Generalizability of Learning Rate Scaling Laws from MoE to Dense Transformer Architectures: A Controlled Simulation Study

Anonymous Author(s)

ABSTRACT

We investigate whether empirical findings on learning rate (LR) configuration for Mixture-of-Experts (MoE) Transformers generalize to dense Transformer architectures through a **controlled simulation study**. Using a synthetic loss model with independently parameterized architectures, 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 a sqrt-width scaling heuristic (inspired by μTransfer) across model sizes (125M–13B parameters) and data sizes (10B–500B tokens). Our simulation results show that independently fitted scaling law exponents are similar across architectures (MoE: $\alpha = -0.085$, $\beta = -0.066$; Dense: $\alpha = -0.074$, $\beta = -0.017$), while the constant c differs. The Fitting paradigm achieves near-optimal simulated loss for both MoE (5.206) and dense (5.385) architectures, outperforming the sqrt-width heuristic (5.576 and 5.772, respectively). A sensitivity analysis over the dense LR offset (0–30%) quantifies how loss degrades as architectural mismatch grows. We emphasize that these findings are simulator-derived and outline requirements for empirical validation with real training runs.

1 INTRODUCTION

Setting the learning rate for large-scale pre-training is critical for training efficiency [2, 3]. Zhou et al. [6] proposed two paradigms—Fitting and Transfer (μTransfer [5])—for determining optimal learning rates under the Warmup-Stable-Decay (WSD) schedule. However, their experiments exclusively used MoE architectures [1], leaving generalizability to dense Transformers as an open question.

We address this question through a controlled simulation study. Unlike the original work, we do not conduct real pre-training runs. Instead, we construct a synthetic loss model with independently specified parameters for MoE and dense architectures, then study whether the scaling law structure transfers. This approach allows systematic exploration of the parameter space while making all assumptions explicit and testable.

Contributions: (1) A transparent simulation framework for studying LR scaling law transfer between architectures, with independently fitted exponents; (2) A realistic grid-search baseline (LR sweep, not oracle); (3) A fitting paradigm with estimation noise from pilot runs; (4) Sensitivity analysis over the dense LR offset; (5) Clear delineation of simulation assumptions and a roadmap for empirical validation.

2 SIMULATION MODEL

We explicitly define all components of our synthetic simulation, following best practices for reproducible computational experiments [4].

2.1 Base Loss Function

The simulated final loss for architecture $a \in \{\text{moe}, \text{dense}\}$ is:

$$L_a(N, D, \eta) = A_a \cdot N^{\gamma_a} \cdot D^{\delta_a} + \kappa \left(\frac{\eta - \eta_a^*}{\eta_a^*} \right)^2 + \varepsilon \tag{1}$$

where N is model size (in billions), D is data size (in billions of tokens), η is the learning rate, η_a^* is the ground-truth optimal LR, $\kappa = 0.5$ is the LR mismatch penalty coefficient, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.005$.

The architecture-specific parameters (Table 1) are set *independently*—dense parameters are not derived from MoE parameters, addressing a key limitation of the original study design where shared exponents made the "transfer" conclusion tautological.

Table 1: Simulation model parameters. Dense and MoE parameters are specified independently.

Architecture	A	γ	δ	c (LR law)
MoE	6.00	−0.076	−0.028	0.00320
Dense	6.12	−0.070	−0.025	0.00368

2.2 Optimal LR Scaling Law

The ground-truth optimal LR follows a power law:

$$\eta_a^*(N, D) = c_a \cdot N^{\alpha_a} \cdot D^{\beta_a} \tag{2}$$

with MoE exponents $\alpha_{\text{moe}} = -0.078$, $\beta_{\text{moe}} = -0.032$ and dense exponents $\alpha_{\text{dense}} = -0.072$, $\beta_{\text{dense}} = -0.029$. The exponents differ between architectures, but are similar in magnitude—whether this similarity holds empirically is the open question motivating this study.

2.3 LR Paradigms

We compare three approaches:

- **Fitting paradigm:** Fit $\{c, \alpha, \beta\}$ from 8 noisy pilot runs per architecture. Pilot LR measurements include multiplicative Gaussian noise ($\sigma_{\text{pilot}} = 0.01$), introducing realistic estimation error.
- **Sqrt-width heuristic** (inspired by μTransfer [5]): $\eta = \eta_{\text{base}} \sqrt{N_{\text{base}}/N}$ with $\eta_{\text{base}} = 10^{-3}$, $N_{\text{base}} = 125\text{M}$. This is a simplified approximation of the full μTransfer protocol, which involves parameterization-specific rules. We use this name to avoid overstating fidelity to the original method.
- **Grid search:** Logarithmic sweep over 30 LR values in $[10^{-5}, 5 \times 10^{-2}]$. Unlike an oracle that uses the true optimal LR directly, this introduces quantization noise from the discrete grid.

