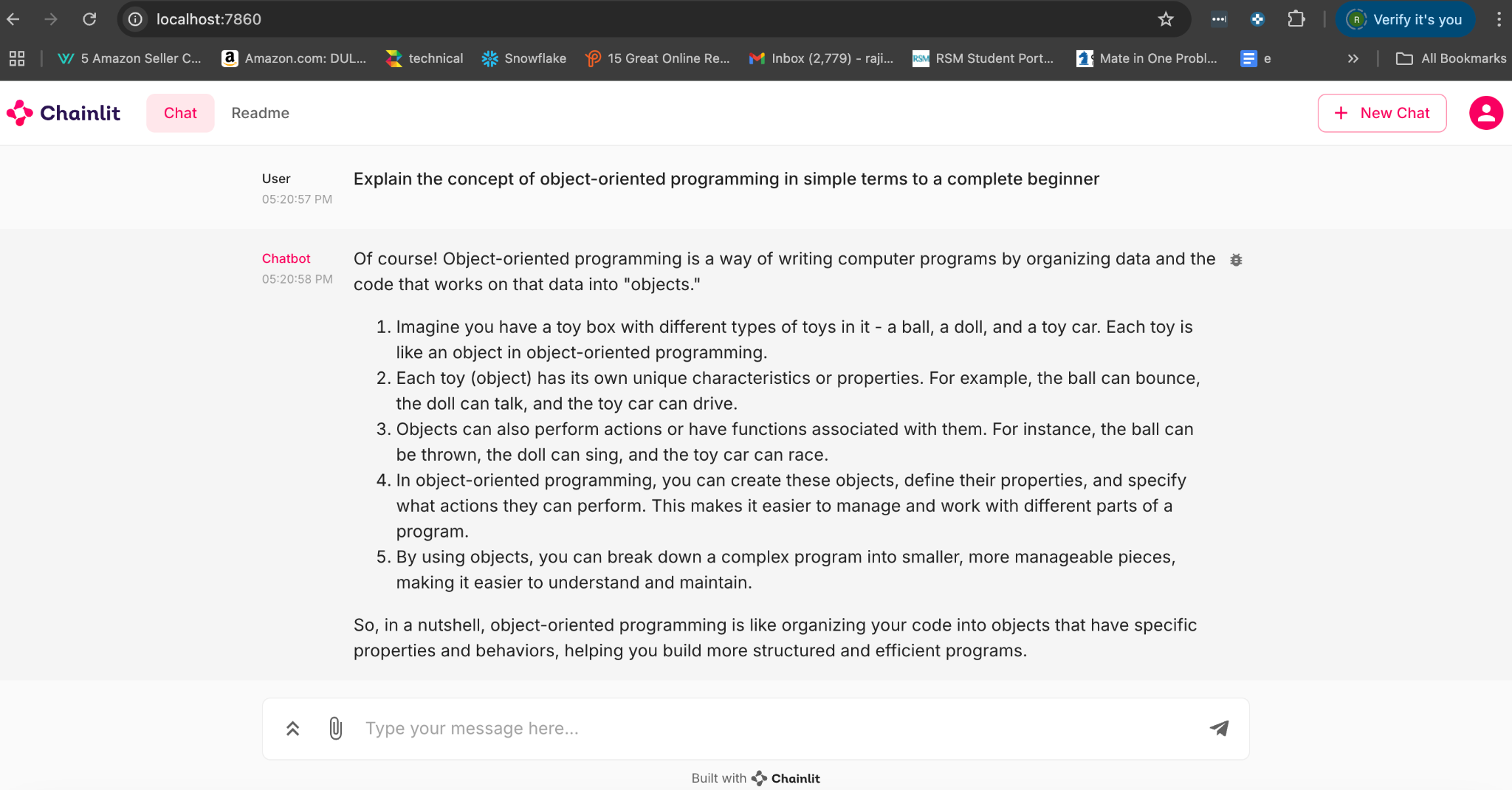
## Vibe Test with the original Application

