From Atomic to Composite: Reinforcement Learning Enables Generalization in Complementary Reasoning

Sitao Cheng, Xunjian Yin, Ruiwen Zhou, Yuxuan Li, Xinyi Wang, Liangming Pan, William Yang Wang, Victor Zhong

Question 1: RL as a skill synthesizer or probability amplifier

Question 2: Is there any prerequisites for RL to generalize
Or How to schedule SFT and RL

Knowledge-intensive Reasoning

Why not Math/Coding ?

1) Can't disentangle memory and reasoning
2) Narrow definition of generalization
3) Limited operations (+,-,*,/)

- QA w/ parametric knowledge → Parametric Reasoning (Mem)
- QA w/ contextual knowledge → Contextual Reasoning (Ctx)
- QA w/ parametric + contextual knowledge → Complementary Reasoning (Comp)

Behavioral Study with Synthetic Data

Why not existing benchmarks?

1) Data contamination
2) Unground knowledge for LLMs
3) Hard evaluation – knowledge sufficiency

- Synthetic human biographies
  - Control the parametric or contextual knowledge
- Multi-hop QA-pairs
  - Control the sufficiency of knowledge
- Control the generalization difficulty


```
"Allison Hill": {
    "name": "Allison Hill",
    "birth_date": "1942-04-29",
    "occupation": "Civil engineer, consulting",
    "email": "garzaanthony@example.org",
    "phone": "538.990.8386",
    "new": true,
    "died_on": "2024-11-01",
    "child": "Donald Marsh",
    "pet": "Whiskers",
    "wrote": "Baby administration",
    "influenced_by": "Matthew Cooper",
    "mentoring": "Daniel Watkins",
    "hobby": "painting",
    "classmate": "Adam Villanueva",
    "first_language": "Finnish",
    "roommate": "Shannon Krause",
    "university": "University of Chicago",
    "service": "Habitat for Humanity",
    "known_for": "painting",
    "died_in": "Brownbury",
    "boss": "Lindsey Johnson",
    "favorite_food": "tacos"
}
```


Allison Hill has a pet named Whiskers. Allison Hill spoke Finnish as their first language. A favorite activity of Allison Hill is painting. Lindsey Johnson is the boss of Allison Hill. Allison Hill was born on 1942-04-29. Allison Hill died on 2024-11-01. Allison Hill shared a room with Shannon Krause. Allison Hill penned Baby administration. Allison Hill was inspired by Matthew Cooper. The contact email for Allison Hill is garzaanthony@example.org. Allison Hill was famous for painting. Allison Hill was a member of Habitat for Humanity. Allison Hill mentors Daniel Watkins. Allison Hill's place of death was Brownbury. Allison Hill's phone number is 538.990.8386. Allison Hill works as a Civil engineer, consulting. Allison Hill is the parent of Donald Marsh. Allison Hill was a classmate of Adam Villanueva. Allison Hill loved eating tacos. Allison Hill went to University of Chicago.

Experiment Setups

- Different training strategies
  - SFT, RL
- Different training data
  - Mem+Ctx, Comp
- Different generalization levels
  - IID, Composition, Zero-shot
- Finding
  - ***RL generalizes only with sufficient atomic skills***

