Gemini 2.5 Pro is here with bigger numbers and great vibes



Just a few months after releasing its first Gemini 2.0 AI models, Google is upgrading again. The company says the new Gemini 2.5 Pro Experimental is its “most intelligent” model yet, offering a massive context window, multimodality, and reasoning capabilities. Google points to a raft of benchmarks that show the new Gemini clobbering other large language models (LLMs), and our testing seems to back that up—Gemini 2.5 Pro is one of the most impressive generative AI models we’ve seen.

Gemini 2.5, like all Google’s models going forward, has reasoning built in. The AI essentially fact-checks itself along the way to generating an output. We like to call this “simulated reasoning,” as there’s no evidence that this process is akin to human reasoning. However, it can go a long way to improving LLM outputs. Google specifically cites the

model’s “agentic” coding capabilities as a beneficiary of this process. Gemini 2.5 Pro Experimental can, for example, generate a full working video game from a single prompt. We’ve tested this, and it works with the publicly available version of the model.

[embedded content]

Gemini 2.5 Pro builds a game in one step.

Google says a lot of things about Gemini 2.5 Pro; it’s smarter, it’s context-aware, it *thinks*—but it’s hard to quantify what constitutes improvement in generative AI bots. There are some clear technical upsides, though. Gemini 2.5 Pro comes with a 1 million token context window, which is common for the big Gemini models but massive compared to competing models like OpenAI GPT or Anthropic Claude. You could feed multiple very long books to Gemini 2.5 Pro in a single prompt, and the output maxes out at 64,000 tokens. That’s the same as Flash 2.0, but it’s still objectively a lot of tokens compared to other LLMs.

Naturally, Google has run Gemini 2.5 Experimental through a battery of benchmarks, in which it scores a bit higher than other AI systems. For example, it squeaks past OpenAI’s o3-mini in GPQA and AIME 2025, which measure how well the AI answers complex questions about science and math, respectively. It also set a new record in the Humanity’s Last Exam benchmark, which consists of 3,000 questions curated by domain experts. Google’s new AI managed a score of 18.8 percent to OpenAI’s 14 percent.

https://arstechnica.com/ai/2025/03/google-says-the-new-gemini-2-5-pro-model-is-its-smartest-ai-yet/