Grokking of Implicit Reasoning in Transformers: A Mechanistic Journey to the Edge of Generalization

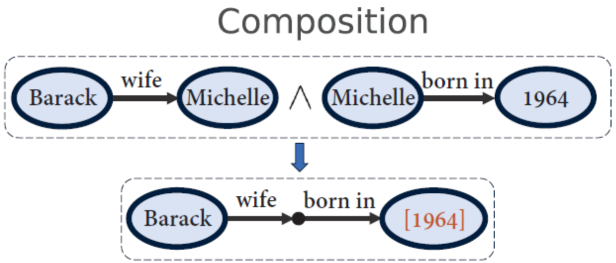
Boshi Wang, Xiang Yue, Yu Su, Huan Sun

LLMs Struggle at Implicit Reasoning w/ Parametric Memory

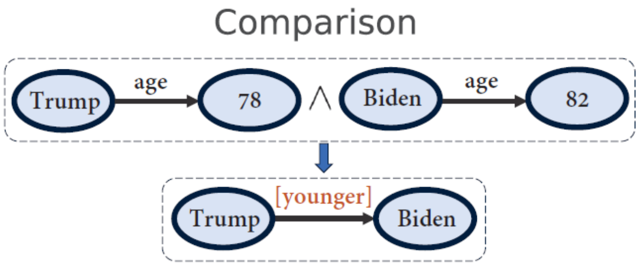
- Implicit Reasoning
  - Reasoning *without* explicit verbalization of intermediate steps
- Parametric Memory
  - Facts & rules stored in weights

LLMs Struggle at Implicit Reasoning w/ Parametric Memory

- Composition
  - LLMs only show substantial evidence in first hop reasoning (Yang et al. 2024)
  - "Compositionality gap" does not decrease with scale (Press et al. 2023)



- Comparison
  - GPT-4 struggles at implicitly comparing entity attributes despite knowing them perfectly (Zhu et al. 2023)



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. arXiv-23.

Why Does it Matter?

- Implicit Reasoning
  - Reasoning *without* explicit verbalization of intermediate steps
  - 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)

- Parametric Memory
  - Facts & rules stored in weights
  - Unique power in *compressing and integrating information at scale*
    - Important for tasks with large intrinsic complexity (example later)

Research Questions

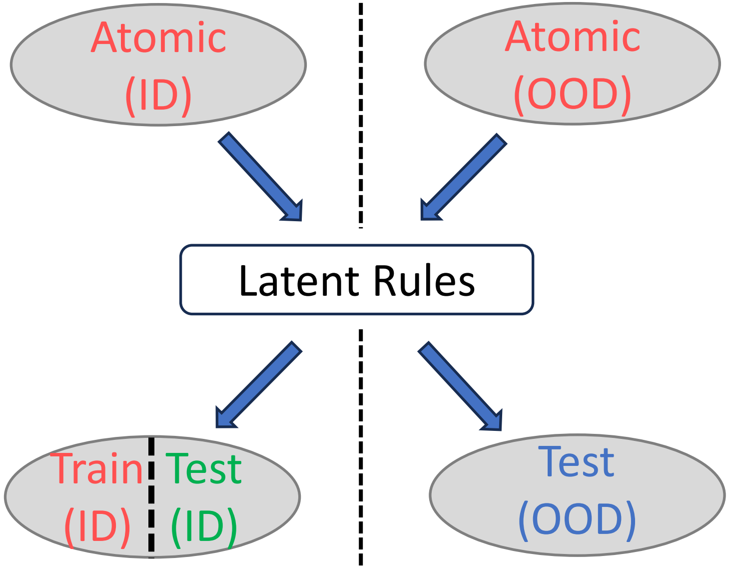
- Is implicit reasoning doomed given that even the most capable models struggle?
- Can it be resolved by further scaling data and compute, or are there fundamental limitations of transformers that prohibit robust acquisition of this skill?

