Is Grokking Worthwhile? Functional Analysis and Transferability of Generalization Circuits in Transformers

Anonymous ACL submission

Abstract

While Large Language Models (LLMs) excel at factual retrieval, they often struggle with the "curse of two-hop reasoning" in compositional tasks. Recent research suggests that parameter-sharing transformers can bridge this gap by forming a "Generalization Circuit" during a prolonged "grokking" phase. A fundamental question arises: Is a grokked model superior to its non-grokked counterparts on downstream tasks? Furthermore, is the extensive computational cost of waiting for the grokking phase worthwhile? In this work, we conduct a mechanistic study to evaluate the Generalization Circuit's role in knowledge assimilation and transfer. We demonstrate that: (i) The inference paths established by non-grokked and grokked models for in-distribution compositional queries are identical. This suggests that the "Generalization Circuit" does not represent the sudden acquisition of a new reasoning paradigm. Instead, we argue that grokking is the process of integrating memorized atomic facts into an naturally established reasoning path. (ii) Achieving high accuracy on unseen cases after prolonged training and the formation of a certain reasoning path are not bound; they can occur independently under specific data regimes. (iii) Even a mature circuit exhibits limited transferability when integrating new knowledge, suggesting that "grokked" Transformers do not achieve a full mastery of compositional logic.

1 Introduction

Implicit reasoning refers to a model's capacity to derive new information from its stored parameters without the aid of externalized tools. A primary failure mode, often termed the "curse of two-hop reasoning,"(Balesni et al., 2025) occurs when standard Transformers fail to compose separately learned facts. For example, a model may successfully recall that "*Barack's wife is Michelle*" and "*Michelle was born in 1964*," yet it remains unable to answer "*Barack's wife was born in?*" purely through internal computation.

This limitation is often attributed to a **representational mismatch** in the model's internal information flow (see Figure 1). In standard Transformers, atomic facts are typically retrieved and resolved in shallow layers. When performing compositional reasoning, the bridge entity b (e.g., *Michelle*) emerges as an *output* of these early layers. However, the subsequent deeper layers, which contain the parameters for the second hop, have not been trained to use this specific latent representation of b as an *input* for further retrieval (Wang et al., 2024).

Existing research has demonstrated that parameter-sharing Transformers, when trained significantly beyond the point of training-label saturation, can achieve correct implicit compositional reasoning on unseen cases—a phenomenon known as **"grokking"**. Subsequent mechanistic studies have revealed that this transition is driven by the emergence of a **"Generalization Circuit"**: a specific internal computation path where the model learns to internalize the reasoning chain by explicitly recovering necessary intermediate information within its middle layers (Power et al., 2022; Wang et al., 2024).

While prior work has sought to explain the mechanisms behind grokking (Wang et al., 2024; Varma et al., 2023; Nanda et al., 2023; Merrill et al., 2023; Huang et al., 2024; Gu et al., 2025; Furuta et al., 2024; AlquBoj et al., 2025; Doshi et al., 2024; Lv et al., 2025), leveraged the phenomenon to improve model performance (Abramov et al., 2025; Lyle et al., 2025; Murty et al., 2023; Furuta et al., 2024; Africa et al., 2025), and explored diverse methodologies to induce it (Lin et al., 2025; Lee et al., 2024; Xu et al., 2025; Zhou et al., 2025; Park et al., 2024; Carvalho et al., 2025), a central question remains: Is a transformer trained through grokking better than its counterpart? Furthermore, does the

