Google

Gemini 3.5 Flash-Lite Model evaluation

Approach, methodology & results

Approach

We evaluated Gemini 3.5 Flash-Lite 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-3.5-flash-lite with high thinking and 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 5.4 Mini and Claude Haiku 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 following official provider methodology, using an internal version of the Antigravity harness. We report the full public set results for completeness.
  - *Terminal-Bench 2.1* results for Gemini models and 4.5 Haiku 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.
- **Expert Tasks:** *GDPVal-AA v2* results are sourced from the Artificial Analysis public leaderboard.

- **UI Control:** *OSWorld-Verified* scores for Gemini models 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:** *CharXiv Reasoning* results for all models are self-computed. Gemini's "with tools" score is with 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 all models are self-computed. 1M version of the eval does not fit into the context window of non-Gemini models. Available scores for other models are also sourced from https://contextarena.ai/.

Results

Gemini 3.5 Flash-Lite results as of July, 2026 are below:

Benchmark		Gemini 3.5 Flash-Lite	Gemini 3.1 Flash-Lite	GPT-5.4 mini	Claude Haiku 4.5
Input price $/1M tokens		$0.30	$0.25	$0.75	$1.00
Output price $/1M tokens		$2.50	$1.50	$4.50	$5.00
SWE-Bench Pro (Public) Diverse agentic coding tasks		54.2%	38.3%	**54.4%**	39.5%
Terminal-bench 2.1 Agentic terminal coding	Terminus-2 harness	54.0%	31.0%	**59.2%**	44.2%
MLE-Bench Machine Learning Engineering		**39.2%**	22.0%	-	-
GDPVal-AA v2 Knowledge work	Elo	1140	642	**1171**	907
OSWorld-Verified Computer use		**74.0%**	54.3%	72.1%	50.7%
CharXiv Reasoning Information synthesis from complex charts	no tools	74.5%	73.2%	**80.3%**	61.7%
	with tools	**76.5%**	75.6%	-	-
GDM-MRCR v2 (8-needle) Long context performance	128k (average)	**72.2%**	60.1%	42.7%	35.3%
	1M (pointwise)	**21.3%**	12.3%	-	-

Methodology: deepmind.google/models/evals-methodology/gemini-3-5-flash-lite