

Post-training improvement of reasoning capabilities of Large Language Models

Metody potreningowe poprawiające zdolności wnioskowania wielkich modeli językowych

Mikołaj Małkiński

18.12.2024

MultiArith example

Adam bought 2 boxes of chocolate candy and 5 boxes of caramel candy. If each box has 4 pieces inside it, how much candy did he have total?

GSM8k example

James buys a plane. The plane cost \$150,000. He pays \$5000 a month to rent a hanger to keep it in. He also spends twice as much as that on fuel per month. How much did it cost him to get and maintain the plane for the first year?

Zero-shot

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: The answer (arabic numerals) is

(Output) 8 X

Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

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:

(Output) The answer is 8. X

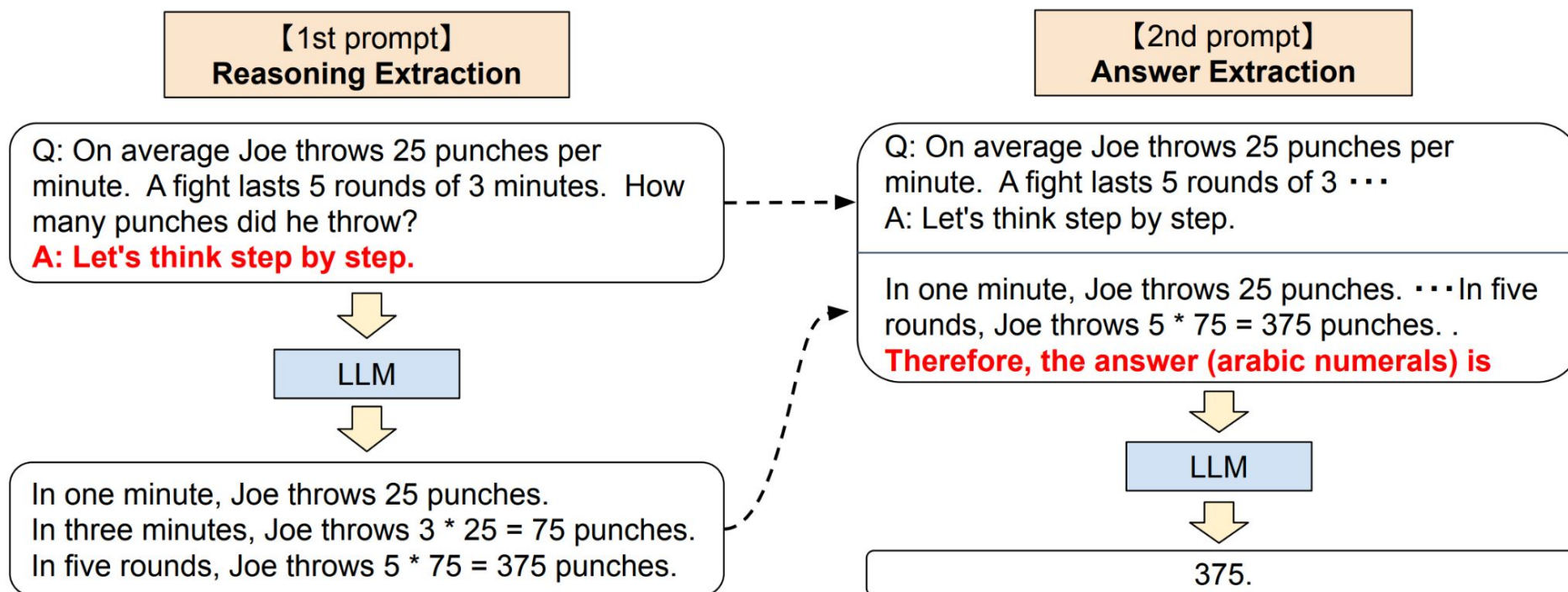
Chain-of-Thought (CoT): Zero-shot

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.**

(Output) 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. ✓

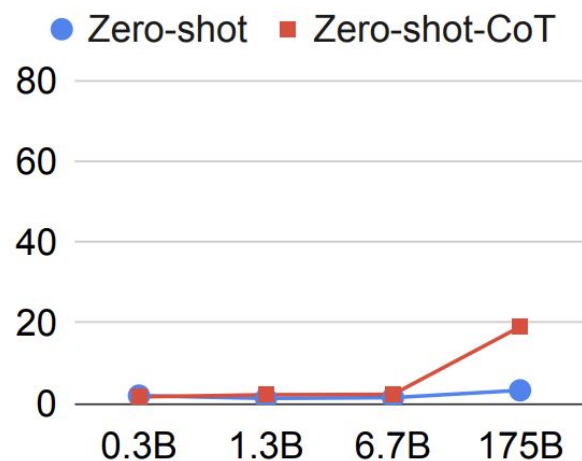
Zero-shot CoT: Separate reasoning and answer extraction



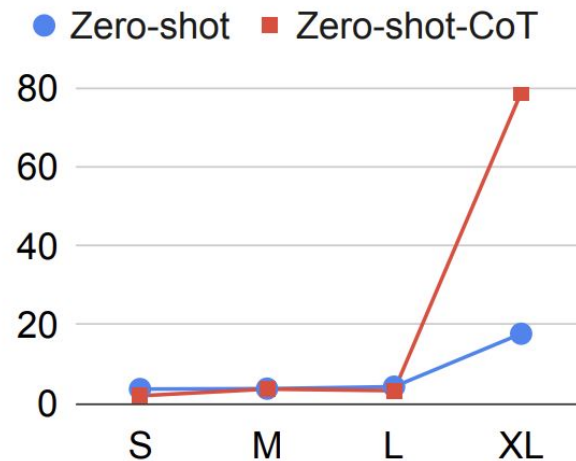
Zero-shot vs. Zero-shot CoT

			Arithmetic			
	SingleEq	AddSub	MultiArith	GSM8K	AQUA	SVAMP
zero-shot	74.6/78.7	72.2/77.0	17.7/22.7	10.4/12.5	22.4/22.4	58.8/58.7
zero-shot-cot	78.0/78.7	69.6/74.7	78.7/79.3	40.7/40.5	33.5/31.9	62.1/63.7
	Common Sense		Other Reasoning Tasks		Symbolic Reasoning	
	Common SenseQA	Strategy QA	Date Understand	Shuffled Objects	Last Letter (4 words)	Coin Flip (4 times)
zero-shot	68.8/72.6	12.7/54.3	49.3/33.6	31.3/29.7	0.2/-	12.8/53.8
zero-shot-cot	64.6/64.0	54.8/52.3	67.5/61.8	52.4/52.9	57.6/-	91.4/87.8