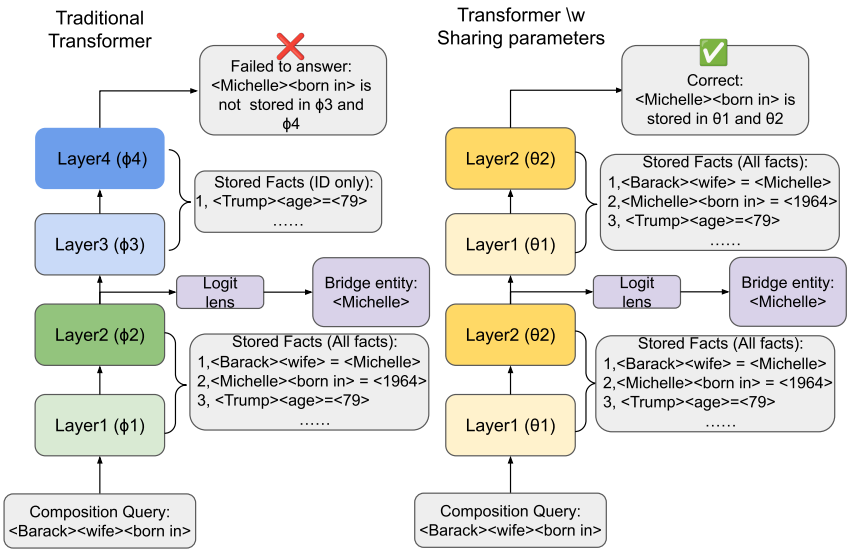

Figure 1: **Mechanistic Comparison Between traditional transformer and parameter-sharing transformer.** (Left) Standard Transformers struggle with representational mismatch across independent layers and are unable to generalize to compositional reasoning. (Right) Transformer with shared parameter resolves the bridge entity b and is able to use shallow parameters to get the final answer.

reasoning capability acquired through grokking effectively transfer to downstream tasks or novel facts?

In this work, we leverage the state-of-the-art parameter-sharing Transformer architecture (Jolicoeur-Martineau, 2025) on reasoning tasks to investigate the transferability of the grokked transformers on new facts. Our findings reveal that a grokked transformer with a Generalization Circuit may not be generalized as we expected:

Finding 1: Non-grokked transformer follows identical reasoning path as the grokked one on In-Distribution data, and grokking is the process of integrating memorized atomic facts into an naturally established reasoning structure. By providing sufficient atomic and compositional supervision, models can establish a stable Generalization Circuit and achieve equivalent transfer performance without undergoing the time-consuming grokking.

Finding 2: Grokking and the formation of a Generalization Circuit are not intrinsically linked, A "grokked" model may not necessarily reason through a Generalization Circuit, and a Generalization Circuit may not necessarily be obtained through grokking.

Finding 3: Following a Generalization Circuit does not imply a full mastery of compositional reasoning, while it accelerates the assimilation of new facts in specific setups, its ability to generalize reasoning patterns across domains remains limited compared to human-like reasoning.

2 Task definition & Training data

We follow the compositional task definition and data format established in recent literature (Wang et al., 2024; Ye et al., 2025; Guo et al., 2025; Brinkmann et al., 2024). Models are trained on two types of queries: **atomic queries**, such as *<Barack> <wife>?*", and **compositional queries**, such as *<Barack> <wife> <born in>?*". To evaluate systematic generalization rather than surface-level memorization, we partition the set of all atomic facts into In-Distribution (ID) and Out-of-Distribution (OOD) subsets. During training, the model is exposed to all atomic queries but only those compositional queries constructed from ID facts. At test time, the model must answer compositional queries that involve OOD facts. Success in this setting indicates that the model has acquired a genuine compositional reasoning ability. Following the schema of Wang et al. (2024), we construct a synthetic dataset to ensure precise control over reasoning paths. The dataset comprises 2,000 entities and 200 unique relations, with each entity associated with 20 relations, totaling 40,000 atomic facts. We reserve 5% (2,000 facts) as the OOD subset. More details on the construction of the dataset are shown in Appendix B.

