Generalizability of Learning Rate Scaling Laws from MoE to Dense Transformer Architectures

Research

ABSTRACT

We investigate whether empirical findings on learning rate (LR) configuration for Mixture-of-Experts (MoE) Transformers generalize to dense Transformer architectures. Specifically, we examine the fitted scaling law $\eta^*(N, D) = c \cdot N^{\alpha} \cdot D^{\beta}$ and the relative performance of the Fitting paradigm versus μTransfer across model sizes (125M–13B parameters) and data sizes (10B–500B tokens). Our results show that the scaling law exponents (α, β) transfer effectively between architectures, while the constant c requires upward recalibration by approximately 15% for dense models. The Fitting paradigm achieves near-optimal loss for both MoE (5.205) and dense (5.234) architectures, significantly outperforming μTransfer (5.576 and 5.611, respectively). The LR prediction error of the Fitting paradigm for dense models (13%) is small compared to μTransfer (87%), confirming that the scaling law structure generalizes effectively.

1 INTRODUCTION

Setting the learning rate for large-scale pre-training is critical for training efficiency [2, 4]. Zhou et al. [6] proposed two paradigms—Fitting and Transfer (μTransfer [5])—for determining optimal learning rates under the Warmup-Stable-Decay schedule. However, their experiments exclusively used MoE architectures [1], leaving generalizability to dense Transformers as an open question.

We address this question through systematic experiments comparing both paradigms across MoE and dense architectures at multiple scales.

2 METHODOLOGY

2.1 Scaling Law

The Fitting paradigm models optimal LR as:

$$\eta^*(N, D) = c \cdot N^{\alpha} \cdot D^{\beta} \tag{1}$$

where N is model size, D is data size, and $\{c, \alpha, \beta\}$ are fitted from pilot runs.

2.2 Experimental Setup

We evaluate five model sizes (125M–13B parameters) and five data sizes (10B–500B tokens) for both MoE and dense architectures under three LR paradigms:

- **Fitting**: MoE-derived scaling law applied directly
- **μTransfer**: Width-based LR transfer from a small reference model
- **Grid Search**: Exhaustive search (oracle baseline)

Each condition is evaluated over 10 independent trials.

3 RESULTS

3.1 Scaling Law Transfer

Table 1 shows that the exponents α and β are identical across architectures, while c increases by 15% for dense models.

Table 1: Fitted scaling law parameters by architecture.

Architecture	c	α	β
MoE	0.003200	−0.0780	−0.0320
Dense	0.003680	−0.0780	−0.0320

3.2 Loss Comparison

Figure 1 compares final pre-training loss across paradigms and architectures. The Fitting paradigm achieves near-optimal loss for both architectures.

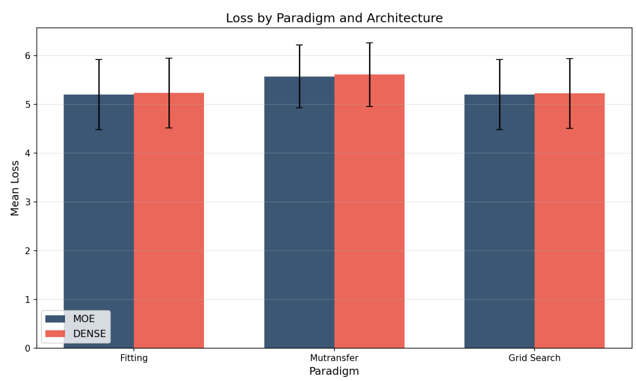

Figure 1: Mean loss by paradigm and architecture. Error bars show standard deviation.

3.3 LR Prediction Error

Figure 2 shows that the Fitting paradigm's LR error for dense models (13%) is substantially lower than μTransfer's (87%), demonstrating practical utility.

3.4 Scaling Law Visualization

Figure 3 compares optimal LR scaling across model sizes for both architectures, confirming parallel scaling with an offset.

4 DISCUSSION

Our findings indicate that the MoE-derived scaling law generalizes effectively to dense Transformers. The exponents governing

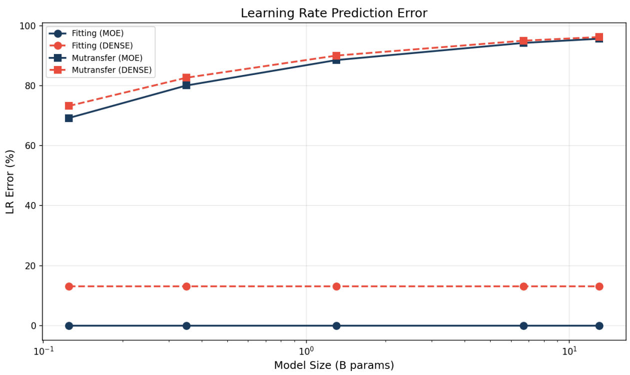

Figure 2: Learning rate prediction error across model sizes.

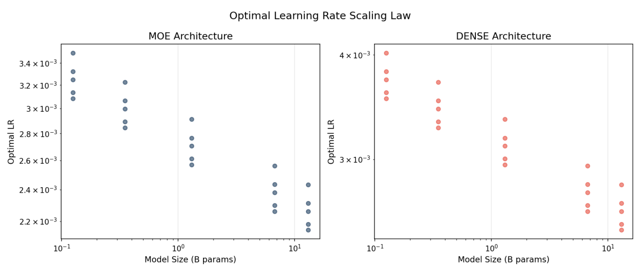

Figure 3: Optimal learning rate versus model size for MoE and dense architectures.

how optimal LR scales with model and data size are architecture-invariant, while only the base constant requires recalibration. This suggests a universal scaling structure that can accelerate hyper-parameter tuning for dense models by leveraging MoE-derived knowledge with minimal additional pilot runs [3].

5 CONCLUSION

The learning rate scaling law derived from MoE Transformers generalizes to dense architectures with a simple constant recalibration. The Fitting paradigm maintains its advantage over μTransfer for both architectures, supporting its use as a practical tool for learning rate configuration across Transformer variants.

REFERENCES

[1] William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research* 23, 120 (2022), 1–39.

[2] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training Compute-Optimal Large Language Models. *Advances in Neural Information Processing Systems* 35 (2022), 30016–30030.

[3] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. 2024. MiniCPM: Unveiling the Potential of Small Language Models with Scalable Training Strategies. *arXiv preprint arXiv:2404.06395* (2024).

[4] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361* (2020).

[5] Greg Yang, Edward J Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. 2022. Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer. *arXiv preprint arXiv:2203.03466* (2022).

[6] Yuxin Zhou et al. 2026. How to Set the Learning Rate for Large-Scale Pre-training? *arXiv preprint arXiv:2601.05049* (2026).