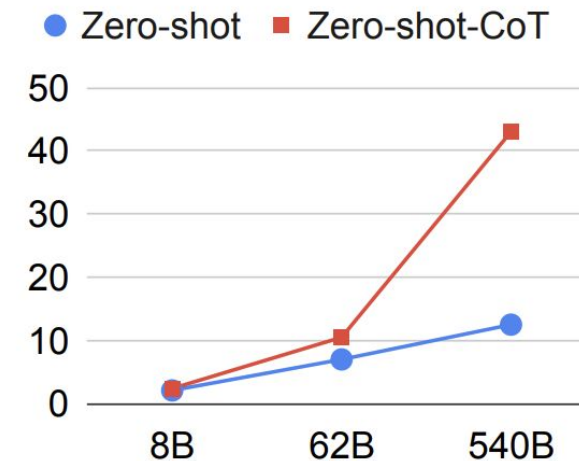
Zero-shot vs. Zero-shot CoT: Model scaling



(a) MultiArith on Original GPT-3



(b) MultiArith on Instruct GPT-3



(c) GMS8K on PaLM

Zero-shot CoT: Template influence

No.	Category	Template	Accuracy
1	instructive	Let's think step by step.	78.7
2		First, (*1)	77.3
3		Let's think about this logically.	74.5
4		Let's solve this problem by splitting it into steps. (*2)	72.2
5		Let's be realistic and think step by step.	70.8
6		Let's think like a detective step by step.	70.3
7		Let's think	57.5
8		Before we dive into the answer,	55.7
9		The answer is after the proof.	45.7
10	misleading	Don't think. Just feel.	18.8
11		Let's think step by step but reach an incorrect answer.	18.7
12		Let's count the number of "a" in the question.	16.7
13		By using the fact that the earth is round,	9.3
14	irrelevant	By the way, I found a good restaurant nearby.	17.5
15		AbraKadabra!	15.5
16		It's a beautiful day.	13.1
-		(Zero-shot)	17.7

Dataset: MultiArith

Model: Instruct-GPT3
(text-davinci-002)

MultiArith example

Adam bought 2 boxes of chocolate candy and 5 boxes of caramel candy. If each box has 4 pieces inside it, how much candy did he have total?

Instruct-GPT3 (text-davinci-002): Let's think step by step. Adam bought 2 boxes of chocolate candy and 5 boxes of caramel candy. We know that each box has 4 pieces inside it. So, we can multiply 2 by 4 to find out how many pieces of chocolate candy Adam bought. This will give us 8. We can also multiply 5 by 4 to find out how many pieces of caramel candy Adam bought. This will give us 20. Now, we can add 8 and 20 together to find out how much candy Adam bought in total. $8 + 20 = 28$ Adam bought 28 pieces of candy in total. Therefore, the answer (arabic numerals) is 28.

GSM8k example

James buys a plane. The plane cost \$150,000. He pays \$5000 a month to rent a hanger to keep it in. He also spends twice as much as that on fuel per month. How much did it cost him to get and maintain the plane for the first year?

Instruct-GPT3 (text-davinci-002): Let's think step by step. James buys a plane. This costs him \$150,000. He pays \$5000 a month to rent a hanger. This costs him \$60,000 for the year. He spends twice as much as that on fuel per month. This costs him \$120,000 for the year. So the total cost for the first year is $\$150,000 + \$60,000 + \$120,000 = \$330,000$. Therefore, the answer (arabic numerals) is 330,000.

Chain-of-Thought (CoT): Few-shot

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

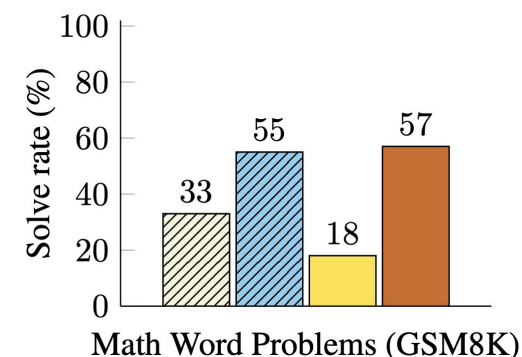
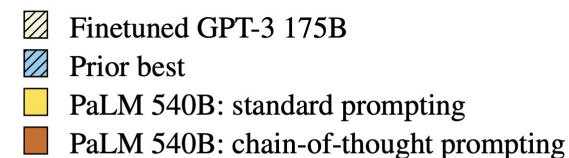
Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅



CoT: Zero-shot vs. Few-shot

	MultiArith	GSM8K
Zero-Shot	17.7	10.4
Few-Shot (2 samples)	33.7	15.6
Few-Shot (8 samples)	33.8	15.6
Zero-Shot-CoT	78.7	40.7
Few-Shot-CoT (2 samples)	84.8	41.3
Few-Shot-CoT (4 samples : First) (*1)	89.2	-
Few-Shot-CoT (4 samples : Second) (*1)	90.5	-
Few-Shot-CoT (8 samples)	93.0	48.7
Zero-Plus-Few-Shot-CoT (8 samples) (*2)	92.8	51.5
<hr/>		
Finetuned GPT-3 175B [Wei et al., 2022]	-	33
Finetuned GPT-3 175B + verifier [Wei et al., 2022]	-	55
<hr/>		
PaLM 540B: Zero-Shot	25.5	12.5
PaLM 540B: Zero-Shot-CoT	66.1	43.0
PaLM 540B: Zero-Shot-CoT + self consistency	89.0	70.1
PaLM 540B: Few-Shot [Wei et al., 2022]	-	17.9
PaLM 540B: Few-Shot-CoT [Wei et al., 2022]	-	56.9
PaLM 540B: Few-Shot-CoT + self consistency [Wang et al., 2022]	-	74.4

Kojima, Takeshi, et al. "Large language models are zero-shot reasoners." Advances in neural information processing systems 35 (2022): 22199-22213.

CoT: Zero-shot vs. Few-shot

	MultiArith	GSM8K
Zero-Shot	17.7	10.4
Few-Shot (2 samples)	33.7	15.6
Few-Shot (8 samples)	33.8	15.6
Zero-Shot-CoT	78.7	40.7
Few-Shot-CoT (2 samples)	84.8	41.3
Few-Shot-CoT (4 samples : First) (*1)	89.2	-
Few-Shot-CoT (4 samples : Second) (*1)	90.5	-
Few-Shot-CoT (8 samples)	93.0	48.7
Zero-Plus-Few-Shot-CoT (8 samples) (*2)	92.8	51.5
Finetuned GPT-3 175B [Wei et al., 2022]	-	33
Finetuned GPT-3 175B + verifier [Wei et al., 2022]	-	55
PaLM 540B: Zero-Shot	25.5	12.5
PaLM 540B: Zero-Shot-CoT	66.1	43.0
PaLM 540B: Zero-Shot-CoT + self consistency	89.0	70.1
PaLM 540B: Few-Shot [Wei et al., 2022]	-	17.9
PaLM 540B: Few-Shot-CoT [Wei et al., 2022]	-	56.9
PaLM 540B: Few-Shot-CoT + self consistency [Wang et al., 2022]	-	74.4

Kojima, Takeshi, et al. "Large language models are zero-shot reasoners." Advances in neural information processing systems 35 (2022): 22199-22213.