3 Finding 1: Grokked and non-grokked transformers follow the same reasoning path

Using the logit lens (Belrose et al., 2025), we visualize the maturation of Generalization Circuits in

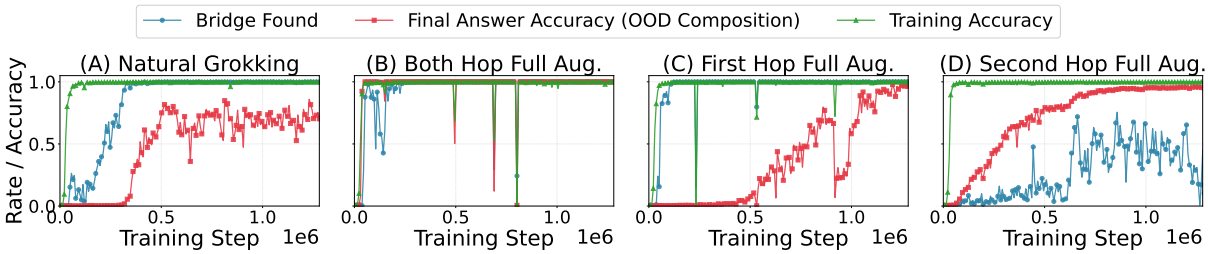

Figure 2: Transformer model "Grokking" under different data regime.

parameter-shared Transformers as they learn compositional reasoning. Specifically, we decode the hidden states across all layers with logit lens to identify the entities they represent. An inference trace is categorized as 'Bridge found' if the decoded hidden state matches the ground-truth bridge entity. We define the Generalization Circuit as the internal reasoning trace that achieves compositional reasoning by explicitly recovering the bridge entity within the model's intermediate layers (Wang et al., 2024). As shown in Figure 2 (A), a model exhibiting natural grokking establishes three distinct stages during training on OOD compostional queries. The **first stage** is the saturation of training targets, occurring at approximately 0.1M steps. In the **second stage**, which persists for an additional 0.2M steps, the model begins to consistently recover the correct bridge entity within its intermediate layers for OOD compositional queries. The **third stage** involves the model learning to effectively utilize these intermediate representations to derive the final answer, a process that stabilizes after 1M steps.

However, when providing compositional supervision on OOD facts, the final model trained with fewer steps follows the same reasoning path as naturally grokked models. Specifically, for **every** OOD facts appearing in the first hop of test queries, we randomly pair them with ID facts to construct *OOD Facts + ID Facts* queries (**First Hop Full Augmentation**); for **every** OOD facts appearing in the second hop of test queries, we construct *ID Facts + OOD Facts* queries (**Second Hop Full Augmentation**). Adding these queries ensures the model observes each OOD fact in both compositional positions, effectively equalizing the level of supervision between OOD and ID facts. Surprisingly, this supervision does not prevent the formation of the Generalization Circuit. As shown in Figure 2(B), within 0.3M training steps, every correct prediction on OOD compositions is achieved through successful recovery of the bridge entity in intermediate layers. Furthermore, we show in §5 and Figure 4 in Appendix A that the model trained under full supervision exhibits identical transferability when applied to new facts compared to a natural grokked one.

These findings suggest that the "Generalization Circuit" does not represent the sudden acquisition of a generalized reasoning paradigm. Instead, such a reasoning path is established in the early stages of training for ID facts. Crucially, this circuit is not directly applicable to all memorized facts; rather, the subsequent grokking phase progressively integrates sparsely supervised OOD facts into this established circuit. Consequently, a model trained extensively under sparse compositional supervision and a model trained rapidly under full supervision converge on the same solution: they are **functionally equivalent** in their mechanistic approach to compositional reasoning.

4 Finding 2: Behavioral Grokking or True Grokking?

First defined by Power et al. (2022), grokking is a phenomenon where a model suddenly achieves high accuracy on unseen cases long after saturating its performance on training targets. Previous studies have leveraged this phenomenon to improve model performance or have employed data augmentation to accelerate grokking (Abramov et al., 2025). However, we demonstrate that not all training processes that satisfy the behavioral description of grokking actually involve the formation of the intended underlying mechanism.