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 evaluations

Approach: Synthetic Data & Training from Scratch

- Test whether the model can
  - Induce latent rules from a mixture of **atomic** facts and **inferred** facts (deduced via latent rules)
  - Deduce novel facts by applying the acquired rules
    - *Test (ID)*: unseen inferred facts deduced from the same set of atomic facts underlying the observed inferred facts
    - *Test (OOD)/systematic generalization*: unseen inferred facts derived from a different set of atomic facts (Lake et al., 2018)



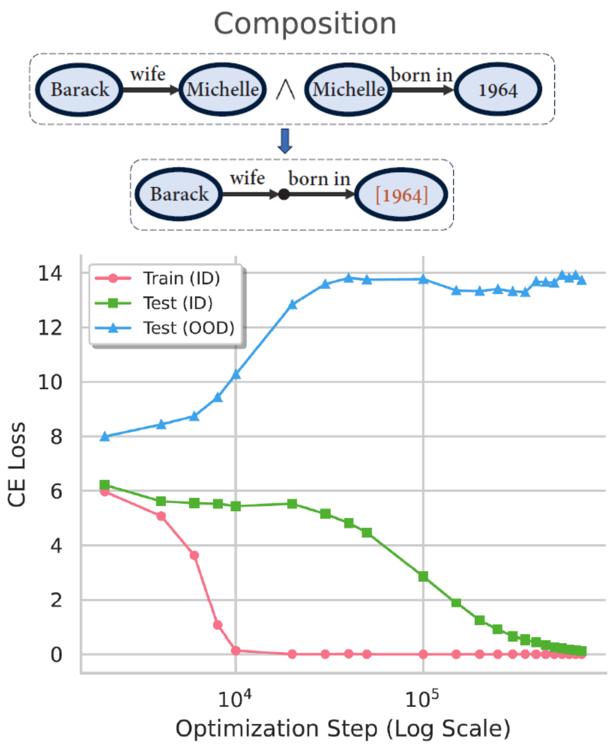
Model & Optimization

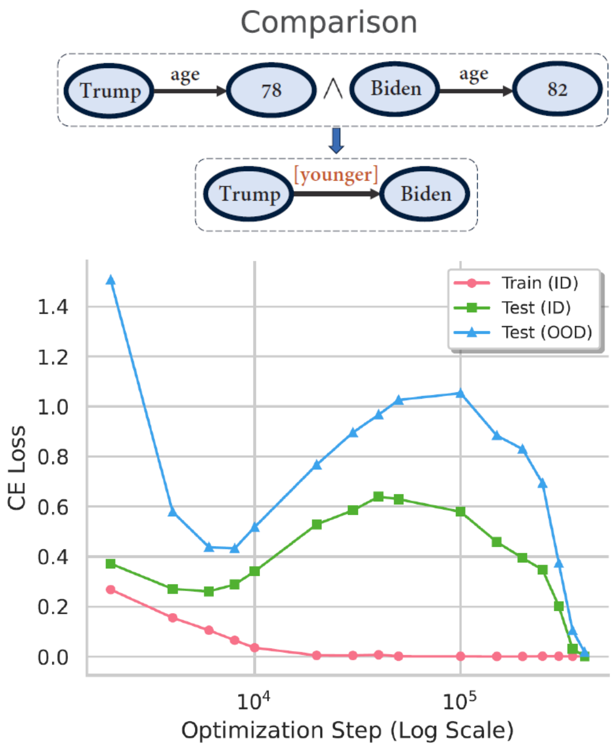
- Standard decoder-only transformer as in GPT-2
  - 8 layers, 768 hidden dimensions and 12 attention heads
  - Results robust to different model scales

- AdamW with learning rate 1e-4, batch size 512, weight decay 0.1 and 2000 warm-up steps

