Grokked Transformers are Implicit Reasoners: A Mechanistic Journey to the Edge of Generalization

Boshi Wang   Xiang Yue   Yu Su   Huan Sun

The Ohio State University

LLMs struggle at Implicit Reasoning w/ Parametric Memory

Implicit reasoning: reasoning without explicit verbalization of intermediate steps (e.g., Chain-of-Thought)

Parametric memory: facts & rules stored in weights

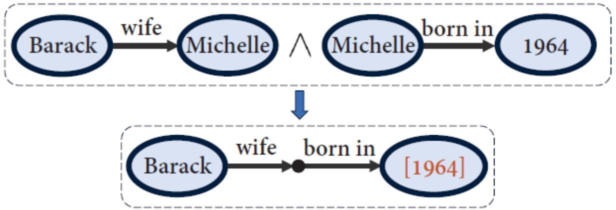
Press et al. & Yang et al.

- LLMs only show substantial evidence in resolving the first hop
- Scaling only improves the first hop; "compositionality gap" does not decrease

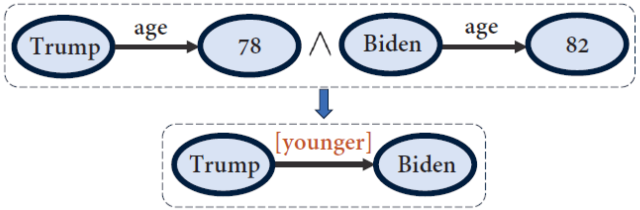
Zhu et al.

- GPT-4 cannot do implicit composition or comparison well

Composition



Comparison



Press et al. *Measuring and Narrowing the Compositionality Gap in Language Models.* Findings of EMNLP-23.
Yang et al. *Do Large Language Models Latently Perform Multi-Hop Reasoning?* ACL-24.
Zhu et al. *Physics of Language Models: Part 3.2, Knowledge Manipulation.* ICML-24 Tutorial.

Why Implicit Reasoning? (can't we just "CoT" everything?)

- The default mode of large-scale (pre-)training
- Fundamentally determines how well LLMs acquire structured representations of facts and rules from data
- Propagateble knowledge updates & systematic generalization (more later)

Why Parametric Memory? (can’t we do retrieval & long-context?)

- Unique power in compressing and integrating information at scale

- Important for tasks with large intrinsic complexity
  - E.g., reasoning problems with large search spaces (example later)

Research Questions

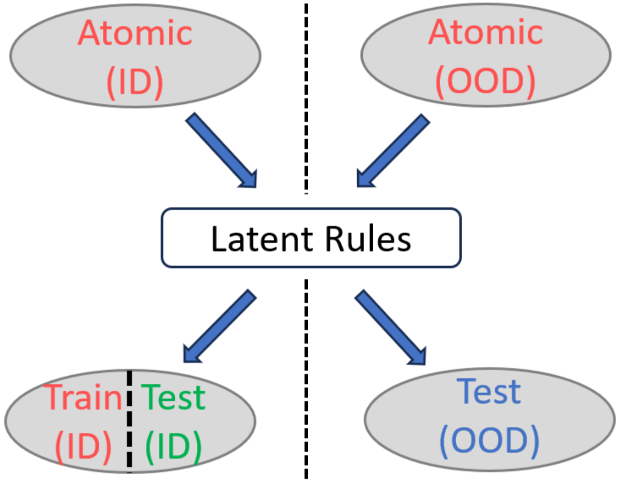
- Is implicit reasoning doomed given that even the most capable models struggle?
- Can it be resolved by further scaling data and compute, or are there fundamental limitations of Transformers that prohibit robust acquisition of this skill?

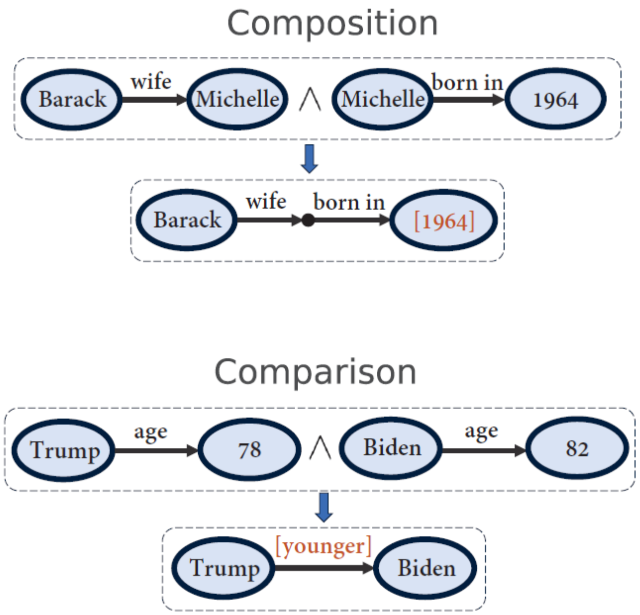
Approach: Synthetic Data & Training from Scratch

- Allows us to control the data and perform clean evaluations
- Important nowadays as pretraining/fine-tuning corpora keeps penetrating downstream evals

Reasoning as Rule Induction & Application

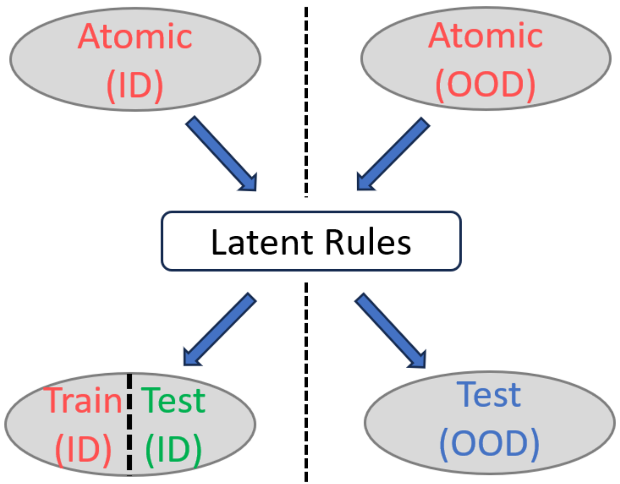
- Induce latent rules from a mixture of **atomic** facts and **inferred** facts (deduced via latent rules)
- Deduce novel facts by applying the acquired rules

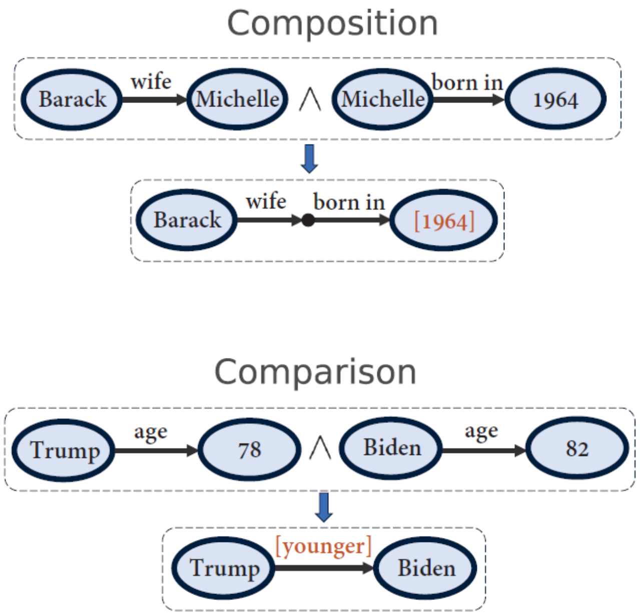




Reasoning as Rule Induction & Application

- ID: unseen inferred facts deduced from the **same** set of atomic facts underlying the observed inferred facts
- OOD (systematicity): unseen inferred facts from a **different** set of atomic facts (Lake et al.)