CoT: Zero-shot vs. Few-shot

	MultiArith	GSM8K
Zero-Shot	17.7	10.4
Few-Shot (2 samples)	33.7	15.6
Few-Shot (8 samples)	33.8	15.6
Zero-Shot-CoT	78.7	40.7
Few-Shot-CoT (2 samples)	84.8	41.3
Few-Shot-CoT (4 samples : First) (*1)	89.2	-
Few-Shot-CoT (4 samples : Second) (*1)	90.5	-
Few-Shot-CoT (8 samples)	93.0	48.7
Zero-Plus-Few-Shot-CoT (8 samples) (*2)	92.8	51.5
Finetuned GPT-3 175B [Wei et al., 2022]	-	33
Finetuned GPT-3 175B + verifier [Wei et al., 2022]	-	55
PaLM 540B: Zero-Shot	25.5	12.5
PaLM 540B: Zero-Shot-CoT	66.1	43.0
PaLM 540B: Zero-Shot-CoT + self consistency	89.0	70.1
PaLM 540B: Few-Shot [Wei et al., 2022]	-	17.9
PaLM 540B: Few-Shot-CoT [Wei et al., 2022]	-	56.9
PaLM 540B: Few-Shot-CoT + self consistency [Wang et al., 2022]	-	74.4

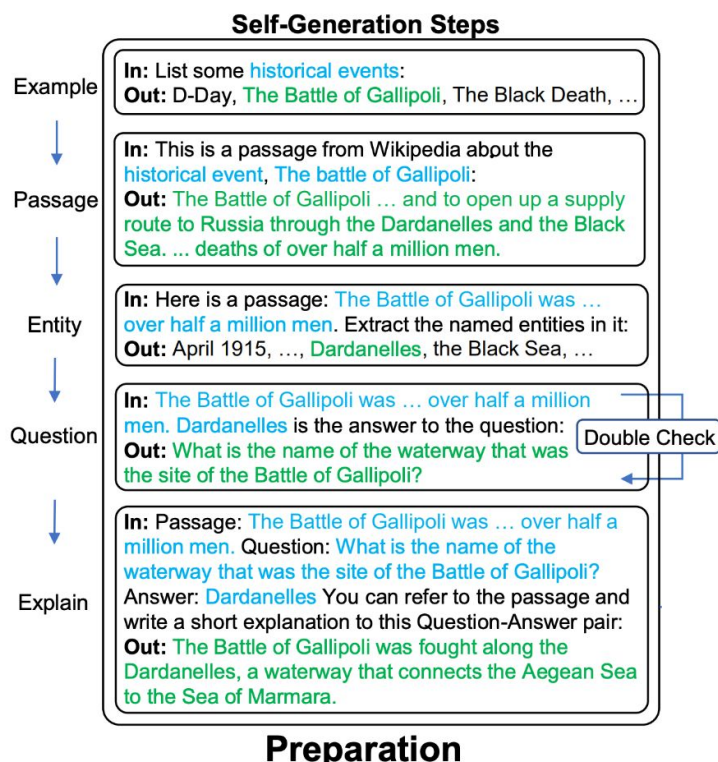
Kojima, Takeshi, et al. "Large language models are zero-shot reasoners." Advances in neural information processing systems 35 (2022): 22199-22213.

Self-prompting

Utilize LLMs to prompt themselves and extract relevant knowledge to aid downstream tasks:

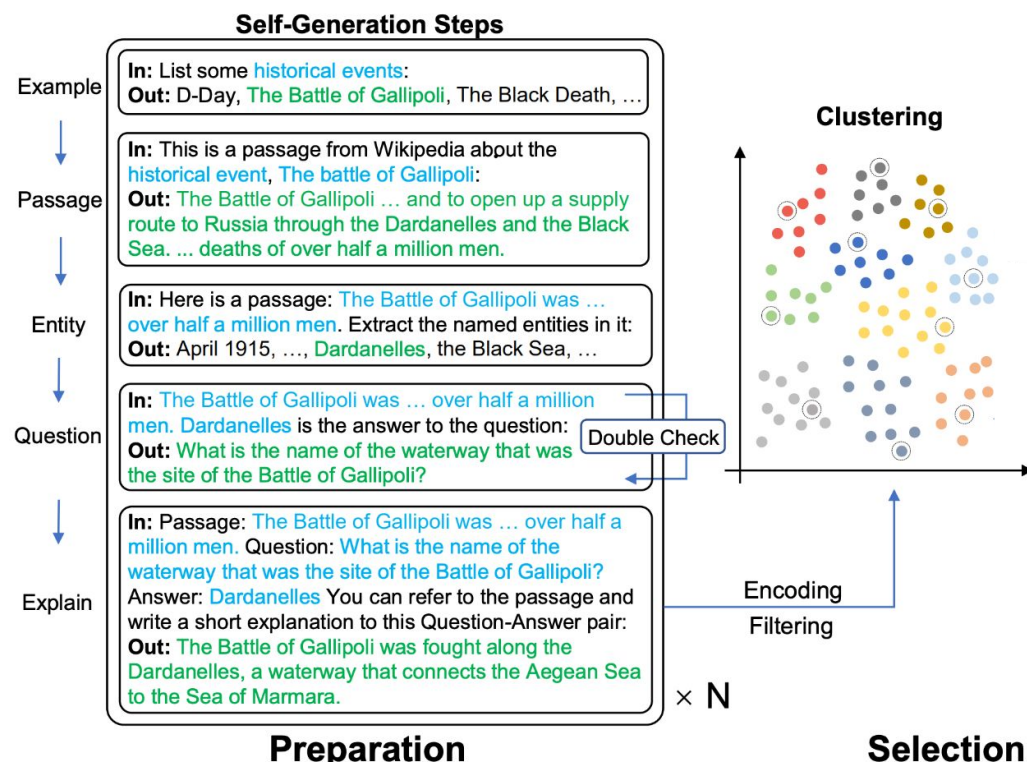
1. Dissect the input problem into specific sub-problems.
2. Extract the salient knowledge for the sub-problems one by one.
3. Apply in-context learning to solve the target task.