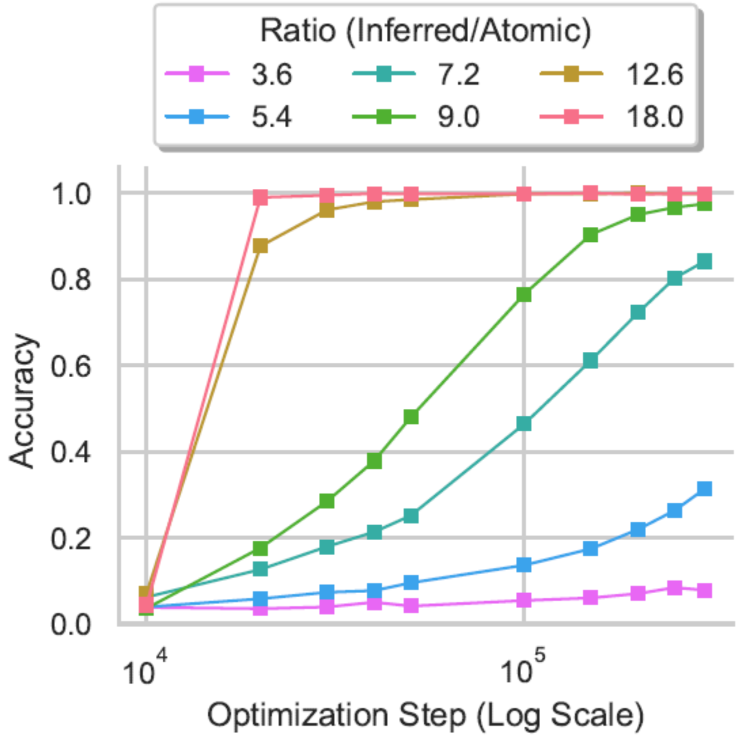
Results

- 1) Unique role of grokking 2) Difference in systematicity in generalization



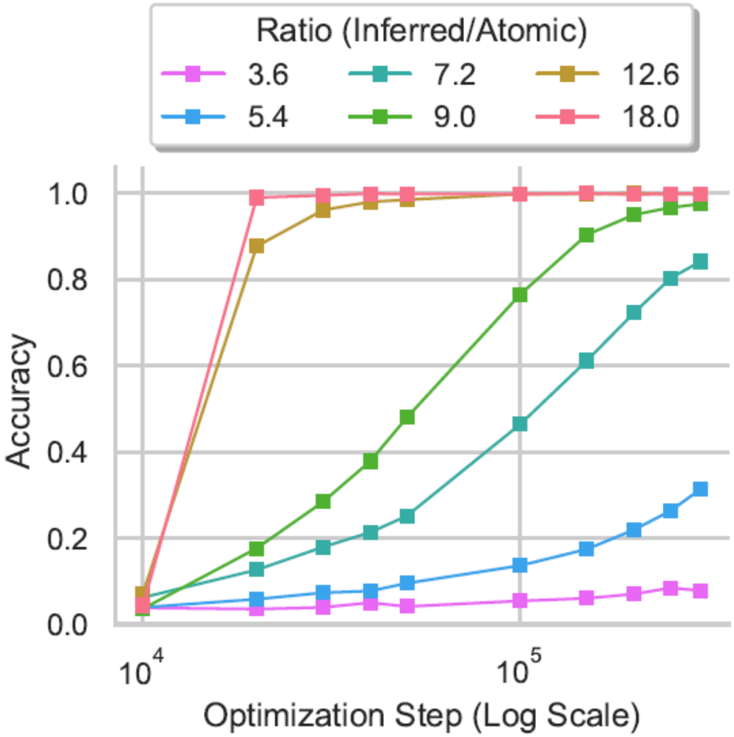


Critical Data Size?

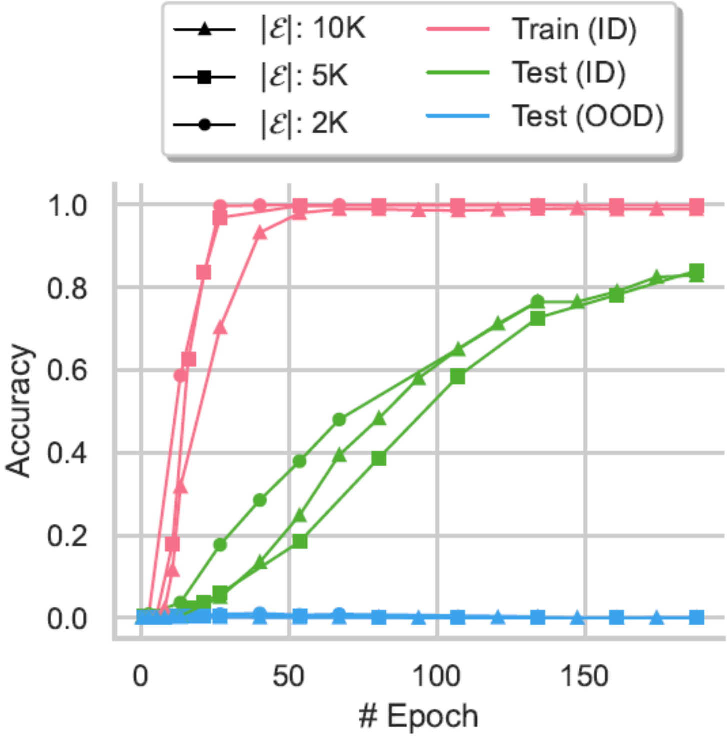


(a) Effect of the inferred/atomic ratio ϕ.

Critical Data ~~Size~~ Distribution?



