Novelty Assessment Report

Paper: Optimal Sparsity of Mixture-of-Experts Language Models for Reasoning Tasks
PDF URL: https://openreview.net/pdf?id=XFw2EPRUUR
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

Empirical scaling laws have driven the evolution of large language models (LLMs), yet their coefficients shift whenever the model architecture or data pipeline changes. Mixture-of-Experts (MoE) models, now standard in state-of-the-art systems, introduce a new sparsity dimension that current dense-model frontiers overlook. We investigate how MoE sparsity influences two distinct capability regimes: memorization skills and reasoning skills. By training MoE families that vary total parameters, active parameters, and top-K routing under fixed compute budgets, we disentangle pre-training loss from downstream accuracy. Our results reveal two principles. First, Active FLOPs: models with identical training loss but greater active compute achieve higher reasoning accuracy. Second, Total tokens per parameter (TPP): memorization tasks improve with more parameters, while reasoning tasks benefit from optimal TPP, indicating that reasoning is data-hungry. Neither reinforcement learning post-training (GRPO) nor increased test-time compute alters these trends. We therefore argue that optimal MoE sparsity must be determined jointly by active FLOPs and TPP, revising the classical picture of compute-optimal scaling. All code, data sources, and logs are released to facilitate reproducibility and future work.

Core Task Landscape

This paper addresses: **Optimal Sparsity Selection in Mixture-of-Experts Language Models**
A total of **50 papers** were analyzed and organized into a taxonomy with **22 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Scaling Laws and Compute-Optimal Design**
- **Routing Strategies and Expert Selection**
- **Model Compression and Sparsification**
- **Architecture Design and Model Construction**
- **System Optimization and Deployment**
- **Empirical Studies and Open Models**
- **Domain-Specific Applications**

Complete Taxonomy Tree

- Optimal Sparsity Selection in Mixture-of-Experts Language Models Survey Taxonomy
- Scaling Laws and Compute-Optimal Design
  - Parameter-FLOPs Trade-offs and Scaling Principles ★ (3 papers)
  - [0] Optimal Sparsity of Mixture-of-Experts Language Models for Reasoning Tasks (Anon et al., 2026) View paper
  - [2] Toward inference-optimal mixture-of-expert large language models (Yun, 2024) View paper
  - [3] Parameters vs flops: Scaling laws for optimal sparsity for mixture-of-experts language models (Abnar, 2025) View paper
  - Sparsity and Superposition Mechanisms (1 papers)
  - [21] Sparsity and Superposition in Mixture of Experts (Marmik Chaudhari, 2025) View paper
- Routing Strategies and Expert Selection
  - Dynamic and Adaptive Routing Mechanisms (4 papers)
  - [7] Routing experts: Learning to route dynamic experts in existing multi-modal large language models (Wu Qiong, 2025) View paper
  - [11] Adamoe: Token-adaptive routing with null experts for mixture-of-experts language models (Zihao Zeng, 2024) View paper
  - [12] Routing experts: Learning to route dynamic experts in multi-modal large language models (Wu Qiong, 2024) View paper
  - [23] Harder tasks need more experts: Dynamic routing in moe models (Huang, 2024) View paper
  - Static Routing Approaches (2 papers)
  - [16] Maximum score routing for mixture-of-experts (Dong Bowen, 2025) View paper
  - [42] Mixture-of-experts with expert choice routing (Zhou, 2022) View paper
  - Shared and Cross-Layer Routing (1 papers)
  - [35] Omni-Router: Sharing Routing Decisions in Sparse Mixture-of-Experts for Speech Recognition (Gu, 2025) View paper
  - Specialized Routing for Multimodal and Multilingual Models (1 papers)
  - [22] Sparse moe with language guided routing for multilingual machine translation (X Zhao, 2024) View paper
  - Routing Stability and Load Balancing (1 papers)
  - [48] Dense Backpropagation Improves Routing for Sparsely-Gated Mixture-of-Experts (A Panda, 2024) View paper
- Model Compression and Sparsification
  - Expert Pruning Methods (5 papers)
  - [8] Not all experts are equal: Efficient expert pruning and skipping for mixture-of-experts large language models (Huang, 2024) View paper

