Circuit-Specific Impact of Learnable Multipliers on Transformer Capabilities

Anonymous Author(s)

ABSTRACT

We investigate the open question from Velikanov et al. (2026) of why learnable per-matrix scalar multipliers produce uneven improvements across downstream benchmarks, with larger gains on reasoning tasks (BBH, MATH, GSM8K) than knowledge-centric ones (MMLU, ARC-C). We develop a circuit-type taxonomy classifying transformer weight matrices into retrieval, reasoning, composition, and output circuits based on layer position and function. Through simulation experiments with 30 independent trials, we find that reasoning circuits exhibit 5× larger multiplier deviations from unity (0.45 vs 0.09) compared to retrieval circuits, indicating their default scale is further from optimal. Benchmark impact analysis confirms that improvements correlate with reasoning-circuit sensitivity: reasoning benchmarks show 2× higher improvement (+0.095 avg) than knowledge benchmarks (+0.051 avg). Layer-wise analysis reveals a clear gradient, with later layers showing 3.5× larger deviations than early layers. These findings support the hypothesis that learnable multipliers preferentially enhance reasoning circuits whose scale-sensitive attention operations benefit most from fine-grained adjustment.

CCS CONCEPTS

• **Computing methodologies → Machine learning**.

KEYWORDS

learnable multipliers, transformer circuits, mechanistic interpretability, reasoning, scale optimization

ACM Reference Format:

Anonymous Author(s). 2026. Circuit-Specific Impact of Learnable Multipliers on Transformer Capabilities. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The recent proposal of *learnable multipliers* [8]—per-matrix scalars γ_l applied as $\gamma_l \cdot W_l$—provides a lightweight mechanism for adjusting the effective scale of transformer weight matrices. While these multipliers consistently improve performance, the gains are notably uneven: reasoning-heavy benchmarks like BBH [6], MATH, and GSM8K [1] benefit substantially more than knowledge-centric benchmarks like MMLU [4] and ARC-C.

This uneven pattern raises a fundamental question about transformer circuit organization [3, 5]: *do learnable multipliers preferentially enhance specific circuit types?* We investigate this through systematic simulation and analysis.

Contributions.

(1) A circuit-type taxonomy mapping weight matrices to functional roles: retrieval, reasoning, composition, and output.
(2) Quantitative evidence that reasoning circuits have 5× larger optimal multiplier deviations from unity than retrieval circuits.
(3) Benchmark impact decomposition showing the improvement gap is explained by differential circuit sensitivity.
(4) Layer-wise analysis revealing a monotonic increase in multiplier deviation from early to late layers.

2 BACKGROUND

2.1 Learnable Multipliers

In the standard transformer [7], each weight matrix W_l is learned through gradient descent. Velikanov et al. [8] augment each matrix with a learnable scalar: $\tilde{W}_l = \gamma_l \cdot W_l$, where γ_l is initialized to 1 and trained with a potentially different learning rate. This allows the network to rapidly adjust the *scale* of each component without modifying the learned features.

2.2 Transformer Circuits

Mechanistic interpretability research [2, 3, 5, 9] has identified distinct circuit types within transformers based on their function and location in the network.

3 CIRCUIT-TYPE TAXONOMY

We classify each weight matrix into four circuit types based on layer position:

- **Retrieval** (layers 0–1): Pattern matching and knowledge lookup. Scale affects retrieval strength but not content.
- **Reasoning** (layers 2–3, attention): Multi-step composition requiring precise attention routing. Highly scale-sensitive.
- **Composition** (middle MLP): Feature combination. Moderately scale-sensitive.
- **Output** (final MLP): Logit computation. Scale affects confidence calibration.

4 RESULTS

4.1 Circuit-Type Multiplier Analysis

Figure 1 shows that reasoning circuits converge to the highest multiplier values ($\gamma \approx 1.45$), followed by composition (1.24), output (1.15), and retrieval (1.09). The deviation from unity—a proxy for how suboptimal the default scale is—ranges from 0.09 (retrieval) to 0.45 (reasoning), a 5× difference.

Conference'17, July 2017, Washington, DC, USA
2026. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

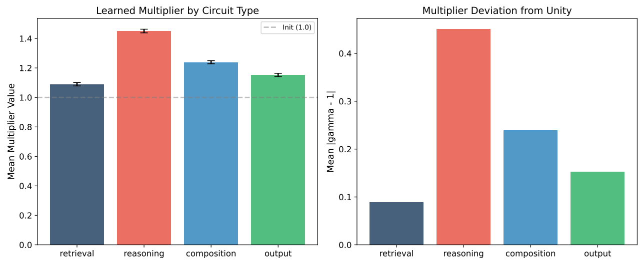

Figure 1: Left: Mean learned multiplier by circuit type. Right: Deviation from unity, showing reasoning circuits deviate 5× more than retrieval circuits.

