Choosing Prompting Strategy

Gurnoor Singh Khurana (366788)

May 2023

1 Introduction

I referred to the papers P1 [1] and P2 [2]for developing my prompting strategy. [1] is a pre print available on arxiv whereas P2 [2] is a published paper in ICLR 2022. In addition to these, I also read the paper given in suggested readings on "Let's think step by step" and incorporated some ideas from there.
I outlined a few key observations from P1 [1]:

- **Using a task-specific prompt is better than using a generalized prompt.**

  Since in the dataset given to our team, there are multiple choice questions and the task is question answering, I make sure to design the prompt such that it incorporates this fact that we have a question answering dataset.

- **Prompts with choices present outperform the prompts without choices**

  Therefore, I include choices in my prompt.

- **Presenting the options as multiple distinct choices improves the results**

  Therefore, I include choices as different options in my prompts.

Additionally, section 4 of P2 [2] mentions some points that the prompt designers were required to adhere to while designing their prompts:

- Prompts should be gramatically correct and understandable by a fluent English speaker with no prior knowledge of the task.
- Prompts should not require explicit counting or numbering. Following this, I rule out usage of any prompts of the form: `Given this {{question}} and {{answer_choices}}, how many of these choices are correct`

I also referred to Appendix G of P2 [2] , which contains prompts designed for various tasks. Specifically, I looked at the section 1.7 of Appendix G as it deals with multiple choice questions.

I would like to thank my teammate Sebastian Breguel for pointing me to the course on Prompt Engineering by Deep Learning AI. Using ideas from that course, and from papers P1 [1] and P2 [2] mentioned above, I iterate to obtain the following final prompt.

2 Final Prompt

```
You are given the following multiple choice question from a neuroscience exam.  Follow the steps
below to help a student who has zero idea about neuroscience

Step 1:  Choose an appropriate list of topics for the question, so that the student can read about
it if needed.  Include no more than 3 items in your response and separate them by commas, with
the first item being the most relevant
Step 2:  Generate a paragraph that provides student adequate information to answer the question
Step 3:  For each answer choice, indicate whether it is correct or not.  Give reason to support
your decision

Question:
{question}
Answer Choices:
{options}

Give your output in the following format
```

```
Topic: [ result of step 1 ]
Relevant context: [ result of step 2 ]
Answer: [ result of step 3 ]
```

We now highlight some key features of this prompt. Note that we arrived at these features by taking inspiration from the readings and the course from deep learning AI. The process was iterative, and whether or not to include a specific instruction in the prompt was decided based on the output the model gives with and without that instruction. If the model performed well, then that instruction was kept, else removed.

Key Features

1. It asks to summarise the topics relevant to the question. This is inspired from forums such as stackoverflow which have the concept of tags for questions. By asking our model to give these *tags*, we are achieving two goals. First, we are conditioning the model to generate its response later on based on these topics. This will help the model to generate a more relevant response. Secondly, it satisfies the purpose of creating an AI tutor. It gives the student a hint, which topics he/she can read to solve this question.

2. In our prompt, we mention that the question is taken from a neuroscience dataset. This is done to provide model with some context regarding the question, so that it can generate a more accurate response by conditioning on this text.

3. We give the model *time to think*. As outlined in the course by Deep Learning AI, we should give model time to think. To do that, we add steps in our prompt where we ask the model to follow a step wise approach. This approach mimics human thinking of breaking a problem into smaller manageable steps. This approach is inspired from teh *Let's think step by step* paper. Since we are not doing a deductive reasoning task, but more of a factual questions task, *steps* implicitly do not have a well defined meaning for the problem. Therefore we outline the steps to be followed explicitly.

4. We ask the model to generate some context relevant to the question, which can help the student to answer that question. This prevents the model from hallucinating, and also at the same time helps the student since they get some context for the question.

5. We instruct the model to indicate for each option whether it is correct or incorrect, and give reason for the same. This ensures that the model does not overlook important details of the text in the answer choices and base its answer on the order of answer choices for instance.

6. We ask the model to provide the output in a structured manner. This ensures the model follows all the instructions outlined in the first portion of the prompt.

7. Note that in the second line of the prompt, we state that the student has zero idea about neuroscience. This was added so that the model explains the relevant context of the question (step 2) in layman terms. This addition was in fact made towards the very end, and I saw considerable improvement in the explanation given by the model, so chose to add it in the prompt.

8. Another thing to note, we specify the answer choices without any labels, to avoid the model from using spurious correlations between the answer label and its correctness. Specifically, we provide the answer choices with a `-` symbol rather than using an ordering such as *Choice 1*, *Choice 2* and so on.

For some experimental results demonstrating the comparison of this prompt with other strategies and the significance of some features of this prompt, please refer to appendix A.

