Google

Gemini 3.6 Flash
Model evaluation

Approach, methodology & results

Approach

We evaluated Gemini 3.6 Flash across a range of benchmarks, including coding, expert tasks, UI control, multimodal capabilities, and long-context.

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-3.6-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 Luna, Sonnet 5, and Grok 4.5 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 July, 2026:

- **Coding**
  - *SWE-Bench Pro* results for Gemini models are self-computed, follow official provider methodology, using an internal version of the Antigravity harness. We report full results for completeness.
  - *DeepSWE v1.1* results for Gemini models 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, which is medium reasoning for Gemini 3.5 and high reasoning for Gemini 3.6 Flash.
  - *Terminal-Bench 2.1* results for Gemini 3.6 Flash and 3.5 Flash 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.
  - *MLE-Bench* results are self computed. We evaluate on the canonical MLE-bench Partial 30 subset, reporting the Average Position Score across $k = 2$ independent runs per problem. The Position Score normalizes performance per competition by measuring the agent's percentile rank against the historical Kaggle private leaderboard via the formula $\frac{N-r+1}{N}$, where r is the agent's achieved rank among N human submissions (with 1.0 representing 1st place and 0.0 representing a missing/failing submission). The agent interacts through an interactive Bash terminal harness featuring a 1M token context window, with each problem instance executing inside an isolated sandbox container equipped with an NVIDIA H100 GPU and internet access. While budgeted for up to 256

tool calls per problem with a 20-minute command timeout, most rollouts complete within 12 hours. All the reported numbers are self computed.

- **Expert Tasks:** *GDPVal-AA v2* results are sourced from the Artificial Analysis public leaderboard.
- **UI Control:** *OSWorld-Verified* scores for Gemini models and GPT-5.6 Luna are self-computed, averaged over 5 runs with a single attempt per run. We use the github OSWorld docker, official methodology, default 1080p resolution, and max step length of 100. We use pyautogui for actuation, and for tools we use UI-specific function declarations.
- **Image & Video:** *CharXiv Reasoning* results for Gemini models, GPT 5.6 Luna and Grok 4.5 are self-computed. Gemini's "with tools" score uses search and code execution.
- **Long Context:** *GDM-MRCR v2* includes 128k results as a cumulative score to ensure they can be comparable with other models and a pointwise value for 1M context window to show the capability of the model at full length. The full dataset is available in our repository. Results for Gemini models and Grok 4.5 are self-computed. Available scores for other models are sourced from https://contextarena.ai/. The 1M version of the eval does not fit into the context window of non-Gemini models.

Results

Gemini 3.6 Flash results as of July, 2026 are below:

Benchmark		Gemini 3.6 Flash	Gemini 3.5 Flash	Gemini 3.1 Pro	GPT-5.6 Luna	Grok 4.5	Claude Sonnet 5
Input price $/1M tokens		$1.50	$1.50	$2.00	$1.00	$2.00	$3.00 (full-price) $2.00 (temp discount)
Output price $/1M tokens		$7.50	$9.00	$12.00	$6.00	$6.00	$15.00 (full-price) $10.00 (temp discount)
SWE-Bench Pro (Public) Diverse agentic coding tasks		58.7%	55.1%	54.2%	62.7%	**64.7%**	63.2%
DeepSWE v1.1 Long-horizon software engineering		49%	37%	12%	**67%**	54%	54%
Terminal-bench 2.1 Agentic terminal coding	Terminus-2 harness	78.0%	76.2%	73.8%	**84.7%**	83.3%	80.4%
MLE-Bench Machine Learning Engineering		63.9%	49.7%	42.6%	47.6%	43.2%	**66.9%**
GDPVal-AA v2 Knowledge work	Elo	1421	1349	965	1584	1535	**1607**
OSWorld-Verified Computer use		**83.0%**	78.4%	76.2%	72.6%	-	81.2%
CharXiv Reasoning Information synthesis from complex charts	no tools	**85.2%**	84.2%	83.3%	82.7%	81.6%	77.0%
	with tools	**89.4%**	84.9%	83.2%	-	-	88.3%
GDM-MRCR v2 (8-needle) Long context performance	128k (average)	**91.8%**	77.3%	84.9%	74.8%	81.4%	71.6%
	1M (pointwise)	**54.0%**	26.6%	26.3%	-	-	-

Methodology: deepmind.google/models/evals-methodology/gemini-3-6-flash