To investigate this, we conducted additional augmentation experiments on OOD facts. Similar to augmentation in § 3, but this time we only did the **Second Hop Full Augmentation**. As shown in Figure 2 (D), although the training curve fits the behavior description of grokking. Its underlying mechanism is different from Figure 2 (A) and (B), their final answer to composition queries does not come through a correctly recovered bridge entity during the middle layer. And we will show in the § 5 that such "Fake grokked" transformer lacks

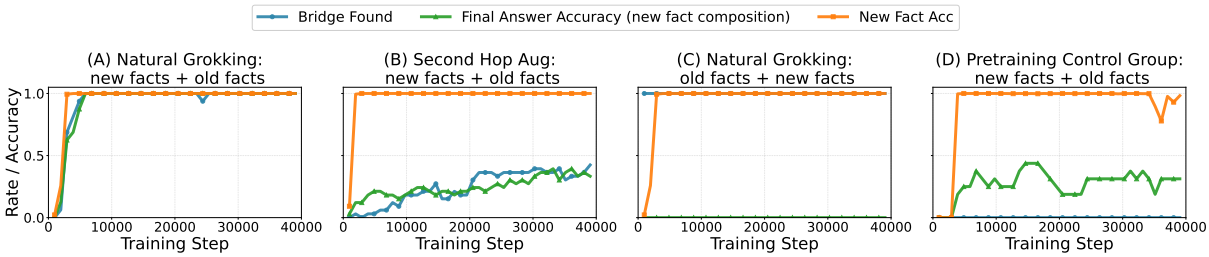

Figure 3: "Grokked" Transformer learning new knowledge, detailed visualization on all settings can be seen in Figure 4 in Appendix A

the same level of transferability compared to those transformers that really established Generalization Circuits.

5 Finding 3: Transferability of Generalization Circuit

Most evaluations of "Grokking" have been conducted using static datasets where the atomic facts are static. Intuitively, if a model has truly mastered a specific reasoning logic, it should exhibit human-like adaptability; once the reasoning rule is learned, it should be immediately applicable to novel facts. Despite this expectation, the downstream transferability of "grokked" Transformers remains largely underexplored.

5.1 Finetuning data

In the pre-training dataset, there are 200 relations in total, and each entity is connected to 20 relations. For each entity, we randomly sample relations that have not been previously used (i.e., unseen (entity, relation) pairs), and we instantiate new triples by assigning a target entity. Repeating this procedure yields entirely $2,000$ new atomic facts.

To prevent catastrophic forgetting of the original knowledge, we also retain a subset of 8,000 atomic facts from the original in-distribution (ID) training set. We then build the finetuning training split as follows: (1) the retained $8,000$ original ID atomic facts, (2) the $2,000$ newly added atomic facts, and (3) all compositional queries whose supporting atomic facts are included in the retained $8,000$ ID subset (so that these compositional queries remain answerable from the included old facts).

5.2 Limited transferability of Generalization Circuit

We use the pre-trained checkpoints obtained in § 3 and § 4 as base models for further fine-tuning on finetuning data. This setup enables us to assess the models' capacity to assimilate new atomic facts and their efficiency in resolving associated downstream compositional queries. The resulting learning curves are illustrated in Figure 3. We find that the Generalization Circuit learned during pretraining provides limited benefits. Models that have developed the intended Circuit can often recover the correct bridge entity for compositional queries involving new facts; it then leverages patterns acquired during pretraining to complete the second-hop reasoning instantly (See in Figure 3 (A)).

However, when a novel fact is required in the second hop, models fail to complete the inference despite correctly recovering the bridge entity, indicating that the second component of the Generalization Circuit is less transferable. To establish a reliable circuit for new knowledge, models still require either (i) an additional grokking phase, or (ii) substantially more finetuning data covering the new facts and their compositions. Finally, "Fake grokked" transformers, despite achieving comparable performance on original OOD facts, fail to adapt to compositional reasoning mixing new and old facts. As shown in Figure 3(B), these models memorize new atomic facts but do not learn to reason over them compositionally.