3 Confidence scores

While assigning confidence scores to the generated answers, I observed that in the dataset provided, some questions are repeated and the model generated different answers different time with the same prompt. So, I prompted the model for all the questions twice, and compared the answers to assign confidence scores. In case the model assigned the same answer in both cases, the assigned score was 5. Otherwise it depended on type of question and the number of common answers in the 2 versions.
Also, I observed that in some cases, such as the the question with id 37849, the question referred to a study which is not mentioned in the question. However the model still confidently provides an answer to the user, whereas ideally it should let the user know that the information is insufficient. In such cases I assign a confidence score of 1.
The model is very good at producing convincing explanations, even if they are wrong. So the idea of using this way of generating two outputs and comparing them helps in the cases when the model is hallucinating.
The files containing the two versions of model outputs can be found here.

References

[1] G. Orlanski, "Evaluating prompts across multiple choice tasks in a zero-shot setting," 2022.

[2] V. Sanh, A. Webson, C. Raffel, S. H. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, T. L. Scao, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. V. Nayak, D. Datta, J. Chang, M. T. Jiang, H. Wang, M. Manica, S. Shen, Z. X. Yong, H. Pandey, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Févry, J. A. Fries, R. Teehan, S. Biderman, L. Gao, T. Bers, T. Wolf, and A. M. Rush, "Multitask prompted training enables zero-shot task generalization," *CoRR*, vol. abs/2110.08207, 2021. [Online]. Available: https://arxiv.org/abs/2110.08207

A Comparison of strategies

We demonstrate the difference in output for the last 2 points of key features in Section 2, as these were my own ideas rather and not taken from literature.

First we look the effect of adding *who has zero idea about neuroscience* to the prompt. We choose question with id 37949 from the dataset. The question is as follows:

Question:

What are the essential parts of a brain atlas?

Answer Choices:

- A morphological description of the neurons in each region
- A coordinate space defined by a 3D template
- A gene expression profile, defining each region
- An ontology defining names of and relationships between regions
- The parcellation: 3D boundaries of brain regions
- An electrical definition of neurons in each region

Figure 1: Question 37949 of the dataset

Now we compare the two answers by the gpt model. The first answer is without the use of *who has zero idea about neuroscience* and second with the use of this. We only provide the content in the field *Relevant Context*, since we expect this field to be affected the most.

Answer 1: *A brain atlas is a collection of maps of the brain that provides information about the structure and function of different regions. It is used to understand the organization of the brain and to help identify regions that are involved in specific functions. Brain atlases typically include different types of information, such as the morphology of neurons, gene expression profiles, and 3D boundaries of brain regions.*

Answer 2: *A brain atlas is a map of the brain that helps researchers and doctors identify and study different regions of the brain. It provides information about the structure, function, and connectivity of the brain. There are different types of brain atlases available, such as anatomical, functional, and genetic atlases, each providing different types of information about the brain. In this question, we are asked about the essential parts of a brain atlas, which are necessary to identify and study different regions of the brain.*

We see that *Answer 2* is a bit more detailed than the *Answer 1* and provides more context than it. This is expected since the model was instructed to explain to somebody who has no idea about neuroscience.

Now we demonstrate the effect of including labels in answer choices. For the same question above, we present it in 2 different ways.

Question:

What are the essential parts of a brain atlas?

Answer Choices:

Choice 1: A morphological description of the neurons in each region

Choice 2: A coordinate space defined by a 3D template

Choice 3: A gene expression profile, defining each region

Choice 4: An ontology defining names of and relationships between regions

Choice 5: The parcellation: 3D boundaries of brain regions

Choice 6: An electrical definition of neurons in each region

(a) With choice labels

Question:

What are the essential parts of a brain atlas?

Answer Choices:

- A morphological description of the neurons in each region
- A coordinate space defined by a 3D template
- A gene expression profile, defining each region
- An ontology defining names of and relationships between regions
- The parcellation: 3D boundaries of brain regions
- An electrical definition of neurons in each region

(b) Without choice labels

We see that when we include numbered choices in the prompt, we get the answer in the following form

Choice 1: Incorrect. A morphological description of neurons is not an essential part of a brain atlas.

Choice 2: Correct. A coordinate space defined by a 3D template is an essential part of a brain atlas.

Choice 3: Incorrect. A gene expression profile is not an essential part of a brain atlas.

Choice 4: Correct. An ontology defining names of and relationships between regions is an essential part of a brain atlas.

Choice 5: Correct. The parcellation: 3D boundaries of brain regions is an essential part of a brain atlas.

Choice 6: Incorrect. An electrical definition of neurons is not an essential part of a brain atlas.