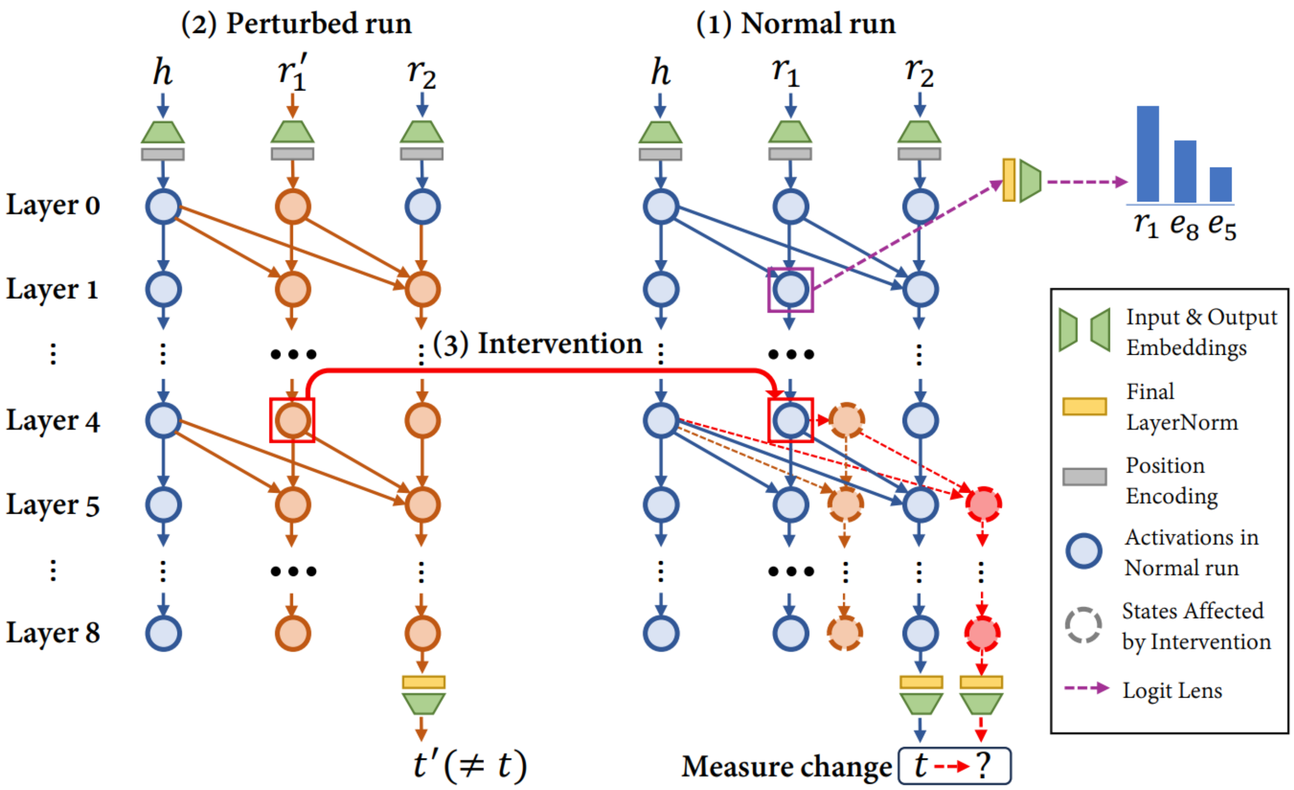
(a) Effect of the inferred/atomic ratio ϕ.



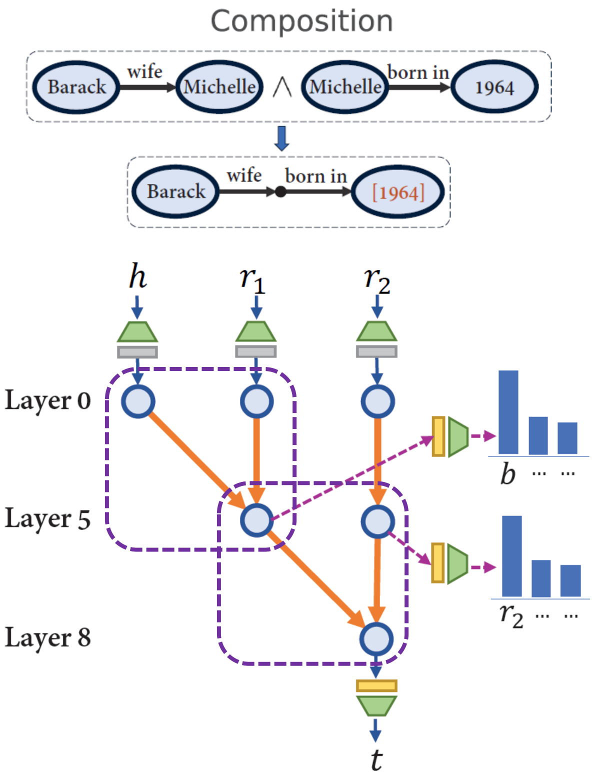
(b) Effect of changing $|\mathcal{E}|$ ($\phi = 9.0$).

