Multi-Hop Knowledge Composition is Bound by Pretraining Exposure

Anonymous ACL submission

Abstract

Large Language Models fail at implicit multi-hop reasoning: a model answers *"When was X born?"* and *"Who is Y's closest friend?"* correctly but fails on *"When was Y's closest friend born?"* in a single forward pass, even when both facts are perfectly memorized and individually retrievable. We study this failure in a controlled natural language setting with a strict separation between individuals exposed to compositional contexts during pretraining and those that never appear in any such context. We confirm that compositional failure persists even at 97% 1-hop accuracy, establishing the gap as a pretraining failure rather than a knowledge absence. We propose and test nine data-centric augmentation formats and find that compositional pretraining transfers to unseen questions for exposed individuals, but never to individuals absent from compositional pretraining, suggesting that exposure to compositional contexts during pretraining is a necessary condition for implicit multi-hop reasoning. Our code is available online.

1 Introduction

Large language models (LLMs) store and retrieve factual knowledge with surprising fidelity (Allen-Zhu and Li, 2024; Petroni et al., 2019), yet struggle on the simplest form of implicit multi-hop reasoning. A model that correctly answers *"When was Thierry born?"* and *"Who is Zinedine's closest friend?"* may fail on *"When was Zinedine's closest friend born?"*, a query where the bridge entity *Thierry* is absent, even when both constituent facts are individually extractable. This failure, the compositionality gap (Press et al., 2023), is robust to model scale and persists across architectures (Xu et al., 2024). Existing solutions externalize composition at inference time via chain-of-thought, decomposition, or model patching (Wei et al., 2022; Press et al., 2023; Biran et al., 2024), leaving open whether compositional reasoning can emerge from pretraining itself. On proprietary models, the cause remains undiagnosable as pretraining data is undisclosed. Allen-Zhu and Li (2024) show that knowledge memorization and extraction are distinct in controlled biography settings but study only single-hop retrieval. Ye et al. (2026) show that second-hop generalization requires query-level exposure but rely on symbolic tokens (Wang et al. (2024a)) without any augmentation.

In this short paper, we ask whether data-centric pretraining augmentations can induce implicit multi-hop composition in a controlled natural language setting. We extend the synthetic biography framework of Allen-Zhu and Li (2024) with inter-individual relations and partition individuals into a composed population and a held-out population restricted to atomic 1-hop biographies only. We then finetune on a subset of exposed individuals and evaluate on the remainder, separating transfer across unseen questions from transfer across individuals never seen in any compositional context during pretraining. Compositional augmentation substantially improves multi-hop QA on unseen questions for exposed individuals, but never transfers to those absent from compositional pretraining, establishing pretraining exposure as a necessary condition for implicit multi-hop composition. Our setup is summarized in Figure 1.

Our contributions are as follows: *(i)* we release an open reproduction of the biography framework (Allen-Zhu and Li, 2024) extended with inter-individual relations and multi-hop QA, *(ii)* we confirm in natural language that a model at 97% 1-hop accuracy scores near 0% on 2-hop queries, establishing the compositionality gap as a pretraining failure, *(iii)* across 9 augmentation formats, compositional pretraining transfers to unseen questions for exposed individuals but never to the held-out population.

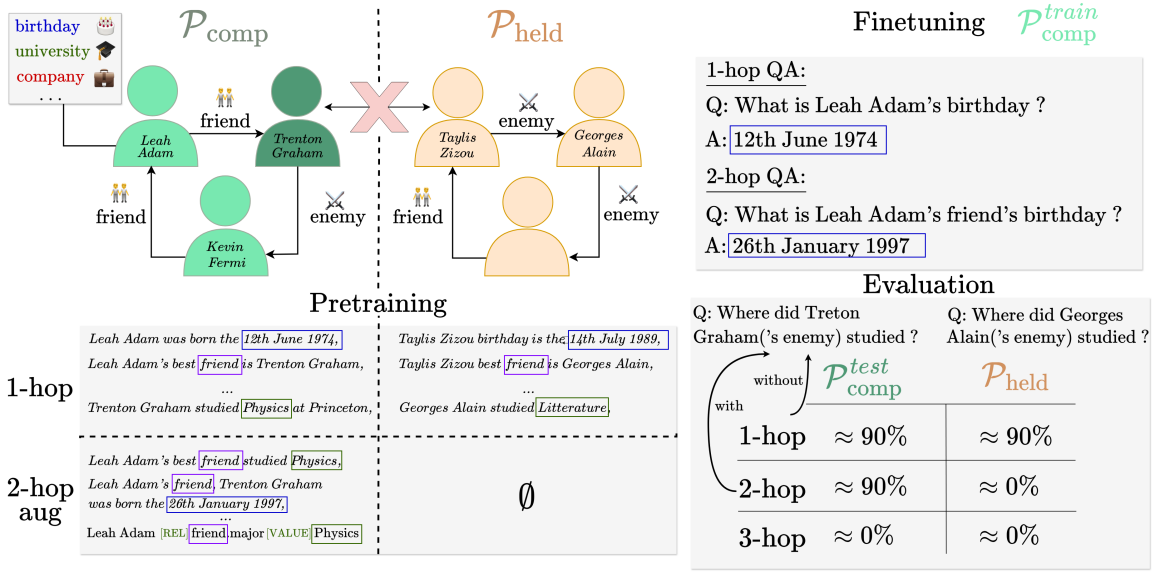

Figure 1: **Overview.** *(i)* We study the compositionality gap in a controlled natural language setting, extending the synthetic biography framework of Allen-Zhu and Li (2024) with `friend` and `enemy` relations to enable implicit reasoning on multi-hop QA. *(ii)* We pretrain on all 1-hop biographies ($\mathcal{P}_{\text{comp}} + \mathcal{P}_{\text{held}}$) and perform 2-hop data augmentation only on $\mathcal{P}_{\text{comp}}$. *(iii)* We then perform multi-hop QA fine-tuning on $\mathcal{P}_{\text{comp}}^{\text{train}}$, which transfers well to $\mathcal{P}_{\text{comp}}^{\text{test}}$ depending on the augmentation strategy, but never, under any tested condition, to $\mathcal{P}_{\text{held}}$.

