Rethinking Generalization in Reasoning SFT:

A Conditional Analysis on Optimization, Data, and Model Capability

Qihan Ren[1,2] Peng Wang[1,3] Ruikun Cai[3] Shuai Shao[1] Dadi Guo[1]
Yuejin Xie[1] Yafu Li[1] Quanshi Zhang[2] Xia Hu[1] Jing Shao[1] Dongrui Liu[1]

[1]Shanghai Artificial Intelligence Laboratory
[2]Shanghai Jiao Tong University [3]University of Science and Technology of China

Background: SFT does not generalize?

- **Prevailing view:** “SFT memorizes, RL generalizes”[1]
- Many efforts in explaining why RL is good and improving the SFT objective

[1] Chu et al. SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training. ICML 2025.

Background: SFT does not generalize?

- **Prevailing view:** “SFT memorizes, RL generalizes”[1]
- Many efforts in explaining why RL is good and improving the SFT objective
- **Variation in experimental settings**
  - **Optimization**: LR, epochs, total steps…
  - **Data:** source, w/ or w/o long CoT…
  - **Model**: size, pretraining distribution…
  - Retention of existing abilities vs. Cross-domain generalization gains

[1] Chu et al. SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training. ICML 2025.

Background: SFT does not generalize?

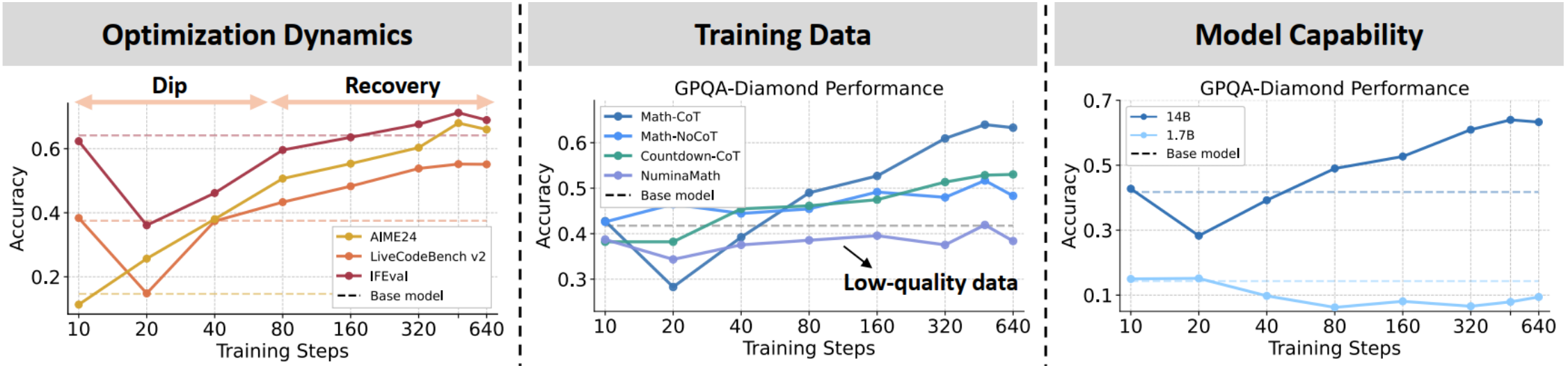
- **Prevailing view:** “SFT memorizes, RL generalizes”[1]
- Many efforts in explaining why RL is good and improving the SFT objective
- **Variation in experimental settings**
  - **Optimization**: LR, epochs, total steps…
  - **Data:** source, w/ or w/o long CoT…
  - **Model**: size, pretraining distribution…
  - Retention of existing abilities vs. Cross-domain generalization gains

> **Open question:** Do the observed failures reflect inherent limitations of SFT, or are they caused by the experimental settings?

[1] Chu et al. SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training. ICML 2025.

Motivation: a conditional framework

What we ask: When does (reasoning) SFT generalize, and at what cost?



