Learning from Synthetic Data Improves Multi-hop Reasoning

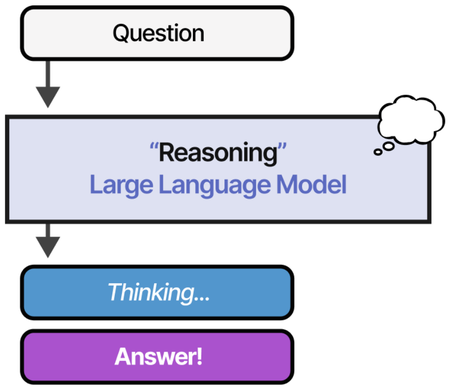
ICLR 2026 + DATA-FM workshop + VerifAI workshop

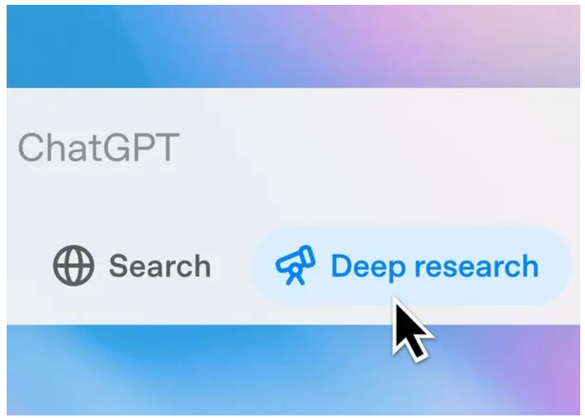
Anmol Kabra, Yilun Yin, Albert Gong, Kamilė Stankevičiūtė, Dongyoung Go, Johann Lee, Katie Z. Luo, Carla P. Gomes, Kilian Q. Weinberger

Cornell University, University of Cambridge, Stanford University

Training for LLM Reasoning

Recent advances in LLMs thanks to “reasoning”





deepseek

LLMs generally fine-tuned with
Reinforcement Learning from Verifiable Rewards (RLVR)

Training for LLM Reasoning

- Key challenge for Reinforcement Learning fine-tuning: **high-quality data**
  - Expensive
  - Scarce
  - Noisy

Real vs. Synthetic Data

Real-world Reasoning Data

- **Questions are complex**: require navigating real-world knowledge and rich language
- **Example**: "*Aside from Yodobashi, what other towns were merged into the ward which gave the major Japanese electronics retail chain its name?*"

Expensive and scarce ✗

Well-aligned to real-world problems ✓

Real vs. Synthetic Data

Real-world Reasoning Data

- **Questions are complex**: require navigating real-world knowledge and rich language
- **Example**: "*Aside from Yodobashi, what other towns were merged into the ward which gave the major Japanese electronics retail chain its name?*"

Expensive and scarce ✗

Well-aligned to real-world problems ✓

Rule-generated Synthetic Data

- **Questions are simple**: generated from templates & programs on fictional identities
- **Example**: "*Who is the nephew of the friend of the person whose hobby is birdwatching?*"

Free and infinite ✓

Large gap to real-world problems ✗

Real vs. Synthetic Data

Real-world Reasoning Data

- **Questions are complex**: require navigating real-world knowledge and rich language
- **Example**: *"Aside from Yodobashi, what other towns were merged into the ward which gave the major Japanese electronics retail chain its name?"*

Expensive and scarce ✗

Well-aligned to real-world problems ✓

Rule-generated Synthetic Data

- **Questions are simple**: generated from templates & programs on fictional identities
- **Example**: *"Who is the nephew of the friend of the person whose hobby is birdwatching?"*

Free and infinite ✓

Large gap to real-world problems ✗

Can LLMs develop transferrable reasoning from synthetic data alone?

RULE-GENERATED

Synthetic Data

Fictional • Templated • Zero real knowledge

PhantomWiki

GSM-∞

ReasoningGym

EXAMPLE

Who is the nephew of the friend of the person whose hobby is birdwatching?

- ✦ Fully verifiable by construction
- ✦ Free, infinite scale, on any machine
- ✦ Precise difficulty control

RL Fine-tuning

RULE-GENERATED

Synthetic Data

Fictional • Templated • Zero real knowledge

PhantomWiki

GSM-∞

ReasoningGym

EXAMPLE

Who is the nephew of the friend of the person whose hobby is birdwatching?