2 Related Work

Compositionality gap and implicit reasoning. Language models struggle to implicitly compose individually retrievable facts within a single forward pass (Press et al., 2023), and scaling can worsen this by favoring memorization shortcuts (Wang et al., 2025). This setting is particularly challenging because the required reasoning is structurally simple, yet composition must occur entirely in parametric knowledge without external context. Compositional failures in parametric settings do not extend to in-context learning (Zhu et al., 2024; Allen-Zhu and Li, 2025), motivating separate study of these regimes. Existing approaches externalize composition through chain-of-thought, decomposition, or external modules (Wei et al., 2022; Press et al., 2023; Chen et al., 2024).

Mechanistic approaches. Mechanistic studies trace compositional failure to insufficient propagation of intermediate entities (Li et al., 2024; Biran et al., 2024; Hou et al., 2023). Biran et al. (2024) filter behavioral shortcuts at evaluation time, but co-occurrence of constituent facts during pretraining remains uncontrolled, leaving open whether success reflects genuine composition or joint memorization. Controlled studies isolate the phenomenon but remain limited: Ye et al. (2026) use symbolic tokens without augmentation, while Wang et al. (2024a) find implicit reasoning emerges via grokking with systematic OOD failure. Balesni et al. (2024) show that facts learned in separate documents fail to compose. Wang et al. (2024b) show that augmentingpretraining with random-walk paths over knowledge graphs improves multi-hop reasoning, but without a strict separation between training and test entities. We extend this line of work to natural language, with strict co-occurrence control and a systematic evaluation of data-centric pretraining interventions.

3 Controlled Multi-Hop Setting

We present here our controlled natural language setting to study implicit multi-hop knowledge reasoning. We describe here successively the dataset and population partition, the pretraining setup, our new data augmentation strategy, as well as the fine-tuning and evaluation protocol.

Dataset construction. We build on the synthetic biography framework of Allen-Zhu and Li (2024): $N = 100$K individuals, each described by six attributes (`birthday`, `birthcity`, `university`, `major`, `company`, `workcity`). We add a unique directional `friend` and `enemy` relation per individual, enabling multi-hop queries such as *What is the birthday of X's friend's enemy?*, where the bridge entity $Y = X$.friend is absent from the query. We partition individuals into $\mathcal{P}_{\text{comp}}$, exposed to compo-

# Exp.	Format	Natural Language (NL)		RDF		
		Explicit 2-hop	Implicit 2-hop	1-hop	Explicit 2-hop	Implicit 2-hop
Exp. 0	Baseline (no augmentation)	✗	✗	✗	✗	✗
Exp. 1	PT RDF	✗	✗	✓	✗	✗
Exp. 2	PT 2-hop implicit NL	✗	✓	✗	✗	✗
Exp. 3	PT 2-hop explicit NL	✓	✗	✗	✗	✗
Exp. 4	PT 2-hop implicit RDF	✗	✗	✗	✗	✓
Exp. 5	PT 2-hop explicit RDF	✗	✗	✗	✓	✗
Exp. 6	PT 2-hop implicit + explicit RDF	✗	✗	✗	✓	✓
Exp. 7	PT 2-hop implicit + explicit NL	✓	✓	✗	✗	✗
Exp. 8	PT 2-hop implicit + explicit NL-RDF	✓	✓	✗	✓	✓
Exp. 9	All formats	✓	✓	✓	✓	✓

Table 1: Data augmentation strategy for pretraining (PT) over $\mathcal{P}_{\text{comp}}$. All conditions include natural-language 1-hop biographies with `multi5p-permute` augmentation for every individuals. Explicit settings include the bridge entity.

sitional pretraining contexts, and $\mathcal{P}_{\text{held}}$, restricted to atomic biographies only. Relations are defined exclusively within each population ensuring that $\mathcal{P}_{\text{held}}$ individuals never appear as intermediate entities in any compositional chain.

Pretraining setup. We pretrain GPT-2 Small and evaluate scalability on GPT-2 Medium and Large (Radford et al., 2019), all trained from scratch with rotary positional embeddings (Su et al., 2024). Training batches mix atomic biographies from all N individuals with compositional augmentation sequences drawn exclusively from $\mathcal{P}_{\text{comp}}$. Atomic biographies follow the `multi5p-permute` format of Allen-Zhu and Li (2024): five paraphrased versions per individual with permuted attribute order, shown to be necessary for reliable 1-hop extraction. Dataset statistics and token counts are reported in Table 7.

Augmentation strategy. We study 9 augmentation conditions varying two axes: data format (natural language vs. RDF) and bridge verbalization (explicit vs. implicit), summarized in Table 1. Prior work uses only implicit symbolic contexts (Ye et al., 2026). We additionally hypothesize that explicitly verbalizing the bridge entity may align representations between explicit and implicit compositions, making the intermediate entity easier to retrieve during inference. Since composition becomes substantially easier once the bridge entity is identified (Biran et al., 2024), explicit and implicit formulations may reinforce a shared compositional representation. As we later show, explicit augmentations improve bridge-entity localization, although this alone does not yield compositional transfer (Section C.5). RDF isolates relational structure from lexical variation. Mix ratios and format examples are in Section B.5 and Table 5.

