A Prompting Assignment for Exploring Pretrained LLMs

Carolyn Jane Anderson
Wellesley College
Wellesley, MA
carolyn.anderson@wellesley.edu

1 Introduction

As the scale of publicly-available large language models (LLMs) has increased, so has interest in few-shot prompting methods. This paper presents an assignment that asks students to explore three aspects of large language model capabilities (commonsense reasoning, factuality, and wordplay) with a prompt engineering focus.

The assignment consists of three tasks designed to share a common programming framework, so that students can reuse and adapt code from earlier tasks. Two of the tasks also involve dataset construction: students are asked to construct a simple dataset for the wordplay task, and a more challenging dataset for the factuality task. In addition, the assignment includes reflection questions that ask students to think critically about what they observe.

2 Course Context

This assignment was designed for an advanced undergraduate Natural Language Processing course. The corresponding lectures cover prompting techniques like chain-of-thought reasoning and continuous prompting, as well as limitations of LLMs. By this point in the semester, students are familiar with the mechanics of LLMs, from byte-pair tokenization (Gage, 1994; Sennrich et al., 2016) to multi-head attention (Vaswani et al., 2017). Students had one week to complete the assignment.

3 Learning Goals

This assignment is designed to allow students to explore three different aspects of LLM capabilities by experimenting with prompting techniques. The learning goals for the assignment are as follows:

- Build programs that interface with LLMs
- Explore various ways of constructing prompts
- Construct datasets to explore LLM capabilities related to factuality and wordplay
- Critically analyze LLM capabilities

Pig Latin
papaya -> apayapay
Commonsense Reasoning (from Roemmele et al. (2011))

1. Premise: The man broke his toe.
   Question: What was the CAUSE of this?
   (a) He got a hole in his sock.
   (b) He dropped a hammer on his foot.

Notable Scientist Facts

1. What is Barbara Partee's field of study?
   (a) Linguistics
   (b) Physics

Figure 1: Example items from the three main tasks

4 Assignment Design

This assignment consists of three core tasks, each exploring a different aspect of LLM capability.

4.1 Task 1: Wordplay

In Task 1, students explore the ability of a pretrained LLM to solve one kind of wordplay puzzle: Pig Latin. Pig Latin is a language game in which the initial consonants of a word are removed and appended to the end along with the syllable "ay" (Figure 1). Although the pattern is simple, the subword tokenization used by contemporary LLMs may make it more challenging to recognize.

This task consists of five subtasks, plus a set of reflection questions:

1. Create a Pig Latin dataset of 20 words
2. Write a function to generate prompts
3. Write a function to submit a single prompt to the model
4. Write a function to post-process a completion and extract the answer
5. Write a function to run the prompting experiment on the entire dataset and report the model's performance

Proceedings of the Sixth Workshop on Teaching NLP, pages 81–84
August 15, 2024 ©2024 Association for Computational Linguistics

Students were required to experiment with three aspects of the prompt: providing examples (few-shot prompting), describing the task in different ways, and varying the format of the examples.

The analysis questions asked students to make observations about the effect of different prompt formats, and to comment on factors that might affect the model's performance. I was particularly hoping that students might pick up on the fact that subword tokenization makes this task more challenging, since they were familiar with byte-pair encoding tokenization.

4.2 Task 2: Commonsense Reasoning

In Task 2, students explore pretrained LLM performance on a commonsense reasoning benchmark: the Choice of Plausible Alternatives (COPA) task (Roemmele et al., 2011) from the SuperGLUE suite of LLM benchmarks (Wang et al., 2019). The dataset targets model understanding of real world cause and effect relationships (Figure 1).

The subtasks for this part were similar to those in Part 1, except that students did not have to construct their own dataset. However, some students did find the JSONL format of COPA more challenging to work with, particularly because there were multiple ways of incorporating the cause/effect label for each question into the prompt.

4.3 Task 3: Factuality

Task 3 explores the use of pretrained LLMs as knowledge bases. In this part, each student constructs a dataset of 20 multiple choice questions about a notable female scientist and uses it to explore the LLM's knowledge of the scientist.[1]

This task was more open-ended. The prompting task was not autograded, to allow more freedom in the structure of the dataset and program. However, students were encouraged to follow the format of the COPA dataset so that they could reuse their code from the previous task as much as possible.

The reflection questions for this task asked students to reflect on the limitations of the task (many brought up the small sample size) and to make observations about which kinds of questions were easier or harder for the model. One trend that emerged across submissions was that the model performed better for scientists born more recently, perhaps because they had bios on many different websites.

4.4 Intellectual Curiosity Points

The original assignment contains an additional section entitled Intellectual Curiosity. A key aspect of my course design is that 10 points from each assignment are reserved for demonstrating intellectual curiosity. I implemented this policy after observing how many CS students approach assignments like a checklist and expect full credit for completing all items. The curiosity points system is my way of encouraging open-ended exploration and independent learning. Table 1 in Appendix A summarizes how these points were awarded in one version of CS 333.

5 LLM Access Practicalities

I have used two LLMs with this assignment in the past: OpenAI's GPT-3 model (Brown et al., 2020), and Meta's LLaMA (13B) model (Touvron et al., 2023). These models worked well since they performed above chance performance on all tasks, but still made many mistakes for students to analyze. In the starter code, I provide a Python program to be used as a library that passes a prompt to the model and returns a completion; this makes it easy to substitute a different model.

When using GPT-3, I created the impression that I could track student's individual usage by sending API keys individually; in reality, each key was shared by several students. I had no issues with students sending too many queries, but this might be challenging in a larger class. For a class of 24 students, the assignment cost around $50 USD.

I ran LLaMA (13B) on a server with an Nvidia A6000 GPU using a Gradio app that allowed web requests from Wellesley IP addresses. Gradio handles request queuing. However, in a large class, the latency for a single model could be significant. I have included the code for the LLaMA model and Gradio app in my materials.

6 Conclusion

This assignment allows students to experiment with prompting techniques in the context of exploring three aspects of pretrained LLMs: their ability to solve wordplay, their grasp of commonsense reasoning, and their use as knowledge bases. Some aspects of this assignment may not scale well to larger class sizes: for instance, the two ways of setting up access to the pretrained LLM that I used both pose problems at scale.

[1] The individual datasets can be combined together for use in a later assignment.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of NAACL-HLT 2019*.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*.

A Intellectual Curiosity Points

10 points	Ran few-shot prompting experiments with a novel task
	Read additional papers and did more few-shot prompting experiments
	Set up an antonym probe task
	Read about SuperGLUE (Wang et al., 2019) and ran a prompting experiment with the BoolQA (Clark et al., 2019) subset
7 points	Ran the Pig Latin task on another LLM
	Tested LLaMA on another language game
5 points	Reversed the Pig Latin experiment
	Tested statistical significance
4 points:	Extra research on prompting

Table 1: Examples of intellectual curiosity point allocation