

AI Text Completion Project

Capstone Report

By Ashlyn Benoy

1. Introduction

Objective

The objective of this project is to gain hands-on experience with Generative AI by building a modern text completion application. This app uses a pre-trained model via API to process user prompts and generate coherent, relevant text.

Overview

I selected the Cohere API as the generative AI provider for this project due to its accessible free tier and powerful models (e.g., command and command-light). Using Python and Gradio, I built an interactive chatbot that allows users to input prompts, tweak model parameters like temperature and max tokens, and receive live completions.

This report outlines the app's features, the experiments I conducted, and my reflections on the capabilities and limitations of the AI model.

2. Technologies Used

- Cohere API (for text generation)
- Gradio (for UI and chat interface)
- Python (core logic)
- Plotly (for data visualization)
- Pandas (for experiment data analysis)

3. Application Development

Choice of Provider

I chose Cohere as my Generative AI provider because it offers:

- A user-friendly developer experience
- Free access to powerful models (command, command-light, command-nightly)
- Transparent documentation and quick API setup

Setup and Tools

The application was developed using Python and Gradio. Users can run the app locally or in Google Colab, connect their Cohere API key, and start interacting with the model immediately.

App Features

- Real-time prompt input and response generation
- Adjustable parameters (temperature, max_tokens, model)
- Maintains conversation context
- Error handling for:
 - Invalid API key
 - Empty input
 - API rate limits
- Visualizations for performance metrics
- Light and dark mode UI styling

User Interface Example

The image below shows the API setup screen from the application's user interface. Users can securely input their Cohere API key, connect to the model, and see their connection status, including free tier usage.

