Chain-of-thought: Extensions and Analysis

- Just like standard in-context learning, we can ask what properties of the training examples are effective for chain-of-thought

- Extensions: how can we make chain-of-thought even more effective?

What makes explanations effective?

- Do LMs “follow” explanations?
- We can check both perturbing the “computation trace” (blue) and the natural language expression of that computation (green)

Question

Take the last letters of the words in "Bill Gates” and concatenate them.

Gold Explanation

Trace **NL**

The last letter of "Bill" is letter"l". The last of "Gates" is "s". Concatenating “l" and "s" is “ls". So the answer is ls.

Perturbing Trace

The last letter of "Bill" is letter " ". The last of "Gates" is " ". Concatenating “l" and "s" is “ls". So the answer is ls.

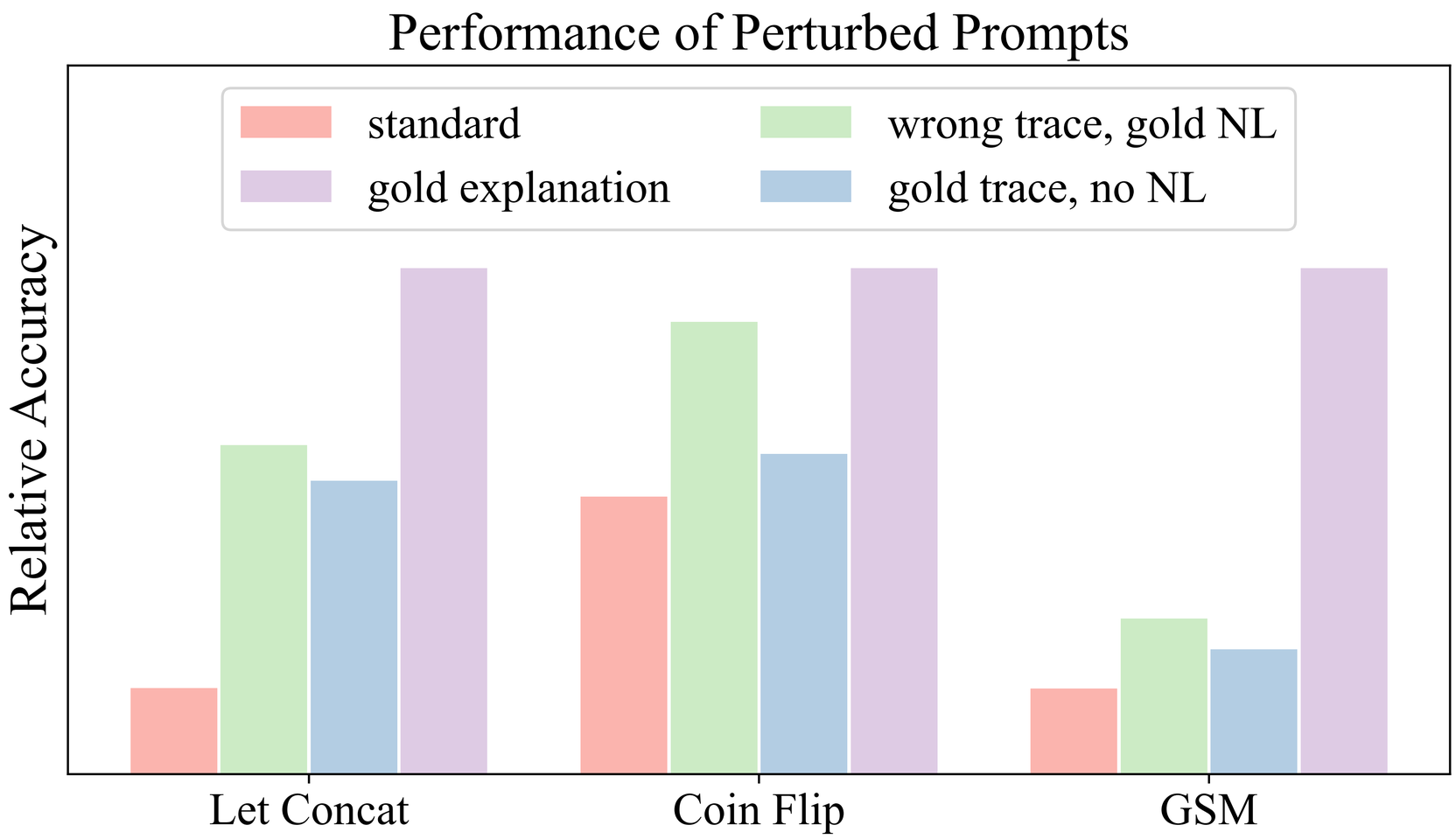
Perturbing NL

"Bill","l","Gates","s","l","s","ls". So the answer is ls.

Xi Ye et al. (2023)

What makes explanations effective?

- Perturbing either the NL or the computation trace causes reduced performance on three tasks. Both of these are important!



Xi Ye et al. (2023)

Extension: Program-aided LMs

- For math: why are we doing the arithmetic in the LLM itself?

- Instead: generate code fragments and actually execute them to get an answer (how most earlier math word problem systems worked)

- Many flavors of this: "Faithful Chain-of-thought", "Program-of-thought", Toolformer, etc.

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
```

Luyu Gao et al. (2022)

Extension: Self-ask

- Similar idea but with QA/a search engine in the loop
- Demonstration shows sub-questions and sub-answers, can potentially do search at these intermediate points
- Bing Chat / Google Bard can do this

GPT-3

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?
Are follow up questions needed here: Yes.
Follow up: How old was Theodor Haecker when he died?
Intermediate answer: Theodor Haecker was 65 years old when he died.
Follow up: How old was Harry Vaughan Watkins when he died?
Intermediate answer: Harry Vaughan Watkins was 69 years old when he died.
So the final answer is: Harry Vaughan Watkins

Question: Who was president of the U.S. when superconductivity was discovered?
Are follow up questions needed here: Yes.
Follow up: When was superconductivity discovered?
Intermediate answer: Superconductivity was discovered in 1911.
Follow up: Who was president of the U.S. in 1911?
Intermediate answer: William Howard Taft.
So the final answer is: William Howard Taft.

Ofir Press et al. (2022)

Frontiers

- Many efforts to integrate additional tools beyond programmatic execution (program-aided LMs) and search (self-ask):
  - ChatGPT “plugins”
  - Toolformer
- Future versions of these models will liely be even more tightly integrated with other capabilities

- Another line of work: verifying that chain-of-thought reasoning is correct. One baseline: ask an LLM to check its own work!