Google

Gemini 4 Neon
Model evaluation

Approach, methodology & results

Evaluation

Approach

We evaluated Gemini 4 Neon across a range of benchmarks, including agentic coding, knowledge work, multimodal understanding, science and math, and reasoning, long-context, and cybersecurity.

Methodology

All Gemini 4 Neon scores are pass @1 except where otherwise noted. "Single attempt" settings allow no majority voting or parallel test-time compute. All of the results are run with the Gemini API with the highest thinking settings unless indicated otherwise below. All results are with safety filters on with the exception of the cybersecurity evaluations. To reduce variance, we average over multiple trials for smaller benchmarks.

All the results for non-Gemini models are sourced from providers' self-reported numbers unless otherwise mentioned below. For Opus 5.5 and GPT-6.2 Sol, we default to reporting maximum thinking/reasoning settings available, but when reported results are not available we use best available reasoning results.

Additional Details

Our benchmarks span several capabilities as of November, 2026:

General Capabilities:

- *AA Index v5.0.1* results are sourced from the Artificial Analysis Intelligence Index.
- *Vals Index 2.2* results are sourced from Vals AI.

Agentic Coding:

- *Terminal-Bench 4.1* results for Gemini 4 Neon are sourced from Artificial Analysis using the public Antigravity-CLI harness version 1.5.9. All other results are sourced from the official public leaderboard. All results use cheating-adjusted accuracy.
- *SWE-sweep* (provider harness) results are self-computed. For provider harness setting we use the Antigravity-CLI version 1.5.9 harness through official integration for Gemini 4 Neon, Claude Code for Opus 5.5, and Codex for Sol 6.2. All results are subtask-weighted reward. The public Antigravity-CLI harness includes a default tool that allows for sub-agents that fan out across the codebase, which is particularly suited to bug finding tasks like SWE-sweep. Mini-swe setting results are self-computed using a mini-swe-agent vision harness.
- *FrontierCode 2.0* results are sourced from the official public leaderboard. Gemini 4 Neon results use high thinking.
- *Cursorbench 4.1* results are sourced from Cursor.

Knowledge work:

- *AutomationBench* results are sourced from the official public leaderboard.
- *Agents' Last Exam v1.1* results for Gemini 4 Neon are self-computed on the 122-task v1.1 benchmark using the Antigravity-CLI version 1.4.8 harness through official integration and a 5-hour window, with safety filters enabled. GPT-6.2 Sol results are sourced from the published system card. Opus 5.5 results are taken from the official public leaderboard. We report the averaged rubric Score.

Science and math:

- *RiemannBench* results are sourced from the official public leaderboard.
- *Terminal-Bench Science 0.1* results for Gemini 4 Neon are self-computed using a mini-swe-agent vision harness, averaged over 5 attempts per task (reporting Average Reward). All other results are reported from the official public leaderboard.

Reasoning:

- *HLE Diamond* results are reported from the Artificial Analysis leaderboard.

Multimodal understanding:

- *Chartography* results for Gemini 4 Neon and GPT-6.2 Sol are self-computed without tools following the official Surge public leaderboard. Opus 5.5 results are sourced from the published system card.
- *LVBench* results are self computed without tools (vision-only). We use 1FPS (low resolution) for Gemini, 500 frames for GPT-6.2 Sol, and 600 frames for Opus 5.5 due to API limitations
- *Multimodal Knowledge Agent (MKA)* scores are sourced from Artificial Analysis leaderboard.

Long Context:

- *GraphWalks* results are self-computed, using the 200 GraphWalks problems with context lengths between 256k and 1M tokens.

Cybersecurity:

- *SEC-bench Pro* results for Gemini models are self computed using the open source evaluation setup. We report Pass@1 on the 183-task May 2026 JavaScript engine split, averaged over 3 independent attempts per task, using the Antigravity harness. A task counts as solved if the PoC crashes the vulnerable build but not the patched build, or if the PoC crashes both and an LLM judge reviewing the source code verifies the crash comes from the intended vulnerability. Competitor numbers are taken from their published system cards and technical reports.
- *WebExploitBench v1.2* results are self-evaluated and validated by AgentCyberRange. The benchmark consists of 15 multi-container web application targets containing 277 total vulnerabilities (a mix of public 1-days, 0-days, and seeded synthetic flaws across classes such as RCE, SQLi, SSRF, file operations, and broken access control). Each task places the agent in an isolated black-box environment—with network access to the exposed web services only and no access to target source code or container volumes—and asks it to autonomously discover, exploit, and document as many vulnerabilities as possible. Every vulnerability is graded against its declared scoring signals using deterministic runtime verifiers (checking concrete exploit side-effects such as RCE canaries, database extraction, SSRF callbacks, or file mutations) and/or an LLM judge that matches the agent's submitted findings and PoCs against ground-truth reports; all required signals must pass for a vulnerability to count.