- ✦ Fully verifiable by construction
- ✦ Free, infinite scale, on any machine
- ✦ Precise difficulty control

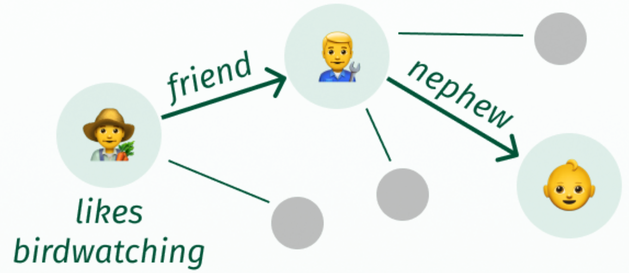
RL Fine-tuning

teaches

LEARNED SKILL

Knowledge Composition

Chaining information across multiple steps



LLMs learn to compose knowledge from synthetic data alone

- ✅ Generalizes to OOD difficulties
- ✅ Improves intermediate reasoning steps
- ✅ Scales with synthetic data
- ❌ SFT fails to teach this skill

Synthetic-to-real transfer by learning to **compose knowledge**

RULE-GENERATED

Synthetic Data

Fictional • Templated • Zero real knowledge

PhantomWiki GSM-∞ ReasoningGym

EXAMPLE

Who is the nephew of the friend of the person whose hobby is birdwatching?

- ✦ Fully verifiable by construction
- ✦ Free, infinite scale, on any machine
- ✦ Precise difficulty control

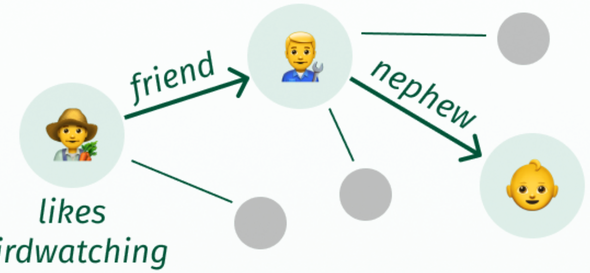
RL Fine-tuning

teaches

LEARNED SKILL

Knowledge Composition

Chaining information across multiple steps



LLMs learn to compose knowledge from synthetic data alone

- ✅ Generalizes to OOD difficulties
- ✅ Improves intermediate reasoning steps
- ✅ Scales with synthetic data
- ❌ SFT fails to teach this skill

