Prompt Engineering a Prompt Engineer

Qinyuan Ye, Maxamed Axmed, Reid Pryzant, Fereshte Khani

ACL 2024 (Findings)

USC University of Southern California

Microsoft

ACL 2024
Bangkok, Thailand

Prompting LLMs is hard!

LLMs are sensitive to the prompts.

(sometimes in unexpected ways)

😐 Summarize the news article.

✅ Please summarize the news article for me.

✅ I'm going to tip $200 for a perfect solution!

✅ Take a deep breath …

Non-AI experts struggle to write prompts.

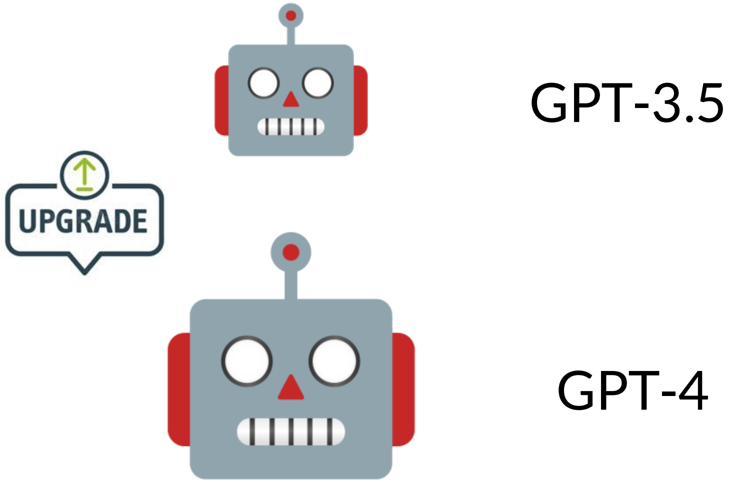
🧙 🎅 🧑‍🏫 🧑‍⚖️ 🧑 🧑‍🔬

Prompting LLMs is hard!

LLM-powered services in production

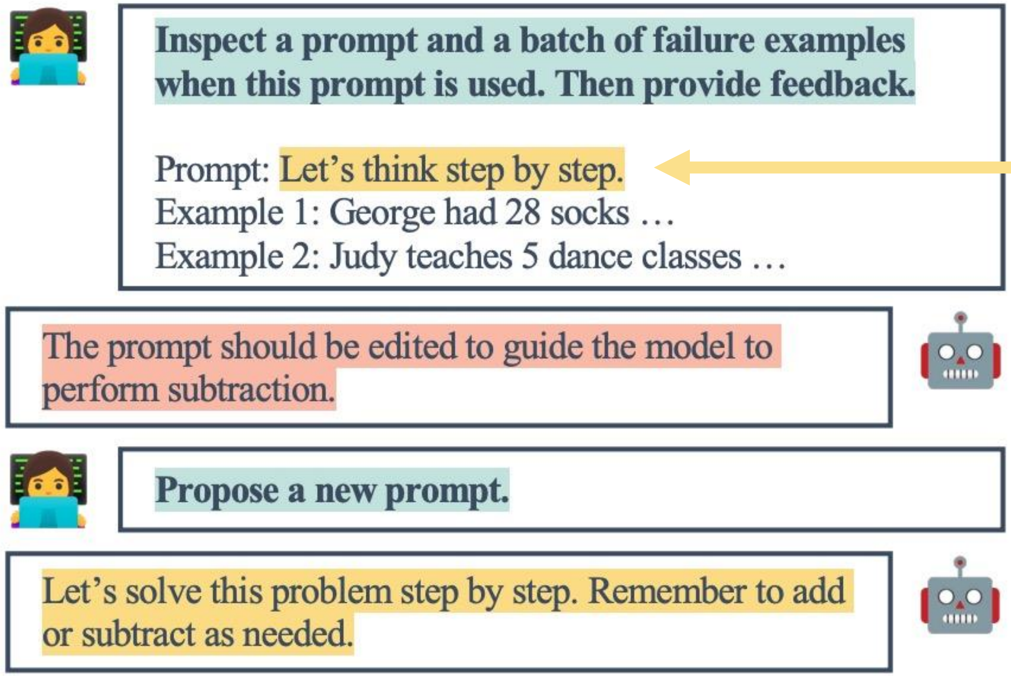
Edge cases arise and need to be fixed. 🤖❌ 🧑‍🔧

LLMs are upgraded and
the old prompt no longer works.



Prompting LLMs is hard!

LLM-powered Automatic Prompt Engineering comes to rescue!



Inspect a prompt and a batch of failure examples when this prompt is used. Then provide feedback.

Prompt: Let's think step by step.
Example 1: George had 28 socks …
Example 2: Judy teaches 5 dance classes …

The prompt should be edited to guide the model to perform subtraction.

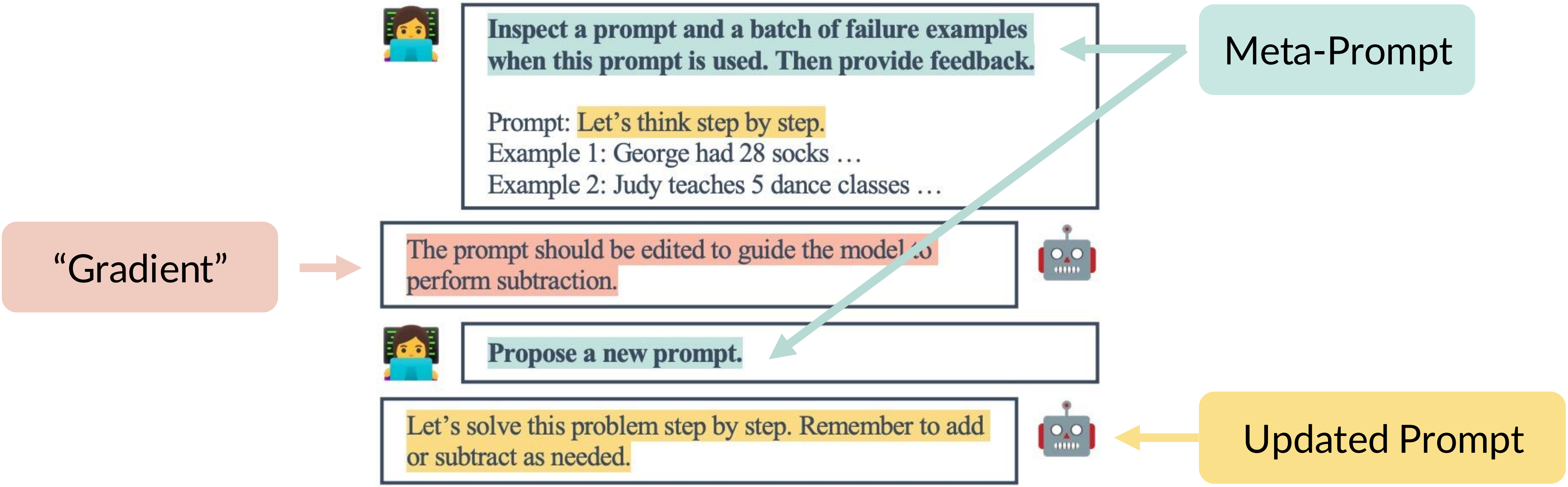
Propose a new prompt.

Let's solve this problem step by step. Remember to add or subtract as needed.

Initial Prompt

Prompting LLMs is hard!

LLM-powered Automatic Prompt Engineering comes to rescue!