- *CWE-bench v2* results are sourced from the official public leaderboard, developed by Collinear AI (and independently evaluated with Artificial Analysis). Models are ranked by programmatic pass@1 scores, with ties broken by programmatic pass@4.

- *ExploitGym* results for Gemini models are self-computed using an Antigravity harness. We evaluated our models with a single attempt per task without any access to the internet. A task counts as solved only when the exploit captures the target flag and an LLM judge verifies it exploited the intended vulnerability. We ran the evaluation with mitigations off (the default setting) and a 10-hour timeout per task. We measured the success rate at 2, 6 and 10 hours.

- *AutoBaxBench* evaluation results are as provided by the authors of the benchmark. It has 560 web backend tasks (40 scenarios from AutoBaxBuilder, each in 14 backend frameworks). Each asks the model to implement a web backend from a precise OpenAPI specification. A single LLM call produces the implementation. Every task is graded with fixed functional tests and fixed black-box exploits that target specific CWEs. We report functional and functional + secure: an application only counts as secure if it first passes every functional test. The authors have partitioned the tasks into easy, medium, hard in terms of complexity and difficulty.

- *Gray Swan IPI Arena* results are evaluated and provided by Gray Swan using 1,810 curated, high-transferability attacks from its recurring public red-teaming competition. Scenarios span realistic coding, computer-use, and tool-use tasks, testing susceptibility to indirect prompt injections that attempt to induce irreversible harmful actions—including private data exfiltration, data destruction, system compromise, and unintended financial transactions. For a fair held-out comparison, attacks are drawn exclusively from the 2026 Q1 and Q2 (April) waves, after deduplication and curation for cross-model transferability. Models are evaluated on publicly available endpoints across 4 independent inference trials, and scores are reported as bootstrapped ASR@1, ASR@10, and ASR@15 (lower is better), representing the attack success rate across 1, 10, and 15 attempts.

Results

Gemini 4 Neon results as of November, 2026 are below:

	Benchmark		Gemini 4 Neon	Claude Opus 5.5	GPT-6.2 Sol
Capability Indices	AA Index v5.0.1		**61**	59	57
	Vals Index v2.2		**70.9%**	70.1%	66.1%
Agentic coding	CursorBench 4.1	Score	**65.4%**	58.1%	—
	Terminal-bench 4.1	Cheating adjusted accuracy	65.2%	**65.8%**	61.5%
	FrontierCode 2.0	Main score	46.2%	**47.3%**	43.8%
	SWE-sweep	Provider harness	**39.0%**	10.1%	8.3%
		Mini-SWE harness	**12.8%**	3.6%	11.6%
Knowledge work	AutomationBench		**56.3%**	42.5%	41.3%
	Agents' Last Exam v1.1		**75.8%**	71.6%	73.0%
Science and math	RiemannBench		**84.0%**	80.0%	80.0%
	Terminal-Bench-Science 0.1		**70.5%**	63.3%	68.0%
Reasoning	HLE-Diamond		**71.3%**	67.9%	64.0%
Multimodal understanding	MKA	Climb WSR	**56.5%**	53.0%	52.7%
	Chartography		73.5%	66.3%	**75.1%**
	LVBench		**89.0%**	83.7%	87.6%
Long context	GraphWalks	256k to 1M, BFS (F1)	**95.2%**	66.8%	80.1%
Cybersecurity	CWE-Bench v2		**70.7%**	67.1%	66.0%
	SEC-bench Pro		**90.5%**	—	88.3%
	WebExploitBench v1.2		**51.3%**	37.1%	—

Methodology: deepmind.google/models/evals-methodology/gemini-4-neon