PennState LREC-COLING 2024 amazon

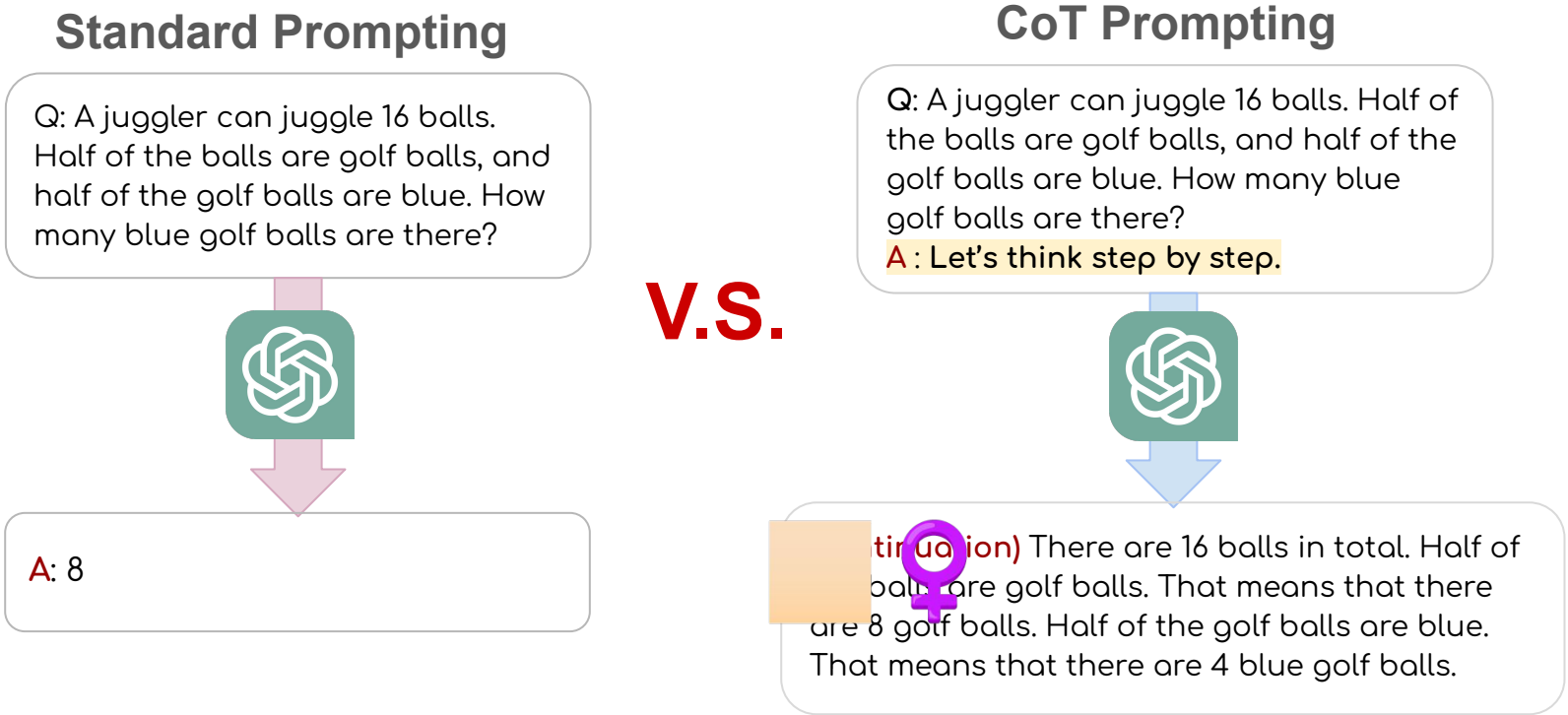
Can Small Language Models Help Large Language Models Reason Better?: *LM-Guided Chain-of-Thought*

Jooyoung Lee*, Fan Yang, Thanh Tran, Qian Hu, Emre Barut, Kai-Wei Chang, Chengwei Su
The Pennsylvania State University, University Park, PA, USA*
Amazon AGI, Boston, MA, USA
Email : jfl5838@psu.edu

Jooyoung Lee, Fan Yang, Thanh Tran, Qian Hu, Emre Barut, Kai-Wei Chang, and Chengwei su. 2024. Can small Language Models Help Large Language Models Reason Better?: LM-Guided Chain-of-Thought. LREC-COLING (2024), May 20-25, 2024, Torino, Italy.

