Flashlight LLM Test

David R. Miller

{dave}@millermattson.com

August 9, 2025

1 Introduction

OpenAI recently released GPT-5 with mixed reviews. Here is my first comparison test between Anthropic Sonnet 4 and OpenAI GPT-5. The models are comparably priced per token. All tests were through their respective APIs.

2 Test prompt

```
Imagine that every person on Earth was given a common typical
flashlight. Imagine that on a Tuesday at 9 a.m., every person aimed
their flashlight east, parallel to the ground, turned on their
flashlight, and left it on for 24 hours. How much would that change
the rotational period of the Earth after those 24 hours? Show your
calculations.
```

3 The models

I gave the same prompt to:

- Claude Sonnet 4
- Claude Sonnet 4 “thinking mode”
- GPT-5
- GPT-5 “high reasoning mode”

4 Results

All four models explicitly assumed that a typical flashlight emits 1W of energy in a tight beam and tacitly assumed that batteries don't deplete. (Tie)

Default Claude Sonnet 4 and default GPT-5 assumed everybody lives on the equator. (Tie)

In "thinking mode" and "high reasoning mode," both models assumed a uniform distribution of people across all latitudes. (Tie)

Both models in default and thinking/reasoning mode made similar assumptions and calculations for the photon momentum and planetary torque calculations. (Tie)

Claude Sonnet 4, in default and thinking mode, tacitly assumed no atmosphere so that all the photon energy escapes into space causing a reaction force. GPT-5, in both default and high reasoning mode, correctly pointed out that only some of the photon energy will escape the atmosphere. In default mode, GPT-5 quantified the estimated effects of atmosphere. (GPT-5 is the winner here)

5 Bottom-line answers

Model	Result
Sonnet 4 default	-1.8×10^{-16}s
Sonnet 4 "thinking mode"	-1.4×10^{-16}s
GPT-5 default	-1.4×10^{-22}s (with atmosphere) -2.2×10^{-16}s (no atmosphere)
GPT-5 "high reasoning mode"	"orders of magnitude below 10^{-16}s" (with atmosphere) -2×10^{-16}s (no atmosphere)

6 Final thoughts

Only GPT-5 mentioned the effect of atmospheric absorption, which is a huge factor in calculating torque in this scenario. Otherwise their calculations and answers are comparable.

Sonnet 4 has been my go-to LLM for a while, but GPT-5 is the winner in this test. I'm going to give more attention to GPT-5.