Evaluation and Output Analysis

In evaluating the performance of the GPT-2 model from Hugging Face, five prompts were tested to assess the model's coherence, relevance, and ability to follow instructions. The first prompt, *"What is the capital of Senegal?"*, yielded a repetitive and irrelevant output, looping around the phrase "It's a question that has been asked in the past" without ever providing a factual answer. This suggests that GPT-2 struggles with basic factual questions and lacks instruction-following capability.

The second prompt, *"Explain photosynthesis to a 10-year-old"*, produced an unrelated response discussing NIH and NSF grants, revealing the model's tendency to hallucinate contextually inappropriate content. Similarly, the third prompt, *"Explain recursion like I'm five"*, used a very high temperature setting (6.0), which resulted in a chaotic and incoherent response composed of fragmented and nonsensical language. This demonstrates how excessively high temperature leads to loss of semantic control in generation.

The fourth test, *"Continue this story: the dog was playing football"*, showed that while GPT-2 can stay somewhat on topic, the response became overly repetitive ("He was playing with his owner's dog" repeated several times), lacking narrative depth and progression. Finally, a summarization prompt based on a traumatic personal anecdote failed to produce a meaningful summary, instead repeating text and adding hallucinated emotional responses, illustrating the model's lack of summarization training.

Across all five prompts, it became clear that GPT-2 performs poorly on instruction-based tasks, especially when prompts require reasoning, summarization, or factual knowledge. The model also demonstrated a tendency to produce repetitive or off-topic content. Temperature and token length adjustments slightly affected output creativity and verbosity, but did not substantially improve relevance or coherence. This reinforces GPT-2's limitations as a general-purpose instruction-following model.

To improve this application, switching to a more recent instruction-tuned model—such as `flan-t5-base`, `mistral-7b-instruct`, or models from Hugging Face's `text2text-generation` pipeline—would provide better comprehension of natural language commands and generate more accurate and coherent responses. Additionally, implementing a post-processing filter to detect repetition or nonsense, and possibly integrating fact-checking layers, would enhance the overall output quality and reliability.