transfers

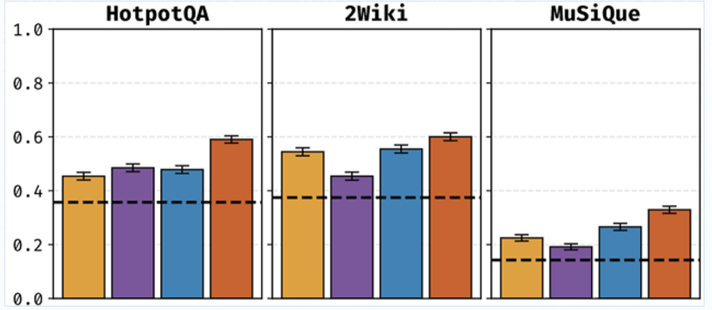
REAL-WORLD

Multi-Hop Reasoning

Knowledge-intensive • Linguistically rich

HotpotQA MuSiQue SynthWorlds 2Wiki CofCA

Improved LLM Reasoning

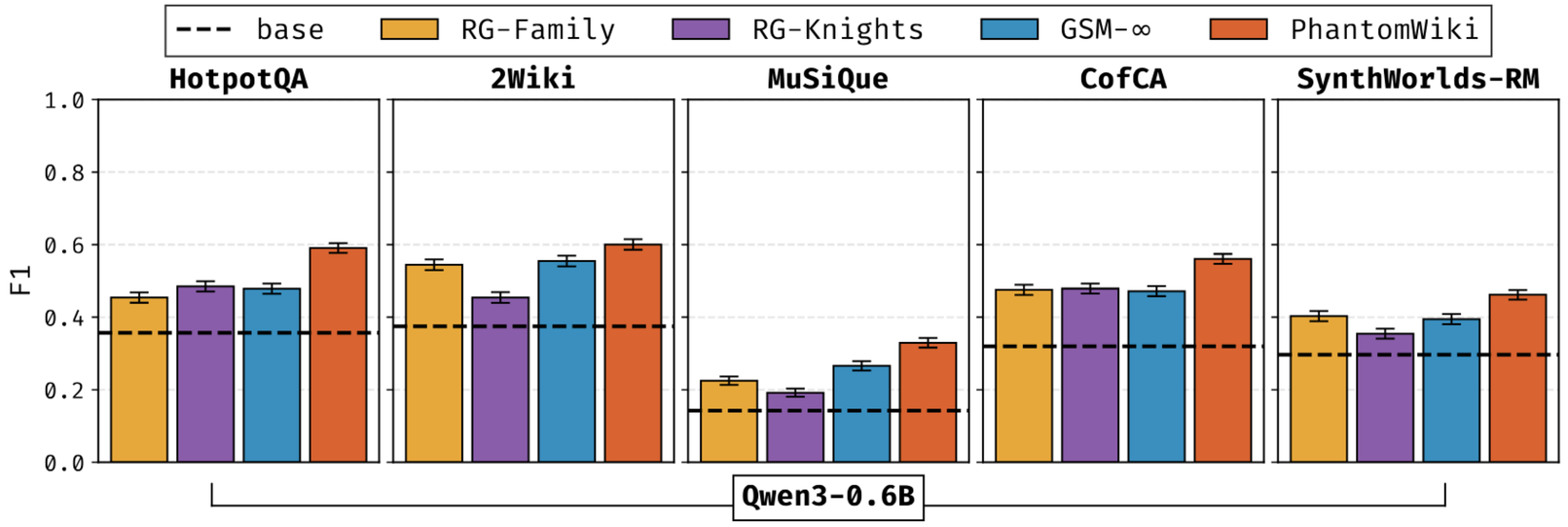


Broad Performance Transfer

Across:

- From all synthetic training data
- To all real-world evaluation benchmarks
- LLM families
- LLM sizes

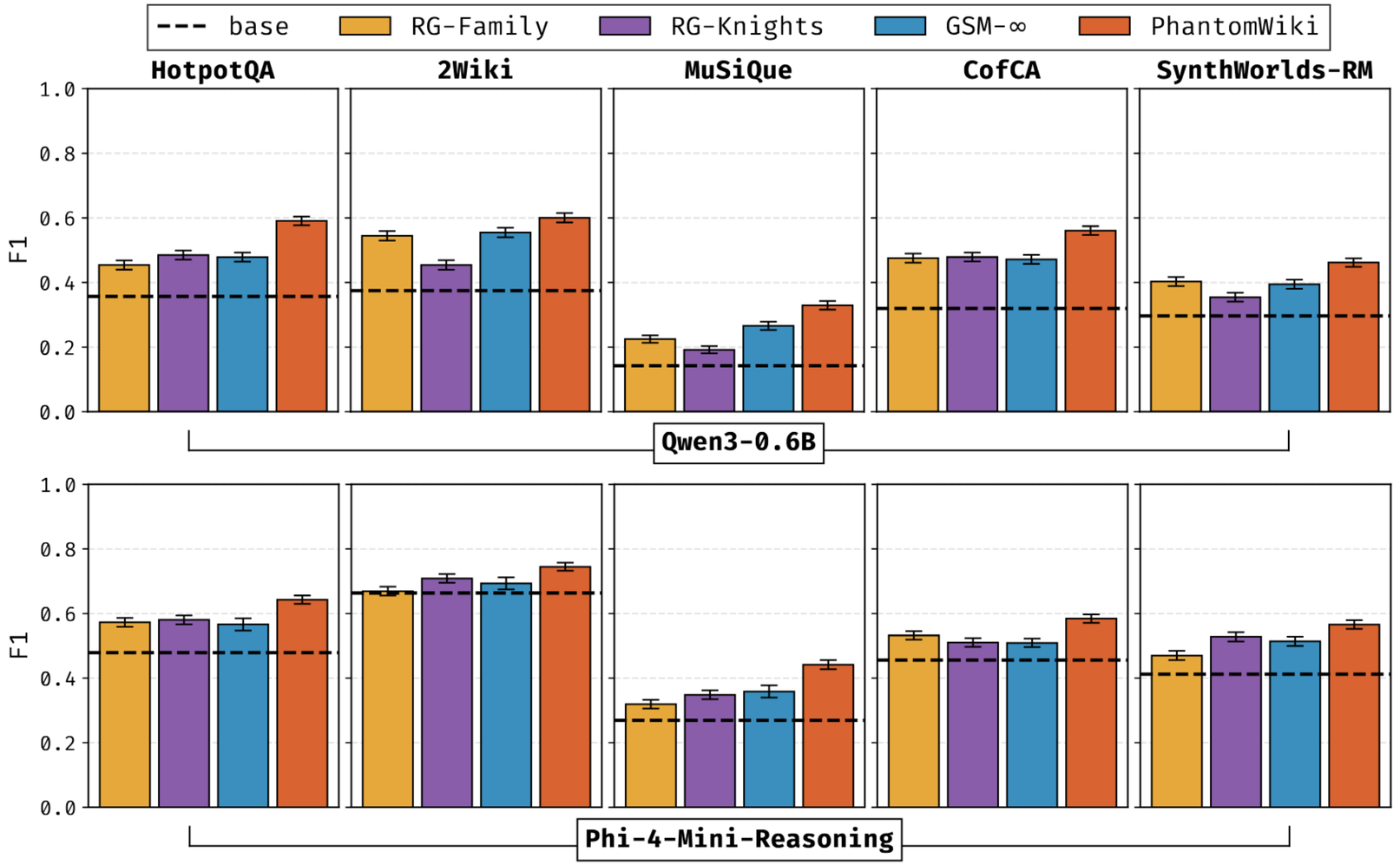
Broad Performance Transfer



Across:

- From all synthetic training data
- To all real-world evaluation benchmarks
- LLM families
- LLM sizes

Broad Performance Transfer



Across:

- From all synthetic training data
- To all real-world evaluation benchmarks
- LLM families
- LLM sizes

Why is transfer working?

