Mariyah Jamali

Generative AI, Cognizant

Final Project Report

Out of curiosity, I completed this project for 2 API's: Hugging Face and Open AI.

Project can be viewed: https://github.com/mjamali1/ai-text-completion-project/tree/main

Hugging Face

Parameters:

Max_length = 30

Top_p = 0.2

Temperature = 0.7

Prompts:

Creative: "The big bear jumped over the…"

Informational: "Tell me about hibernation"

Instructional: "Explain mitosis to me"

Open AI

Parameters:

Max_length = 100

Top_p = 0.9

Temperature = 0.7

Prompts:

Creative: "Finish this sentence: The bear ate…"

Informational: "Explain photosynthesis to me like I'm 5 years old."

Instructional: "How do I make a burger"

Reflection

When does the model perform well?

Both models perform badly more than they give a sufficient output. However, when asked how to make a burger, the Open AI model was able to give cooking instructions. This output included potatoes from somewhere and still included some repetition but it still had the basic structure of a recipe. Overall, Open AI's model performed better than Hugging Face. Hugging Face (gpt2) gave more repetition and unclear answers if the output wasn't plainly repeating the prompt.

When does it struggle (e.g., logical reasoning, niche topics)?

The Hugging Face gpt2 model struggles with more complex or knowledge-intensive tasks. The output to the prompt often had lots of repetition even though its max_length was shorter. The prompt "The big bear jumped over the…" led to the same phrase being repeated dozens of times without any conclusion to the sentence. The Hugging Face model struggled with the creative prompt. In both models' creative prompts I added finish the sentence in one and not the other to see the difference but the same output came in each trial.

In other trials, the model produced insufficient or irrelevant responses. When gpt2 was asked to explain photosynthesis or mitosis, it either veered into strange personal narratives or repeated meaningless phrases like "I'm not sure if I'm being honest or not." These patterns reveal the model's difficulty with logical reasoning, factual accuracy, and staying on topic, particularly when handling educational content or niche subjects.

How might you improve the application (e.g., filtering outputs, validating facts)?

To improve the application, the most obvious switch would be switching to more capable models such as gpt-neo, Mistral-7B, or OpenAI's gpt-3.5-turbo would provide better coherence, factual

grounding, and output control. Additionally, changing the parameters could help control the output. This can be done by lowering the temperature, setting top_p to reduce randomness. If there were a way to detect repetition before it is outputted, that would be helpful as well. Finally, prompt engineering techniques such as few-shot examples or more structured queries can help guide the model's behavior.