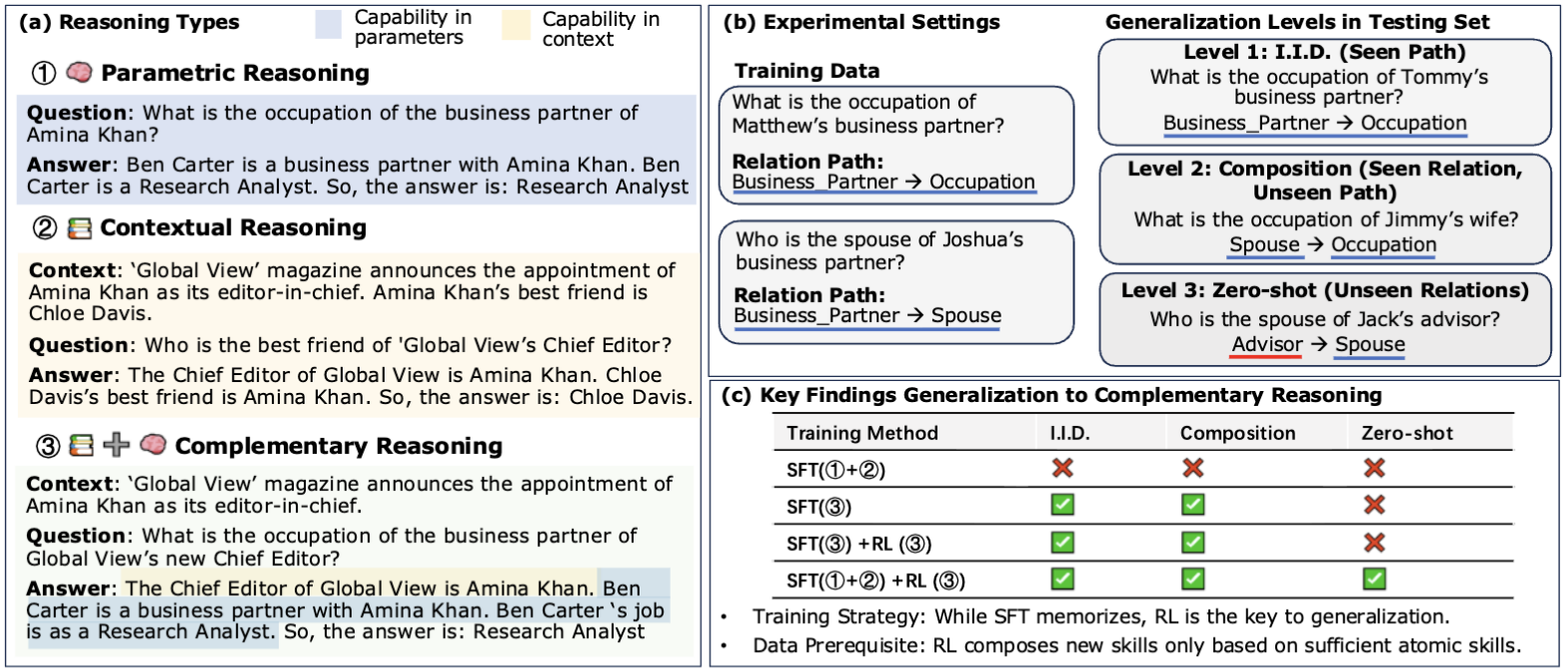

Figure 1. Our settings and findings. (a) Examples of Complementary Reasoning requiring both Parametric and Contextual skills. (b) Evaluation protocol across three levels of difficulty. _ and _ denotes seen and unseen pattern, respectively. (c) The SFT Generalization Paradox: Models trained on atomic skills generalize better via RL than models trained directly on the composite task.

