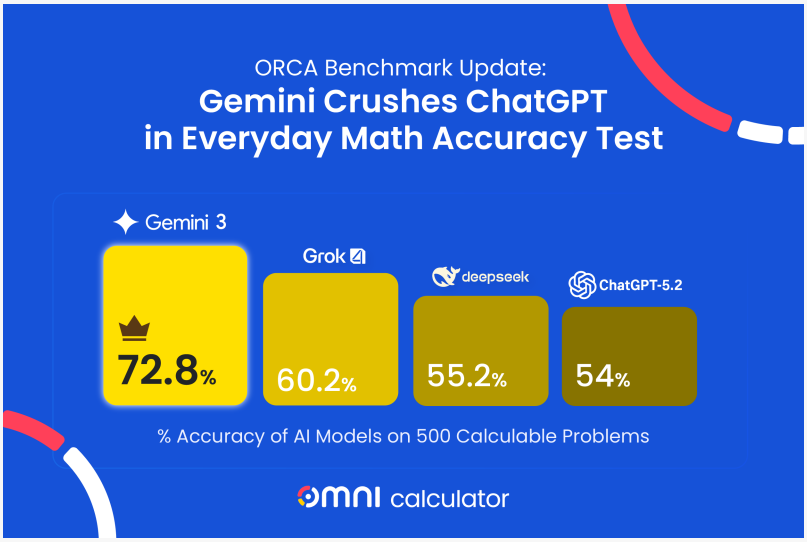
Gemini 3 Flash Crushes ChatGPT-5.2 in Accuracy Test - ORCA Benchmark Update

New ORCA results show Gemini leading in practical math, but no AI matches the consistency of a simple calculator.



Gemini crashes chat GPT in everyday math

KRAKOW, POLAND, March 3, 2026 /EINPresswire.com/ -- The results are in for the second ORCA (Omni Research on Calculation in AI) Benchmark, and the leaderboard looks very different than it did two months ago. Gemini 3 Flash has surged to the top, becoming the first model to solve nearly three-quarters of real-world math and logic problems correctly.

While the industry often focuses on academic tests, the ORCA Benchmark uses 500 practical questions and the kind of "messy" math people actually deal with every day. In this latest run, Gemini 3 Flash hit an accuracy rate of 72.8%, a significant jump from its previous performance. Meanwhile, ChatGPT-5.2 and DeepSeek V3.2 showed modest, steady gains, while Grok 4.1 saw its scores slip.

> Calculators are predictable, always giving the same answer. AI is different; Mathematically, a model can get a question right today and wrong tomorrow."
>
> *Dawid Siuda, Researcher at ORCA*

The "Calculator" Problem

Despite the high scores, the study highlights a lingering frustration for AI users: inconsistency. Unlike a standard calculator, which gives the same answer every time, these AI models are probabilistic. They "predict" answers rather than calculating them through fixed rules.

The ORCA team tracked this using a new "instability metric." It turns out that even when models are wrong, they aren't even consistently wrong.

- ChatGPT-5.2 changed its answer on 65% of persistent errors.
- Gemini 3 Flash proved more "stubborn," changing only 46% of the time.
- DeepSeek was the most erratic, shifting its response 69% of the time.

"A calculator is predictable. Ask it the same question today or next year, and the answer stays the same," says Dawid Siuda, researcher at ORCA. "AI doesn't work that way. These systems are predicting the next likely word based on patterns. Mathematically, it's possible for a model to get a question right today and wrong tomorrow."

Omni Calculator Logo

Winners and Losers by Category
The progress wasn't even across the board. Gemini saw massive gains in biology, chemistry, and physics, but actually dropped slightly in engineering. DeepSeek V3.2, now out of its "alpha" phase, saw its biology scores skyrocket from 11% to 44%. On the flip side, Grok 4.1 struggled, losing ground in health, sport, and statistics.

What This Means for Users
The data shows that while AIs are getting better at rounding numbers and formatting results, they still trip up on core arithmetic. Calculation errors now account for 39.8% of all mistakes made across the models tested.

The takeaway? AI is a powerful assistant, but it’s not a replacement for a human eye or a calculator.

About ORCA Benchmark
The ORCA Benchmark (Omni Research on Calculation in AI) is an initiative by Omni Calculator designed to provide a genuine assessment of the mathematical capabilities of today’s leading AI chatbots. For our second iteration, we tested four prominent models: ChatGPT-5.2 (OpenAI), Gemini 3 Flash (Google), Grok-4.1 (xAI), and DeepSeek V3.2.
Our methodology prioritizes accessibility; we only test new models that offer a free tier to the public. This is why Anthropic’s Claude 4.5 Sonnet was not retested this round (as it has not been updated since our first report), and why DeepSeek V3.2 was included again—while the name remains the same, the model has transitioned from an alpha version to a stable release, resulting in a performance jump of over 3 percentage points. By avoiding paid-tier or private prototypes, ORCA provides a genuine assessment of the mathematical tools available to the average user right now.

For the full report, visit https://www.omnicalculator.com/reports/orca-ai-benchmark-2026-update

Reyhaneh Mansouri

Omni Calculator sp. z o.o.
+48 730 061 124
email us here
Visit us on social media:
LinkedIn