Finetuning and evaluation task. After pretraining we finetune on 75% of individuals composed during pretrain $\mathcal{P}_{\text{comp}}^{\text{train}}$ We evaluate 1-hop, 2-hop, and 3-hop reasoning in a single forward pass without access to intermediate entities in context. Following Allen-Zhu and Li (2024), we measure performance on $\mathcal{P}_{\text{comp}}^{\text{train}}$ and $\mathcal{P}_{\text{comp}}^{\text{test}}$ using first-token accuracy, and on $\mathcal{P}_{\text{held}}$ using exact-match accuracy on a single run with fixing seed. All training and optimization configurations are provided in Section B.

4 Empirical Results

We evaluate multi-hop QA accuracy across populations and augmentation conditions. We diagnose the compositional gap under standard training, then systematically evaluate the 9 data augmentations. Additional experiments are presented in Section C.

(i) Compositional gap diagnosis. Under 1-hop finetuning (Table 3), models reach near-perfect 1-hop accuracy on both populations, confirming `multi5p` resolves 1-hop extraction (Allen-Zhu and Li, 2024), but 2-hop and 3-hop remain at chance. Adding 2-hop questions to finetuning marginally improves $\mathcal{P}_{\text{comp}}$ (0.08) but not $\mathcal{P}_{\text{held}}$, despite perfectly memorized 1-hop facts. Our LoRA sweep ($r_{\text{qv}} \in \{8, 16, 32\}$, $r_{\text{emb}} \in \{32, 64, 128\}$) yields no improvement, and full finetuning converges on $\mathcal{P}_{\text{comp}}$ at the cost of catastrophic forgetting on $\mathcal{P}_{\text{held}}$ (Table 8), indicating that compositional information is absent from pretrained representations. This pattern persists across model scales (Table 4), as well as the guarantee that we are not in an undertraining scheme (Table 7), this is consistent with findings that the compositionality gap is robust to model scale (Press et al., 2023). The compositional gap is a pretraining failure, not a capacity limit.

# Exp.	Data augmentation	$\mathcal{P}_{\text{comp}}^{\text{test}}$			$\mathcal{P}_{\text{held}}$		
		1-hop	2-hop	3-hop	1-hop	2-hop	3-hop
Exp. 1	PT RDF	0.97	0.08	0.08	0.97	0.01	0.01
Exp. 2	PT 2-hop implicit NL	0.88	0.62	0.05	0.75	0.01	0.01
Exp. 3	PT 2-hop explicit NL	0.97	0.08	0.08	0.89	0.01	0.01
Exp. 4	PT 2-hop implicit RDF	0.97	0.79	0.05	0.40	0.01	0.01
Exp. 5	PT 2-hop explicit RDF	0.98	0.08	0.08	0.38	0.02	0.02
Exp. 6	PT 2-hop implicit + explicit RDF	0.98	0.79	0.15	0.50	0.01	0.01
Exp. 7	PT 2-hop implicit + explicit NL	0.91	0.73	0.06	0.79	0.01	0.01
Exp. 8	PT 2-hop implicit + explicit NL-RDF	0.99	0.83	0.04	0.83	0.01	0.01
Exp. 9	All formats	0.99	0.79	0.14	0.80	0.01	0.01

Table 2: **Results across augmentation conditions.** First-token accuracy for 1, 2 and 3-hop queries on $\mathcal{P}_{\text{comp}}^{\text{test}}$ (unseen at finetuning) and $\mathcal{P}_{\text{held}}$ (never compositionally exposed). Cell color encodes accuracy from low to high. 2-hop accuracy on $\mathcal{P}_{\text{held}}$ is at chance across all conditions under both first-token and exact-match evaluation.

Finetuning	Pop.	1-hop	2-hop	3-hop
1-hop only	$\mathcal{P}_{\text{comp}}$	1.00	0.01	0.01
	$\mathcal{P}_{\text{held}}$	0.97	0.01	0.01
1+2-hop	$\mathcal{P}_{\text{comp}}$	1.00	0.08	0.01
	$\mathcal{P}_{\text{held}}$	0.93	0.01	0.01

Table 3: **Baseline LoRA finetuning.** Near-perfect 1-hop accuracy on both populations under both regimes. 2-hop stays marginal on $\mathcal{P}_{\text{comp}}$ and at chance on $\mathcal{P}_{\text{held}}$.

Model	1-hop	2-hop	3-hop
GPT-2 small (124M)	0.93	0.01	0.01
GPT-2 medium (354M)	0.94	0.01	0.01
GPT-2 large (774M)	0.94	0.01	0.01

Table 4: **Ablation: scale analysis** on $\mathcal{P}_{\text{held}}$ under 1+2-hop finetuning. The compositional gap persists regardless of model size.

