

Capstone Project Report: AI-Powered Text Completion

By Sarina Iqbal

This project is an interactive AI text completion app powered by Cohere's API. It allows users to input various prompts, adjust the length of tokens to be generated, regulate the level of randomness, and generate text-based responses.

The following table shows the various prompts input to the AI app and the responses generated. The respective settings are also mentioned.

Settings	Prompt	Response
max_tokens = 300 temperature = 0.9	Explain osmosis to a 2-year-old.	Osmosis is a process where water moves from one place to another. Imagine you have a balloon filled with air and a bowl of water. If you put the balloon in the water, the air will start to move out of the balloon and into the water. This is like osmosis, where water moves from a place with a lot of water to a place with less water.
max_tokens = 600 temperature = 0.5	Write a story about dreams	There once was a little girl named Anne who dreamed of becoming a dancer. Every night, she danced with imaginary friends in her dreams. They danced across glass bridges spanning canyons, through the dense branches of mystical forests, and in wide-open fields of blooming wildflowers. Her dreams felt so real that she could almost feel the wind blowing through her hair and the floor beneath her feet.
max_tokens = 100 temperature = 0.3	Continue this story: Sally woke up in a different world..	She woke up in a different world, one in which she had always existed, but one she had never seen before. Everything was new and yet so familiar. The air was crisp and clean, the sky was a vibrant blue, and the sun shone brightly overhead. The trees were green and tall, with a mix of colors in their leaves. The grass was lush and soft, inviting her to lay down and take a nap. But something was wrong. This world was empty. There were no buildings

max_tokens = 600 temperature = 0.7	Share a brief history of Japan	Ancient Japan was a civilization and cultural hub rooted in the Neolithic era, spanning thousands of years until the end of the Edo period in the mid-19th century. Its extensive history is categorized into diverse eras and periods defined by the significant events, cultural movements, and changes in leadership that define them.
max_tokens = 200 temperature = 0.4	Who is the President of USA	I cannot provide you with a definitive answer because my data is limited to information prior to May 2022. However, at the time of my data cutoff, Joe Biden was the President of the United States of America.
max_tokens = 100 temperature = 0.4	What is the derivative of $3x$?	The derivative of $3x$ is 3. Can I help you with any other derivative problems?
max_tokens = 700 temperature = 0.4	What is the integral of $3x$? Show steps.	The integral of $3x$ is $x^2 + x$. Here is a breakdown of the steps:
max_tokens = 700 temperature = 0.4	Show steps to solve integral of $3x$	To find the integral of $3x$, we will use the power rule. The power rule states that the integral of x^n , where n is a real number, is equal to $(x^{(n+1)})/(n+1)$.

Notes:

All of the generated outputs produced grammatically correct and coherent sentences. The various types of prompts, starting from logical questions to story continuations, produced relevant responses. However, in general, inputs with higher temperatures produced more creative and imaginative responses, and those lower temperatures were more factual and straightforward. Prompts with higher numbers of tokens had more in-depth/detailed explanations, and ones with lower numbers were shorter and more focused. One other noticeable pattern was that inputs with larger tokens also seemed to begin with much older, less relevant examples, then approach the point but end before formulating a proper conclusion.

One major limitation of the model is that its training data is made up of information up till 2022, as indicated by the response for factual prompt #5, leading to it being unaware of recent events. For other historical and factual prompts such as prompt #4, the response became too detailed when only a brief history of the country was asked. In most of the cases the response did not reach the maximum number of tokens, leading to somewhat under-explanations. It also struggled with scientific and metaphorical prompts, such as prompt #1, where the response had the correct idea but failed to explain it using real-life or simple examples.

However, the model's logical, grammatical, and quantitative reasoning was satisfactory and seemed highly developed, as there was no significant error, misinformation, or bias in most of the responses. The model, therefore, showed versatility in its ability to handle a diverse range of inputs and produce mostly accurate and creative responses to those.

One interesting finding can be seen from the last two responses: using a prompt with the same objective but different wording and structure produced very different outcomes (without altering the token and temperature settings). The first prompt which was input as two-part questions generated a response that showed the result of the calculation but stopped right before showing the breakdown of the problem. The second prompt, which was designed as a single order/request resulted in a longer output bearing the calculation result as well as the problem breakdown. This depicts the importance of prompt engineering and designing prompts in meaningful ways.

One possible improvement could be incorporating a fact-checking algorithm to prevent any obvious errors, such as scientific and factual errors. Output filters can help in the prevention of offensive language and sensitive content in the generated response, making the app more accessible. Requesting feedback from users on the quality of responses generated can also allow greater sentiment analysis. Keeping track of previous prompts and respective responses can enable the user to continue working on the same problem and receive more tailored responses to newer inputs on the same topic, and even navigate through various options while focusing on a single task.