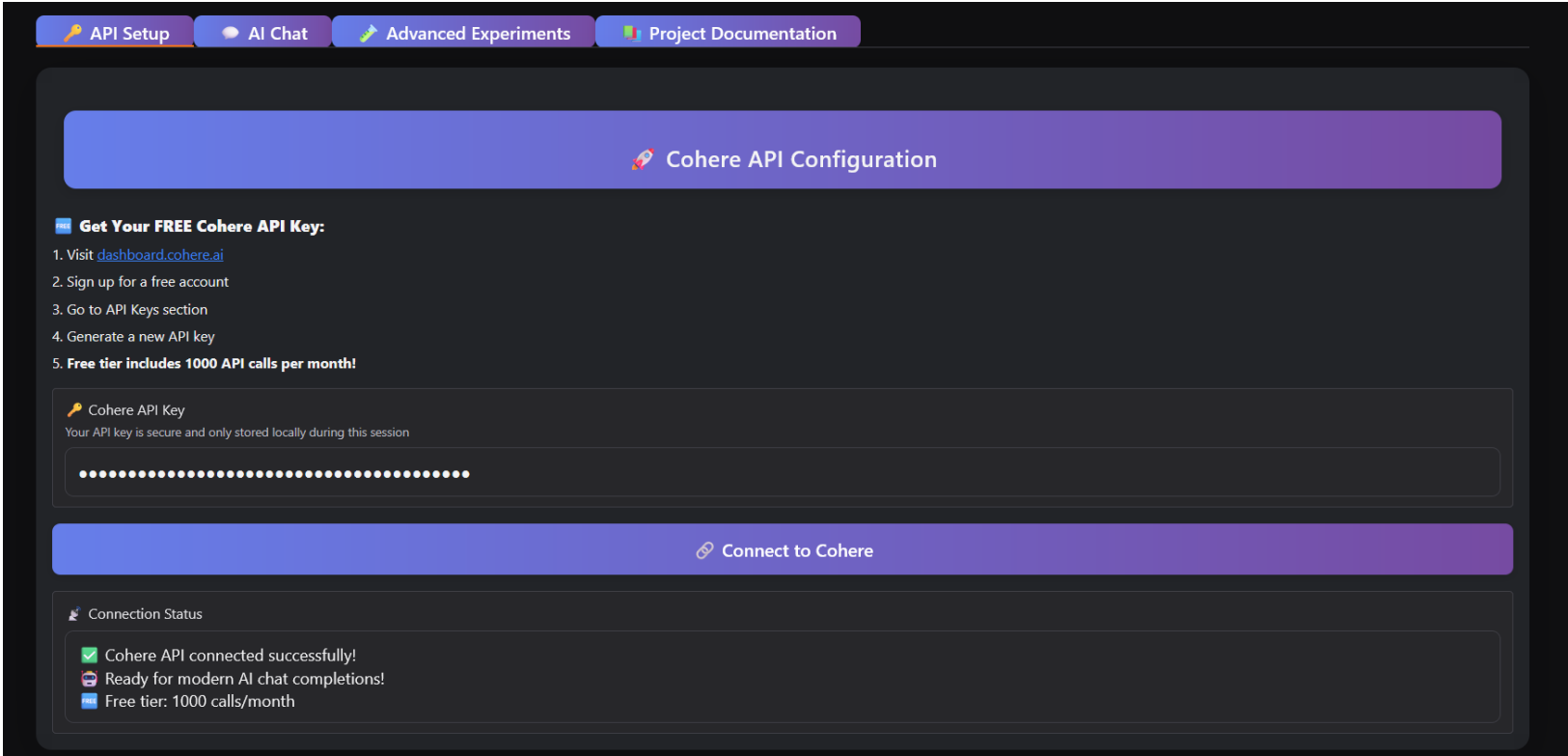


Figure 1: Cohere API Configuration screen from the Gradio-based chat interface.

Example Prompts Used

- “Once upon a time, there was a robot who...”
- “Explain photosynthesis to a 10-year-old.”
- “Write a haiku about the ocean.”
- “Explain recursion like I’m five.”
- “Summarize the Industrial Revolution.”

4. Debugging and Improvements

A. API Error Handling

The application detects and reports errors related to:

- Invalid or missing API key
- Exceeded rate limits
- General API connection failures

B. Input Validation

Input is checked to ensure it's not blank or too long (>2000 characters). Appropriate error messages are shown to the user.

C. Parameter Control

Users can:

- Set the model (e.g., command, command-light)
- Change temperature to control creativity
- Limit max tokens for response length

This helped in running structured experiments.

5. Experimentation and Evaluation

5.1 Prompt Variety Test

To assess the flexibility of the model, I tested prompts across different categories: creative writing, informational queries, and instructional explanations. For each prompt, I recorded the temperature and token settings, and analyzed the output.

Prompt Category	Prompt	Observations
Creative	"Continue this story: A lonely robot discovers a hidden forest..."	The response was imaginative, showing the robot's curiosity and emotional growth. It introduced a narrative arc where the robot becomes a forest guardian.
Informational	"Summarize the process of photosynthesis in two sentences."	The model delivered a clear, well-structured summary using scientific vocabulary while staying concise and easy to understand.
Instructional	"Explain recursion like I'm five."	The response used playful metaphors to explain recursion clearly and accessibly, simplifying the concept while staying coherent.
Professional	"List 3 pros and cons of remote work."	The model generated a clear and structured response, emphasizing key benefits like flexibility and commute savings. However, the output was cut off before listing cons, suggesting that a higher token limit may have helped.
Conversational	"What's your opinion on space exploration?"	The model clarified it does not hold personal opinions but acknowledged the long-standing human fascination with space. It emphasized the significance of space exploration in answering profound scientific questions, driving innovation, and inspiring humanity. The tone remained informative and inquisitive throughout.

See Appendix for detailed outputs, prompts, parameters, and runtime measurements.