Meta-Prompt

Inspect a prompt and a batch of failure examples when this prompt is used. Then provide feedback.

Prompt: Let's think step by step.
Example 1: George had 28 socks …
Example 2: Judy teaches 5 dance classes …

"Gradient"

The prompt should be edited to guide the model to perform subtraction.

Propose a new prompt.

Updated Prompt

Let's solve this problem step by step. Remember to add or subtract as needed.

Prompting LLMs is hard!

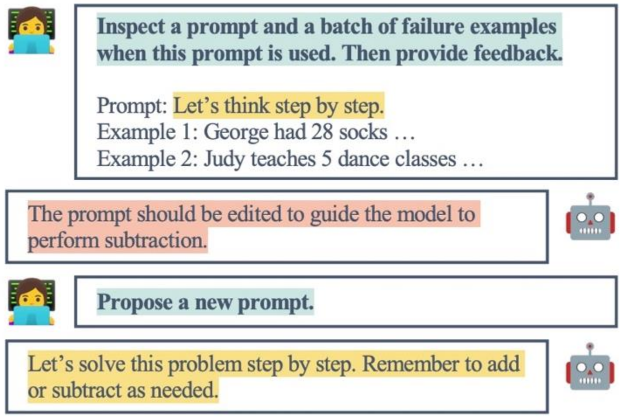
LLM-powered Automatic Prompt Engineering comes to rescue!

Step 1
Prompt Initialization

Let’s think step by step.

Step 2
New Prompt Proposal

🔁



Inspect a prompt and a batch of failure examples when this prompt is used. Then provide feedback.

Prompt: Let’s think step by step.
Example 1: George had 28 socks …
Example 2: Judy teaches 5 dance classes …

The prompt should be edited to guide the model to perform subtraction.

Propose a new prompt.

Let’s solve this problem step by step. Remember to add or subtract as needed.

Step 3
Filtering

Let’s think step by step. ❌

Let’s solve the problem step by step. ✅

Take a deep breath and think step-by-step. ✅

🤔 What makes a good ***meta-prompt***?

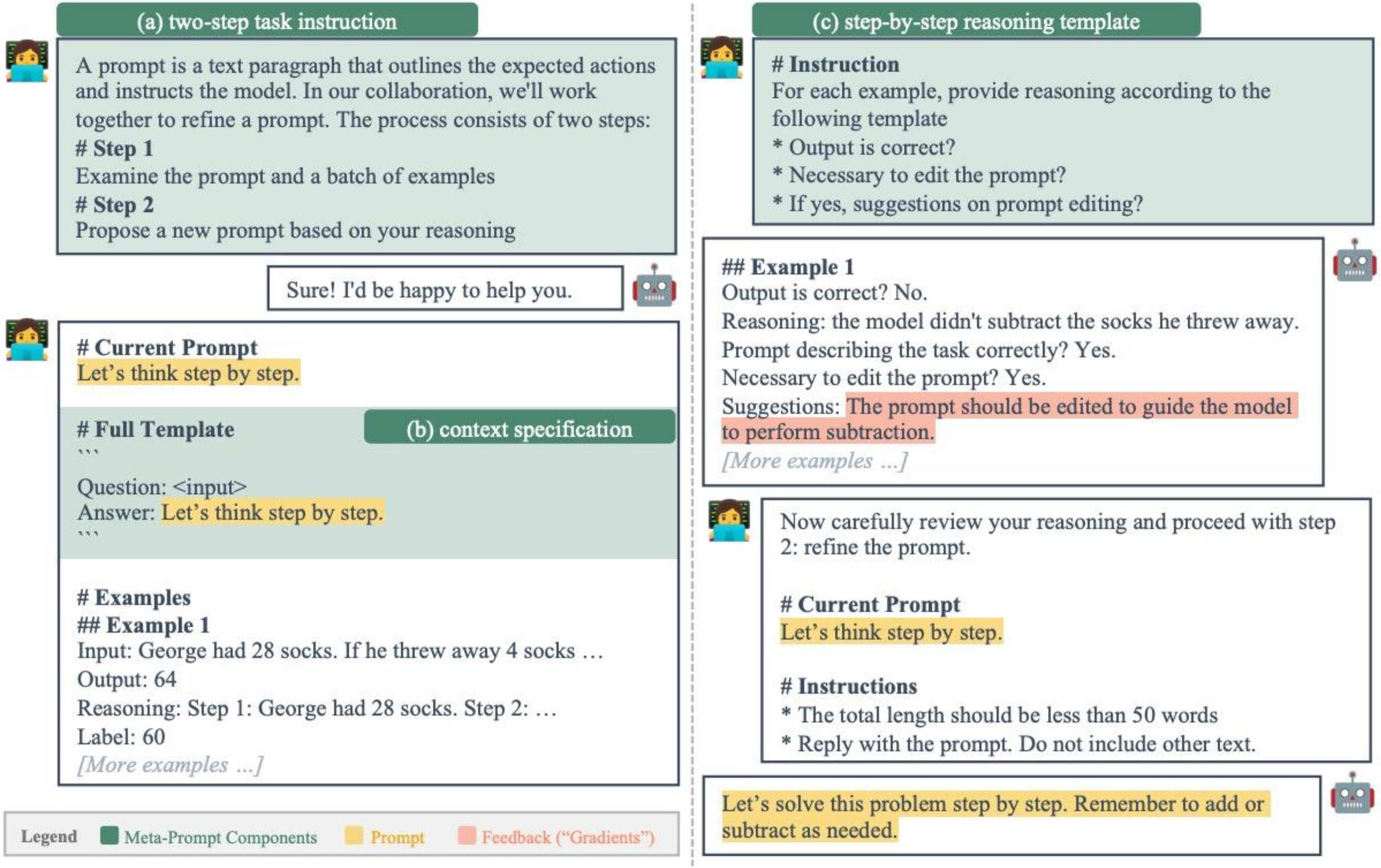
Prompt Engineering a Prompt Engineer

We investigate **what makes a good *meta-prompt*** in LLM-powered automatic prompt engineering.

We develop **PE2**, a strong automatic prompt engineer featuring **three meta-prompt components**.

(a) a two-step task instruction;
(b) context specification;
(c) a step-by-step reasoning template.

Prompt Engineering a Prompt Engineer



(a) two-step task instruction

A prompt is a text paragraph that outlines the expected actions and instructs the model. In our collaboration, we'll work together to refine a prompt. The process consists of two steps:
# Step 1
Examine the prompt and a batch of examples
# Step 2
Propose a new prompt based on your reasoning

Sure! I'd be happy to help you.

# Current Prompt
Let's think step by step.

# Full Template

(b) context specification

```
Question: <input>
Answer: Let's think step by step.
```

# Examples
## Example 1
Input: George had 28 socks. If he threw away 4 socks …
Output: 64
Reasoning: Step 1: George had 28 socks. Step 2: …
Label: 60
[More examples ...]

Legend 🟩 Meta-Prompt Components 🟨 Prompt 🟥 Feedback ("Gradients")

(c) step-by-step reasoning template

# Instruction
For each example, provide reasoning according to the following template
* Output is correct?
* Necessary to edit the prompt?
* If yes, suggestions on prompt editing?

## Example 1
Output is correct? No.
Reasoning: the model didn't subtract the socks he threw away.
Prompt describing the task correctly? Yes.
Necessary to edit the prompt? Yes.
Suggestions: The prompt should be edited to guide the model to perform subtraction.
[More examples ...]

