Documentation: GenAI Model Testing Log

Overview

This document outlines the evaluation results of two Generative AI language models—OpenAI's `gpt-4o` and Groq's `deepseek-r1-distill-llama-70b`. The models were tested on four types of prompts to assess performance across factual accuracy, creativity, comprehension, and code translation. The tests were conducted using the `prompt_test.ipynb` notebook.

Models Evaluated

1. OpenAI `gpt-4o`

- Strengths: Consistently accurate, concise, and relevant responses.

- Latency: ~2–3 seconds across different prompt types.

- Notable Behavior: Clean and usable output with no additional markup or verbosity.

2. Groq `deepseek-r1-distill-llama-70b`

- Strengths: Competent and informative, with reasoning blocks.

- Latency: ~2–4 seconds depending on prompt complexity.

- Notable Behavior: Outputs are prefixed with `<think>` blocks, offering insight into reasoning but adding verbosity.

Prompt Categories & Evaluation Metrics

Each model was tested on the following categories:

- Factual Question: General knowledge or scientific fact.

- Creative Writing: A storytelling or imagination-based prompt.

- Long Input/Complex Data: Analytical task on large contextual inputs.

- Code Base Prompt: Cross-language code translation.

Each prompt was assessed using:

- Accuracy

Model: OpenAI `gpt-4o`

Prompt Type	Prompt	Expected Output	Output Summary	Accuracy	Latency	Bugs/Glitches	Rating & Notes
Factual Question	What is the speed of light in a vacuum?	Speed of light ≈ 299,792,458 m/s or ~300,000 km/s.	Correct value returned concisely.	High	~2 sec	None	9 - Clear and concise.
Creative Writing	A bustling city under the sky.	Short 150-word story about a vibrant city and the sky.	Story about artist Elara capturing sunset & city lights.	High	~2 sec	None	9 - Imaginative and well-written.
Long Input	Analyze tech history (computers, programming, internet).	Summary with main themes, key points, and a follow-up question.	Summarized tech evolution, noted lack of stats, asked AI/ML question.	High	~3 sec	None	9 - Comprehensive analysis.
Code Translation	Translate Python: print('Hello, World!') to JavaScript.	console.log('Hello, World!');	Correct translation from Python to JavaScript.	High	~3 sec	None	10 - Correct and direct.

Model: Groq `deepseek-r1-distill-llama-70b`

Prompt Type	Prompt	Expected Output	Output Summary	Accuracy	Latency	Bugs/Glitches	Rating & Notes
Factual Question	What is the speed of light in a vacuum?	Speed of light ≈ 299,792,458 m/s or ~300,000 km/s.	Correct value with `<think>` block explanation.	High	~2 sec	<think> meta block	8 - Verbose due to <think>
Creative Writing	A bustling city under the sky.	Short 150-word story about a vibrant city and the sky.	Vivid story of city square, prefixed with `<think>` block.	High	~2 sec	<think> meta block	8 - Good story but verbose.
Long Input	Analyze tech history (computers, programming, internet).	Summary with main themes, key points, and a follow-up question.	Detailed summary with trends and `<think>` intro.	High	~4 sec	<think> meta block	8 - Detailed but verbose.
Code Translation	Translate Python: print('Hello, World!') to JavaScript.	console.log('Hello, World!');	Correct code with reasoning in `<think>` block.	High	~3 sec	<think> meta block	8 - Accurate but verbose.

Final Insights

- GPT-4o is ideal for production-ready, clean outputs, especially in user-facing applications.
- Groq's LLaMA 70B excels in interpretability, useful for educational and research contexts where model reasoning transparency is valuable.
- Consider post-processing Groq output to remove `<think>` blocks if concise responses are desired.

Recommendations for Improvement

Model	Suggested Improvements
GPT-4o	Already concise and performant. Optional: include optional explanation if requested.
Groq	Provide toggles to disable `<think>` meta-output for cleaner integration in pipelines.