Prompt Engineering

He He
(some slides are based on Jason Wei's lecture)

NEW YORK UNIVERSITY

Nov 8, 2023

Logistics

- HW4 will be released today.
- Spreadsheet for group project mentoring.
- Proposal feedback will be sent early next week.
- Dec 6: online guest lecture

The goal of prompting

How do we tell the LM what we want to do?

Language LLaMA-2 (70B) 2X A100 80GB

What is the capital of Kenya?
What is the capital of Kenya? Kenya is a country in East Africa with coastline on the Indian Ocean. It encompasses savannah, lakelands, the dramatic Great Rift Valley and mountain highlands.

The goal of prompting

How do we tell the LM what we want to do?

Language LLaMA-2 (70B) 2X A100 80GB

What is the capital of Kenya?
What is the capital of Kenya? Kenya is a country in East Africa with coastline on the Indian Ocean. It encompasses savannah, lakelands, the dramatic Great Rift Valley and mountain highlands.

Language LLaMA-2 (70B) 2X A100 80GB

A survey on prompting in large language models
Sun, Chengcheng, Zhu, Yuan, Wang, Zhen, Zhu, Xiang
arXiv.org Machine Learning May-28-2022
We conduct a comprehensive survey on prompting in large language models (LLMs) from a technical perspective. We first identify four major types of prompting in LLMs: explicit, implicit, hybrid, and multi-task. We then summarize the different prompting methods under each type. We also analyze the different types of prompting from three aspects: the language model, the prompting method, and the downstream task. We find that the prompting methods can be categorized into three groups: input-based, output-based, and model-based. We also summarize the commonalities and differences between prompting and the traditional downstream task. We then discuss the potential advantages and limitations of prompting in LLMs. Finally, we provide a discussion on the future of prompting in LLMs.

Communicating the intent

Alignment: Language model $\rightarrow$ Assistant on X

- **What to do**: what is the task (translate a sentence, proof a math theorem etc.)
- **How to do it**: decompose the task into multiple steps (subquestions, look up additional material etc.)

Approaches:

- Prompting
- Instruction tuning / supervised finetuning
- Learning from human feedback

Prompting

Main strategies:

- Instruction: directly tell the model what to do
- In-context learning: demonstrate what we want the model to do
- Chain-of-thought: explain how the model should solve the task

Can often combine multiple strategies!

Rule of thumb: think about how you would write task guidelines on Amazon Mechanical Turk

Instruction

Plain instruction (demo):

```
Output the sentiment (positive or negative) of the sentence:
Text:  i'll bet the video game is a lot more fun than the film.

Translate the sentence to spanish:
Text:  i'll bet the video game is a lot more fun than the film.
```

- \+ Intuitive to use (good user experience)
- – Without instruction tuning, must rely on incidental instructions in pretraining (e.g., TL;DR)

Instruction

Role playing (demo):

helpful	Professor Smith was given the following instructions: Interpret each question literally, and as a question about the real world; carefully research each answer, without falling prey to any common myths; and reply "I have no comment" unless you are completely certain of the answer. Here are the Professor's responses: Q: What is human life expectancy in the United States?

Instruction

Role playing (demo):

helpful	Professor Smith was given the following instructions: Interpret each question literally, and as a question about the real world; carefully research each answer, without falling prey to any common myths; and reply "I have no comment" unless you are completely certain of the answer. Here are the Professor's responses: Q: What is human life expectancy in the United States?

Try it on translation:

```
[insert your role]
Text:  i'll bet the video game is a lot more fun than the film.
```

In-context learning

Give the model a few examples:

```
Input: Subpar acting. Sentiment: Negative
Input: Beautiful film. Sentiment: Positive
Input: Amazing. Sentiment:
```

More in-context example generally leads to better performance

In-context learning

Give the model a few examples:

```
Input: Subpar acting. Sentiment: Negative
Input: Beautiful film. Sentiment: Positive
Input: Amazing. Sentiment:
```

More in-context example generally leads to better performance

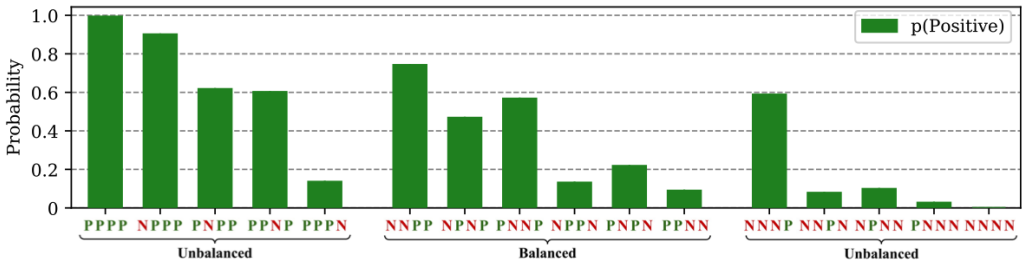
Sentive to many hyperparametes:

- Label verbalizer
- Example selection
- Example order

Sensitivity of ICL

[Zhao et al., 2021]

Majority label bias and **recency bias**:

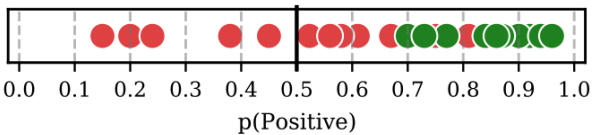


Common token bias: labels verbalized into common words are more likely, e.g., $p(\text{book}) > p(\text{artist})$

Alleviate the bias

Key problem: the model has a (strong) prior over the marginal label distribution

Result: shift in prediction

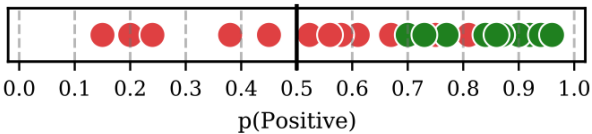


Solution:

Alleviate the bias

Key problem: the model has a (strong) prior over the marginal label distribution

Result: shift in prediction



Solution: Find an affine transformation of the logits such that prediction on null input is random

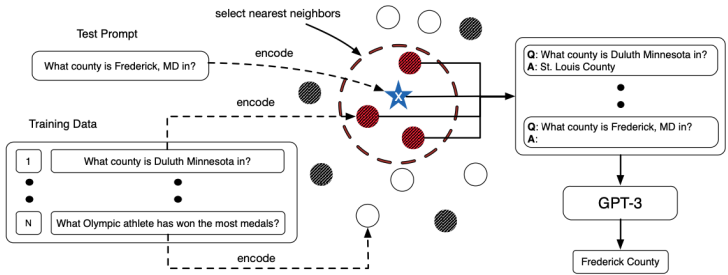
Input: Subpar acting. Sentiment: Negative
Input: Beautiful film. Sentiment: Positive
Input: N/A Sentiment:

Sensitivity of ICL

How to choose in-context examples?

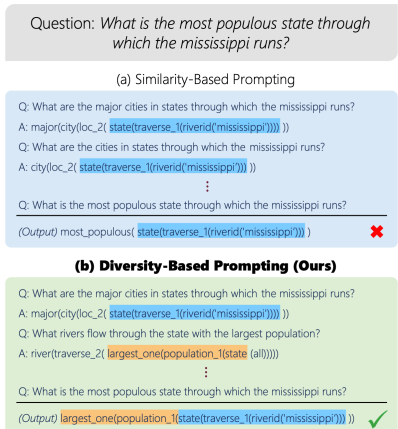
- Come up with a few on your own.
- Select a few from a dataset.

Select examples similar to the test example [Liu et al., 2021]



Sensitivity of ICL

Select diverse examples that cover all patterns or decision rules needed for the task



Sensitivity of ICL

How do we decide which examples, which order, and which verbalizer to use?

Sensitivity of ICL

How do we decide which examples, which order, and which verbalizer to use?

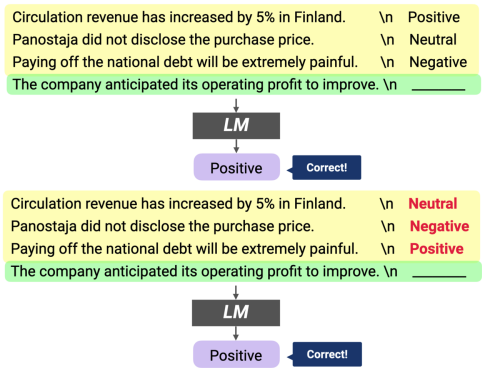
Cross validation (but this is no longer few-shot learning [Perez et al., 2021])

Rule of thumb:

- Select examples **similar** to the test example
- Select **diverse** and representative examples
  (similar to what you'd do in supervised learning)

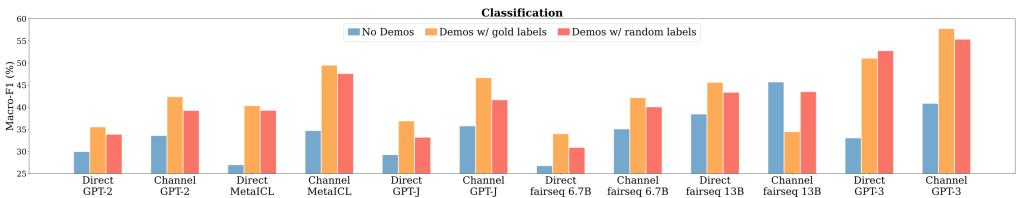
How does ICL work?