UNIVERSITY OF WATERLOO

Sufficiency – RL as a synthesizer of atomic skills

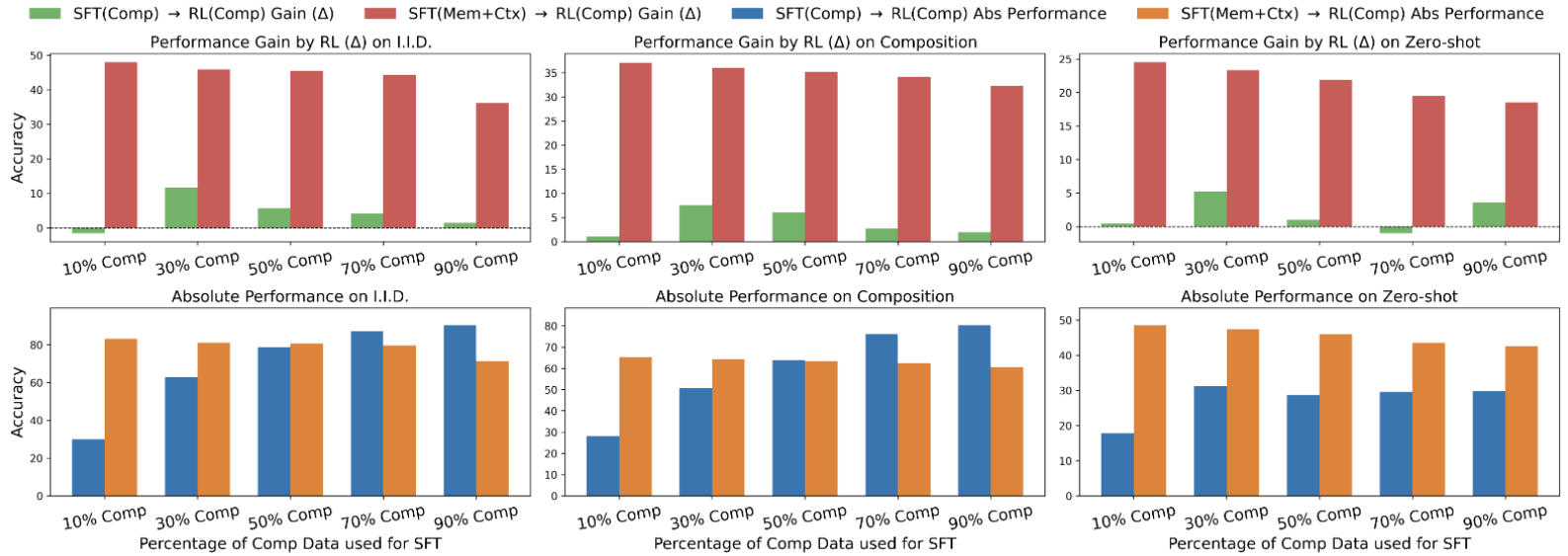

Figure 2. Comparison of training with different complementary data proportions. The top row compares the gain from RL, while the bottom row compares the absolute performance after RL. It shows that LLMs generalize to Complementary Reasoning only from the model with both Parametric and Contextual Reasoning skills.

- Varying portions of Comp data for SFT and RL Left → Right
  - Same amount of RL data for SFT(Mem+Ctx) and SFT(Comp)
- RL efficiently synthesizes from atomic skills
- SFT Memorization Paradox: Distinguishing Generalization from Memorization