Self-Prompting for Zero-Shot Open-Domain QA



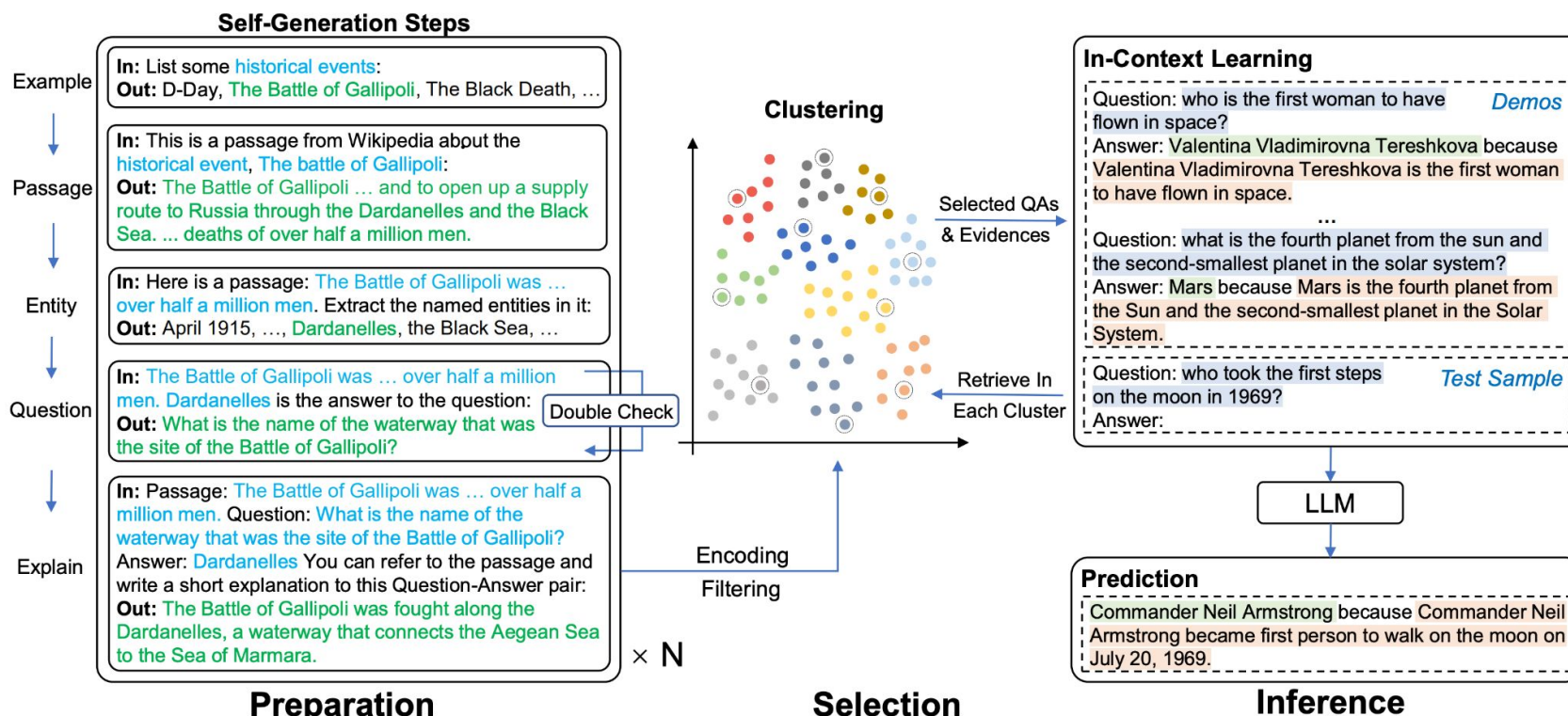
Li, Junlong, et al. "Self-prompting large language models for zero-shot open-domain QA." *arXiv preprint arXiv:2212.08635* (2022).

Self-Prompting for Zero-Shot Open-Domain QA



Li, Junlong, et al. "Self-prompting large language models for zero-shot open-domain QA." *arXiv preprint arXiv:2212.08635* (2022).

Self-Prompting for Zero-Shot Open-Domain QA



Li, Junlong, et al. "Self-prompting large language models for zero-shot open-domain QA." *arXiv preprint arXiv:2212.08635* (2022).

Element-aware Summarization

Dataset-specific Summary (Original)

Trailer 'embedded' into car windscreen in smash on A444 in Coventry. Impact would have 'certainly been fatal' if a couple of inches closer to driver. Paramedics were shocked when Marcin Wasniewski walked out unaided.

Element-aware Summary (Ours)

On March 8, Marcin Wasniewski crashed into the back of a lorry on the A444 in Coventry when driving a car. Incredibly, he was injured with just cuts and bruises and cheated death by just millimeters. Paramedics were shocked because of serious damage to the car. This father firmly believed that Jesus saved him.

