



Anmol Kabra



Yilun Yin



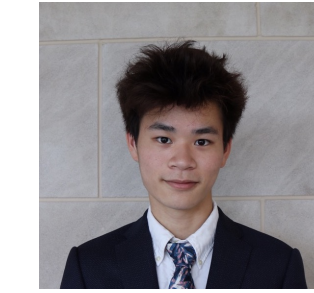
Albert Gong



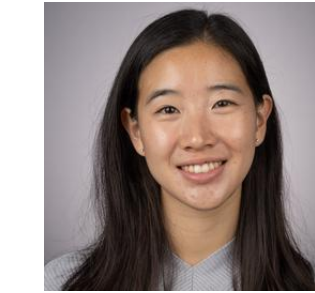
Kamilė Stankevičiūtė



Dongyoung Go



Johann Lee



Katie Z. Luo

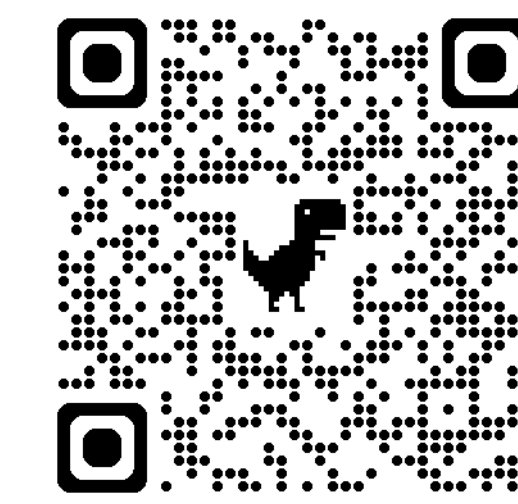


Carla P. Gomes

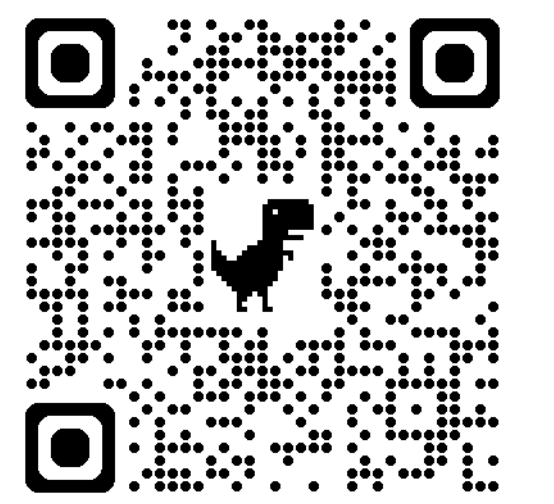


Kilian Q. Weinberger

Video



Paper



ICLR 2026 conference
DATA-FM workshop
VerifAI workshop

Learning from Synthetic Data Improves Multi-hop Reasoning

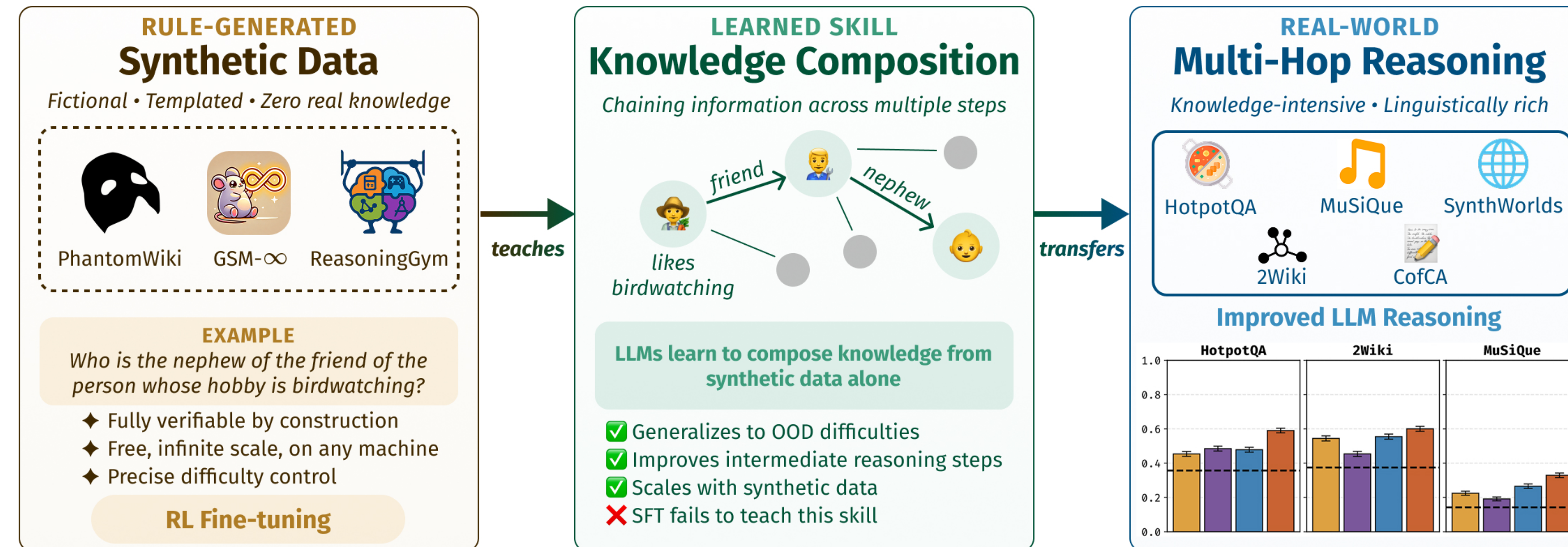
Takeaways

RL fine-tuning on synthetic data teaches LLMs a fundamental skill:

knowledge composition

1. Broad synthetic-to-real performance transfer