6 Conclusion

In this work, we mechanistically re-evaluate grokking and Generalization Circuits in Transformers. Our findings challenge the view that grokking represents genuine acquisition of generalized reasoning. We show that grokked and non-grokked models follow identical reasoning paths, revealing that grokking merely integrates memorized facts into naturally established circuits. We further demonstrate a dissociation between behavioral grokking and circuit formation, and expose limited transferability to novel facts. These results suggest that the decision to pursue grokking involves a trade-off between data scarcity and computational resources: training until grokking may be necessary under sparse supervision, but with sufficient compositional data, models can stop at training saturation and achieve equivalent performance.

7 Limitations

Despite the mechanistic insights provided by our study, several limitations remain.

Computational Constraints on Exploring Fake Grokking: First, our investigation into "fake grokking" (behavioral grokking without circuit formation) is not exhaustive. Since the emergence of grokking phenomena typically requires extremely prolonged training cycles (often exceeding 1M steps in our settings), conducting a large-scale grid search to identify all possible augmentation strategies that might trigger circuit collapse was computationally infeasible. Consequently, while we have demonstrated that certain augmentation regimes lead to fake grokking, the precise boundary conditions and the universal principles governing this transition remain an open question. Future research is needed to systematically categorize which data distributions favor circuit formation over pattern memorization.

Architectural Scope: Second, our findings are specifically constrained to parameter-sharing Transformer architectures. As established in recent literature, these models are currently unique in their ability to resolve the representational mismatch that typically hinders implicit compositional reasoning. Standard Transformers with independent layer weights inherently struggle to compose separately learned facts; their deeper layers are not trained to recognize the latent representations produced by earlier layers as valid inputs for subsequent retrieval. Therefore, while our results regarding circuit formation and collapse are robust within the context of shared-parameter models, they do not necessarily extend to traditional architectures, which appear unable to achieve this form of implicit generalization even with extensive training.

References

Roman Abramov, Felix Steinbauer, and Gjergji Kasneci. 2025. Grokking in the wild: Data augmentation for real-world multi-hop reasoning with transformers. In *Forty-second International Conference on Machine Learning*.

David Demitri Africa, Sara M. Kapoor, Theo Simon Sorg, and Challenger Mishra. 2025. Learning modular exponentiation with transformers. *Preprint*, arXiv:2506.23679.

H. V. AlquBoj, Hilal AlQuabeh, Velibor Bojkovic, Munachiso Nwadike, and Kentaro Inui. 2025. Mechanistic insights into grokking from the embedding layer. *Preprint*, arXiv:2505.15624.

Mikita Balesni, Tomek Korbak, and Owain Evans. 2025. Lessons from studying two-hop latent reasoning. *Preprint*, arXiv:2411.16353.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. 2025. Eliciting latent predictions from transformers with the tuned lens. *Preprint*, arXiv:2303.08112.

Jannik Brinkmann, Abhay Sheshadri, Victor Levoso, Paul Swoboda, and Christian Bartelt. 2024. A mechanistic analysis of a transformer trained on a symbolic multi-step reasoning task. *Preprint*, arXiv:2402.11917.

Breno W. Carvalho, Artur d'Avila Garcez, Luis C. Lamb, and Emílio Vital Brazil. 2025. Inducing grokking with distribution shifts. In *2025 IEEE 37th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 222–229.

Darshil Doshi, Aritra Das, Tianyu He, and Andrey Gromov. 2024. To grok or not to grok: Disentangling generalization and memorization on corrupted algorithmic datasets. *Preprint*, arXiv:2310.13061.

Hiroki Furuta, Gouki Minegishi, Yusuke Iwasawa, and Yutaka Matsuo. 2024. Towards empirical interpretation of internal circuits and properties in grokked transformers on modular polynomials. *Transactions on Machine Learning Research*.