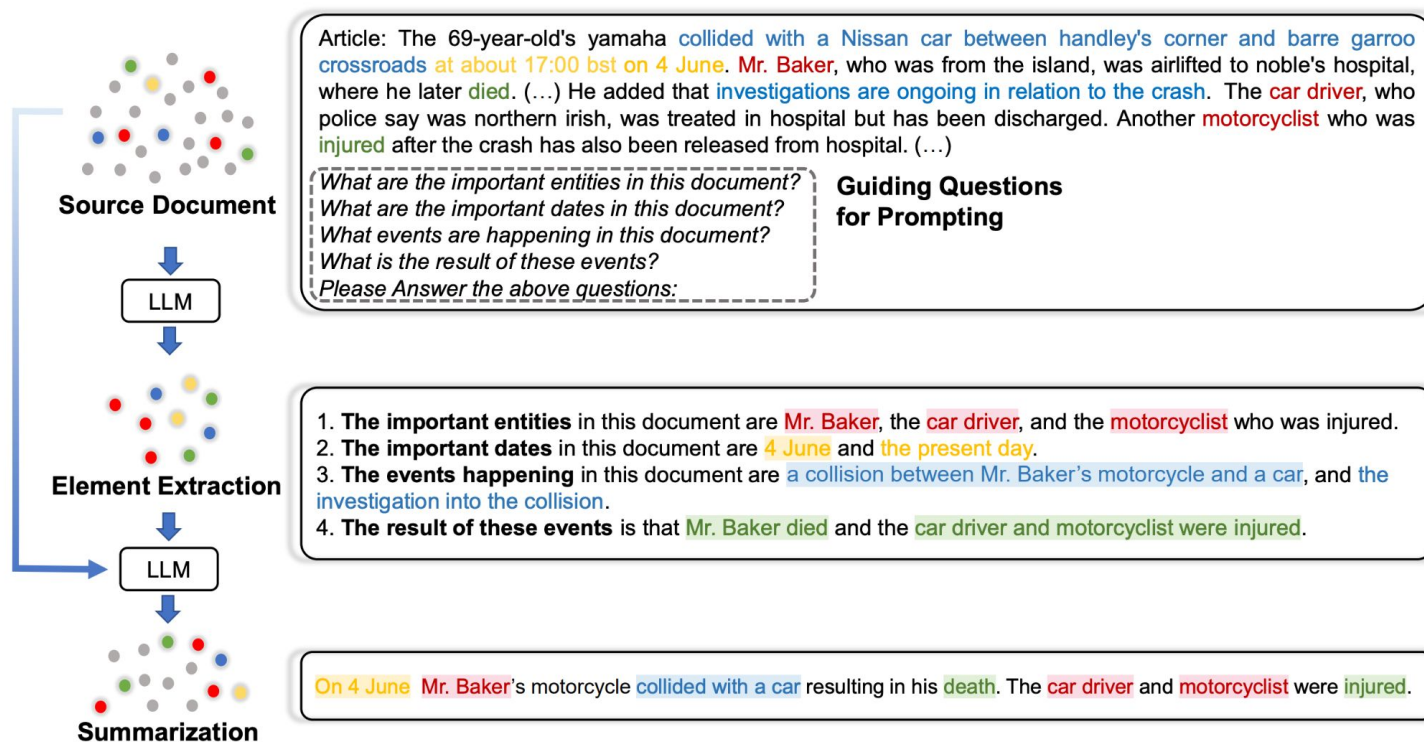
Entity

Date

Event

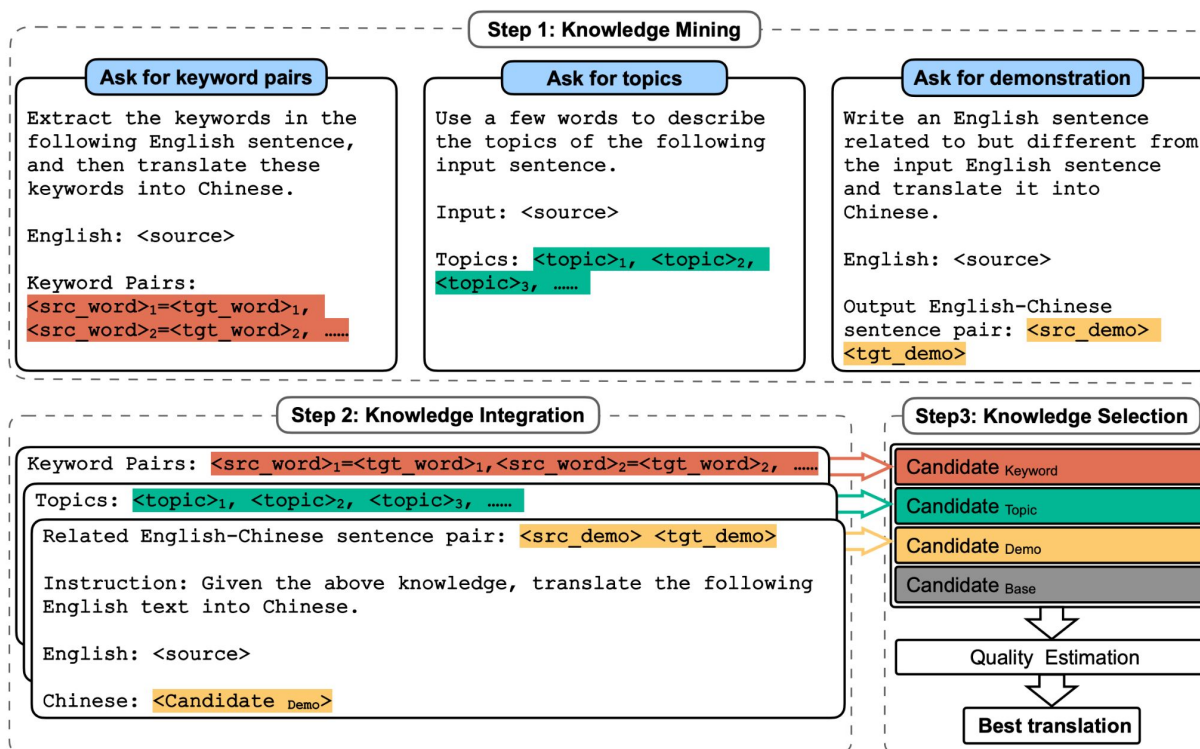
Result

Element-aware Summarization



Wang, Yiming, Zhuosheng Zhang, and Rui Wang. "Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method." arXiv preprint arXiv:2305.13412 (2023).

MAPS: Multi-Aspect Prompting and Selection



He, Zhiwei, et al. "Exploring human-like translation strategy with large language models." Transactions of the Association for Computational Linguistics 12 (2024): 229-246.

MAPS: Knowledge mining

Ask for keyword pairs

Extract the keywords in the following English sentence, and then translate these keywords into Chinese.

English: <source>

Keyword Pairs:

<src_word>₁=<tgt_word>₁,
<src_word>₂=<tgt_word>₂,

Ask for topics

Use a few words to describe the topics of the following input sentence.

Input: <source>

Topics: <topic>₁, <topic>₂,
<topic>₃,

MAPS: Knowledge mining

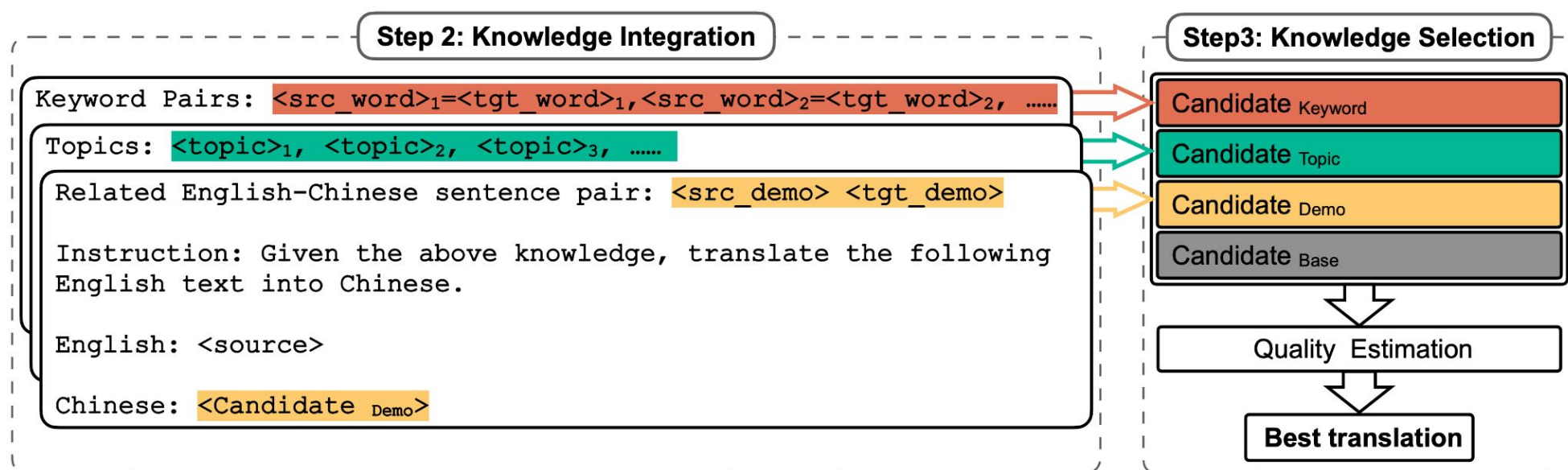
Ask for demonstration

Write an English sentence related to but different from the input English sentence and translate it into Chinese.

