Practical Guide to Prompt LLM

Vijay Parthasarathy, Head of AI at Zoom
Agenda

- Introduction
- Quick tips and possibly some refreshers
- Advance Prompting
- LLM as Judge
- Q&A
Quick Tips
For Starters

Like any ML problem:

- If you CAN write a rule, do that.
- Models might not follow instructions
- Pre and post processing
- Setup evaluation datasets
  - Generate Data: Use LLM’s if you can’t collect
- Tests: Set up regression on edge cases
- Experiment with temperature among other settings
Don't vs Do

Real world prompt: (Redacted)

* The summary should capture the content of the discussion when possible, not just the topic.
* The summary should attribute statements to whoever made them whenever possible.
* If a question is asked, the summary should include the answer, not just the question.
* The summary must always start with "In the meeting,"

[Transcript]
{text}
[End of transcript]
GPT reads forward
Designed output format

- Showing the model the format.

Prompt (Redacted):

5. Extract good questions asked by specific person, extract no more than 2 most important good questions
6. Give a very short title for the text within 6 words,
8. Give a list of sentiment score for "John Doe" and "Vijay", score within 0-100, followed by your reason

Format Response like below:

```json
{
    "specific topics": [],
    "general topics": [],
    "summary": 
    "next steps": [],
    "good questions": [],
    "title": 
    "highlighted sentences": [],
    "sentiment scores": 
}
```

...
Multi Lingual

- Prompting the LM with the target language yields better results
- Models are sensitive to the context,
  - Option 1: Make sure context is all in same language
  - Option 2: Be explicit what the outcome needs to be
- Tools or API response can be in different language
- Language models are not equally good in all languages
  - Experiment, experiment, experiment
Advance Prompting
Zero-shot CoT Prompting

**Prompt:** I have 10 items in my hand and 2 of them are red and others might be any color, one of it was returned to the store and one i have loaned to a friend and how many red balls do i have?

**ChatGPT:** It seems there might be a bit of confusion in your question. You mentioned having 10 items, but then referred to “red balls.” If you have 10 items, and 2 of them are red, it implies that you have 2 red items, not necessarily balls. Now, considering that one item was returned to the store and one was loaned to a friend, the total number of red items in your possession would still be 2. The fact that you returned one and loaned another doesn’t change the initial count of red items you have.

**Prompt:** I have 10 items in my hand and 2 of them are red and others might be any color, one of it was returned to the store and one i have loaned to a friend and how many red balls do i have?

**ChatGPT:** Let’s break down the information:
- You have 10 items.
- 2 of them are red.
- One item was returned to the store.
- One item is loaned to a friend.

Let’s consider the red items first. Initially, you have 2 red items. Now, one of them was returned to the store, so subtract 1. Now you have 1 red item left. Since the loaned item’s color is not specified, we don’t know if it’s red or not. Therefore, you could have either 1 or 0 red items depending on the color of the loaned item.

In conclusion, you either have 1 or 0 red items in your hand.
Designed output format - Chain of Thought

- How do you parse these answer?

""
<Initial Prompt>
....
</Initial Prompt>

Use this format:

Think step by step,
<steps>give reasoning</steps>
Therefore, the answer is: <final_answer>answer</final_answer>
""
**Few Shot Prompting - CoT and Manual CoT**

**Prompt:** 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 and borrowed 2 more?

**A:**

ChatGPT: 7

**Prompt:** 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 and borrowed 2 more?

**A:**

ChatGPT The cafeteria started with 23 apples. They used 20 apples to make lunch, so they have $23 - 20 = 3$ apples left. Then, they bought 6 more apples, making the total $3 + 6 = 9$ apples. Finally, they borrowed 2 more apples, bringing the final count to $9 + 2 = 11$ apples.
Self-Prompting + CoT

SYSTEM

You are a system to write instruction/prompt for large language model given task.
- Your instruction will include chain-of-thought examples

[Instruction]

For this task, you will be provided with a sales answer and an expected answer. Your job is to evaluate the quality of the sales answer by comparing it to the expected answer and generate a matching quality score on a scale of 0 to 10. A score of 0 indicates no match or poor quality, while a score of 10 indicates a perfect match or excellent quality.