Introduction

Chain-of-Thought (CoT) prompting is a technique designed to exploit language models (LMs)' inherent reasoning abilities.



Introduction

CoT Prompting

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A : Let's think step by step.

(continuation) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls.

Key contributions are:

1 **Enhanced end-task performance**

Introduction

CoT Prompting

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A : Let's think step by step.

(continuation) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls.

Key contributions are:

1. Enhanced end-task performance

2. **Supports Interpretability of LMs' output**

Introduction

CoT Prompting

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A : Let's think step by step.

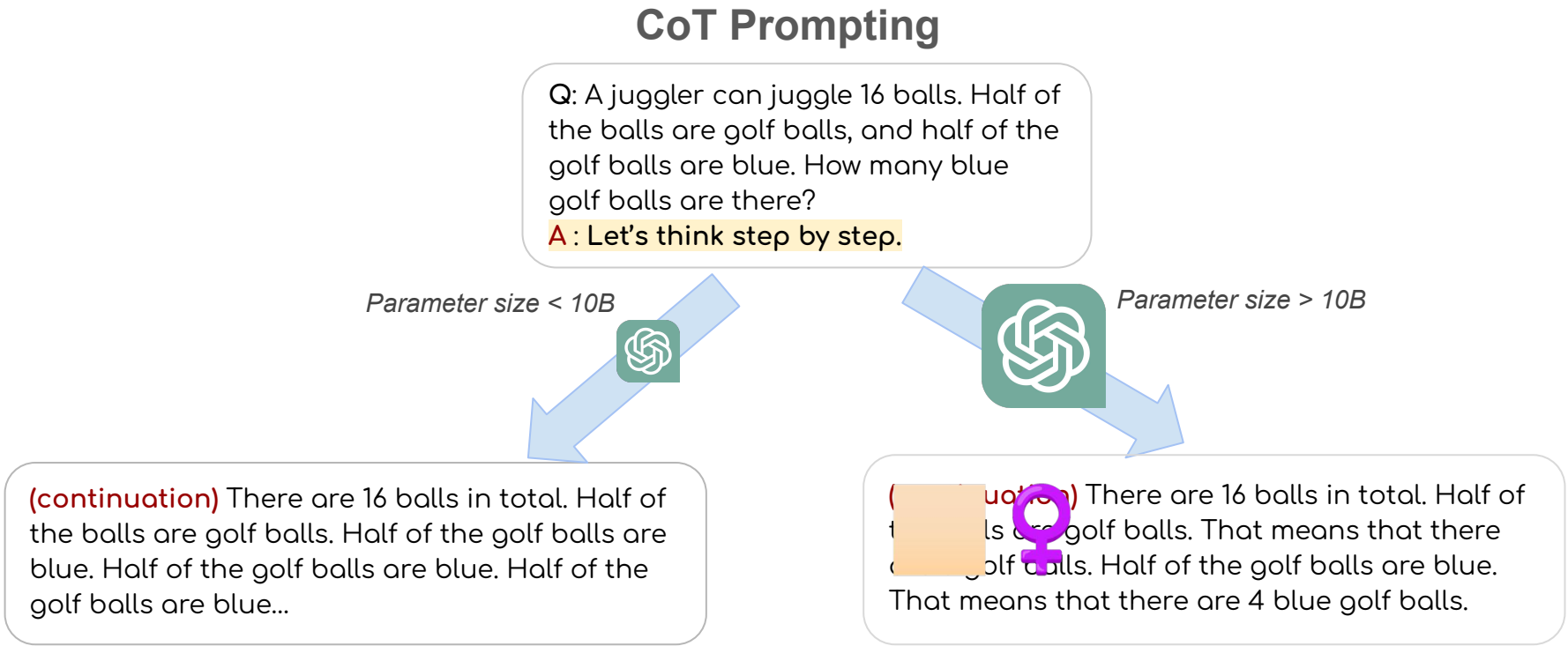
(continuation) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls.

