Comparing Google's Gemma 4 and OpenAI's GPT-5.3 Chat

This article will compare Google's Gemma 4 26B A4B and OpenAI's GPT-5.3 Chat based on key metrics including cost, context length, and other model features.

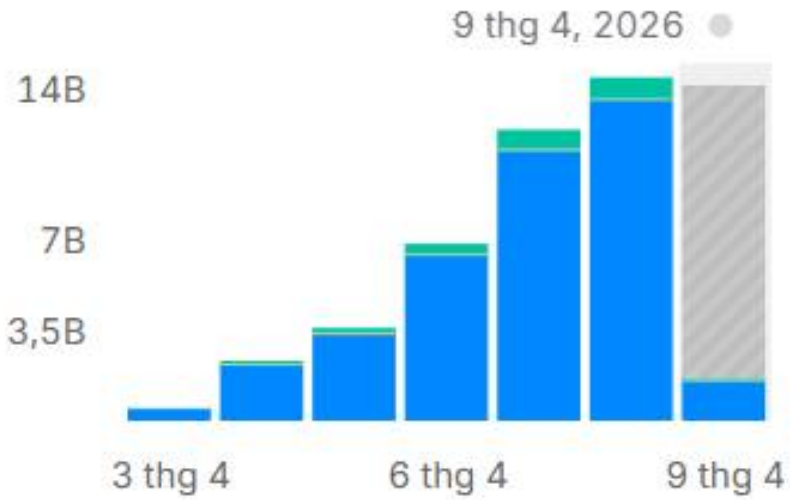
This article will compare Google's Gemma 4 26B A4B and OpenAI's GPT-5.3 Chat based on key metrics including cost, context length, and other model features.

Basic information about Gemma 4 26B A4B and GPT-5.3 Chat

Gemma 4 26B A4B	GPT-5.3 Chat
1. Author: Google 2. Context length: 262K 3. Reasoning ability: Yes	1. Author: OpenAI 2. Context length: 128K 3. Reasoning ability: None
Gemma 4 26B A4B IT is a Mixture-of-Experts (MoE) model tuned according to instructions from Google DeepMind. Despite having a total of 25.2 billion parameters, only 3.8 billion parameters are activated per token during the inference process – delivering the quality of nearly 31 billion parameters at a much lower computational cost. Supports multimodal input including text, images, and video (up to 60 seconds at 1 frame per second). It has a 256K token context window, native function calls, configurable thinking/reasoning modes, and support for structured output. Released on the Apache 2.0 operating system.	GPT-5.3 Chat is an update to the most widely used ChatGPT model, making everyday conversations smoother, more helpful, and more direct. It provides more accurate answers with better context and significantly reduces unnecessary denials, notes, and overly cautious language that can disrupt the flow of conversation.

Work

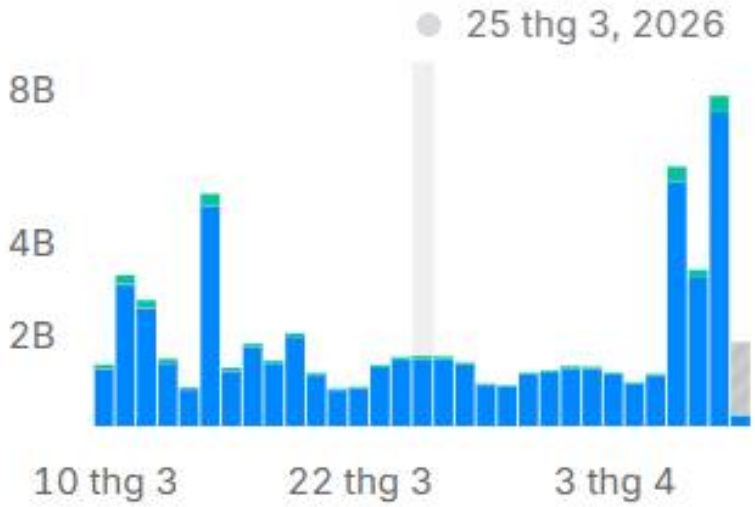
Gemma 4 26B A4B



1. Prompt: 6.44B
2. Completed: 405M
3. Inference: 54.3M
4. Latency (p50): 0.75s
5. Throughput (p50): 46.0 tok/s
6. Visual performance: Completed 2050 tokens in 45.3 seconds

```
Attention is all you need. The dominant sequence transduction models are based o
```

GPT-5.3 Chat



1. Prompt: 206M

2. Completed: 6.04M
3. Inference: 258M
4. Latency (p50): 1.40 seconds
5. Throughput (p50): 55.0 tokens/second
6. Visual performance: Completed 2050 tokens in 38.7 seconds

```
Attention is all you need. The dominant sequence transduction models are based o
```

Price

Gemma 4 26B A4B	GPT-5.3 Chat
1. Input: $0.13/million tokens 2. Output: $0.40/million tokens	1. Input: $1.75/million tokens 2. Output: $14/million tokens

Features

Features	Gemma 4 26B A4B	GPT-5.3 Chat
Input method	Images, text, videos	Text, images, files
Output method	Document	Document
Quantum	bf16	---
Maximum number of tokens (input + output)	262K	128K
Maximum number of tokens to output	262K	16K
Cancel stream	Have	Have
Tool support	Are not	Have
No prompt training reminder needed.	Have	Have
Save cache	Are not	Have