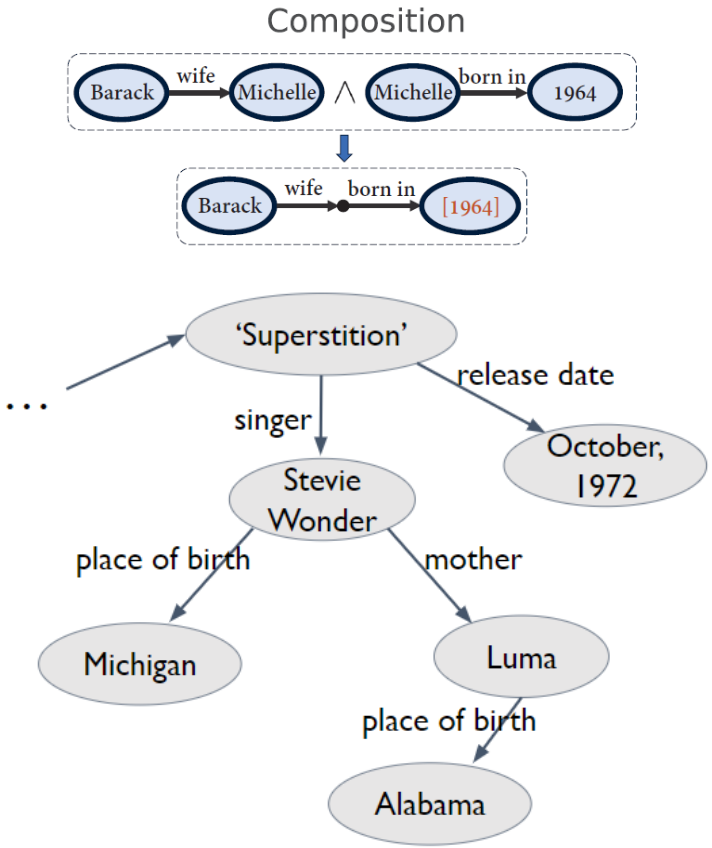
Lake et al. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. ICML-18.

Reasoning as Rule Induction & Application

Composition

- Atomic facts
  - Random KG consisting of |R| = 200 relations
  - Randomly split into ID & OOD atomic facts
- Inferred facts: two-hop compositions

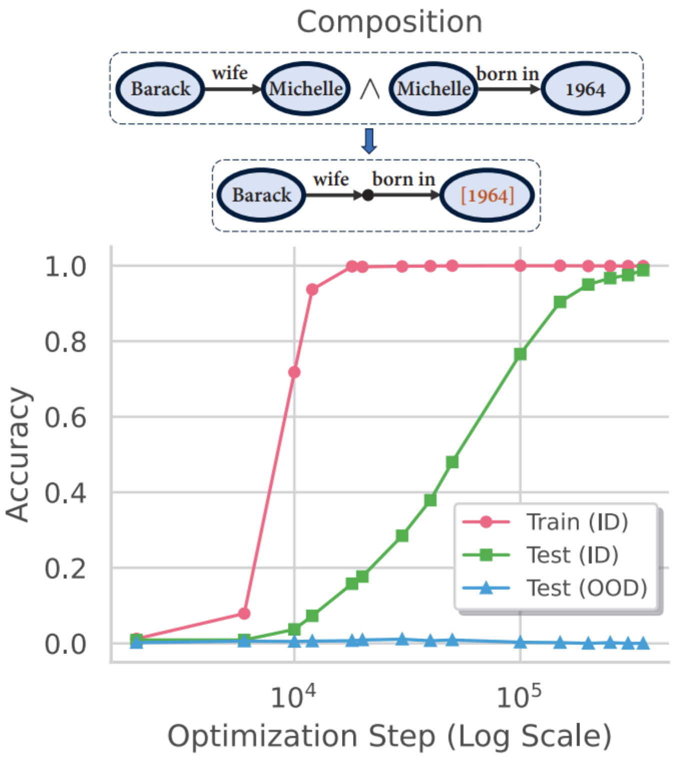
$$(h, r_1, b) \wedge (b, r_2, t) \implies (h, r_1, r_2, t)$$