Consider the following factors when generating the matching quality score:

1. Relevance: Does the sales answer address the main points or concerns raised in the expected answer?
2. Completeness: Is the sales answer comprehensive and cover all aspects mentioned in the expected answer?
3. Clarity: Is the sales answer clear, concise, and easy to understand?
4. Accuracy: Is the information provided in the sales answer accurate and consistent with the expected answer?

Chain-of-thought examples:

Example 1:
Sales Answer: Our product offers a 30-day money-back guarantee, free shipping, and 24/7 customer support.
Expected Answer: The product comes with a 30-day money-back guarantee, free shipping, and round-the-clock customer support.
Matching/Quality Score: 9 (The sales answer is highly relevant, complete, clear, and accurate, but there's a slight difference in wording.)

Example 2:
Sales Answer: Our software has a user-friendly interface and can be easily integrated with other tools.
Expected Answer: The software provides advanced data analytics and customizable reporting features.
Matching/Quality Score: 2 (The sales answer is not relevant to the expected answer and does not cover the main points.)

Example 3:
Sales Answer: The subscription plan includes access to all premium features, regular updates, and priority support.
Expected Answer: The subscription plan covers all premium features, frequent updates, and priority customer support.
Matching/Quality Score: 9.5 (The sales answer is very relevant, complete, clear, and accurate, with only minor differences in wording.)

Remember to provide a matching/quality score between 0 and 10 based on the factors mentioned above.
ReAct Prompting

- (Optional) Memory, multiple options
- Open AI - Function calling
  - [https://platform.openai.com/docs/assistants/tools/function-calling](https://platform.openai.com/docs/assistants/tools/function-calling)
  - OpenAI implementation seems simpler and faster
  - GPT-3.5 is faster and more accurate now a days
  - There are multiple tool_choice
    - Auto, None (Always call)
Further experimentation

- Hallucination: Ask the model for citation
  - Quote the transcript for this action item, Provide links, etc
- CoT not enough? How about Tree of thought
  - Explore various possibilities and hypotheses, much like human cognitive processes where multiple scenarios are considered
- Prompt chaining
- Dialog-Enabled Resolving Agents
LLM as Judge
Limitation of Human Evaluation

- Hard to scale
- Cumbersome UX
- We need a robust, scalable, explainable, and automated method to evaluate LLM alignment with human preferences
History of LLM-as-a-Judge

- Berkeley: Vicuna
- Microsoft: Sparks of Artificial General Intelligence: Early experiments with GPT-4
- OpenAI: Evals is a framework for evaluating llms and llm systems, and an open-source registry of benchmarks.
- UW: Qlora: Efficient fine tuning of quantized llms
- Meta: LIMA: Less Is More for Alignment, Humpback, etc
- PKU: Large Language Models are not Fair Evaluators
- NTU: Can Large Language Models Be an Alternative to Human Evaluation?
- Berkeley: Judging LLM-as-a-judge with MT-Bench and Chatbot Arena
Humans vs LLM on complex tasks

(a) All votes

Option 1: Self-Consistency Prompting

- If you have the luxury of Prompting multiple times, do it.
  - Return consistent responses
  - Semantic similarity

Source: https://arxiv.org/abs/2203.11171
Option 2: Reflection

- Ask LLM to revise their own outputs based on reasoning
  - Given prompt and the output
- Ask it to reflect as a expert on a topic
  - Chain of thought in action
Option 3: LLM as a Judge

- Pairwise comparison:
  - Generate two outputs with different LLM’s
  - For single LLM, Make sure you have the right temperature
- Single answer grading:
  - Ask the model to evaluate its generation with CoT
  - Evaluation is a simpler task than Generation
  - (There is ton of research and papers on this topic, might be out of scope for this class)
- Chained grading:
  - Send to larger model, if the score is below a threshold

<table>
<thead>
<tr>
<th></th>
<th>Win</th>
<th>Tie</th>
<th>Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zoom_LLM vs GPT-4</td>
<td>89</td>
<td>11</td>
<td>111</td>
</tr>
<tr>
<td>FedAI vs GPT-4</td>
<td>118</td>
<td>23</td>
<td>70</td>
</tr>
<tr>
<td><strong>One_on_one</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zoom_LLM vs GPT-4</td>
<td>63</td>
<td>3</td>
<td>55</td>
</tr>
<tr>
<td>FedAI vs GPT-4</td>
<td>75</td>
<td>8</td>
<td>38</td>
</tr>
</tbody>
</table>
Discussion