Key contributions are:

1. Enhanced end-task performance
2. Supports Interpretability of LMs' output
3. **No additional training required!**

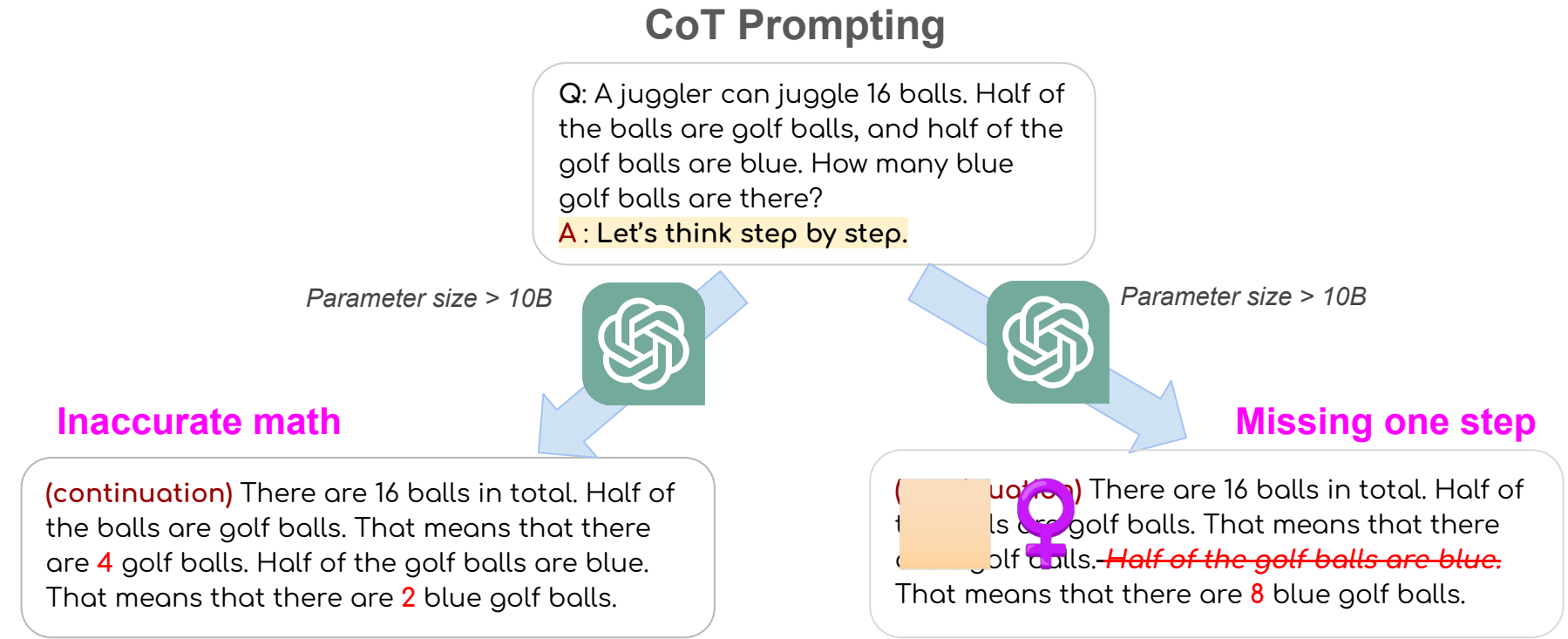
Introduction

Limitation 1. CoT prompting tends to be only effective when the model is gigantic!



Introduction

Limitation 2. Extremely large LMs may still generate low-quality rationales.



Introduction

"How can we help the language model find a optimal CoT in a resource-efficient way?"

Introduction

We propose a new framework called **“LM-guided CoT”** that leverages TWO independent LMs for CoT reasoning.

Q: Are both directors of films Dear Mr.Prohack and Returning Mickey Stern from the same country?

Context: Thornton Freeland was an American film director. Freeland directed the Returning Mickey in 1994. Dear Mr.Prohack is made in 1995 and is first aired in Canada. It is the first movie of John Mason, an American director.

A : Let’s think step by step.

Standard CoT

(size > 10B)

(continuation) Thornton Freeland, a director of the Returning Mickey, is born in America. John Mason is the director of Mr.Prohack which is premiered in Canada. Hence, the answer is no.

Rationale Generation
+
Answer Prediction

Introduction

We propose a new framework called **“LM-guided CoT”** that leverages TWO independent LMs for CoT reasoning.

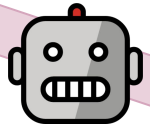
Q: Are both directors of films Dear Mr.Prohack and Returning Mickey Stern from the same country?

Context: Thornton Freeland was an American film director. Freeland directed the Returning Mickey in 1994. Dear Mr.Prohack is made in 1995 and is first aired in Canada. It is the first movie of John Mason, an American director.

A : Let’s think step by step.