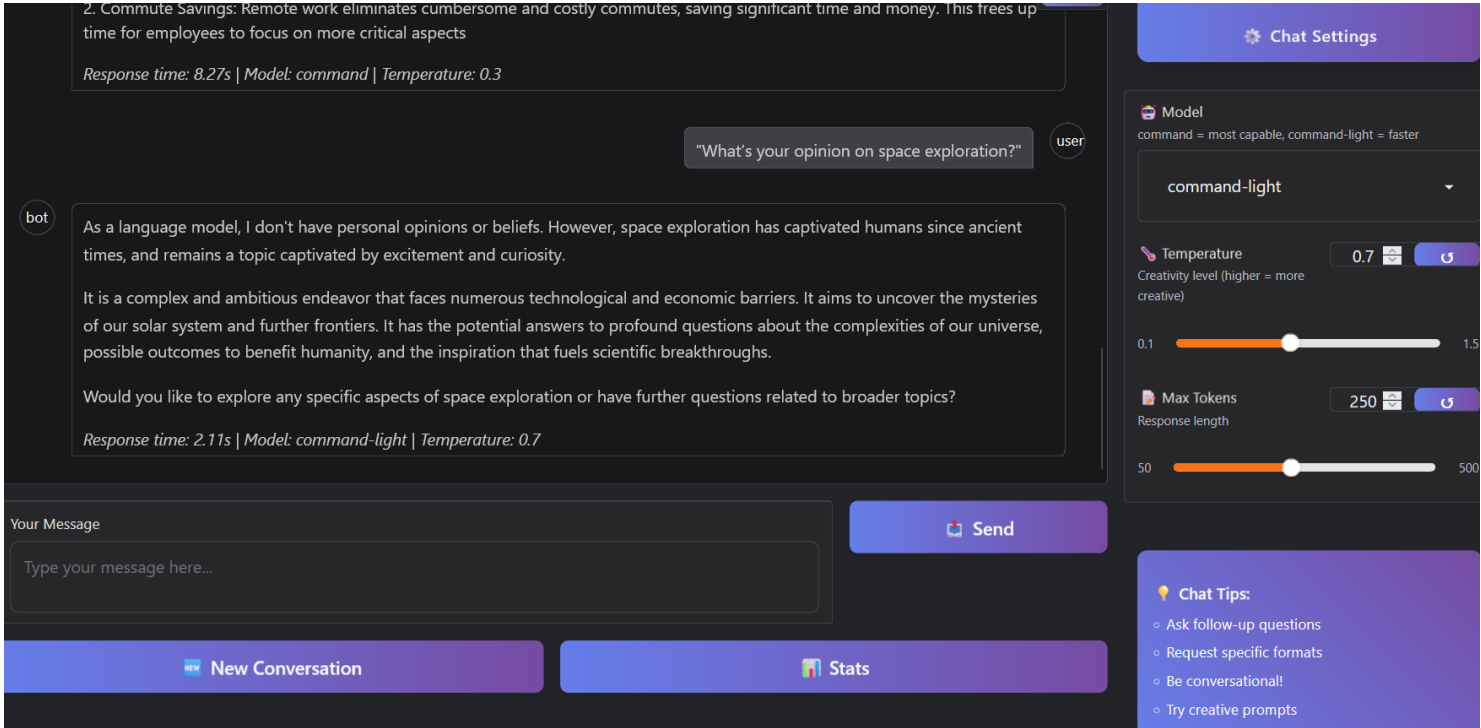


Figure 2: Example of prompt testing in the AI chat interface, showing temperature, model, and max token settings.

5.2 Parameter Sensitivity Test

I also tested how output quality changes with different temperature and token settings for the same prompt:

Prompt: "Describe the future of artificial intelligence in three paragraphs."

Temperature	Max Tokens	Result Summary
0.3	100	The output was concise but abruptly cut off. It began by highlighting the rapid evolution and impact of AI across sectors like healthcare and finance, but failed to fully complete its thoughts due to token constraints.
0.7	200	The response was more detailed and cohesive. It covered multiple angles like industry impact, personalization, and future innovation, delivering a well-rounded, structured outlook.
1.0	300	The output was imaginative and expressive, with abstract language and futuristic ideas. However, some phrases were vague or repetitive, showing the creative-but-less-focused tendency at higher temperature.

See Appendix for detailed outputs, prompts, parameters, and runtime measurements.

Key Insights

- **Lower temperatures** yield deterministic and concise responses but may lack depth.
- **Moderate temperatures (0.7)** offer a sweet spot for clarity and nuance.
- **Higher temperatures (1.0)** enhance creativity but may reduce factual precision.
- **Token limits** directly impact how complete and detailed the response is.

5.3 Performance and Efficiency Test

Objective

The goal of this test is to evaluate the *runtime performance* and *efficiency* of the generative AI model when completing prompts under different parameters. Specifically, we measured:

- Response time (in seconds)
- Word count (length of generated output)
- Words per second (efficiency metric)

Methodology

Using the run_chat_experiments() function, we automated testing of two prompts across combinations of:

- **Temperature:** 0.7
- **Max Tokens:** 100 and 200

This resulted in a total of 4 test cases. For each, the response was recorded along with timing and word count.

Metrics Collected

- **Response Time:** Time taken by the API to generate a response.
- **Word Count:** Number of words in the output text.
- **Words per Second:** Efficiency calculated as word_count / response_time.