    - [10] Efficient expert pruning for sparse mixture-of-experts language models: Enhancing performance and reducing inference costs (Liu, 2024) View paper
    - [14] Unveiling Hidden Collaboration within Mixture-of-Experts in Large Language Models (Tang Yan, 2025) View paper
    - [33] Cluster-Driven Expert Pruning for Mixture-of-Experts Large Language Models (Guo, 2025) View paper
    - [49] Revisiting smoe language models by evaluating inefficiencies with task specific expert pruning (Sarkar, 2024) View paper
    - Quantization and Mixed-Precision Techniques (1 papers)
    - [20] EAC-MoE: Expert-Selection Aware Compressor for Mixture-of-Experts Large Language Models (Yuanteng Chen, 2025) View paper
    - Activation Sparsity Exploitation (3 papers)
    - [41] Samoyeds: Accelerating MoE Models with Structured Sparsity Leveraging Sparse Tensor Cores (Wu Chenpeng, 2025) View paper
    - [44] SeerAttention: Learning Intrinsic Sparse Attention in Your LLMs (Gao Yizhao, 2024) View paper
    - [46] Turbo Sparse: Achieving LLM SOTA Performance with Minimal Activated Parameters (Song Yixin, 2024) View paper
- Architecture Design and Model Construction
    - Dense-to-Sparse Upcycling (2 papers)
    - [17] DIVE into MoE: Diversity-Enhanced Reconstruction of Large Language Models from Dense into Mixture-of-Experts (Yuchen Feng, 2025) View paper
    - [43] Upcycling Large Language Models into Mixture of Experts (Khattar, 2024) View paper
    - Hybrid Dense-Sparse Training Regimes (1 papers)
    - [15] Dense training, sparse inference: Rethinking training of mixture-of-experts language models (Pan Bowen, 2024) View paper
    - Parameter-Efficient Fine-Tuning with MoE (5 papers)
    - [25] A case study of instruction tuning with mixture of parameter-efficient experts (O Ostapenko, 2023) View paper
    - [28] LLaVA-MoLE: Sparse Mixture of LoRA Experts for Mitigating Data Conflicts in Instruction Finetuning MLLMs (Chen Shaoxiang, 2024) View paper
    - [30] DynMoLE: Boosting Mixture of LoRA Experts Fine-Tuning with a Hybrid Routing Mechanism (Dengchun Li, 2025) View paper
    - [32] Let the Expert Stick to His Last: Expert-Specialized Fine-Tuning for Sparse Architectural Large Language Models (Zihan Wang, 2024) View paper
    - [45] Mixture of Routers (Zhang Jia-Chen, 2025) View paper
- System Optimization and Deployment
    - Training Systems and Frameworks (6 papers)
    - [5] Fsmoe: A flexible and scalable training system for sparse mixture-of-experts models (Xinglin Pan, 2025) View paper
    - [6] Megablocks: Efficient sparse training with mixture-of-experts (Gale, 2023) View paper
    - [9] Deepspeed-moe: Advancing mixture-of-experts inference and training to power next-generation ai scale (Rajbhandari, 2022) View paper
    - [18] Locmoe: A low-overhead moe for large language model training (Li Jing, 2024) View paper
    - [19] LIBMoE: A Library for comprehensive benchmarking Mixture of Experts in Large Language Models (Nam V. Nguyen, 2024) View paper
    - [29] Fastermoe: modeling and optimizing training of large-scale dynamic pre-trained models (Jiaao He, 2022) View paper
    - Inference Optimization and Serving (6 papers)
    - [1] A survey on inference optimization techniques for mixture of experts models (Liu Jiaâ□□Cheng, 2024) View paper
    - [24] GRACE-MoE: Grouping and Replication with Locality-Aware Routing for Efficient Distributed MoE Inference (Han Yu, 2025) View paper
    - [27] Moetuner: Optimized mixture of expert serving with balanced expert placement and token routing (Mahajan, 2025) View paper
    - [38] D2MoE: Dual Routing and Dynamic Scheduling for Efficient On-Device MoE-based LLM Serving (Wang Haodong, 2025) View paper
    - [39] MoESD: Unveil Speculative Decoding's Potential for Accelerating Sparse MoE (Zhu Lei, 2025) View paper
    - [40] Faster MoE LLM Inference for Extremely Large Models (Shi Luohe, 2025) View paper
    - Expert Parallelism and Communication Optimization (1 papers)
    - [34] Advancing MoE Efficiency: A Collaboration-Constrained Routing (C2R) Strategy for Better Expert Parallelism Design (Zhang, 2025) View paper
    - Near-Data and Memory-Centric Computing (1 papers)
    - [50] MoNDE: Mixture of Near-Data Experts for Large-Scale Sparse Models (Taehyun Kim, 2024) View paper
    - Distributed and Edge Deployment (2 papers)
    - [31] MoE: Optimizing Collaborative Inference for Edge Large Language Models (L Jin, 2025) View paper
    - [47] Optimal Expert Selection for Distributed Mixture-of-Experts at the Wireless Edge (Wu Hai, 2025) View paper
- Empirical Studies and Open Models
    - Open-Source MoE Model Releases (2 papers)
    - [4] Openmoe: An early effort on open mixture-of-experts language models (Xue, 2024) View paper
    - [13] Olmoe: Open mixture-of-experts language models (Muennighoff, 2024) View paper
    - Mechanistic Interpretability and Expert Collaboration (1 papers)
    - [37] Unveiling Instruction-Specific Neurons & Experts: An Analytical Framework for LLM's Instruction-Following Capabilities (Zhang Junyan, 2025) View paper
- Domain-Specific Applications
    - Multimodal MoE Systems (1 papers)
    - [36] Scaling and Enhancing LLM-based AVSR: A Sparse Mixture of Projectors Approach (Cappellazzo, 2025) View paper
    - Latent Space and Manifold Learning (1 papers)
    - [26] Unraveling the Localized Latents: Learning Stratified Manifold Structures in LLM Embedding Space with Sparse Mixture-of-Experts (LI Xin, 2025) View paper

