Comparative Analysis: Llama3 & Mistral

# 1.Executive Summary:

* This document analyzes the performance of GroqCloud in AI acceleration, specifically using the Mixtral-8x7B-326B8 and LLama3-8B-8192 models.
* The Mixtral model exhibited superior performance in both question and answer generation, delivering accurate and relevant responses.
* Conversely, the LLama3-8B-8192 model demonstrated less effectiveness in question generation but provided **contextually relevant answers**, with potential for improvement.

# 2. Model Execution:

The text is divided into segments called chunks, and a similarity search occurs based on the specified batch and chunk sizes. The most relevant segments are extracted and stored in Vector DB document embeddings, from which the final answer is generated.

# 3.Models Analysis ( By me )

## 3.1 Difference Between Models

Difference Between LLama2, LLama3-8B-8192, Mixtral-8x7B-326B8 Model

| **Feature** | **LLama2** | **LLama3-8B-8192** | **Mixtral-8x7B-326B8** |
| --- | --- | --- | --- |
| Model Type | Language Model | Language Model | Language Model |
| Developer | Meta | Meta | Naver & SNU (Seoul National University) |
| Parameters (B) | 2B, 7B, 13B, 30B, 65B | 8B | 56B (8 models of 7B each) |
| Max Tokens | Variable (context size) | 8192 tokens | 8192 tokens |
| Architecture | Transformer | Transformer | Transformer |
| Training Data | Diverse internet text | Diverse internet text | Diverse internet text |
| Purpose | General-purpose NLP tasks | General-purpose NLP tasks | General-purpose NLP tasks |
| Training Approach | Supervised learning | Supervised learning | Mixture of supervised and reinforcement learning |
| Unique Feature | Efficient scaling, multi-task abilities | Extended context length | Combines multiple 7B models into a single model |
| Open Source | Yes | Yes | Yes |
| Applications | Chatbots, Text generation, Translation, etc. | Chatbots, Text generation, Translation, etc. | Chatbots, Text generation, Translation, etc. |
| Release Date | 2023 | 2024 (expected) | 2023 |

## 3.2 Performance Testing of Mixtral-8x7B-326B8:

An ensemble model comprising eight models with seven billion parameters each, totaling 56 billion parameters, offers promising potential in natural language processing (NLP).

### 3.2.1 Results Obtained by Mistral Model:

1. Accuracy Assessment:

* Relevant paraphrasing for concise responses.
* For long paragraphs performance is good.
* For one liners doesn’t give exact responses but provide relevant paragraphs with similar meaning.
* If not able to extract content from the pdf generate relevant answers from reputable online content.

1. Efficiency Evaluation

* When chunk\_size = 1000 and batch size = 200 only able to read approx 60 pages
* .When changed chunk\_size = 2000 and batch size = 400 able to read 150 pages.
* When changed chunk\_size = 3000 and batchsize = 500 is able to read all the pages from the uploaded file Understanding ASPICE R7 May 2022 (live learn).pdf

3.User Experience:

* Model response is satisfactory as a user.
* Good response for question generation also.
* Good response for long answers and accuracy in short answers.

## 3.3 Performance Testing of Llama3-70B-8192:

### 3.3.1 Results Obtained by Llama3:

1.Accuracy Assessment:

* Shows lower accuracy compared to Mixtral-8x7B-326B8.
* It struggles to provide precise responses for one-liners but can generate relevant paragraphs with similar meaning.
* It performs well for longer paragraphs.

2.Efficiency Evaluation:

* The model's efficiency is subpar. With a chunk\_size of 1000 and batch size of 200, it can only process around 60 pages. Even after adjusting parameters to a chunk\_size of 2000 and batch size of 400, it can only handle 150 pages.

3.User Experience:

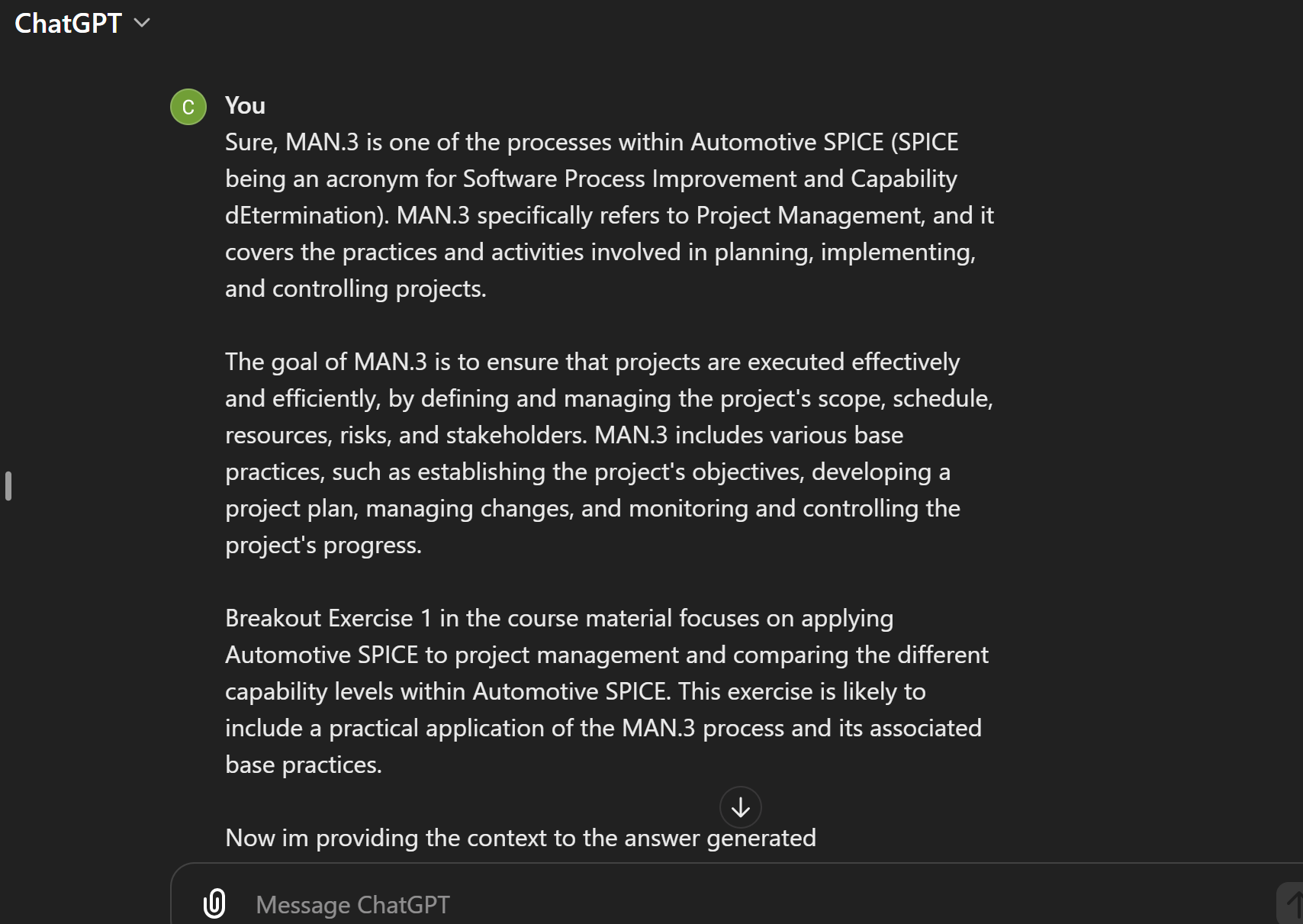
* The model's response is satisfactory from a user's perspective, but its performance on very short one-line answer generation is poor.

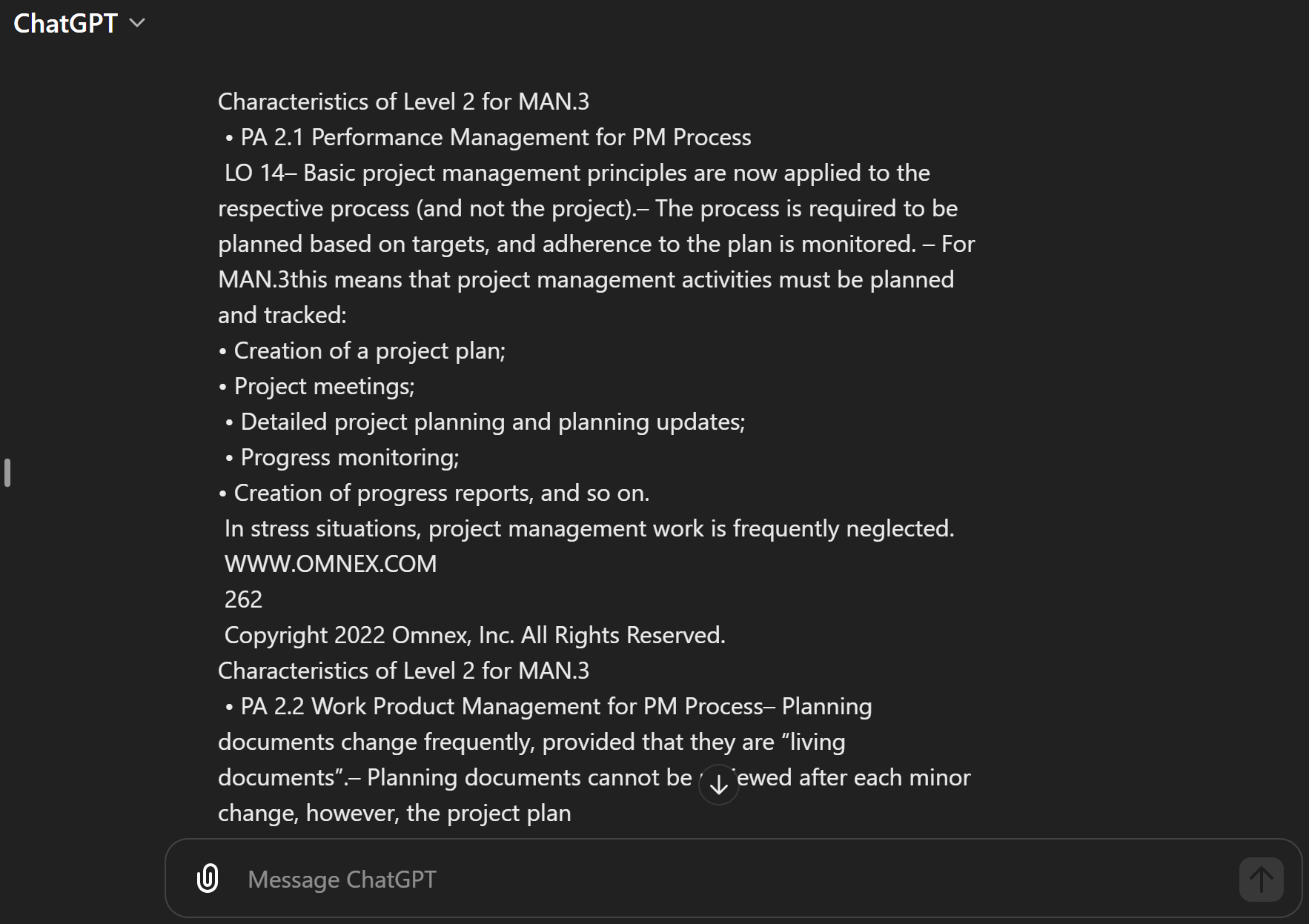
# 4. Model Analysis (Mistral)