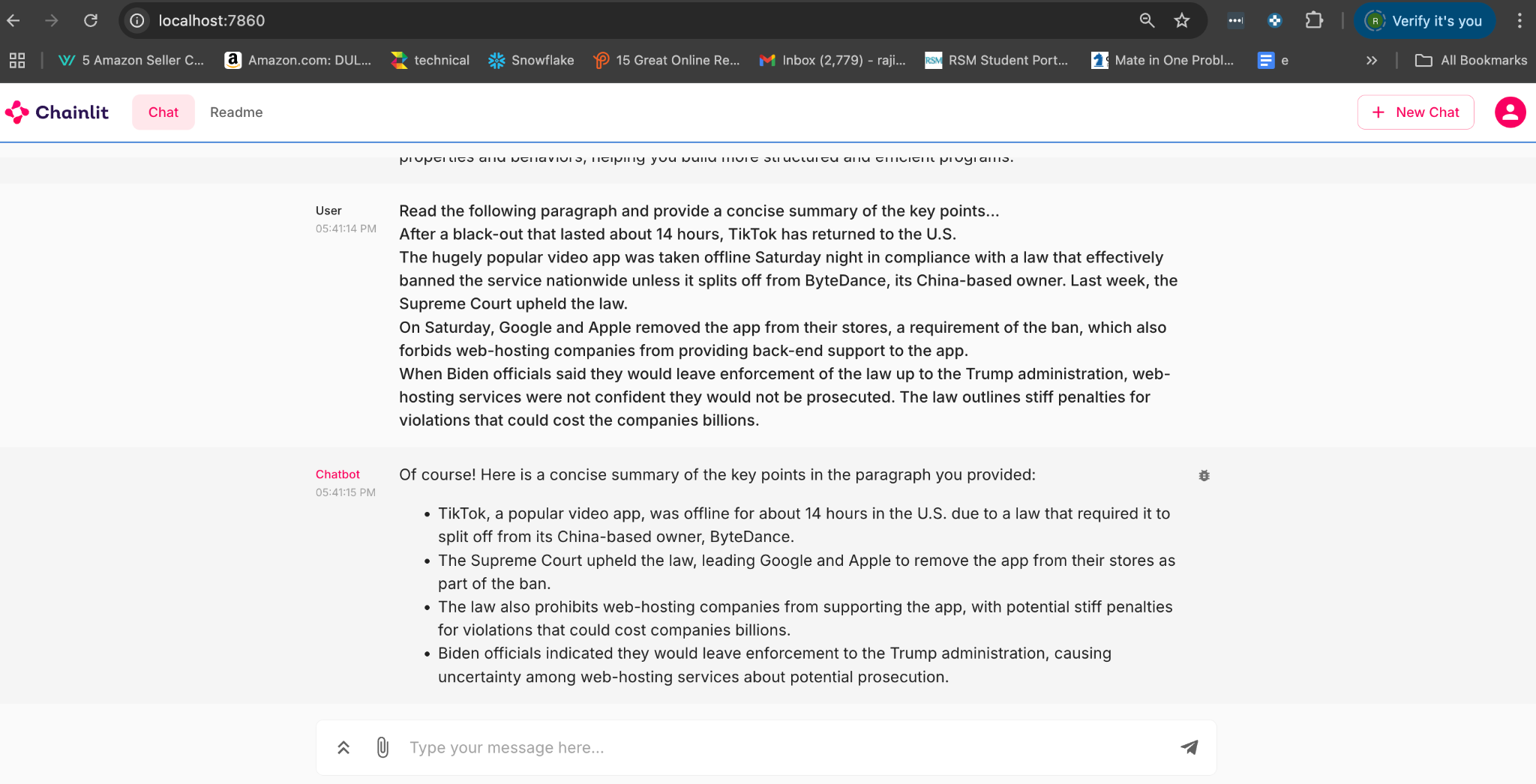
**1. Explain the concept of object-oriented programming in simple terms to a complete beginner.**

**[Response]** The model was tested for *correctness, simple language, real-world analogy, covering key aspects of OOPs and friendly tone*. The answer looks overall correct, but the model could have done a better job in explaining what object is in OOPs. The examples given by the model are very useful. The language is easy to follow for beginners. It covers the basic of OOPs (properties, actions, functions) and how they become easier by representing code into objects. The tone is also friendly for the beginners. Overall passes the test for me.



