Model Evaluation – Approach, Methodology & Results

Gemini 3.5 Flash

Approach: Gemini 3.5 Flash was evaluated across a range of benchmarks, including coding, reasoning, multimodal capabilities, agentic tool use, 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 all run with the Gemini API for the model-id gemini-3.5-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 Claude Opus 4.7, Sonnet 4.6, and GPT-5.5 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 May, 2026:

- **Coding**
  - *Terminal-Bench 2.1* results for Gemini 3.5 Flash and 3 Flash are self computed and for other models are reported from the public leaderboard. Results are reported for the default agent harness (Terminus 2) only.
  - *SWE-Bench Pro* results for Gemini models are self-computed. Scores are averaged over 5x runs, follow official provider methodology, using an internal version of the Antigravity harness. We report full results for completeness.
- **Agentic**
  - *MCP Atlas* results are reported from the ScaleAI official leaderboard.
  - *Toolathlon* results for the Gemini models are provided by the benchmarks authors at HKUST.
- **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.
- **Expert tasks**
  - *Finance Agent v2* results are sourced from the vals.ai public leaderboard.
  - *GDPval-AA* results are sourced from the Artificial Analysis public leaderboard.
- **Image**
  - *CharXiv Reasoning* results for Gemini models and GPT 5.5 are self-computed.
  - *MMMU-Pro* scores are averaged across the Standard (10 options) and Vision settings. Results for Gemini models and Opus 4.7 are self-computed.
  - *Blueprint Bench 2* results are sourced from the Andon Labs Blueprint-Bench 2 public leaderboard. Leaderboard scores are normalized to a 0-100 scale for ease of comparison.
- **Long Context**
  - *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.
- **Reasoning and Academic Knowledge:**
  - *Humanity's Last Exam* results for Gemini models are self-computed.
  - *ARC-AGI-2* results are sourced from the ARC Prize leaderboard and are ARC Prize Verified. The set reported is semi-private.

Results: Gemini 3.5 Flash results as of May, 2026 are below:

	Benchmark		Gemini 3.5 Flash	Gemini 3 Flash	Gemini 3.1 Pro	Claude Sonnet 4.6	Claude Opus 4.7	GPT-5.5
Coding	Terminal-bench 2.1 Agentic terminal coding	Terminus-2 harness	76.2%	58.0%	70.3%	-	66.1%	**78.2%**
	SWE-Bench Pro (Public) Diverse agentic coding tasks	Single attempt	53.9%	48.4%	54.2%	53.0%	**64.3%**	58.6%
Agentic	MCP Atlas Multi-step workflows using MCP		**83.6%**	62.0%	78.2%	69.5%	79.1%	75.3%
	Toolathlon Real-world general tool use		**56.5%**	49.4%	-	-	-	55.6%
UI control	OSWorld-Verified Agentic computer use		78.4%	65.1%	76.2%	72.5%	78.0%	**78.7%**
Expert tasks	Finance Agent v2 Financial analysis and decision-making		**57.9%**	42.6%	43.0%	51.0%	51.5%	51.8%
	GDPval-AA Economically valuable knowledge work	Elo	1656	1204	1314	1674	1753	**1773**
Multimodal	CharXiv Reasoning Information synthesis from complex charts	No tools	**84.2%**	80.3%	83.3%	70.5%	82.1%	84.1%
	MMMU-Pro Multimodal understanding and reasoning	No tools	**83.6%**	81.2%	80.5%	74.5%	75.2%	81.2%
	Blueprint-Bench 2 Agentic spatial reasoning	Normalized score	33.6%	0.0%	26.5%	6.7%	24.5%	**36.2%**
Long context	MRCR v2 (8-needle) Long context performance	128k (average)	77.3%	67.2%	84.9%	84.9%	59.3%	**94.8%**
		1M (pointwise)	**26.6%**	22.1%	26.3%	Not supported	Not supported	Not supported
Reasoning	Humanity's Last Exam Academic reasoning (full set, text + MM)		40.2%	33.7%	44.4%	33.2%	**46.9%**	41.4%
	ARC-AGI-2 Abstract reasoning puzzles		72.1%	33.6%	77.1%	58.3%	75.8%	**85.0%**

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