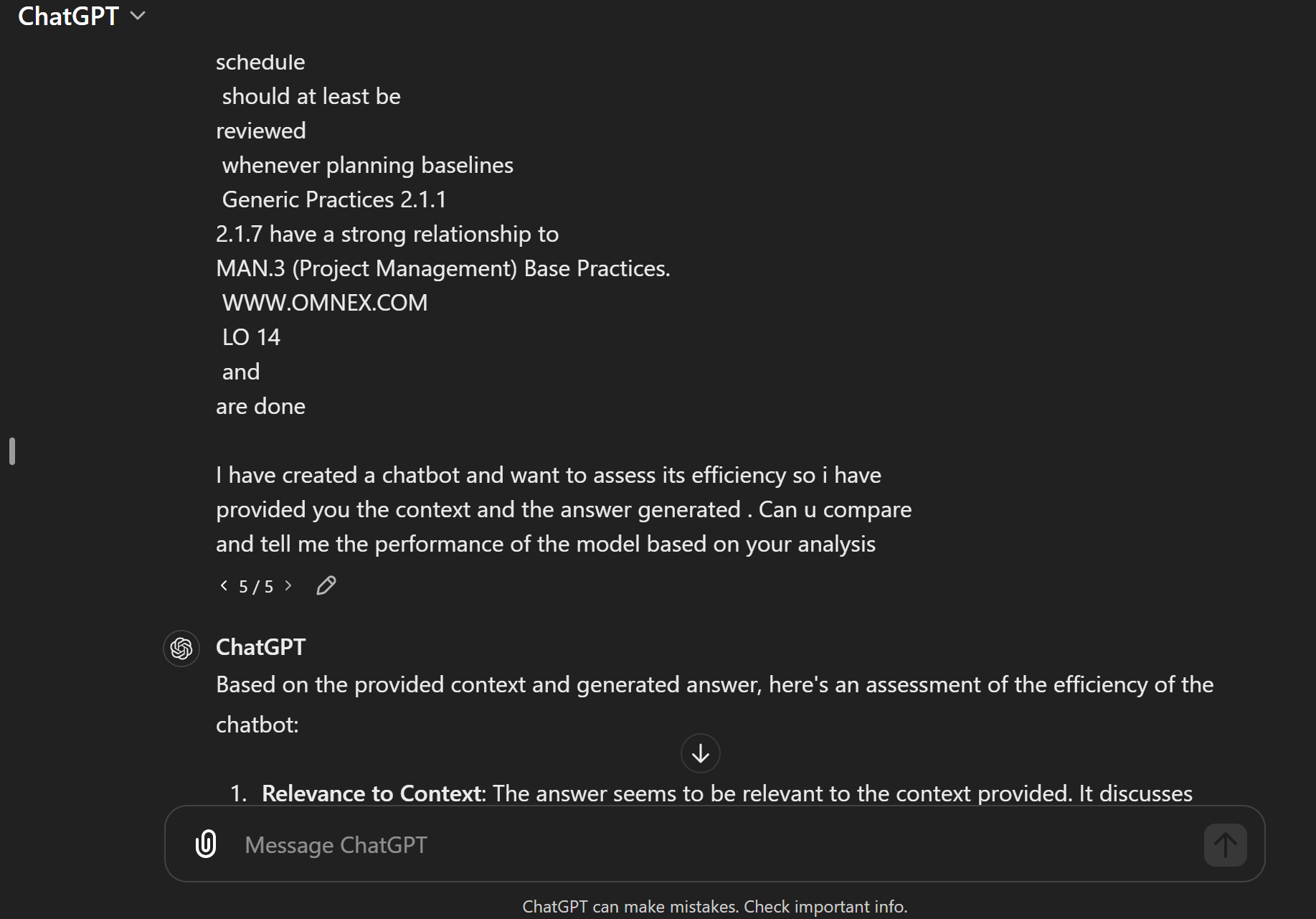
## 4.1 Answer Generation Analysis on Mistral:

## 

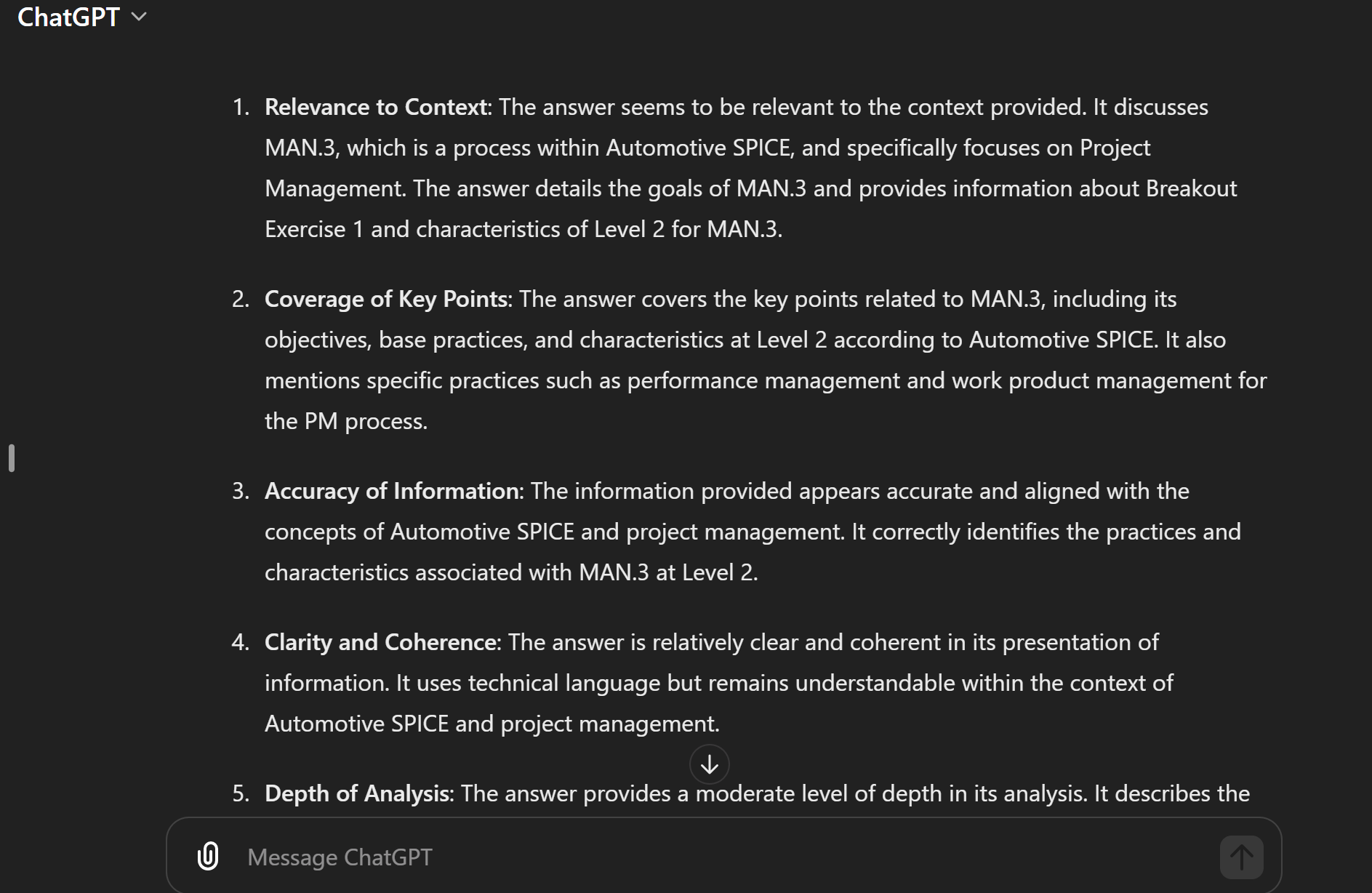
### 4.1.1 Question Analysis By Chatgpt:

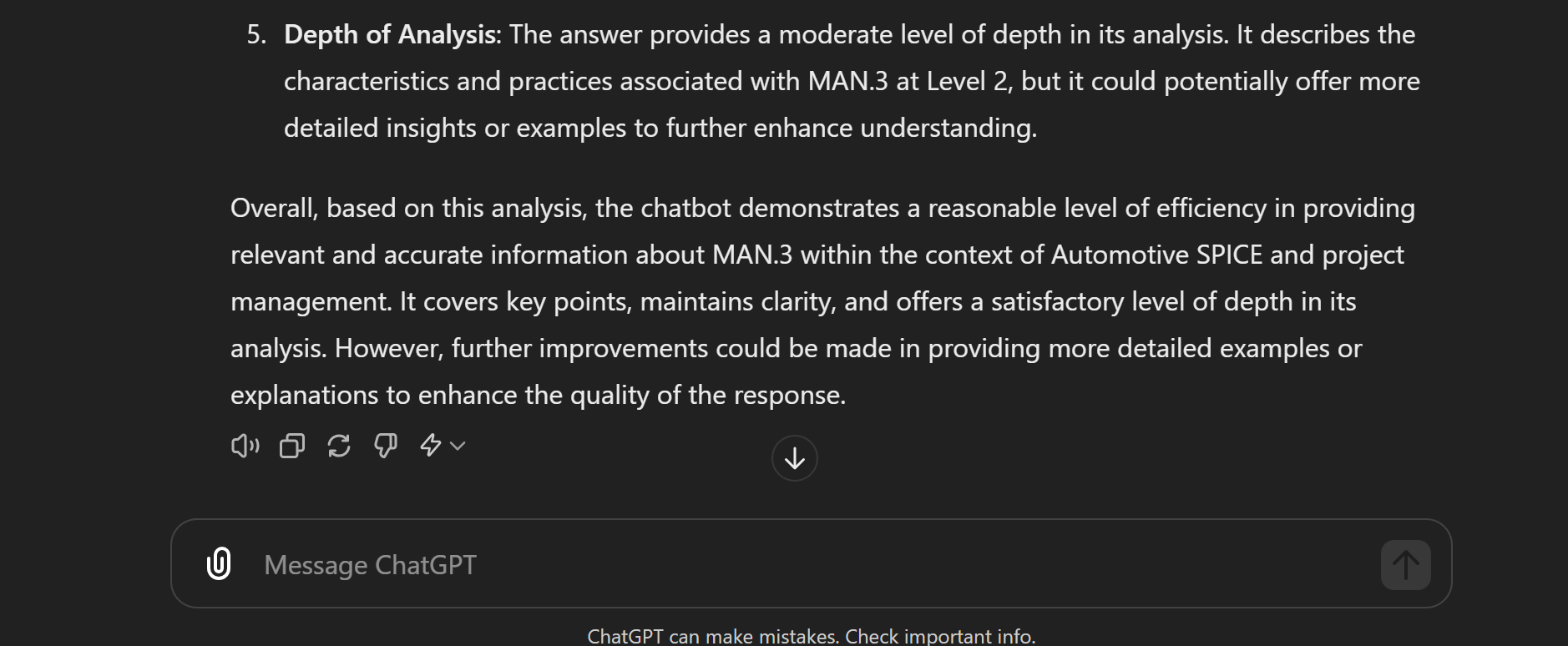
****





### 4.1.2 Answer generated by Chatgpt:





### 

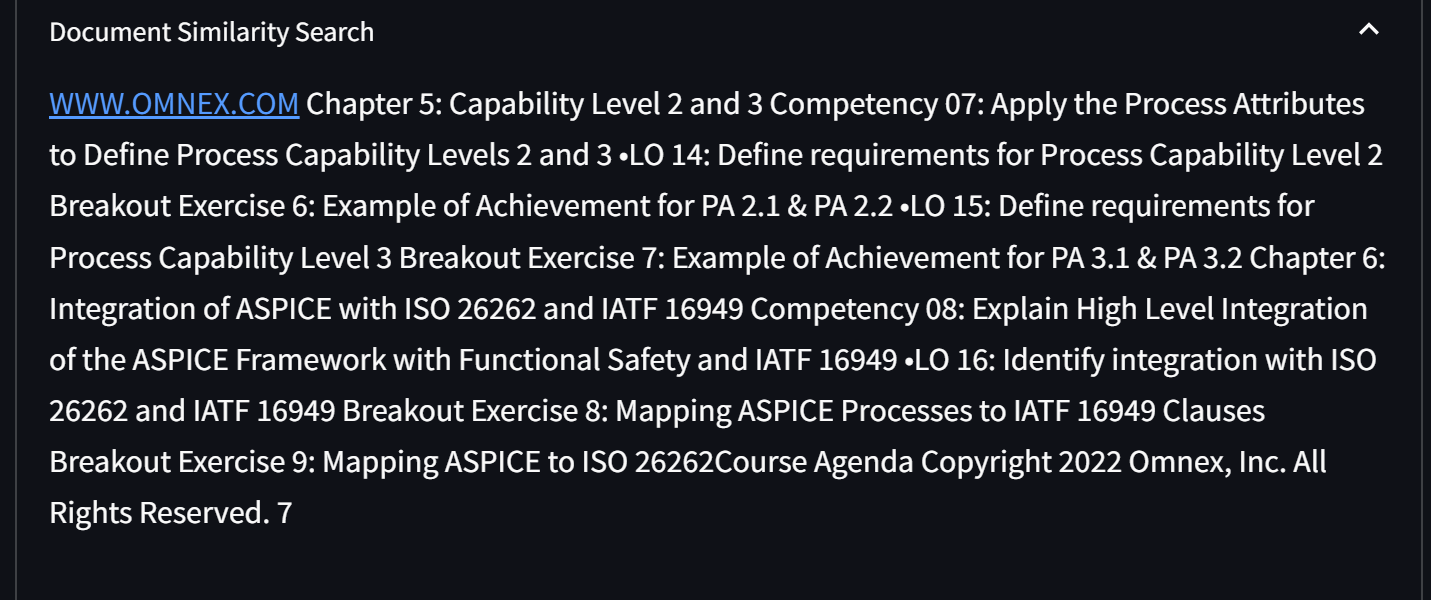
### 4.1.3 Conclusion on Answer Generation Using Mistral:

Chatbot demonstrates a reasonable level of efficiency in providing relevant accurate information within the context. Maintains clarity and offers a satisfactory level of depth in its analysis. However , further improvement could be made in providing more detailed examples or explanations to enhance the quality of response .

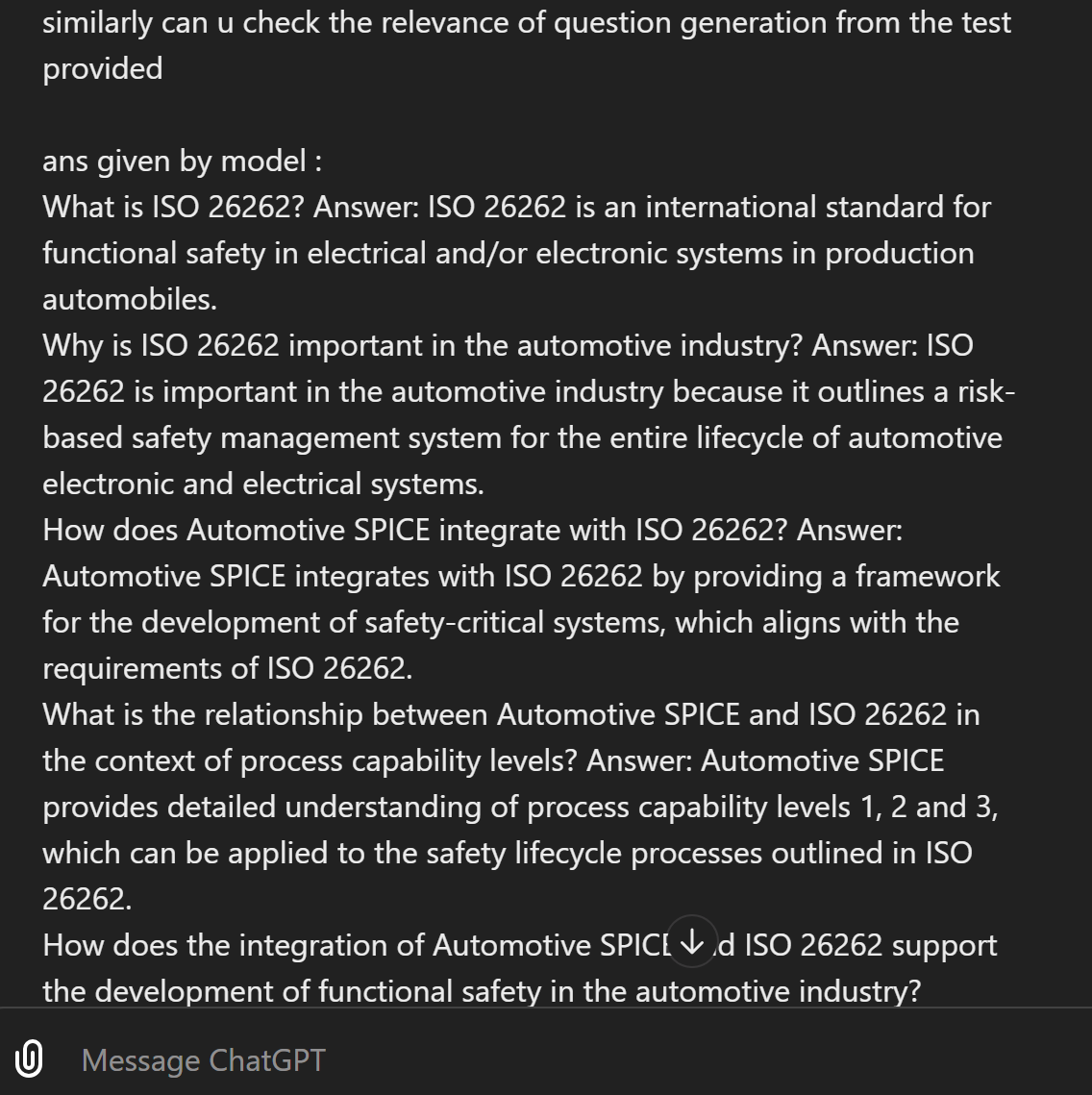
## 4.2 Question Generation Analysis of Mistral

