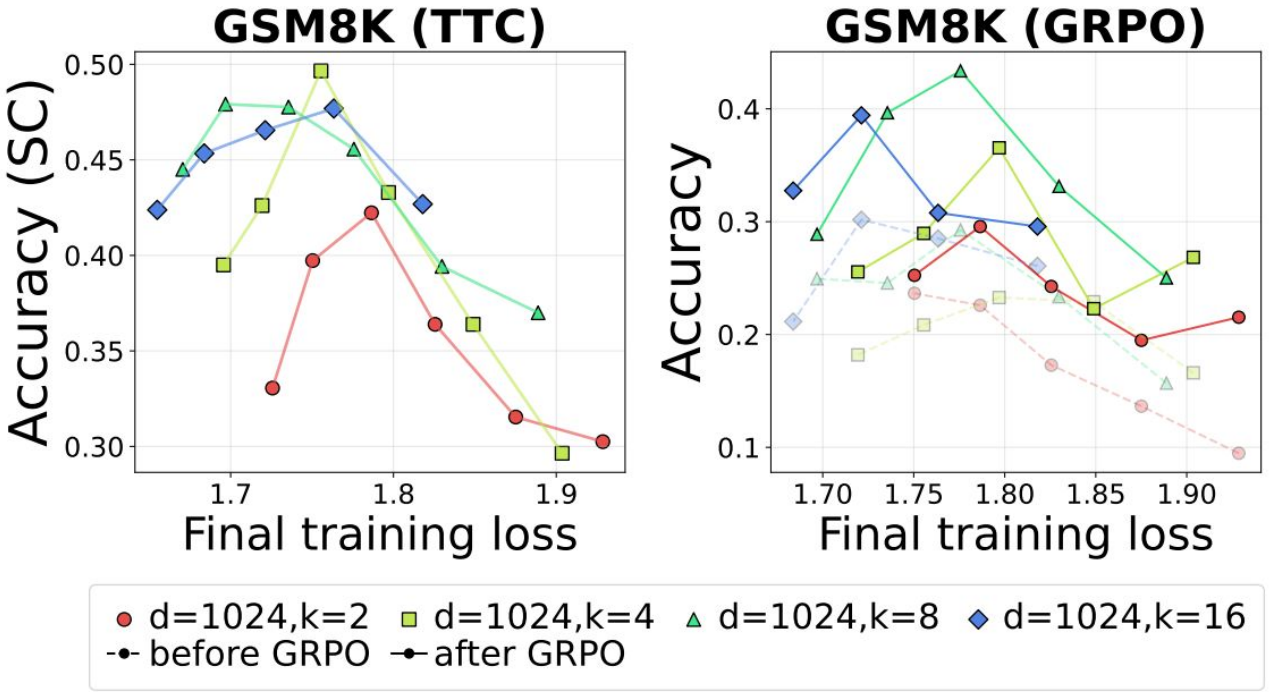
How Model Depth Affects Reasoning?

Adding 32-layer models does not change the U-shaped dependence on TPP



Does Post-Training Fix Overly Sparse MoE Models?

- Test-time compute (Self-Consistency) and GRPO improve accuracy
- But do not eliminate the effect of sparsity



Conclusion

- Loss-based scaling does not characterize reasoning performance in MoE models
- Optimal sparsity exists for reasoning under a fixed compute budget
- **Total tokens per parameter and active parameter allocation are crucial for reasoning**
- Post-training and test-time compute do not change the optimal sparsity trend