Narrative

Core task: Optimal sparsity selection in mixture-of-experts language models. The field organizes around several complementary perspectives on how to design, deploy, and optimize MoE architectures. Scaling Laws and Compute-Optimal Design investigates fundamental trade-offs between parameter count and computational cost, seeking principles that guide when and how much sparsity to introduce—works like Inference-Optimal MoE[2] and Parameters vs FLOPs[3] exemplify efforts to balance model capacity with inference

efficiency. Routing Strategies and Expert Selection focuses on mechanisms that decide which experts process each token, ranging from learned gating functions to more sophisticated schemes like Expert Choice Routing[42] and Maximum Score Routing[16]. Model Compression and Sparsification explores techniques to reduce active parameters through pruning or dynamic expert selection, as seen in Efficient Expert Pruning[10] and Expert Pruning Skipping[8]. Architecture Design examines structural choices—expert granularity, layer configurations, and hybrid dense-sparse patterns—while System Optimization addresses practical deployment challenges such as memory management and parallelization strategies exemplified by MegaBlocks[6] and DeepSpeed-MoE[9]. Empirical Studies and Open Models, including OpenMoE[4] and OLMoE[13], provide reproducible benchmarks, and Domain-Specific Applications adapt MoE principles to specialized tasks.

A central tension runs through the literature: increasing sparsity reduces computation but risks underutilizing model capacity or destabilizing training, while denser activation preserves expressiveness at higher cost. Many studies explore adaptive or learned routing to strike this balance dynamically, and recent work investigates how expert collaboration and token-level specialization interact with sparsity choices. Optimal Sparsity Reasoning[0] sits squarely within the Scaling Laws branch alongside Inference-Optimal MoE[2] and Parameters vs FLOPs[3], emphasizing principled selection of sparsity levels based on compute budgets and downstream performance. Where Inference-Optimal MoE[2] targets deployment-time efficiency and Parameters vs FLOPs[3] examines broad parameter-compute frontiers, Optimal Sparsity Reasoning[0] focuses on reasoning through the interplay of expert utilization, routing entropy, and task-specific demands to prescribe sparsity configurations. This positioning reflects a shift from purely empirical tuning toward theory-driven guidelines that inform architecture decisions before large-scale training begins.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Toward inference-optimal mixture-of-expert large language models

Authors: Yun, Longfei, Zhuang, Yonghao, Longfei Yun, et al. (13 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Mixture-of-Expert (MoE) based large language models (LLMs), such as the recent Mixtral and DeepSeek-MoE, have shown great promise in scaling model size without suffering from the quadratic growth of training cost of dense transformers. Like dense models, training MoEs requires answering the same question: given a training budget, what is the optimal allocation on the model size and number of tokens? We study the scaling law of MoE-based LLMs regarding the relations between the model performance,...

Relationship Analysis

Both papers belong to the Parameter-FLOPs Trade-offs and Scaling Principles category, investigating optimal resource allocation in MoE models under fixed compute budgets. They overlap in examining how total parameters, active parameters, and training tokens interact to determine model performance, with both deriving scaling laws that extend beyond dense models. The key difference is that the original paper focuses on task-specific optimal sparsity for reasoning versus memorization skills and introduces metrics like Active FLOPs and TPP, while the candidate paper emphasizes inference efficiency as a constraint, proposing over-trained configurations with more experts to balance training cost and serving cost.