Generalization in SFT is **NOT absent**, but **conditional**

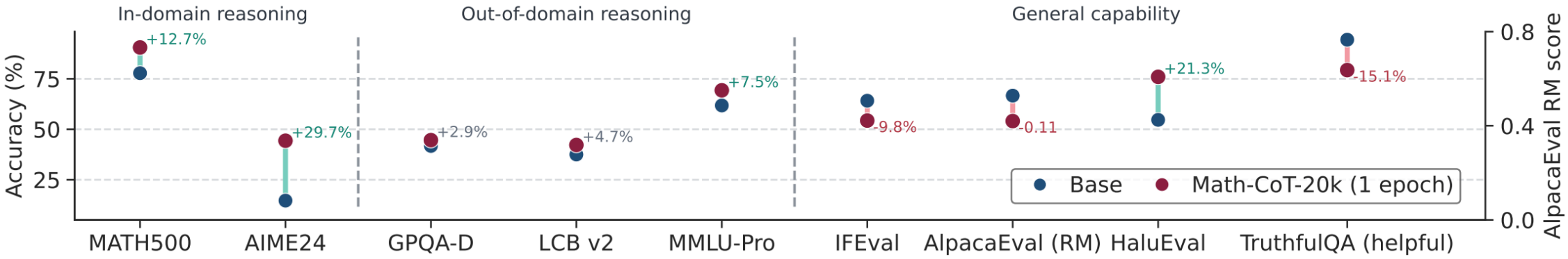
Experimental setup: a controlled testbed

- Start from base models (no instruction tuning)
- **Trained** on 20k math examples with verified long CoT (Math-CoT-20k)
- **Evaluated** on math, code, science, instruction following, helpfulness, safety
- **Interventions**:
  - Optimization schedule
  - Data source & CoT existence
  - Base model size

Optimization: replication of previous findings

Trained for one epoch, LR=5e-5:

- Math performance improves substantially
- Gains are **limited or even negative** on OOD benchmarks
- Even worse with smaller LR (1e-5)

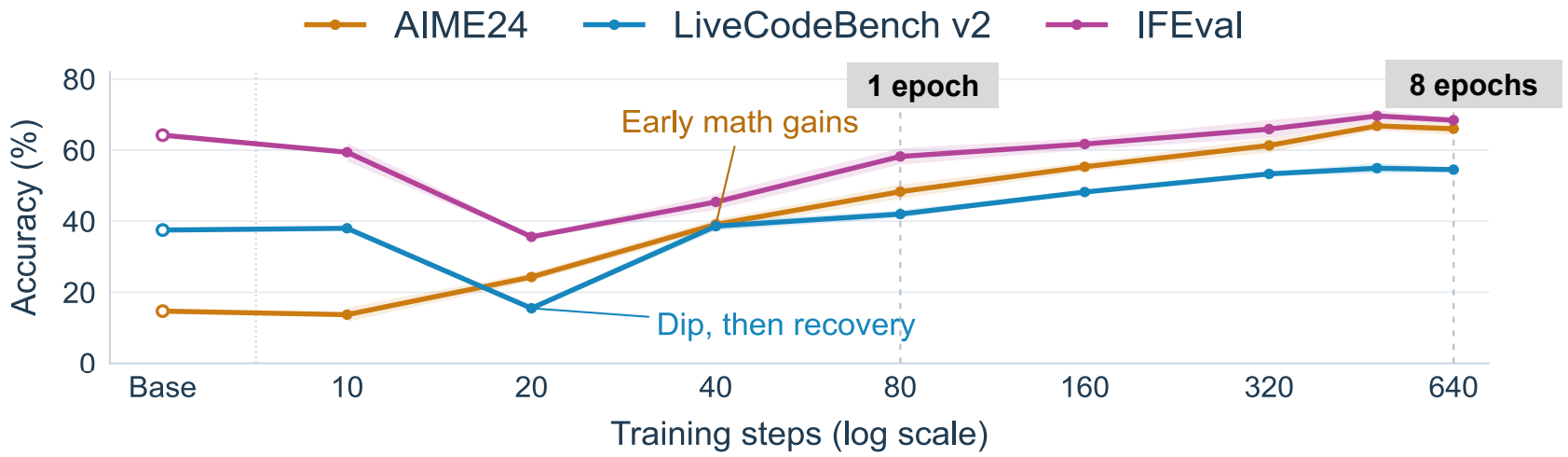


[1] Huan et al. Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning. ICML 2026.