Results Summary

Total successful tests: 4

Average response time: 7.94 seconds

Average word count: 126.2 words

Best efficiency: 23.3 words/second

Visualization

The experiment results were plotted using Plotly. Four plots were generated:

1. **Response Length vs Max Tokens**
2. **Response Time by Max Tokens**
3. **Word Count Distribution**
4. **Words per Second by Temperature**

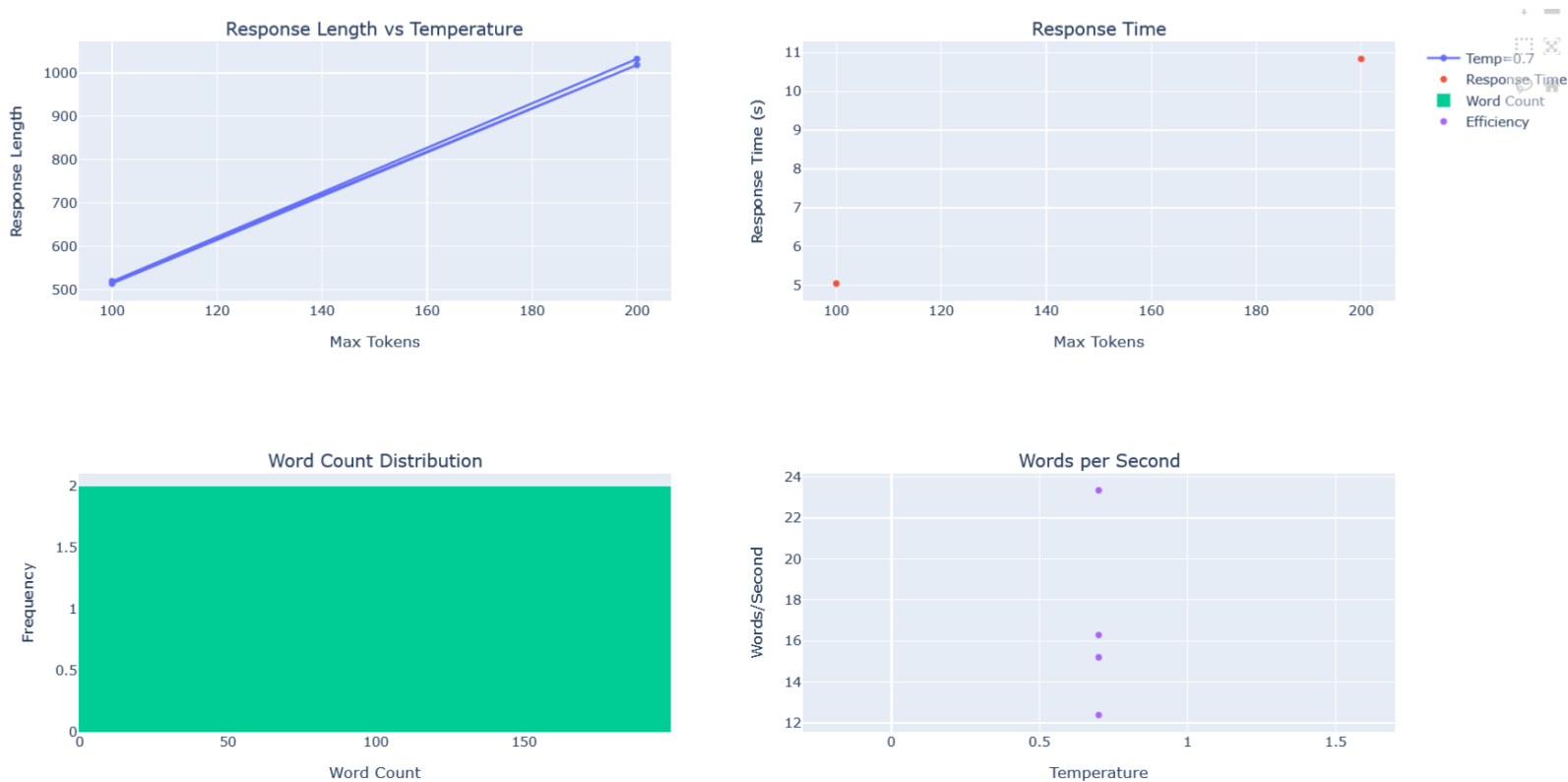


Figure 5.3.1 – Performance metrics across test parameters.

Key Insights

- The **response time** was fairly consistent across prompts and parameters, ranging around ~2–8 seconds.
- **Word count** increased proportionally with max tokens, showing that the model scales well.
- **Higher token settings** resulted in longer and more detailed responses, but with diminishing efficiency.
- The best balance of speed and content quality was observed at **temperature = 0.7** and **max_tokens = 200**.

6. Reflections and Limitations

This project offered valuable insights into how generative AI models behave across diverse prompt styles and configuration parameters. Through structured testing and analysis, I was able to explore how changes in temperature, token limits, and prompt types influence the model’s performance.