2. Parameters vs flops: Scaling laws for optimal sparsity for mixture-of-experts language models

Authors: Abnar, Samira, Shah, Harshay, Samira Abnar, et al. (16 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Scaling the capacity of language models has consistently proven to be a reliable approach for improving performance and unlocking new capabilities. Capacity can be primarily defined by two dimensions: the number of model parameters and the compute per example. While scaling typically involves increasing both, the precise interplay between these factors and their combined contribution to overall capacity remains not fully understood. We explore this relationship in the context of sparse Mixture-o...

Relationship Analysis

Both papers belong to the Parameter-FLOPs Trade-offs and Scaling Principles category, investigating how to optimally allocate compute between total parameters, active parameters, and training tokens in MoE models. The original paper focuses on how optimal sparsity differs between memorization and reasoning tasks, identifying that reasoning tasks require balanced TPP (tokens per parameter) around 20 while memorization benefits from higher sparsity, whereas the candidate paper (Abnar et al.) derives general scaling laws for MoE sparsity across compute budgets without distinguishing task types, finding that optimal sparsity increases with model size and that sparser models are more compute-efficient during pretraining but may underperform on certain downstream reasoning tasks.

Contributions Analysis

Overall novelty summary. The paper proposes two principles for MoE sparsity selection: Active FLOPs (models with identical training loss but greater active compute achieve higher reasoning accuracy) and Total tokens per parameter (TPP, distinguishing memorization from reasoning tasks). It resides in the 'Parameter-FLOPs Trade-offs and Scaling Principles' leaf alongside two sibling papers—Inference-Optimal MoE and Parameters vs FLOPs—within the broader 'Scaling Laws and Compute-Optimal Design' branch. This leaf contains only three papers total, indicating a relatively sparse research direction focused on theoretical scaling relationships rather than empirical system benchmarks or routing mechanisms.

The taxonomy reveals that most MoE research concentrates on routing strategies (six sub-leaves), system optimization (five sub-leaves), and compression techniques (three sub-leaves), while scaling law investigations remain comparatively underexplored. The sibling papers address inference-time efficiency and broad parameter-compute frontiers, whereas neighboring branches examine routing stability, expert pruning, and training systems. The paper's focus on disentangling active compute from total parameters through controlled experiments positions it at the intersection of scaling theory and architectural design, diverging from the field's dominant emphasis on deployment optimization and routing policy refinement.

Among thirty candidates examined, none clearly refuted any of the three contributions. The Active FLOPs principle examined ten candidates with zero refutable matches; the TPP principle similarly found no overlapping prior work across ten candidates; the revised compute-optimal framework also encountered no refutations among ten examined papers. This absence of refutation reflects either genuine novelty within the limited search scope or insufficient coverage of closely related scaling law studies. The sibling papers in the same taxonomy leaf establish parameter-compute trade-offs but do not explicitly separate memorization from reasoning or quantify active FLOPs effects, suggesting the contributions address gaps within this sparse research direction.

Based on the top-thirty semantic matches and taxonomy structure, the work appears to introduce distinct principles within an underexplored corner of MoE research. The limited search scope and sparse taxonomy leaf suggest the analysis captures the most relevant prior work but cannot guarantee exhaustive coverage of all scaling law investigations. The absence of refutations across all contributions, combined with the leaf's small size, indicates the paper may be advancing a relatively novel theoretical framework, though broader literature searches could reveal additional related efforts.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Active FLOPs principle for MoE reasoning performance

Description: The authors demonstrate that downstream reasoning quality in MoE models is determined not solely by pre-training loss, but critically by the number of active FLOPs during both training and inference. Models with larger top-k consistently outperform those with smaller top-k even when pre-training loss is matched.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Openmoe: An early effort on open mixture-of-experts language models

URL: View paper

Brief Assessment

OpenMoE[4] focuses on architectural design, routing mechanisms, and training strategies for MoE models, but does not investigate the relationship between active FLOPs and downstream reasoning performance that the original paper establishes.

---

2. Scaling laws across model architectures: A comparative analysis of dense and MoE models in large language models

URL: View paper

Brief Assessment

Scaling Laws Architectures[74] focuses on resource allocation strategies and hyperparameter scaling (batch size, learning rate) across dense and MoE architectures, but does not investigate the relationship between active FLOPs and downstream reasoning performance that the original paper demonstrates.