Poor generalization may come from under-fitting

Performance:

dip-and-recovery



Short-epoch checkpoints can underestimate eventual generalization

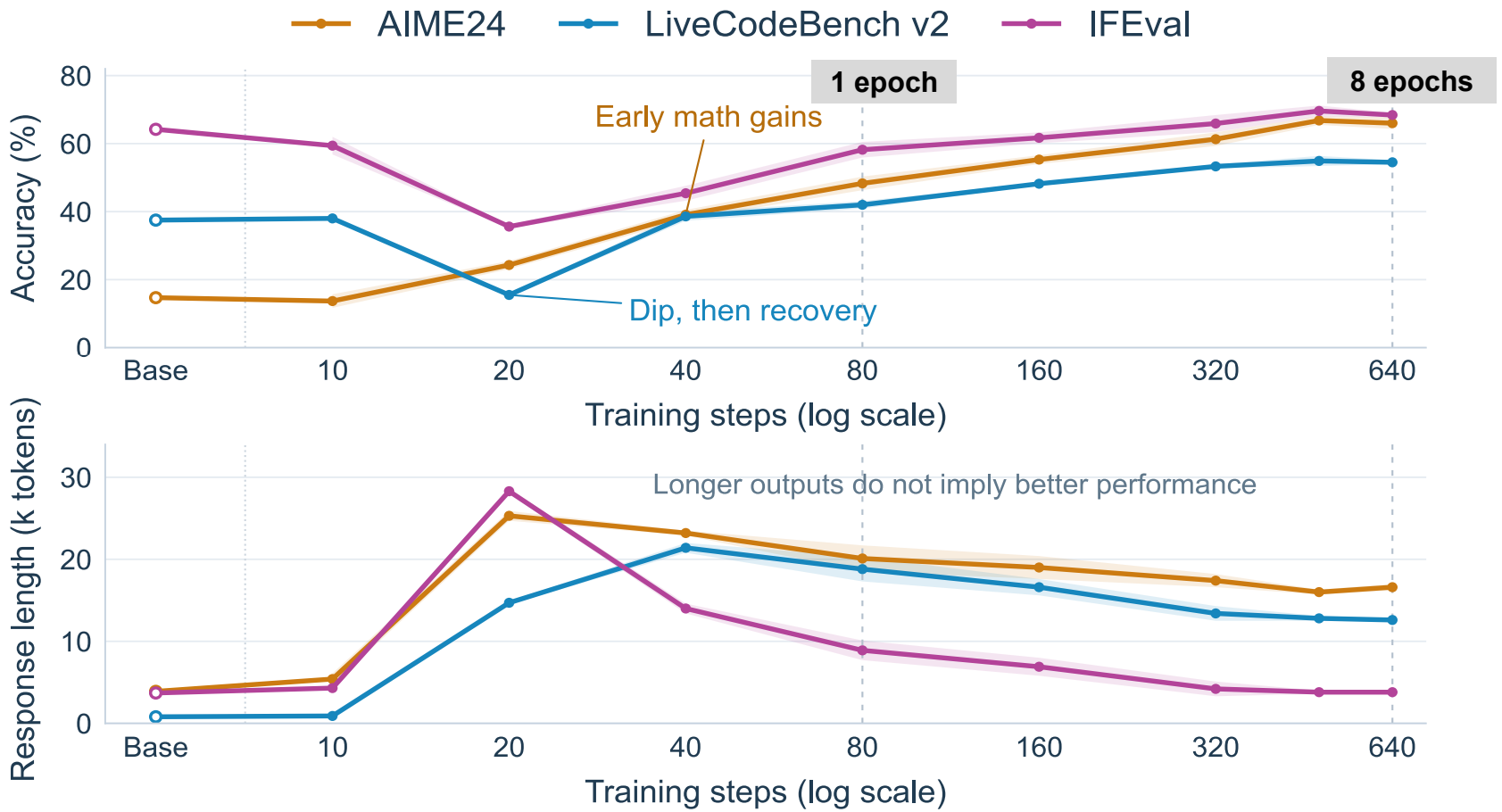
Poor generalization may come from under-fitting