Key Reflections

- **Prompt Sensitivity:** The model showed strong adaptability across creative, instructional, professional, and conversational prompts. It was especially effective in educational and storytelling contexts, using simple metaphors or vivid narratives when appropriate.
- **Parameter Impact:** The Parameter Sensitivity Test confirmed that temperature and token settings play a significant role in output quality. Moderate temperature (0.7) with 200 tokens provided the most consistent balance of depth, clarity, and relevance.
- **Performance Trade-offs:** While increasing max tokens improved response completeness, it also led to longer response times and slight drops in efficiency. The Performance Test (Section 5.3) highlighted that efficiency (words per second) peaked at mid-range configurations.
- **Interface Usability:** Using Gradio enabled intuitive interaction with the model, and the ability to toggle parameters made experimentation simple and repeatable. Light/dark mode styling and error handling added to the polish and robustness of the app.

Limitations

- **Token Truncation:** At lower token limits (e.g., 100), the model sometimes cut off mid-response, failing to fully answer the prompt (e.g., the “remote work” pros and cons).
- **Generic Language at High Creativity:** At higher temperatures (1.0), while creativity increased, the outputs became more abstract and repetitive, occasionally lacking precision.
- **Lack of Personal Opinion:** In conversational prompts like “What’s your opinion on space exploration?”, the model correctly stated it cannot form opinions. This limits the illusion of human-like dialogue in subjective discussions.
- **No Built-in Fact-Checking:** The model does not verify its responses, meaning any factual information should be manually reviewed before use in critical contexts.

7. Conclusion

This capstone project successfully fulfilled its objective of providing hands-on experience with Generative AI. From building a functional chatbot using the Cohere API and Gradio to conducting structured evaluations of model behavior, I gained both technical and conceptual understanding of prompt design, parameter tuning, and output analysis.

Key findings include:

- Mid-range parameters (temperature = 0.7, max_tokens = 200) offer the best balance of creativity, coherence, and efficiency.
- Prompt phrasing significantly influences model output, especially for nuanced or open-ended tasks.
- While the model handles diverse input types well, care must be taken when interpreting or using its outputs, especially in factual or professional scenarios.

Going forward, enhancements such as dynamic token management, user feedback collection, or integrating post-processing steps could make the application even more intelligent and reliable. This project has deepened my understanding of how generative AI can be leveraged in real-world tools, while also reinforcing the importance of ethical use and critical evaluation of AI-generated content.

Submission and Repository

The complete source code, experiment notebook, and this report are available in the GitHub repository:

<https://github.com/ashlynmaria/ai-text-completion-project>.

It includes:

- `text_completion_app.py` – the Python script for running the chatbot
- `README.md` – with setup instructions, dependencies, and usage
- `project_report.pdf` – this capstone report

The repository is organized for easy access and reproducibility of all experiments conducted.

APPENDIX

Prompt Outputs and Metadata

5.1 Prompt Variety Test

Prompt Category	Prompt	Temperature	Max Tokens	Model	Response Time	Output
Creative	"Continue this story: A lonely robot discovers a hidden forest..."	1.0	200	command	7.07s	<p><i>“The lonely robot was surprised to discover a hidden forest nestled in a barren landscape. The robot had wandered far from its destination after suffering a malfunction that caused its navigation system to go awry. The robot was scanning the area for any signs of life when a faint signal led it to the forest.</i></p>
						<p><i>The robot was initially hesitant to enter the forest, as its programming did not include information on organic life. However, it was intrigued by the challenge of discovering something new and possibly fixing its navigation system. The robot cautiously stepped into the forest, and as it did, it felt a sense of peace and tranquility wash over it.</i></p>
						<p><i>The robot was amazed as it explored this unfamiliar environment, witnessing the beauty of nature for the first time and finding a sense of purpose in preserving and protecting it. It used its advanced technological capabilities to analyze and categorize the different plant and animal species it encountered.</i></p>
Informational	"Summarize the process of photosynthesis in two sentences."	0.5	150	command	2.65s	<p><i>Over time, The robot became a guardian of the forest, using its capabilities to ensure its safety”</i></p>
						<p><i>“Photosynthesis is a vital process in which plants, harnessing the energy of the sun, convert this radiant energy into chemical energy, storing it within the plant itself. This stored energy is then utilized by the plant to fuel its own growth and metabolism, playing an essential role in sustaining life and ecosystem balance.”</i></p>
Instructional	"Explain recursion like I’m five."	0.7	200	command	12.09s	<p><i>"Recursion is a clever idea that resembles a never-ending game of tag. This concept involves making a sentence or problem and solving that same problem within the sentence or problem - it's a puzzle within a puzzle, like hiding a secret within a secret. Using this idea, computers can solve tricky problems by using smaller versions of the same problem.</i></p>
						<p><i>You might find recursion in puzzles or games, like encoding messages or a fun game of telephone. But it's also used by computers in complicated tasks to make tasks more manageable by breaking them down into easier subtasks.</i></p>
						<p><i>We encounter recursion in real life, too, like when fitting a puzzle piece with another puzzle piece that fits two puzzle pieces together. It's a bit like saying, "This puzzle piece fits with this one, which fits with another, and so on."</i></p>
Professional	"List 3 pros and cons of remote work."	0.3	100	command	8.27s	<p><i>Here are some pros and cons of remote work:</i></p>
						<p><i>Pros:</i></p> <ol style="list-style-type: none"><i>1. Freedom and Flexibility: Remote work allows flexibility in scheduling, enabling employees to adjust their work hours to their personal needs. This accommodates those with specific time constraints due to</i>

