Gemini 4 Neon explained: pricing, thinking levels and what Google reports about its benchmarks

25 November 2026 · 8 min read · blogs.iictp.org/gemini-4-neon-frontier-model-pricing-thinking-levels-explained

Short answer: Gemini 4 Neon is a Google DeepMind model announced on 18 November 2026 that Google positions as frontier-level reasoning with faster, cheaper task completion for software engineering, knowledge work and science. Google lists it at $3 per million input tokens and $15 per million output tokens with a 1 million-token context window, and offers four thinking levels so teams can trade quality against token spend. Most benchmark figures in the announcement come from third-party indexes quoted by Google, so treat them as vendor-reported until you test the model on your own tasks.

TL;DR The short version

- Gemini 4 Neon, announced by Google DeepMind on 18 November 2026, is priced at $3 per million input tokens and $15 per million output tokens, with a 1 million-token context.
- Four thinking levels let teams trade quality against token spend, and Google reports a 32% lower cost per task than Gemini 4 Carbon at the default level.
- Most benchmark figures are vendor-reported, so test 20 to 50 real tasks and record pass rate, cost and time per thinking level.

Read online, listen to it, or change the text size: https://blogs.iictp.org/gemini-4-neon-frontier-model-pricing-thinking-levels-explained

What is Gemini 4 Neon?

According to Google DeepMind's launch post, Neon is a model in the Gemini 4 family, sitting beside Gemini 4 Carbon and Gemini 4 Hydrogen. The post describes it as built for long-running agent workflows that use many tools, either on its own or as an orchestrator directing parallel Hydrogen subagents in Google Antigravity. It rolls out in the Gemini API in Google AI Studio, in Gemini for business, in Antigravity for paid AI Pro and Ultra subscribers, and in the Gemini app for Pro and Ultra subscribers.

How it works: thinking levels and cost control

- **Four thinking levels.** Google says each level scales steadily so teams can set how much reasoning, and therefore how many tokens, a job gets. Default is "high"; "xhigh" adds deeper multi-step reasoning for the hardest engineering and science work.
- **Cached input.** The post lists cached input at $0.15 per million tokens, which it describes as a 95% discount against the standard input price.

- **Agent behaviour.** The post says Neon cuts redundant file reads, reports progress while it works and handles context compaction in long sessions.

What is new compared with Gemini 4 Carbon

- At the default high level, Google reports a 32% lower cost per task than Carbon on a third-party index, and roughly 50% to 80% lower cost per task than top competing frontier models on the Vals Index and Agent Arena index.
- Google says Neon produces more secure code than Carbon across the easy, medium and hard tiers of AutoBaxBench.
- It reports a reduction of 43% to 56% in under-service and 13% to 35% in overreach on its agentic behaviour evaluation, depending on effort level.

Results reported by the source

- **Intelligence indexes (Google's figures):** a score of 61 on the Artificial Analysis Intelligence Index at xhigh, ranked #3 overall; #2 on the Vals Index at 70.9%.
- **Coding and agents:** 65.2% on Terminal-Bench 4.1 and 46.2% on FrontierCode 2.0 at the default high level; #1 on the Artificial Analysis Coding Agent Index at 58.6.
- **Professional work:** Google says Neon is #1 on the Legal Research Bench and the Tax Agent Bench, and scores 73.5% on Chartography.
- **Security:** 90.5% on SEC-bench Pro, 51.26% on WebExploitBench v1.2 and 70.7% on Collinear's CWE-bench v2, all quoted by Google.
- **Prompt injection:** with real-time safeguards at its highest thinking level, Google says Neon resisted every attack in the Gray Swan indirect prompt injection benchmark.

Limitations and open questions

- These are Google's summaries of indexes run by others. The post does not give the full evaluation settings in the text, so you cannot reproduce them from the announcement alone.
- A separate Hugging Face post on CWE-bench v2 by Collinear AI lists 11 models with Gemini 4 Argon leading at 67.8% pass@1. Neon's 70.7% appears in Google's post, not in the leaderboard table I read, so the two are not directly comparable.
- "Neon for cyber" is described as rolling out to trusted defenders in the coming days through the Fairwind Program, so it is not generally available. Sources were read on 25 November 2026. Products, rules and standards change, so check the linked originals before you rely on any detail.

How to evaluate it for your workload

1. Pick 20 to 50 real tasks from your own backlog, with a clear pass or fail check for each.
2. Run them at the default high level first and record pass rate, cost per task and wall-clock time.
3. Repeat at one level lower and one level higher to find where extra reasoning stops paying for itself.
4. Test prompt-injection resilience with your own documents and tools, not only the vendor benchmark.

What it means for engineers and platform teams

The practical change is that model choice is now two decisions: which model and which thinking level. Cost per task, not price per token, is the number to track, because a cheaper model that needs more tokens can cost more. See Gemini 4 Carbon: vendor claims and your own evals for a method you can reuse.

Frequently asked questions

- **What is Gemini 4 Neon?** A Gemini 4 family model from Google DeepMind aimed at fast, lower-cost agentic coding, knowledge work and science tasks.
- **How much does it cost?** Google lists $3 per million input tokens, $0.15 per million cached input tokens and $15 per million output tokens, with a 1 million-token context window.
- **Where can I use it?** The Gemini API in Google AI Studio, Gemini for business, Google Antigravity for Pro and Ultra subscribers, and the Gemini app for Pro and Ultra subscribers.
- **Are the benchmark numbers independent?** They come from third-party indexes, but Google chose and reported them. Run your own tests.

Related reading

- Gemini 4 Carbon: vendor claims and your own evals
- Antigravity /teamwork: multi-agent verification
- AI Practitioner certification

> Related credential: AI Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. Google DeepMind. Introducing Gemini 4 Neon. 18 November 2026 (page last updated 23 November 2026). deepmind.google/blog/introducing-gemini-4-neon
2. Gonzalez, Jagannath, Rajani, Habib, Collinear AI. CWE-bench v2: More coverage, quality, and transparency. Hugging Face, 3 November 2026. huggingface.co/blog/cwe-bench

Published by IICTP, an independent technical credential body, on 25 November 2026. The current version of this post is at https://blogs.iictp.org/gemini-4-neon-frontier-model-pricing-thinking-levels-explained. More guides: blogs.iictp.org.