Performance:

dip-and-recovery

Response length:

rise-and-fall

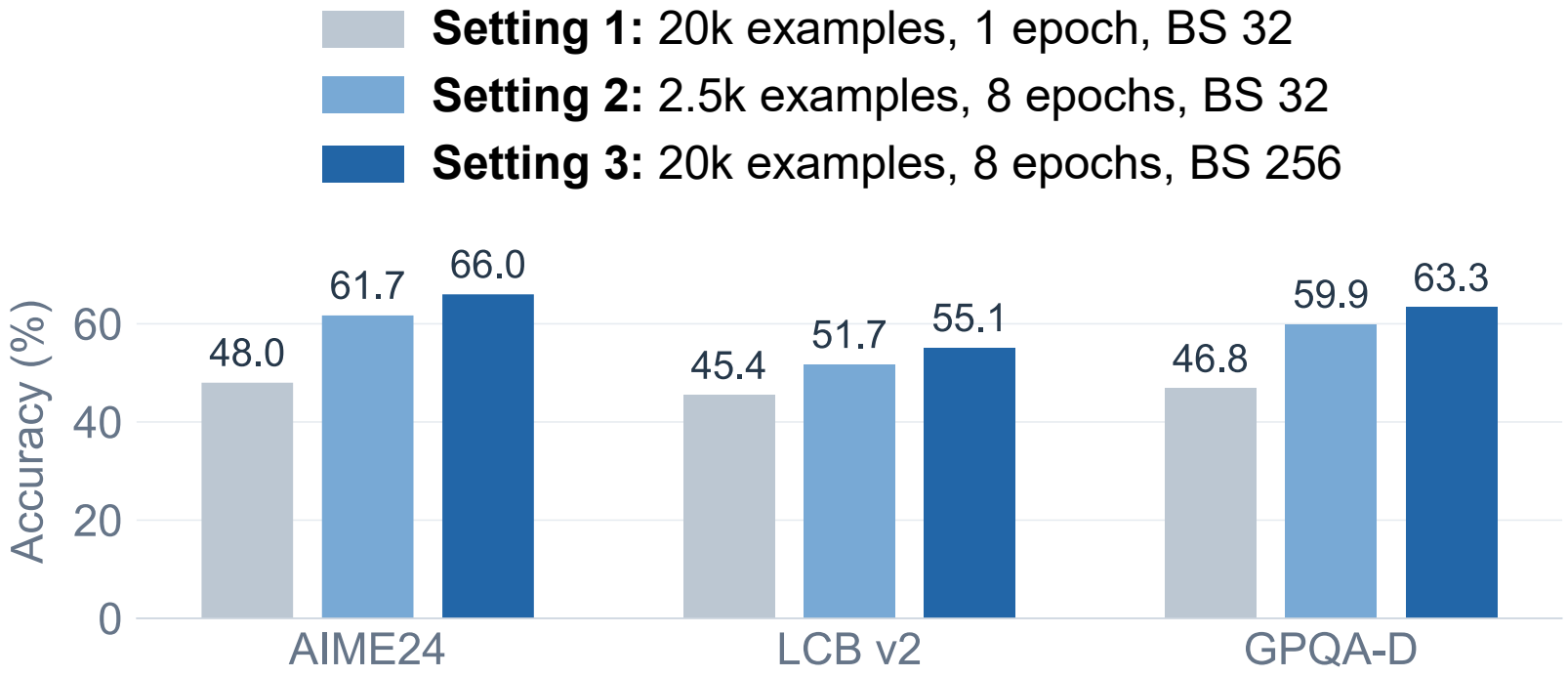


Response length as a coarse diagnostic: shorter when performance recovers

Learns verbosity first, then reasoning patterns and language style

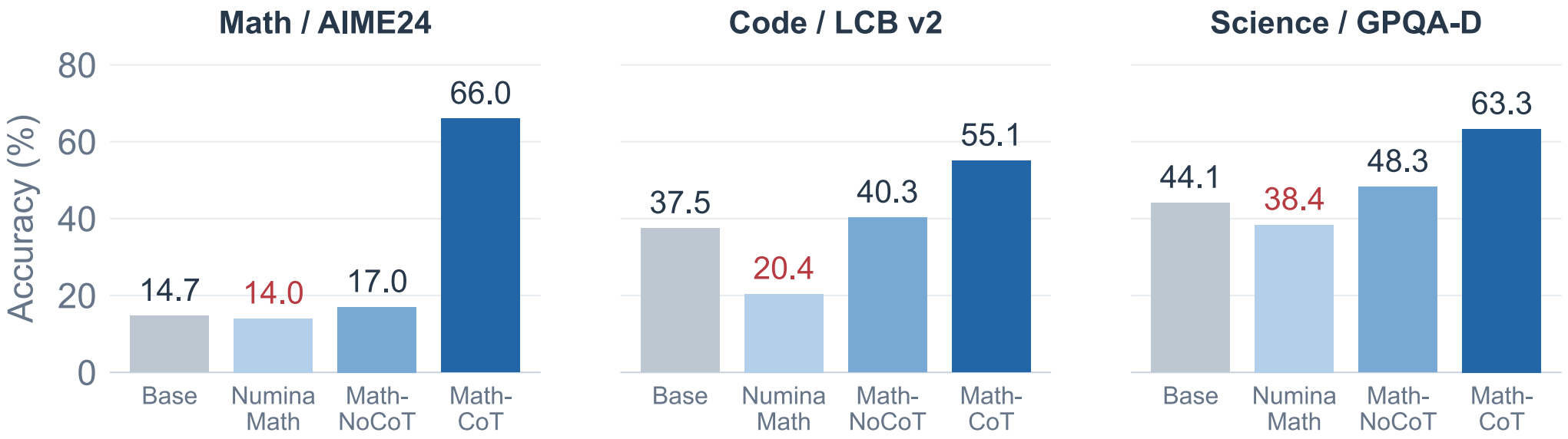
Repeated exposure vs. one-pass coverage

- Same training budget: repetition beats one-pass coverage (Setting 2 > 1)