Why is transfer working?

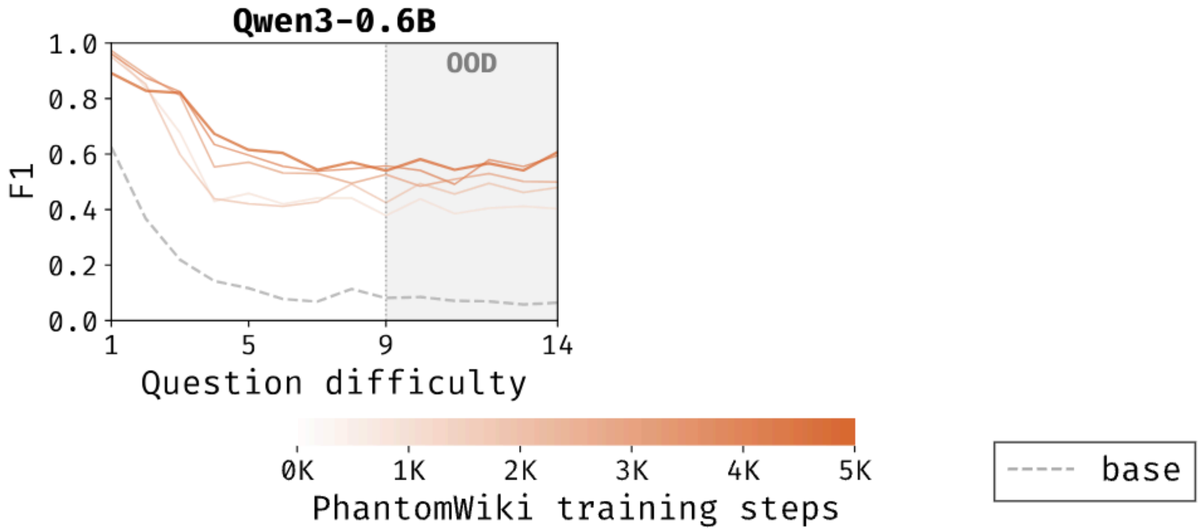
- Are LLMs just surfacing pretrained knowledge?

Why is transfer working?

- Are LLMs just surfacing pretrained knowledge?
  - Evaluate on fictional synthetic worlds

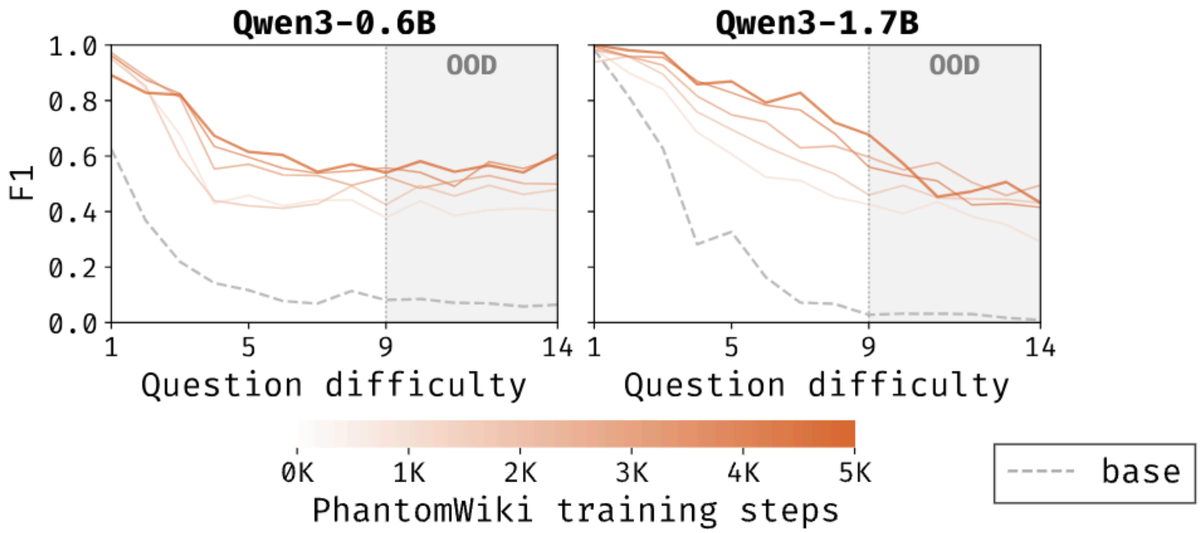
Why is transfer working?

