Google

Gemini 3.8 Flash
Model evaluation

Approach, methodology & results

Approach

We evaluated Gemini 3.8 Flash across a range of benchmarks, including coding, knowledge work, multimodal capabilities, long-context, computer use and scientific reasoning.

Methodology

All Gemini 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 for the model-id gemini-3.8-flash with default sampling 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-5.6 Terra and Sonnet 5, 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 September, 2026:

- **Coding**
  - *DeepSWE v1.1* results are reported from the public leaderboard from deepswe.datacurve.ai. We use the highest scoring thinking level for each model as reported by Datacurve on the main leaderboard. Results for Gemini 3.8 Flash are self computed, and use a mini-swe agent harness with high thinking. Note: We originally incorrectly reported Opus 5's score as 74% due to rounding on the Datacurve public leaderboard.
  - *Terminal-Bench 2.1* results for Gemini models are self computed and for other models are reported from the public leaderboard and Artificial Analysis. Results are reported for the default agent harness (Terminus 2) only.
  - *Terminal-Bench 4.0* results are reported from the official public leaderboard. We use the highest scoring thinking level for each model as reported by the Terminal Bench authors.
- **Knowledge work:**
  - *GDPVal-AA v2* results are sourced from the Artificial Analysis public leaderboard.
  - *Vals Finance Agent v2* results are sourced from Vals.AI.
  - *Harvey's Legal Agent Benchmark* results are sourced from Vals.AI.
- **Multimodal:**
  - *GDP.PDF* results for all models are self computed.
  - *CharXiv Reasoning* results for Gemini models, GPT-5.6 Terra and Sol, and Opus 5 are self computed.
  - *LVBench* results are self computed without tools. We use 1024 frames for Gemini and GPT-5.6 models and 300 frames for Claude models due to API limitations.
- **Long Context:** *GDM-MRCR v2* includes 128k results as a cumulative score. The full dataset is available in our repository. Results are self computed.

- **Computer use:**
    - *OSWorld 2.0* scores for Gemini models and Sonnet 5 are self computed, maxed over 3 runs with a single attempt per run. We report the partial score. We use the github OSWorld 2.0 docker, official evaluator, default 1080p resolution, and max step length of 500 and added batched tool call in our harness for all models. We use the Anthropic and Gemini CUA harness provided in the OSWorld 2.0 repo. We use pyautogui for actuation with screenshot-only observation space, and for tools we use UI-specific function declarations. Runs were completed before the official OSWorld 2.0 team's 08.08 patch to be compatible with competitor's self-reported numbers. GPT-5.6 Terra results are taken from the official blog post, and Opus 5 results are taken from the Fable 5.1 blog post.
- **Scientific Reasoning:**
    - *HLE-Verified* results for all models are self computed. We report accuracy on the full 1,811 item verified set from https://arxiv.org/pdf/2602.13964, including the 668 verified items from the original HLE set, and the 1143 revised items. We exclude the 689 items from the original HLE set identified as uncertain. Note: a significant proportion of questions were blocked by content policy filters for Sonnet 5 and a small number for Opus 5.
    - *BioMysteryBench* results for Gemini models and GPT-5.6 Terra and Sol are self computed. Models are given access to a Linux terminal with pre-installed bioinfo tools, Python, and R, as well as internet access. Internet access is restricted to a limited allowed domain list specified by the benchmark authors.
    - *LABBench2* results are self computed. Models are given access to a Linux terminal with pre-installed bioinfo tools, Python, and R, as well as internet access. Overall performance is calculated as the macro-average across 11 sub-tasks—DbQA2, FigQA2-img, FigQA2-pdf, LitQA3, PatentQA, ProtocolQA2, SourceQuality, SuppQA2, TableQA2-img, TableQA2-pdf, and TrialQA.

Results

Gemini 3.8 Flash results as of September, 2026 are below:

	Gemini 3.8 Flash	Gemini 3.7 Flash	Claude Opus 5	Claude Sonnet 5	GPT-5.6 Sol	GPT-5.6 Terra
Input price $/1M tokens, no caching	**$0.75** ($1.50 regular)	**$0.75** ($1.50 regular)	$5.00	$2.00	$4.00	$2.00
Output price $/1M tokens	**$3.75** ($7.50 regular)	**$3.75** ($7.50 regular)	$25.00	$10.00	$20.00	$12.00
DeepSWE v1.1 Long-horizon software engineering	73.7%	65.3%	**74.0%**	53.8%	72.7%	69.6%
GDPVal-AA v2 Knowledge work Elo	1545	1482	**1824**	1584	1710	1528
Vals Finance Agent v2 Financial analyst tasks	**61.4%**	59.0%	58.6%	53.9%	53.8%	54.4%
Harvey's Legal Agent Benchmark Complex legal workflows All pass rate	**10.0%**	8.8%	6.7%	5.0%	2.5%	0.8%
Terminal-bench 2.1 Agentic terminal coding	**89.4%**	85.8%	89.1%	80.4%	88.8%	87.4%
Terminal-bench 4.0 General agent capabilities	19.1%	11.2%	**51.8%**	12.4%	37.3%	23.6%
GDP.PDF Expert PDF document comprehension All pass rate	35.0%	34.0%	37.0%	28.0%	**40.0%**	29.0%
CharXiv Reasoning Information synthesis from complex charts No tools	**86.2%**	84.5%	83.7%	70.1%	85.8%	85.9%
LVBench Long video understanding	**87.8%** (agentic) **87.1%** (static)	85.4%	75.4%	68.5%	82.1%	78.9%
HLE-Verified Multidisciplinary expert reasoning	**54.9%**	53.6%	54.4%	31.0%	54.5%	51.1%
OSWorld-2.0 Agentic computer use Partial score batch tool enabled	59.0%	50.6%	**75.4%**	42.6%	62.6%	50.2%
BioMysteryBench Bioinformatics research workflows Human Solvable	88.8%	87.1%	**90.1%**	87.5%	79.5%	83.8%
BioMysteryBench Human Difficult	**56.5%**	43.5%	49.4%	34.1%	44.7%	49.4%
LABBench2 Biology real-world research tasks	**86.2%**	82.1%	84.2%	80.1%	82.1%	81.2%

Methodology: deepmind.google/models/evals-methodology/gemini-3-8-flash

* For 3.7 and 3.8 Flash, introductory price expires on December 31, 2026. Starting January 1, 2027, $1.50/1M input tokens and $7.50/1M output tokens will apply.