(ii) Data-centric augmentation. Table 2 reports accuracy across the 9 augmentation conditions, highlighting four key dynamics. First, pretraining exposure is a strict prerequisite for multi-hop composition. 2-hop accuracy on $\mathcal{P}_{\text{held}}$ remains at chance in first token accuracy or exact match (0.01) across all conditions, showing that no data-centric augmentation compensates for missing exposure, consistent with symbolic findings on OOD triplets (Ye et al., 2026). Second, implicit augmentation consistently outperforms explicit augmentation in isolation. Although explicit formats yield stronger bridge-entity signals (Figure 2), this does not translate into compositional gains. Explicit-only setups (Exp. 3, 5) match baseline performance (0.08) on $\mathcal{P}_{\text{comp}}^{\text{test}}$, whereas implicit NL (Exp. 2) reaches 0.62 and implicit RDF (Exp. 4) 0.79. This suggests that explicit supervision encourages direct association learning, while implicit supervision better matches the inference setting where the bridge entity is absent. Explicit augmentation becomes beneficial when combined with implicit formats: implicit+explicit RDF (Exp. 6) reaches 0.79 and NL+RDF (Exp. 8) peaks at 0.83 on $\mathcal{P}_{\text{comp}}^{\text{test}}$, suggesting that mixing explicit and implicit signals improves representation alignment across NL and RDF formats. Notably, implicit RDF alone (Exp. 4, 0.79) matches this performance at a fraction of the data cost (Table 7), suggesting structured triples as a lightweight and effective format for compositional pretraining augmentation. Transfer to $\mathcal{P}_{\text{held}}$ remains zero under all conditions. Finally, gains on $\mathcal{P}_{\text{comp}}^{\text{test}}$ come at a cost to 1-hop retention (Exp. 4: 0.40), likely due to dilution of atomic biographies in training batches. We fix the atomic/compositional ratio at 30/70 following prior compositional supervision regimes (Ye et al., 2026; Wang et al., 2024a). Mixing details in Section B.5.

5 Conclusion

We demonstrate that data-centric pretraining augmentation can induce multi-hop composition for exposed individuals, but never transfers to individuals absent from compositional pretraining contexts, confirming the fundamental limit of this approach beyond symbolic settings (Ye et al., 2026). Models fail on compositions of length two for unexposed individuals, suggesting augmentation induces entity-specific associations rather than a reusable composition logic. We also show that mixing graph with natural language enables composition over knowledge graph entities, consistent with recent efforts to unify LLMs and knowledge graphs (Pan et al., 2024). We release code, data, and our reproduction of Allen-Zhu and Li (2024) to support future investigation into training objectives or architectural modifications that may overcome this limit.

Limitations

Due to limitied computational ressources our augmentation experiments use GPT-2 small to large. We argue scale would not change our central claim, as the compositionality gap persists up to 175B parameters (Press et al., 2023) and our own baseline confirms near-zero 2-hop accuracy from 124M to 774M. More fundamentally, our claim concerns exposure, not capacity: $\mathcal{P}_{\text{held}}$ fails because its individuals never appear in a compositional context during pretraining, and no amount of parameters supplies a missing training signal. However, LLMs trained on long-context naturally co-occurring text raise a separate question: whether naturally co-occurring related facts in long-context pretraining corpora could implicitly substitute for explicit compositional supervision remains an open question our synthetic setting cannot address.

We focus on data-centric interventions under standard pretraining and supervised finetuning, and do not evaluate alternative training objectives such as reinforcement learning (Hatamizadeh et al., 2026) or knowledge distillation (Yu et al., 2024), which may offer complementary solutions to the compositional gap. We evaluate a single mixing ratio per condition, argued to be a conservative choice in Section B.5. Our setting is fully synthetic, and generalization to naturally occurring corpora remains untested. We restrict evaluation to two relation types (`friend`, `enemy`), leaving open whether findings extend to richer relational structures.

References

Zeyuan Allen-Zhu and Yuanzhi Li. 2024. Physics of language models: Part 3.1, knowledge storage and extraction. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 1067–1077. PMLR / OpenReview.net.

Zeyuan Allen-Zhu and Yuanzhi Li. 2025. Physics of language models: Part 3.2, knowledge manipulation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Mikita Balesni, Tomasz Korbak, and Owain Evans. 2024. The two-hop curse: LLMs trained on A->B, B->C fail to learn A->C. *CoRR*, abs/2411.16353.

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. 2024. Hopping too late: Exploring the limitations of large language models on multi-hop queries. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 14113–14130. Association for Computational Linguistics.

Jiaao Chen, Xiaoman Pan, Dian Yu, Kaiqiang Song, Xiaoyang Wang, Dong Yu, and Jianshu Chen. 2024. Skills-in-context: Unlocking compositionality in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Findings of ACL, pages 13838–13890. Association for Computational Linguistics.

Ali Hatamizadeh, Syeda Nahida Akter, Shrimai Prabhumoye, Jan Kautz, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, and Yejin Choi. 2026. RLP: Reinforcement as a pretraining objective. In *The Fourteenth International Conference on Learning Representations*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Yifan Hou, Jiaoda Li, Yu Fei, Alessandro Stolfo, Wangchunshu Zhou, Guangtao Zeng, Antoine Bosselut, and Mrinmaya Sachan. 2023. Towards a mechanistic interpretation of multi-step reasoning capabilities of language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4902–4919. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Zhaoyi Li, Gangwei Jiang, Hong Xie, Linqi Song, Defu Lian, and Ying Wei. 2024. Understanding and patching compositional reasoning in LLMs. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Findings of ACL, pages 9668–9688. Association for Computational Linguistics.

nostalgebraist. 2020. Interpreting GPT: the logit lens. LessWrong. Accessed: May 26, 2026.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Findings of ACL, pages 5687–5711. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. 2024a. Grokking of implicit reasoning in transformers: A mechanistic journey to the edge of generalization. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Xinyi Wang, Alfonso Amayuelas, Kexun Zhang, Liangming Pan, Wenhu Chen, and William Yang Wang. 2024b. Understanding reasoning ability of language models from the perspective of reasoning paths aggregation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Xinyi Wang, Shawn Tan, Shenbo Xu, Mingyu Jin, William Yang Wang, Rameswar Panda, and Yikang Shen. 2025. Do Larger Language Models Generalize Better? A Scaling Law for Implicit Reasoning at Pretraining Time.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Zhuoyan Xu, Zhenmei Shi, and Yingyu Liang. 2024. Do large language models have compositional ability? an investigation into limitations and scalability. *CoRR*, abs/2407.15720.

