TRAINING / CORECTION LLM *and* RL

by **Adhilsha Ansad**

THE PAPER

Google DeepMind

2024-10-7

Training Language Models to Self-Correct via Reinforcement Learning

Aviral Kumar[*+,1], Vincent Zhuang[*+,1], Rishabh Agarwal[*,1], Yi Su[*,1], JD Co-Reyes[1], Avi Singh[1], Kate Baumli[1], Shariq Iqbal[1], Colton Bishop[1], Rebecca Roelofs[1], Lei M Zhang[1], Kay McKinney[1], Disha Shrivastava[1], Cosmin Paduraru[1], George Tucker[1], Doina Precup[1], Feryal Behbahani[†,1] and Aleksandra Faust[†,1]

[1]Google DeepMind, [*]Equal Contribution, [+]Randomly ordered via coin flip, [†]Jointly supervised.

CONTENT

CONTEXT

- LLMs for reasoning
  - math, logic and coding
- need to “*self-correct*” applicable to test-data
- limitations in “***intrinsic self-correction***”
  - lack of prompt - limiting use of underlying knowledge
  - distribution shift
  - behavior collapse
- Prior attempts:
  - Prompt engineering (less meaningful)
  - Fine-tuning (multi-model inf., refinement models, Teacher-model)

CONTRIBUTIONS

- proposes ***SCoRe*** *(Self-Correction via Reinforcement Learning)*
- single model - response and correction
- No oracle feedback
- on-policy, multi-turn RL
- rewards “progress” towards self-correction (vs correctness)