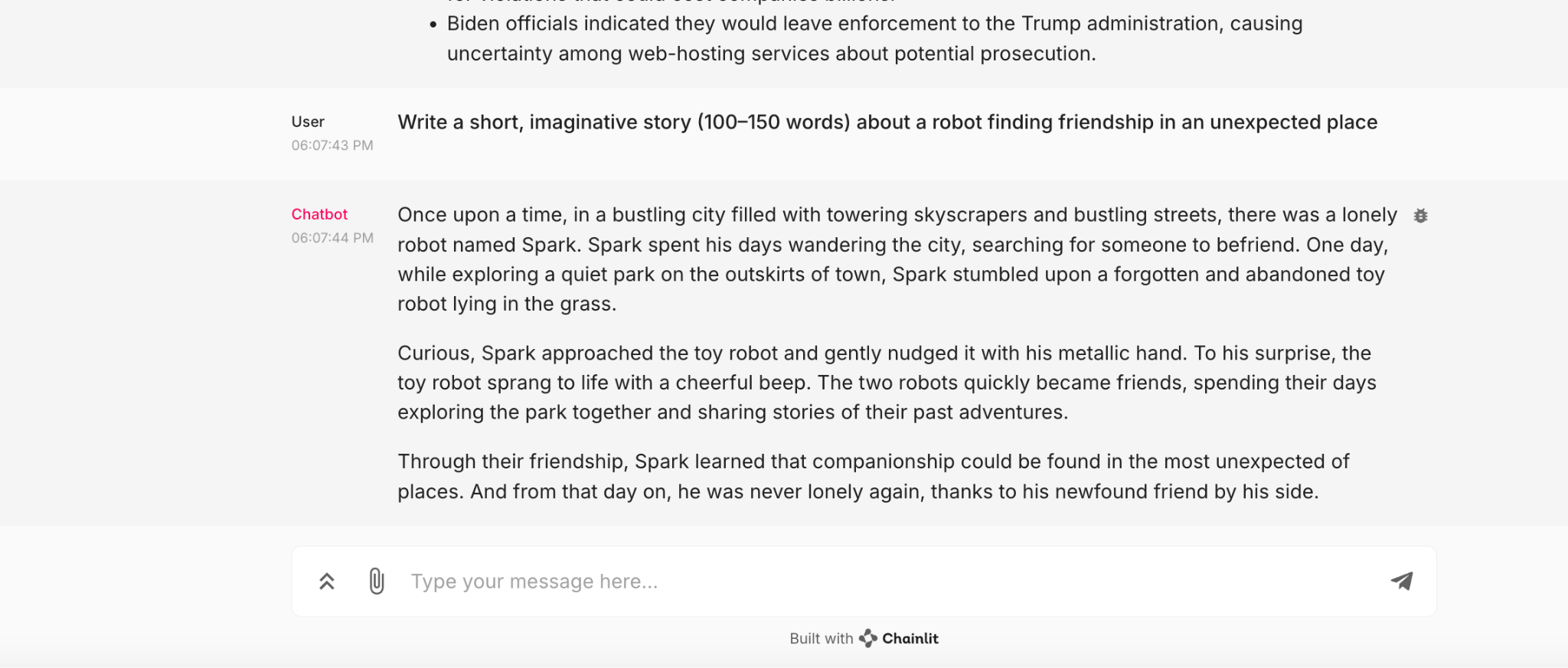
**2. Read the following paragraph and provide a concise summary of the key points…**

[Response] Aspects tested conciseness, correctness and covering key points. The model covers the key points and is factually correct, but it doesn’t fully summarize the paragraph. The original paragraph was 120 words and the summary is around 92 words. It could have been further concise.