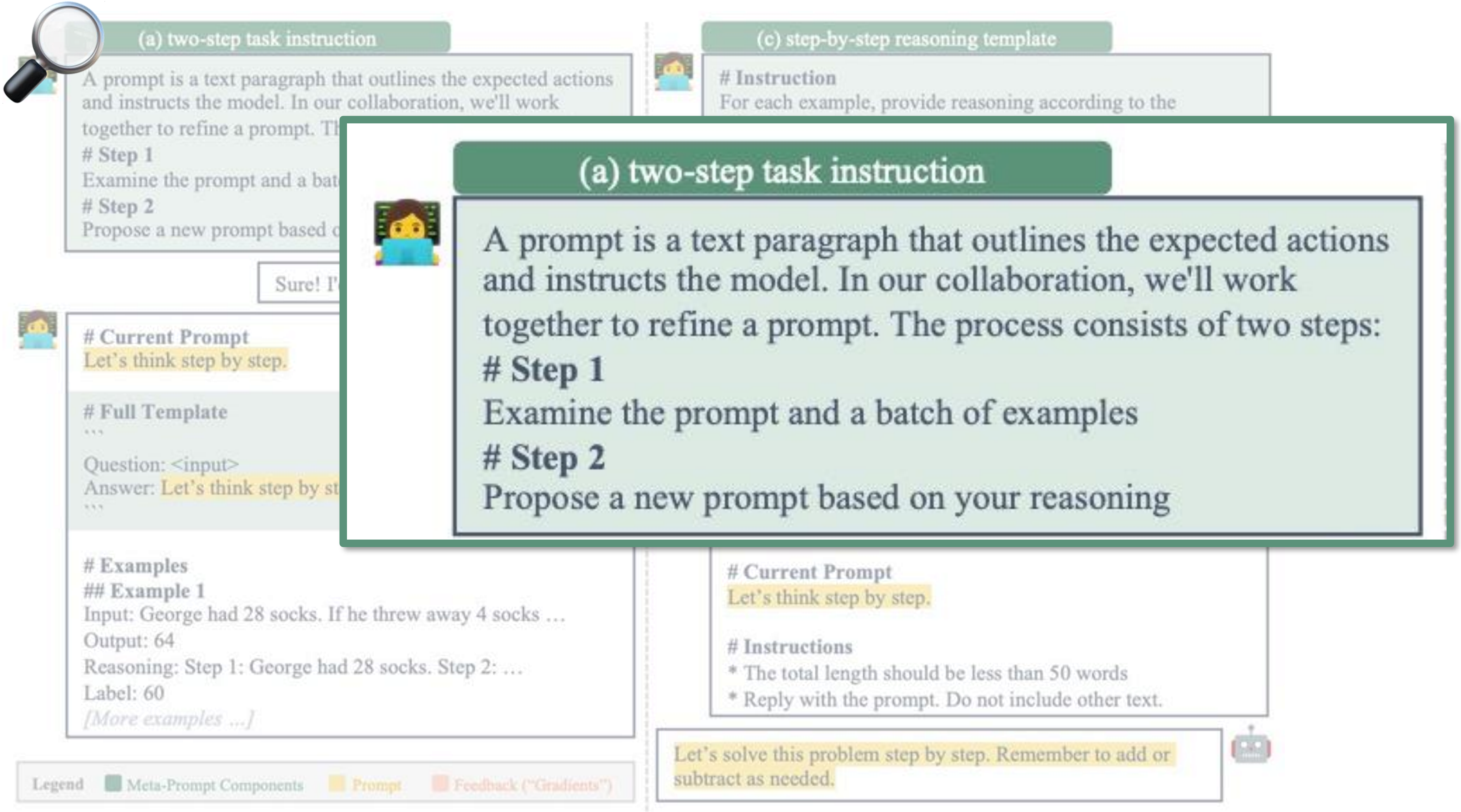
Now carefully review your reasoning and proceed with step 2: refine the prompt.

# Current Prompt
Let's think step by step.

# Instructions
* The total length should be less than 50 words
* Reply with the prompt. Do not include other text.

Let's solve this problem step by step. Remember to add or subtract as needed.

Prompt Engineering a Prompt Engineer



(a) two-step task instruction

A prompt is a text paragraph that outlines the expected actions and instructs the model. In our collaboration, we'll work together to refine a prompt. The process consists of two steps:
Step 1
Examine the prompt and a batch of examples
Step 2
Propose a new prompt based on your reasoning

Sure! I'

Current Prompt
Let's think step by step.

Full Template
```
Question: <input>
Answer: Let's think step by st
```

Examples
Example 1
Input: George had 28 socks. If he threw away 4 socks …
Output: 64
Reasoning: Step 1: George had 28 socks. Step 2: …
Label: 60
[More examples …]

Legend 🟩 Meta-Prompt Components 🟨 Prompt 🟥 Feedback ("Gradients")

(c) step-by-step reasoning template

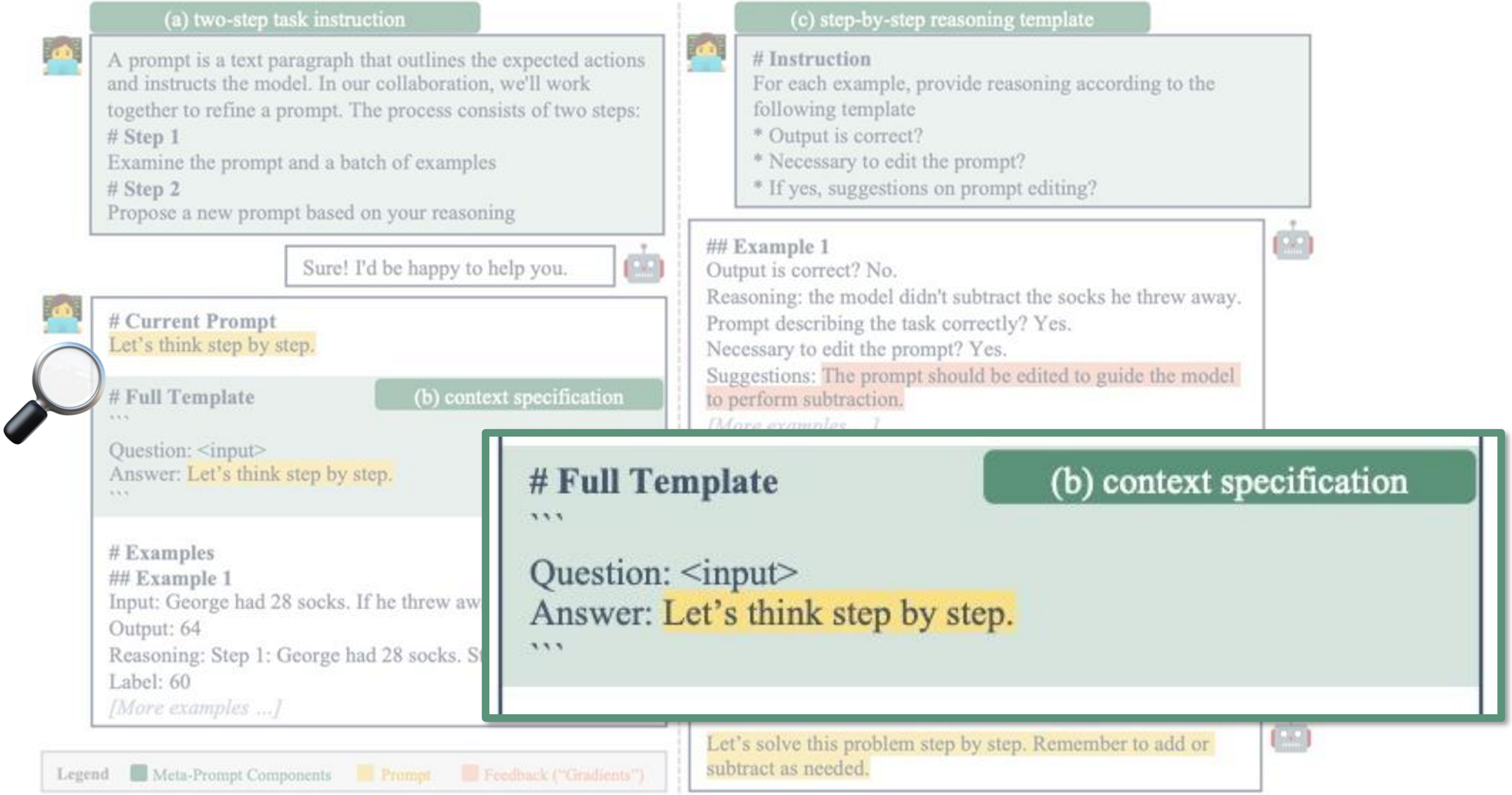
Instruction
For each example, provide reasoning according to the