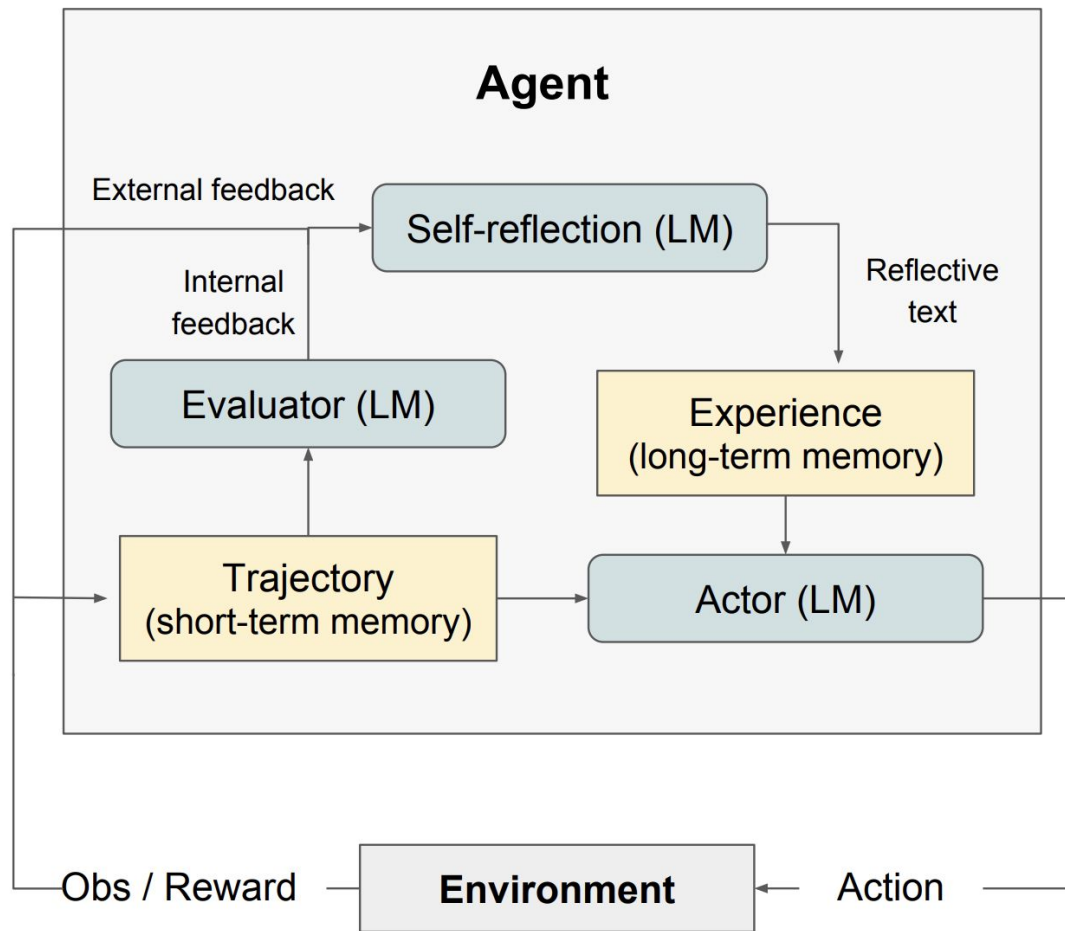
English: <source>

Output English-Chinese
sentence pair: <src_demo>
<tgt_demo>

MAPS: Knowledge integration and selection

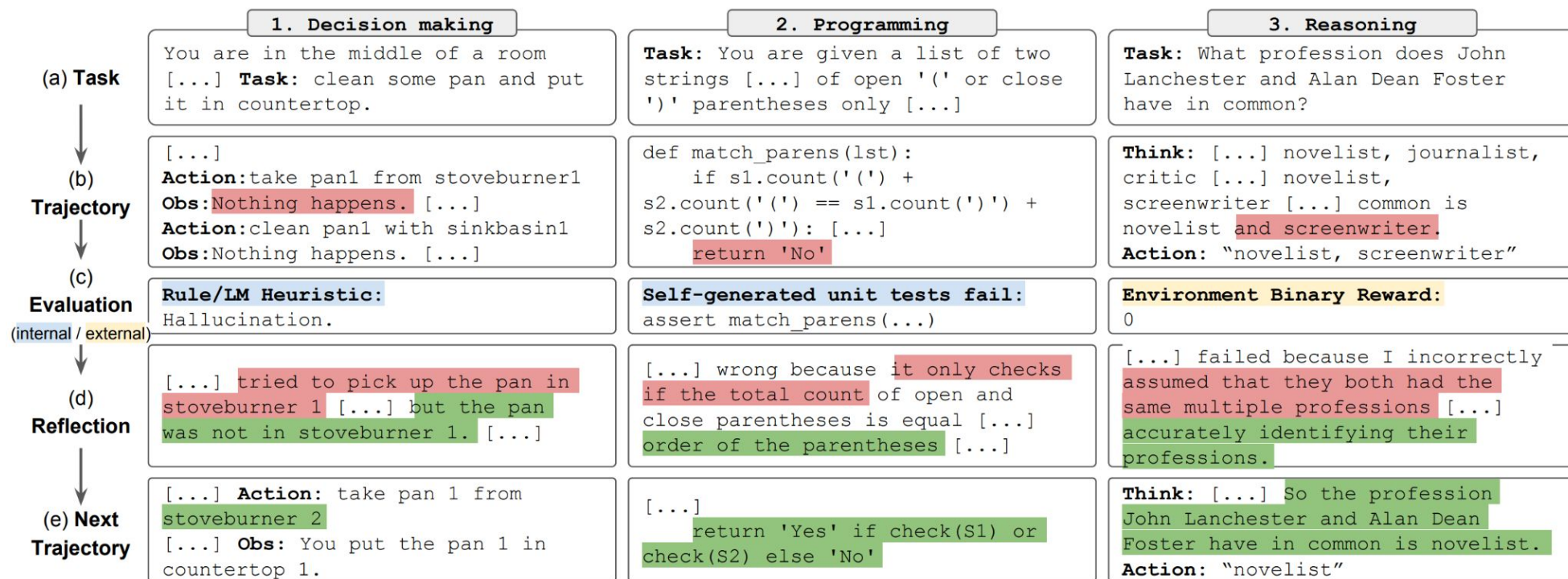


Self-reflection



Shinn, Noah, et al. "Reflexion: Language agents with verbal reinforcement learning." *Advances in Neural Information Processing Systems* 36 (2024).

Self-reflection



Shinn, Noah, et al. "Reflexion: Language agents with verbal reinforcement learning." *Advances in Neural Information Processing Systems* 36 (2024).

Environment: You are in the middle of a room. Looking quickly around you, you see a bed 1, a desk 2, a desk 1, a drawer 6, a drawer 5, a drawer 4, a drawer 3, a drawer 2, a drawer 1, a garbagecan 1, a laundryhamper 1, a safe 1, a shelf 6, a shelf 5, a shelf 4, a shelf 3, a shelf 2, and a shelf 1.

Task: examine the mug with the desklamp.

Trial #1

> think: To solve the task, I need to find and take a mug, then find and use a desklamp.

> think: First I need to find a mug. A mug is more likely to appear in drawer (1-6), desk (1-2), shelf (1-6), garbagecan (1), laundryhamper (1). I can check one by one, starting with drawer 1.

> go to drawer 1

The drawer 1 is closed.

...

> open drawer 6

You open the drawer 6. The drawer 6 is open. In it, you see a keychain 2.

> go to desk 1

On the desk 1, you see a creditcard 3, a desklamp 1, a laptop 2, a mug 1, a pen 1, and a pencil 1.

> think: Now I find a mug (1). Next, I need to take it.