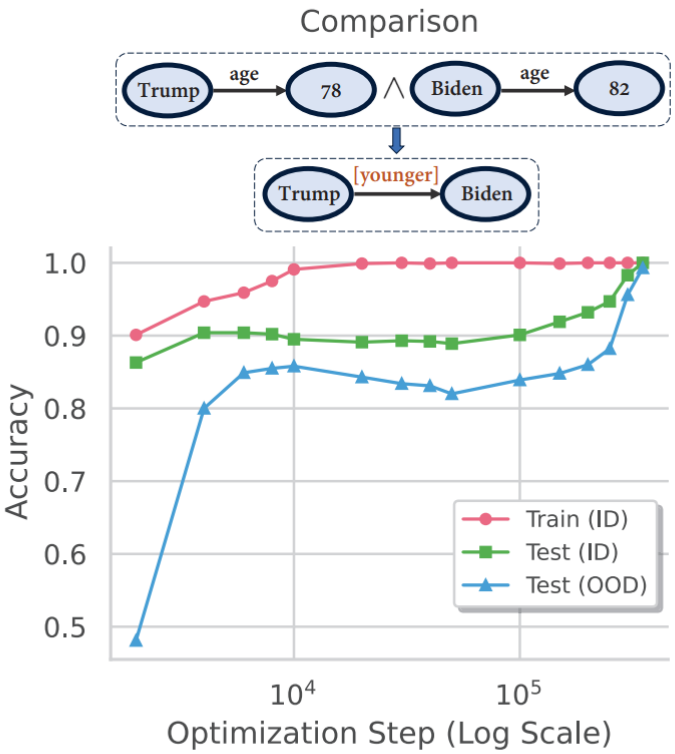


Model & Optimization

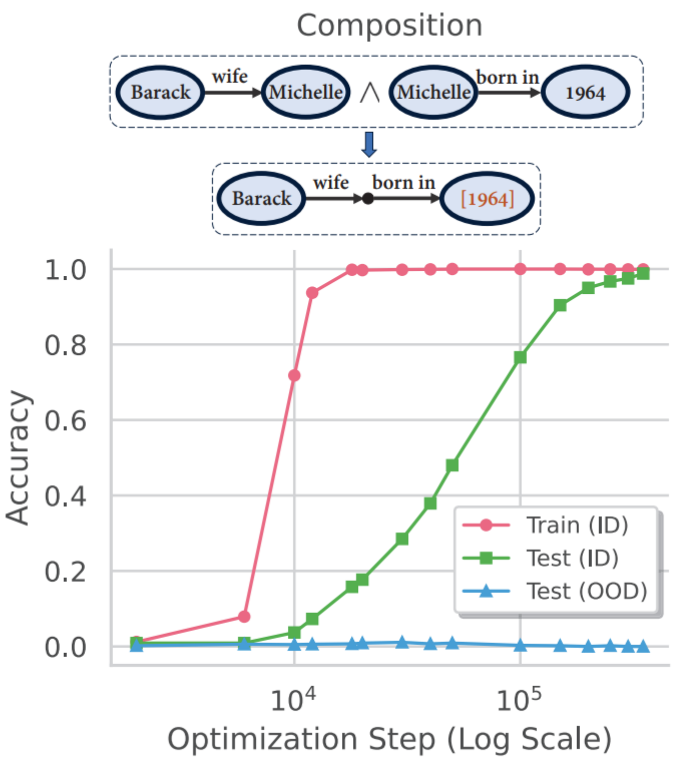
- Standard decoder-only transformer as in GPT-2
  - 8 layers, 768 hidden dimensions and 12 attention heads
- AdamW with learning rate 1e-4, batch size 512, weight decay 0.1 and 2000 warm-up steps
- “Concept-level” inputs: each entity/relation has its own learnable embedding
- More variants later

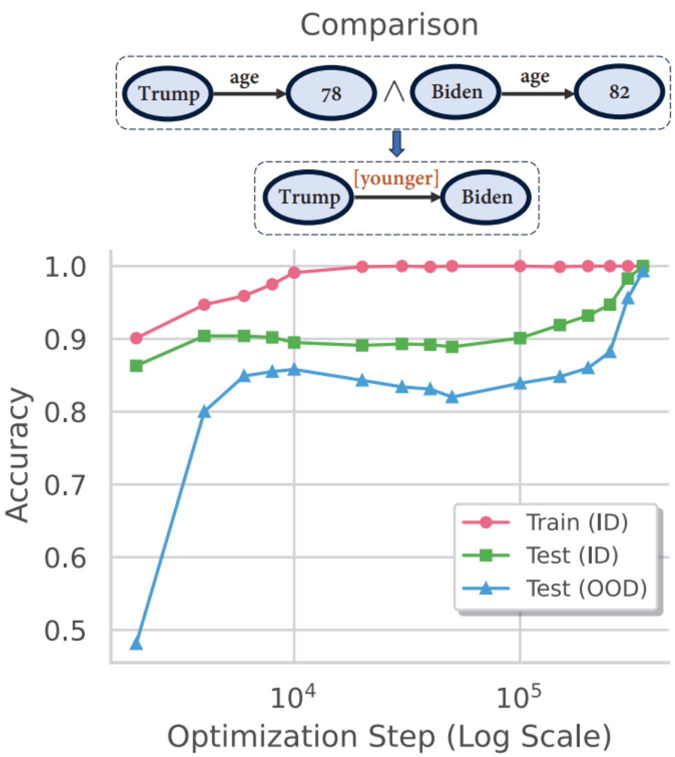
#1: “Grokking” in ID generalization



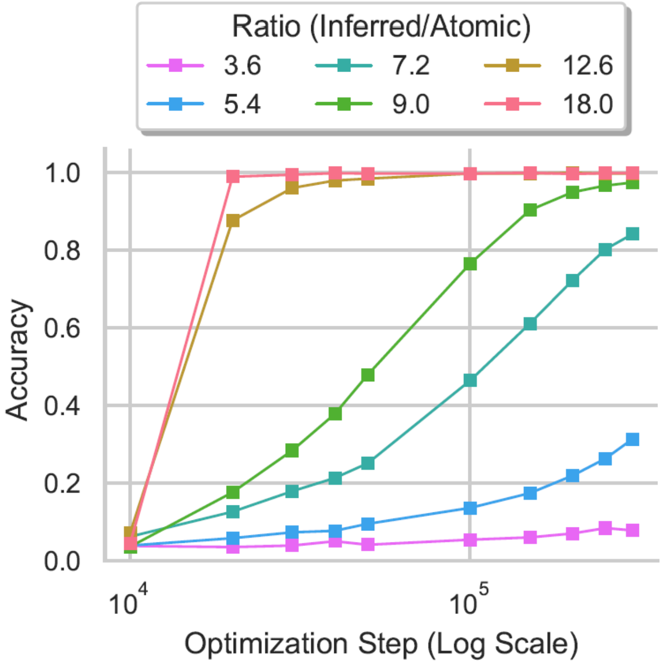


#2: Difference in OOD generalization

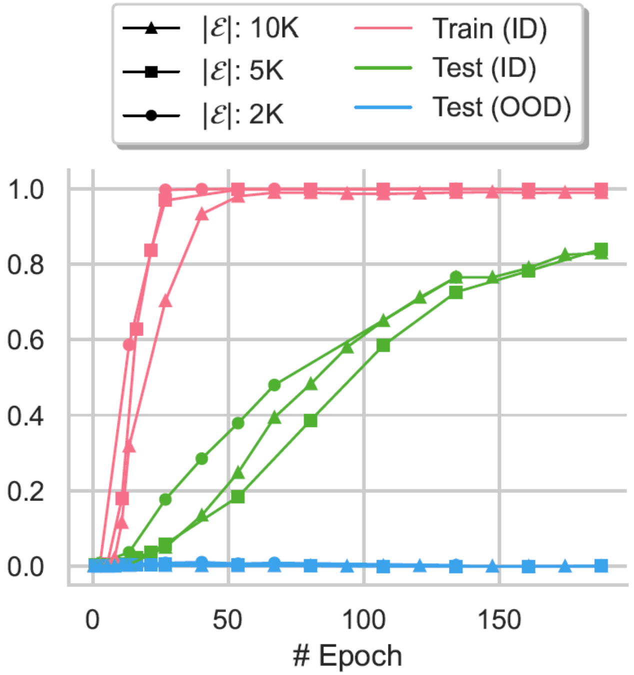




#3: Data **distribution**, not data size, drives generalization



(a) Effect of the inferred/atomic ratio ϕ.



(b) Effect of changing $|\mathcal{E}|$ ($\phi = 9.0$).