Zihan Gu, Ruoyu Chen, Hua Zhang, Yue Hu, and Xiaochun Cao. 2025. Beyond progress measures: Theoretical insights into the mechanism of grokking. *Preprint*, arXiv:2504.03162.

Tianyu Guo, Hanlin Zhu, Ruiqi Zhang, Jiantao Jiao, Song Mei, Michael I. Jordan, and Stuart Russell. 2025. How do llms perform two-hop reasoning in context? *Preprint*, arXiv:2502.13913.

Yufei Huang, Shengding Hu, Xu Han, Zhiyuan Liu, and Maosong Sun. 2024. Unified view of grokking, double descent and emergent abilities: A perspective from circuits competition. *Preprint*, arXiv:2402.15175.

Alexia Jolicoeur-Martineau. 2025. Less is more: Recursive reasoning with tiny networks. *Preprint*, arXiv:2510.04871.

Jaerin Lee, Bong Gyun Kang, Kihoon Kim, and Kyoung Mu Lee. 2024. Grokfast: Accelerated grokking by amplifying slow gradients. *Preprint*, arXiv:2405.20233.

Pengxiao Lin, Zheng-An Chen, and Zhi-Qin John Xu. 2025. Identity bridge: Enabling implicit reasoning via shared latent memory. *Preprint*, arXiv:2509.24653.

Ang Lv, Ruobing Xie, Xingwu Sun, Zhanhui Kang, and Rui Yan. 2025. Language models "grok" to copy. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 735–741, Albuquerque, New Mexico. Association for Computational Linguistics.

Clare Lyle, Gharda Sokar, Razvan Pascanu, and Andras Gyorgy. 2025. What can grokking teach us about learning under nonstationarity? *arXiv preprint arXiv:2507.20057*.

William Merrill, Nikolaos Tsilivis, and Aman Shukla. 2023. A tale of two circuits: Grokking as competition of sparse and dense subnetworks. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher D. Manning. 2023. Grokking of hierarchical structure in vanilla transformers. *CoRR*, abs/2305.18741.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. *Preprint*, arXiv:2301.05217.

Yeachan Park, Minseok Kim, and Yeoneung Kim. 2024. Acceleration of grokking in learning arithmetic operations via kolmogorov-arnold representation. *Preprint*, arXiv:2405.16658.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. Grokking: Generalization beyond overfitting on small algorithmic datasets. *Preprint*, arXiv:2201.02177.

Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. 2023. Explaining grokking through circuit efficiency. *Preprint*, arXiv:2309.02390.

Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. 2024. Grokking of implicit reasoning in transformers: A mechanistic journey to the edge of generalization. In *Advances in Neural Information Processing Systems*, volume 37, pages 95238–95265. Curran Associates, Inc.

Zhiwei Xu, Zhiyu Ni, Yixin Wang, and Wei Hu. 2025. Let me grok for you: Accelerating grokking via embedding transfer from a weaker model. In *The Thirteenth International Conference on Learning Representations*.

Jiaran Ye, Zijun Yao, Zhidian Huang, Liangming Pan, Jinxin Liu, Yushi Bai, Amy Xin, Liu Weichuan, Xiaoyin Che, Lei Hou, and Juanzi Li. 2025. How do transformers learn implicit reasoning? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Xinyu Zhou, Simin Fan, Martin Jaggi, and Jie Fu. 2025. Neuralgrok: Accelerate grokking by neural gradient transformation. *Preprint*, arXiv:2504.17243.