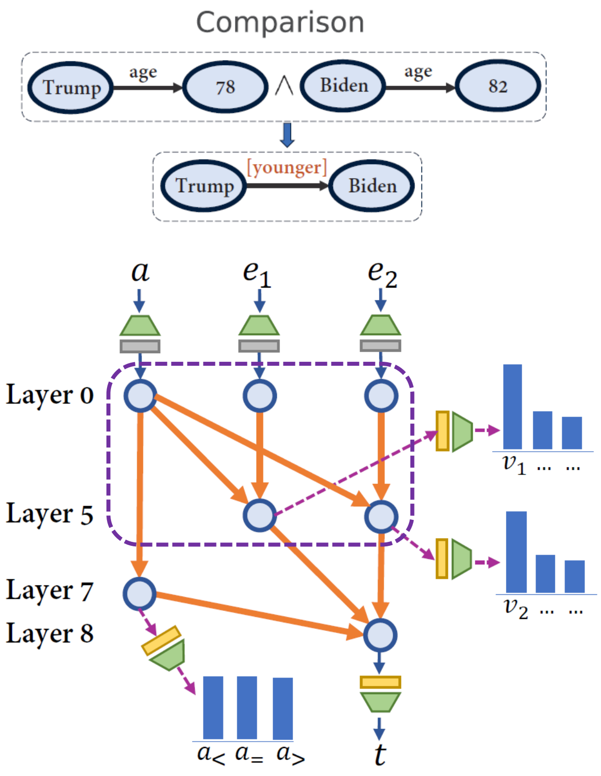
Analyzing the (change) in Inner Workings during Grokking