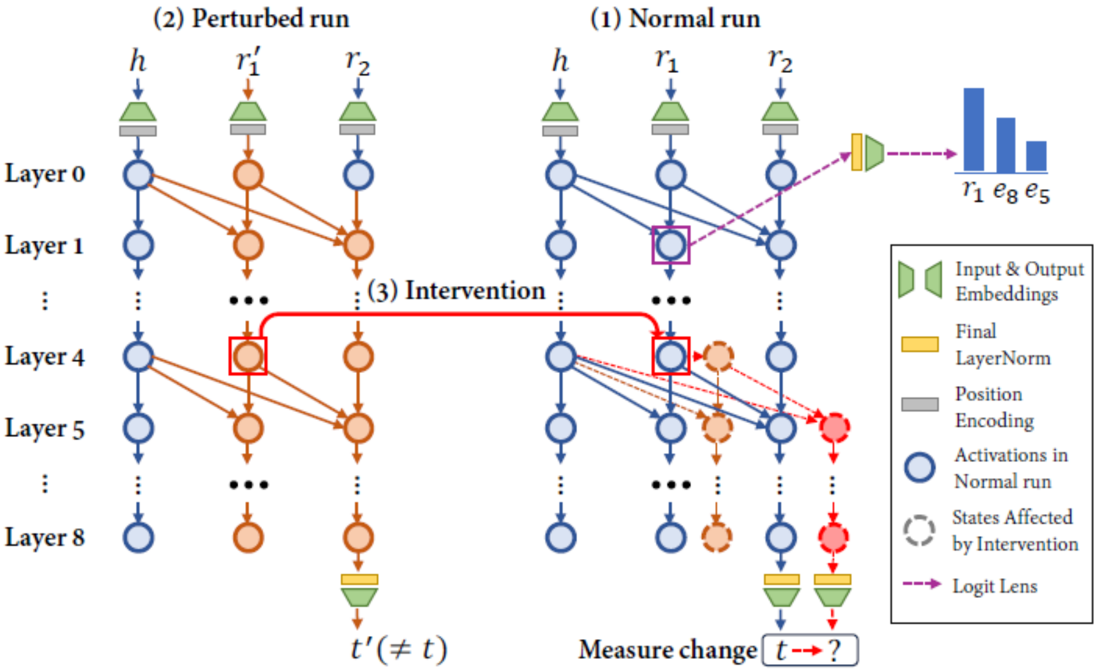
Important Questions Remain

- What happens during grokking?
- Why does grokking happen?
- Why no systematic generalization?

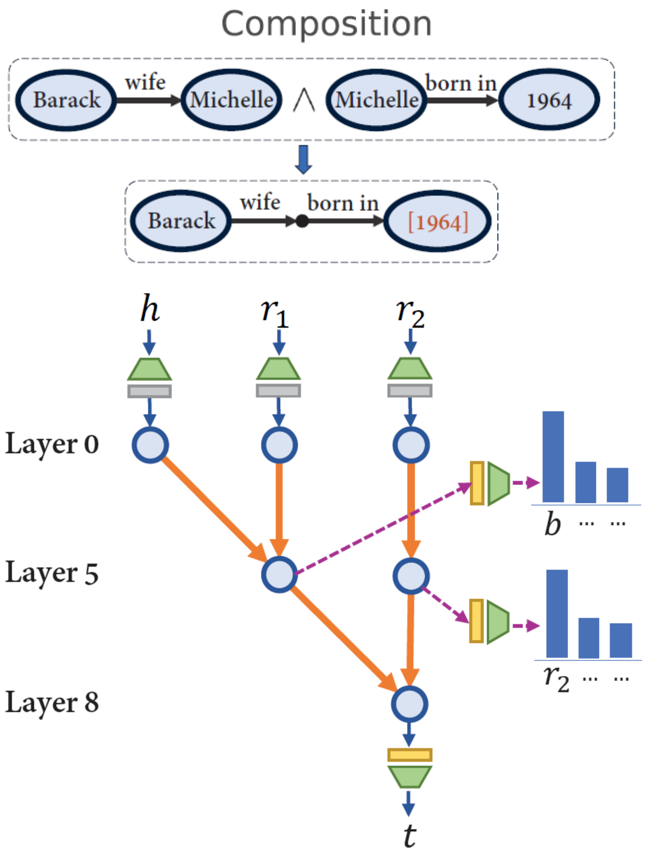
These require a deeper look inside of the model

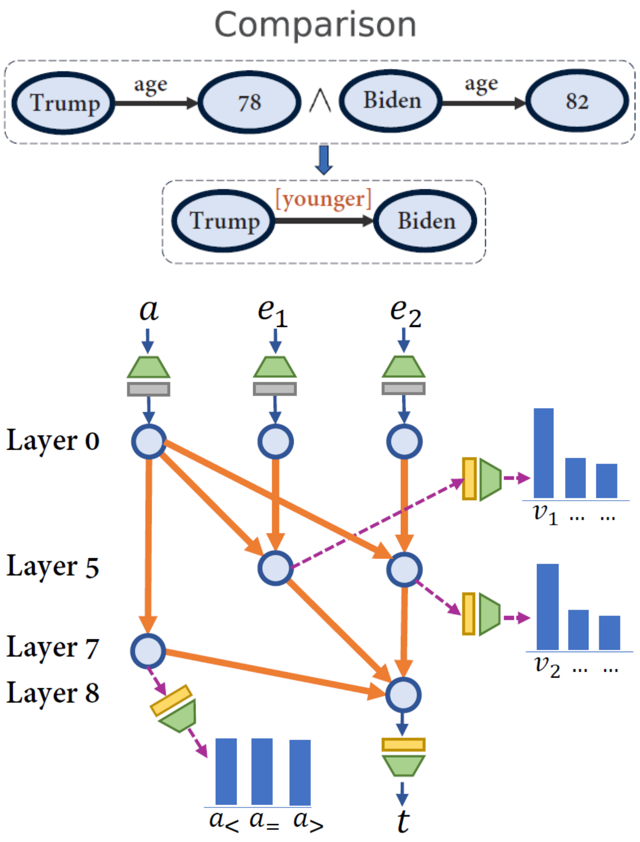
Analyzing the (change) in inner workings during grokking

- Logit lens
- Causal tracing

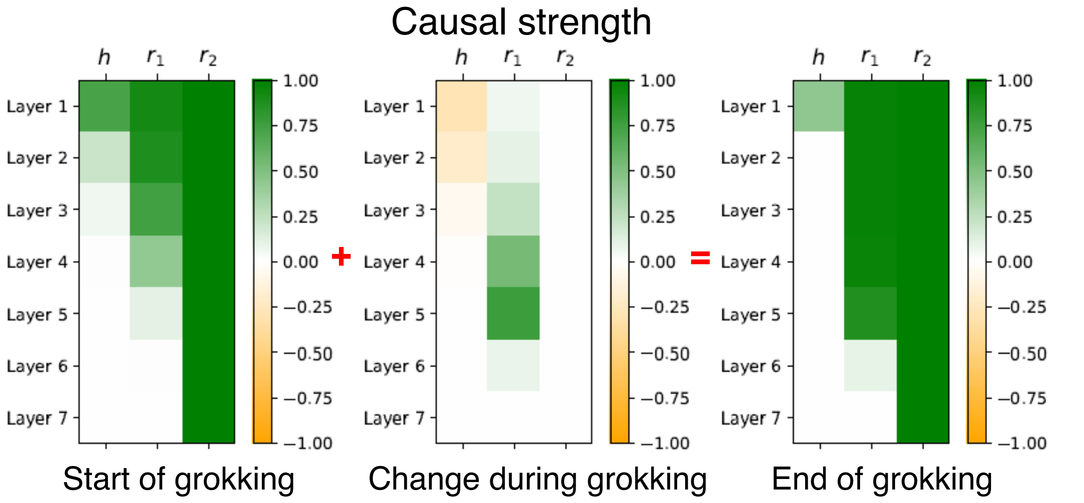


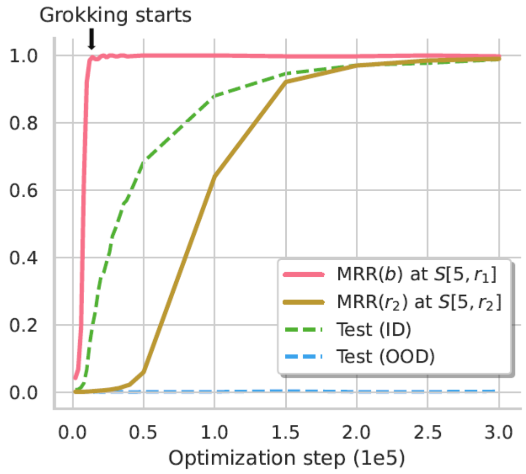
Generalizing Circuits (after Grokking)