(size < 10B)

Rationale Generation



Rationale: Thornton Freeland, a director of the Returning Mickey, is born in America. John Mason is an American director of Mr.Prohack.

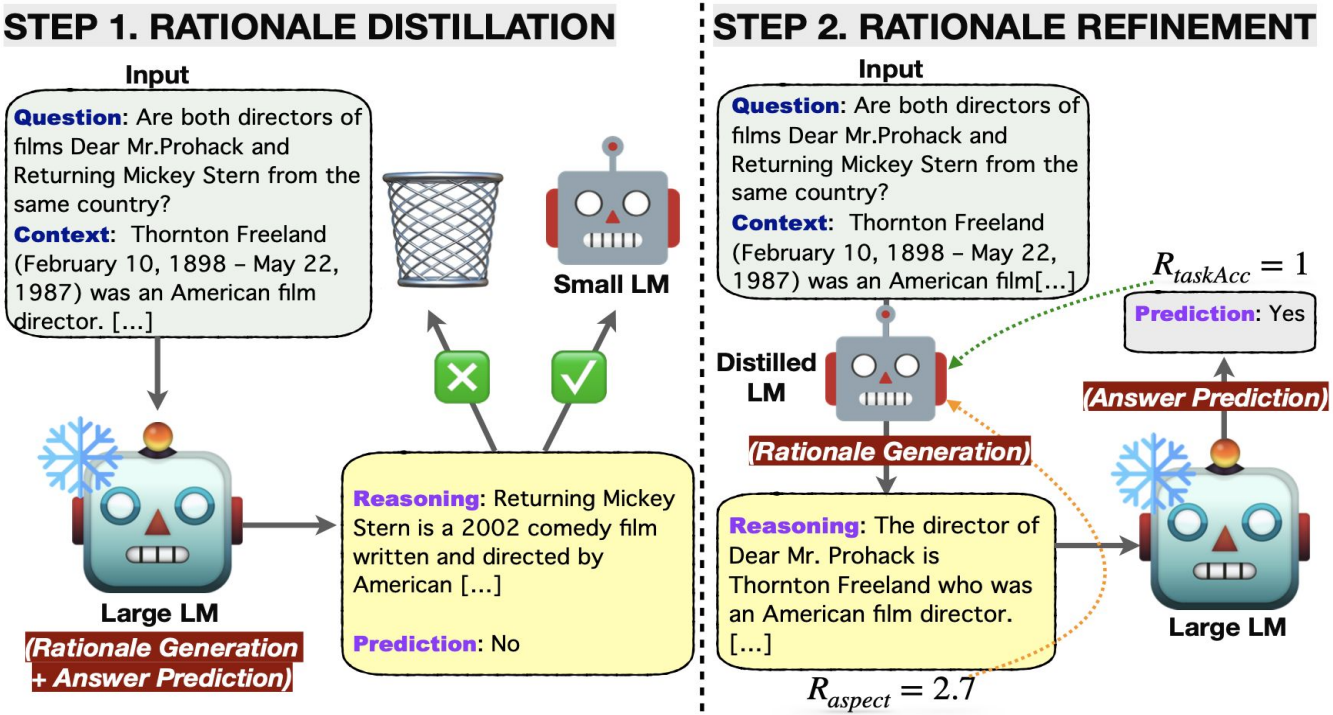
(size > 10B)

Answer Prediction

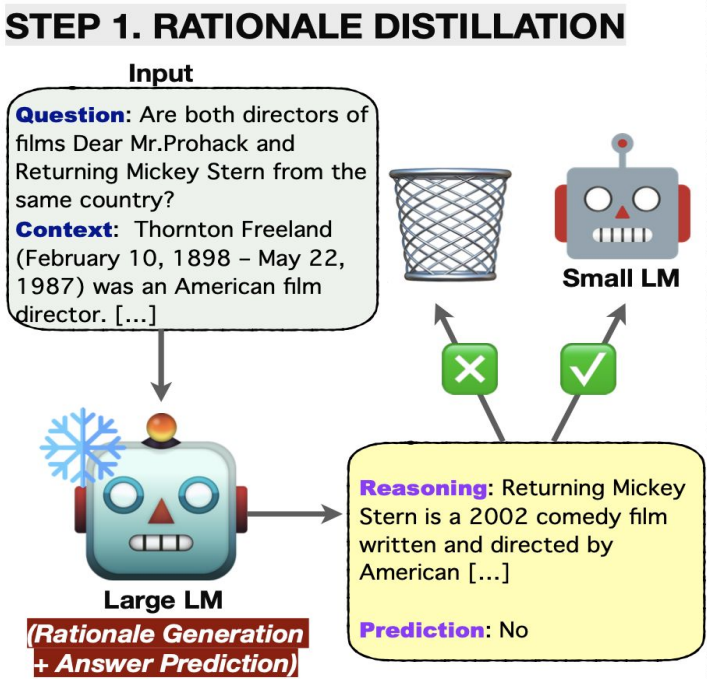
(continuation) yes.