- Logit Lens
- Causal Tracing



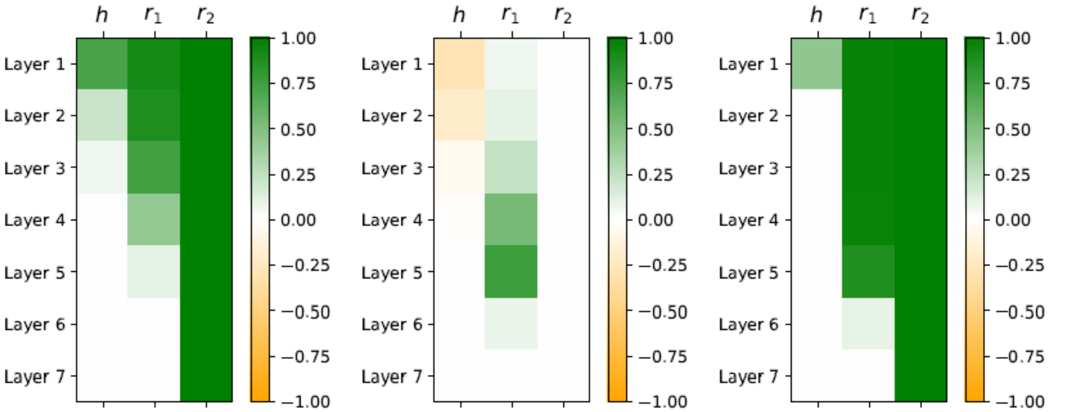
Generalizing Circuits

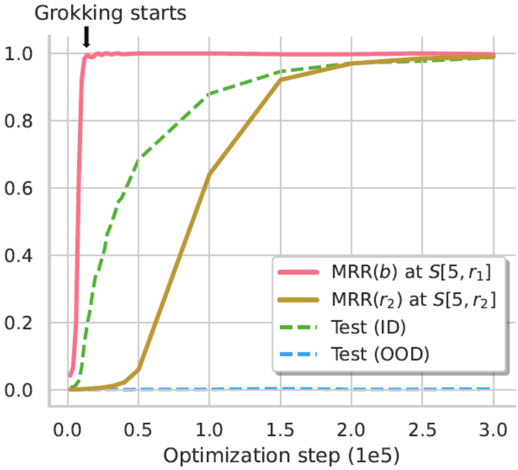


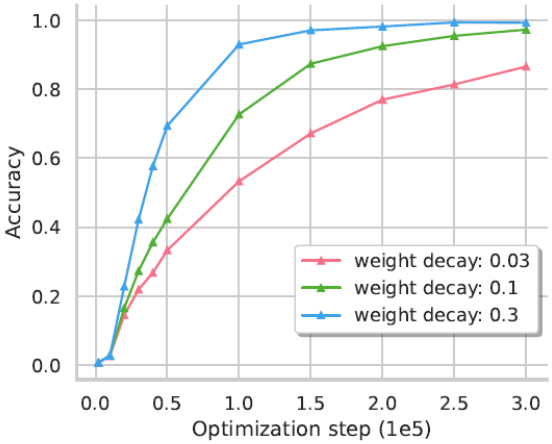


Changes during Grokking

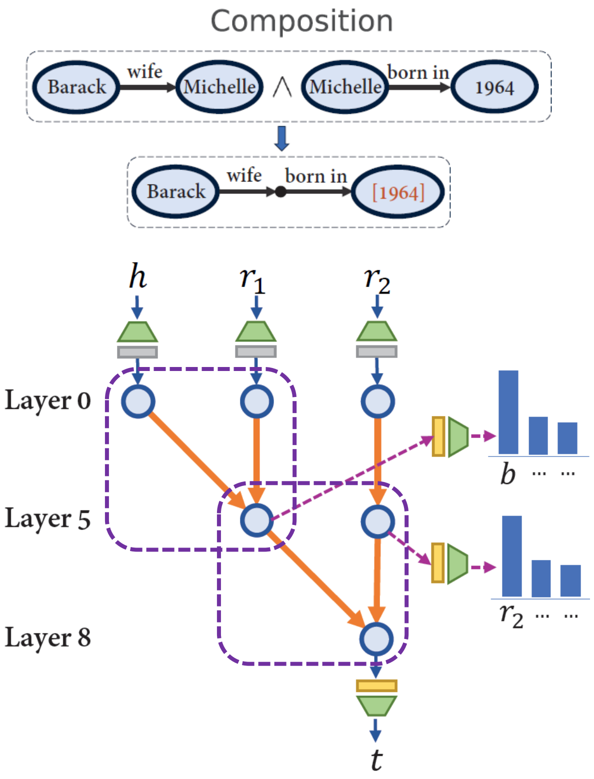
- Explanation via circuit efficiency
  - Amount of facts stored by memorizing & generalizing circuits
- Effects from regularization