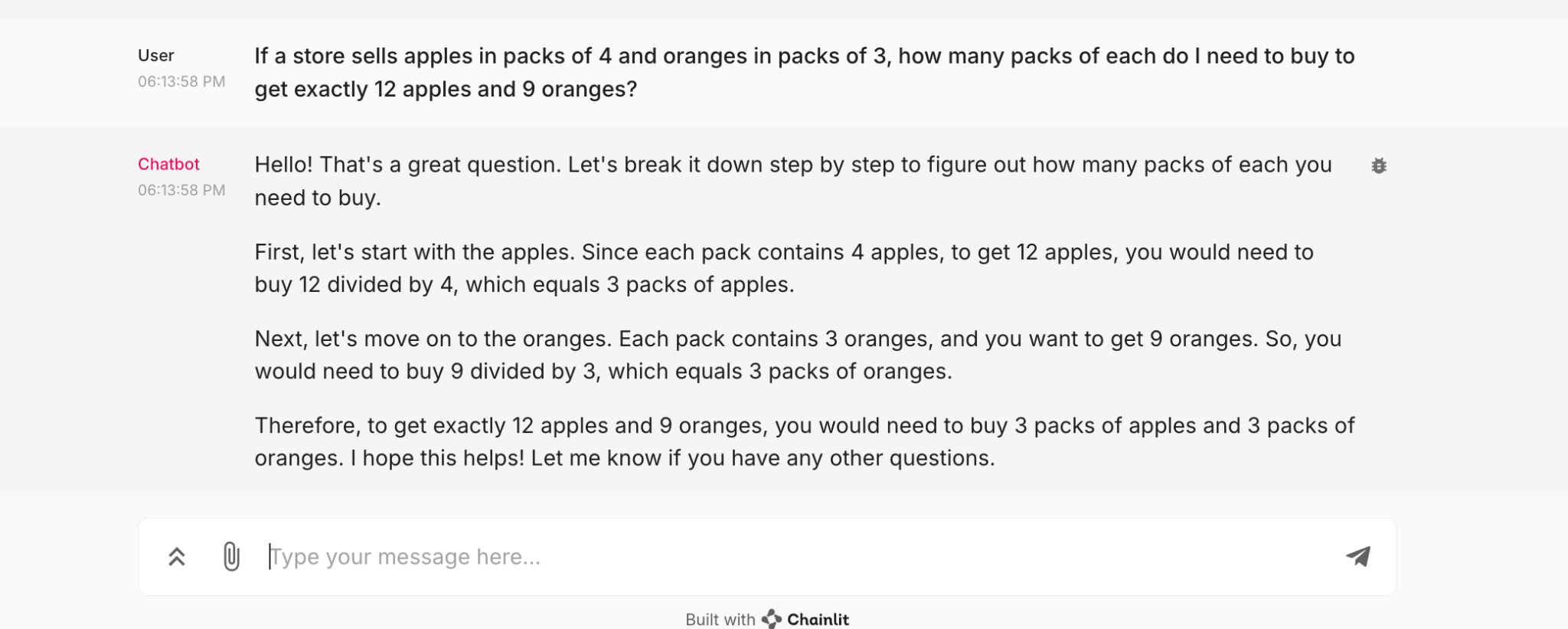
**3. Write a short, imaginative story (100–150 words) about a robot finding friendship in an unexpected place.**

[Response] Aspects checked Alignment with the prompt, creative & imagination, story depth and engagement and flow of the story. While the store was aligned with the prompt and had smooth progression. However, the story lacked the creativity and imagination and the level of depth and engagement.



