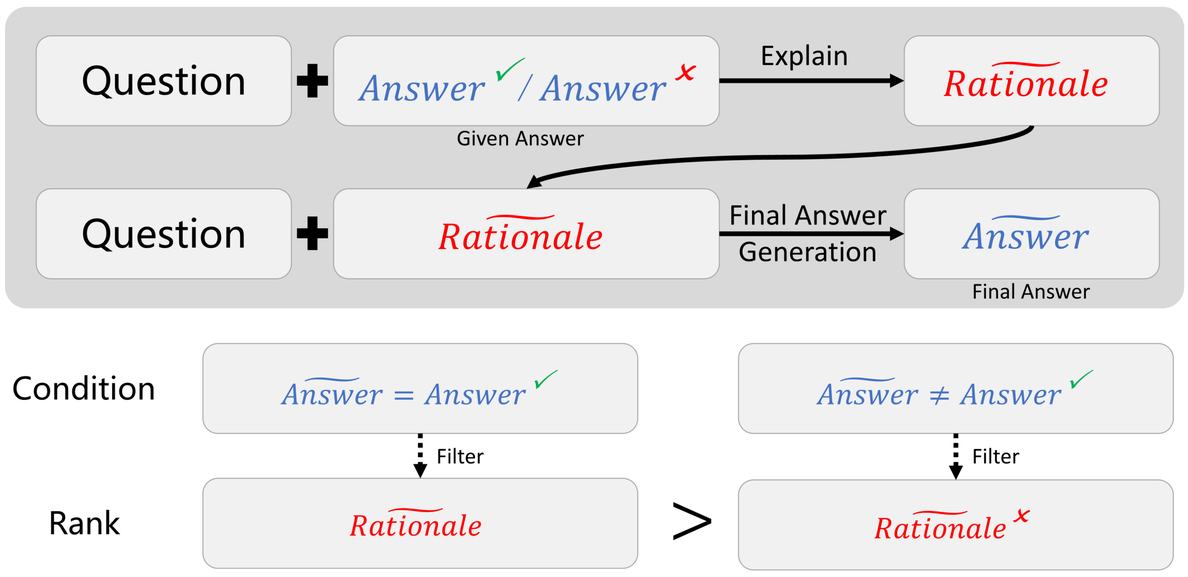
Improving Language Model Reasoning with Self-motivated Learning

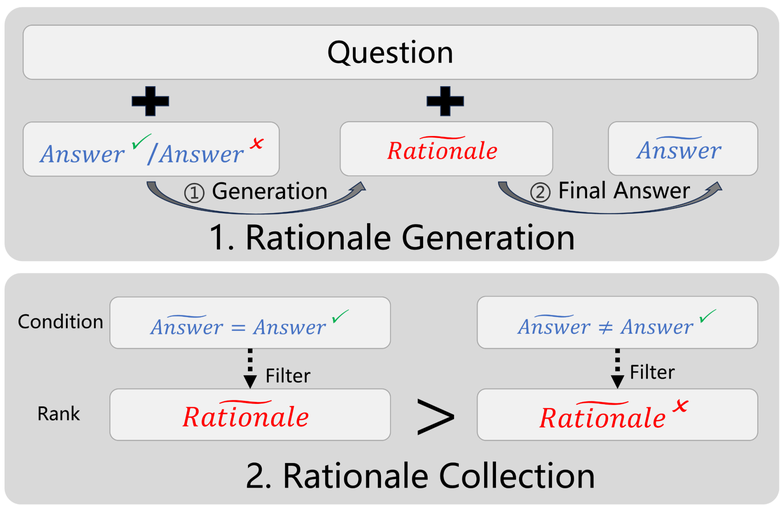
Yunlong Feng, Yang Xu, Libo Qin, Yasheng Wang, Wanxiang Che
Harbin Institute of Technology, China
Huawei Noah's Ark Lab

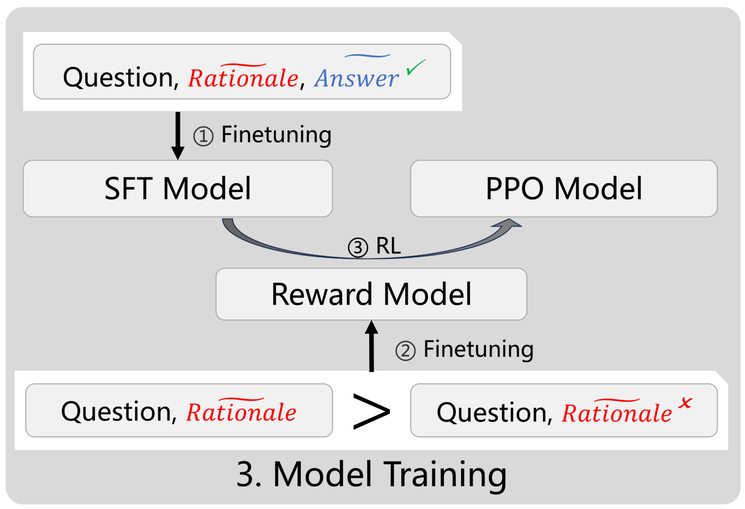
Motivation & Background

- Models gain reasoning capability
  - After trained with data that has rationales
- Lacks of datasets with high-quality rationales
  - high annotation cost
- How to use existing datasets without rationales?
  - Utilize the correctness
  - Correct is better than Wrong