Jiaran Ye, Zijun Yao, Zhidian Huang, Liangming Pan, Jinxin Liu, Yushi Bai, Amy Xin, Liu Weichuan, Xiaoyin Che, Lei Hou, and 1 others. 2026. How do transformers learn implicit reasoning? *Advances in Neural Information Processing Systems*, 38:65810–65838.

Ping Yu, Jing Xu, Jason E Weston, and Ilia Kulikov. 2024. Distilling system 2 into system 1. In *The First Workshop on System-2 Reasoning at Scale, NeurIPS'24*.

Hanlin Zhu, Baihe Huang, Shaolun Zhang, Michael Jordan, Jiantao Jiao, Yuandong Tian, and Stuart Russell. 2024. Towards a theoretical understanding of the 'reversal curse' via training dynamics. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

A Dataset Details

A.1 Population partition and graph structure

We use $N = 100$K individuals, following Allen-Zhu and Li (2024), who show that training on larger populations does not change their conclusion on knowledge storage, extraction, or manipulation. The $N = 100$K individuals are split into two disjoint populations of 50K each. $\mathcal{P}_{\text{comp}}$ contains individuals whose `friend` and `enemy` relations point exclusively to other members of $\mathcal{P}_{\text{comp}}$, forming a closed relational component over which compositional chains can be constructed. $\mathcal{P}_{\text{held}}$ contains individuals whose relations point exclusively within $\mathcal{P}_{\text{held}}$. This strict containment guarantees that no member of $\mathcal{P}_{\text{held}}$ ever appears as a head, bridge, or target entity in any compositional sequence seen during pretraining. $\mathcal{P}_{\text{comp}}$ is further split for finetuning: 75% form $\mathcal{P}_{\text{comp}}^{\text{train}}$ (2-hop questions seen during QA finetuning) and the remaining 25% form $\mathcal{P}_{\text{comp}}^{\text{test}}$ (2-hop questions held out).

A.2 Attributes and relations

Each individual is described by six atomic attributes following Allen-Zhu and Li (2024): `birthday`, `birthcity`, `university`, `major`, `company`, and `workcity`. We add two directional inter-individual relations, `friend` and `enemy`, each pointing from an individual to exactly one other individual in the same population. Relations are directional: X being the friend of Y does not imply Y being the friend of X.

A.3 Augmentation format examples

Table 5 illustrates each augmentation format on a single running example. Atomic 1-hop biographies (shared across all conditions) use the `multi5p-permute` format of Allen-Zhu and Li (2024). RDF sequences encode the same facts

as structured triples with dedicated special tokens [ENTITY], [RELATION], and [VALUE]. Explicit formats name the bridge entity, implicit formats omit it.

Format	Example sequence
1-hop NL (base)	*Marcus Halloway was born on June 14, 1967. He studied Linguistics at Northgate University.*
1-hop RDF	[ENTITY] Marcus Halloway [RELATION] birthday [VALUE] June 14, 1967
NL 2-hop implicit	*Marcus Halloway's friend was born in Ashford.*
NL 2-hop explicit	*Marcus Halloway's friend Delia Crane was born in Ashford.*
RDF 2-hop implicit	[ENTITY] Marcus Halloway [RELATION] friend.birthcity [VALUE] Ashford
RDF 2-hop explicit	[ENTITY] Marcus Halloway [RELATION] friend [VALUE] Delia Crane [RELATION] birthcity [VALUE] Ashford

Table 5: **Augmentation format examples.** Each format expresses the same underlying 2-hop fact (the birth city of Marcus Halloway's friend). Implicit formats omit the bridge entity (Delia Crane) and explicit formats include it.

A.4 Question counts

Table 6 reports the number of QA pairs used for finetuning and evaluation, per population and hop count. Questions for $\mathcal{P}_{\text{comp}}^{\text{train}}$ are seen during finetuning, questions for $\mathcal{P}_{\text{comp}}^{\text{test}}$ and $\mathcal{P}_{\text{held}}$ are used only at evaluation.

	$\mathcal{P}_{\text{comp}}^{\text{train}}$	$\mathcal{P}_{\text{comp}}^{\text{test}}$	$\mathcal{P}_{\text{held}}$
1-hop	300K	100K	400K
2-hop	600K	200K	800K
3-hop	1.2M	400K	1.6M

Table 6: **QA pair counts** per population and hop count. Counts scale with hop depth as the number of relational paths grows.

B Training and Optimization Details

B.1 Architecture and pretraining configuration

We pretrain a GPT-2 small architecture (12 layers, 12 heads, 768 hidden dimension, 124M parameters) from scratch using rotary positional embeddings (Su et al., 2024). Optimization uses AdamW with peak learning rate 10^{-3}, minimum learning rate 10^{-4}, 1000-step warmup, cosine decay, and gradient clipping at 1.0. Training uses 512-token context windows and a batch size of 49 152 tokens, for 800 000 steps. We train our models on 4-H100 GPU, pretraining take around 8 hours and finetuning 2 hours. The vocabulary size $|\mathcal{V}|$ for experiments without RDF is $50,264$, and for experiments with RDF is 50304 (a number divisible by a power of 2 to speed up training).