---

3. Efficient scaling of large language models with mixture of experts and 3D analog in-memory computing

URL: View paper

Brief Assessment

3D Analog Computing[64] focuses on hardware implementation of MoE models using analog in-memory computing architectures. The candidate does not address the relationship between active FLOPs and reasoning performance in MoE models.

---

4. Every FLOP Counts: Scaling a 300B Mixture-of-Experts LING LLM without Premium GPUs

URL: View paper

Brief Assessment

Every FLOP Counts[73] focuses on cost-efficient training of MoE models on lower-performance hardware, not on the relationship between active FLOPs and reasoning performance quality. The paper does not investigate how active compute during training/inference determines downstream reasoning accuracy.

---

5. Parameters vs flops: Scaling laws for optimal sparsity for mixture-of-experts language models

URL: View paper

Brief Assessment

Parameters vs FLOPs[3] focuses on optimal sparsity trade-offs during pretraining under fixed compute budgets, finding that sparser models achieve lower pretraining loss. The original paper specifically examines how active FLOPs during both training and inference determine downstream reasoning accuracy even when pretraining loss is matched, which is a distinct claim about reasoning performance that Parameters vs FLOPs[3] does not directly address.

---

6. Mixture of Lookup Experts

URL: View paper

Brief Assessment

Mixture Lookup Experts[72] focuses on efficient MoE inference through lookup table re-parameterization and offloading strategies, not on the relationship between active FLOPs and reasoning performance during training and inference.

---

7. Infrastructure Economics of Sparse Mixture-of-Experts in Cloud-Native NLP: Benchmarking Cost, Accuracy, and Performance

URL: View paper

Brief Assessment

Infrastructure Economics[71] focuses on cloud deployment costs, infrastructure benchmarking, and economic analysis of MoE systems across cloud platforms. It does not investigate the relationship between active FLOPs and downstream reasoning performance, nor does it examine how top-k routing affects reasoning accuracy independent of pre-training loss.

---

8. BlackMamba: Mixture of Experts for State-Space Models

URL: View paper

Brief Assessment

BlackMamba[75] focuses on combining SSM (state-space models) with MoE architecture for efficiency gains, not on analyzing how active FLOPs during training/inference affect downstream reasoning quality across different top-k configurations.

---

9. Efficient Diffusion Transformer Policies with Mixture of Expert Denoisers for Multitask Learning

URL: View paper

Brief Assessment

Diffusion Transformer Policies[76] focuses on imitation learning for robotics tasks using mixture-of-denoising experts, not on language model reasoning or the relationship between active FLOPs and reasoning accuracy in MoE language models.

---

10. CLIP-UP: A Simple and Efficient Mixture-of-Experts CLIP Training Recipe with Sparse Upcycling

URL: View paper

Brief Assessment

CLIP-UP[77] focuses on converting dense CLIP models to sparse MoE architectures for vision-language tasks, not on analyzing how active FLOPs during training/inference affect reasoning performance in language models. The paper does not discuss reasoning tasks, active compute principles, or the relationship between top-k routing and downstream reasoning quality.

Contribution 2: Total tokens per parameter (TPP) principle distinguishing memorization from reasoning

Description: The work establishes that memorization tasks are parameter-hungry and benefit from lower TPP (more parameters), whereas reasoning tasks exhibit a non-monotonic relationship with TPP, peaking around 20 tokens per parameter. This reveals that reasoning skills require careful balancing of data and parameters.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Learning associative reasoning towards systematicity using modular networks

URL: View paper

Brief Assessment

Associative Reasoning[57] focuses on associative reasoning tasks using modular networks with TPR (Tensor Product Representation) as external memory. The candidate does not address tokens-per-parameter ratios, data-parameter scaling relationships, or the distinction between memorization and reasoning tasks in the context of language model training efficiency.

---

2. Projected Compression

URL: View paper

Brief Assessment

Projected Compression[59] focuses on model compression through trainable projection modules for reducing transformer size, not on analyzing tokens-per-parameter ratios or their differential effects on memorization versus reasoning tasks.

---

3. Surprising effectiveness of pretraining ternary language model at scale

URL: View paper

Brief Assessment

Ternary Language Model[51] focuses on low-bitwidth model training and scaling laws in terms of model size (bits) and parameters, not on the tokens-per-parameter ratio's differential effects on memorization versus reasoning tasks.

---

4. A Theory of Inference Compute Scaling: Reasoning through Directed Stochastic Skill Search

URL: View paper

Brief Assessment

Inference Compute Scaling[54] focuses on inference-time compute scaling and reasoning through skill search, not on the TPP ratio's differential effects on memorization versus reasoning tasks during pretraining.