2. LLMs generalize to OOD knowledge and difficulties
3. SFT overfits to synthetic data task, but RL generalizes to real-world multi-hop reasoning
4. Grounded reasoning traces from just outcome-reward
5. Performance scales with synthetic data samples

RL fine-tuning teaches LLMs a fundamental skill: *knowledge composition*



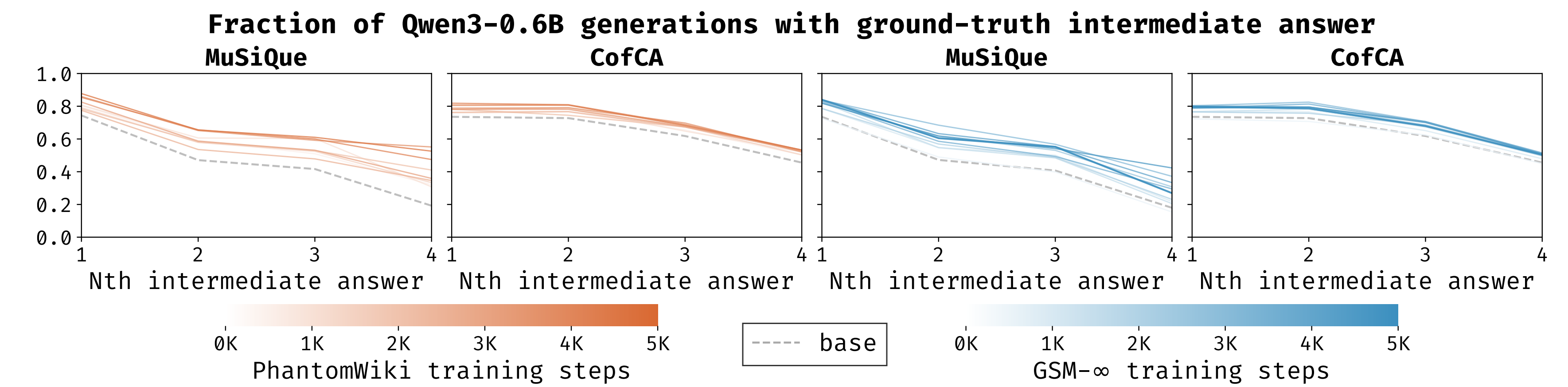
Broad performance transfer, across:

- From all synthetic datasets, to all real-world evaluation benchmarks
- On various LLM families (Qwen2.5, Qwen3, Phi) + model sizes (0.6-7B)

Synthetic data fine-tuning leads to grounded generations in real-world

How is transfer working?

