AI-Powered Text Completion Using Cohere: A Technical Report

By Chris Malit

1. Introduction

Generative AI has made significant advancements in text automation, offering new possibilities for content creation, education, and information retrieval. This project explores text generation using Cohere's AI model, focusing on prompt design, model interaction, and response evaluation.

The objective was to build a Python-based application that interacts with a trained AI model to generate meaningful and coherent text. By adjusting model parameters such as temperature and token length, the study aimed to analyze the impact of these settings on response quality.

2. Development Process

Environment Setup

The project was developed in Python, utilizing Cohere's API for text completion. The setup involved:

- Installing the required library:
- pip install cohere
- Securing the API key as an environment variable to enhance security.
- Writing a Python script that accepts user input, interacts with the Cohere model, and returns generated text.

Application Features

The final implementation included:

- A user-friendly interface that accepts multiple prompts within a session.
- An option for users to exit the program when needed.
- Error handling mechanisms to prevent crashes caused by invalid input.
- Adjustable parameters such as temperature and max tokens to refine responses.

3. Experimentation and Parameter Tuning

Impact of Temperature

Temperature controls how deterministic or creative the AI's response is. Lower values generate structured and predictable outputs, while higher values increase variety and unpredictability.

Temperature Observations

0.3	Responses were highly structured and lacked diversity.
0.7	Balanced coherence and creativity, resulting in relevant yet engaging text.
1.0	Increased randomness, producing imaginative responses but occasionally lacking focus.

Lower temperatures provided clear and factual responses, while higher values encouraged originality, sometimes at the expense of consistency.

Impact of Max Tokens

Max tokens define the length of AI-generated responses. Adjusting this parameter affects the depth and detail of the output.

Max Tokens Observations

600	Responses were brief and to the point.
1200	More detailed responses but occasionally repetitive.

A setting between 800-1000 tokens generally produced balanced responses without excessive verbosity.

---

4. Debugging and Refinements

Throughout development, several improvements were made:

- **Input validation** to prevent empty prompts from causing errors.
- **Enhanced user experience** by allowing multiple prompts in a single session.
- **Improved error handling** to display clear messages in case of API failures.

The refined application ensures seamless execution with clear output and minimal interruptions.

---

5. Evaluating AI Performance

Strengths

- Strong ability to generate **engaging creative content**, including stories and poetry.

- Provides **coherent summaries** when tasked with condensing information.

- Delivers **structured explanations** for educational topics.

Limitations

- Occasional **logical inconsistencies**, particularly in reasoning-heavy tasks.

- Some **factual inaccuracies**, especially when handling complex queries.

- Long responses sometimes contain **redundant phrasing**.

While the AI produces high-quality results across various domains, further refinement is needed to improve **accuracy and logical reasoning**.

---

6. Potential Enhancements

Future improvements could include:

- **Fact-checking mechanisms** to improve accuracy.

- **User-controlled customization** allowing real-time parameter adjustments.

- **Pre-processing prompts** for enhanced relevance and clarity.

These refinements would strengthen the reliability and usability of AI-generated text.

---

7. Conclusion

This project demonstrated the practical applications of **AI-driven text completion**, showcasing Cohere's ability to generate structured and engaging responses. By experimenting with prompt design and model parameters, the study highlighted the strengths and limitations of AI-generated text. While the model excels in creativity and summarization, improvements in logical consistency and factual accuracy would further enhance its capabilities.

Generative AI continues to evolve, providing valuable tools for education, content creation, and automated communication. Future development efforts should focus on **optimizing model performance** for real-world applications.