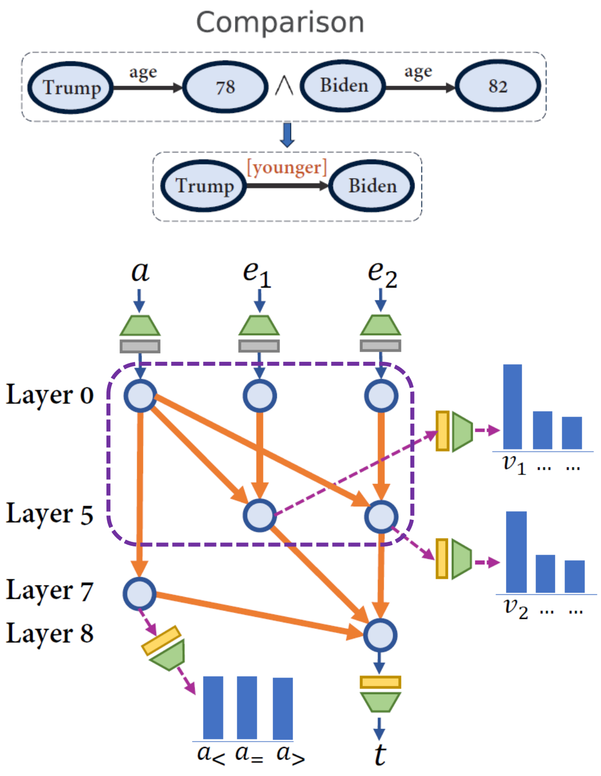




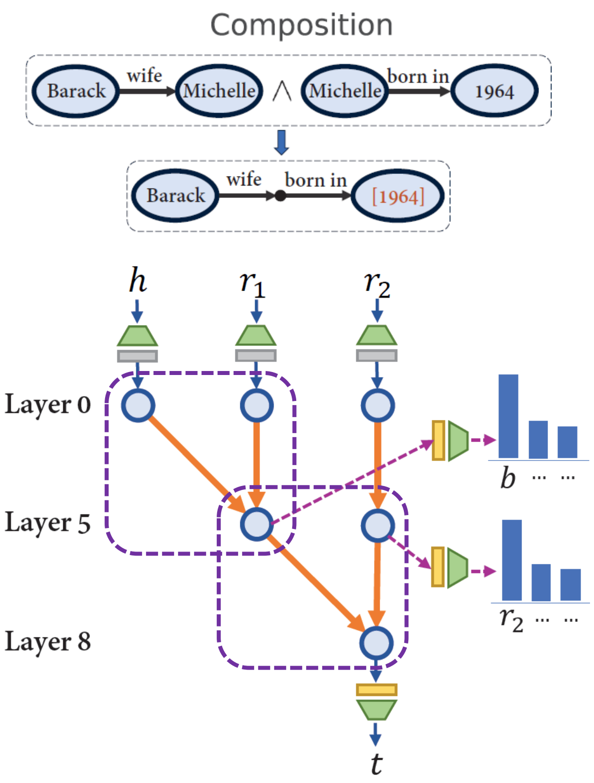


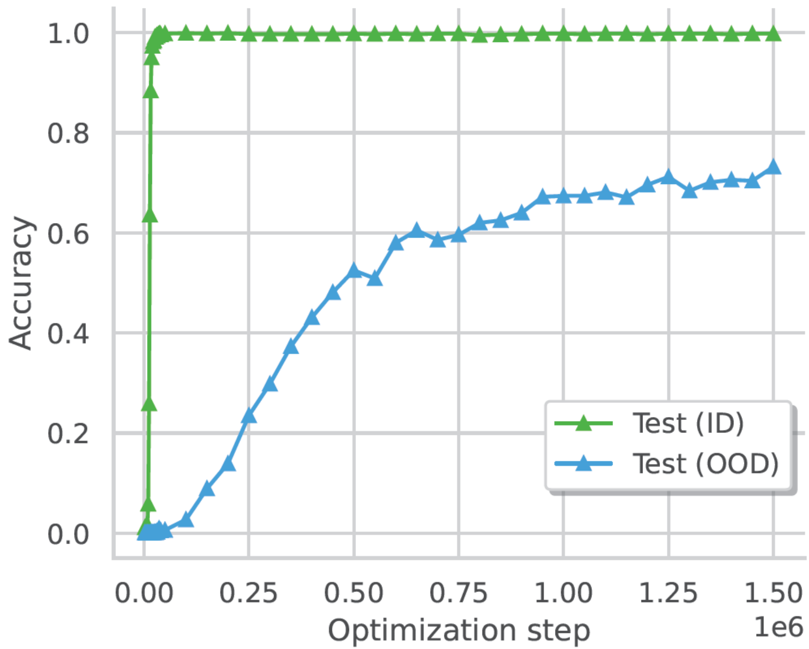
Generalizing Circuits & Systematic Generalization





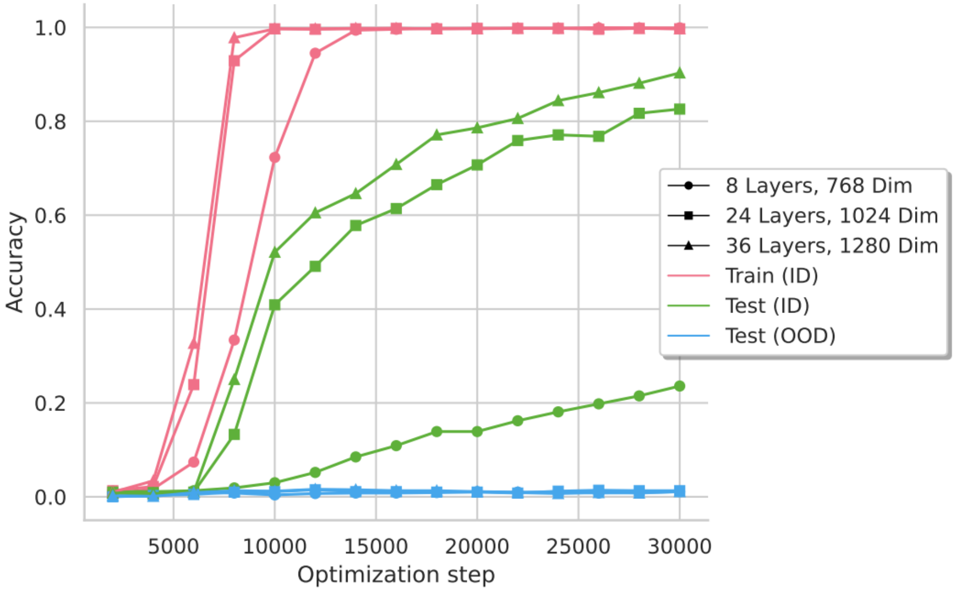
Generalizing Circuits & Systematic Generalization