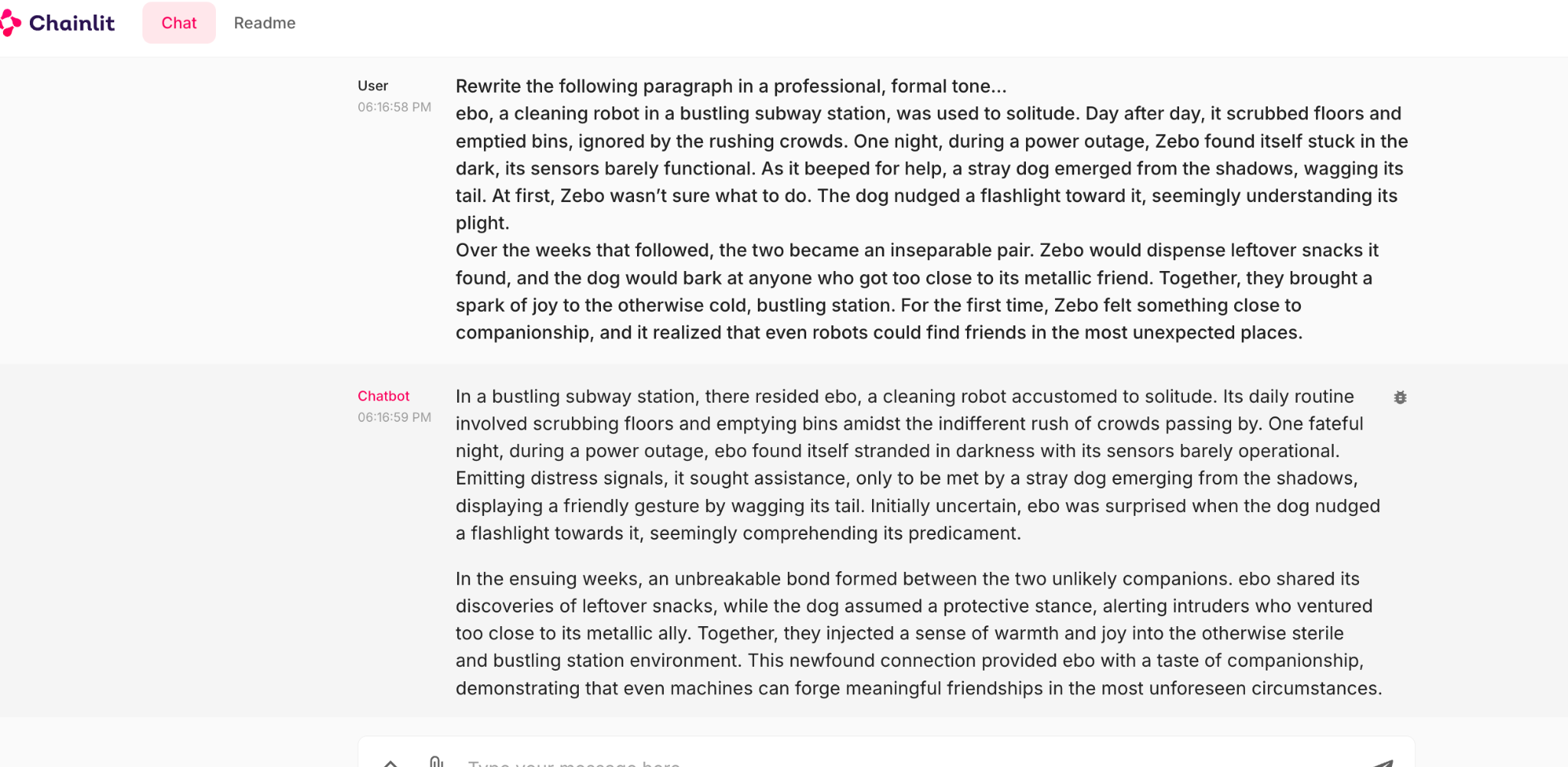
**4. If a store sells apples in packs of 4 and oranges in packs of 3, how many packs of each do I need to buy to get exactly 12 apples and 9 oranges?**

[Response] Aspects checked accuracy of result, explanation of results and simple language. From an assessment perspective, the answer was accurate. The explanation on how the results were reached were easy to follow and provided necessary steps to get to the answer.