Necessity – Sufficient Atomic Skills are Necessary

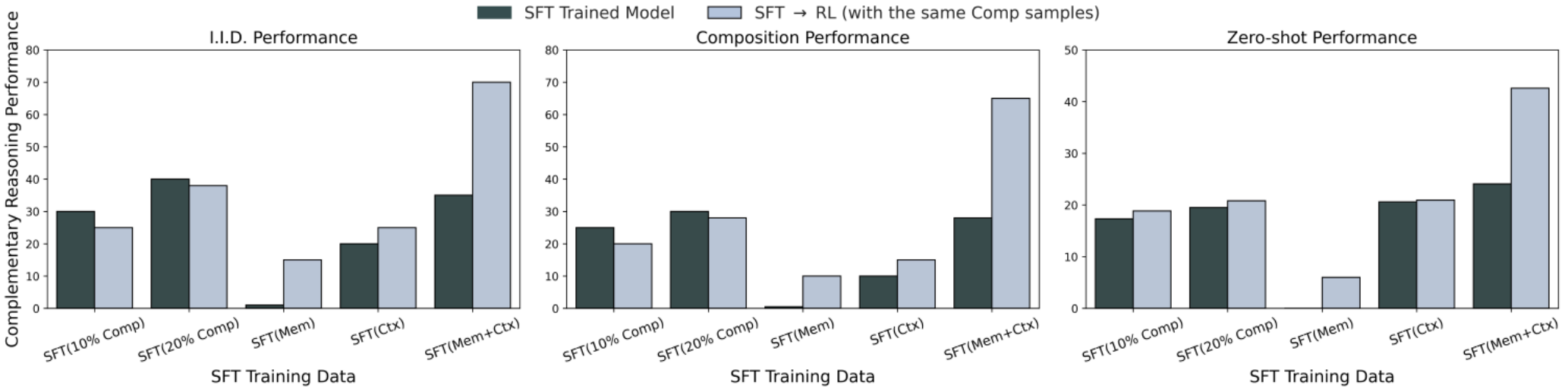

Figure 3: Necessity of atomic skills for RL generalization. We conduct RL with the same amount of COMP data from different SFT trained models. Only $\text{SFT}_{\text{MEM+CTX}}$ generalizes well in all levels.

- Removing any atomic skill collapses generalization
- RL-driven generalization does not directly related to initial model performance

Necessity – RL is Necessary for Generalization

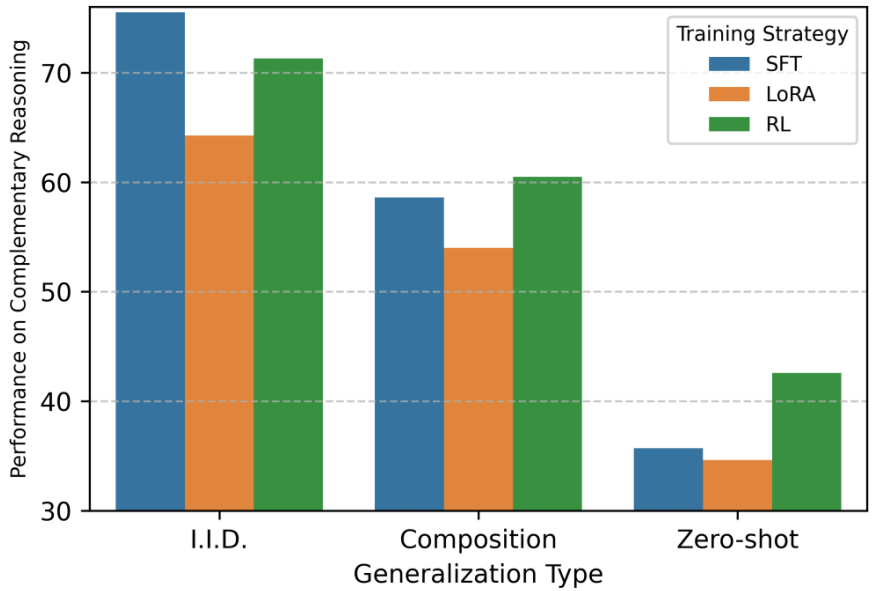

Figure 4: Performance of training with different strategies over 12.8k COMP samples.

- All training strategies are good in IID and Composition
- SFT memorizes (IID), RL generalizes (Zero-shot)