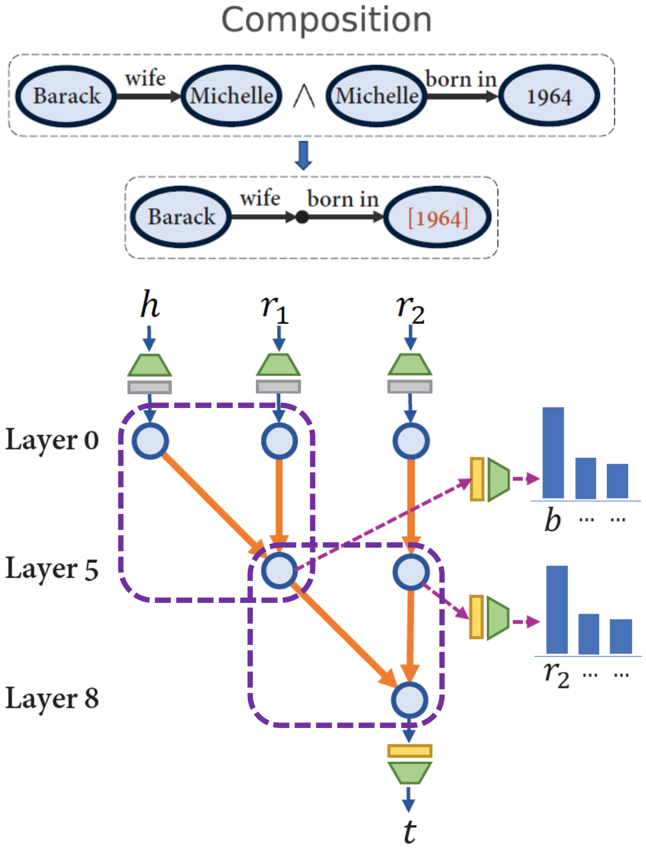
Changes during grokking

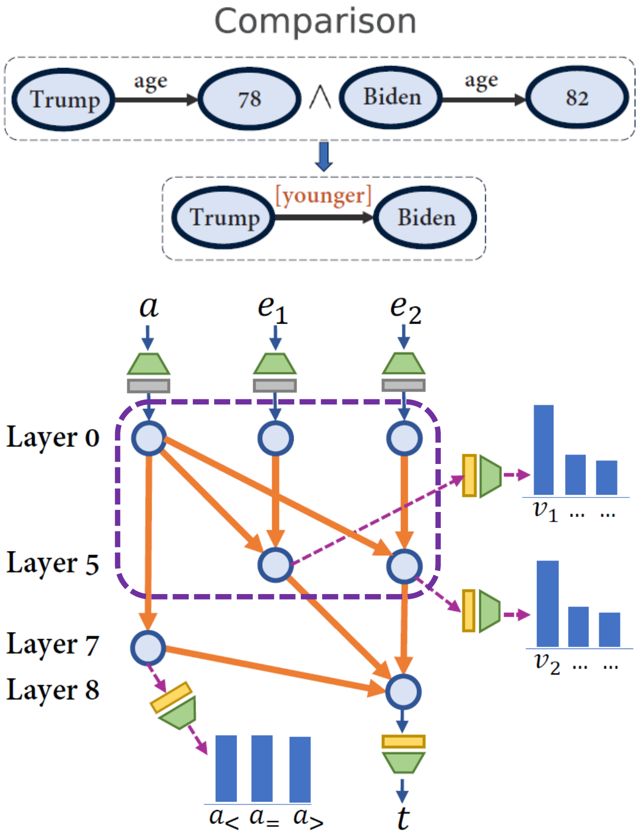




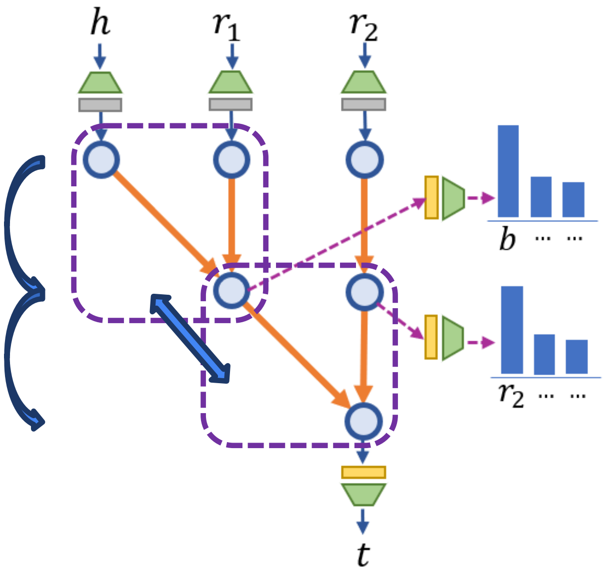
- Causal connection between S[5, r1] and the final prediction t grows significantly
- MRR(r2) gradually improves as S[5, r2] (via logit lens); S[5, r1] represents b throughout
- => Model gradually forms the second hop in the upper layers
- When grokking starts, very likely directly associates (h, r1, r2) with t, mostly memorization

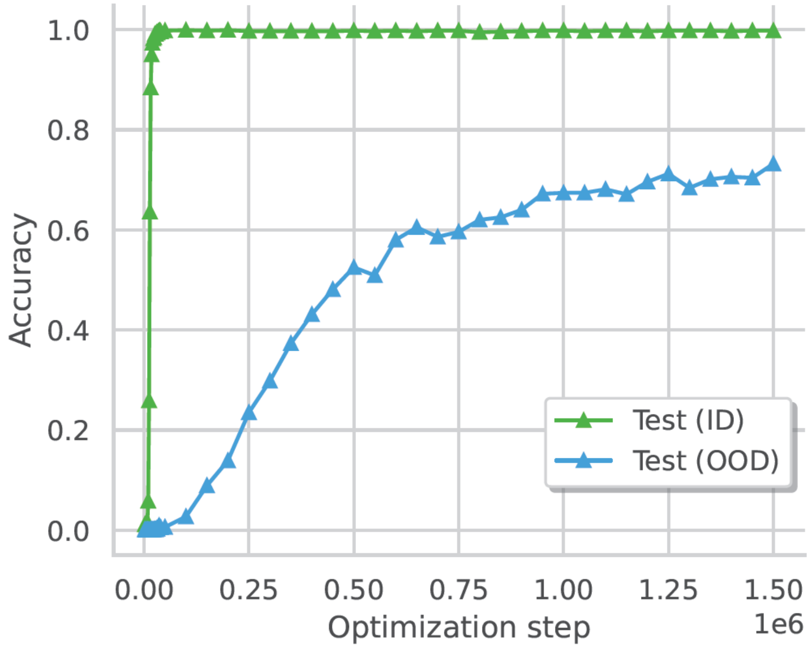
Understanding & Improving OOD generalization