- We only reward final answer i.e. outcome. What's the effect on reasoning process?
 - In reasoning traces, count # of times they contain golden intermediate answers
- LLMs' reasoning traces show higher fractions of correct intermediate steps
- **Bonus** of outcome-reward on synthetic data: **more grounded reasoning traces**



Data bottleneck in RL

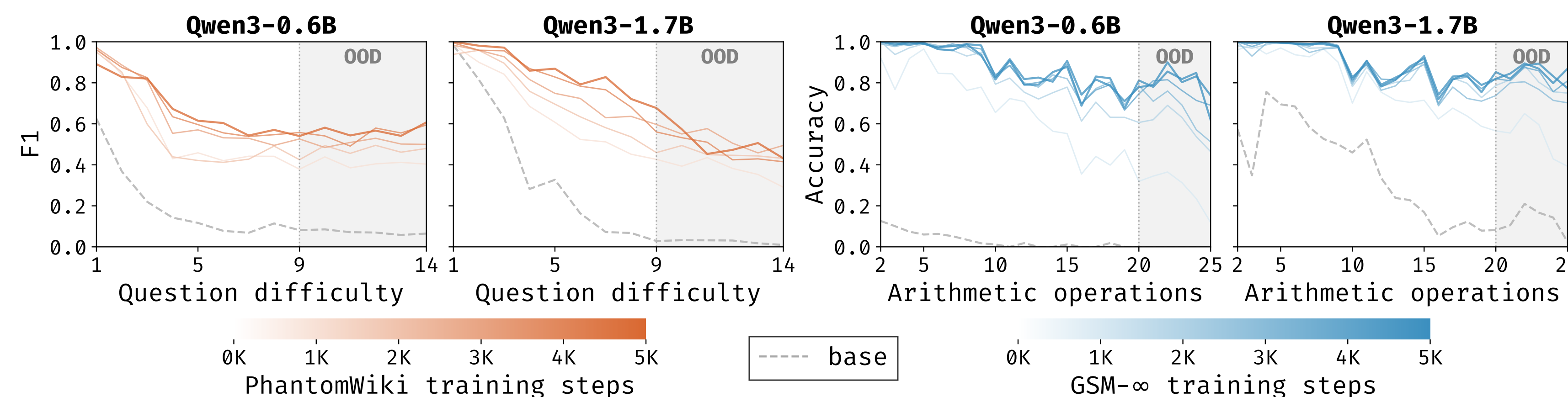
- RL from Verifiable Rewards (RLVR) is data hungry
- High-quality reasoning data from humans / frontier LLMs:
 - **Expensive, scarce, noisy**

Can LLMs develop reasoning skills from synthetic data alone?

