MoEP: Modular Expert Paths for Sample-Efficient Language Modeling

Joonas Tapaninaho [1]

[1]University of Oulu, CMVS

Motivation

Dense decoder-only Transformers (e.g., GPT-2) activate all parameters for every token, which is inefficient for small-scale data such as the BabyLM strict-small track. MoEP introduces sparse modular routing to improve **sample efficiency** and **model modularity**. Each token follows selective computation paths through experts and blocks.

MoEP Architecture

MoEP integrates two levels of sparsity:

- **Top-k routing** across parallel Transformer blocks
- **Mixture-of-Experts (MoE)** feed-forward projections (Linear or SwiGLU)

Each token activates only a subset of experts and blocks, creating diverse computation paths while reducing redundancy. A load-balancing term prevents expert collapse.

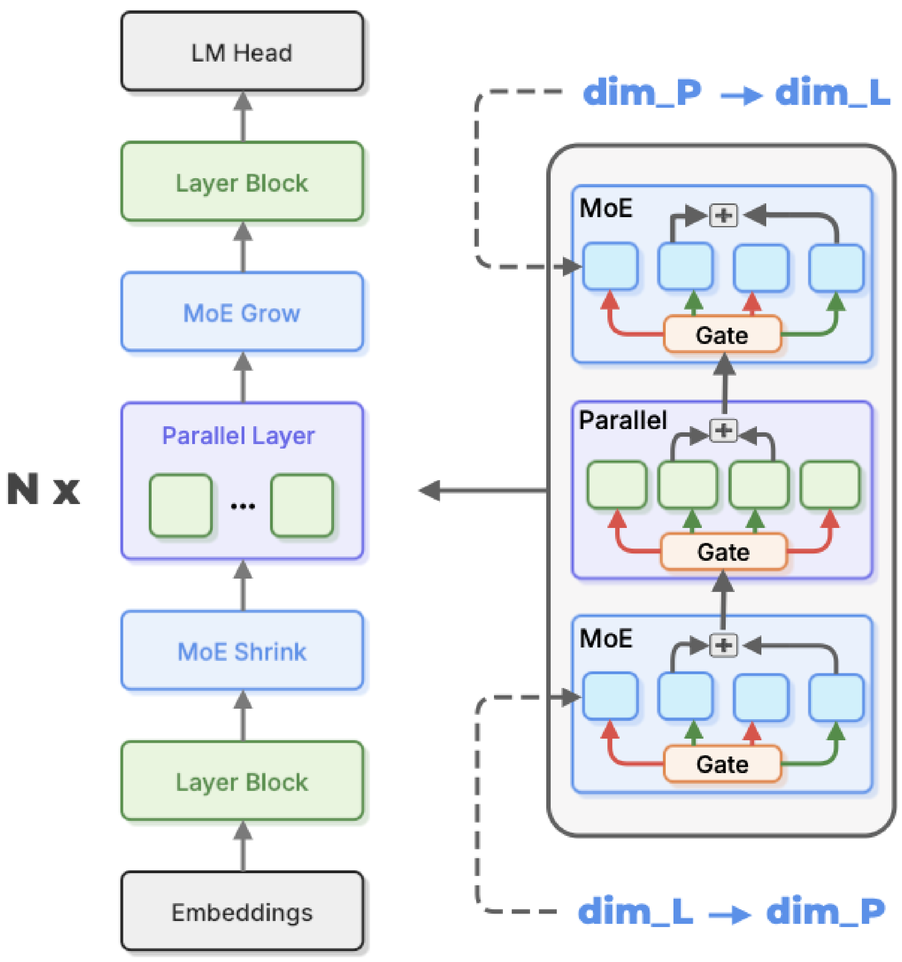

Figure 1. MoEP routing structure: sparse expert and parallel block selection.

Model Overview

Structure: Layer Block → MoE (shrink) → Parallel Layer Stack (top-k) → MoE (grow) → Layer Block

Experimental Setup

All models were trained on the **BabyLM strict-small** dataset (10M words) using a **GPT-2 BPE tokenizer** with a 16K vocabulary. Training was conducted for **10 epochs** using the AdamW optimizer with a cosine learning rate schedule on a single **NVIDIA A100 GPU**. Evaluation followed the official **BabyLM pipeline**, covering BLiMP, WUG, AoA, and related benchmark tasks.

Results

MoEP outperformed all BabyLM strict-small baselines, including GPT-2, while using the same parameter count.

Model	Macro Avg.	Peak Checkpoint
GPT-2	48.1	30M
MoEP	**49.0**	**30M**
MoEP-SwiGLU	47.7	80M

Observation: MoEP reaches its best score at 30M words (seen), indicating **faster early learning**. SwiGLU variant improves task specialization but introduces instability.

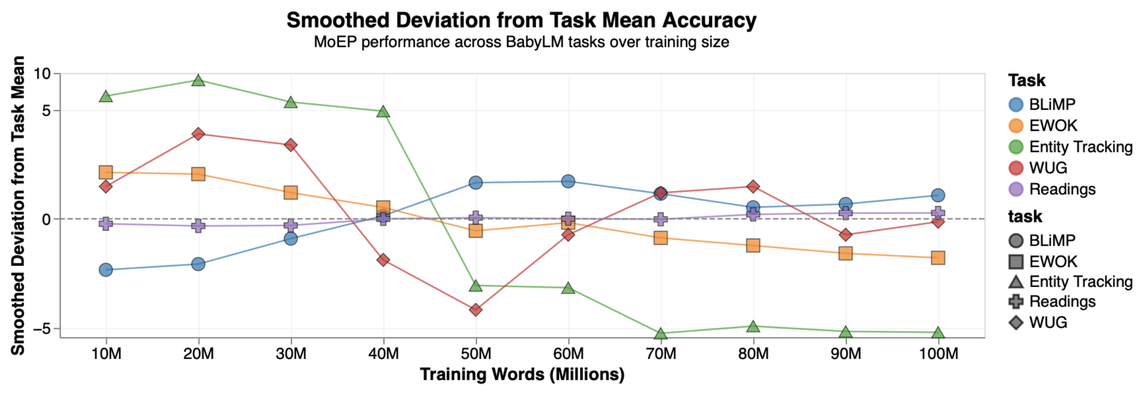

Figure 2. Fast-eval trends: MoEP stabilizes earlier than GPT-2.

Architectural Takeaways

- Sparse modular routing provides a balance between compute efficiency and learning dynamics.
- SwiGLU adds specialization but increases training variance.
- Modular experts enable fine-grained control over capacity allocation.

Key Findings

- Sparse routing accelerates early learning but may reduce late-stage stability.
- Linear experts outperform SwiGLU in low-data regimes.
- Parallel sparse paths can match or exceed dense GPT-2 and other baselines under BabyLM constraints.

Figure 3. Scan for code and model