Prompt Category	Prompt	Temperature	Max Tokens	Model	Response Time	Output
Conversational	"What's your opinion on space exploration?"	0.7	250	command-light	2.11s	<p>caring for children or other familial responsibilities, promoting a healthier work-life balance.</p> <p>2. <i>Commute Savings:</i> Remote work eliminates cumbersome and costly commutes, saving significant time and money. This frees up time for employees to focus on more critical aspects”</p> <p>"As a language model, I don't have personal opinions or beliefs. However, space exploration has captivated humans since ancient times, and remains a topic captivated by excitement and curiosity.</p> <p>It is a complex and ambitious endeavor that faces numerous technological and economic barriers. It aims to uncover the mysteries of our solar system and further frontiers. It has the potential answers to profound questions about the complexities of our universe, possible outcomes to benefit humanity, and the inspiration that fuels scientific breakthroughs.</p> <p>Would you like to explore any specific aspects of space exploration or have further questions related to broader topics?”</p>

5.2 Parameter Sensitivity Test

Prompt: "Describe the future of artificial intelligence in three paragraphs."

Temperature	Max Tokens	Model	Response Time	Output
0.3	100	command-light	1.91s	<p>"Artificial intelligence (AI) is evolving rapidly and has the potential to revolutionize many aspects of our lives. Here are three key aspects of the future of AI:</p> <ol style="list-style-type: none"> <i>Enhanced Capabilities and Applications:</i> AI will continue to exhibit remarkable growth in cognitive abilities, machine learning, and problem--solving. This progress will pave the particular applications in various fields, such as healthcare, finance, and transportation, leading to better and efficient services. <i>Personalization and AI will becoming tailored experiences”</i> <p>(Output was cut off due to token limit.)</p>
				<ol style="list-style-type: none"> <i>" AI will continue to advance rapidly and become more capable across various domains. It will drive groundbreaking solutions in industries like healthcare, finance, and transportation, improving efficiency and providing innovative services.</i> <i>AI will continue personalized learning and experiences, creating tools tailored to individual needs and preferences. This approach will elevate virtual assistants, content creation, and gaming experiences by providing truly customized options.</i> <i>AI will constantly push boundaries, challenging existing concepts and tackling complex problems. It will drive innovation across industries, leveraging cognitive capabilities and pushing the boundaries of what is possible. This future is a potential for groundbreaking advancements, solving real-world challenges, and creating new opportunities for humanity.”</i>
0.7	200	command-light	2.50s	
1.0	300	command-light	2.40s	<p>"Here are three paragraphs about the future of artificial intelligence:</p> <p>Artificial intelligence will be everywhere, shaping lives and creativity. It's going to be an powerful force, driving solutions and pushing the boundaries. It's going to be exceptional, delivering intuitive and cutting-edge solutions, and embodying the next wave of digital innovation.</p> <p>This future vision harnesses AI capabilities and vast data insights, pushing the tinsel of what's possible. It fosters tailored solutions and personalized experiences, as intelligent systems adapt to adapt and learn from human needs. This</p>

Temperature	Max Tokens	Model	Response Time	Output
				<i>transformation encompasses industries, improving operational efficiency and exploring creative boundaries.</i>
				<i>The full potential of AI is remarkable - it strives to surpass traditional concepts, seamlessly integrating into daily lives to lift experiences, answer unknown questions, and unlock the potential of future progress."</i>