Google

Gemini 4 Hydrogen Model evaluation

Approach, methodology & results

Evaluation

Approach

We evaluated Gemini 4 Hydrogen across a range of benchmarks, including agentic coding, knowledge work, computer use, multimodal understanding, science and math, and reasoning.

Methodology

All Gemini 4 Hydrogen 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 Sonnet 5.5 and GPT-6.1 Terra, 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:

Agentic Coding:

- *SWE-sweep* results for Gemini 4 Hydrogen are self-computed using the Antigravity-CLI (agy) version 1.4.8 harness through official integration with safety filters enabled. (reporting the subtask-weighted reward). GPT-6.1 Terra, and Sonnet 5.5 results are reported from the official public leaderboard.
- *Terminal-Bench 4.1* results for Gemini 4 Hydrogen are self-computed using a mini-swe-agent vision harness averaged over 5 attempts per task (reporting Average Reward) using high effort level. Claude Sonnet 5.5 results are taken from the official public leaderboard. All results use cheating-adjusted accuracy.
- *Terminal-Bench 4.0* results for Gemini 4 Hydrogen are self-computed using a mini-swe-agent vision harness averaged over 5 attempts per task (reporting Average Reward) using high effort level. GPT-6.1 Terra and Claude Sonnet 5.5 results are taken from the official public leaderboard. We report cost vs performance results for 4.0 because competitor data was not available for 4.1 across thinking levels.

Knowledge work:

- *Agents' Last Exam v1.1* results for Gemini 4 Hydrogen are self-computed on the 122-task v1.1 benchmark using the Antigravity-CLI (agy) version 1.4.8 harness through official integration and a 5-hour window, with safety filters enabled. Sonnet 5.5 results are computed using the ALE-Claw under the same setup. 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-6.1 Sol, GPT-6.1 Terra results are taken from the official public leaderboard. We report the averaged rubric Score.
- *AutomationBench* results are sourced from the Artificial Analysis leaderboard.
- *Vending-Bench 2* results were reported from Andon Labs official public leaderboard.

Computer use:

- *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 Hydrogen and Sonnet 5.5 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-6.1 Terra offline subset scores are sourced from the published system card.

Multimodal understanding:

- *MKA* results for Gemini 4 Hydrogen are self-computed. All other results are sourced from the Artificial Analysis leaderboard.
- *Chartography* results for Gemini 4 Hydrogen and GPT-6.1 Terra are self-computed without tools following the official Surge public leaderboard. Sonnet 5.5 results are sourced from the published system card.

Science and math:

- *Terminal-Bench Science 0.1* results for Gemini 4 Hydrogen 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.

Results

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

	Benchmark		Gemini 4 Hydrogen	Claude Sonnet 5.5	GPT-6.1 Terra
Agentic coding	SWE-sweep		**13.5%**	3.1%	5.8%
	Terminal-Bench 4.1	Cheating-adjusted	57.0%	**63.9%**	—
Knowledge work	Agents' Last Exam v1.1	Score	**71.6%**	60.4%	69.8%
	AutomationBench-AA		**76.1%**	71.8%	60.2%
	Vending-Bench 2	Bank balance ($)	**$14,549**	$11,165	$9,107
Computer use	OSWorld-2.1	Offline subset	77.6%	**83.9%**	74.2%
Multimodal understanding	MKA	Climb WSR accuracy	**53.4%**	52.0%	43.8%
	Chartography		**66.6%**	61.6%	63.2%
Science and math	Terminal-Bench-Science 0.1		51.4%	**59.9%**	58.3%
Reasoning	HLE-Diamond		**58.8%**	53.3%	47.3%

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