- Are LLMs just surfacing pretrained knowledge?
  - Evaluate on fictional synthetic worlds
  - Evaluate on more difficult questions (out-of-domain)



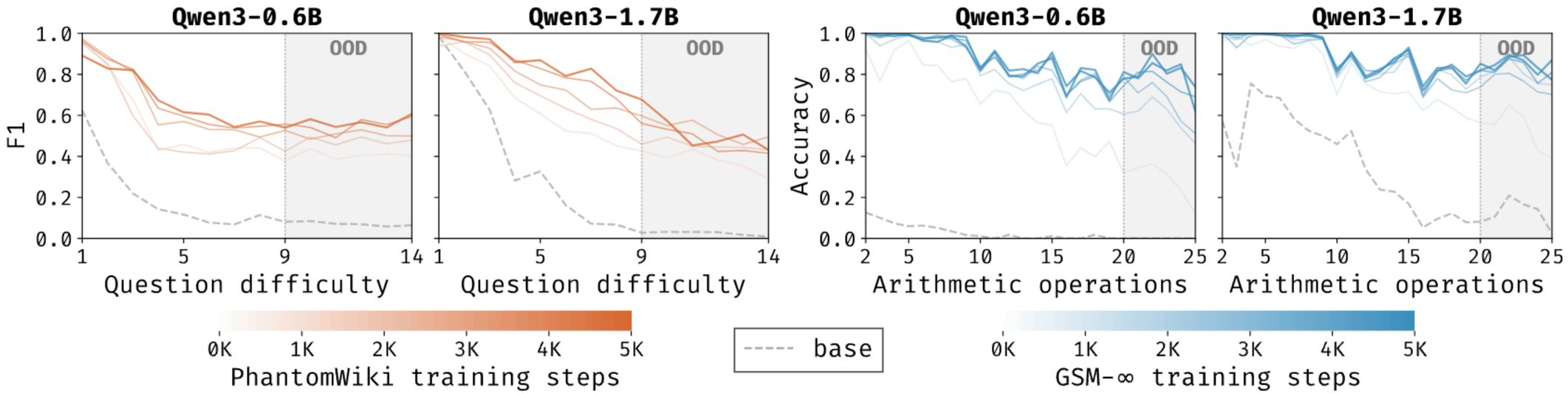
Why is transfer working?

- Are LLMs just surfacing pretrained knowledge?
  - Evaluate on fictional synthetic worlds
  - Evaluate on more difficult questions (out-of-domain)



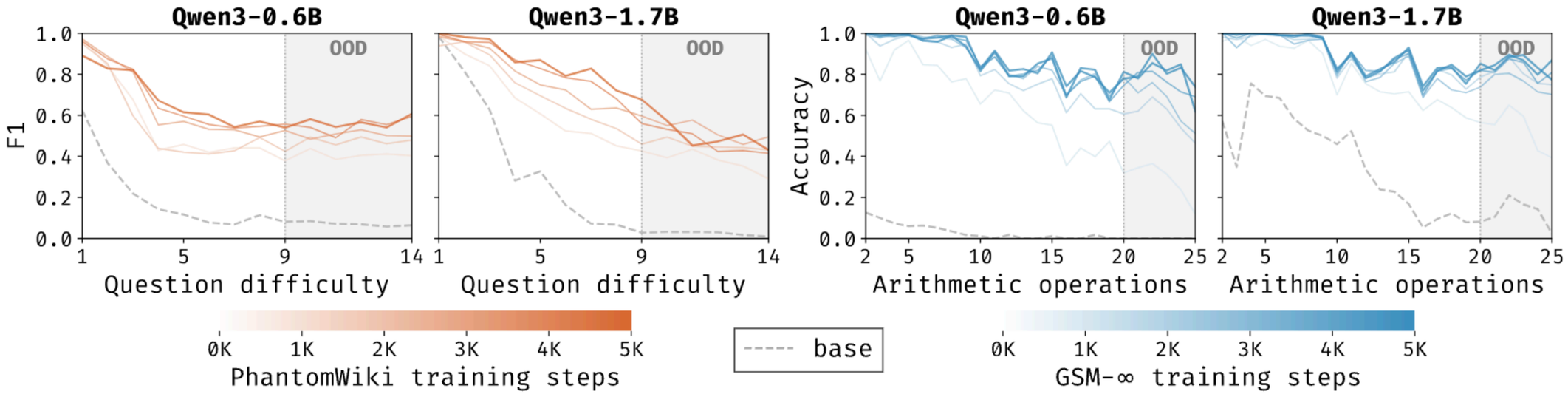
Why is transfer working?

