

The lead TA for this assignment is Risako Owan ([owan0002@umn.edu](mailto:owan0002@umn.edu)). Please communicate with her via Slack, email, or office hours.

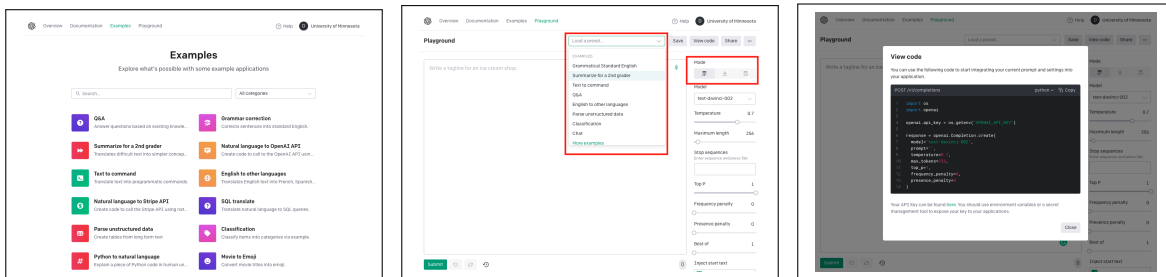
The computational trend of NLP research is shifting from feature engineering to representation learning to pretraining-finetuning to very recently prompt engineering with large language models (LLM). Large language models (or other generative models trained on other modalities) allow the extraction of diverse and intrinsic knowledge from human-written texts/images/videos and their pairs. This assignment requires you to **explore the limits and capabilities of large language models** by designing your own prompts to interact with LLMs, observing their outputs, and understanding their shortcomings. Before you start the homework, you are encouraged to understand the course material on prompting\* taught on March 22. Please go through the three steps described below and submit your prompt spreadsheet and report (both as PDFs).

## Step 1: Getting used to LLMs APIs

You first need to choose which LLM to play with: GPT3 or ChatGPT.

ChatGPT (<https://chat.openai.com/chat>) is currently free and you should be able to login using your usual google credentials. You may also choose to use the ChatGPT API, which costs \$0.002 for 1K tokens (about 750 words).

If you choose GPT3 or ChatGPT API, you'll need to create an account on OpenAI to get access. Please make your account in OpenAI's playground: <https://beta.openai.com/playground>. You should receive \$5 worth of credits that you can use during your first 3 months unless you already have an account†. To illustrate what you can do with \$5, if you use the gpt-3.5-turbo model (\$0.002 per 1K tokens), you can generate a total of 2.5 million tokens for free.



Once you've logged in, please take a look at existing example tasks in Examples tab‡, such as Question Answering, Summarization, and Text-to-Command, and load preset prompts in Playground tab§, as described in the figures above. You can also choose three types of modes, *Complete*, *Chat*, *Insert*, and *Edit* in the Mode tab on right panel (red rectangle box on the right figure above). Experiment with one or more modes for each prompt.

For those using free ChatGPT, take a look at this resource for examples: <https://lifearchitect.ai/chatgpt-prompt-book/>

Play around with whichever LLM you choose. You will be asked to generate prompts for submission in step 3. Note that for your submission, you are not allowed to use any online examples, whether provided here or not.

\*[https://dykang.github.io/classes/csci5541/S23/slides/0327\\_prompting.pdf](https://dykang.github.io/classes/csci5541/S23/slides/0327_prompting.pdf)

†<https://openai.com/pricing/>

‡<https://beta.openai.com/examples/>

§<https://beta.openai.com/playground>

## Step 2: Understand Current Prompting Techniques.

This assignment requires you to understand the state-of-the-art prompting techniques. Refer to the following articles and papers (you can find links to the original articles in the Reference section):

- Discrete prompting methods: Auto-prompting method [SRLI<sup>+</sup>20] and analysis on the limitation and tricks in discrete prompting [ZWF<sup>+</sup>21]
- Soft prompting methods: Prefix-tuning [LL21] and Prompt-tuning [LARC21].
- Stress test of GPT3 on various aspects: commonsense reasoning [MD20], hypes and ethics [BGMMS21], and planning [VOSK22].
- Some prompting tricks like Chain-of-thoughts [WWS<sup>+</sup>22] and applications to human-GPT3 collaboration for text editing [SDYJ<sup>+</sup>22] and poetry writing [CPH22].