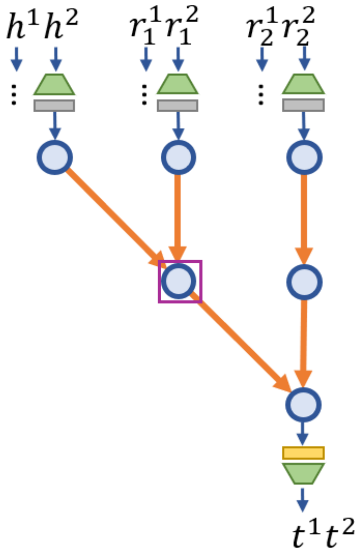
Understanding & Improving OOD generalization

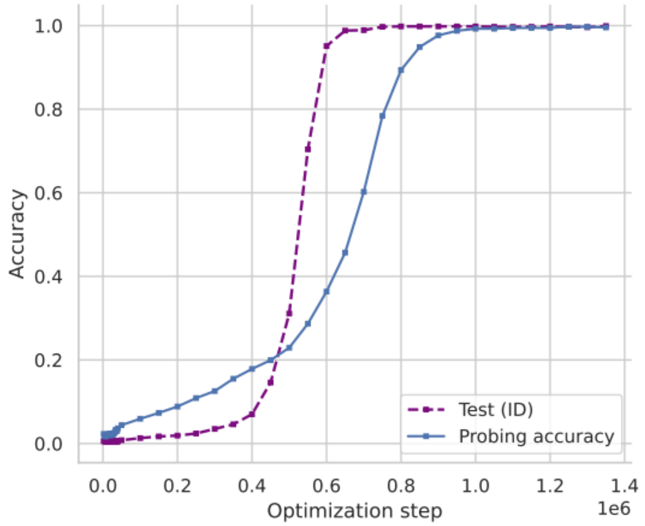




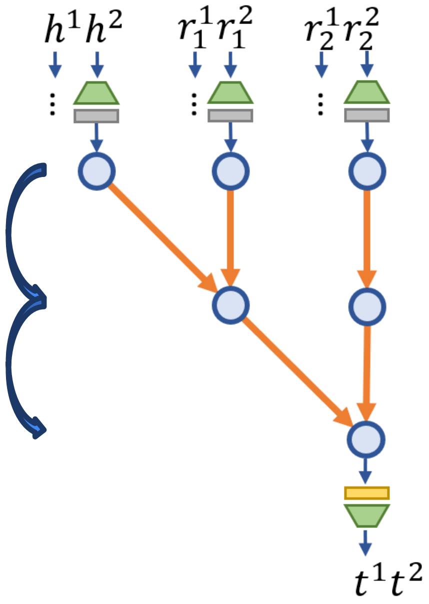
When inputs are at the **surface** level…

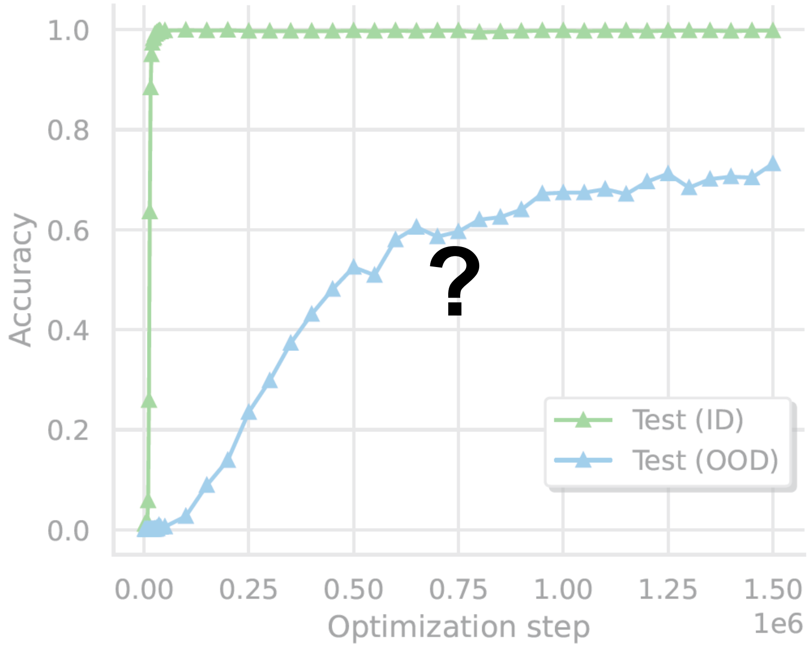
During grokking, the model seems to gradually stores the later surface-name tokens in the bridge hidden state