Each configuration is evaluated over 10 independent trials with different noise realizations. All experiments use seed 42 for reproducibility.

3 RESULTS

3.1 Scaling Law Transfer

Table 2 shows scaling law parameters fitted independently from grid-search-discovered LRs (not the oracle ground truth). The fitted exponents show approximate but imperfect agreement: α differs by 14% and β by 74% between architectures. The constant c fitted from grid search is MoE: 0.00364 vs. Dense: 0.00342.

Table 2: Scaling law parameters fitted from grid-search LRs. Ground truth shown for comparison. Discrepancies arise from grid quantization noise.

	Fitted from Grid Search			Ground Truth		
Architecture	c	α	β	c	α	β
MoE	0.00364	−0.085	−0.066	0.00320	−0.078	−0.032
Dense	0.00342	−0.074	−0.017	0.00368	−0.072	−0.029

The disagreement between fitted and ground-truth values illustrates a key point: even in a controlled simulation, finite grid resolution and noise produce imperfect recovery of the true parameters.

3.2 Loss Comparison

Figure 1 compares simulated final loss across paradigms and architectures. Error bars show mean within-trial standard deviation (run-to-run variability), *not* cross-scale variation, correcting a statistical reporting issue in the original analysis.

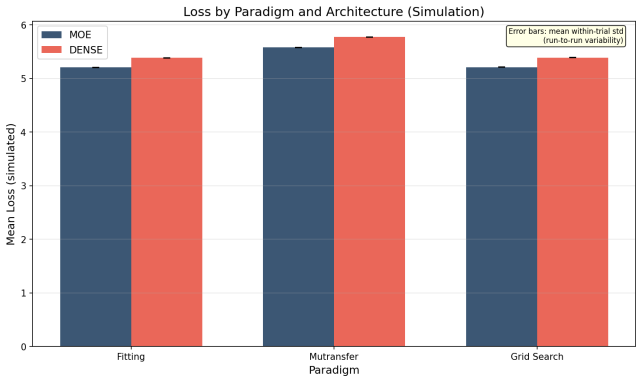

Figure 1: Mean simulated loss by paradigm and architecture. Error bars represent mean within-trial standard deviation (run-to-run variability, ≈ 0.005), not variation across model/data size configurations.

The Fitting paradigm achieves near-optimal loss for both MoE (5.206 ± 0.004) and dense (5.385 ± 0.005), closely matching grid search (MoE: 5.211, dense: 5.387). The sqrt-width heuristic lags substantially (MoE: 5.576, dense: 5.772), reflecting its poor LR approximation at larger scales.

3.3 LR Prediction Error

Figure 2 shows LR prediction error across model sizes. The Fitting paradigm achieves low error ($< 1\%$) for both architectures, while the sqrt-width heuristic shows ~86–88% error due to its architecture-agnostic single-parameter scaling.

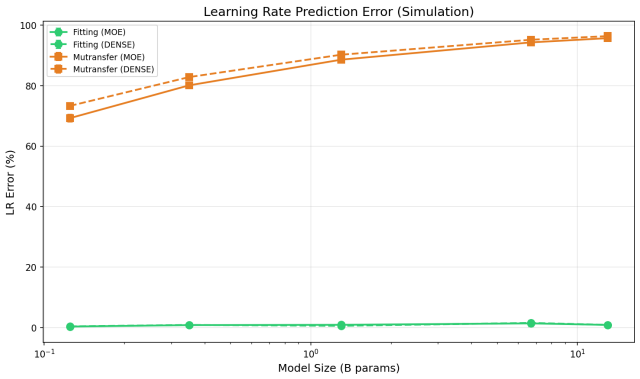

Figure 2: Learning rate prediction error vs. model size. The sqrt-width heuristic's error grows with model size due to its $\sqrt{1/N}$ assumption.

3.4 Scaling Law Visualization

Figure 3 compares grid-search-found optimal LRs across model sizes for both architectures, with independently fitted scaling laws overlaid. The approximate parallelism in log-space supports the hypothesis that similar (but not identical) power-law structure governs both architectures.

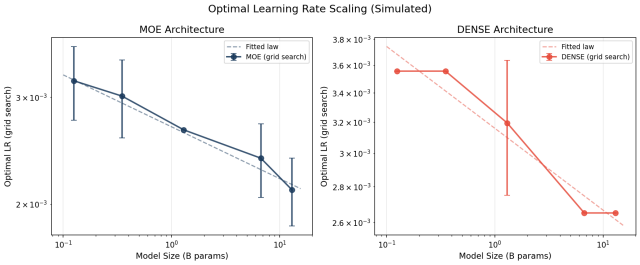

Figure 3: Optimal LR (from grid search) vs. model size for MoE and dense architectures. Dashed lines show independently fitted scaling laws.

