Google

Gemini 4 Carbon
Model evaluation

Approach, methodology & results

Evaluation

Approach

We evaluated Gemini 4 Carbon across a range of benchmarks, including agentic coding, computer use, knowledge work, science and math, multimodal capabilities, long-context, cybersecurity, and the Artificial Analysis Intelligence Index.

Methodology

All Gemini 4 Carbon and Gemini 4 Argon 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. 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 GPT-6 Astra, Opus 5.5, and Fable 5.1, 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 October 2026:

General capabilities:

- *AA Index v5.0 (with tools)* results are sourced from the Artificial Analysis Intelligence Index.

- *Vals Index v2.2* results are sourced from Vals AI.

Agentic Coding:

- *SWE-sweep* results for Gemini 4 Carbon are self-computed using a mini-swe-agent harness (reporting the subtask-weighted reward). GPT-6 Astra, Opus 5.5, and Fable 5.1 results are reported from the official SWE-sweep public leaderboard.

- *FrontierSWE 2* results are reported from Proximal's official public leaderboard.

- *Terminal-Bench 4.0 (AA)* results are sourced from Artificial Analysis.

Computer use:

- *Agents' Last Exam v1.1* results for Gemini 4 Carbon and Gemini 4 Argon are self-computed on the 122-task v1.1 benchmark using the default ALE-Claw harness and a 5-hour window, with safety filters enabled. We use the task version released on Oct 17, 2026, without applying the Oct 26, 2026 update. Responses flagged by safety filters are returned as empty strings, though the episode can continue. GPT-6 Astra and Opus 5.5 results are taken from the official public leaderboard. Fable 5.1 results are missing from the public leaderboard so are not included. We report the averaged rubric Score.

- *OSWorld 2.1* scores are reported on the 82-task offline (no-internet) subset for apples-to-apples comparison across models. Scores for Gemini 4 Carbon and Gemini 4 Argon are self-computed, maxed over 3 runs with a single attempt per run. Responses flagged by safety filters are returned as empty strings, though the episode can continue. We report the partial score from the offline subset. We use the official GitHub OSWorld docker, official evaluator, default 1080p resolution, and max step length of 500. We use an updated Gemini CUA harness, with parallel batch tool calling enabled, that uses all screenshots in history and uses a compaction procedure following the same setup as reported in the Opus 5.5 system card. We use PyAutoGUI for actuation with screenshot-only observation space, and for tools, we use UI-specific function declarations. GPT-6 Astra, Opus 5.5, and Fable 5.1 offline subset scores are sourced from their respective published evaluations and system cards.

Knowledge work:

- *AutomationBench* results use the private set and are sourced from the official Zapier public leaderboard.

- *Vending-Bench 2* results were reported from Andon Labs official public leaderboard.

Multimodal understanding:

- *Chartography* results are evaluated without tools following the official Surge public leaderboard.

- *Blueprint-Bench 2* results were reported from Andon Labs official public leaderboard.

Long Context:

- *GraphWalks* results for all models are self-computed. The “256k to 1M” subset uses the 200 GraphWalks problems with context lengths between 256k and 1M tokens (reporting the BFS F1 score).

Science and math:

- *RiemannBench* results are reported from the official Surge public leaderboard.
- *Terminal-Bench Science 0.1 (AA)* results are sourced from Artificial Analysis.

Cybersecurity:

- *CWE-bench v1* (shown as CWE-bench in the table) 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 are self computed using the open source evaluation setup. We evaluated our models with a single attempt per task using the Antigravity harness, 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 in the default benchmark setting (with benchmark mitigations off), and measured the success rate under both the default 6 hour timeout per task, as well as an extended 10 hour timeout. Competitor numbers are taken from their published system cards and technical reports.
- *SEC-bench Pro* results 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.

Results

Gemini 4 Carbon results as of October 2026 are below (cybersecurity benchmark results reflect performance when accessed through the Fairwind Program):

	Benchmark		Gemini 4 Carbon	Gemini 4 Argon	GPT-6 Astra	Claude Opus 5.5	Claude Fable 5.1
General capabilities	AA Index v5.0		**59**	50	57	58	49
	Vals Index v2.2		**70.4%**	69.2%	65.0%	70.2%	—
Agentic coding	SWE-sweep		**9.6%**	4.3%	9.5%	3.6%	2.9%
	FrontierSWE 2		61.2%	55.0%	**65.5%**	62.3%	56.3%
	Terminal-Bench 4.0 (AA)		**65.2%**	57.1%	59.1%	59.6%	52.0%
Computer use	Agents' Last Exam v1.1	Score	**77.4%**	69.0%	73.1%	71.6%	—
	OSWorld-2.1	Offline subset	81.5%	74.7%	80.9%	**87.2%**	76.8%
Knowledge work	AutomationBench	Zapier	**53.9%**	51.3%	41.4%	42.5%	31.4%
	Vending-Bench 2	Bank balance ($)	**$16,484**	$13,718	$15,515	$9,235	$5,422
Multimodal understanding	Chartography		**72.5%**	71.6%	71.0%	66.3%	46.2%
	Blueprint-Bench 2	Similarity score	**58.6%**	54.4%	49.7%	51.2%	41.9%
Long context	GraphWalks	256k to 1M, BFS (F1)	**96.2%**	84.2%	71.8%	66.8%	65.0%
Science and math	RiemannBench		**80.0%**	76.0%	72.0%	69.6%	65.6%
	Terminal-Bench-Science 0.1 (AA)		**64.8%**	52.9%	63.3%	59.0%	43.3%
Cybersecurity	CWE-bench		**71.0%**	68.0%	68.0%	67.0%	58.0%
	ExploitGym		**54.8%**	42.0%	42.4%	34.5%	—
	SEC-bench Pro		**88.2%**	83.4%	85.4%	—	—

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