- Are LLMs just surfacing pretrained knowledge?
  - Evaluate on fictional synthetic worlds
  - Evaluate on more difficult questions (out-of-domain)



Why is transfer working?

- Are LLMs just surfacing pretrained knowledge?
  - Evaluate on fictional synthetic worlds
  - Evaluate on more difficult questions (out-of-domain)



***Conclusion**: "pure" knowledge composition*

Is RL fine-tuning necessary?

- Supervised fine-tuning (SFT) is standard transfer technique
- SFT on golden solutions from GSM-infinite

Is RL fine-tuning necessary?

- Supervised fine-tuning (SFT) is standard transfer technique
- SFT on golden solutions from GSM-infinite

	Accuracy on GSM-∞		F1 score on HotpotQA	
	base	SFT	base	SFT
Qwen3-0.6B	0.0241	**0.7735**	0.3569	0.3995
Qwen3-1.7B	0.1354	0.7742	0.5935	0.5761

Is RL fine-tuning necessary?

- Supervised fine-tuning (SFT) is standard transfer technique
- SFT on golden solutions from GSM-infinite

	Accuracy on GSM-∞			F1 score on HotpotQA		
	base	SFT	RL	base	SFT	RL
Qwen3-0.6B	0.0241	**0.7735**	0.6452	0.3569	0.3995	**0.4786**
Qwen3-1.7B	0.1354	0.7742	**0.8532**	0.5935	0.5761	**0.6664**

How is transfer working?

How is transfer working?

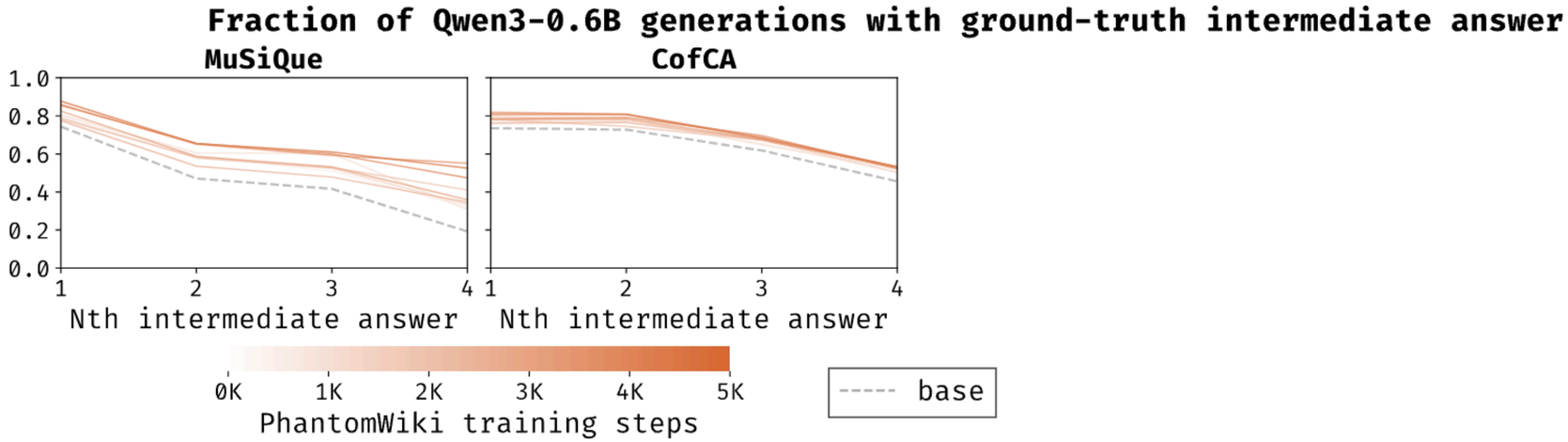
- In RL, we only reward LLM's final answer. What's going under the hood?

How is transfer working?

- In RL, we only reward LLM's final answer. What's going under the hood?
- In reasoning traces, count # of times they contain intermediate answers