B.2 Finetuning configuration

We use LoRA (Hu et al., 2022) applied to query, value, and embedding layers with AdamW ($lr = 3 \times 10^{-4}$, weight decay 0.01). We sweep LoRA ranks $r_{\text{qv}} \in \{8, 16, 32\}$ and $r_{\text{emb}} \in \{32, 64, 128\}$ and report in Table 2 the best result per condition.

B.3 Training statistics

In Table 7, we detail the number of unique tokens for each experiment and compare our training budget against Chinchilla scaling laws (Hoffmann et al., 2022). Following the methodology described in (Allen-Zhu and Li, 2024), we perform multiple passes over the dataset until the model converges, ensuring complete memorization of the training distribution.

# Exp.	Unique Tokens	Total Tokens	Chinchilla Ratio
Bas. GPT-S	60 M.	39.3 B	$\times$ 15.9
Bas. GPT-M	60 M.	39.3 B	$\times$ 5.23
Bas. GPT-L	60 M.	39.3 B	$\times$ 3.54
Exp. 1	111.4 M.	39.3 B	$\times$ 15.9
Exp. 2	129.6 M.	39.3 B	$\times$ 15.9
Exp. 3	142.1 M.	39.3 B	$\times$ 15.9
Exp. 4	127.8 M.	39.3 B	$\times$ 15.9
Exp. 5	121.4 M.	39.3 B	$\times$ 15.9
Exp. 6	249.2 M.	39.3 B	$\times$ 15.9
Exp. 7	271.0 M.	39.3 B	$\times$ 15.9
Exp. 8	385.1 M.	39.3 B	$\times$ 15.9
Exp. 9	396.4 M.	39.3 B	$\times$ 15.9

Table 7: **Token statistics.** Comparison of unique token counts, total tokens seen during training, and the compute ratio relative to Chinchilla-optimal scaling laws.

B.4 Full Fine-tuning and Catastrophic Forgetting

In the experiments presented in Table 3, we observe that the model struggles to converge, even when using the following LoRA parameters: $r_{\text{qv}} \in \{8, 16, 32\}$ and $r_{\text{emb}} \in \{32, 64, 128\}$. Furthermore, we demonstrate that full fine-tuning—by forcing the model to converge on 2-hop relations, leads to catastrophic forgetting on the $\mathcal{P}_{\text{held}}$ population.

B.5 Data augmentation ratios

Each batch mixes atomic biographies from all individuals with compositional augmentation sequences drawn exclusively from $\mathcal{P}_{\text{comp}}$. We fix the atomic/compositional ratio at 30/70, placing

Fine-tuning	Pop.	1-hop	2-hop	3-hop
1+2-hop	$\mathcal{P}_{\text{comp}}$	1.00	0.99	0.01
	$\mathcal{P}_{\text{held}}$	0.01	0.01	0.01

Table 8: **Full fine-tuning.** Full fine-tuning on 1 and 2-hop questions leads to catastrophic forgetting on 1-hop generalization.

# Exp	$\mathcal{P}^{\text{train}}_{\text{comp}}$	$\mathcal{P}^{\text{test}}_{\text{comp}}$	$\mathcal{P}^{\text{held}}_{\text{comp}}$
Exp 1	0.10	0.00	0.00
Exp 2	0.77	0.49	0.00
Exp 3	0.00	0.00	0.00
Exp 4	0.92	0.51	0.00
Exp 5	0.00	0.00	0.00
Exp 6	0.81	0.50	0.00
Exp 7	0.79	0.44	0.00
Exp 8	0.81	0.60	0.00
Exp 9	0.93	0.51	0.00

Table 9: **Conditional analysis.** Success rate of 2-hop composition given that both constituent 1-hop sub-questions are answered correctly.

our experiments in the high-supervision regime shown to be necessary for implicit composition to emerge (Ye et al., 2026; Wang et al., 2024a). This is a deliberately favorable setting for composition: the absence of transfer to $\mathcal{P}_{\text{held}}$ under this regime cannot be attributed to insufficient augmentation. Multi-format conditions divide the compositional portion equally across constituent formats. Full mixing ratios are in Table 10.

# Exp.	Setting	Ratio
Exp. 1	1-hop RDF	50–50
Exp. 2	NL 2-hop (imp.)	30–70
Exp. 3	NL 2-hop (exp.)	30–70
Exp. 4	RDF 2-hop (imp.)	30–70
Exp. 5	RDF 2-hop (exp.)	30–70
Exp. 6	RDF 2-hop (exp+imp)	30–35–35
Exp. 7	NL 2-hop (exp+imp)	30–35–35
Exp. 8	NL+RDF 2-hop (exp+imp)	30–17.5×4
Exp. 9	Full (NL+RDF+1-hop RDF)	15–15–17.5×4

Table 10: **Pretraining mixtures.** Ratios indicate the proportion of atomic 1-hop biographical text and compositional augmentation sequences in each training batch. The baseline (Exp. 0) uses only `multi5p-permute` biographical text and therefore has no mixing ratio.

C Additional Experimental Results

C.1 Full main results

Table 12 reports 1-hop, 2-hop, and 3-hop accuracy across all 9 augmentation conditions and the baseline, for the three evaluation populations including $\mathcal{P}^{\text{train}}_{\text{comp}}$, which is omitted from the main text. Accuracy on $\mathcal{P}^{\text{train}}_{\text{comp}}$ reflects performance on 2-hop questions seen during finetuning. The gap between $\mathcal{P}^{\text{train}}_{\text{comp}}$ and $\mathcal{P}^{\text{test}}_{\text{comp}}$ isolates QA memorization from genuine compositional transfer. 2-hop accuracy on $\mathcal{P}_{\text{held}}$ remains at chance across every condition.