METHODOLOGY

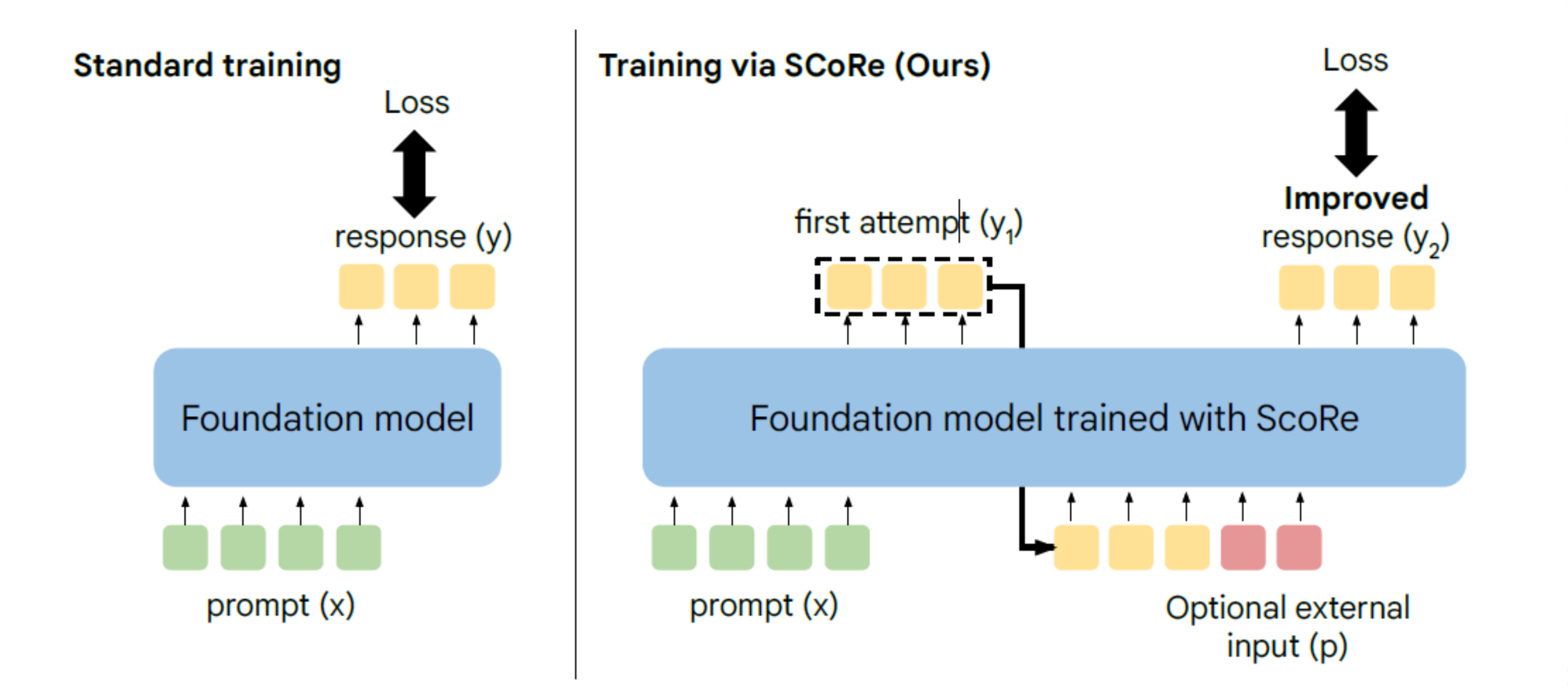
- Objective

Concretely, given a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i^*)\}_{i=1}^N$ of problems $\boldsymbol{x}_i$ and responses $\boldsymbol{y}_i^*$, we will train an LLM policy $\pi_\theta(\cdot|[\boldsymbol{x}, \hat{\boldsymbol{y}}_{1:l}, \boldsymbol{p}_{1:l}])$ that, given the problem $\boldsymbol{x}$, previous l model attempts $\hat{\boldsymbol{y}}_{1:l}$ at the problem, and auxiliary instructions $\boldsymbol{p}_{1:l}$ (e.g., instruction to find a mistake and improve the response), solves the problem $\boldsymbol{x}$ as correctly as possible.

Self-correction instruction. There might be an error in the solution above because of lack of understanding of the question. Please correct the error, if any, and rewrite the solution.

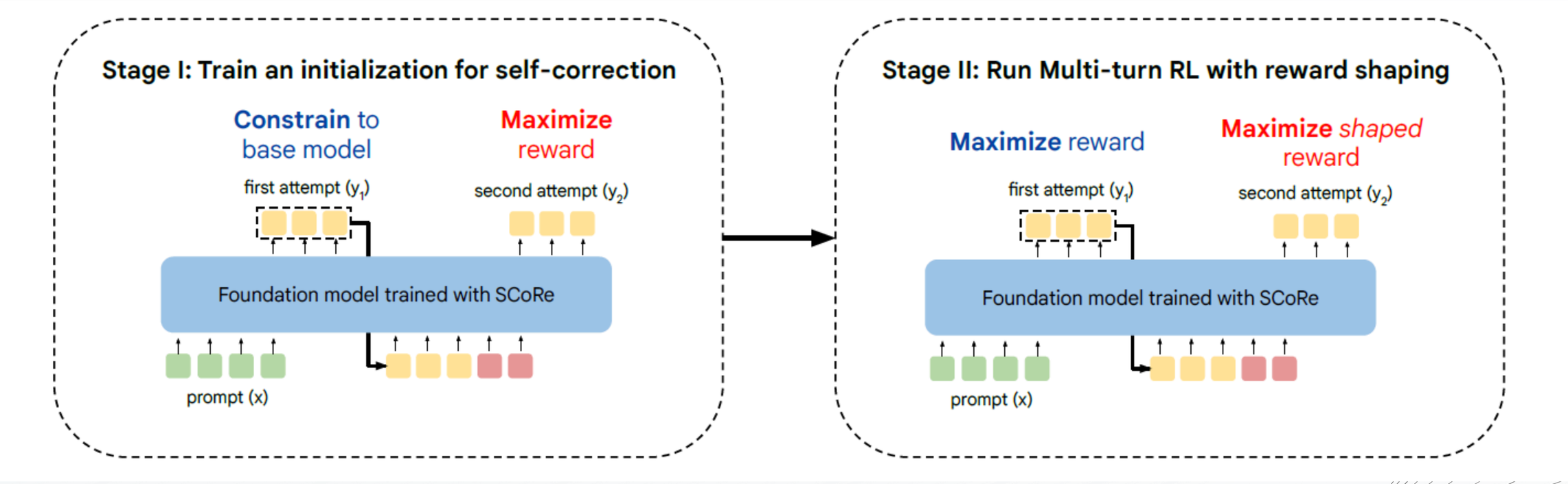
- Oracle reward disabled at test-time
- No majority voting