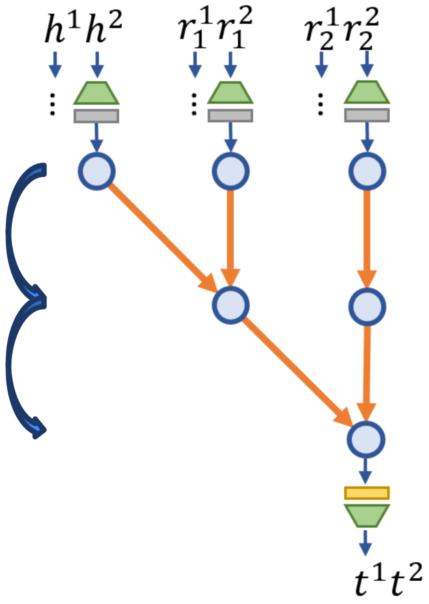


When inputs are at the **surface** level…



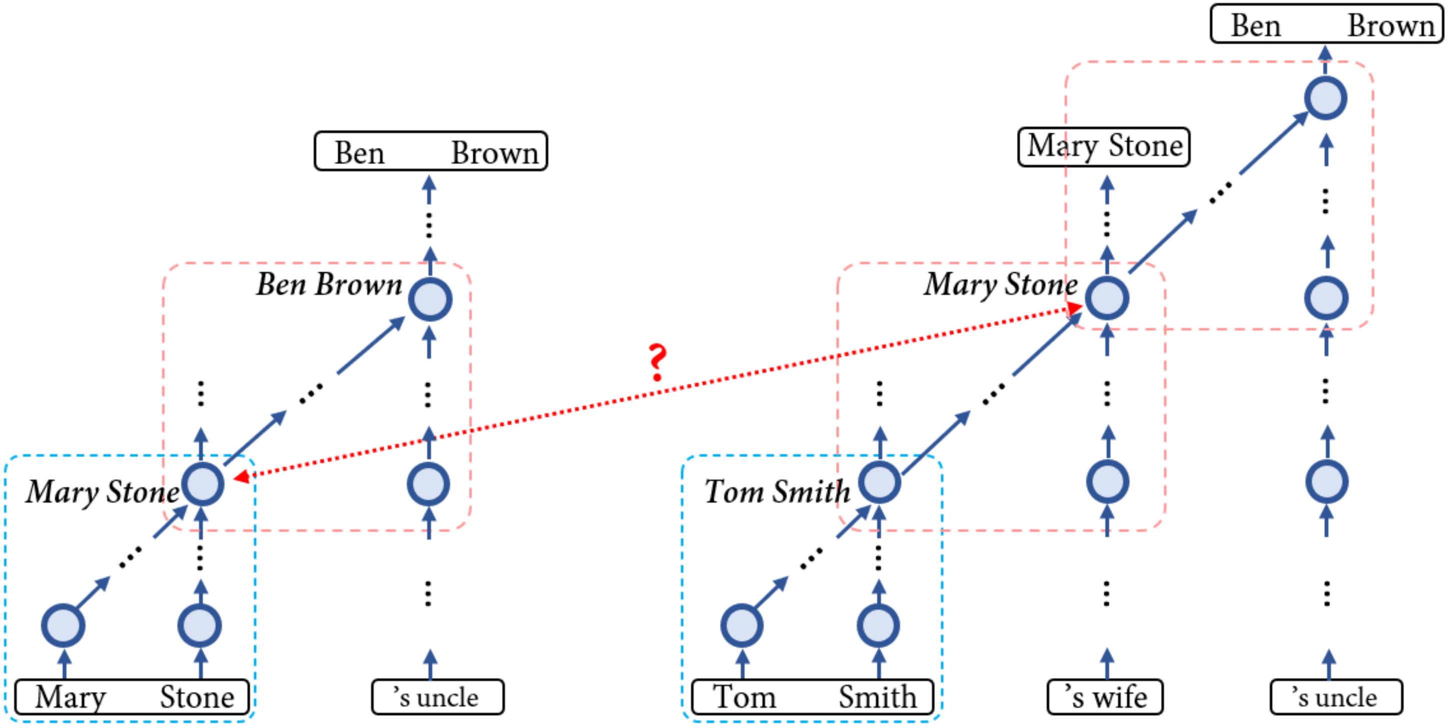


When inputs are at the **surface** level…

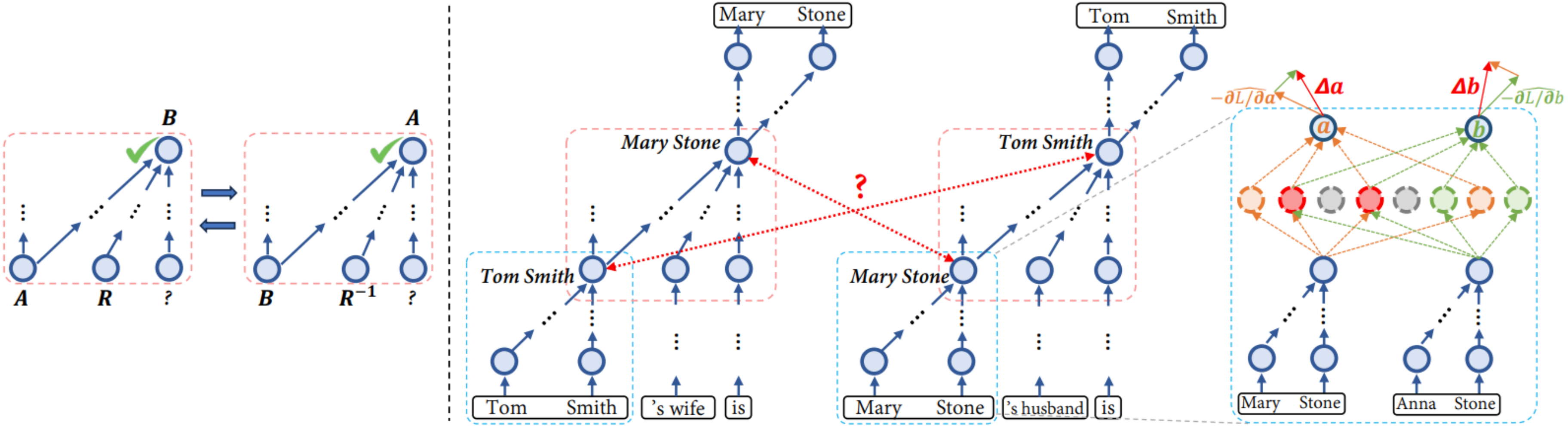


No OOD Generalization!

Surface-level Inputs & Binding



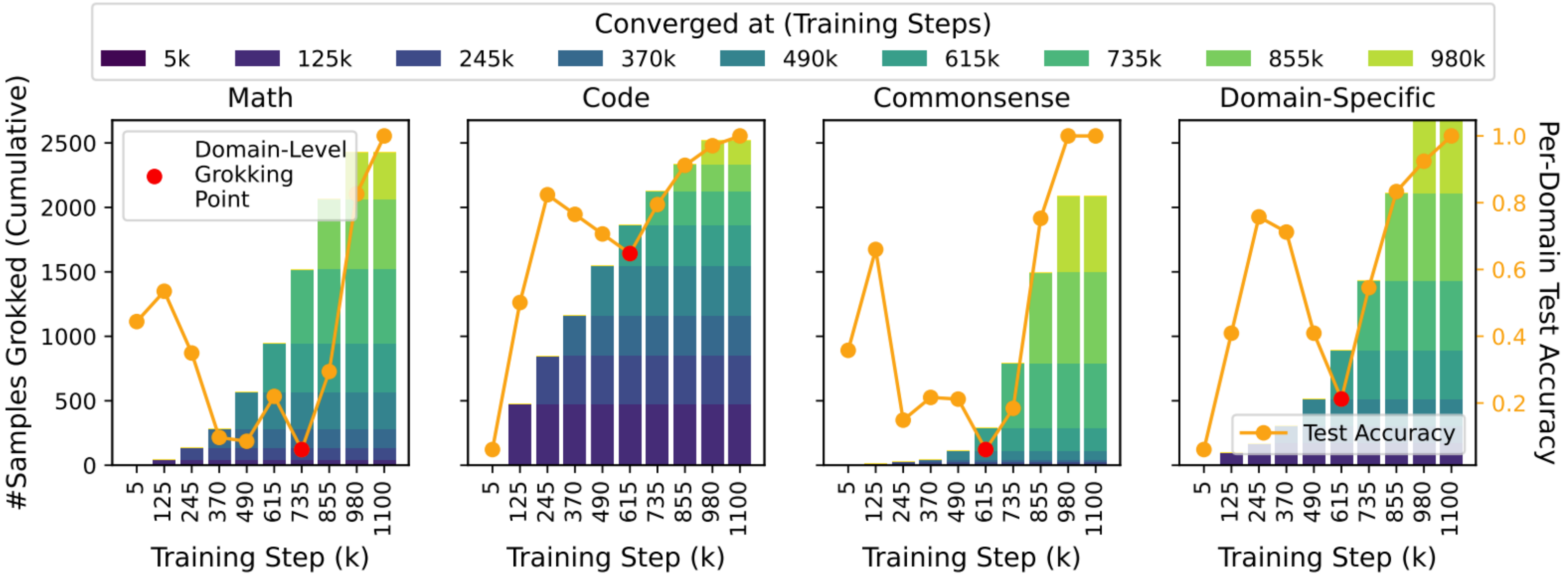
The binding problem & the “Reversal Curse”



- Inconsistent entity representations when switching roles between perceived subjects and predicted objects
- Representational entanglements cause interferences on learning dynamics and impede generalization

Wang et al. *Is the Reversal Curse a Binding Problem? Uncovering Limitations of Transformers from a Basic Generalization Failure*. arXiv-25.