A Figures

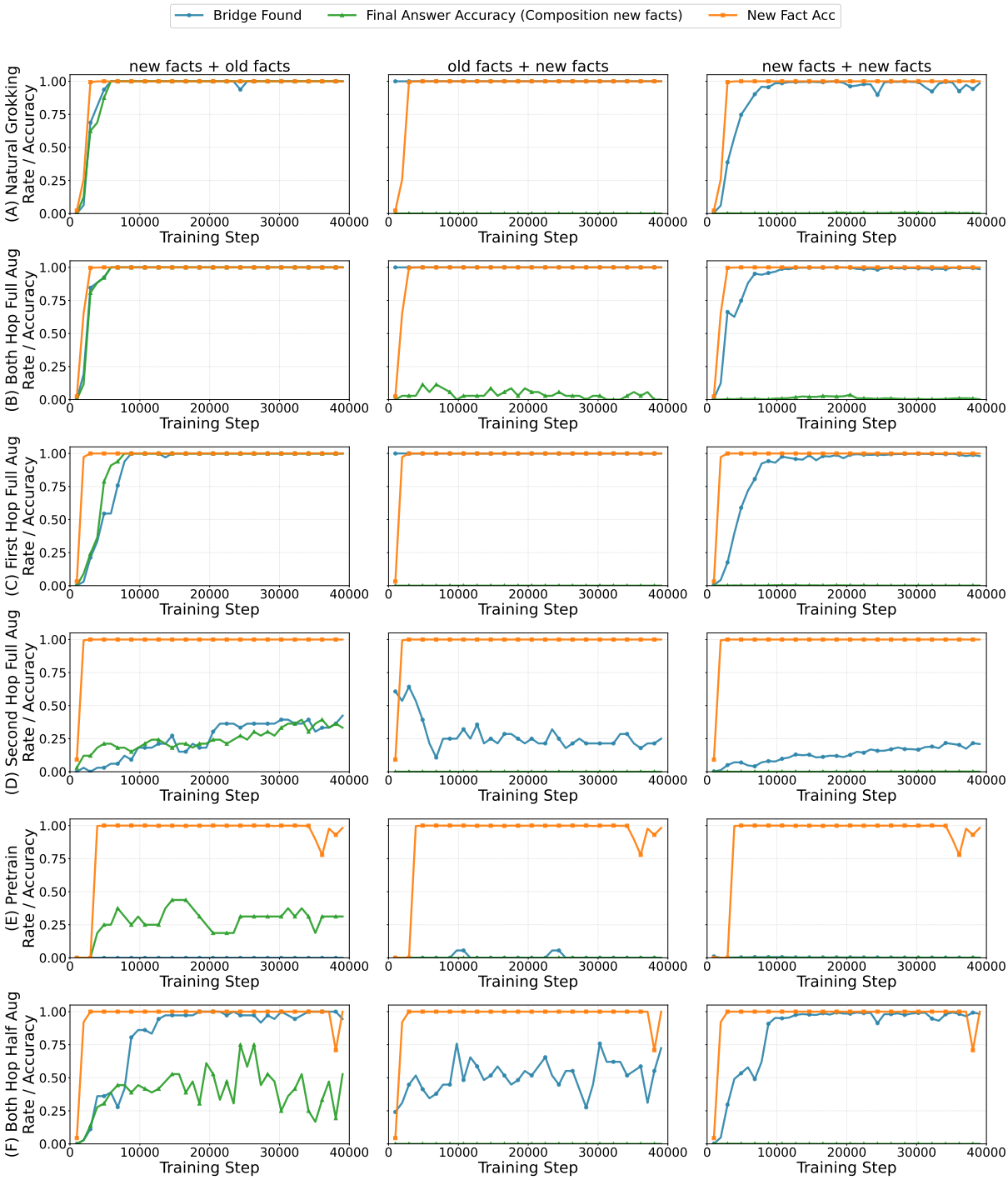

Figure 4: “Grokked” transformer finetuned on new facts.

B Pre-training Data Construction

We adopt a synthetic data generation framework following Wang et al. (2024) to ensure precise control over reasoning paths and data distributions.

B.1 Synthetic Dataset Generation

The dataset is modeled as a directed knowledge graph G, consisting of $|\mathcal{E}| = 2,000$ entities and $|\mathcal{R}| = 200$ unique relations. Each entity serves as a subject for 20 unique relations, resulting in a total of $40,000$ atomic triplets.