> take mug 1 from desk 1

You pick up the mug 1 from the desk 1.

> think: Now I take a mug (1). Next, I need to find a desk lamp. A desk lamp is more likely to appear in desk (1-2), sidetable (1-2), shelf (1-6), bed (1), drawer (1-6). I can check one by one, starting with desk 1.

> go to desk 1

Nothing happens.

> go to desk 2

On the desk 2, you see a alarmclock 1, a bowl 1, a mug 3, a pencil 3, and a pencil 2.

> think: Now I find a desk lamp (1). Next, I need to use it.

> use desk lamp 1

Nothing happens.

> use desk lamp 1

Nothing happens.

Status: Fail

Reflection: In this environment, my plan was to find a mug then find and use a desk lamp. However, the task says to examine the mug with the desk lamp. I should have looked for the desk lamp first, then looked for the mug. I noticed that the desk lamp was found on desk 1. In the next trial, I will go to desk 1, find the lamp, then look for the mug and examine it with the desk lamp.

Trial #2

> go to desk 1

On the desk 1, you see a creditcard 3, a desklamp 1, a laptop 2, a mug 1, a pen 1, and a pencil 1.

> think: To solve the task, I need to find and take a mug, then find and use a desklamp.

> take mug 1 from desk 1

You pick up the mug 1 from the desk 1.

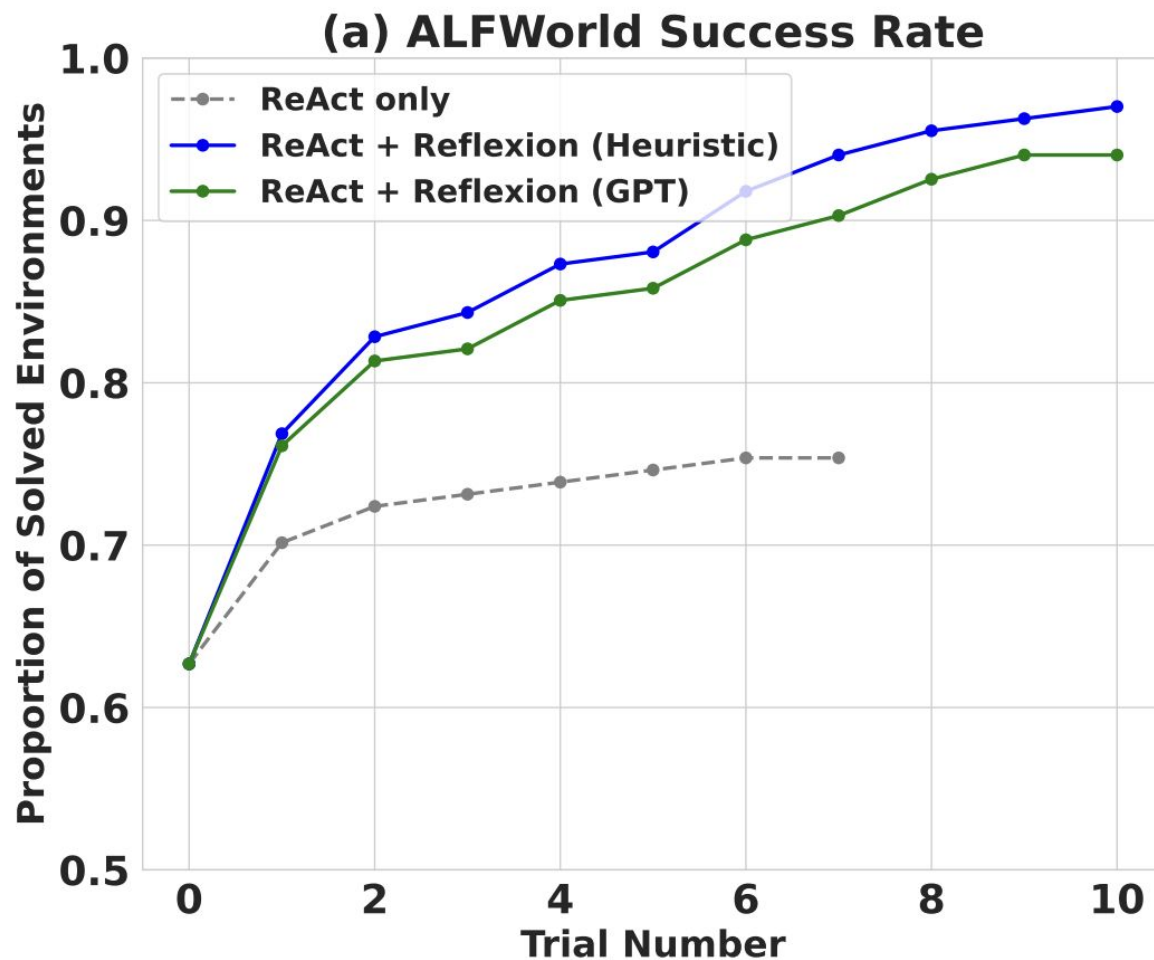
> think: To solve the task, I need to find and take a mug, then find and use a desklamp.

> use desklamp 1

You turn on the desklamp 1.

Status: Success

Self-reflection



Shinn, Noah, et al. "Reflexion: Language agents with verbal reinforcement learning." *Advances in Neural Information Processing Systems* 36 (2024).

Self-reflection: Degeneration-of-Thought

“Once the LLM-based agent has established confidence in its answers, it is unable to generate novel thoughts later through self-reflection even if the initial stance is incorrect.”

Multi-agent debate

Counter-Intuitive Question

The two circles are externally tangent and there is no relative sliding. The radius of circle A is 1/3 the radius of circle B. Circle A rolls around circle B one trip back to its starting point. How many times will circle A revolve in total?

Chain-of-Thought

Let the radius of circle A be r and the radius of circle B be $3r$. The distance circle A travels along the circumference of circle B is equal to the circumference of circle B, which is $6\pi r$. The number of revolutions = $(6\pi r) / (2\pi r) = 6\pi r / 2\pi r = 3$. So, circle A will revolve 3 times in total as it rolls around circle B once. ❌

Self-Reflection

Reflection:

- (1) I have a clear understanding of the problem.
- (2) I have used the correct formula for the circumference of a circle and applied it to both circles to find the number of revolutions.
- (3) I have correctly calculated the number of revolutions of circle A.

Based on the results of my self-reflection, I am confident that my answer is correct. Therefore, circle A will revolve 3 times in total as it rolls around circle B once. ❌

Multi-Agent Debate

I disagree with you. To find the total number of revolutions, we need to consider both the rotation around circle B and the rotation of circle A itself. Therefore, circle A will revolve 3 times around its own center and 1 time around circle B, making a total of 4 revolutions.

I see your point, but...

That's a valid point, however...