METHODOLOGY



METHODOLOGY

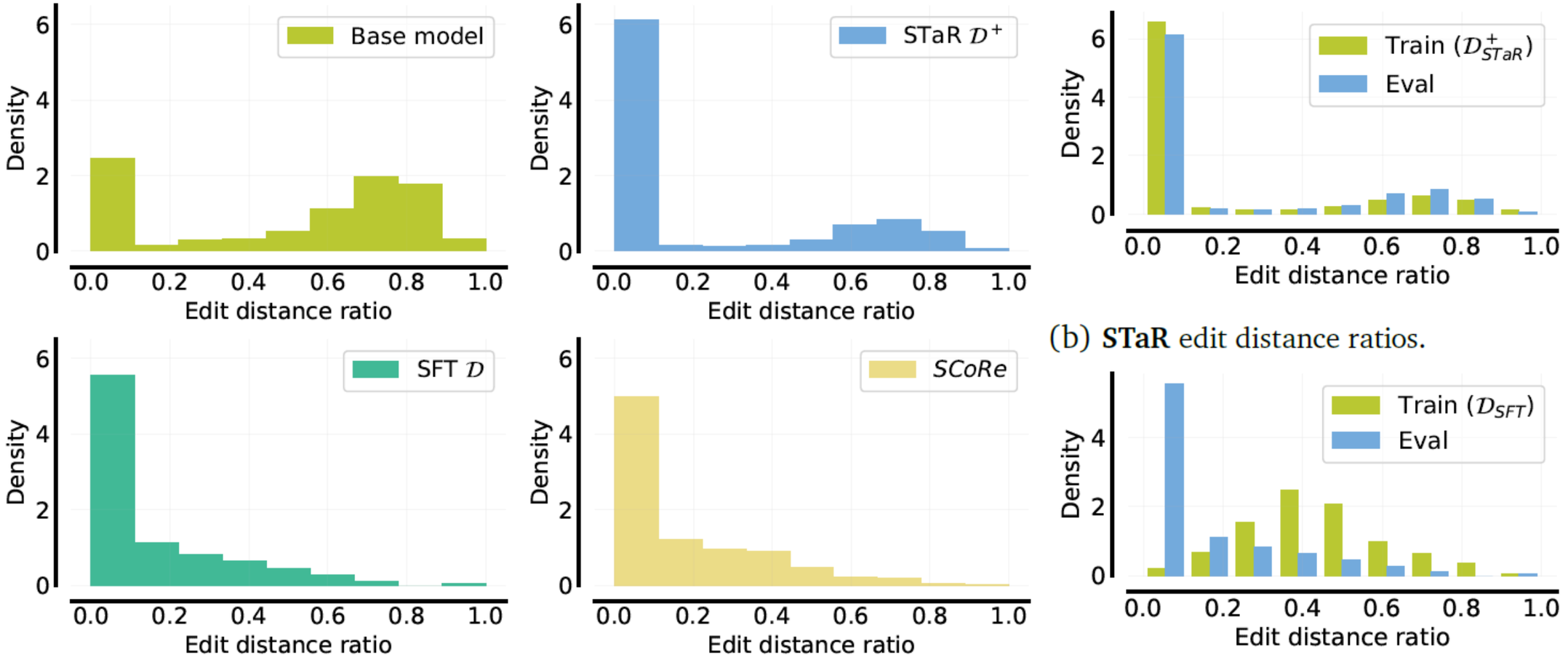
- REINFORCE policy gradient training
- KL-divergence penalty (primarily used in single-turn RLHF)