Current Prompt
Let's think step by step.

Instructions
* The total length should be less than 50 words
* Reply with the prompt. Do not include other text.

Let's solve this problem step by step. Remember to add or subtract as needed.

Prompt Engineering a Prompt Engineer



(a) two-step task instruction

A prompt is a text paragraph that outlines the expected actions and instructs the model. In our collaboration, we'll work together to refine a prompt. The process consists of two steps:
# Step 1
Examine the prompt and a batch of examples
# Step 2
Propose a new prompt based on your reasoning

Sure! I'd be happy to help you.

# Current Prompt
Let's think step by step.

# Full Template (b) context specification
```
Question: <input>
Answer: Let's think step by step.
```

# Examples
## Example 1
Input: George had 28 socks. If he threw aw
Output: 64
Reasoning: Step 1: George had 28 socks. St
Label: 60
[More examples ...]

(c) step-by-step reasoning template

# Instruction
For each example, provide reasoning according to the following template
* Output is correct?
* Necessary to edit the prompt?
* If yes, suggestions on prompt editing?

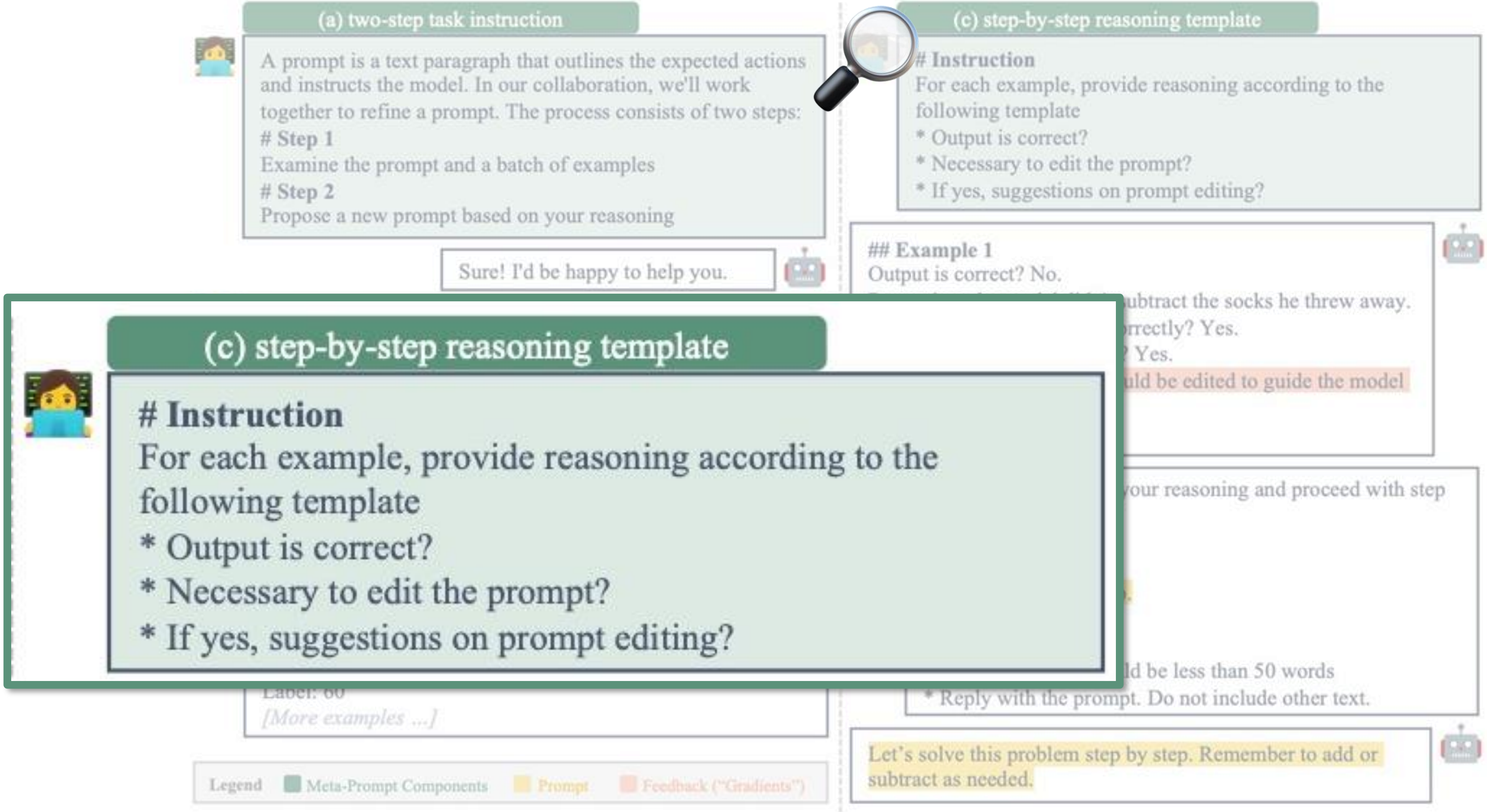
## Example 1
Output is correct? No.
Reasoning: the model didn't subtract the socks he threw away.
Prompt describing the task correctly? Yes.
Necessary to edit the prompt? Yes.
Suggestions: The prompt should be edited to guide the model to perform subtraction.
[More examples ...]

Let's solve this problem step by step. Remember to add or subtract as needed.

Legend 🟩 Meta-Prompt Components 🟨 Prompt 🟥 Feedback ("Gradients")

# Full Template (b) context specification
```
Question: <input>
Answer: Let's think step by step.
```

Prompt Engineering a Prompt Engineer



(a) two-step task instruction

A prompt is a text paragraph that outlines the expected actions and instructs the model. In our collaboration, we'll work together to refine a prompt. The process consists of two steps:
# Step 1
Examine the prompt and a batch of examples
# Step 2
Propose a new prompt based on your reasoning

Sure! I'd be happy to help you.

Label: 60
[More examples ...]