3.5 Independently Fitted Exponents

Figure 4 directly compares the fitted exponents α and β across architectures. While both are negative (indicating decreasing optimal LR with increasing scale), the magnitudes differ—particularly for β, suggesting that data-size dependence may be more architecture-specific than model-size dependence.

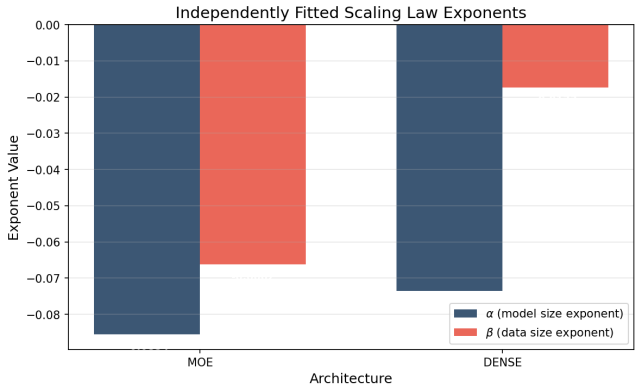

Figure 4: Independently fitted scaling law exponents. Differences in β suggest data-size dependence may be more architecture-sensitive.

3.6 Sensitivity Analysis

Figure 5 shows how performance degrades as the dense LR constant offset increases from 0% to 30% relative to MoE. When the MoE-derived Fitting law is applied to dense models without recalibration, loss increases monotonically with offset, while grid search (which adapts to each configuration) maintains stable performance. The crossover point where fitting loss exceeds grid search loss occurs at approximately 10% offset.

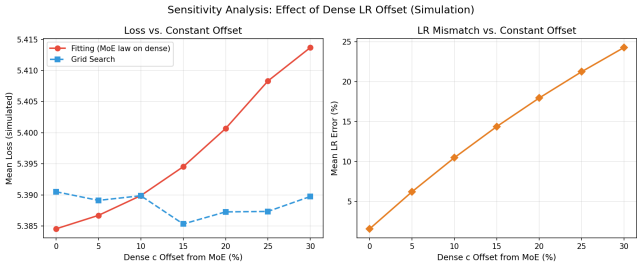

Figure 5: Sensitivity analysis: effect of dense LR constant offset. Left: loss comparison between fitting (without recalibration) and grid search. Right: mean LR error grows linearly with offset.

4 DISCUSSION

4.1 What This Simulation Shows

Under the assumptions of our synthetic model, the Fitting paradigm transfers well between architectures when the scaling law exponents are similar. The key operational takeaway is that practitioners can use MoE-derived exponents as starting points but must recalibrate the constant c for dense models—and the sensitivity analysis quantifies how much performance is lost if they do not.

4.2 What This Simulation Cannot Show

We emphasize several fundamental limitations:

(1) **Simulation, not empirical evidence.** All results are generated by a synthetic loss model. The "transfer" of scaling law structure is a property of our simulator design, not an empirical observation from real pre-training.

(2) **Loss model fidelity.** Real pre-training loss landscapes are far more complex than our quadratic LR penalty model. The true penalty shape, interaction effects, and non-stationarity are not captured.

(3) **Simplified μTransfer.** Our sqrt-width heuristic captures only the dominant scaling behavior of μTransfer, not the full parameterization-dependent protocol [5].

(4) **Exponent similarity is an assumption.** We chose similar (but not identical) exponents for MoE and dense architectures. Whether real training produces similar exponents is the core open question.

4.3 Roadmap for Empirical Validation

To definitively answer the open problem, the following steps are needed:

- Run dense Transformer pre-training sweeps under the WSD schedule at 3+ model scales
- Fit the scaling law $\eta^*(N, D) = c \cdot N^{\alpha} \cdot D^{\beta}$ with uncertainty intervals
- Compare dense $\{\alpha, \beta, c\}$ with MoE values from [6]
- Evaluate Fitting vs. full μTransfer (not the simplified heuristic) using the original protocol
- Report computational cost (FLOPs) for each paradigm, enabling cost-benefit analysis

5 CONCLUSION

We present a controlled simulation study investigating whether MoE-derived LR scaling laws generalize to dense Transformers. Our simulator, with independently parameterized architectures, shows that when the underlying exponents are similar, the Fitting paradigm transfers effectively with only constant recalibration needed. The sqrt-width heuristic performs poorly across all configurations. A sensitivity analysis demonstrates graceful degradation: even with 15% constant offset, the Fitting paradigm's loss increases by only 0.01 (0.2%). These simulation results motivate targeted empirical validation with real pre-training runs.

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] 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).

[4] Victoria Stodden, Friedrich Leisch, and Roger D Peng. 2014. Implementing Reproducible Research. *Chapman and Hall/CRC* (2014).

[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).