LLM Eye Blink Test

David R. Miller

{dave}@millermattson.com

August 26, 2025

1 Introduction

This is a comparison of OpenAI GPT-5 vs. Anthropic Claude Sonnet 4. These are comparably priced commercial LLMs and are available in reasoning and non-reasoning varieties.

2 The prompt

```
On Tuesday morning at 7:00 a.m., Jane Wilkenson of Akron, Ohio woke up
and blinked. Her eye blink generated a gravitational wave. Calculate
the strain h of that gravitational wave as it passes through the
Andromeda galaxy two million years later. State your assumptions and
show the calculations.
```

3 The contestants

- GPT-5 minimal reasoning
- Claude Sonnet 4 non-thinking
- GPT-5 high reasoning
- Claude Sonnet 4, 10,000-token thinking budget

4 The winner

GPT-5 in reasoning mode wins by applying more precise modeling of the eye blink dynamics, more precise math calculations, and a slightly more complete modeling of the gravitational wave radiation pattern.

5 The details

All four models realized that this non-relativistic problem involves finding the second time derivative of the quadrupole moment of the eye blink. The models started with similar formulas for the dimensionless strain h, expressed in different forms:

Model	**Basic strain formula**
GPT-5 minimal reasoning	$h \sim (G/c^4)(mA^2\omega^2)/D$
Claude Sonnet 4 non-thinking	$h \approx (2GmaL)/(c^4 r)$
GPT-5 high reasoning	$h \sim (G/c^4) \times (1/r) \times (d^2 I/dt^2)$
Claude Sonnet 4 thinking	$h = (2G/c^4) \times (1/r) \times \lvert d^2 I/dt^2 \rvert$

All four models assumed that the quadrupole moment scales with the mass and the square of the length (amplitude or orbital radius) of the eyelid movement. All assumed that the second time derivative scales with the inverse of the timescale squared.

All four models pointed out that the distance to Andromeda is closer to 2.5 Mly, not 2 Mly as stated in the prompt. Three of the models used the more precise 2.5 Mly, while GPT-5 in minimal reasoning mode went with 2.0 Mly.

The GPT-5 models assumed a combined eyelid mass of 4 g to 10 g for two eyelids moving in phase. The Claude models assumed a mass of 1.5 g to 5 g and assumed only a single eyelid. Points for GPT-5for assuming that Jane has two eyes.

All models assumed an eyelid mass displacement of about 1 cm.

Claude in thinking mode assumed a blink duration of 200 ms while the other models went with 100 ms. Both estimates are reasonable.

GPT-5 in reasoning mode was the only model to correctly point out that the frequency of a gravitational wave is twice the frequency of the object's mechanical motion.

One of the biggest differences is how they evaluated the second derivative. Both GPT-5 models used a more precise sinusoidal formula while both Claude models applied a duration-acceleration approximation that dropped a factor of $(2\pi)^2$.

GPT-5 in reasoning mode was the only model to mention and apply prefactors assuming an optimal orientation of the radiation pattern for maximum strain.

All models appropriately mentioned the absurdity of this problem.

GPT-5's results are larger in value due mainly to its sinusoidal derivative calculation, doubled eyelid mass, and more precise prefactors. Here are their final order-of-magnitude answers for the dimensionless strain h:

Model	**Final result**
GPT-5 minimal reasoning	10^{-69}
Claude Sonnet 4 non-thinking	10^{-72}
GPT-5 high reasoning	10^{-69}
Claude Sonnet 4 thinking	10^{-72}

For its more thorough treatment, GPT-5 in reasoning and non-reasoning modes wins over Claude Sonnet 4