

Experimentation and Evaluation Report

By Roman Nekrasov

1. Introduction

The goal of this part of the project was to test the text-completion application across a variety of prompt types and evaluate the quality, coherence, and limitations of the model's outputs. We used five different prompts—creative, informational, instructional, opinionated, and technical—all with the same generation settings (temperature = 0.7, max_tokens = 150, top_p = 1.0).

2. Prompts, Responses, and Observations

Prompt (settings)	Response	Observations
Creative Continue this story: "The old lighthouse keeper opened the door and saw..."	<p>a young woman standing on the doorstep, drenched from head to toe. Her hair was plastered to her face, and her clothes clung tightly to her body. She looked up at him with big, pleading eyes and said, "Please, you have to help me. My boat capsized and I've been stranded out at sea for hours."</p> <p>The lighthouse keeper's heart went out to her and he quickly ushered her inside, wrapping a thick blanket around her shoulders. He made her a cup of hot tea and listened as she recounted her harrowing ordeal.</p> <p>As she spoke, the lighthouse keeper couldn't help but feel a sense of unease. Something about her story didn't quite add up. But he pushed aside his doubts</p>	Excellent narrative continuation; builds suspense. Coherent and on-theme.
Informational Summarize in one sentence: "Photosynthesis is the process by which plants..."	Photosynthesis is the process through which plants utilize sunlight to transform carbon dioxide and water into glucose and oxygen.	Perfectly accurate and concise summary.

<p>Instructional Explain recursion like I'm five years old.</p>	<p>Recursion is like when you use a magic mirror to see yourself, but then you see another mirror inside that mirror, and another mirror inside that one, and it keeps going on forever. It's like looking at a picture of yourself in a picture of yourself in a picture of yourself, and so on.</p>	<p>Great analogy, clear for a child. Engaging and understandable.</p>
<p>Opinionated What are the pros and cons of remote work?</p>	<p>Pros of remote work:</p> <ol style="list-style-type: none"> 1. Flexibility: Remote work allows employees to have more control over their schedules and work environment. 2. Cost savings: Companies can save money on office space and utilities by allowing employees to work remotely. 3. Increased productivity: Many employees report being more productive when working remotely due to fewer distractions and interruptions. 4. Better work-life balance: Remote work can help employees better balance their work and personal lives. 5. Access to a wider talent pool: Companies can hire employees from anywhere in the world, increasing diversity and access to specialized skills. <p>Cons of remote work:</p> <ol style="list-style-type: none"> 1. Communication challenges: Remote work can make it harder for employees to collaborate and communicate effectively with colleagues. 2. Feelings of isolation: Remote workers may 	<p>Balanced structure, but truncated. Needed more tokens to finish the list completely.</p>

<p>Technical Difference between list and tuple in Python.</p>	<p>In Python, a list is a mutable data type, meaning that it can be changed after it is created. Lists are created using square brackets [] and can contain elements of different data types. Elements in a list can be modified, added, or removed using various built-in methods.</p> <p>On the other hand, a tuple is an immutable data type, meaning that it cannot be changed after it is created. Tuples are created using parentheses () and can also contain elements of different data types. Once a tuple is created, its elements cannot be modified, added, or removed.</p> <p>In summary, the main difference between a list and a tuple in Python is that a list is mutable while a tuple is immutable.</p>	<p>Accurate and clear; informative for beginners.</p>
--	---	---

3. Analysis of Outputs

- **Strengths:**

- Handles creative storytelling with coherent details.
- Produces accurate definitions and child-friendly analogies.
- Correctly formats lists and paragraphs.

- **Weaknesses & Limitations:**

- Token limits led to truncated lists (opinion prompt).
- Over-verbose on technical explanations.
- Occasional tone shifts between formal and informal.

- **Biases & Inaccuracies:**

- No major factual errors detected.
- Tends to generalize viewpoints rather than deep nuance.

4. Reflection & Improvement Ideas

- **When it shines:** Creative and factual summaries.
- **Where it struggles:** Long enumerations under token constraints; very niche or highly technical topics.
- **Improvements:**
 1. Increase `max_tokens` for completeness.
 2. Add post-processing to detect and re-prompt when outputs are truncated.
 3. Integrate retrieval-augmented generation (RAG) for grounded factual responses.

5. Conclusion

This experiment demonstrates that the GPT-3.5-turbo model in our application can generate coherent, contextually appropriate text across multiple domains. While it excels at creative and educational prompts, careful tuning of generation parameters and additional post-processing are needed for consistent completeness and factual reliability.