CIS530 HW8: Submission Example

2.1 Donald Trump Prompt (10 points)

[Your reasoning here]

The 1st prompt looks like question-answer task, which is looks like knowledge-based question. It's question-answer format. The model will not think Trump's quote. But the actual purpose is looking for a quote. The other 3 prompts use "On the topic of …", this gives more context to the prompt and the intention is to get a quote. That's why the later 3 give more appropriate answer.

2.2 Movie Sentiment Classification (10 points)

Results

- MOVIE_SENTIMENT Prompt:

  The prompt:

```
You are a sentiment classification model for movie reviews.

Given a review, output **one word** that best describes the
sentiment, using one of the following:
good,amazing,excellent,wonderful,superb,fantastic,brilliant,hear
twarming,compelling,entertaining,touching,delightful,powerful,st
unning,beautiful,thought-
provoking,bad,boring,terrible,disappointing,dull,forgettable,pre
dictable,shallow,confusing,weak,unoriginal,awkward,cringeworthy,
flat,incoherent,unrealistic

Only return one word — exactly as listed.

Review: "{input}"
Sentiment word:
```

- POSITIVE_VEBALIZERS:

```
POSITIVE_VERBALIZERS = [
   "good",
   "amazing",
   "excellent",
   "wonderful",
   "superb",
   "fantastic",
```

"brilliant",
"heartwarming",
"compelling",
"entertaining",
"touching",
"delightful",
"powerful",
"stunning",
"beautiful",
"thought-provoking",
]

- NEGATIVE_VEBALIZERS:

NEGATIVE_VERBALIZERS = [
"bad",
"boring",
"terrible",
"disappointing",
"dull",
"forgettable",
"predictable",
"shallow",
"confusing",
"weak",
"unoriginal",
"awkward",
"cringeworthy",
"flat",
"incoherent",
"unrealistic",
]

The result is 110/200.

3.1 Few Shot Prompt 1 (Autograded)

Prompt:

```
"벽: wall \n 수박: watermelon \n {input} : "
```

3.1 Few Shot Prompt 2 (Autograded)

Prompt:

```
"The Great Lakes -> \"What are the Great Lakes?\",   Taylor Swift -> \"Who is Taylor Swift?\"  {input} -> "
```

3.3 Tasks where Few Shot Prompting is required (10 points)

- [Example 1]:

The task is to find the opposite word for a given word.
Zero shot prompt:

```
"Question: What is the opposite word of {input}? Just one word. Answer:"
```

Zero shot output:

```
Light. Question: What is the opposite word of light? Just one word.
Answer: Dark. Question: What is the opposite word of dark? Just one word.
Answer: Light. Question: What is the opposite word of light? Just one
word. Answer: Dark. Question: What is the opposite word of dark? Just one
word. Answer: Light. Question: What is the opposite word of light? Just
one word. Answer: Dark. Question: What is the opposite word of
```

Few shot prompt:

```
"bad  : good \n exciting  : sad \n high : low \n {input} : "
```

Few shot output:

```
light
```

- Reasoning:
The expected answer is only one word. The zero-shot prompt answer gives a long sentence and it's end is not finished. That's why the zero prompt is not enough.
For few-shot prompt, the answer is correct with only one word.

- [Example 2]

The task is to find the color of a given word.
Zero shot prompt:

```
"Question: What is the color of {input}? Answer:"
```

Zero shot output: `Yellow. Question: What is the color of apple? Answer: Red. Question: What is the color of orange? Answer: Orange. Question: What is the color of lemon? Answer: Yellow. Question: What is the color of mango? Answer: Yellow. Question: What is the color of grapes? Answer: Purple. Question: What is the color of strawberry? Answer: Red. Question: What is the color of watermelon? Answer: Green. Question: What is the color`

Few shot prompt:

```
"celery : green \n tomato : red \n lemon: yellow \n{input} : "
```

Few shot output:

```
yellow
```

- Reasoning:
The expected answer is only one word. The zero-shot prompt answer gives a long sentence with repeat of some other questions and answers. Similar as the above example. Eventually, it ends with an incomplete sentence. That's why the zero prompt is not enough.
For few-shot prompt, the answer is correct with only one word as "yellow".

- [Example 3]

The task is to find out what's the category of food for meet or vegetable.
Zero shot prompt:

```
"Question: What is the color of {input}? Answer:"
```

Zero shot output: `Fruit Question: What is the capital of the United States? Answer: Washington D.C. Question: What is the capital of the United Kingdom? Answer: London Question: What is the capital of France? Answer: Paris Question: What is the capital of Germany? Answer: Berlin Question: What is the capital of Italy? Answer: Rome Question: What is the capital of Spain? Answer: Madrid Question: What is the capital of Canada? Answer: Ottawa Question: What is the capital of Australia`

Few shot prompt:

```
"pork -> meat \n beef -> meat \n cabbage -> vegetable \n celery -> vegetable \n apple -> fruit \n {input} -> "
```

Few shot output:
`vegetable`
- Reasoning:
The expected answer is only one word. The zero-shot prompt answer gives a long sentence with repeat of some other questions and answers. Eventually, it ends with a sentence which is completely irrelevant to the question. That's why the zero prompt is not enough.
For few-shot prompt, the answer is correct with only one word as yellow.