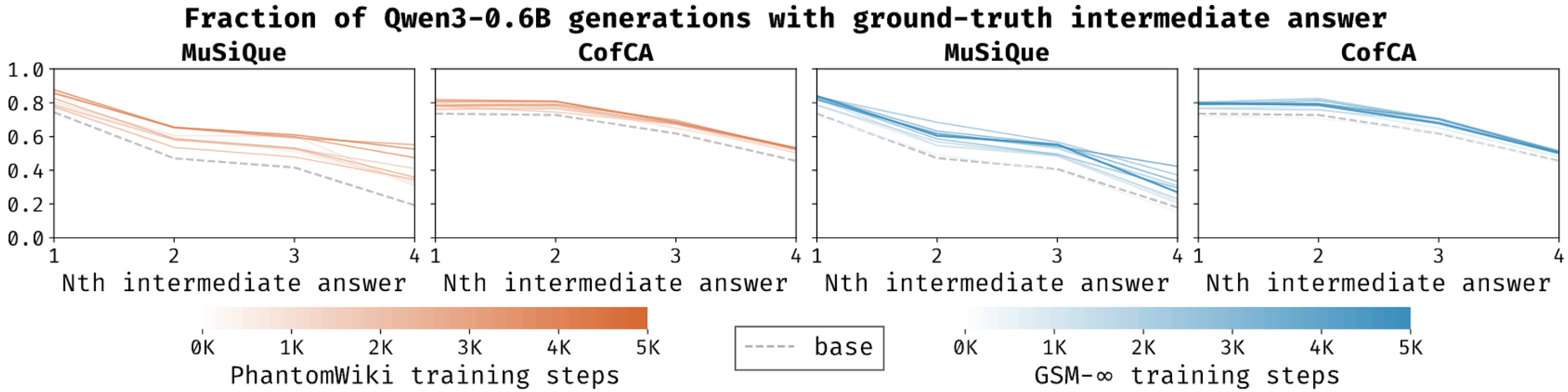
How is transfer working?

- In RL, we only reward LLM's final answer. What's going under the hood?
- In reasoning traces, count # of times they contain intermediate answers
- LLMs are learning to take correct intermediate steps



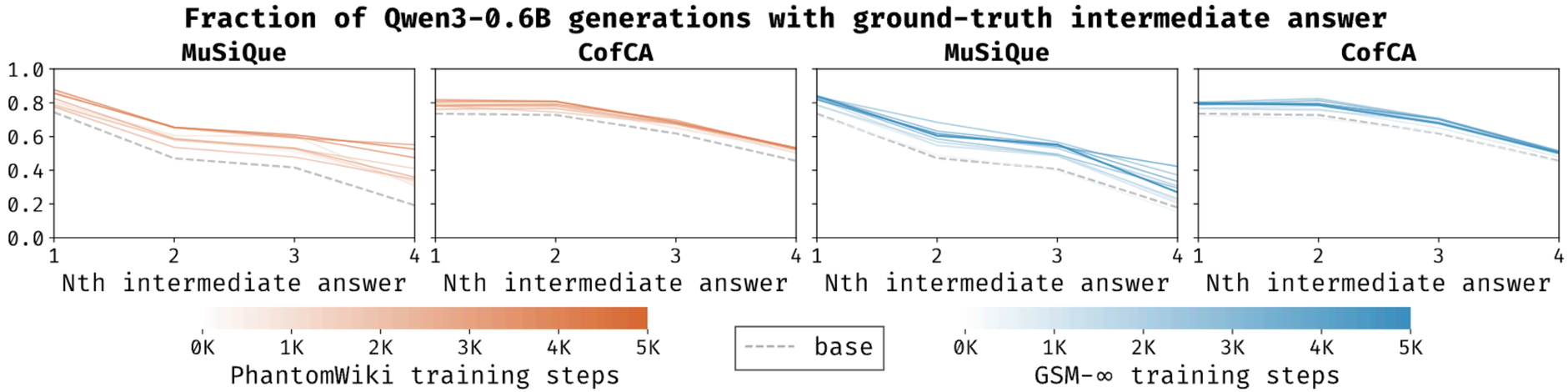
How is transfer working?

- In RL, we only reward LLM's final answer. What's going under the hood?
- In reasoning traces, count # of times they contain intermediate answers
- LLMs are learning to take correct intermediate steps



How is transfer working?

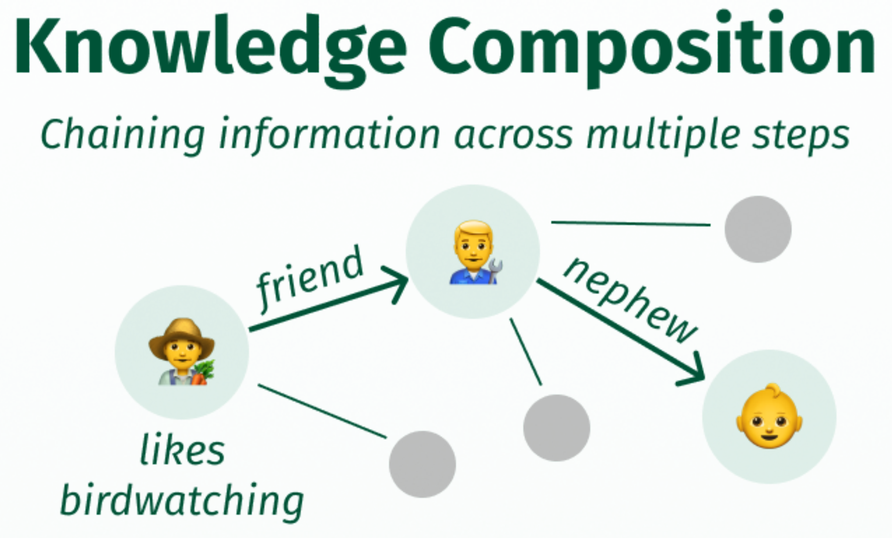
- In RL, we only reward LLM's final answer. What's going under the hood?
- In reasoning traces, count # of times they contain intermediate answers
- LLMs are learning to take correct intermediate steps