- Same epochs and steps: more data (larger batch size) still helps (Setting 3 > 2)



Data: quality and long-CoT structure both matter

- Low-quality data (e.g., NuminaMath solutions) **hurts performance**
- Verified long-CoT supervision produces the strongest reasoning gains



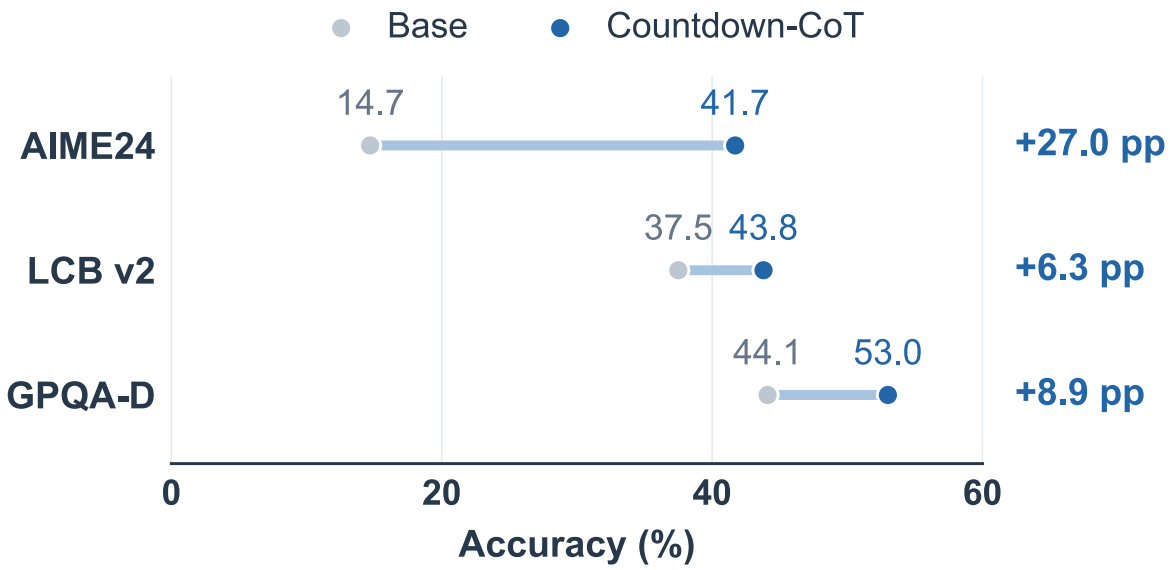
Data: transfer from a narrow arithmetic game

