Model Evaluation – Approach, Methodology & Results

Gemini 3 Flash

Approach: Gemini 3 Flash was evaluated to analyze its core capabilities across domains.

Methodology: All Gemini scores are pass @1 and run with Gemini API for the model id gemini-3-flash-preview using the default API sampling settings unless indicated otherwise. "Single attempt" settings allow no majority voting or parallel test-time compute. 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 mentioned otherwise for individual evals below. For Claude Sonnet 4.5, and GPT-5.2 we default to reporting high and xhigh reasoning results respectively, but when reported results are not available we use best available reasoning results. For Grok 4.1 Fast we use the reasoning settings and since the self reported numbers were not available we resort to using eval results from 3p providers (AIME2025, GPQA, SWE-Bench from Vals.ai and Humanity's Last Exam and MMMU-Pro from Artificial Analysis). Otherwise, where self reported or official leaderboard numbers were not available (MMMU-Pro, ScreenSpot-Pro, CharXiv Reasoning, OmniDocBench 1.5, Video-MMMU, MMMLU, Global PIQA) they were computed by Google DeepMind using official provider APIs.

Setup: Our benchmarks span several capabilities, details below:

- **Reasoning and Academic Knowledge:** We test the model's ability to draw logical conclusions, and reason about mathematical, scientific, and common-sense problems.
    - *Humanity's Last Exam* results for Gemini 3 Pro, 2.5 Pro, 2.5 Flash and Claude Sonnet 4.5 are from Scale AI leaderboard & GPT-5.2 and Gemini 3 Flash from self reported numbers,
    - *ARC-AGI* results are sourced from the ARC Prize website and are using the semi-private set.

- **Image**
    - *MMMU-Pro* scores are averaged across the Standard (10 options) and Vision settings and from self-reported numbers and are without the use of any tools
    - *ScreenSpotPro* results for Gemini 3 require setting the media_resolution to "ultra_high". GPT-5.2 results use python so should not be directly comparable with others.
    - *CharXiV Reasoning* results are on 1000 reasoning questions from the validation split of CharXiv.
    - *OmniDocBench1.5* results are the average Edit Distance across the Text, Formula, Table, and ReadingOrder sub-metrics using the official OmniDocBench code and data, following the exact methodology from DeepSeekOCR (https://arxiv.org/abs/2510.18234). Lower score is better. For OpenAI 5.2 models we have been unable to obtain results on the xhigh setting due to high error rate so the score reported is for the high setting.

- **Video:**
  - *Video-MMMU* results for Gemini models are computed with the recommended setting using media_resolution=HIGH (280 tokens per frame) and temperature = 0.

- **Code**
  - *LiveCodeBench Pro*: We report ELO Rating in the table. Scores for existing models are from the public LiveCodeBench Pro leaderboard.
  - *Terminal-Bench 2.0* results are reported from the public leaderboard and follow the default agent harness (Terminus-2). Gemini 3 Flash submission is pending at the time of publication.
  - *SWE-bench Verified* numbers follow official provider reports, using different scaffoldings and infrastructure. Our scaffolding is single-attempt only, composed of a bash tool to run shell commands, file operation tools to make actions such as editing and undoing easier, and a submit tool. Averaged over 5x runs.

- **Tool Use**
  - *τ2-bench* results for Gemini use standard Sierra framework with a prompt adjustment to provide instructions relevant to each environment. The user model uses Gemini 3 Pro with a custom system instruction. All scores reported above are the simple average of scores on the three individual categories: Retail, Airline and Telecom. For the Airline domain we adopt the fixes to the domain as proposed in the Claude Opus 4.5 release report. Hence, we report numbers on Retail, Airline (fixed) and Telecom. The difference in reported numbers for older models come from the difference after using the Airline fix.
  - *Toolathlon (Tool Decathlon)* results are obtained from runs completed by the benchmark authors, from the official leaderboard, and from self-reported results in the case of GPT-5.2.
  - *MCP Atlas* results are sourced from Scale AI's MCP Atlas leaderboard, and from self-reported results in the case of GPT-5.2.
  - *Vending-bench 2* results are reported from https://andonlabs.com/evals/vending-bench-2

- **Factuality**
  - *FACTS Benchmark Suite* results are not directly comparable to our previously reported FACTS Grounding results as they represent a more robust set of factuality related benchmarks which were recently launched and the results are available on Kaggle leaderboard.
  - *SimpleQA Verified* results are reported from the official Kaggle leaderboard.