***Conclusion**: Outcome-reward also improving reasoning "process"*

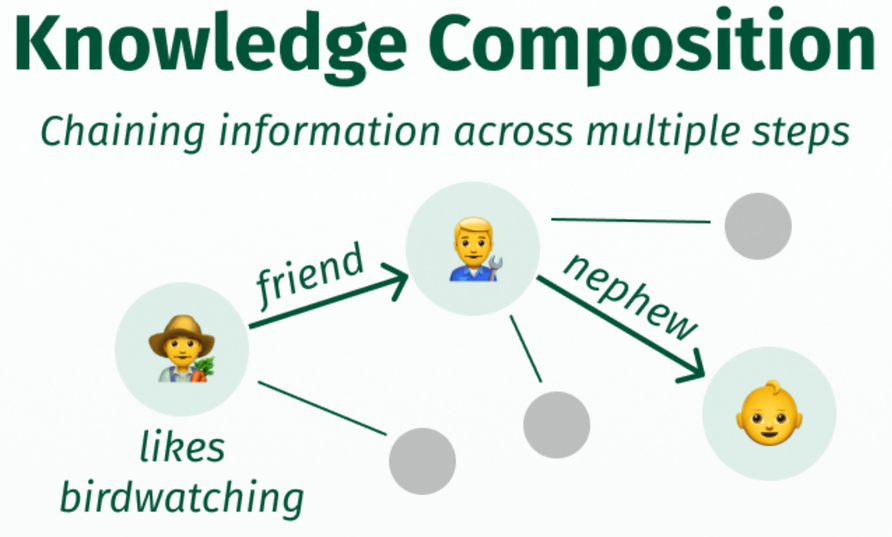
Takeaways

- We can teach LLMs the **fundamental skill** of composing knowledge
  - Fictional worlds
  - Simple language
  - From synthetic data alone

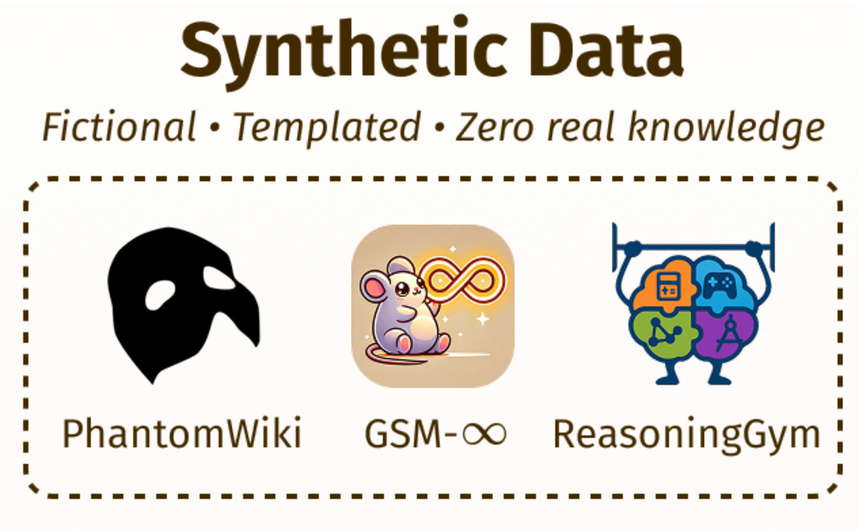


Takeaways

- We can teach LLMs the **fundamental skill** of composing knowledge
  - Fictional worlds
  - Simple language
  - From synthetic data alone



- Rule-generated synthetic data: a **new data source** for LLM reasoning
  - Free, scalable, on any machine
  - Alleviates data bottleneck for scaling RL fine-tuning



Poster session #3
Friday, April 24
10:30am-1pm

Thanks

- NSF NAIRR Pilot, Anvil AI
  - Compute for training
  - Scaling runs to all datasets
  - Feasibility testing
- Early access to NVIDIA DGX Station
  - Scaling runs to bigger LLMs

NAIRR Pilot

NVIDIA

CORNELL UNIVERSITY
FOUNDED A.D. 1865

AI-MI
ARTIFICIAL INTELLIGENCE
MATERIALS INSTITUTE

NSF