Legend 🟩 Meta-Prompt Components 🟨 Prompt 🟥 Feedback ("Gradients")

(c) step-by-step reasoning template

# Instruction
For each example, provide reasoning according to the following template
* Output is correct?
* Necessary to edit the prompt?
* If yes, suggestions on prompt editing?

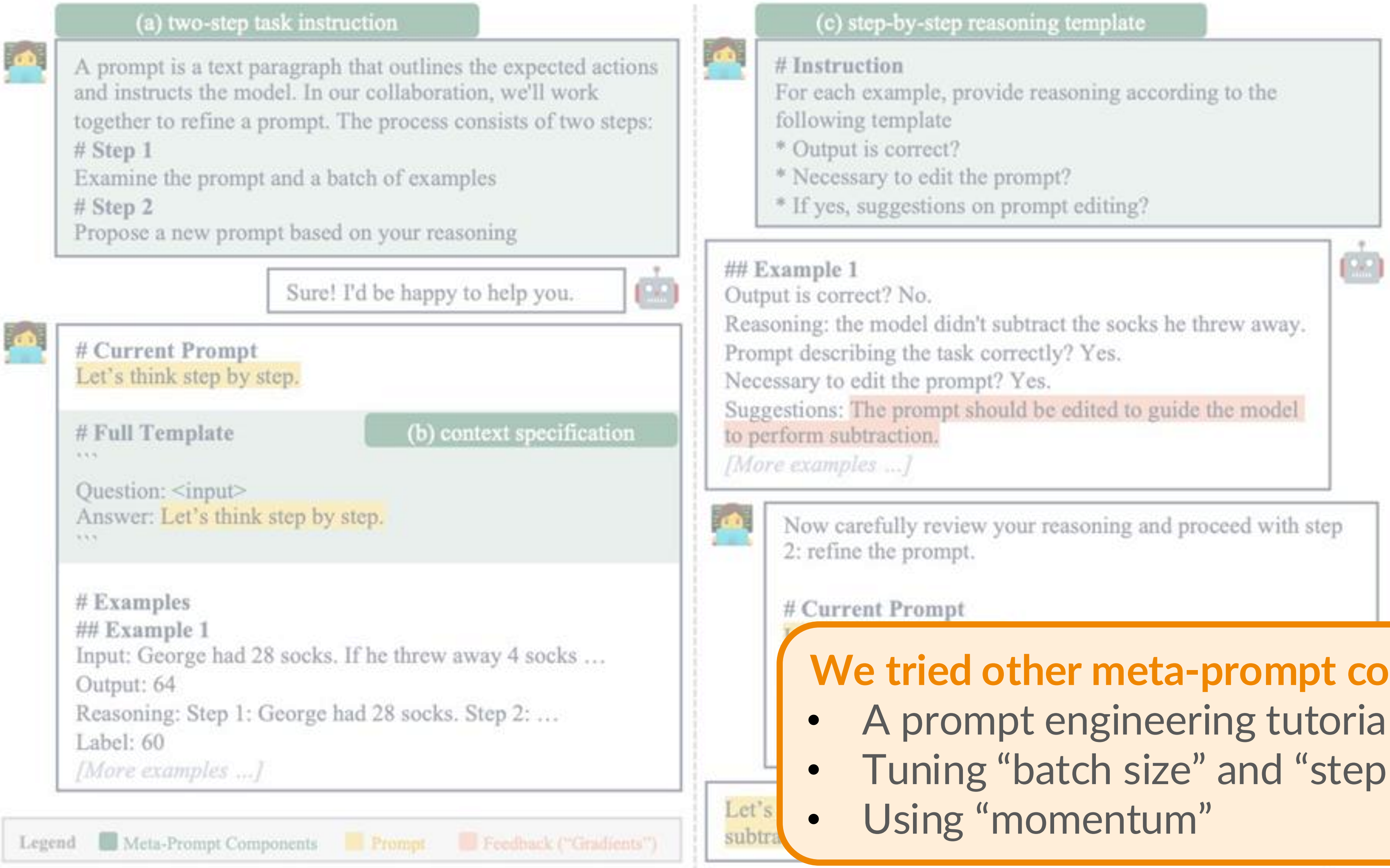
## Example 1
Output is correct? No.
ubtract the socks he threw away.
rrectly? Yes.
? Yes.
uld be edited to guide the model

your reasoning and proceed with step

ld be less than 50 words
* Reply with the prompt. Do not include other text.

Let's solve this problem step by step. Remember to add or subtract as needed.

Prompt Engineering a Prompt Engineer



(a) two-step task instruction

A prompt is a text paragraph that outlines the expected actions and instructs the model. In our collaboration, we'll work together to refine a prompt. The process consists of two steps:
# Step 1
Examine the prompt and a batch of examples
# Step 2
Propose a new prompt based on your reasoning

Sure! I'd be happy to help you.

# Current Prompt
Let's think step by step.

# Full Template

(b) context specification

```
Question: <input>
Answer: Let's think step by step.
```

# Examples
## Example 1
Input: George had 28 socks. If he threw away 4 socks …
Output: 64
Reasoning: Step 1: George had 28 socks. Step 2: …
Label: 60
[More examples …]

(c) step-by-step reasoning template

# Instruction
For each example, provide reasoning according to the following template
* Output is correct?
* Necessary to edit the prompt?
* If yes, suggestions on prompt editing?

## Example 1
Output is correct? No.
Reasoning: the model didn't subtract the socks he threw away.
Prompt describing the task correctly? Yes.
Necessary to edit the prompt? Yes.
Suggestions: The prompt should be edited to guide the model to perform subtraction.
[More examples …]

Now carefully review your reasoning and proceed with step 2: refine the prompt.

# Current Prompt

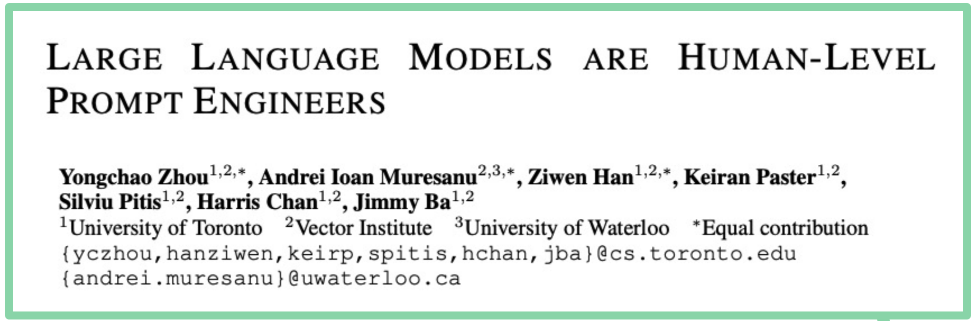
Let's
subtra

Legend 🟩 Meta-Prompt Components 🟨 Prompt 🟥 Feedback ("Gradients")

We tried other meta-prompt components!