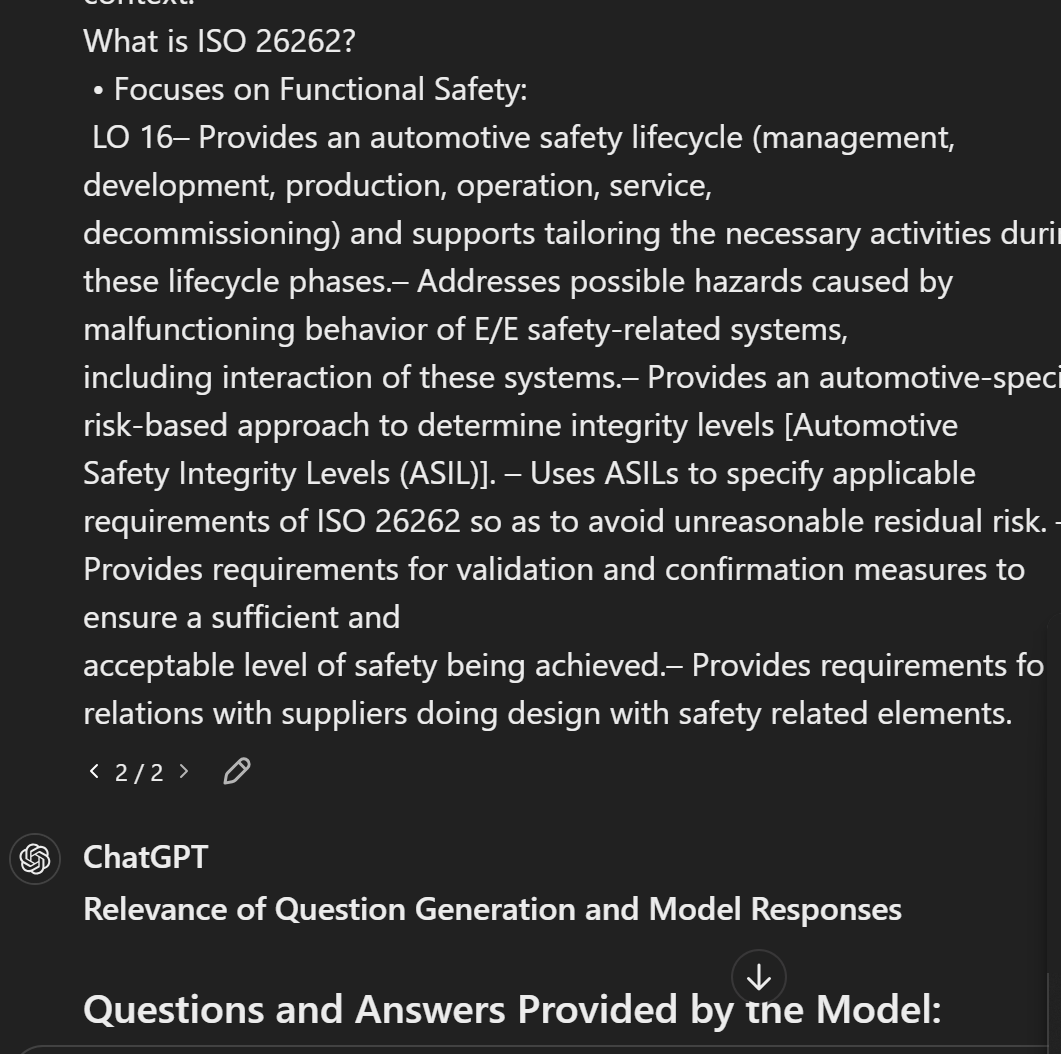


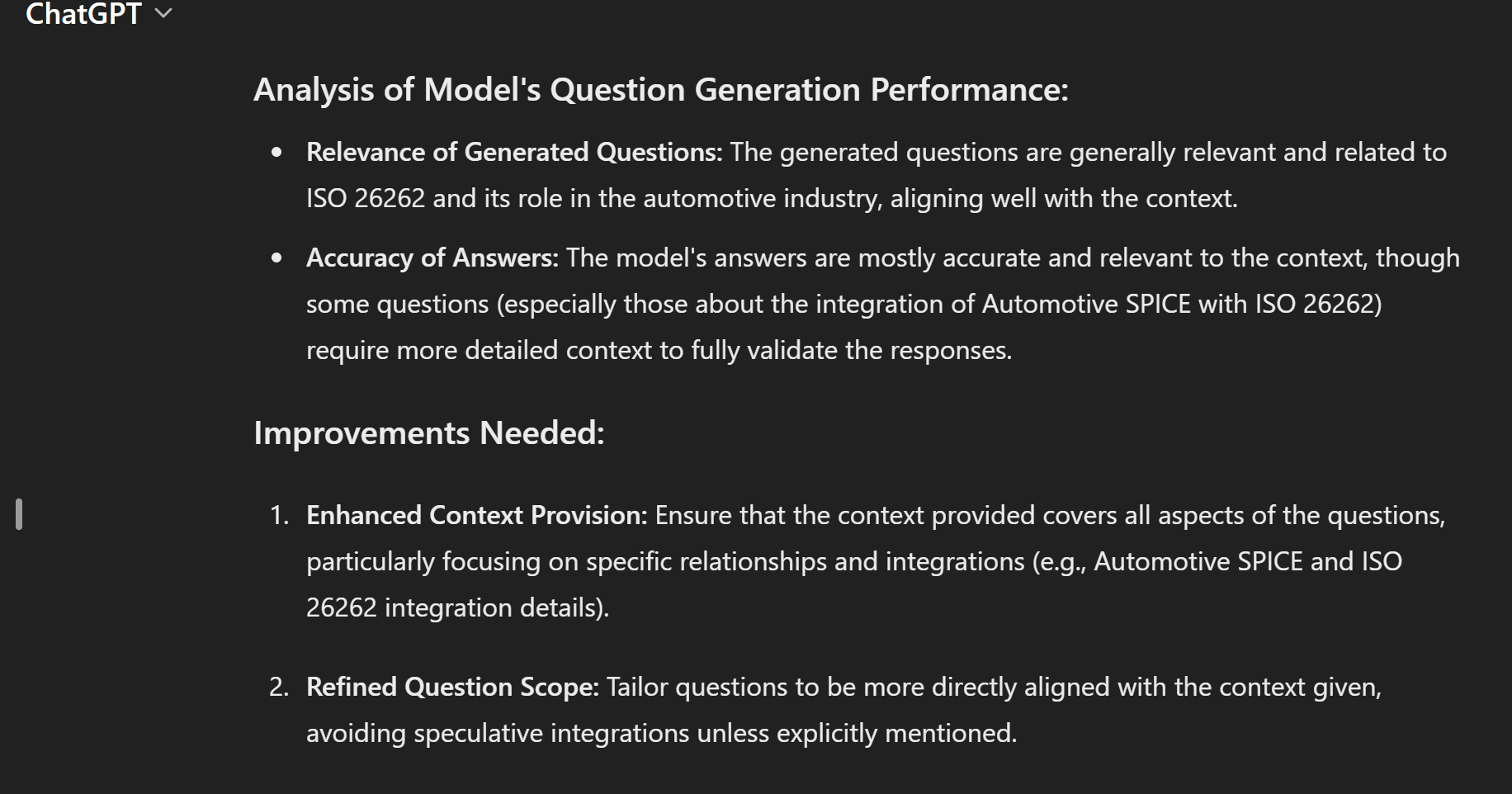
Document Similarity Search:



### 4.2.1 Inference Generated By Chatgpt:





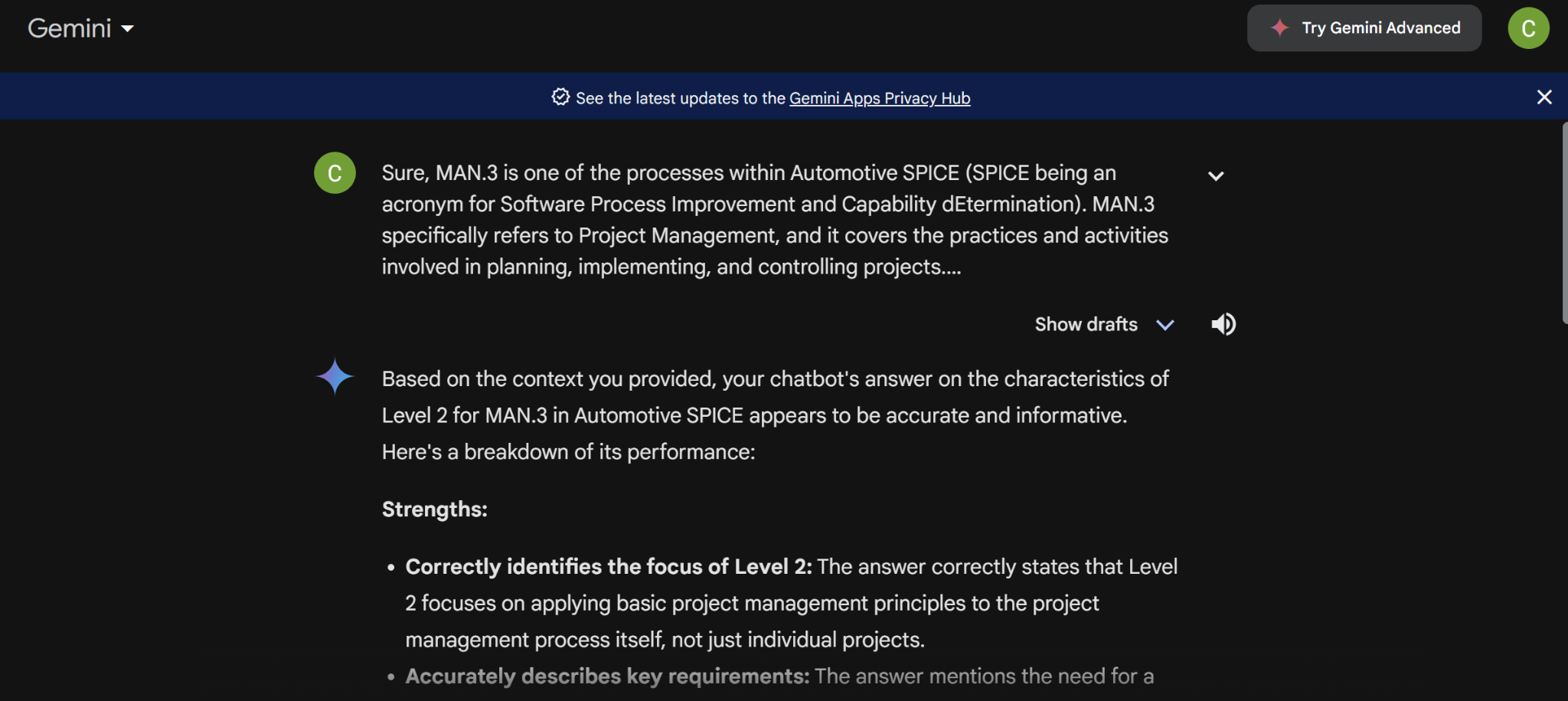


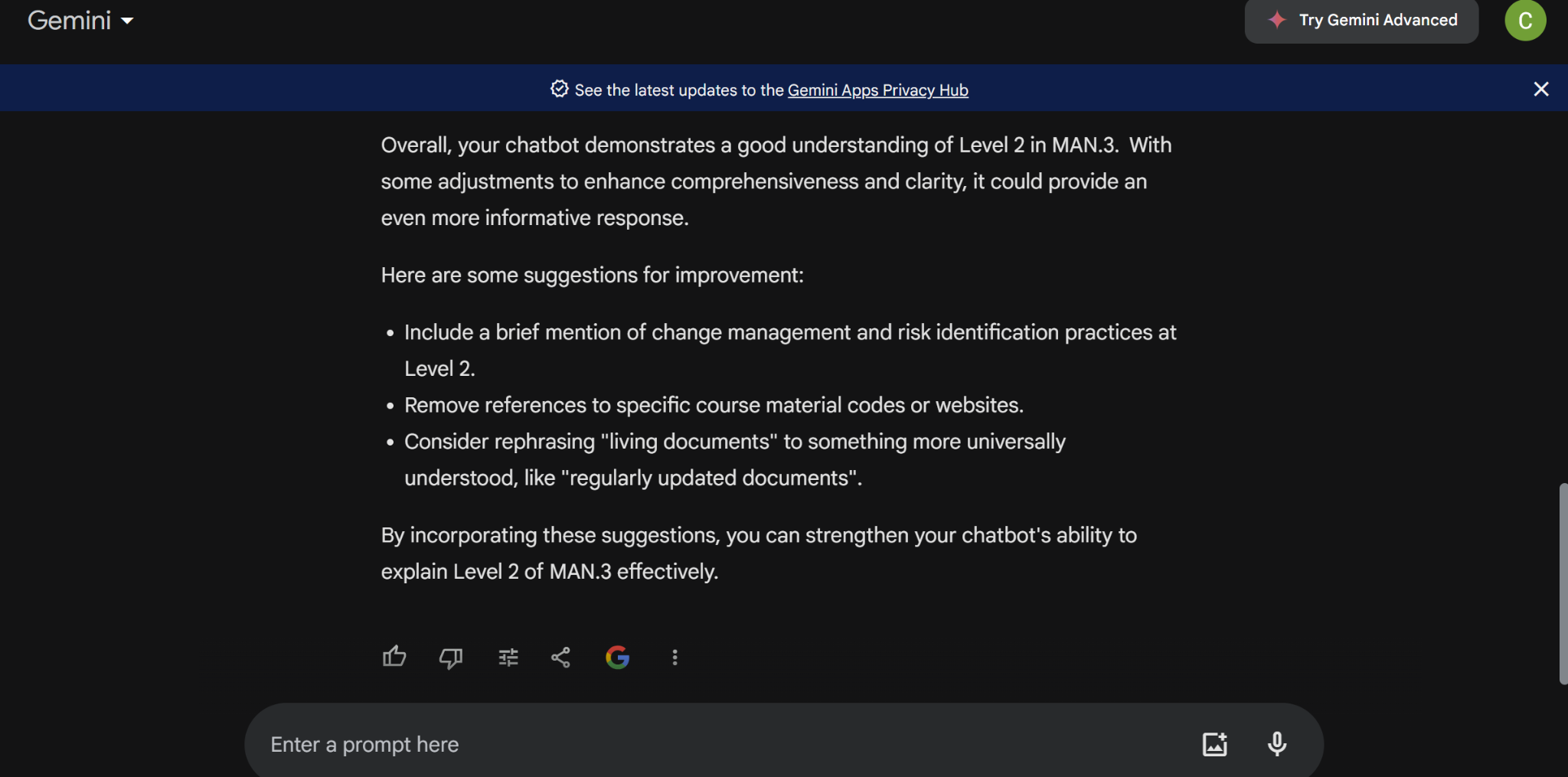
### 4.2.2 Conclusion on Mistral Question Generation :

The generated questions are relevant and related to context. The model answers are mostly accurate . Some questions require more detailed context to fully validate the response.