Figure 3: Answer when choices are presented with labels

We see that this form is prone to errors, as the model might draw correlations between *Choice 1* and the words *Correct* and *Incorrect*. To avoid this, we use - to present answer choices. This way we get the following response:

- A morphological description of the neurons in each region: This answer is incorrect. A morphological description of the neurons in each region is not an essential part of a brain atlas, although it is a useful feature.
- A coordinate space defined by a 3D template: This answer is correct. A coordinate space defined by a 3D template is an essential part of a brain atlas. This template allows researchers to compare data across different brains and studies.
- A gene expression profile, defining each region: This answer is incorrect. A gene expression profile is not an essential part of a brain atlas, although it is a useful feature.
- An ontology defining names of and relationships between regions: This answer is correct. An ontology defining names of and relationships between regions is an essential part of a brain atlas. This ontology allows researchers to communicate about different regions of the brain using a common language.
- The parcellation: 3D boundaries of brain regions: This answer is correct. The parcellation, or 3D boundaries of brain regions, is an essential part of a brain atlas. This information allows researchers to accurately locate and compare different regions of the brain.
- An electrical definition of neurons in each region: This answer is incorrect. An electrical definition of neurons in each region is not an essential part of a brain atlas, although it is a useful feature.

Figure 4: Answer when choices are presented without labels

We see that this is less likely to be prone to errors since the model conditions on the option rather than the label.

Comparison with other prompting strategies

In this section, we outline the comparison of our prompting strategy with 2 other strategies I tried. Note the my final prompting strategy was made by a clever mix of building upon existing work, some task specific tailoring and trial and error.

Strategy 1

First I tried a simple baseline strategy which gives the model the question and the answer choices. The following figure demonstrates the prompt.

How does the W3C provenance standard (W3C PROV) apply to neuroscience datasets?

Choose the correct option among the following

- Entity refers to the who (who performed the experiment or what transformed the data), activity refers to how (what was done on the dataset/how was the data transformed) and agent refers to the datasets themselves
- Entity refers to the dataset itself, activity refers to how (what was done on the dataset/how was the data transformed) and the agent refers to who/what (who perform the experiment/what transformed the data)
- Entity refers to the dataset itself, activity refers to who (who perform the experiment/what transformed the data) and agent refers to how (what was done on the dataset/how was the data transformed)

Figure 5: Prompt using strategy 1

The output we get from the above prompt is *Entity refers to the dataset itself, activity refers to how (what was done on the dataset/how was the data transformed) and the agent refers to who/what (who performed the experiment/what transformed the data).*

Whereas the output we get using our final prompting strategy is
Topic: W3C PROV, Neuroscience datasets, Data provenance
Relevant context: The W3C provenance standard (W3C PROV) is a set of guidelines for describing the history

or lineage of digital objects, including datasets. It is commonly used in the field of neuroscience to track the provenance of datasets, which helps to ensure data quality and reproducibility.
Answer:
- Entity refers to the who (who performed the experiment or what transformed the data), activity refers to how (what was done on the dataset/how was the data transformed) and agent refers to the datasets themselves. (Incorrect)
- Entity refers to the dataset itself, activity refers to how (what was done on the dataset/how was the data transformed) and the agent refers to who/what (who perform the experiment/what transformed the data). (Correct)
- Entity refers to the dataset itself, activity refers to who (who perform the experiment/what transformed the data) and agent refers to how (what was done on the dataset/how was the data transformed). (Incorrect)
Reasoning: In the W3C PROV standard, an entity refers to a digital object, such as a dataset. The activity refers to an action performed on the entity, such as an experiment or data transformation. The agent refers to the person or software that performed the activity. Therefore, the correct answer is the one that correctly identifies the entity as the dataset, the activity as the action performed on the dataset, and the agent as the person or software that performed the activity. Answer choice 1 and 3 are incorrect because they incorrectly identify the entity and agent.
As we can clearly see, the strategy 1 produces a very vague output, whereas our final prompting strategy produces a very detailed output.

Strategy 2
I tried another strategy, where I ask the model to provide reason for its answer. An example of this strategy for the same question in above strategy can be seen in the following figure:

How does the W3C provenance standard (W3C PROV) apply to neuroscience datasets?

Giving reason, state which of the following options answer the question

- Entity refers to the who (who performed the experiment or what transformed the data), activity refers to how (what was done on the dataset/how was the data transformed) and agent refers to the datasets themselves
- Entity refers to the dataset itself, activity refers to how (what was done on the dataset/how was the data transformed) and the agent refers to who/what (who perform the experiment/what transformed the data)
- Entity refers to the dataset itself, activity refers to who (who perform the experiment/what transformed the data) and agent refers to how (what was done on the dataset/how was the data transformed)

Figure 6: Prompt using strategy 2

I get the following output using this strategy:
Option 2: Entity refers to the dataset itself, activity refers to how (what was done on the dataset/how was the data transformed) and the agent refers to who/what (who perform the experiment/what transformed the data). This option correctly identifies the roles of entity, activity, and agent in the W3C PROV standard. In the context of neuroscience datasets, the entity would refer to the dataset itself, the activity would describe how the data was collected or manipulated, and the agent would identify who or what performed the experiment or transformed the data. This information can help to establish the provenance of the dataset and ensure that it can be properly attributed and understood.

As we can see it is much better than strategy 1, but it is still not as comprehensive as the output we get from our final prompting strategy.