LLMs generalize to OOD knowledge and difficulties

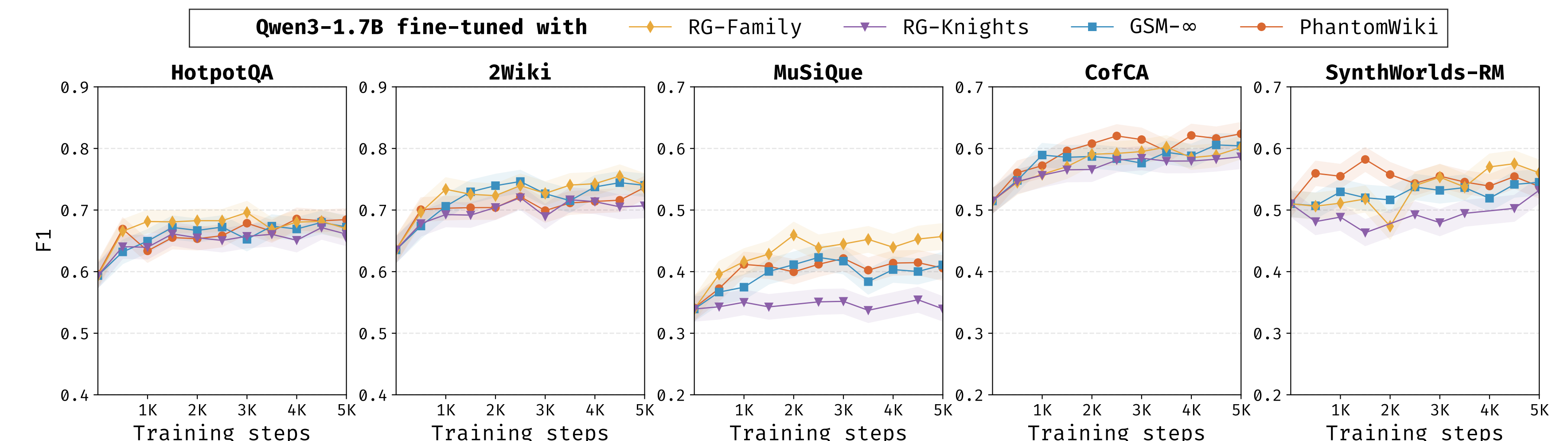
Why is transfer working?

- Are LLMs just surfacing pretrained knowledge?
 - Evaluate on fictional synthetic worlds (OOD knowledge)
 - Evaluate on more difficult questions (OOD difficulties)



Performance scales with synthetic data samples

- Every synthetic data sample is fresh $\implies$ training steps $\approx$ data samples
- Training on synthetic data shows no overfitting to real-world benchmarks
- **New axis for scaling** RL fine-tuning: scaling synthetic data



Transfer not obvious

PhantomWiki (synthetic)

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

HotpotQA (real)

Aside from Yodobashi, what other towns were merged into the ward which gave the major Japanese electronics retail chain its name?

Beyond hacking format reward

- Does RL on binary format reward help?
 - Not for Qwen3/Phi $\implies$ RL on synthetic data is **actually** transferring

		HotpotQA	2WikiMultihopQA	MuSiQue
Qwen3-0.6B	base	0.36 $\pm$ 0.02	0.37 $\pm$ 0.02	0.14 $\pm$ 0.01
	format	0.38 $\pm$ 0.02	0.34 $\pm$ 0.02	0.13 $\pm$ 0.01
Qwen3-1.7B	base	0.59 $\pm$ 0.02	0.64 $\pm$ 0.02	0.34 $\pm$ 0.02
	format	0.64 $\pm$ 0.02	0.67 $\pm$ 0.02	0.35 $\pm$ 0.02
Phi-4-mini-reasoning	base	0.48 $\pm$ 0.02	0.66 $\pm$ 0.02	0.27 $\pm$ 0.02
	format	0.47 $\pm$ 0.02	0.48 $\pm$ 0.02	0.26 $\pm$ 0.02
Qwen2.5-1.5B-Instruct	base	0.02 $\pm$ 0.01	0.14 $\pm$ 0.02	0.04 $\pm$ 0.01
	format	0.43 $\pm$ 0.02	0.30 $\pm$ 0.02	0.20 $\pm$ 0.02

SFT overfits to synthetic, RL generalizes to real

- SFT on golden GSM- ∞ solutions
- Better on GSM- ∞ , **not** HotpotQA
- **RL better** on GSM- ∞ + HotpotQA

	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

Implications + open questions

Implications

- Rule-generated synthetic data is a free + infinite resource for scaling RL fine-tuning
- Knowledge composition is a domain-independent skill

Open questions

- Mix synthetic + real data: how do composition and memorization interact?
- Other fundamental reasoning skills: planning, causal, analogical?
- What makes an LLM “good for RL”?