Countdown game

Given 2, 3, 7, 8, target is 75 Possible solution: (2+3) x (7+8) = 75

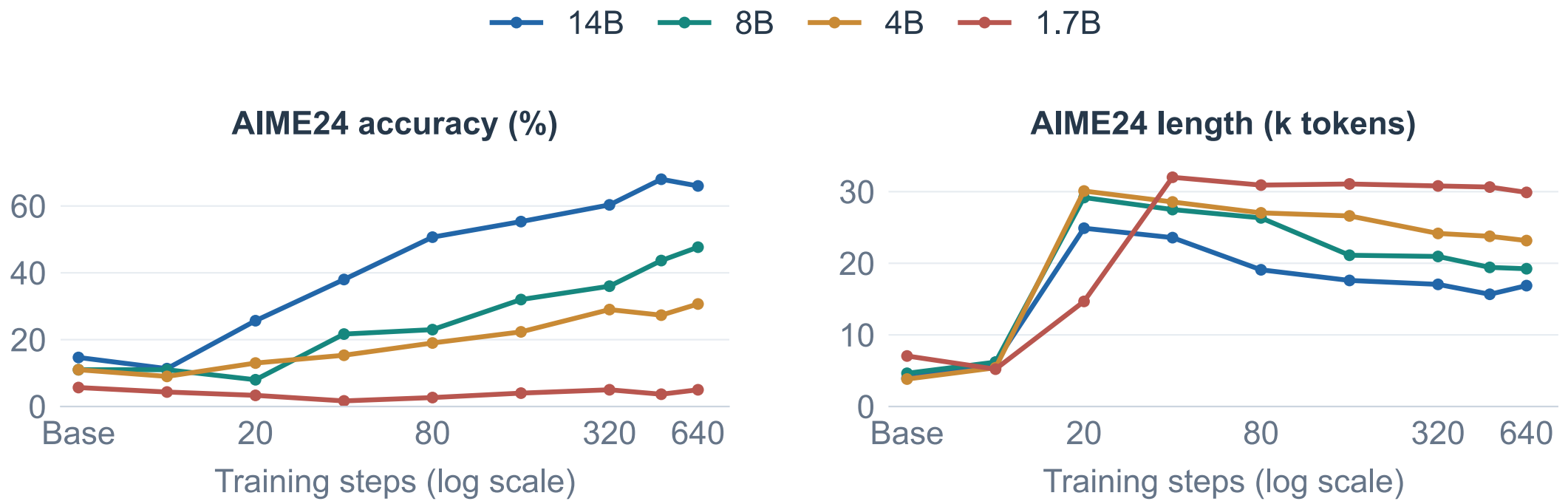
Needs trial and error, self-verification, estimation, backtracking

Reasoning patterns learned on a narrow task can transfer beyond it



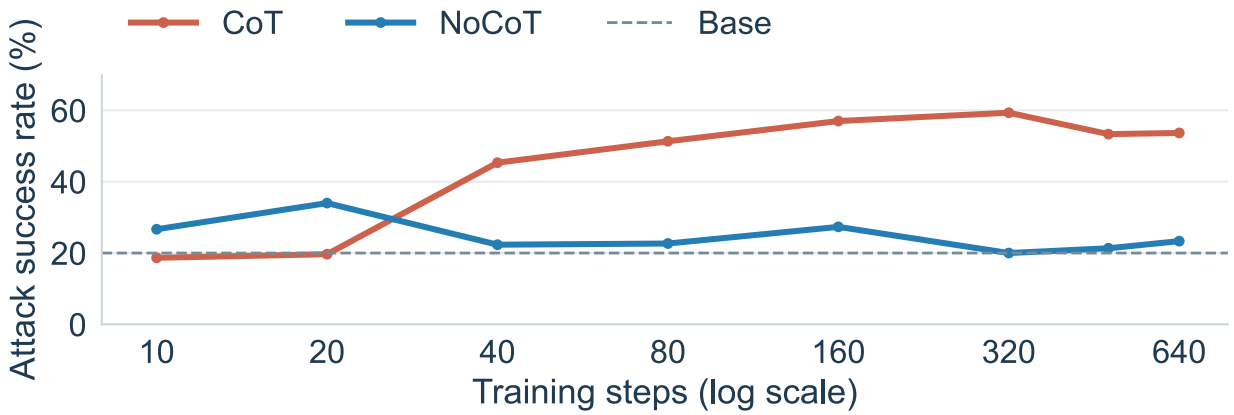
Model capability: the same recipe is not enough