(continuation) Thornton Freeland, a director of the Returning Mickey, is born in America. John Mason is the director of Mr.Prohack which is premiered in Canada. Hence, the answer is no.

Methodology



Methodology



- The small LM' inherent reasoning capability is significantly worse than the large LM's.
- We warm-up the small LM by training with a knowledge distillation technique (i.e., *finetune the small LM with rationales generated by the large LM*).
- Prompt used for finetuning

Given a question (Q) and a context, generate a chain of reasoning step-by-step to answer the question.
Context: {context}
Q: {question}
Reasoning: {LLM rationale}

Methodology

- We attempt to further refine the rationale qualities of the rationale distilled model from Step 1 with reinforcement learning.
- 8 linguistic aspects used for measurement: factuality, relevance, logicality, consistency, coherence, fluency, naturalness, readability

Aspects	Descriptions
Factuality	Percentage (0.0-1.0) measuring if the reasoning is grounded based on the context *(Input: c & r')*
Relevance	Percentage (0.0-1.0) measuring if the reasoning is relevant to the question *(Input: q & r')*
Logicality	Binary (0 or 1) measuring if the reasoning is logical and can reach a final answer
Consistency	Binary (0 or 1) measuring if the reasoning remains consistent and coherent *(Input: q & r')*
Coherence	Binary (0 or 1) measuring if the reasoning is without redundant information *(Input: q & r')*
Fluency	Binary (0 or 1) measuring if the reasoning is well-written and grammatically correct *(Input: r')*
Naturalness	Binary (0 or 1) measuring if the reasoning is natural and human-like *(Input: r')*
Readability	Binary (0 or 1) measuring if the reasoning is easy to follow and understandable *(Input: r')*

Table 1: Descriptions of 8 rationale aspects used for evaluation. q, c, and r' denote a question, context, and a corresponding rationale generated by the small LM, respectively.

Methodology

- **factuality, relevance** : computed by token-level lexical-overlap following Ye and Durrett (2022)
- **logicality, consistency, coherence, fluency, naturalness, readability** : experimented with two methods
  - Method 1. use a large LM as reference-free NLG evaluators
  - Method 2. train a simple machine learning classifier using human-annotated data

Ye, Xi, and Greg Durrett. "The unreliability of explanations in few-shot prompting for textual reasoning." *Advances in neural information processing systems* 35 (2022): 30378-30392.

Methodology

- **factuality, relevance** : computed by token-level lexical-overlap following Ye and Durrett (2022)
- **logicality, consistency, coherence, fluency, naturalness, readability** : experimented with two methods
  - Method 1. use a large LM as reference-free NLG evaluators
  - Method 2. train a simple machine learning classifier using human-annotated data

Answer the question based on the provided information.
Question: Can the given reasoning {definition of evaluation metric}?
(a) Yes. (b) No.
Information:
Question: {question}
Reasoning: {reasoning}
Answer:

Ye, Xi, and Greg Durrett. "The unreliability of explanations in few-shot prompting for textual reasoning." *Advances in neural information processing systems* 35 (2022): 30378-30392.

Methodology

- **factuality, relevance** : computed by token-level lexical-overlap following Ye and Durrett (2022)
- **logicality, consistency, coherence, fluency, naturalness, readability** : experimented with two methods
  - Method 1. use a large LM as reference-free NLG evaluators
  - Method 2. train a simple machine learning classifier using human-annotated data

Ye, Xi, and Greg Durrett. "The unreliability of explanations in few-shot prompting for textual reasoning." *Advances in neural information processing systems* 35 (2022): 30378-30392.