Attribute	Base	Exp. 8
`birthday`	0.92	0.88
`birthcity`	0.97	0.94
`major`	0.96	0.94
`employer`	0.96	0.95
`friend`	0.86	0.60
`enemy`	0.87	0.60

Table 11: **1-hop accuracy by attribute on $\mathcal{P}^{\text{test}}_{\text{comp}}$.** Scalar attributes are retrieved near-perfectly under both conditions. Relational targets (`friend`, `enemy`) are retrieved far less reliably, and degrade markedly under Exp. 8.

C.2 LoRA rank sweep

Table 13 reports the full LoRA rank sweep on $\mathcal{P}^{\text{train}}_{\text{comp}}$. Each cell gives 1-hop / 2-hop first-token accuracy for a given $(r_{\text{qv}}, r_{\text{emb}})$ pair. Increasing LoRA rank yields only marginal variation in 2-hop accuracy within each condition, confirming that the compositional outcome is determined by the pretraining augmentation rather than by finetuning capacity. Main-text results use the best $(r_{\text{qv}}, r_{\text{emb}})$ pair per condition.

C.3 Per-attribute results

Scalar attributes are retrieved near-perfectly under both conditions. Relational targets behave differently: `friend` and `enemy` drop from ∼0.86 at baseline to 0.60 under Exp. 8. The two cases are not symmetric. A scalar attribute is drawn from a small, low-cardinality set, whereas a relational target is one specific individual among 100K, identified by a name. Retrieving it requires discriminating that name from a very large space of similar tokens, where first names recur across many individuals and offer little disambiguating signal. Under Exp. 8, the compositional augmentation dilutes the atomic biographical text in each batch (Section B.5), reducing exposure to exactly the per-individual name associations that relational retrieval depends on. High-cardinality name retrieval is therefore the first capability to degrade when atomic supervision is diluted, while low-cardinality scalar attributes remain robust.

#	$\mathcal{P}_{\text{comp}}^{\text{train}}$			$\mathcal{P}_{\text{comp}}^{\text{test}}$			$\mathcal{P}_{\text{held}}$		
	1-hop	2-hop	3-hop	1-hop	2-hop	3-hop	1-hop	2-hop	3-hop
Exp. 0	1.00	0.08	0.08	0.97	0.08	0.08	0.97	0.01	0.01
Exp. 1	1.00	0.08	0.08	0.97	0.08	0.08	0.97	0.01	0.01
Exp. 2	1.00	0.75	0.05	0.88	0.62	0.05	0.75	0.01	0.01
Exp. 3	1.00	0.08	0.08	0.97	0.08	0.08	0.89	0.01	0.01
Exp. 4	1.00	0.98	0.05	0.97	0.79	0.05	0.40	0.01	0.01
Exp. 5	1.00	0.37	0.37	0.98	0.08	0.08	0.38	0.02	0.02
Exp. 6	1.00	0.92	0.04	0.98	0.79	0.15	0.50	0.01	0.01
Exp. 7	1.00	0.78	0.05	0.91	0.73	0.06	0.79	0.01	0.01
Exp. 8	1.00	0.90	0.14	0.99	0.83	0.04	0.83	0.01	0.01
Exp. 9	1.00	0.98	0.04	0.99	0.79	0.14	0.80	0.01	0.01

Table 12: **Full results across augmentation conditions.** First-token accuracy for 1-hop, 2-hop, and 3-hop queries on $\mathcal{P}_{\text{comp}}^{\text{train}}$ (2-hop questions seen at finetuning), $\mathcal{P}_{\text{comp}}^{\text{test}}$ (2-hop questions held out from finetuning), and $\mathcal{P}_{\text{held}}$ (never compositionally exposed during pretraining). Row 0 is the baseline. Cell color encodes accuracy from low (purple) to high (green). The high accuracy on $\mathcal{P}_{\text{comp}}^{\text{train}}$ together with the accuracy retained on $\mathcal{P}_{\text{comp}}^{\text{test}}$ for the strongest conditions indicates transfer across unseen questions. The persistent chance-level accuracy on $\mathcal{P}_{\text{held}}$ indicates no transfer across populations.

