Prompt Engineering for Narrative Choice Generation

Sophia Rutman, Class of 2024

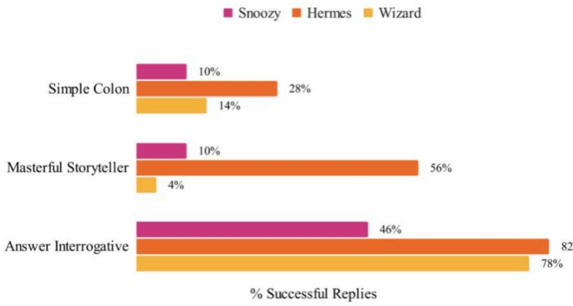
This project explored the potential of Large Language Models (LLMs) in aiding humans with media creation. How can these models, which are a form of machine learning that seek to understand a prompt and return a sensical response, change video games, movies, and other forms of media to adapt to their watchers, readers, or users? We investigated whether machine learning can change media in a logical manner and in real time based on the prior choices or demonstrated interests of a user. For example, can LLMs help the creators of *The Sims*™ change the flow of the game based on a specific user's past decisions? The field of interactive digital storytelling believes that there are many facets of machine learning that can help humans advance media and change the face of entertainment.

To test this hypothesis, we fed prompts into three different types of LLMs. These prompts included the plot of movies up to a "crossroads point," or a point where the character was forced to make a decision for the story to progress. These movies were all released after February 23rd, 2023 and contained no information on any events that occurred before this date.This was to ensure that the LLMs could not gather any prior knowledge on any of the films from the internet.

Each LLM was given three different prompts for each crossroads point. Each prompt contained the same plot description, and then a question asking what the character does next. The question was asked in three separate ways - one for each type of prompt. This process is called "prompt engineering," which attempts to discover the best wording for a prompt. The best wording for a prompt yields the best response.

To quantify "best," we created different categories where an LLM could go wrong or "fail" within a response to each prompt. There were 18 types of failures, and they were categorized into mild, severe, and catastrophic, ordered from least to most harmful. Mild failures included incorrect response formatting or vague answers, severe failures included illogical responses or responses about the wrong character, and catastrophic failures included responses that simply copied the question or asked additional questions instead of giving a response.

I went through each response to check for failures, in a process called data annotation. We found that the LLM titled Hermes and prompt style titled "Answer Interrogative" produced the least failures, as shown in the chart below.



This fellowship allowed Professor Harmon and I to present our work in the International Conference on Interactive Digital Storytelling and attend other presentations. We observed presentations on topics ranging from machine learning to ethics in the field to design patterns.

Faculty Mentor: Dr. Sarah Harmon
Funded by the Fall Research Award