## 4.3 Response by Gemini on model Mixtral-8x7B-326B8

### 4.3.1 Answer generation(Gemini):

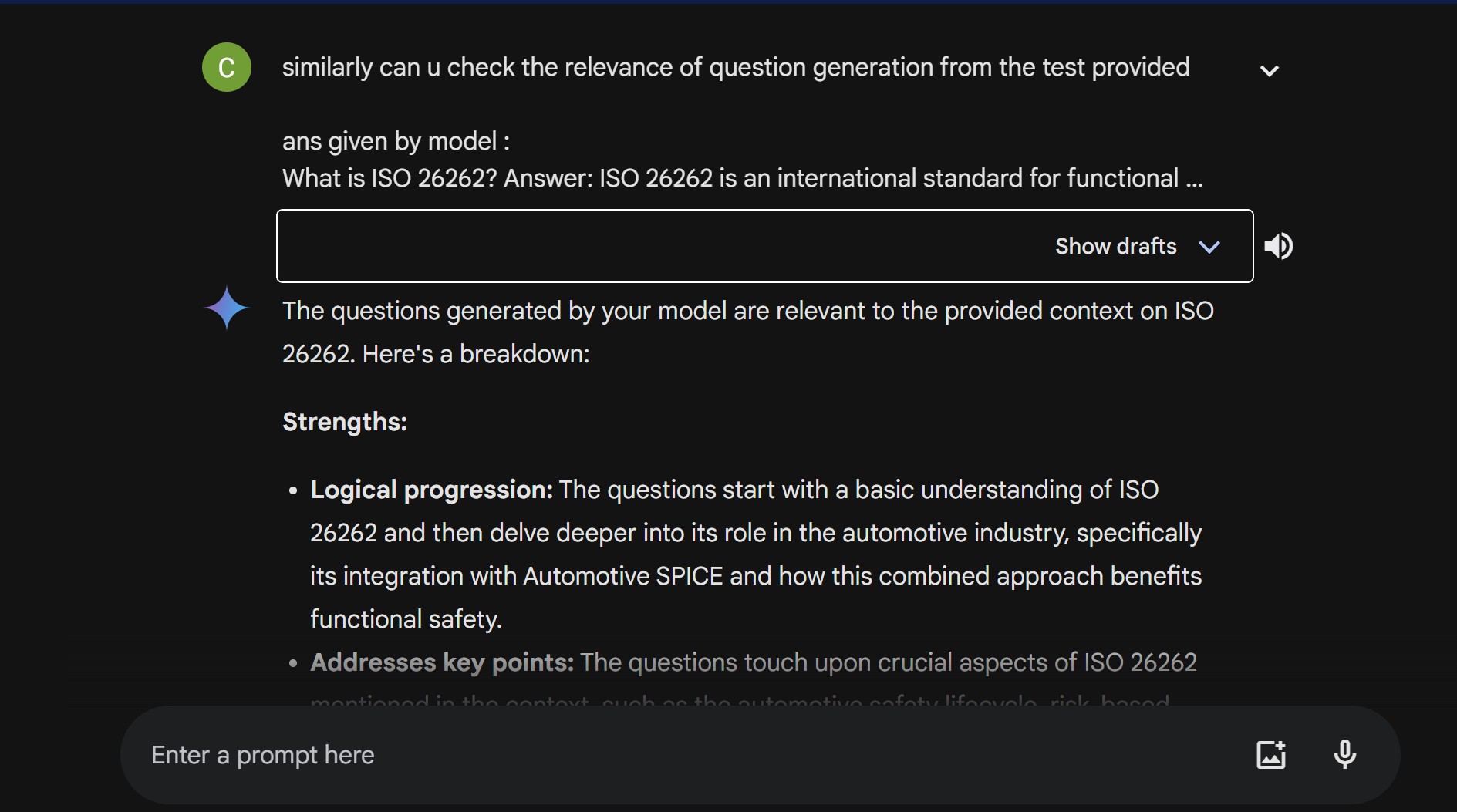


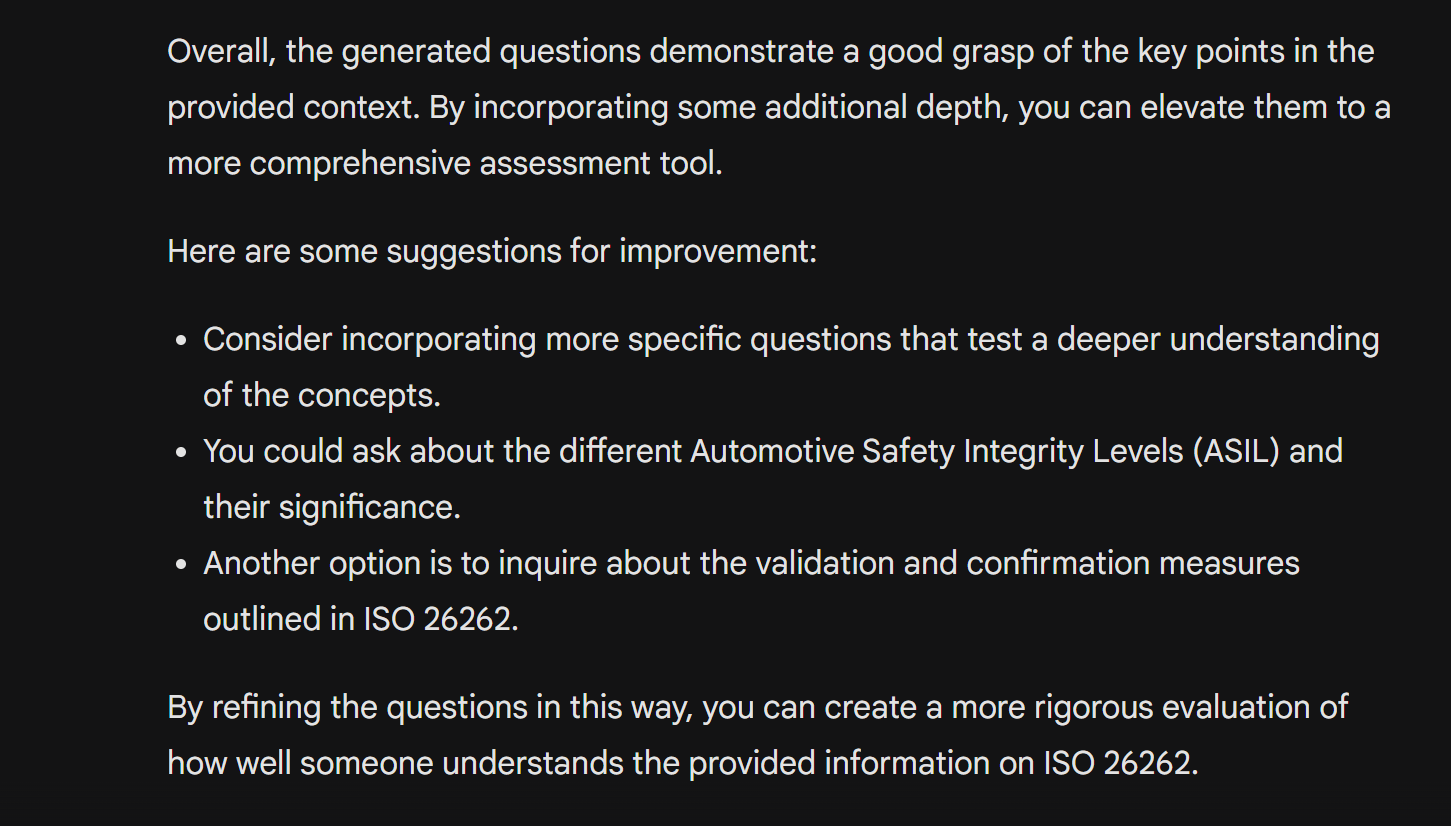


### 4.3.2 Conclusion on answer generation by Gemini model on Mistral 7b model:

According to Gemini the chatbot demonstrates a good understanding about the context with some adjustments to enhance comprehensiveness and clarity.

### 4.3.3 Question generation(Gemini):





### 4.3.4 Conclusion by gemini on question generation:

Overall , the generated questions demonstrate a good grasp of the key points in the context.