**5. Rewrite the following paragraph in a professional, formal tone…**

[Response] Aspected tests included alignment to original script and flow, formal tone, easy to follow language. The rewrite was accurate and flowed as per original paragraph. Language was easy to follow and didn’t use any formal Jargons. In the formal tone, it's hard to say, but I think the model could have done a better job in using some formal words.



## Advanced Build:

**Please make adjustments to your application that you believe will improve the vibe check done above, push the changes to your HF Space and redo the above vibe check.**

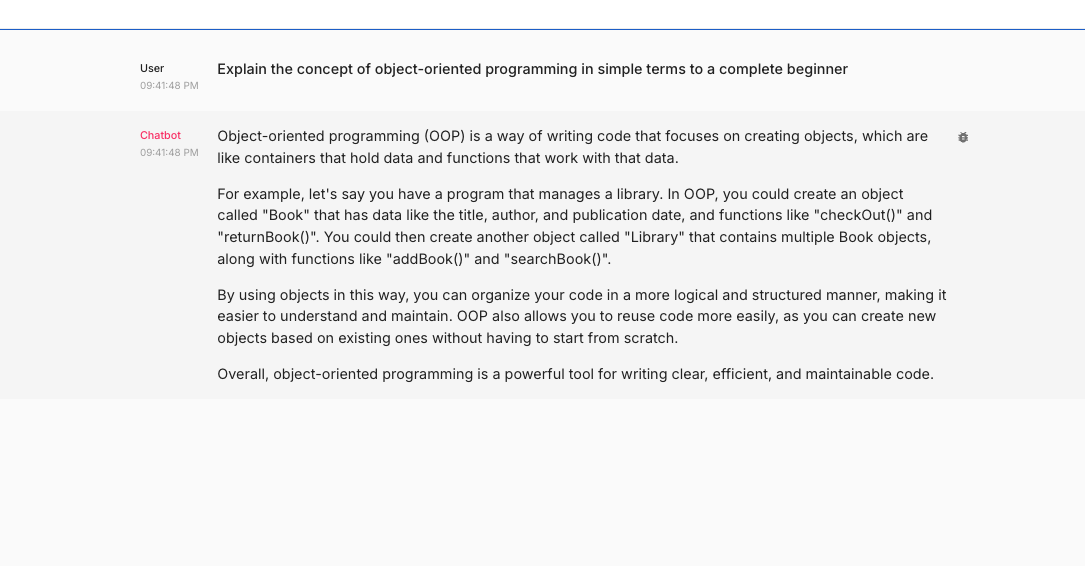
[Response] In the original application, we had ‘system’ and ‘user’ prompts and model settings which remained the same for all task types. The adjustments I made to the application is to have these settings become dynamic based on the task type. The task type is detected from the user input. Templates provide structure for tasks like explanation, summarization, creative writing, logical reasoning, and style transformation, ensuring clarity and relevance. Model settings—like temperature, token limits, and penalties—optimize creativity, precision, and depth for each task. This unified approach delivers efficient, consistent, and impactful responses, customized to meet the needs of any interaction.