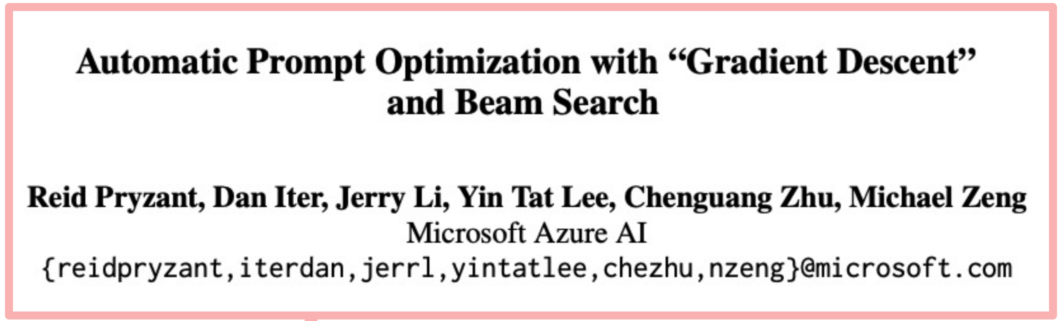
- A prompt engineering tutorial
- Tuning "batch size" and "step size" 😐
- Using "momentum"

PE2 achieves strong empirical performance



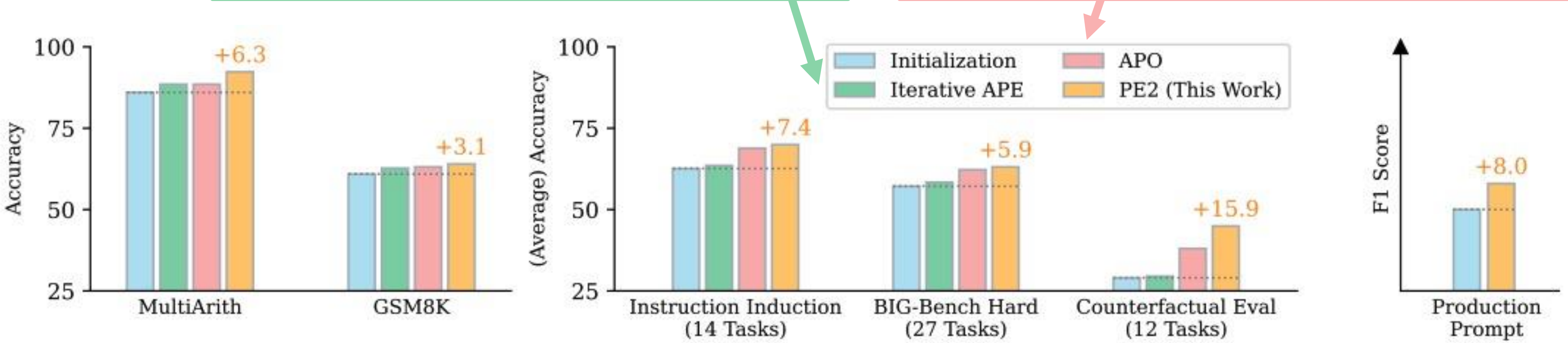
Large Language Models are Human-Level Prompt Engineers

Yongchao Zhou[1,2,*], Andrei Ioan Muresanu[2,3,*], Ziwen Han[1,2,*], Keiran Paster[1,2], Silviu Pitis[1,2], Harris Chan[1,2], Jimmy Ba[1,2]
[1]University of Toronto [2]Vector Institute [3]University of Waterloo *Equal contribution
{yczhou,hanziwen,keirp,spitis,hchan,jba}@cs.toronto.edu
{andrei.muresanu}@uwaterloo.ca

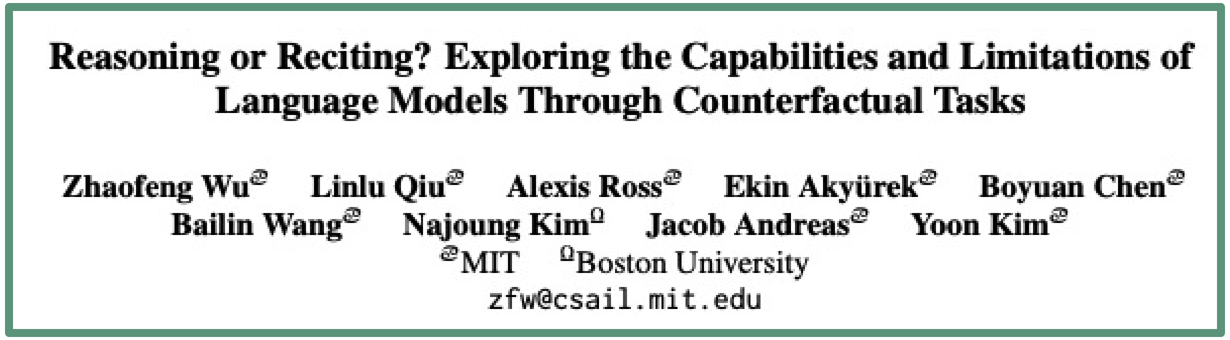


Automatic Prompt Optimization with "Gradient Descent" and Beam Search

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, Michael Zeng
Microsoft Azure AI
{reidpryzant,iterdan,jerrl,yintatlee,chezhu,nzeng}@microsoft.com

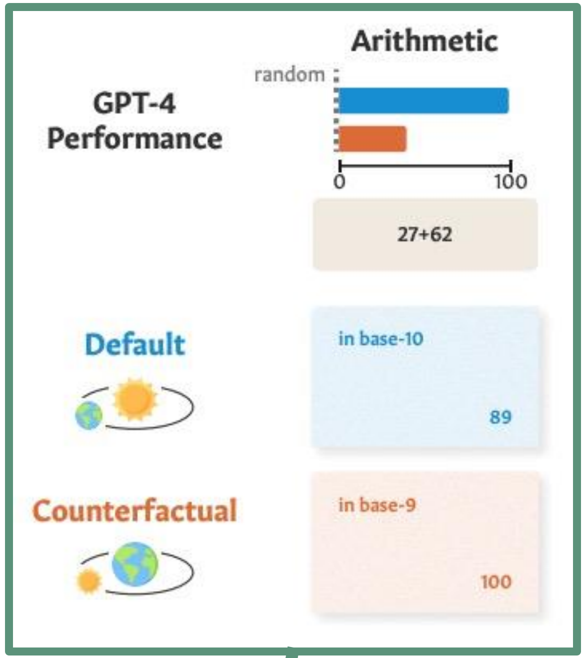


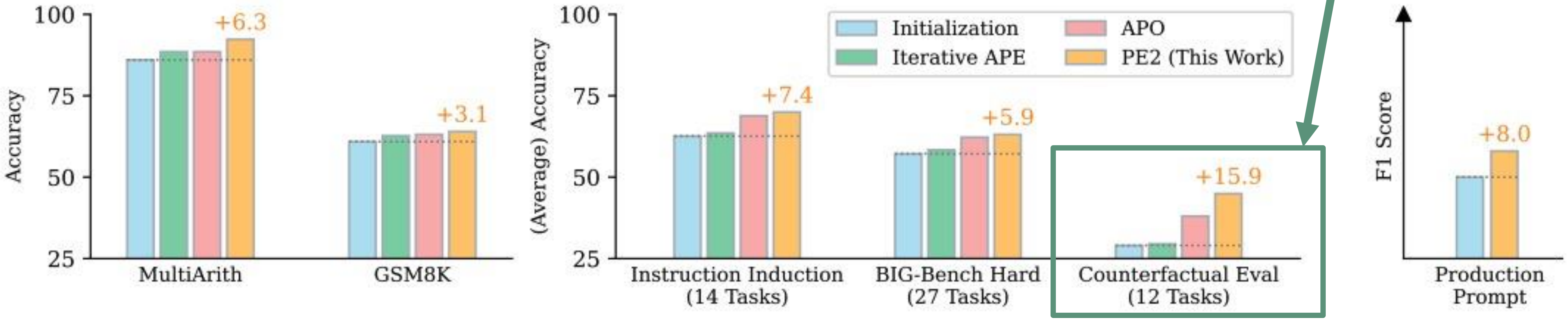
PE2 achieves strong empirical performance



Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks

Zhaofeng Wu[@] Linlu Qiu[@] Alexis Ross[@] Ekin Akyürek[@] Boyuan Chen[@]
Bailin Wang[@] Najoung Kim[Ω] Jacob Andreas[@] Yoon Kim[@]
[@]MIT [Ω]Boston University
zfw@csail.mit.edu

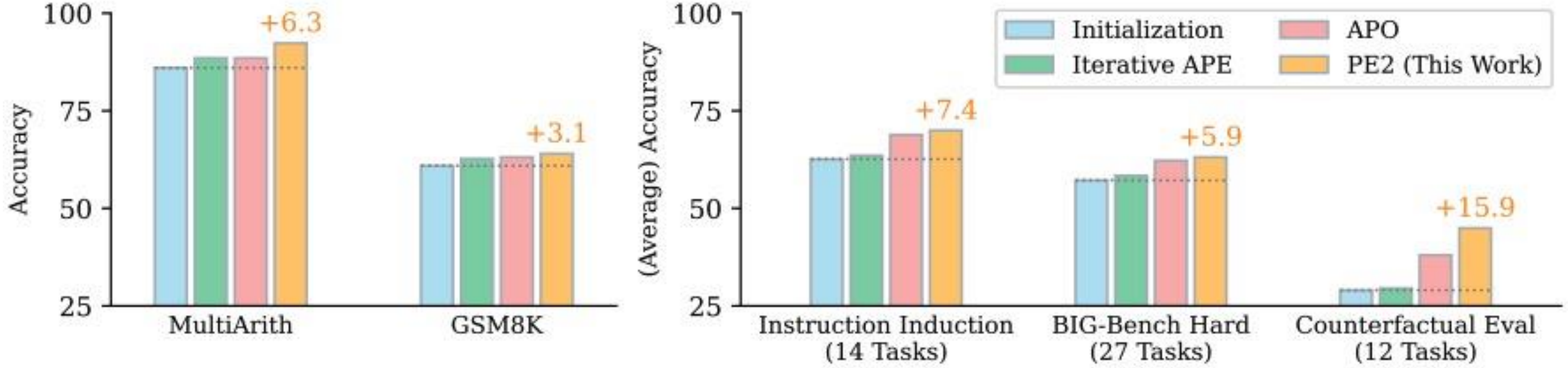


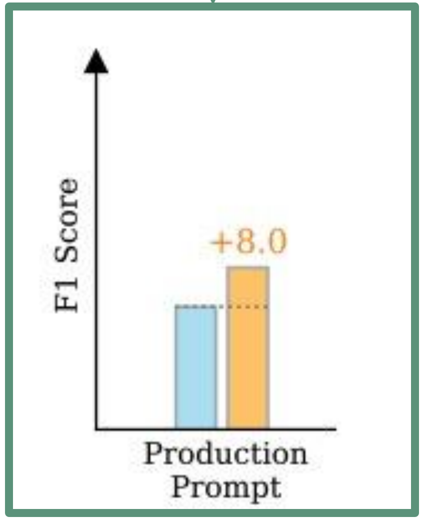


PE2 achieves strong empirical performance

A long production prompt!

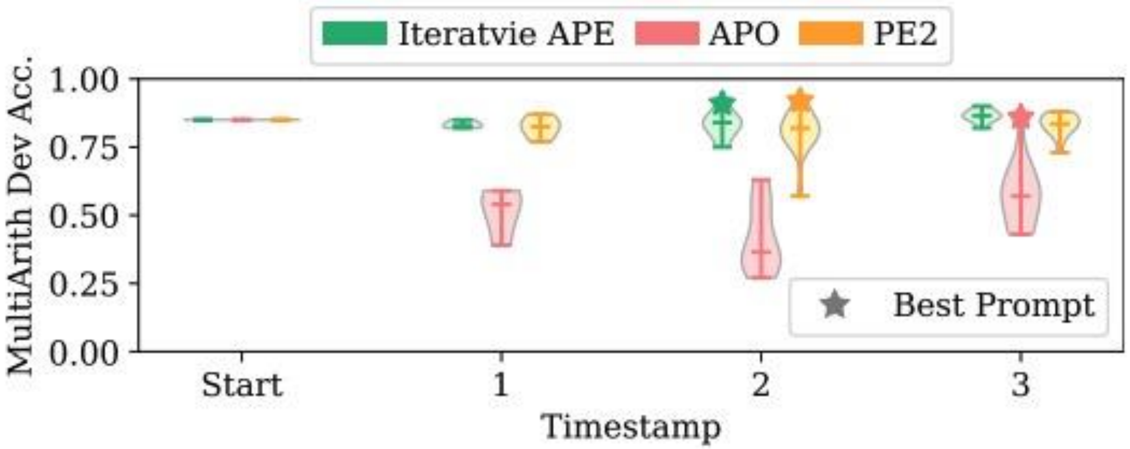
5k+ tokens, written by experts



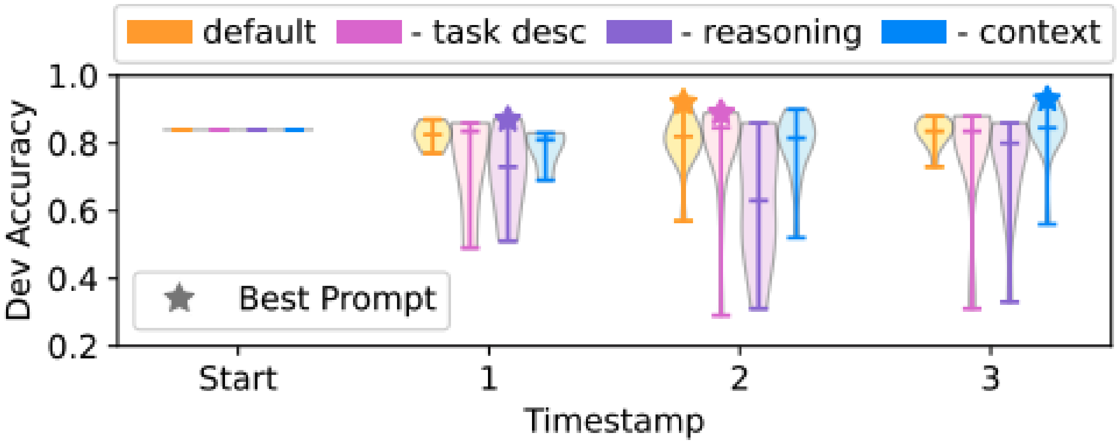


Prompt Optimization Dynamics

Quality (Dev. Acc.) of newly-proposed prompt at each timestamp.



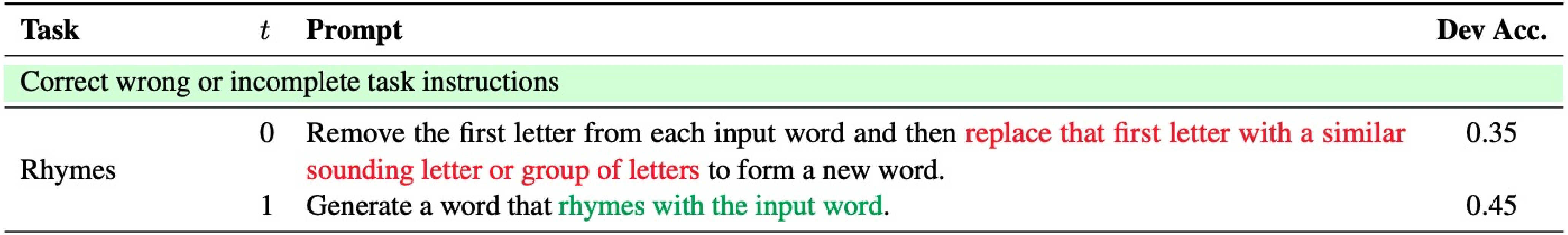
PE2 has a better balance of exploration and stability.