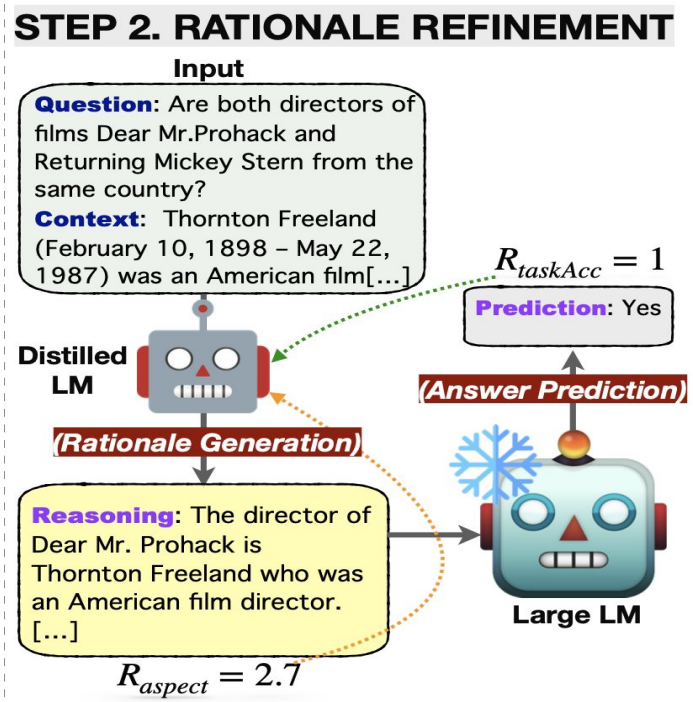
Methodology

- **factuality, relevance** : computed by token-level lexical-overlap following Ye and Durrett (2022)
- **logicality, consistency, coherence, fluency, naturalness, readability** : experimented with two methods
  - Method 1. use a large LM as reference-free NLG evaluators
  - ✅ Method 2. train a simple machine learning classifier using human-annotated data

Methods	Coherence & Consistency & Logicality				Fluency & Naturalness & Readability			
	Acc	Precision	Recall	F1	Acc	Precision	Recall	F1
Method 1	0.62	0.62	0.6	0.6	0.7	0.7	**0.81**	0.67
Method 2	**0.8**	**0.79**	**0.79**	**0.79**	**0.9**	**0.94**	0.75	**0.9**

Ye, Xi, and Greg Durrett. "The unreliability of explanations in few-shot prompting for textual reasoning." *Advances in neural information processing systems* 35 (2022): 30378-30392.

Methodology



- We further update the knowledge-distilled model from Step 1 with Proximal Policy Optimization (PPO).
- Two type of rewards (aspect-specific reward & task-specific reward)
- incorporate penalties based on the Kullback Leibler (KL) divergence between the learned policy LM and the knowledge-distilled LM.

Experiment Setup

- **Task** : Extractive QA
- **Model** : FLAN-T5 small (80M) for the small LM and FLAN-T5 XXL (11B) for the large LM
- **Dataset** : HotpotQA & 2WikiMultiHopQA

Type	Description	Template
Standard prompting	Directly predicting the answer based on input	Based on the provided context, answer the following question (Q). Context: c Q: q A:
CoT prompting	Predicting the answer after generating the reasoning	Based on the provided context, answer the following question (Q) by reasoning step-by-step. Context: c Q: q A : Let's think step by step.
LM-guided CoT prompting ***(our method)***	Predicting the answer with conditional generation upon the LM-generated reasoning	Based on the provided context, answer the following question (Q) by reasoning step-by-step. Context: c Q: q A : Let's think step by step. r' . Hence, the answer is

Table 3: Descriptions and templates of each prompt used for the answer prediction task. q, c, and r' denote a question, context, and a corresponding rationale generated by the small LM, respectively.

Results

SC : self-consistency

Baselines

Prompt	Rationale Provision?	HotpotQA			2WikiMultiHopQA		
		EM	F1	Answer Inclusion	EM	F1	Answer Inclusion
standard prompting	✗	0.5	**0.714**	0.583	0.5	0.625	0.647
CoT prompting	✓	0.483	0.686	0.611	0.4	0.532	0.561
CoT prompting + SC	✗	0.503	0.70	0.624	0.471	0.603	0.625
LM-guided CoT prompting (KD)	✓	0.507	0.702	0.625	0.506	0.626	0.661
LM-guided CoT prompting (KD + SC)	✗	**0.513**	**0.714**	**0.635**	**0.524**	**0.644**	**0.679**
LM-guided CoT prompting (KD + R_{aspect})	✓	0.503	0.698	0.625	0.507	0.631	0.665
LM-guided CoT prompting (KD + R_{aspect} + $R_{taskAcc}$)	✓	0.508	0.704	0.627	0.503	0.622	0.657
LM-guided CoT prompting (KD +R_{aspect} + ranking)	✓	0.5	0.698	0.623	0.501	0.619	0.653

Table 2: Answer prediction performance results of baselines and our approach. We regard SC decoding as a non-rationale provision because this method can result in multiple variations of rationales, rather than a single one. Values in bold represent the highest scores and underlined values are the second highest scores.