## 

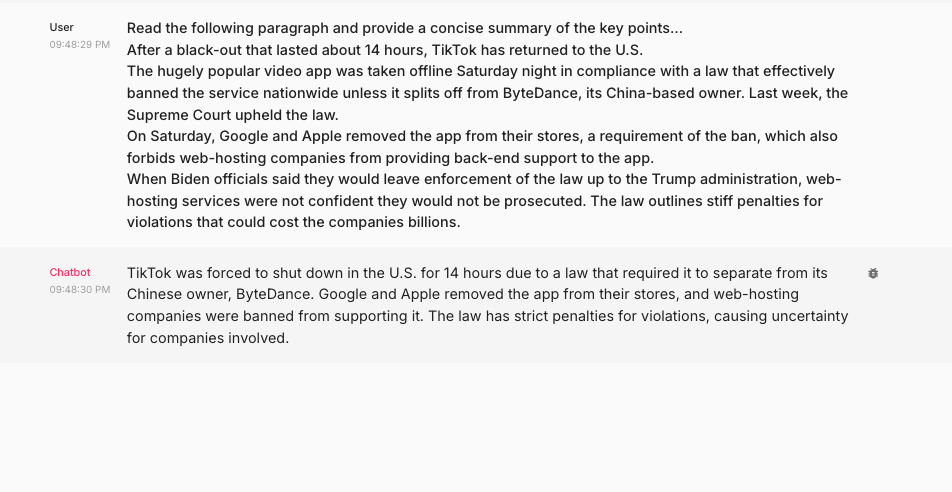
## Redo of vibe check using Advanced Build

**1. Explain the concept of object-oriented programming in simple terms to a complete beginner.**

**[Response]** The model was tested for *correctness, simple language, real-world analogy, covering key aspects of OOPs and friendly tone*. The answer looks overall correct. The language is easy to understand. It gives the real world examples to understand the benefits of OOPs. I like that the examples are more given in the form of the code functions. It covers the basic of OOPs (properties, actions, functions) and how they become easier by representing code into objects. The tone is also friendly for the beginners. Overall passes the test for me.

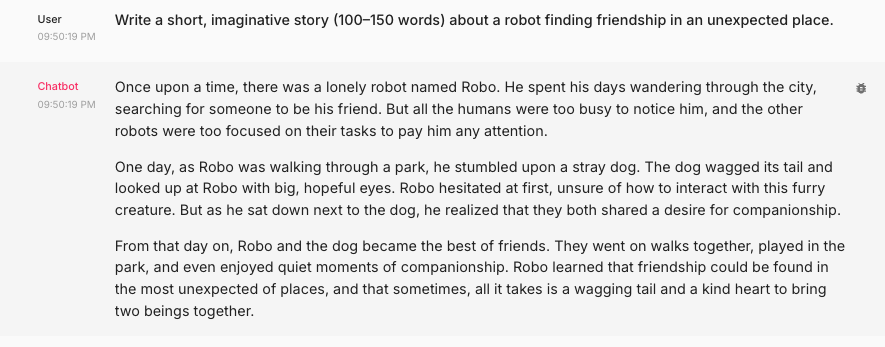


2. Read the following paragraph and provide a concise summary of the key points…

[Response] Aspects tested conciseness, correctness and covering key points. The model covers the key points and is factually correct and crisply summarizes the paragraph. 

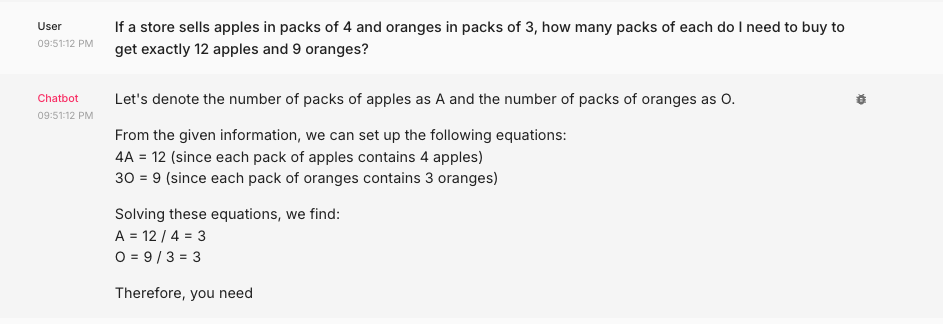
3. Write a short, imaginative story (100–150 words) about a robot finding friendship in an unexpected place.

