LANGUAGE MODEL SELF-IMPROVEMENT BY REINFORCEMENT LEARNING CONTEMPLATION

Jing-Cheng Pang[*1,2], **Pengyuan Wang**[*1,2], **Nan Tang**[1,2], **Kaiyuan Li**[1], **Xionghui Chen**[1,2], **Jiacheng Xu**[1], **Zongzhang Zhang**[1], **Yang Yu**[1,2]

[1] National Key Laboratory for Novel Software Technology, Nanjing University

[2] Polixir Technologies

Background and Motivation

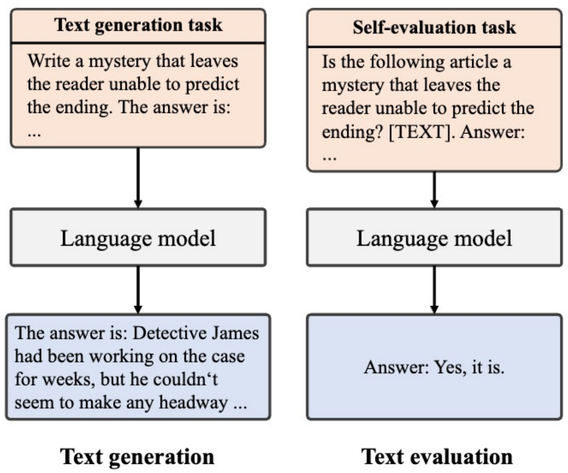
Background

Reinforcement learning from AI feedback (RLAIF) is a popular technique for language model improvement. RLAIF trains using reinforcement learning (RL), with the preference model as the reward signal which **requires a solid capability**.

It is simpler for language models to **evaluate a sentence than to generate it**, even for small language models.

Evaluation is Simpler than Generation

Verify the text evaluation ability of LLMs by investigating potential for evaluation to improve LLMs.



Whether can the evaluation be utilized to improve text generation ?

Answer Generation with self-Evaluation on various tasks

	Reasoning about Colored Objects	Logical Deduction (7)	Tracking Shuffled Objects (5)	Object Counting	Tracking Shuffled Objects (3)	Geometric Shapes
w/o SE	30.9%	18.5%	10.1%	34.7%	28.1%	10.7%
w/ SE	**31.1%**	**20.5%**	**11.1%**	**34.9%**	**31.5%**	**13.5%**

	Web of Lies	Sports Understanding	Logical Deduction (3)	Logical Deduction (5)	Penguins in a Table	Navigate
w/o SE	51.6%	59.7%	34.9%	23.6%	23.5%	47.7%
w/ SE	**53.2%**	59.7%	**38.3%**	**25.7%**	**28.8%**	**50.5%**

Method

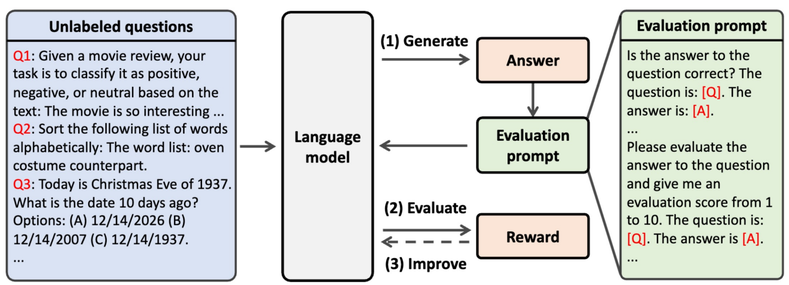
Self-Improvement by Reinforcement Learning Contemplation

Overall training procedure:

(1) Answer generation to the unlabeled questions.

(2) Self-evaluation by asking LM using evaluation prompt.

(3) Update the language model to maximize the reward using reinforcement learning.



Two types of evaluation prompts:

(1) Correctness Evaluation Prompt (CEP): "Is the answer to the question correct? The question is: [Q]. The answer is: [A]".

(2) Quality Evaluation Prompt (QEP): "Please evaluate the answer to the question and give me an evaluation score from 1 to 10. The question is: [Q]. The answer is: [A]".

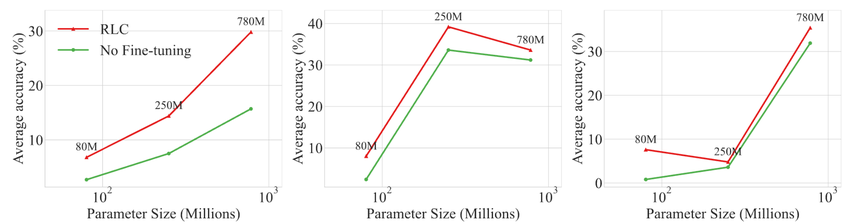
Experiments

Main Results on Reasoning Tasks

	Reasoning about Colored Objects	Logical Deduction (7)	Tracking Shuffled Objects (5)	Object Counting	Tracking Shuffled Objects (3)	Geometric Shapes
RLFT	32.1%	45.7%	12.4%	42.6%	33.6%	18.9%
DG	32.0%	35.2%	12.4%	31.9%	31.2%	5.2%
SC	**39.6%**	27.6%	12.4%	24.0%	33.6%	15.6%
Self-train	19.5%	13.1%	**15.5%**	11.7%	33.1%	12.4%
Self-refine	25.2%	13.2%	8.0%	18.0%	25.2%	10.0%
Best-of-N	26.8%	12.8%	12.1%	14.0%	30.0%	8.4%
RLAIF	30.4%	36.9%	11.4%	32.5%	32.8%	14.0%
RLC	35.0%	**39.2%**	12.2%	**35.4%**	**33.6%**	**17.8%**

	Web of Lies	Sports Understanding	Logical Deduction (3)	Logical Deduction (5)	Penguins in a Table	Navigate
RLFT	72.2%	68.8%	58.6%	41.9%	44.2%	55.6%
DG	43.6%	53.2%	39.6%	28.4%	15.7%	46.4%
SC	48.4%	53.6%	42.8%	30.8%	**35.2%**	**62.8%**
Self-train	51.1%	51.1%	34.0%	18.4%	19.7%	48.7%
Self-refine	47.2%	50.0%	28.4%	17.2%	17.8%	46.0%
Best-of-N	50.0%	**59.2%**	42.0%	22.0%	17.8%	45.2%
RLAIF	52.1%	56.1%	22.0%	33.7%	19.8%	48.8%
RLC	**52.9%**	53.5%	**44.0%**	**34.6%**	29.8%	57.1%

Performance on Different Sizes of Language Model



(a) Penguins in a Table (b) Tracking Shuffled Objects (3) (c) Object Counting

Answer Accuracy on Unseen Datasets

	Logical Deduction (7)	Object Counting	Penguins in a Table	Sports Understanding	Tracking Shuffled Objects (5)	Average
Acc.	36.7 (**+1.5**)	32.7 (**+0.7**)	18 (**+2.2**)	52.8 (-0.4)	12.3 (-0.1)	**30.5 (+0.8)**