Grokking in LLM Pretraining



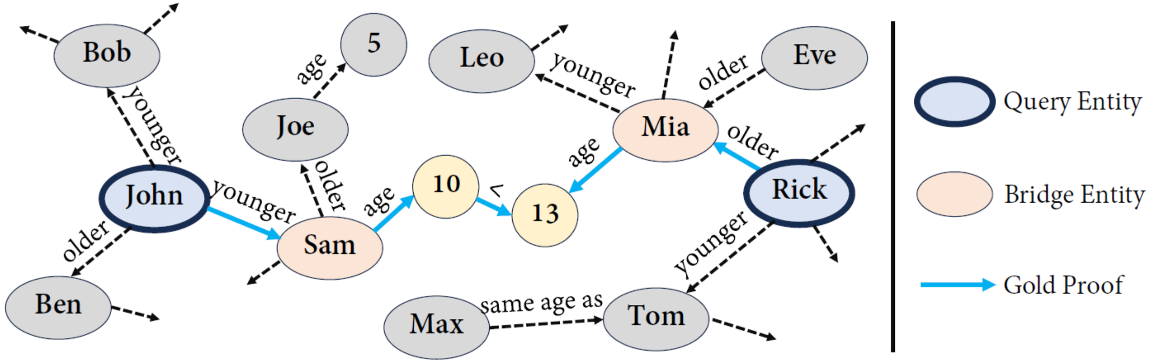
Li et al. *Where to find Grokking in LLM Pretraining? Monitor Memorization-to-Generalization without Test*. arXiv-25.

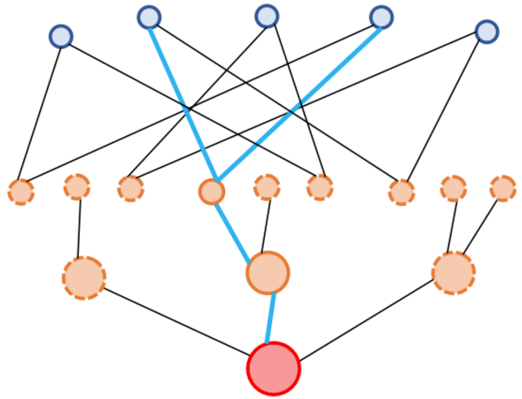
The Power of Parametric Memory for Complex Reasoning

What exactly are we going towards? Why parametric memory?

- Unique ability to compress and integrate information at scale for complex reasoning

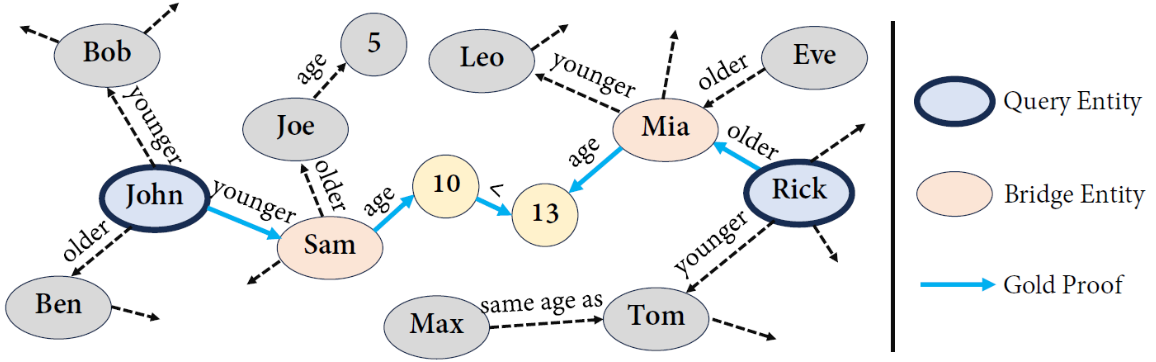
Challenging reasoning tasks with large search space





- Non-parametric memory: information stored in context
  - Explicit (verbalized) reasoning done in context
- Parametric memory: information stored in weights
  - Implicit reasoning done during information internalization

Challenging reasoning tasks with large search space



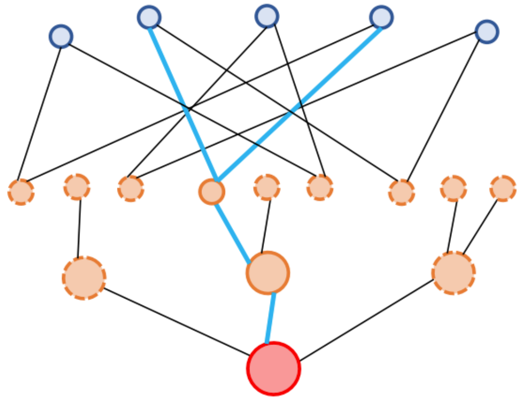
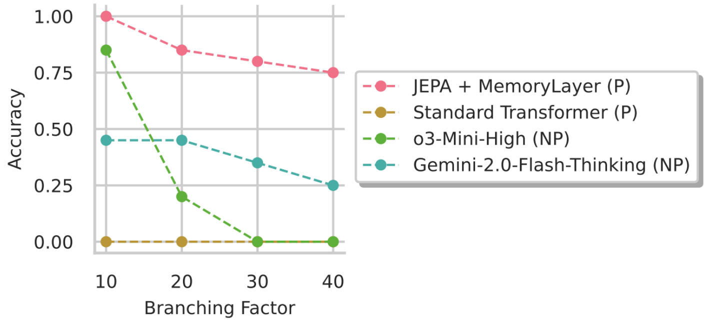


Table 1: Results on the complex reasoning task. Direct/CoT: predict the answer directly/verbalize the reasoning steps. "+R": retrieval augmentation.

	GPT-4-Turbo		Gemini-Pro-1.5				Grokked Transformer
	Direct+R	CoT+R	Direct	CoT	Direct+R	CoT+R	
Accuracy (%)	33.3	31.3	28.7	11.3	37.3	12.0	**99.3**



Summary & Discussion

- Grokking in the acquisition of implicit reasoning skills
- Various levels of generalization across tasks & rules
- The binding problem in Transformer models
    - Both individual concepts & atomic knowledge pieces
    - Need systematic mechanisms with less human scaffolding
- Explicit & implicit reasoning
    - Chain-of-thought & “looped” Transformers
- Non-parametric & Parametric Memory
    - Long-context & “test-time training”

Thanks!

- https://arxiv.org/abs/2405.15071
- https://github.com/OSU-NLP-Group/GrokkedTransformer