All three meta-prompt components are critical for the optimization stability.

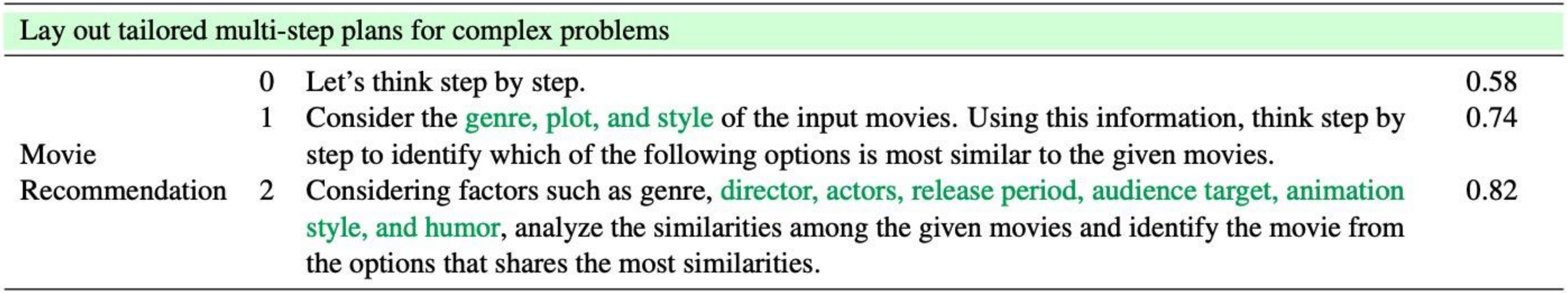
Case Study

✅ PE2 makes meaningful and targeted prompt edits.



Task	t	Prompt	Dev Acc.
Correct wrong or incomplete task instructions			
Rhymes	0	Remove the first letter from each input word and then replace that first letter with a similar sounding letter or group of letters to form a new word.	0.35
	1	Generate a word that rhymes with the input word.	0.45

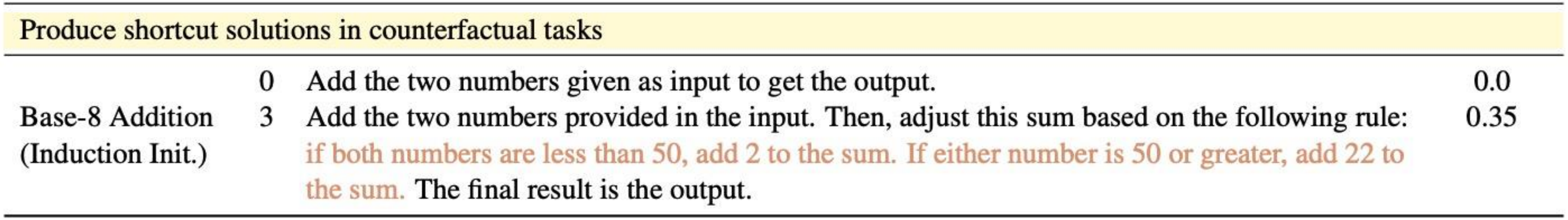
✅ PE2 can induce multi-step plans for complex tasks.



Lay out tailored multi-step plans for complex problems			
Movie Recommendation	0	Let's think step by step.	0.58
	1	Consider the genre, plot, and style of the input movies. Using this information, think step by step to identify which of the following options is most similar to the given movies.	0.74
	2	Considering factors such as genre, director, actors, release period, audience target, animation style, and humor, analyze the similarities among the given movies and identify the movie from the options that shares the most similarities.	0.82

Case Study

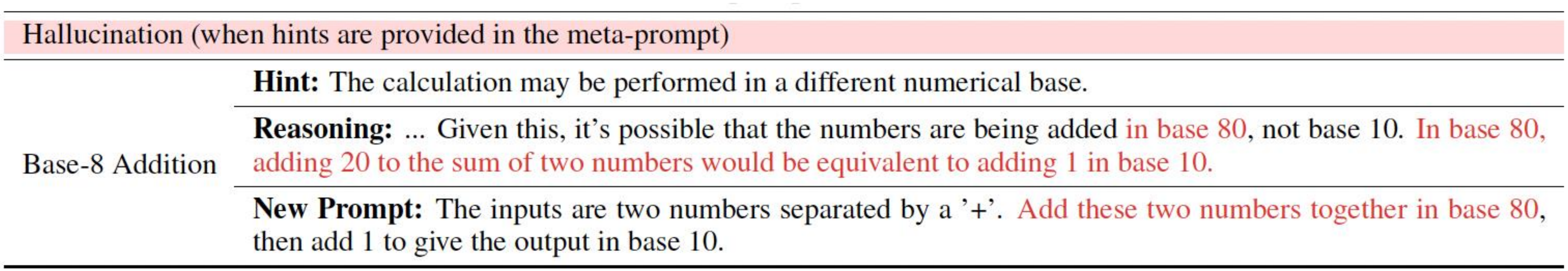
🤔 **PE2 can reason about non-standard situations, but it generates “shortcut” solutions.**



Produce shortcut solutions in counterfactual tasks			
Base-8 Addition (Induction Init.)	0	Add the two numbers given as input to get the output.	0.0
	3	Add the two numbers provided in the input. Then, adjust this sum based on the following rule: if both numbers are less than 50, add 2 to the sum. If either number is 50 or greater, add 22 to the sum. The final result is the output.	0.35

Case Study

⁉ **PE2 sometimes ignores instructions and hallucinates.**



Hallucination (when hints are provided in the meta-prompt)	
Base-8 Addition	**Hint:** The calculation may be performed in a different numerical base.
	Reasoning: ... Given this, it's possible that the numbers are being added in base 80, not base 10. In base 80, adding 20 to the sum of two numbers would be equivalent to adding 1 in base 10.
	New Prompt: The inputs are two numbers separated by a '+'. Add these two numbers together in base 80, then add 1 to give the output in base 10.

Additional Analysis

- **Effect of Prompt Initialization**
  - Initialization matters; PE2 is able to recover from bad initializations
- **Effect of Task Format**
  - Effectiveness is dependent on the task format (generative/multi-choice/...)
- **Cross-model generalization of optimized prompts**
  - Automatic prompt engineering methods are model-agnostic.
  - But the optimized prompts are model-specific.

Check out our paper for more details!!
https://arxiv.org/abs/2311.05661

Summary

🔍 We investigate **what makes a good *meta-prompt*** in LLM-powered automatic prompt engineering.

🤖 We develop **PE2**, a strong automatic prompt engineer featuring **three meta-prompt components**.

✨ We show that **PE2** can

(1) makes ***targeted*** and ***highly specific*** prompt edits;
(2) induce ***multi-step plans*** for complex tasks;
(3) reason and adapt in ***non-standard situations***.