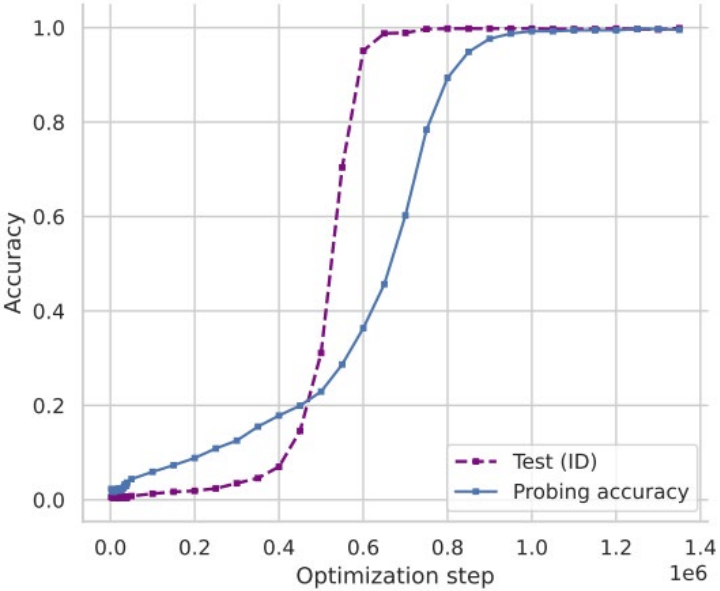


Model Scale & Tokenizations

Larger models converge in less optimization steps (no qualitative differences observed)



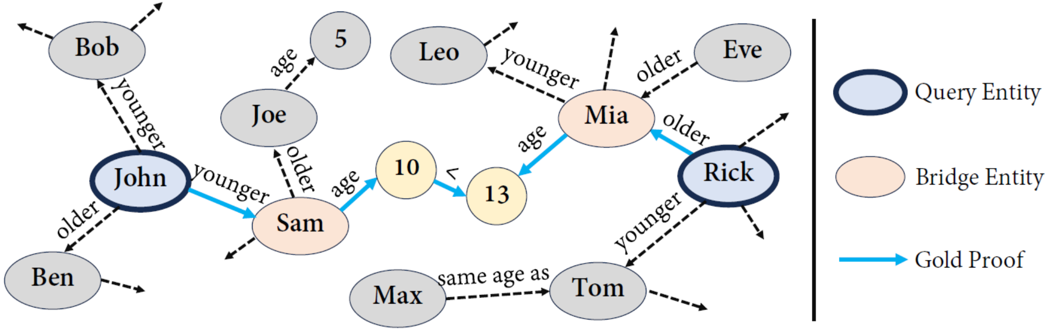
Tokens beyond immediate next token (linearly) encoded in hidden state



Both share with prior findings

The Power of Parametric Memory for Complex Reasoning

- Reasoning task with large search space & no surface form clues



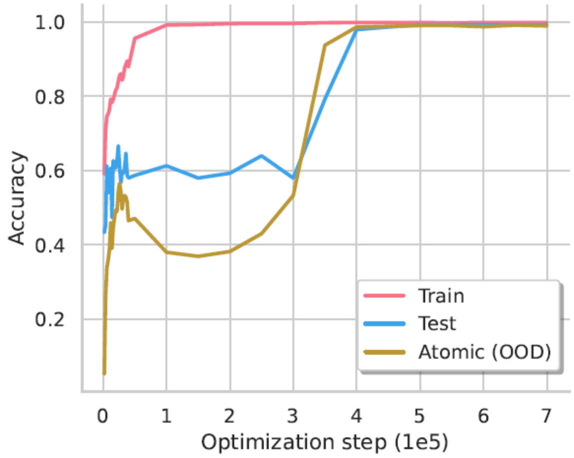


Table 1: Results on the complex reasoning task. Direct/CoT: predict the answer directly/verbalize the reasoning steps. "+R": retrieval augmentation.

	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**

Thanks!