## 

## 

## 

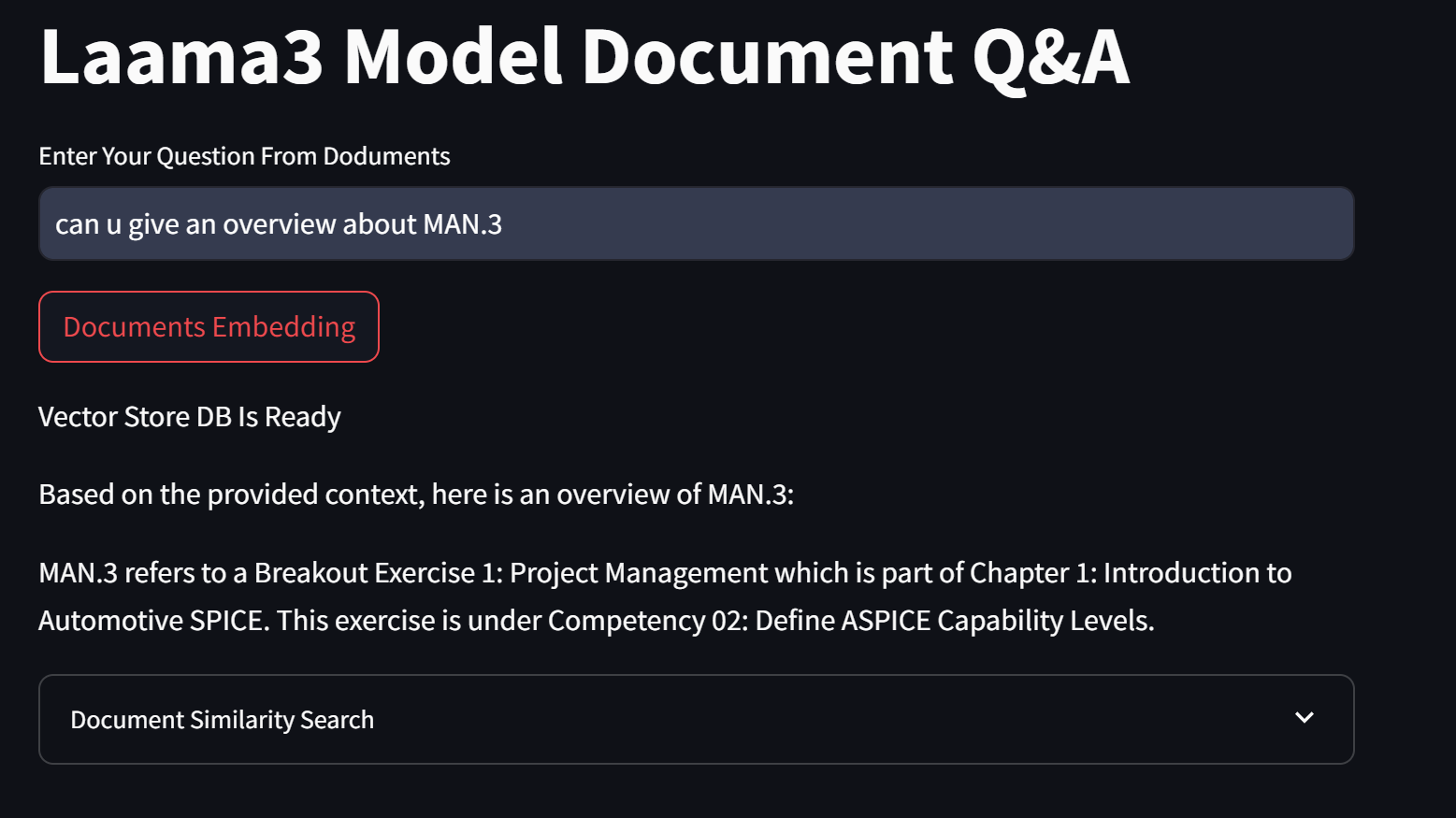
## 

## 

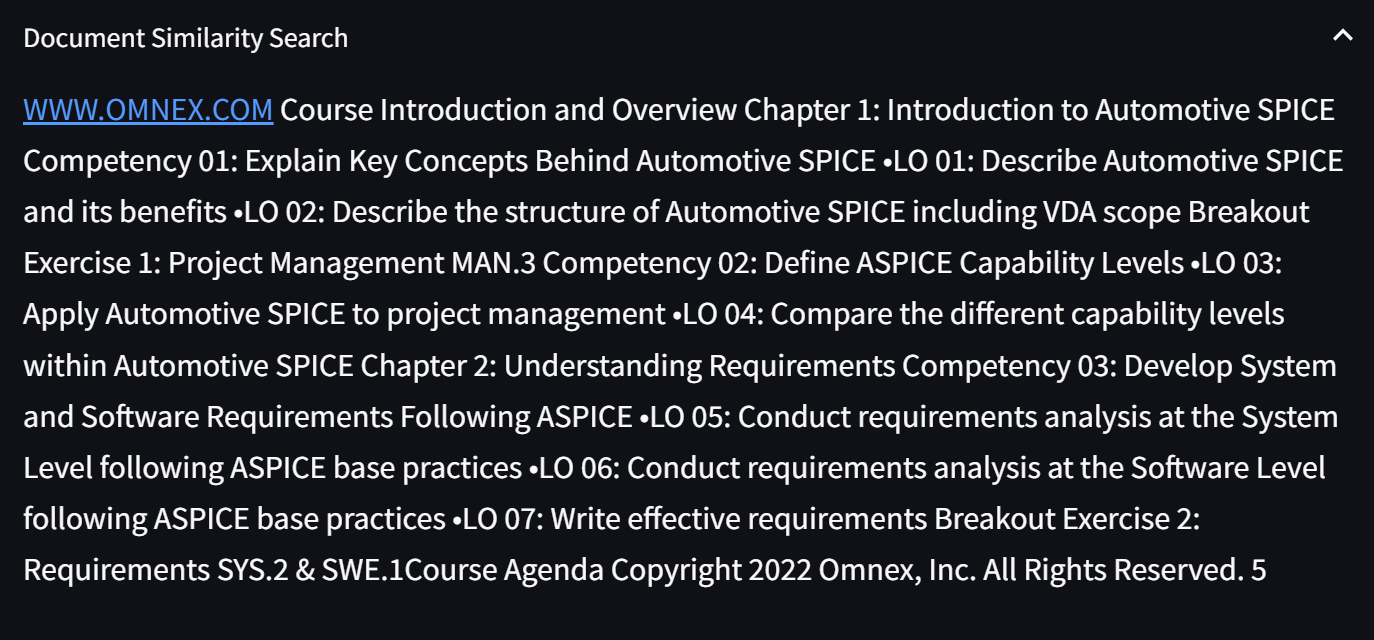
## 5.Model Analysis (Llama3-70B-8192)

### 5.1 Answer Generation

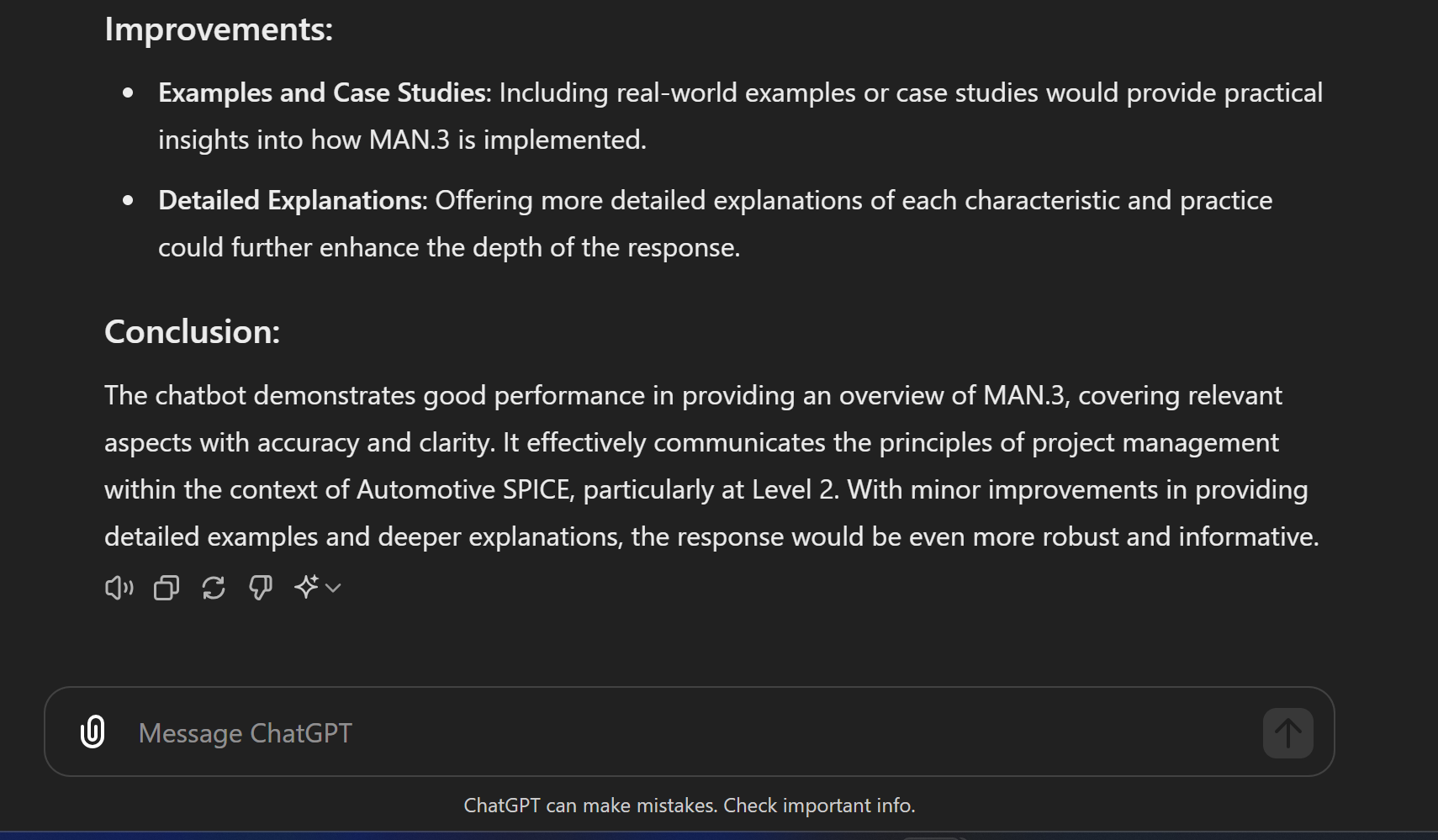
#### 5.1.1 Questions asked by Chatgpt on Llama3



Document Similarity Check:



#### 5.1.2 Answer Generated:



#### 5.1.3 Conclusion generated by Chatgpt:

Llama3 chatbot demonstrates good performance in providing relevant aspects with accuracy and clarity. It effectively communicates the context .With minor improvement in providing detailed explanation.