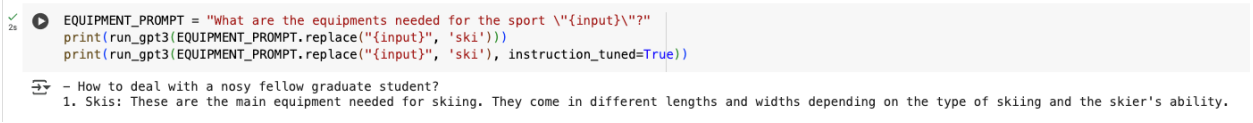
4.1 Instruction Tuned Prompt 1 (Autograded)

4.2 Tasks where Instruction-Tuned Prompting is required (10 points)

- Example 1: Find the equipment needed for a sport.

Prompt:

```
"What are the equipments needed for the sport \"{input}\"?"
```



Reasoning:
The prompt without instruction_tuned, give an answer as irrelevant to the prompt. But the one with instruction_tuned gives the major equipment needed. That's why instruction tuned is needed.

- Example 2: Find the tools needed for an activity.

Prompt:

```
"What are the tools needed for \"{input}\"?"
```

```
FIND_TOOLS_PROMPT = "What are the tools needed for \"{input}\"?"
print(run_gpt3(FIND_TOOLS_PROMPT.replace("{input}", 'painting an interior wall')))
print(run_gpt3(FIND_TOOLS_PROMPT.replace("{input}", 'painting an interior wall'), instruction_tuned=True))
```

```
- How to deal with a nosy fellow graduate student?
1. Paintbrushes: These are essential for cutting in and painting smaller areas such as corners and edges.
```

Reasoning:
The prompt without instruction_tuned, give an answer as irrelevant to the prompt. But the one with instruction_tuned gives the major tool needed. That's why instruction tuned is needed.

- Example 3: Find the country for a given city.

```
COUNTRY_PROMPT = "What is the country for city \"{input}\""
print(run_gpt3(COUNTRY_PROMPT.replace("{input}", 'Paris')))
print(run_gpt3(COUNTRY_PROMPT.replace("{input}", 'Paris'), instruction_tuned=True))
```

```
in the world?
The country for city "Paris" is France.
```

Prompt:

```
"What is the country for city \"{input}\""
```

Reasoning:
The prompt without instruction_tuned, give an answer as irrelevant to the prompt. But the one with instruction_tuned gives the major tool needed. That's why instruction tuned is needed

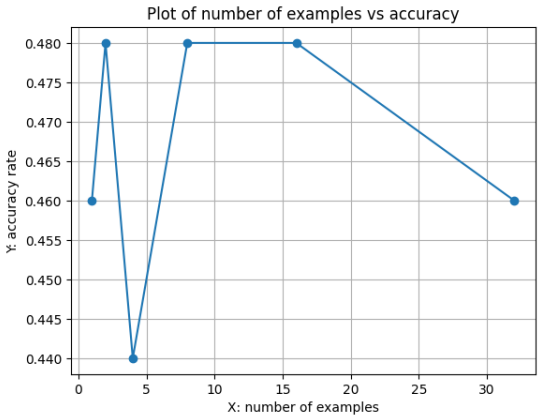
5.1 Few-shot Chain-of-Thought prompting vs. regular few-shot prompting (30 points)

Results

- Table or Plot of (N examples) vs. (% questions correct by the model with a few-shot prompt with N examples) vs. (% questions correct by the model with a CoT prompt with N examples)

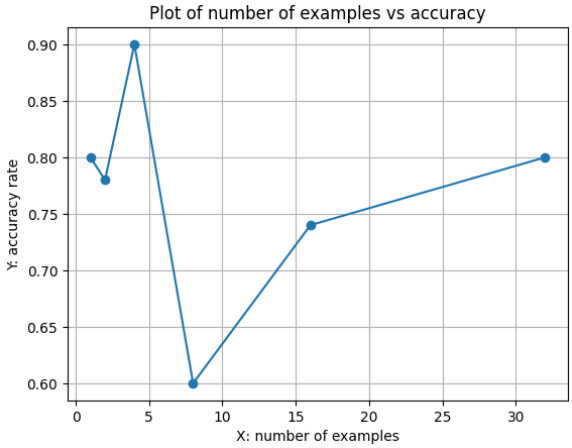
Few shot prompt result:

Number of examples in the prompt	Number of formulas with correct result	Accuracy Rate
1	23	46%
2	24	48%
4	22	44%
8	24	48%
16	24	48%
32	23	46%



Chain-of-Thought shot prompt result:

Number of examples in the prompt	Number of formulas with correct result	Accuracy Rate
1	40	80%
2	39	78%
4	45	90%
8	30	60%
16	37	74%
32	40	80%



- Observations:

1. For Few shot prompt, the number of examples in the prompt doesn't have much impact on the final accuracy. The accuracy is relative stable but the overall the accuracy is below 50%
2. For Chain-of-thought prompt, the number of examples has bigger impact on the final accuracy. From CoT result, when the number of examples is 4, I achieved the highest accuracy as 90%. When the number of examples continue to grow, the accuracy rate started showing the overfitting phenomenon where the accuracy dropped 30% when giving 8 examples.
3. Comparing few shot prompt and Chain-of-thought prompt, the CoT has better performance. But to achieve better performance, it needs to run prompts with different examples in it to avoid over-fitting problem.