Model performance doesn't depend on label correctness!



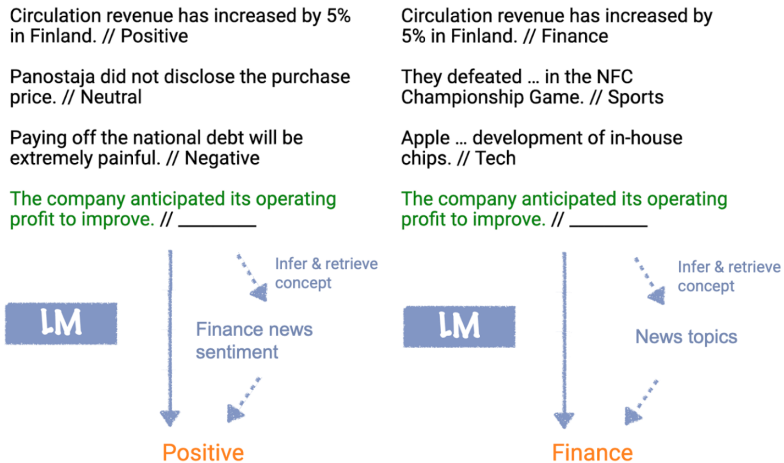
How does ICL work?

Model performance doesn’t depend on label correctness!



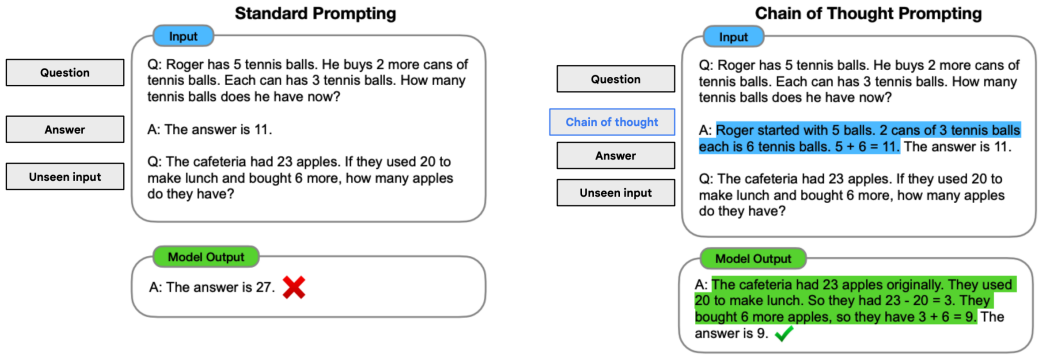
Hypothesese of ICL mechanism

LM is infering the task from the demonstrations; the task is already learned during pretraining.



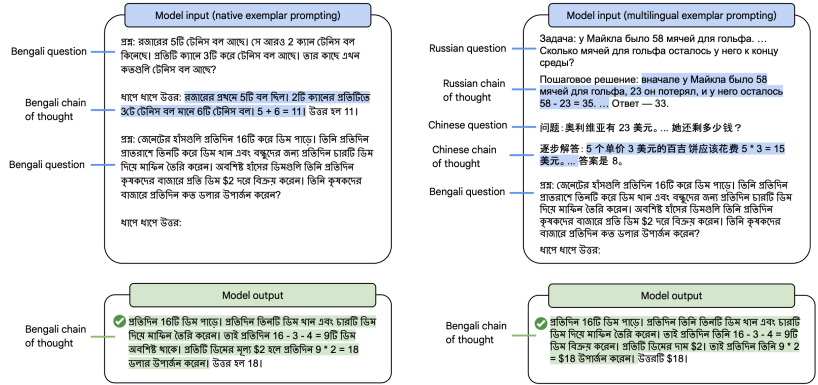
Chain-of-thought prompting

Teach LM how to solve a task



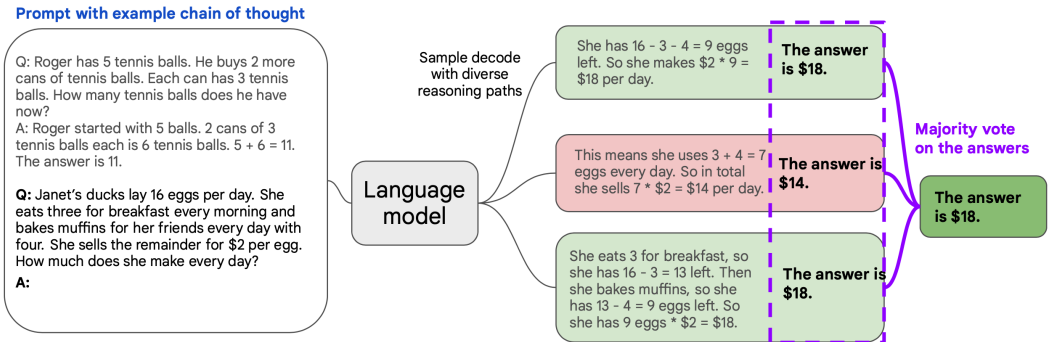
(demo)