---

5. Revisiting the Scaling Properties of Downstream Metrics in Large Language Model Training

URL: View paper

Brief Assessment

Downstream Metrics Scaling[60] focuses on predicting downstream benchmark accuracy from training compute using power laws, without investigating the differential effects of TPP on memorization versus reasoning tasks. The candidate does not address the parameter-hungry vs. data-hungry distinction that is central to the original contribution.

---

6. Resolving discrepancies in compute-optimal scaling of language models

URL: View paper

Brief Assessment

Compute-Optimal Scaling[55] focuses on reconciling discrepancies between Kaplan et al. and Hoffmann et al. scaling laws by analyzing flop counting, warmup duration, and optimizer tuning. It does not investigate how TPP differentially affects memorization versus reasoning tasks, which is the core novelty claim of the original contribution.

---

7. Cola: Compute-efficient pre-training of llms via low-rank activation

URL: View paper

Brief Assessment

Cola[52] focuses on architectural efficiency through low-rank activations in neural network layers, not on the relationship between tokens-per-parameter ratios and task performance types (memorization vs. reasoning).

---

8. Inference Optimal VLMs Need Fewer Visual Tokens and More Parameters

URL: View paper

Brief Assessment

Inference Optimal VLMs[53] focuses on vision-language models and the trade-off between visual tokens and LLM parameters for inference optimization, not on the general TPP principle for distinguishing memorization versus reasoning tasks in language models.

---

9. Evaluation of pre-training large language models on leadership-class supercomputers: J. Yin et al.

URL: View paper

Brief Assessment

Pre-training Supercomputers[56] focuses on performance benchmarking and cost analysis of training LLMs on supercomputers, not on the relationship between TPP ratios and task-specific capabilities (memorization vs. reasoning).

---

10. Scaling Laws and Efficient Inference for Ternary Language Models

URL: View paper

Brief Assessment

Ternary Scaling Laws[58] focuses on ternary quantization and its scaling properties, not on the TPP principle for distinguishing memorization versus reasoning tasks. The candidate examines how ternary models scale with training data and parameters but does not establish TPP-based distinctions between task types.

Contribution 3: Revised compute-optimal scaling framework for MoE models

Description: The authors argue that the classical compute-optimal scaling laws must be revised for MoE architectures to jointly account for active FLOPs and tokens-per-parameter ratio. This framework shows that optimal sparsity is task-dependent: memorization favors higher sparsity while reasoning requires balancing active compute with data intensity.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model

URL: View paper

Brief Assessment

DeepSeek-V2[63] focuses on architectural innovations (MLA, DeepSeekMoE) for efficient inference and training cost reduction, not on deriving compute-optimal scaling laws that jointly account for active FLOPs and tokens-per-parameter ratio across different task types (memorization vs. reasoning).

2. Efficient scaling of large language models with mixture of experts and 3D analog in-memory computing

URL: View paper

Brief Assessment

3D Analog Computing[64] discusses scaling laws for dense and MoE models in the context of hardware efficiency, but does not propose a framework that jointly accounts for active FLOPs and tokens-per-parameter ratio for task-dependent optimal sparsity.

3. Scaling diffusion transformers to 16 billion parameters

URL: View paper

Brief Assessment

Scaling Diffusion Transformers[65] focuses on mixture-of-experts architectures for diffusion models in computer vision, not on compute-optimal scaling laws for language models or the relationship between active FLOPs, tokens-per-parameter ratio, and task-dependent sparsity optimization.

4. Every Activation Boosted: Scaling General Reasoner to 1 Trillion Open Language Foundation

URL: View paper

Brief Assessment

Trillion Open Language[70] focuses on implementing a high-sparsity MoE architecture for reasoning tasks at trillion-parameter scale, but does not present a revised compute-optimal scaling framework that jointly accounts for active FLOPs and tokens-per-parameter ratio across different task types (memorization vs. reasoning).

5. Glam: Efficient scaling of language models with mixture-of-experts

URL: View paper

Brief Assessment

GLaM[61] focuses on demonstrating that sparse MoE models can match dense model performance with lower training costs, but does not propose a revised compute-optimal scaling framework that jointly accounts for active FLOPs and tokens-per-parameter ratio as the original paper does.

6. Scaling laws for native multimodal models

URL: View paper

Brief Assessment

Native Multimodal Scaling[66] focuses on multimodal model architectures (early vs. late fusion) and their scaling properties, not on MoE sparsity optimization for reasoning vs. memorization tasks as claimed in the original paper.