To evaluate systematicity, these facts are partitioned into In-Distribution (C_{ID}) and Out-of-Distribution (C_{OOD}) subsets using a 95%:5% split, yielding 38,000 ID facts and 2,000 OOD facts. While the model is exposed to all atomic facts during training, compositional queries involving OOD facts are reserved strictly for testing true reasoning capability.

B.2 Two-Hop Composition Rule

The core task involves inducing the latent rule of two-hop composition:

$$(h, r_1, b) \wedge (b, r_2, t) \Rightarrow (h, r_1, r_2, t) \qquad (1)$$

where h, b, t denote the head, bridge, and tail entities, respectively. The model is trained to predict t given (h, r_1, r_2) by internally retrieving and composing the two supporting atomic facts.

B.3 Choice of Inferred/Atomic Ratio (ϕ)

The ratio $\phi = |train_inferred_{ID}|/|atomic_{ID}|$ serves as the primary determinant for the speed of grokking. We explicitly set $\phi = 18.0$ for our experiments to facilitate efficient observation of the Generalization Circuit.

According to the circuit efficiency hypothesis, a high ratio like 18.0 increases the relative complexity of a memorizing circuit compared to a generalizing one. This high-density supervision incentivizes the model to transition toward a systematic reasoning circuit more rapidly. Specifically, at $\phi = 18.0$, models have been shown to achieve high generalization accuracy even before training performance saturates, providing a clear window for mechanistic analysis.

C Data Augmentation Detail

To evaluate the impact of supervision on the emergence of the Generalization Circuit, we implement targeted augmentation strategies that incorporate OOD facts into compositional training queries.

Fact-level compositional augmentation (ID–OOD mixing). Given an atomic split $C_{\text{ID}} \cup C_{\text{OOD}}$ and an OOD compositional evaluation set $\mathcal{Q}_{\text{OOD}}$ of queries $(h, r_1, r_2) \mapsto t$, we first *extract the OOD hop facts that actually appear in the evaluation compositions*. For each $(h, r_1, r_2) \mapsto t \in \mathcal{Q}_{\text{OOD}}$, we recover the bridge entity b, so that the corresponding hop facts are (h, r_1, b) and (b, r_2, t); we then collect the subsets $\mathcal{F}^{(1)}_{\text{OOD}} = \{(h, r_1, b) \in C_{\text{OOD}}\}$ and $\mathcal{F}^{(2)}_{\text{OOD}} = \{(b, r_2, t) \in C_{\text{OOD}}\}$ that occur as hop-1 / hop-2 components in $\mathcal{Q}_{\text{OOD}}$.

- **Hop-1 injection (OOD hop-1 + ID hop-2).** For each selected OOD hop-1 fact $(h, r_1, b) \in \mathcal{F}^{(1)}_{\text{OOD}}$, we generate additional training compositions by sampling an *ID* hop-2 fact $(b, r_x, t_x) \in C_{\text{ID}}$ that shares the same bridge b, and add the inferred training example $(h, r_1, r_x) \mapsto t_x$.
- **Hop-2 injection (ID hop-1 + OOD hop-2).** For each selected OOD hop-2 fact $(b, r_2, t) \in \mathcal{F}^{(2)}_{\text{OOD}}$, we generate additional training compositions by sampling an *ID* hop-1 fact $(h_y, r_y, b) \in C_{\text{ID}}$ that ends at the same bridge b, and add the inferred training example $(h_y, r_y, r_2) \mapsto t$.

Both hop full augmentation setting: We expose *all* OOD hop facts extracted from $\mathcal{Q}_{\text{OOD}}$. Each OOD hop fact is injected into compositional training queries. Importantly, the augmentation *does not* add the original OOD test compositions where *both* hops are OOD; instead, every injected query is constructed to contain *exactly one* OOD hop fact and one ID hop fact.