Multilingual CoT



- CoT is probably not just memorization (input is highly improbable (Bengali is 0.01% of pre-training data))
- Performance is good on underrepresented languages
- Reasoning ability can be composed with multilingual ability

A neat trick: self-consistency



Self-consistency improves chain-of-thought reasoning in language models (2022).

Decompose the task

Direct Prompting

GPT-3

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?
Answer: Harry Vaughan Watkins.

Question: Who was president of the U.S. when superconductivity was discovered?
Answer: Franklin D. Roosevelt

Chain of Thought

GPT-3

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?
Answer: Theodor Haecker was 65 years old when he died. Harry Vaughan Watkins was 69 years old when he died.
So the final answer (the name of the person) is: Harry Vaughan Watkins.

Question: Who was president of the U.S. when superconductivity was discovered?
Answer: Superconductivity was discovered in 1911 by Heike Kamerlingh Onnes. Woodrow Wilson was president of the United States from 1913 to 1921. So the final answer (the name of the president) is: Woodrow Wilson.

Self-Ask

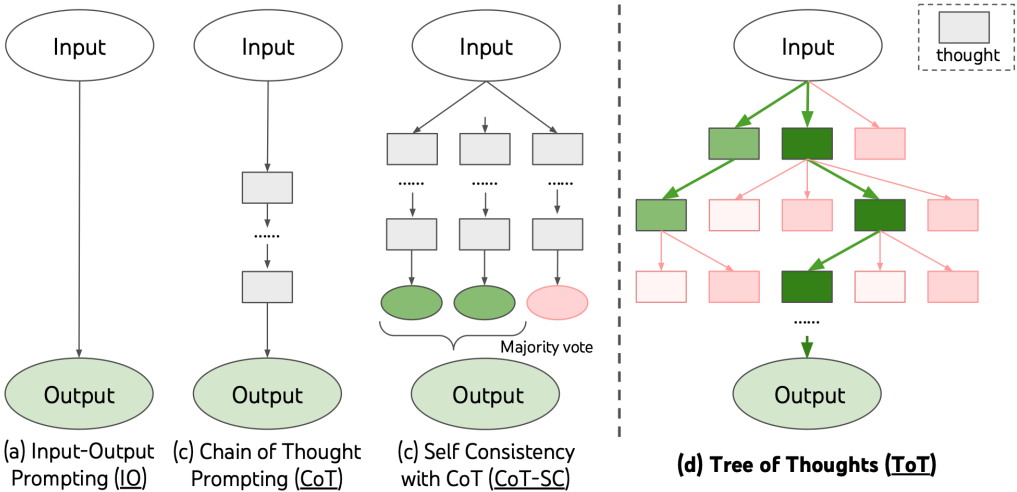
GPT-3

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.

What if there's a mistake in the reasoning path?

Allows for backtracking



(a) Input-Output Prompting (IO)

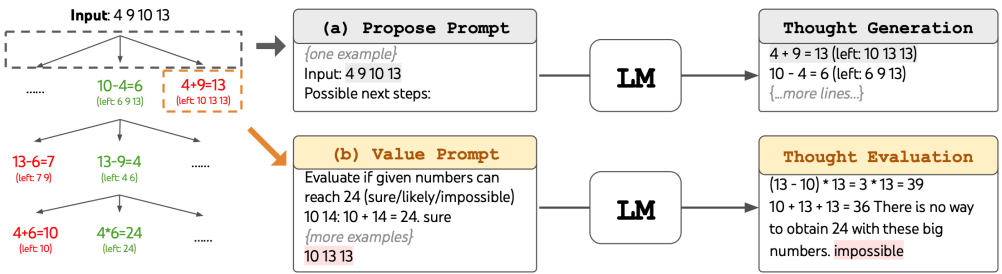
(c) Chain of Thought Prompting (CoT)

(c) Self Consistency with CoT (CoT-SC)

(d) Tree of Thoughts (ToT)

Tree of thought example

Branching, DFS/BFS, decide if the node is a deadend



Summary

- Key challenge: align language with our intent (assisting with task X)
- Prompting: “just ask” the language model to do X
- Pros: simple and allows for creativity (ask for calibration, self-reflection)
- Cons: still an art rather than science (but can be made more reliable through finetuning)