Sample Efficiency – the data before RL to prime generalization

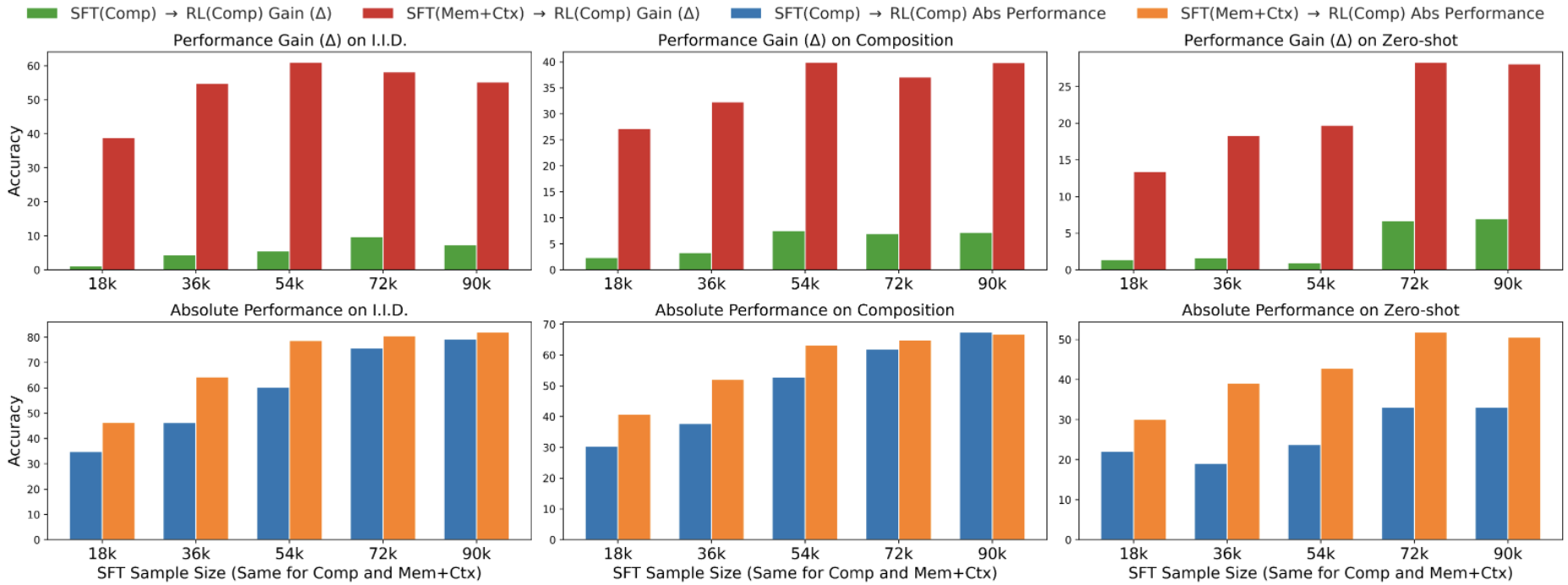

Figure 5. Performance with the same amount of SFT and RL data, comparing $\text{SFT}_{\text{MEM+CTX}}$ and SFT_{COMP}.

- Train with same amount of data and number of hops for SFT and RL
- Atomic capability learning requires less SFT data to prime RL generalization