EXPERIMENTS

- **Datasets**:
  - MATH
  - HumanEval
  - MBPP, MBPP-R
- **model**:
  - FT Gemini 1.5 Flash (MATH)
  - FT Gemini 1.0 Pro

EXPERIMENTS



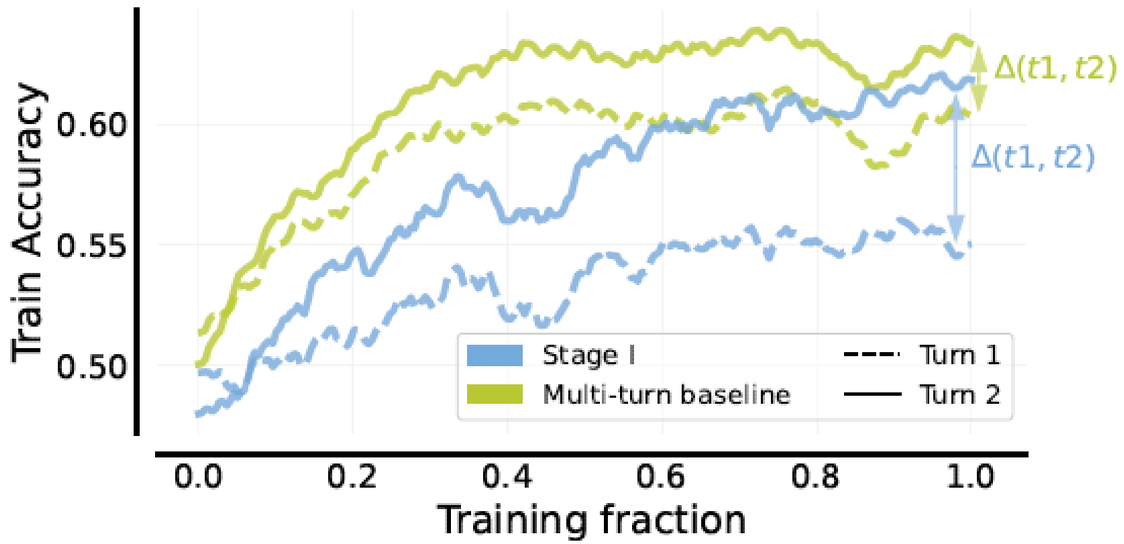
(b) **STaR** edit distance ratios.

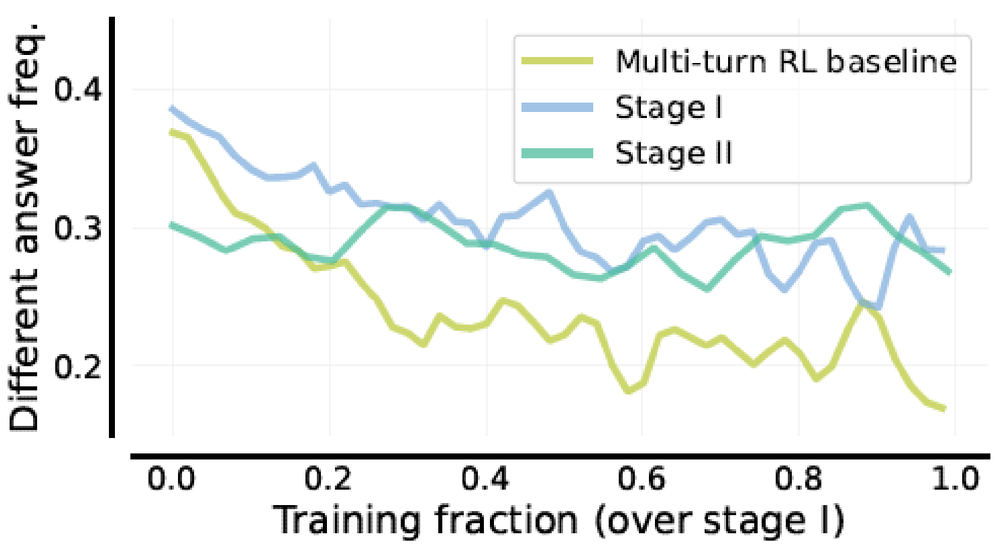
Edit distance between first-attempt and second-attempt responses