7. Scaling physics-informed hard constraints with mixture-of-experts

URL: View paper

Brief Assessment

Physics-Informed MoE[68] focuses on using mixture-of-experts to scale physics-informed hard constraints in neural PDE solvers, not on compute-optimal scaling laws for language models or the relationship between active FLOPs, sparsity, and tokens-per-parameter ratio in general LLM training.

8. Towards greater leverage: Scaling laws for efficient mixture-of-experts language models

URL: View paper

Brief Assessment

Greater Leverage[62] focuses on efficiency leverage (EL) as a metric for MoE computational efficiency relative to dense models, examining how activation ratio, granularity, and compute budget affect this efficiency metric. The original paper investigates optimal MoE sparsity for different task types (memorization vs. reasoning) by analyzing active FLOPs and tokens-per-parameter ratios. These are complementary perspectives on MoE scaling rather than overlapping novelty claims.

9. Mixture of a million experts

URL: View paper

Brief Assessment

Million Experts[69] focuses on architectural design for extreme-scale MoE with product key retrieval and does not propose a revised compute-optimal scaling framework that jointly accounts for active FLOPs and tokens-per-parameter ratio as the original paper does.

10. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity

URL: View paper

Brief Assessment

Switch Transformers[67] focuses on architectural simplification (k=1 routing) and training stability for sparse models, not on compute-optimal scaling laws that jointly account for active FLOPs and tokens-per-parameter ratio across different task types (memorization vs. reasoning).

Appendix: Text Similarity Detection

Textual similarity detection checked 30 papers and found 3 similarity segment(s) across 2 paper(s).

The following **2 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Glam: Efficient scaling of language models with mixture-of-experts