UNIVERSITY OF WATERLOO

Sample Efficiency – the data for RL-driven generalization

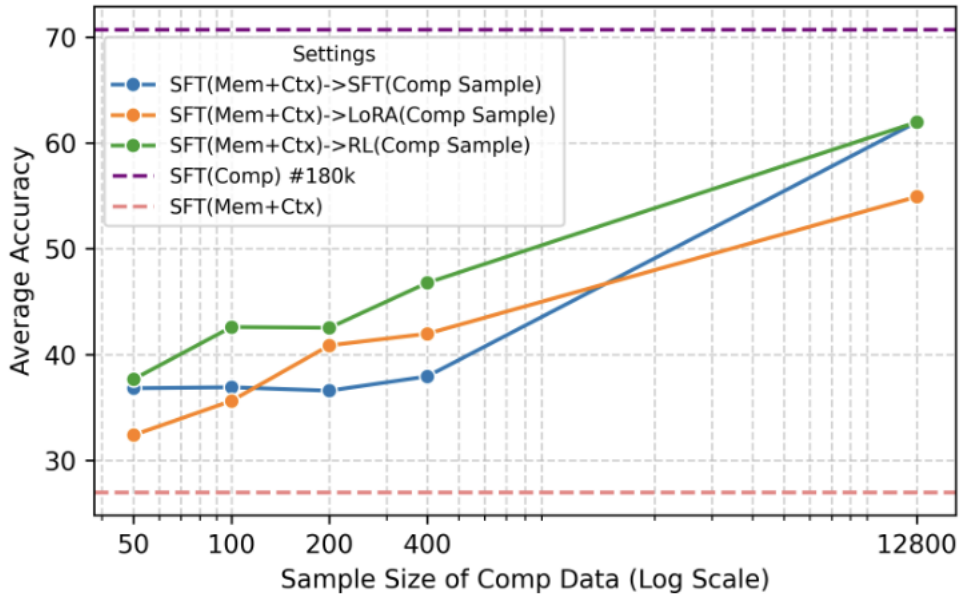

Figure 6. Few-shot adaptation of $\text{SFT}_{\text{MEM+CTX}}$. We show average accuracy over all generalization levels.

- Few-shot training with various training strategies based on SFT(Mem+Ctx)
- Sufficient atomic skills enables few-shot adaptation

Pass@K – RL as a synthesizer or amplifier

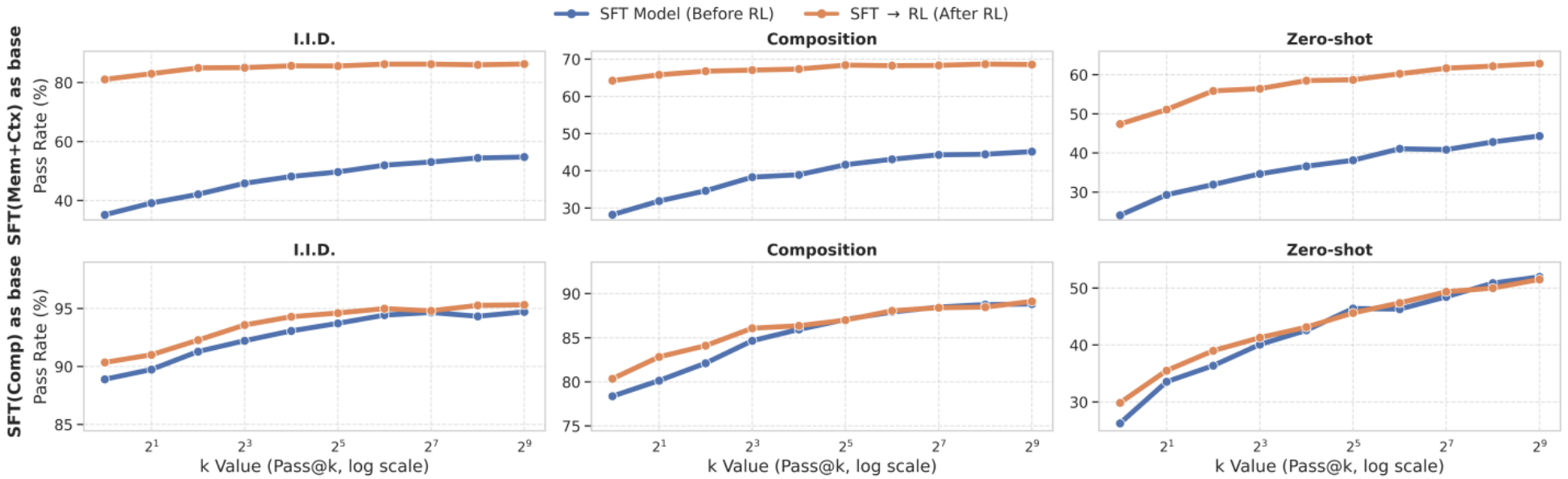

Figure 7: Pass@k comparison for $\text{SFT}_{\text{MEM+CTX}}$ and SFT_{COMP}. It shows that RL synthesizes new compositional skills only based on models with sufficient atomic skills.

- RL synthesizes new pathways for models *with sufficient atomic skills*
- RL only amplifies existing behaviors for models *only with composite skills*

Thanks!