**Your task is to choose one bullet point from the list above and summarize the papers and limitations of current methods.** (Maximum 2 pages; don't forget to properly cite them in your report and do **NOT** use GPT3/ChatGPT or other summarization tools for this.)

## Step 3: Designing your own Prompts

The last step involves designing your own creative and discrete prompts! Your goal is to *fool the system* by finding adversarial examples without any external help or resources and providing reasonable justification of the failure cases. The first step is to **pick three of the following aspects you would like to explore**. You can also think of **your own category** but please describe it clearly in the report:

- **Creativity**, e.g., writing next possible sentences in your own story prompt
- **Generalization** to unseen, novel situations or tasks, e.g., design a totally unseen, new task you can get benefit from GPT3 prompting.
- **Grammatical errors, typo**, and other language fluency measurement: e.g., GPT-3 makes mistakes grammar/different tense with popular phrases and prefers more active speech?
- **Factuality** and memorization of **commonsense knowledge**: e.g., GPT-3 memorizes the phone number of White House? GPT-3 is able to understand Newton's laws of motion?
- **Biases and Ethical Concerns**: e.g., GPT-3 prefers to set higher salaries for men than women? GPT-3 generates more offensive language toward Asian people than white people.
- **Temporal/Spatial reasoning**: e.g., GPT-3 is not able to understand temporal and spatial state transition of objects in the real world?
- **Mathematical reasoning**: e.g., GPT-3 is able to solve Fibonacci sequence?
- **Reasoning on Commonsense, Morality, and Legality**: e.g., GPT-3 can address some commonsense or social issues like morality and legality?
- **Applications on NLP tasks**: e.g., GPT-3 translates low-resource languages? GPT-3 can write code for your assignment? GPT-3 can summarize/rewrite your emails?
- **Others**: Any other aspects you think these large language models are not capable of.

You should make **three adversarial task prompts** for each category you choose from above, where tasks in each aspect are different from each other. Note that you are finding **adversarial** cases in which GPT-3/chatGPT outputs don't match your expected results. Each task prompt should also be tested with a zero-shot setup (no example given) and a few-shot setup (one to three examples given). In total, you need to create **at least three aspects x three tasks x two setup (zero/few) = 18 adversarial prompts**. You can of course try as many prompts as you wish and there would be bonus points based on the quality/creativity of your prompts (See the prompt evaluation criteria below). After prompting, you should provide reasonable reasons for these failures and possible ways to improve them. Additional prompts to support your reasoning/logic are strongly recommended.

Here are some notes and tips for your prompt design:

- You have to provide a reasonable quality of task description and examples in your prompts, and make sure that GPT-3's failure does not come from the quality of your prompt design, but is mainly caused by the lack of inherent capabilities of GPT-3. You can find high-quality prompts through trial-and-error with GPT-3 in Playground or in your python code using OpenAI's API (See View Code tab in Playground interface in the figure above).
- We are scientists! Try different task descriptions and prompt examples, and see if GPT-3 always fails deterministically.
- Once again, you cannot use examples from the Example tab, predefined prompts, or previous papers. It will be treated as *cheating* if I find the same prompt used before. Note that instructors already have a huge list of adversarial prompts from previous classes. Check the class page for our academic integrity policy
- Here are some additional tips you may consider during prompting:
  - Find novel tasks you/GPT3/chatGPT can't do.
  - Find tasks that GPT3/chatGPT can do a better job than humans.
  - Find cases where even humans do not agree with each other and see how GPT3/chatGPT can handle this human disagreement
  - Find unseen (probably not seen in the training data) but realistic cases
  - Find unseen and unrealistic cases

## Deliverable

Please upload your spreadsheet containing your designed prompts and report to [Canvas](#) by **Apr 3, 11:59pm**.