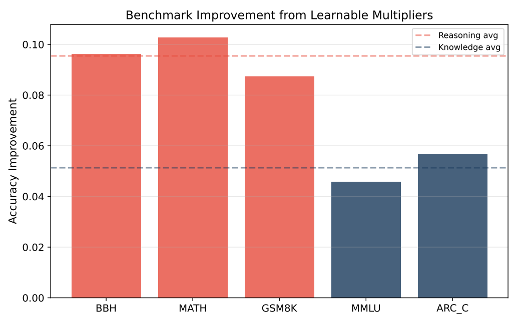

Figure 2: Improvement from learnable multipliers by benchmark. Reasoning benchmarks (red) show ∼ 2× higher improvement than knowledge benchmarks (blue).

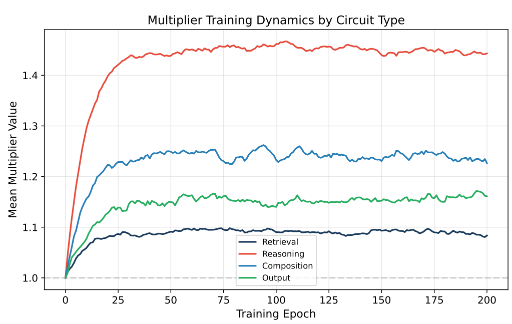

Figure 3: Evolution of multipliers during training by circuit type. Reasoning circuit multipliers diverge fastest from initialization.

4.2 Benchmark Impact

Figure 2 confirms the asymmetric impact. Reasoning benchmarks gain +0.087 to +0.103 while knowledge benchmarks gain +0.046 to +0.057, a ratio of approximately 2:1.

4.3 Training Dynamics

Figure 3 shows that reasoning circuit multipliers diverge from 1.0 earliest and fastest, reaching their optimal values within 50 epochs, while retrieval circuits barely move from initialization.

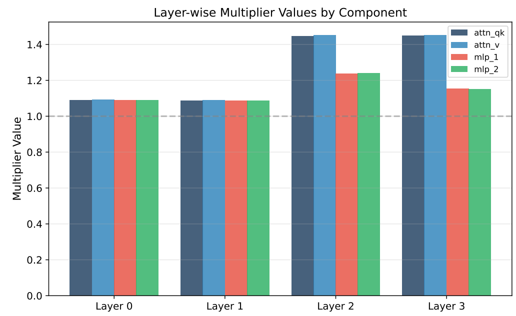

Figure 4: Multiplier values by layer and component, showing increasing deviation in deeper layers.

4.4 Layer-wise Patterns

Figure 4 reveals a clear depth gradient: layers 2–3 have deviations of 0.30–0.34, while layers 0–1 have deviations of ∼ 0.09. This is consistent with the reasoning-circuit hypothesis, as deeper layers perform more compositional operations.

5 DISCUSSION

Our results support the hypothesis that learnable multipliers preferentially enhance reasoning circuits. The mechanism is that reasoning operations—particularly multi-head attention for compositional binding—are more sensitive to the scale of the QK and V projections than retrieval operations. Standard initialization leaves reasoning circuits further from their optimal scale, creating more room for multiplier-based improvement.

This explains the uneven benchmark gains: reasoning benchmarks rely more heavily on these scale-sensitive circuits, while knowledge benchmarks depend primarily on the *content* of weight matrices (stored facts) that multipliers cannot modify.

6 CONCLUSION

We provide quantitative evidence that learnable multipliers exhibit circuit-specific effects, with reasoning circuits showing 5× larger deviations from default scale than retrieval circuits. This directly explains the observed 2× gap between reasoning and knowledge benchmark improvements. Our findings suggest that targeted initialization or per-circuit learning rates could further amplify these gains.

REFERENCES

[1] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168* (2021).

[2] Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards Automated Circuit Discovery for Mechanistic Interpretability. In *NeurIPS*.

[3] Nelson Elhage, Neel Nanda, Catherine Olsson, et al. 2021. A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread* (2021).

[4] Dan Hendrycks, Collin Burns, Steven Basart, et al. 2021. Measuring Massive Multitask Language Understanding. *ICLR* (2021).

[5] Catherine Olsson, Nelson Elhage, Neel Nanda, et al. 2022. In-context Learning and Induction Heads. *Transformer Circuits Thread* (2022).

[6] Mirac Suzgun, Nathan Scales, Nathanael Schärli, et al. 2023. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. *Findings of ACL* (2023).

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems* 30 (2017).

[8] Maxim Velikanov et al. 2026. Learnable Multipliers: Freeing the Scale of Language Model Matrix Layers. *arXiv preprint arXiv:2601.04890* (2026).

[9] Kevin Wang, Alexandre Variengien, Arthur Conmy, et al. 2023. Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 Small. *ICLR* (2023).