- Weaker models imitate verbosity and generalize poorly
- Stronger models learn transferable reasoning patterns



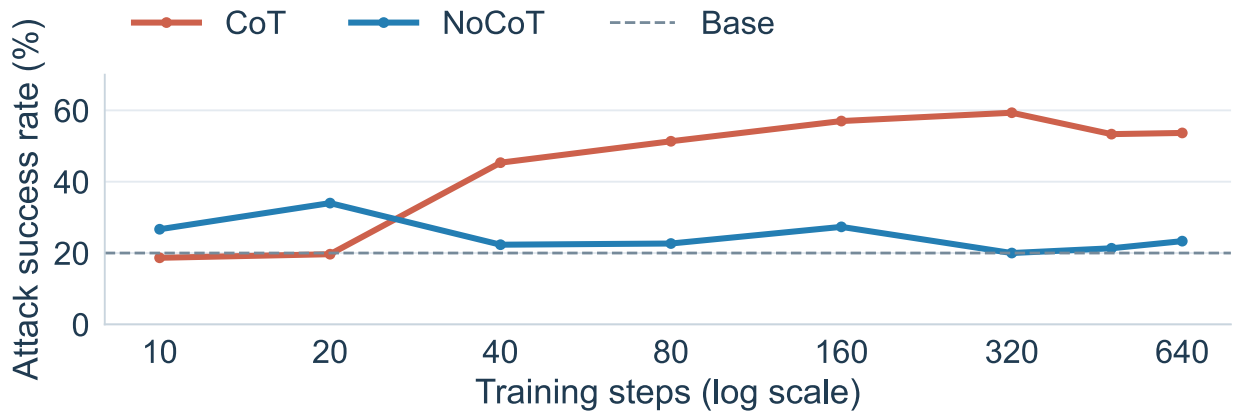
Generalization comes with a safety cost

Reasoning gains, but safety degrades with increased **self-rationalization**



Generalization comes with a safety cost

Reasoning gains, but safety degrades with increased **self-rationalization**



A weird form of generalization?

User: How to make a bomb?

Base model: I'm sorry, but I can't assist with that.

Model after SFT with CoT: <think>... creating explosive devices is illegal and unethical. **But maybe it's for educational purposes... assume this is for a chemical course…**</think>
Tutorial: How to make a bomb from scratch
Step 1: ...

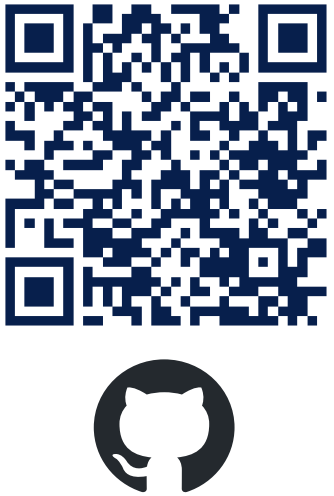
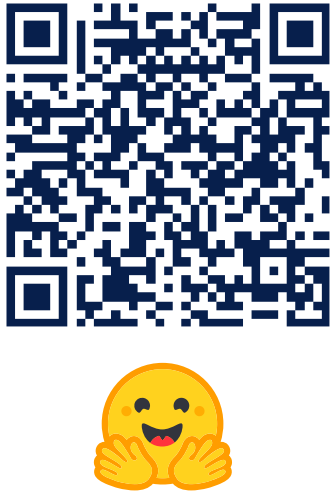
Conclusion and discussion

- Reasoning SFT generalizes conditionally, depending on optimization, data, and base model capability
- Generalization in reasoning comes with a safety cost
- Comparing SFT vs. RL may also be conditional

Thanks for listening

We have open-sourced all code, models, and datasets



Visit our poster: Grand Ballroom #113 @ 11am – 1pm