EXPERIMENTS

Method	Accuracy@t1	Accuracy@t2	Δ(t1, t2)	$\Delta^{i \to c}$(t1, t2)	$\Delta^{c \to i}$(t1, t2)
Base model	52.6%	41.4%	-11.2%	4.6%	15.8%
STaR $\mathcal{D}_{\text{StaR}}$	55.4%	41.2%	-14.2%	5.4%	19.6%
STaR $\mathcal{D}^{+}_{\text{StaR}}$	53.6%	54.0%	0.4%	2.6%	2.2%
Pair-SFT $\mathcal{D}_{\text{SFT}}$	52.4%	54.2%	1.8%	5.4%	3.6%
Pair-SFT $\mathcal{D}^{+}_{\text{SFT}}$	55.0%	55.0%	0%	0%	0%

EXPERIMENTS





RESULTS

Approach	Acc.@t1	Acc.@t2	Δ(t1, t2)	$\Delta^{i\to c}$(t1, t2)	$\Delta^{c\to i}$(t1, t2)
Base model	52.6%	41.4%	-11.2%	4.6%	15.8%
Self-Refine (Madaan et al., 2023)	52.8%	51.8%	-1.0%	3.2%	4.2%
STaR w/ $\mathcal{D}^{+}_{\text{StaR}}$ (Zelikman et al., 2022)	53.6%	54.0%	0.4%	2.6%	2.2%
Pair-SFT w/ $\mathcal{D}_{\text{SFT}}$ (Welleck et al., 2023)	52.4%	54.2%	1.8%	5.4%	3.6%
SCoRe (Ours)	**60.0%**	**64.4%**	**4.4%**	**5.8%**	**1.4%**

Performance of SCoRe on MATH

RESULTS

Method	MBPP-R	Acc.@t1	Acc.@t2	Δ(t1, t2)	$\Delta^{i \to c}$(t1, t2)	$\Delta^{c \to i}$(t1, t2)
Base model	47.3%	53.7%	56.7%	3.0%	7.9%	4.9%
Self-Refine	30.7%	53.7%	52.5%	-1.2%	9.8%	11.0%
Pair-SFT	59.8%	56.1%	54.3%	-1.8%	4.3%	6.1%
SCoRe (Ours)	**60.6%**	52.4%	**64.6%**	**12.2%**	**15.2%**	**3.0%**

Performance of SCoRe on HumanEval

RESULTS

Method	Accuracy@t1	Accuracy@t2	Δ(t1, t2)
SCoRe (Ours)	60.0%	**64.4%**	**4.4%**
w/o multi-turn training	**61.8%**	59.4%	-2.4%
w/o Stage I	59.2%	61.4%	2.2%
w/o reward shaping	60.0%	62.6%	2.6%
w/ STaR instead of REINFORCE Stage II	56.2%	58.4%	2.2%

Ablation studies

OVERVIEW

- **Limitation**:
  - Single round of self-correction
- **Possible works**:
  - Unifying Stages I and II
  - Effective multi-turn RL self-correction
- **Conclusion**:
  - Self-correction without behavioral collapse
  - learning self-correction needs *progress reward*

THANK YOU