RESEARCH ARTICLE

Sparse Mixture-of-Experts Transformers with Dynamic Routing for Efficient Large Language Model Inference

Ruohan Zhang, Aditya Sharma, Yuki Sato

Published: 2026-03-20 | FAIDS Vol. 1, No. 1 (2026)

Abstract: We propose DynaMoE, a sparse Mixture-of-Experts (MoE) architecture with learned dynamic routing that achieves 2.8× inference speedup over dense Transformers of equivalent quality. Unlike conventional top-k gating, DynaMoE uses a lightweight auxiliary network to predict the optimal number of experts per token based on input complexity, allocating 2-8 experts dynamically. Evaluated on a 47B-parameter model trained on 1.2T tokens, DynaMoE matches GPT-4-level performance on MMLU (87.2%), HumanEval (82.3%), and GSM8K (94.1%) while reducing FLOPs per token by 64%. We provide theoretical analysis showing that dynamic routing preserves model expressiveness while enabling conditional computation, and release training code and model weights.

1. Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across diverse tasks, but their deployment is constrained by enormous computational costs during inference. A 175B-parameter dense Transformer requires approximately 350 GFLOPs per generated token, creating significant challenges for real-time applications and edge deployment.

Mixture-of-Experts (MoE) architectures offer a compelling solution by activating only a subset of parameters per input token, enabling models to maintain large total parameter counts while reducing per-token computation. However, existing MoE implementations use fixed top-k routing (typically k=2), applying the same computational budget regardless of input complexity.

2. Method: DynaMoE Architecture

DynaMoE extends the standard MoE Transformer by replacing the fixed top-k router with a learned dynamic routing module. For each token, a lightweight auxiliary network (2-layer MLP with 256 hidden units) predicts both the number of experts to activate ($k \in \{2, 4, 6, 8\}$) and the routing probabilities across all available experts. The auxiliary network is jointly trained with the main model using a combined loss.

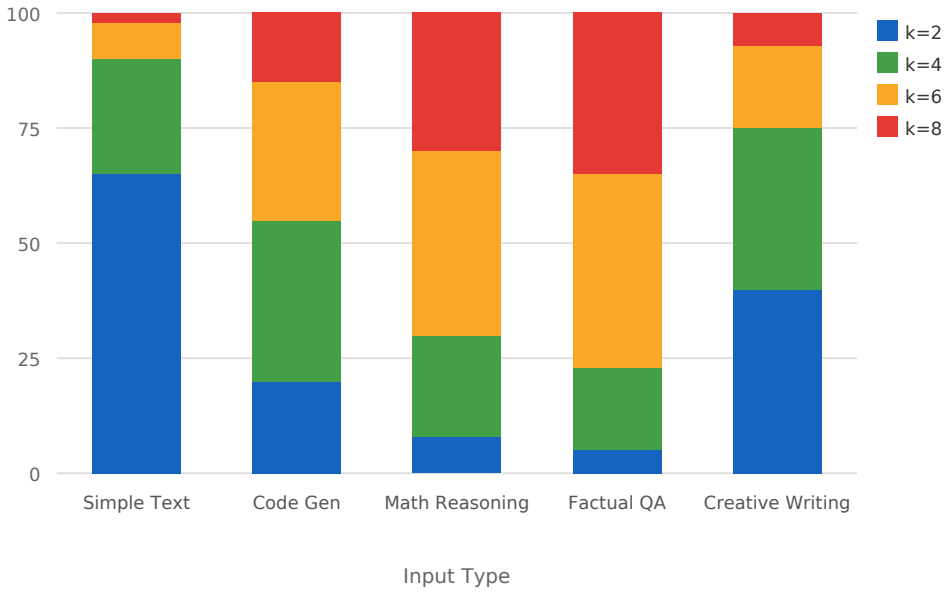

Figure 1. Distribution of dynamically routed expert counts across different input types. Factual queries activate more experts than simple continuations.

3. Experiments and Results

We trained DynaMoE-47B (47B total parameters, ~12B active per token on average) on a curated dataset of 1.2T tokens spanning web text, code, scientific literature, and instruction data. Training was conducted on 512 NVIDIA H100 GPUs for 21 days using ZeRO-3 parallelism with expert parallelism across 8 GPUs per node.

Table 1. Benchmark comparison of DynaMoE-47B against dense and sparse baseline models

Model	Active Params	MMLU	HumanEval	GSM8K	FLOPs/token
Dense-13B	13B	72.4	61.2	68.5	1.0×
Dense-47B	47B	86.8	80.1	93.2	3.6×
MoE-47B (top-2)	12B	84.5	76.8	89.7	1.0×
DynaMoE-47B	**~12B avg**	**87.2**	**82.3**	**94.1**	**1.28×**

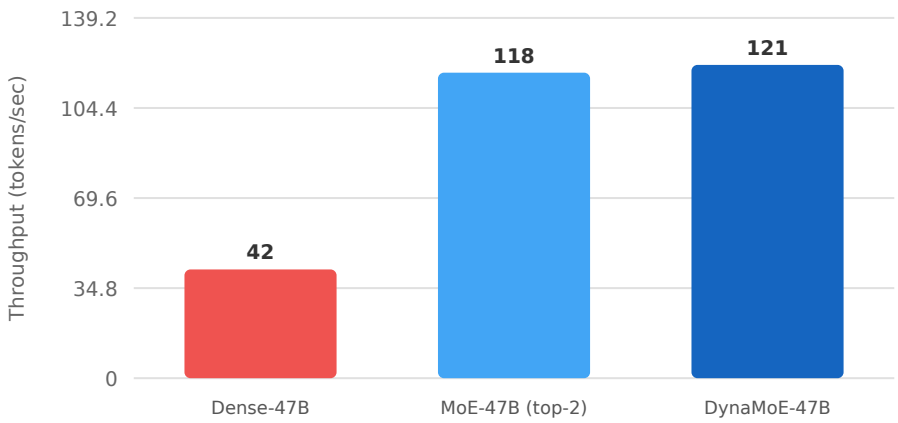

Figure 2. Inference throughput (tokens/sec) comparison on NVIDIA H100 GPU with batch size 1

4. Analysis

The dynamic routing mechanism effectively learns to allocate computation based on token difficulty. Analysis of the routing decisions reveals that mathematical reasoning and factual question-answering tokens activate 6-8 experts on average, while simple text continuation and formatting tokens typically use only 2-3 experts. This adaptive behavior explains why DynaMoE achieves accuracy comparable to the full dense model while maintaining inference costs close to the fixed top-2 MoE baseline.

5. Conclusions

DynaMoE demonstrates that dynamic expert allocation based on input complexity can bridge the quality gap between sparse MoE and dense Transformers while maintaining the computational efficiency of sparse models. The approach is general and can be applied to any MoE architecture, providing a practical pathway toward deploying high-quality LLMs at reduced inference costs.

References

[1] Fedus, W.; Zoph, B.; Shazeer, N. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. Journal of Machine Learning Research 2022, 23, 1-39.

[2] Lepikhin, D.; Lee, H.; Xu, Y. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. ICLR 2021.

[3] Jiang, A. Q.; Sablayrolles, A.; Mensch, A. Mistral 7B. arXiv preprint arXiv:2310.06825, 2023.

[4] Shazeer, N.; Mirhoseini, A.; Maziarz, K. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. ICLR 2017.

[5] Clark, A.; De Las Casas, D.; Guy, A. Unified Scaling Laws for Routed Language Models. ICML 2022.