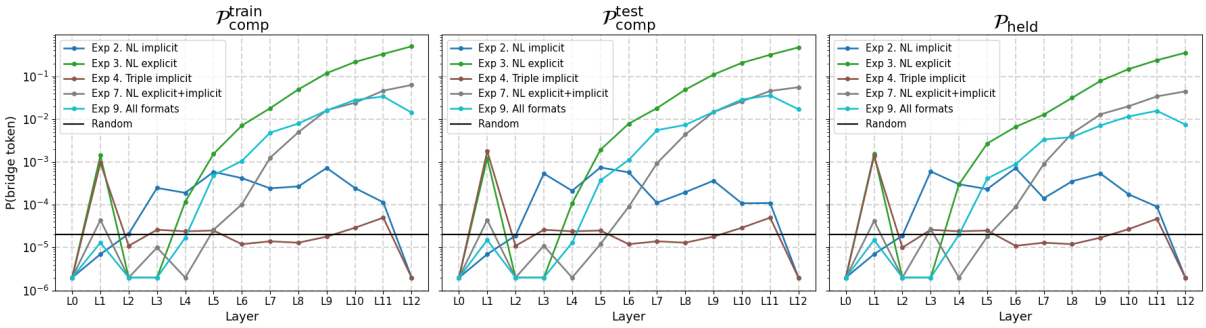

Figure 2: **Layer-wise probability of the bridge-entity token (logit lens).** Mean $P(\text{bridge token})$ at the query position across layers, for five conditions, on $\mathcal{P}_{\text{comp}}^{\text{train}}$, $\mathcal{P}_{\text{comp}}^{\text{test}}$, and $\mathcal{P}_{\text{held}}$. Explicit conditions (Exp. 3, and the explicit-containing Exp. 7, 9) drive the bridge *token* far above the random baseline. Implicit conditions (Exp. 2, 4) stay near it. Crucially, the curves are near-identical across the three populations: surface emission of the bridge token tracks the augmentation format, not whether the individual was compositionally exposed.

r_{qv} / r_{emb}	Exp. 1			Exp. 2			Exp. 3		
	32	64	128	32	64	128	32	64	128
8	1.0/0.09	1.0/0.09	1.0/0.09	0.99/**0.30**	0.99/**0.30**	0.99/0.31	1.0/**0.09**	1.0/**0.09**	1.0/**0.09**
16	1.0/0.09	1.0/0.09	1.0/0.09	1.0/0.32	1.0/0.33	**1.0/0.33**	1.0/**0.09**	1.0/**0.09**	1.0/**0.09**
32	1.0/0.09	1.0/0.09	1.0/0.09	1.0/0.36	1.0/0.36	**1.0/0.37**	1.0/**0.09**	1.0/**0.09**	1.0/**0.09**
	Exp. 4			Exp. 5			Exp. 6		
8	0.99/0.91	0.99/**0.92**	1.0/**0.92**	1.0/0.09	1.0/0.09	1.0/0.09	1.0/**0.87**	1.0/0.88	1.0/0.89
16	1.0/0.93	1.0/0.94	**1.0/0.94**	1.0/0.09	1.0/0.09	1.0/0.09	1.0/**0.95**	1.0/0.91	1.0/0.92
32	1.0/0.96	1.0/0.96	**1.0/0.96**	1.0/0.09	1.0/0.09	1.0/0.09	1.0/0.94	1.0/**0.95**	1.0/0.95
	Exp. 7			Exp. 8			Exp. 9		
8	1.0/0.26	1.0/**0.27**	1.0/**0.27**	1.0/0.90	1.0/**0.97**	1.0/**0.97**	1.0/0.92	1.0/0.92	1.0/**0.93**
16	1.0/0.28	1.0/0.29	**1.0/0.29**	1.0/0.93	1.0/**0.94**	1.0/**0.94**	1.0/0.94	1.0/0.95	**1.0/0.95**
32	1.0/0.33	1.0/0.33	**1.0/0.33**	1.0/0.96	1.0/**0.97**	1.0/**0.97**	1.0/0.97	1.0/**0.97**	**1.0/0.97**

Table 13: **LoRA rank sweep on $\mathcal{P}_{\text{comp}}^{\text{train}}$.** Each cell reports 1-hop / 2-hop first-token accuracy for a $(r_{\text{qv}}, r_{\text{emb}})$ pair. Best 2-hop configuration per condition in bold. Variation across ranks is small, indicating that finetuning capacity is not the limiting factor.

C.4 Conditional Analysis

In Table 9, we address the following question: *"Among instances where both 1-hop sub-questions are answered correctly, what fraction results in a correct 2-hop prediction?"*

Given that our entity bridge is always an individual within a set of 100,000, the task requires precise retrieval. While the overall accuracy on $\mathcal{P}_{\text{held}}$ reaches 97% for the baselines (Table 3) and 83% for Exp. 8, we observe a catastrophic performance drop in compositional reasoning. For instance, consider the query: *"What is the birthday of X's friend?"* Even when the model correctly identifies the friend (Y) and retrieves Y's birthday, it alwats fails to compose these facts for individuals in $\mathcal{P}_{\text{held}}$. As shown in Table 9, despite the guarantee of perfect 1-hop awnsers, the composition success rate remains remarkably low across most experimental settings.

C.5 Intermediate Entity Localization with the Logit Lens

We apply the logit lens (nostalgebraist, 2020) after LoRA finetuning: given a 2-hop query such as *What is the birth date of John's friend? Answer:*, we project each hidden state at the position of the `friend` token through the final layer norm and LM head to track whether the bridge entity emerges across layers. We restrict to single-token bridge names ($\sim$66% of examples) and report mean $P(\text{bridge token})$ against a random baseline.

Figure 2 shows that explicit conditions strongly emit the bridge token, as expected under next-token prediction. Yet emission and composition are dissociated in both directions: Exp. 3 emits the bridge token strongly yet composes in only 8% of cases, while Exp. 4 never emits it yet composes in 79% on $\mathcal{P}^{\text{test}}_{\text{comp}}$. The logit lens measures a surface property, not the compositional operation, instantiating in natural language the observation of Ye et al. (2026) that decodability of an intermediate result does not imply its use. Bridge token probability rises monotonically toward the output rather than resolving at an intermediate layer, a generation trajectory rather than an intermediate variable (Biran et al., 2024). Finally, emission curves are near-identical across populations: what $\mathcal{P}_{\text{held}}$ lacks is not producing the bridge token but using it to retrieve the final attribute, an operation learned only under compositional pretraining exposure.

D Ethics/Transparency Statement

We used AI-assisted tools for language polishing and proofreading of the manuscript. The authors have reviewed, edited, and approved the final content.