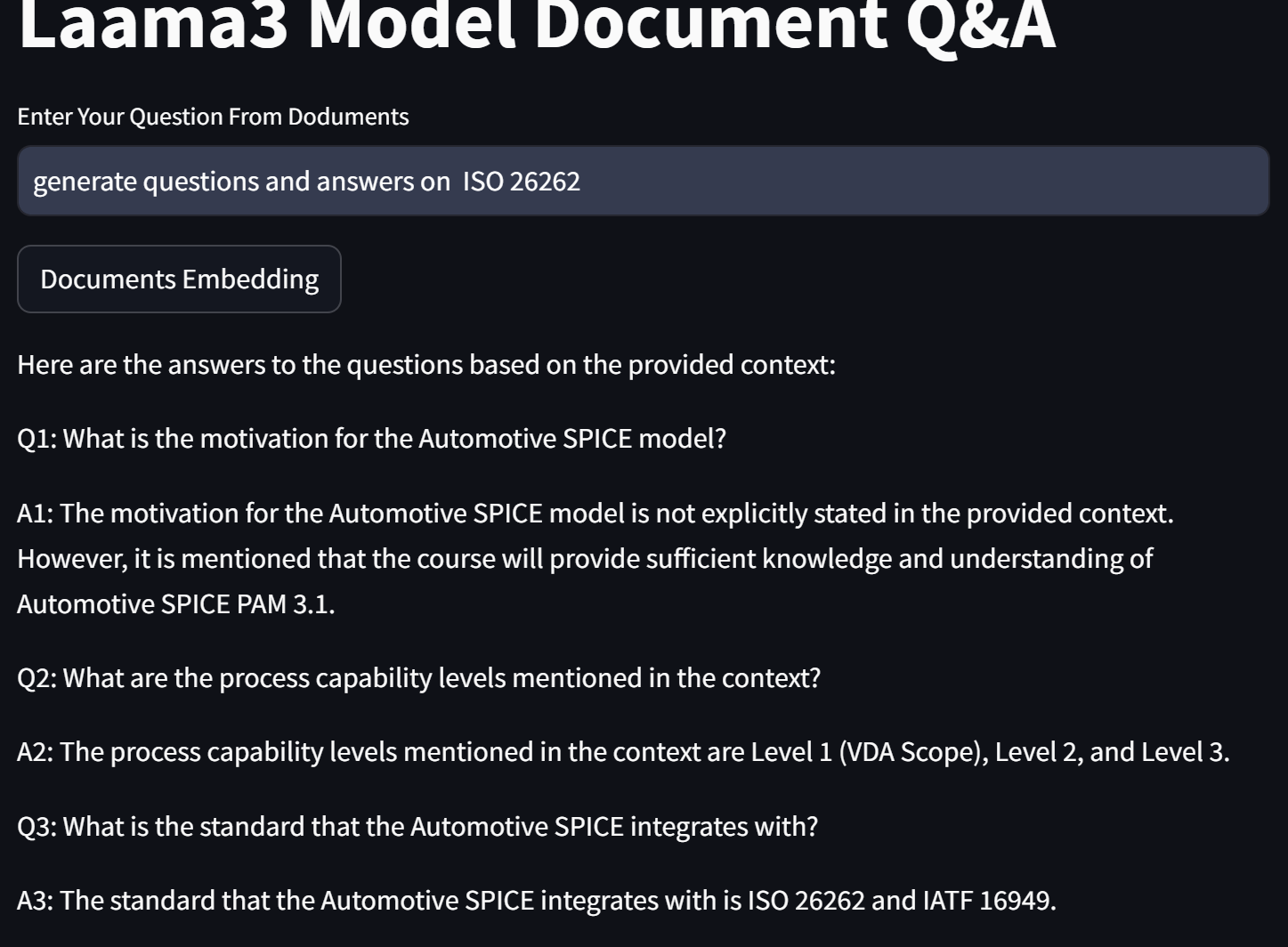
#### 5.1.4 Response By Gemini on Llama3



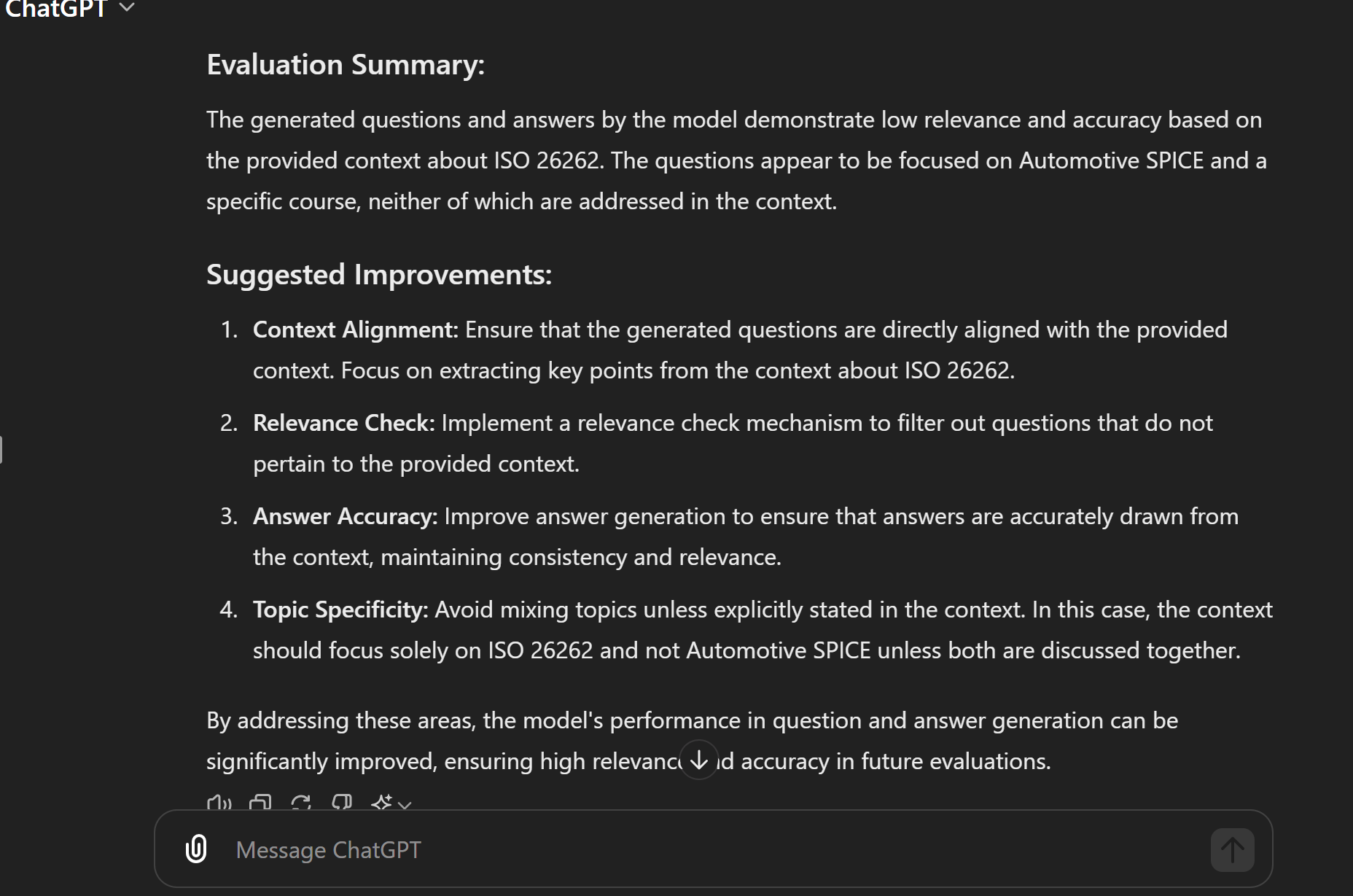
#### 5.1.5 Conclusion on Llama3 answer generation:

According to Gemini the model performance is limited

### 5.2 Question Generation on Llama 3 model:

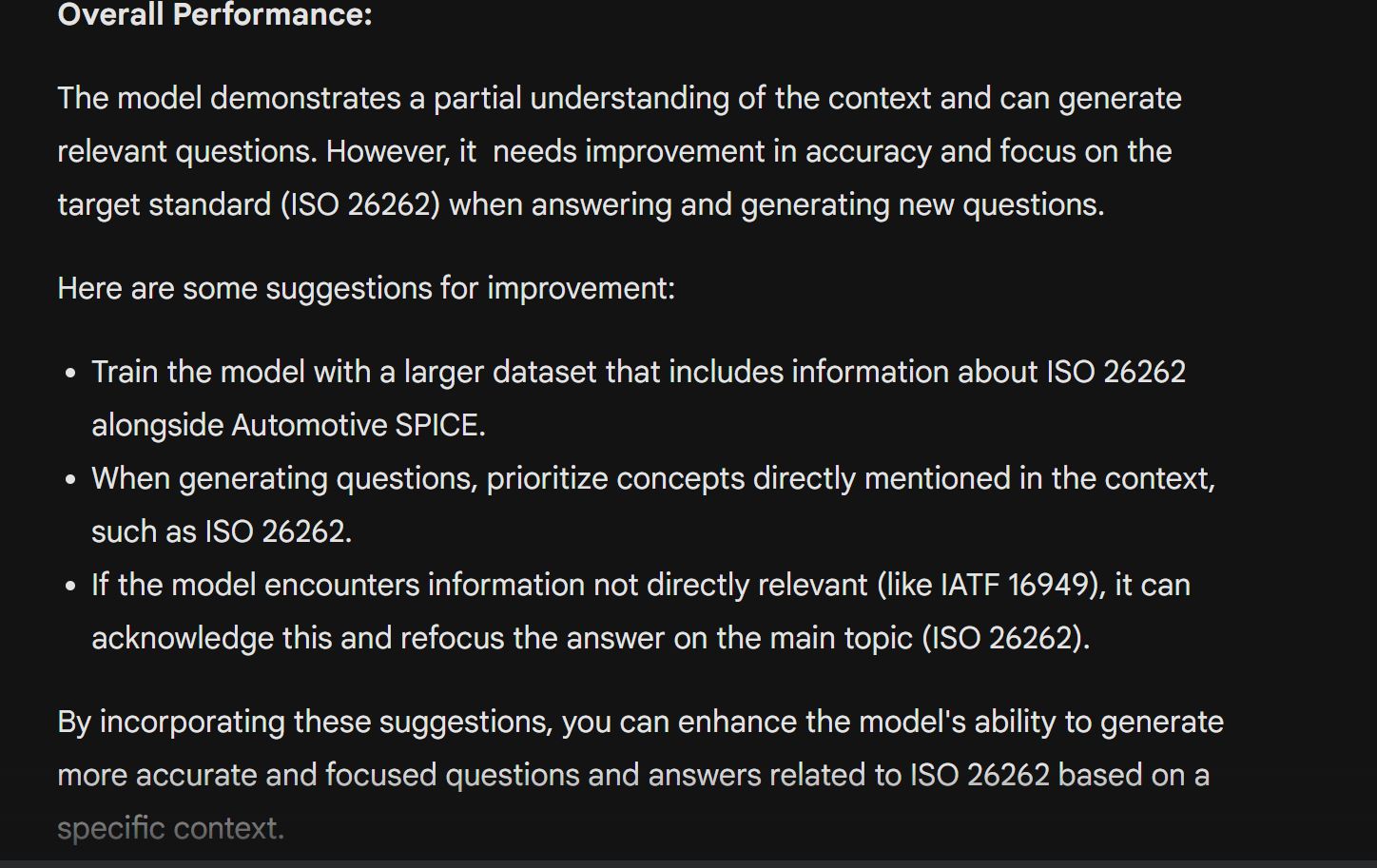


#### 5.2.1 Response generated by Chatgpt



#### 5.2.2 Conclusion on answer generation by Chatgpt The model performance in question and answer generation can be significantly improved , enhancing high relevance accuracy in future evaluations.

#### 5.2.3 Response generated by Gemini



#### 5.2.4 Conclusion by Gemini on Llama3 model

The model demonstrates a partial understanding of the context and can generate relevant questions. However it needs improvement in accuracy and focus on the target standard when answering and generating new questions.

6. GrocCloud Setup for Model Execution

Refer document for model setup.

[Groq Doc](https://docs.google.com/document/d/1ItyLXJ60MDevtFf_xUJAVDpE-2URrhvN3hxDbwcs4_k/edit?usp=sharing)

## 6. Future Work

Individually, my next steps involve tailored strategies to enhance the performance of Mixtral-8x7B-326B8 and LLama3-70B-8192.

For Mixtral-8x7B-326B8, I'll focus on fine-tuning hyperparameters such as learning rates, batch sizes, and optimization algorithms to optimize model performance.Exploring alternative model architectures or ensemble strategies, will try using different embedding models , optimizing resource utilization by reducing complexity. This comprehensive approach aims to elevate accuracy, efficiency, and scalability across diverse tasks.

Concurrently, I'll undertake specific measures to improve LLama3-70B-8192's performance. Will try improving question generation as the performance was poor and improve performance for very short one word answer generation by implementing robust performance monitoring and tuning mechanisms. By executing these strategies, I aim to enhance LLama3-70B-8192's efficiency and adaptability for various applications and domains.

# 

# 