Method





Dataset	Choices	Training Samples	Test Samples	Data Split	License	References
SingleEq	-	356	152	70:30	None	Koncel-Kedziorski et al. (2015a)
AddSub	-	276	119	70:30	Unspecified	Hosseini et al. (2014)
MultiArith	-	420	180	70:30	Unspecified	Roy and Roth (2016)
SVAMP	-	700	300	70:30	MIT	Patel et al. (2021a)
GSM8K	-	7473	1319	Original	MIT	Cobbe et al. (2021)
Date Understanding	5–6	258	111	70:30	Apache-2.0	Srivastava et al. (2022)
CommonSenseQA	5	9741	1221	Original	Unspecified	Talmor et al. (2018)
StrategyQA	2	1603	687	70:30	Apache2.0	Geva et al. (2021a)

Rationale Generation	Final Answer Generation
[Instruction and Question] <q_i> [Answer] <a'> [Rationale] <$\hat{r}_i$>	[Instruction and Question] <q_i> [Rationale] <$\hat{r}_i$> [Answer] <$\hat{a}_i$>

Prompt For Better Rationale	Prompt For Worse Rationale
[Instruction and Question] Would a pear sink in water? [Answer] No [Rationale] *The density of a pear is about 0.6g/cm3, which is less than water. Objects less dense than water float. Thus, a pear would float.*	[Instruction and Question] Would a pear sink in water? [Answer] Yes [Rationale] *The density of a pear is about 0.6g/cm3, which is less than water. Objects less dense than water float. Thus, a pear would sink.*

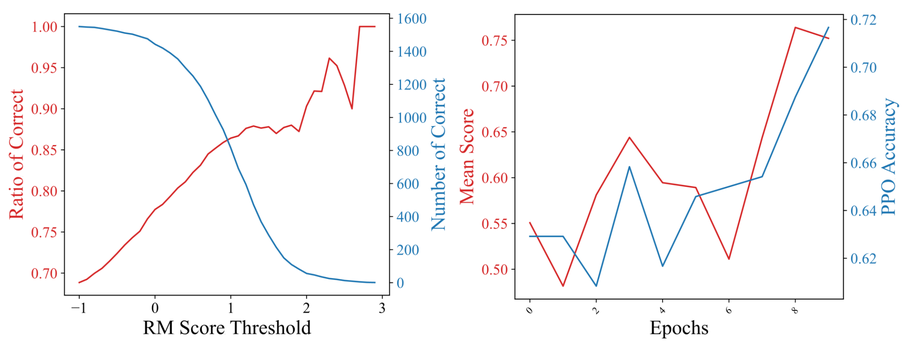
Experiments

Method	Param	Single Eq	Add Sub	Multi Arith	SVAMP	GSM8K	Date Understanding	Common SenseQA	Strategy QA
Close-Source Models									
text-davinci-003	175B	86.4	81.3	83.7	73.6	59.5	77.0	70.0	61.1
text-davinci-002	175B	82.24	78.99	78.89	64.67	40.26	73.87	61.75	53.57
Open-Source Models									
StableVicuna	13B	62.50	57.14	43.33	46.67	40.26	45.95	58.64	41.34
LLama2-Chat	7B	73.03	68.91	67.22	53.67	28.35	35.14	56.67	38.14
Methods on Llama2 7B									
Few-shot-CoT	7B	63.82	54.62	35.00	39.00	14.60	53.15	50.61	61.28
Few-shot-CoT$^{SC=8}$	7B	67.76	67.23	55.56	44.67	15.09	35.13	48.40	62.45
Fine-tune	7B	71.05	63.87	11.67	45.67	12.58	64.87	76.58	65.21
Fine-tune-CoT (text-davinci-002)	7B	70.39	72.27	76.67	47.33	–	73.88	–	58.95
Fine-tune-CoT (STaR)	7B	75.66	67.23	72.78	44.33	17.29	81.98	63.63	64.63
Fine-tune-CoT (Llama2)	7B	71.05	65.55	53.33	40.67	13.72	83.78	69.53	60.84
Self-motivated Learning	7B	**76.32**	**76.47**	**80.00**	**55.33**	**18.88**	**87.39**	**77.97**	**66.08**

- Accuracy (%) in 8 tasks under our different models and methods. Note that the methods based on the LLama2 7B are trained in different datasets separately.
- The Self-motivated Learning applies PPO for reinforcement learning in the training dataset using the Fine-tune-CoT model, resulting in an average increase of 10.68%.

Method	SingleEq	AddSub	MultiArith
Fine-tune-CoT (text-davinci-002)	70.39	72.27	76.67
+RL	71.71	78.16	80.56
Increase	+1.32	+5.89	+3.89

- The reward model trained with rank information exhibits a certain degree of generalization.



- The reward of model can reflect the performance.
- The performance is increasing during training.

Conclusion

• We propose "Self-motivated Learning", a framework is grounded in the idea that a rationale leading to the correct answer is superior to one leading to an incorrect answer.
• We conducted experiments across different datasets encompassing three categories of complex reasoning, demonstrating that our method can significantly enhance model performance without external annotation.

Acknowledgement

We gratefully acknowledge the support of the National Natural Science Foundation of China (NSFC)via grant 62236004 and 62206078.

LREC-COLING 2024