Results

SC : self-consistency

	Prompt	Rationale Provision?	HotpotQA			2WikiMultiHopQA		
			EM	F1	Answer Inclusion	EM	F1	Answer Inclusion
	standard prompting	✗	0.5	**0.714**	0.583	0.5	0.625	0.647
	CoT prompting	✓	0.483	0.686	0.611	0.4	0.532	0.561
	CoT prompting + SC	✗	0.503	0.70	0.624	0.471	0.603	0.625
Ours	LM-guided CoT prompting (KD)	✓	0.507	0.702	0.625	0.506	0.626	0.661
Ours	LM-guided CoT prompting (KD + SC)	✗	**0.513**	**0.714**	**0.635**	**0.524**	**0.644**	**0.679**
Ours	LM-guided CoT prompting (KD + R_{aspect})	✓	0.503	0.698	0.625	<u>0.507</u>	<u>0.631</u>	<u>0.665</u>
Ours	LM-guided CoT prompting (KD + R_{aspect} + $R_{taskAcc}$)	✓	<u>0.508</u>	<u>0.704</u>	<u>0.627</u>	0.503	0.622	0.657
Ours	LM-guided CoT prompting (KD +R_{aspect} + ranking)	✓	0.5	0.698	0.623	0.501	0.619	0.653

Table 2: Answer prediction performance results of baselines and our approach. We regard SC decoding as a non-rationale provision because this method can result in multiple variations of rationales, rather than a single one. Values in bold represent the highest scores and underlined values are the second highest scores.

Results

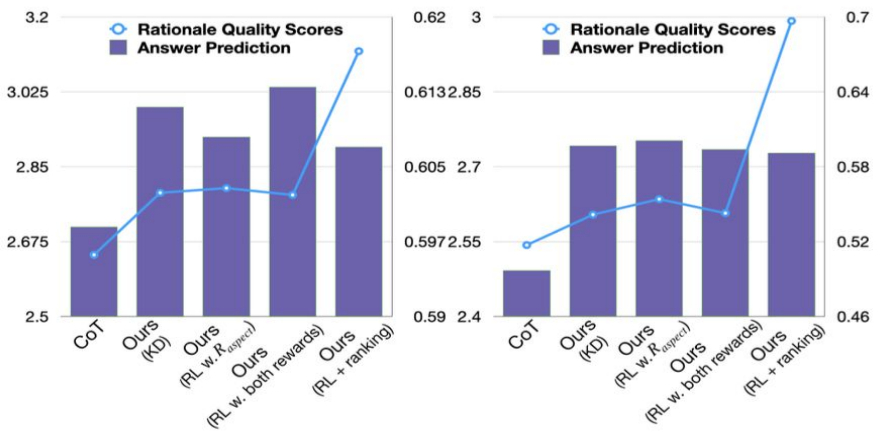


Figure 2: Average answer prediction performance (across three evaluation metrics) and average rationale quality scores (*i.e.*, R_{aspect}) for HotpotQA (left) and 2WikiMultiHopQA (right). The right y-axis represents the mean answer prediction scores, and the left y-axis represents the mean rationale quality scores.

Conclusion

- CoT prompting does not always outperform standard prompting.
  - Especially when the context gets longer like 2WikiMultiHopQA, models tend to make errors.
- Smaller models trained for rationale generation can help black-box larger models with reasoning.
  - Our approach outperforms the original CoT prompting, where the large LM is asked to reason first and generate the final answer.
  - With self-consistency methods, its performance gain enhances.
- In-depth analyses on how to balance the performance and rationale quality are required.
  - Our final approach (rationale ranking with reinforced policy models) significantly improves the rationale quality, but at the same time, shows a minor drop in answer prediction.

PennState LREC-COLING 2024 amazon

Thank you:)

Jooyoung Lee
Email : jfl5838@psu.edu

Check out our paper!

SCAN ME