**Spreadsheet:** Your designed prompts should be contained in a spreadsheet (e.g., CVS, Excel files, Google Spreadsheet) and exported to a PDF in a way that all your input is visible. (Please make sure none of your input is cut off before submitting it since I cannot grade what I don't see.) In each row, you can write a prompt, and in your columns, you can include the following information:

- Mode (Complete, Chat, Edit, or/and Insert). For ChatGPT, you will not have a dropdown to choose from when prompting but you should still specify which mode your prompt belongs to.
- Aspect (e.g., Creativity, Generalization)
- Task Description
- Instructions (if you used Edit mode)
- Number of examples
- Examples
- Input Task (without answer)
- Expected Answer
- Predicted Answer by GPT3
- Setting (e.g., Engine - text-davinci-002, Temperature - 0 (deterministic), Max length - 256, Stop sequences - none, Top P - 1, Frequency penalty - 0, Presence penalty - 0, Best of - 1". If there are no settings for you to modify, enter N/A.)

**Report:** Maximum four pages PDF total. Your report needs to include the following content:

1. Which papers did you choose to read? (Please cite them properly) Summarize the papers and limitations of the current methods. Maximum 2 pages.
2. Explain why you chose your aspect categories and how your prompts were designed. Discuss the outputs (how they differed/matched your expectations and justifications for why GPT3/ChatGPT found them difficult.) Also include challenges you encountered during your homework and your

general thoughts on language model prompting. What are the takeaways or other interesting things you learned through this assignment?

**Rubric:** 15 points total

- Report
  - Properly cites all papers used in the report (1 points)
  - Summarizes the papers clearly. Makes it clear which papers they are summarizing. (4 points)
  - Lists limitations of current methods (2 point)
  - Clearly states which aspect categories they chose (and explains it clearly if they have thought of their own) (1 point)
  - Provides convincing justifications for why the LLM struggled with their adversarial prompts. Offers potential solutions. (2 points)
  - Discusses challenges encountered and takeaways from the assignment. (1 point)
- Spreadsheet
  - Includes all columns specified (1 point)
  - Contains at least 18 adversarial prompts (3 points, -0.5 for each missing/bad-quality prompt)

**Awards:** Your designed prompts will also be considered for the following awards<sup>¶</sup> and you will receive 2 extra points if you win.

- Best Research Application: Discover a finding that may lead to further research or publication
- Best Misuse Case: Discover a scenario where GPT-3 generates harmful content in a way that might be difficult for GPT-3 admins to catch. Ideally, also determine a way one might go about detecting such harmful content
- Best Mistake Case: Discover an interesting case where GPT-3 fails by not giving you the output you want or expect
- Best Prompt: Discover an interesting prompt
- Arbitrary Finding: Find something completely random but interesting

## References

- [BGMMS21] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445922>.
- [CPH22] Tuhin Chakrabarty, Vishakh Padmakumar, and He He. Help me write a poem: Instruction tuning as a vehicle for collaborative poetry writing. *arXiv preprint arXiv:2210.13669*, 2022. <https://arxiv.org/abs/2210.13669>.

---

<sup>¶</sup>Borrowed these criteria from London Lowmanstone’s GPT-3 Edit/Insert Hackathon, March 2022. Thanks, London!

- [LARC21] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. <https://aclanthology.org/2021.emnlp-main.243>.
- [LL21] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online, August 2021. Association for Computational Linguistics. <https://aclanthology.org/2021.acl-long.353>.
- [MD20] Gary Marcus and Ernest Davis. Experiments testing gpt-3’s ability at commonsense reasoning: results, 2020. <https://cs.nyu.edu/~davis/papers/GPT3CompleteTests.html>.
- [SDYJ+22] Timo Schick, Jane Dwivedi-Yu, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, and Sebastian Riedel. Peer: A collaborative language model. *arXiv preprint arXiv:2208.11663*, 2022. <https://arxiv.org/abs/2208.11663>.
- [SRLI+20] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online, November 2020. Association for Computational Linguistics. <https://aclanthology.org/2020.emnlp-main.346>.
- [VOSK22] Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. Large language models still can’t plan (a benchmark for llms on planning and reasoning about change). *arXiv preprint arXiv:2206.10498*, 2022. <https://arxiv.org/pdf/2206.10498.pdf>.
- [WWS+22] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models, 2022. <https://arxiv.org/abs/2201.11903>.
- [ZWF+21] Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models, 2021. <https://arxiv.org/abs/2102.09690>.