The negative side correctly considers both the rotation of circle A around its own center and its rotation around circle B, while the affirmative side only considers the rotation around circle B. Therefore, the answer is 4. ✅

Multi-agent debate: Prompts

Meta prompt: *“You are a debater. Hello and welcome to the debate competition. It’s not necessary to fully agree with each other’s perspectives, as our objective is to find the correct answer. The debate topic is stated as follows: <debate topic>.”*

Affirmative Debater: *“You are affirmative side. Please express your viewpoints.”*

Negative Debater: *“You are negative side. You disagree with the affirmative side’s points. Provide your reasons and answer.”*

Judge: *“You are a moderator. There will be two debaters involved in a debate competition. They will present their answers and discuss their perspectives on the <debate topic>. At the end of each round, you will evaluate both sides’ answers and decide which one is correct.”*

PAL: Program-aided Language Models

Chain-of-Thought (Wei et al., 2022)

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 tennis balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The bakers at the Beverly Hills Bakery baked 200 loaves of bread on Monday morning. They sold 93 loaves in the morning and 39 loaves in the afternoon. A grocery store returned 6 unsold loaves. How many loaves of bread did they have left?

Program-aided Language models (this work)

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 tennis balls.

`tennis_balls = 5`

2 cans of 3 tennis balls each is

`bought_balls = 2 * 3`

tennis balls. The answer is

`answer = tennis_balls + bought_balls`

Q: The bakers at the Beverly Hills Bakery baked 200 loaves of bread on Monday morning. They sold 93 loaves in the morning and 39 loaves in the afternoon. A grocery store returned 6 unsold loaves. How many loaves of bread did they have left?

PAL: Program-aided Language Models

Chain-of-Thought (Wei et al., 2022)

Model Output

A: The bakers started with 200 loaves. They sold 93 in the morning and 39 in the afternoon. So they sold $93 + 39 = 132$ loaves. The grocery store returned 6 loaves. So they had $200 - 132 - 6 = 62$ loaves left.

The answer is 62.



Program-aided Language models (this work)

Model Output

A: The bakers started with 200 loaves

```
loaves_baked = 200
```

They sold 93 in the morning and 39 in the afternoon

```
loaves_sold_morning = 93
```

```
loaves_sold_afternoon = 39
```

The grocery store returned 6 loaves.

```
loaves_returned = 6
```

The answer is

```
answer = loaves_baked - loaves_sold_morning  
- loaves_sold_afternoon + loaves_returned
```

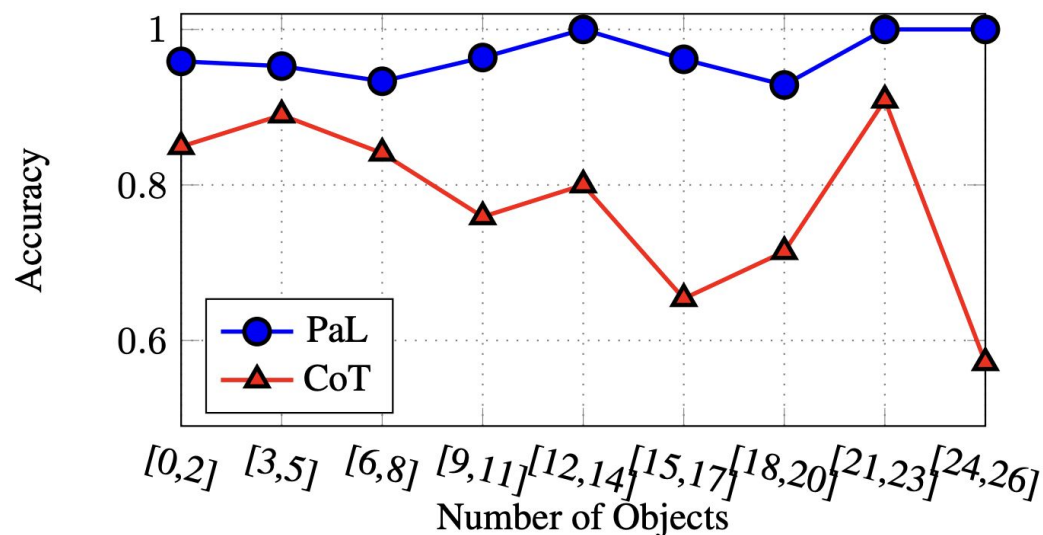
```
>>> print(answer)
```

```
74
```



PAL: Results

> COT generates nearly identical natural language “thoughts” for tasks with small and large numbers, indicating that the primary failure mode is the inability to perform arithmetic accurately.



	GSM8K	GSM-HARD	SVAMP	ASDIV	SINGLEEQ	SINGLEOP	ADDSUB	MULTIARITH
DIRECT _{Codex}	19.7	5.0	69.9	74.0	86.8	93.1	90.9	44.0
COT _{UL2-20B}	4.1	-	12.6	16.9	-	-	18.2	10.7
COT _{LaMDA-137B}	17.1	-	39.9	49.0	-	-	52.9	51.8
COT _{Codex}	65.6	23.1	74.8	76.9	89.1	91.9	86.0	95.9
COT _{PaLM-540B}	56.9	-	79.0	73.9	92.3	94.1	91.9	94.7
COT _{Minerva 540B}	58.8	-	-	-	-	-	-	-
PAL	72.0	61.2	79.4	79.6	96.1	94.6	92.5	99.2

Toolformer: Teach LLMs to use tools

Question answering system

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Calculator

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

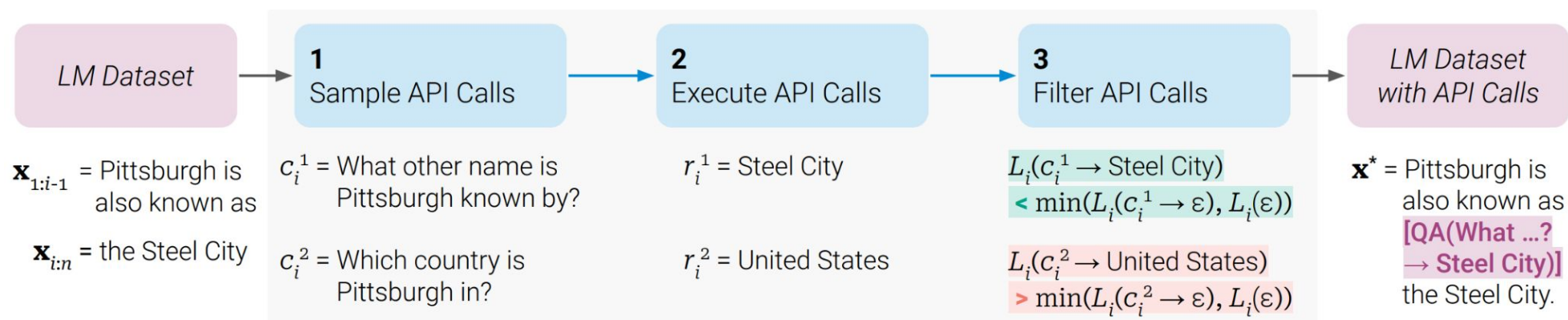
Machine translation system

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

Wikipedia search engine

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

Toolformer: Teach LLMs to use tools



Conclusions

1. Model performance can scale with inference time compute
2. Extended reasoning context is helpful
3. Feedback from the environment is helpful