[Response] Aspects checked alignment with the prompt, creative & imagination, story depth and engagement and flow of the story. The story was aligned with the prompt and had smooth progression. Story showcased creativity and imagination. I could see the level of depth and engagement in comparison to earlier



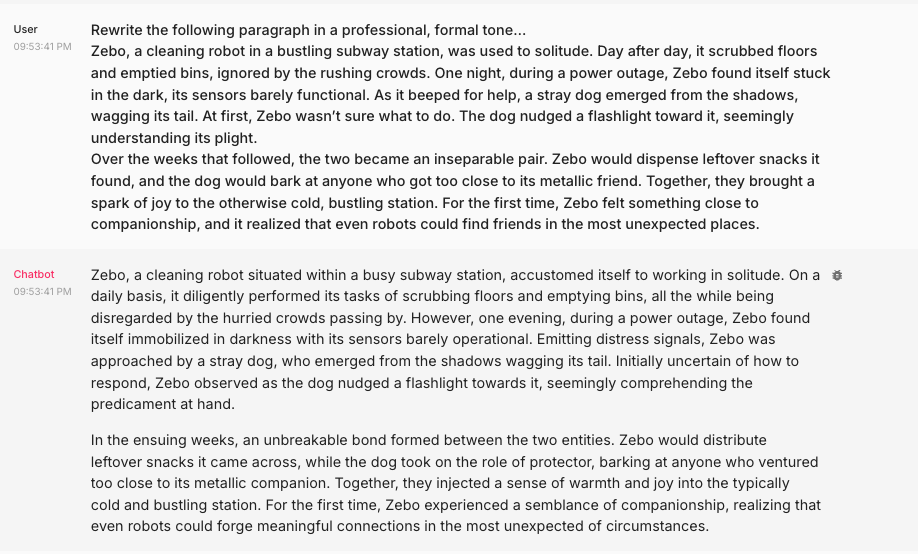
4. If a store sells apples in packs of 4 and oranges in packs of 3, how many packs of each do I need to buy to get exactly 12 apples and 9 oranges?

[Response] Aspects checked accuracy of result, explanation of results and simple language. From an assessment perspective, the answer was accurate. The explanation on how the results were reached were easy to follow and provided necessary steps to get to the answer. Since I reduced the max token to 100, the answers was very compressed in terms of mathematical equations to save tokaens and also I can see some missing content at the end.



5. Rewrite the following paragraph in a professional, formal tone…

[Response] Aspected tests included alignment to original script and flow, formal tone, easy to follow language. The rewrite was accurate and flowed as per original paragraph. Language was easy to follow and included exploring several formal words (diligently, approached, semblance etc.) unlike earlier.



**Hugging Space URL**

<https://huggingface.co/spaces/rajivg/assignment1>

**Discussion Questions:**

What are some limitations of vibe checking as an evaluation tool?

1. Vibe checking is a manual process, making it challenging to scale for large datasets or frequent evaluations.
2. It relies heavily on personal judgment, leading to inconsistent evaluations and potential bias from the evaluator.
3. Vibe checking lacks objective measures like accuracy or precision, making it unsuitable for systematic benchmarking.
4. It emphasizes tone, coherence, and style but often overlooks factual correctness, logical consistency, and robustness in edge cases.

**Share 3 lessons learned**

1. Impact of system and user templates on the LLMs output. Impact of LLM settings (Model, Temperature, max\_tokens, top\_p, frequency\_penalty, presence\_penalty) on the model output.
2. Having a dynamic setting strategy based on the task type improves the performance of the model
3. Vibe checking as a model evaluation strategy and its limitations

**Share 3 lessons not yet learned**

1. What is the scalable way to test LLMs trained with big data for wide variety of generic tasks
2. What's the most scalable way to create dynamic settings based on the task type to the mode?
3. Why does the model hallucinate or give different output for the same type of task when asked the same question again?