Detected in: Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

---

2. Openmoe: An early effort on open mixture-of-experts language models

Detected in: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Optimal Sparsity of Mixture-of-Experts Language Models for Reasoning Tasks View paper
- [1] A survey on inference optimization techniques for mixture of experts models View paper
- [2] Toward inference-optimal mixture-of-expert large language models View paper
- [3] Parameters vs flops: Scaling laws for optimal sparsity for mixture-of-experts language models View paper
- [4] Openmoe: An early effort on open mixture-of-experts language models View paper
- [5] Fsmoe: A flexible and scalable training system for sparse mixture-of-experts models View paper
- [6] Megablocks: Efficient sparse training with mixture-of-experts View paper
- [7] Routing experts: Learning to route dynamic experts in existing multi-modal large language models View paper
- [8] Not all experts are equal: Efficient expert pruning and skipping for mixture-of-experts large language models View paper
- [9] Deepspeed-moe: Advancing mixture-of-experts inference and training to power next-generation ai scale View paper
- [10] Efficient expert pruning for sparse mixture-of-experts language models: Enhancing performance and reducing inference costs View paper
- [11] Adamoe: Token-adaptive routing with null experts for mixture-of-experts language models View paper
- [12] Routing experts: Learning to route dynamic experts in multi-modal large language models View paper
- [13] Olmoe: Open mixture-of-experts language models View paper
- [14] Unveiling Hidden Collaboration within Mixture-of-Experts in Large Language Models View paper
- [15] Dense training, sparse inference: Rethinking training of mixture-of-experts language models View paper
- [16] Maximum score routing for mixture-of-experts View paper
- [17] DIVE into MoE: Diversity-Enhanced Reconstruction of Large Language Models from Dense into Mixture-of-Experts View paper
- [18] Locmoe: A low-overhead moe for large language model training View paper
- [19] LIBMoE: A Library for comprehensive benchmarking Mixture of Experts in Large Language Models View paper
- [20] EAC-MoE: Expert-Selection Aware Compressor for Mixture-of-Experts Large Language Models View paper
- [21] Sparsity and Superposition in Mixture of Experts View paper
- [22] Sparse moe with language guided routing for multilingual machine translation View paper
- [23] Harder tasks need more experts: Dynamic routing in moe models View paper
- [24] GRACE-MoE: Grouping and Replication with Locality-Aware Routing for Efficient Distributed MoE Inference View paper
- [25] A case study of instruction tuning with mixture of parameter-efficient experts View paper
- [26] Unraveling the Localized Latents: Learning Stratified Manifold Structures in LLM Embedding Space with Sparse Mixture-of-Experts View paper
- [27] Moetuner: Optimized mixture of expert serving with balanced expert placement and token routing View paper
- [28] LLaVA-MoLE: Sparse Mixture of LoRA Experts for Mitigating Data Conflicts in Instruction Finetuning MLLMs View paper
- [29] Fastermoe: modeling and optimizing training of large-scale dynamic pre-trained models View paper
- [30] DynMoLE: Boosting Mixture of LoRA Experts Fine-Tuning with a Hybrid Routing Mechanism View paper
- [31] MoE: Optimizing Collaborative Inference for Edge Large Language Models View paper
- [32] Let the Expert Stick to His Last: Expert-Specialized Fine-Tuning for Sparse Architectural Large Language Models View paper
- [33] Cluster-Driven Expert Pruning for Mixture-of-Experts Large Language Models View paper
- [34] Advancing MoE Efficiency: A Collaboration-Constrained Routing (C2R) Strategy for Better Expert Parallelism Design View paper
- [35] Omni-Router: Sharing Routing Decisions in Sparse Mixture-of-Experts for Speech Recognition View paper
- [36] Scaling and Enhancing LLM-based AVSR: A Sparse Mixture of Projectors Approach View paper
- [37] Unveiling Instruction-Specific Neurons & Experts: An Analytical Framework for LLM's Instruction-Following Capabilities View paper
- [38] D2MoE: Dual Routing and Dynamic Scheduling for Efficient On-Device MoE-based LLM Serving View paper
- [39] MoESD: Unveil Speculative Decoding's Potential for Accelerating Sparse MoE View paper
- [40] Faster MoE LLM Inference for Extremely Large Models View paper
- [41] Samoyeds: Accelerating MoE Models with Structured Sparsity Leveraging Sparse Tensor Cores View paper
- [42] Mixture-of-experts with expert choice routing View paper
- [43] Upcycling Large Language Models into Mixture of Experts View paper
- [44] SeerAttention: Learning Intrinsic Sparse Attention in Your LLMs View paper
- [45] Mixture of Routers View paper
- [46] Turbo Sparse: Achieving LLM SOTA Performance with Minimal Activated Parameters View paper
- [47] Optimal Expert Selection for Distributed Mixture-of-Experts at the Wireless Edge View paper
- [48] Dense Backpropagation Improves Routing for Sparsely-Gated Mixture-of-Experts View paper
- [49] Revisiting smoe language models by evaluating inefficiencies with task specific expert pruning View paper
- [50] MoNDE: Mixture of Near-Data Experts for Large-Scale Sparse Models View paper
- [51] Surprising effectiveness of pretraining ternary language model at scale View paper
- [52] Cola: Compute-efficient pre-training of llms via low-rank activation View paper
- [53] Inference Optimal VLMs Need Fewer Visual Tokens and More Parameters View paper
- [54] A Theory of Inference Compute Scaling: Reasoning through Directed Stochastic Skill Search View paper
- [55] Resolving discrepancies in compute-optimal scaling of language models View paper
- [56] Evaluation of pre-training large language models on leadership-class supercomputers: J. Yin et al. View paper
- [57] Learning associative reasoning towards systematicity using modular networks View paper

- [58] Scaling Laws and Efficient Inference for Ternary Language Models View paper
- [59] Projected Compression View paper
- [60] Revisiting the Scaling Properties of Downstream Metrics in Large Language Model Training View paper
- [61] Glam: Efficient scaling of language models with mixture-of-experts View paper
- [62] Towards greater leverage: Scaling laws for efficient mixture-of-experts language models View paper
- [63] DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model View paper
- [64] Efficient scaling of large language models with mixture of experts and 3D analog in-memory computing View paper
- [65] Scaling diffusion transformers to 16 billion parameters View paper
- [66] Scaling laws for native multimodal models View paper
- [67] Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity View paper
- [68] Scaling physics-informed hard constraints with mixture-of-experts View paper
- [69] Mixture of a million experts View paper
- [70] Every Activation Boosted: Scaling General Reasoner to 1 Trillion Open Language Foundation View paper
- [71] Infrastructure Economics of Sparse Mixture-of-Experts in Cloud-Native NLP: Benchmarking Cost, Accuracy, and Performance View paper
- [72] Mixture of Lookup Experts View paper
- [73] Every FLOP Counts: Scaling a 300B Mixture-of-Experts LING LLM without Premium GPUs View paper
- [74] Scaling laws across model architectures: A comparative analysis of dense and MoE models in large language models View paper
- [75] BlackMamba: Mixture of Experts for State-Space Models View paper
- [76] Efficient Diffusion Transformer Policies with Mixture of Expert Denoisers for Multitask Learning View paper
- [77] CLIP-UP: A Simple and Efficient Mixture-of-Experts CLIP Training Recipe with Sparse Upcycling View paper