- **Language**
  - *MMMLU results are a combination of officially reported numbers (Sonnet 4.5 and GPT-5.2) and runs completed by GDM.*
  - *Global PIQA results are obtained from GDM runs.*

- **Long Context**: For MRCR v2 which is not publicly available yet we include 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. We do not report MRCR v2 results at 1M for GPT-5.2 due to its 400k token context window, and for Claude Sonnet 4.5 because our 1M token evaluation setup does not fit into Sonnet 4.5's 1M token context window.

Results: benchmarks as of December, 2025 are below:

Benchmark	Description		Gemini 3 Flash Thinking	Gemini 3 Pro Thinking	Gemini 2.5 Flash Thinking	Gemini 2.5 Pro Thinking	Claude Sonnet 4.5 Thinking	GPT-5.2 Extra high	Grok 4.1 Fast Reasoning
Humanity's Last Exam	Academic reasoning (full set, text + MM)	No tools	33.7%	**37.5%**	11.0%	21.6%	13.7%	34.5%	17.6%
		With search and code execution	43.5%	**45.8%**	—	—	—	45.5%	—
ARC-AGI-2	Visual reasoning puzzles	ARC Prize Verified	33.6%	31.1%	2.5%	4.9%	13.6%	**52.9%**	—
GPQA Diamond	Scientific knowledge	No tools	90.4%	91.9%	82.8%	86.4%	83.4%	**92.4%**	84.3%
AIME 2025	Mathematics	No tools	95.2%	95.0%	72.0%	88.0%	87.0%	**100%**	91.9%
		With code execution	99.7%	**100%**	75.7%	—	**100%**	—	—
MMMU-Pro	Multimodal understanding and reasoning		**81.2%**	81.0%	66.7%	68.0%	68.0%	79.5%	63.0%
ScreenSpot-Pro	Screen understanding	No tools unless specified	69.1%	72.7%	3.9%	11.4%	36.2%	**86.3%** with python	—
CharXiv Reasoning	Information synthesis from complex charts	No tools	80.3%	81.4%	63.7%	69.6%	68.5%	**82.1%**	—
OmniDocBench 1.5	OCR	Overall Edit Distance, lower is better	0.121	**0.115**	0.154	0.145	0.145	0.143	—
Video-MMMU	Knowledge acquisition from videos		86.9%	**87.6%**	79.2%	83.6%	77.8%	85.9%	—
LiveCodeBench Pro	Competitive coding problems from Codeforces, ICPC, and IOI	Elo rating, higher is better	2316	**2439**	1143	1775	1418	2393	—
Terminal-bench 2.0	Agentic terminal coding	Terminus-2 harness	47.6%	**54.2%**	16.9%	32.6%	42.8%	—	—
SWE-bench Verified	Agentic coding	Single attempt	78.0%	76.2%	60.4%	59.6%	77.2%	**80.0%**	50.6%
τ2-bench	Agentic tool use		90.2%	**90.7%**	79.5%	77.8%	87.2%	—	—
Toolathlon	Long horizon real-world software tasks		**49.4%**	36.4%	3.7%	10.5%	38.9%	46.3%	—
MCP Atlas	Multi-step workflows using MCP		57.4%	54.1%	3.4%	8.8%	43.8%	**60.6%**	—
Vending-Bench 2	Agentic long term coherence	Net worth (mean), higher is better	$3,635	**$5,478**	$549	$574	$3,839	$3,952	$1,107
FACTS Benchmark Suite	Factuality benchmark across grounding, parametric, search, and MM		61.9%	**70.5%**	50.4%	63.4%	48.9%	61.4%	42.1%
SimpleQA Verified	Parametric knowledge		68.7%	**72.1%**	28.1%	54.5%	29.3%	38.0%	19.5%
MMMLU	Multilingual Q&A		**91.8%**	**91.8%**	86.6%	89.5%	89.1%	89.6%	86.8%
Global PIQA	Commonsense reasoning across 100 Languages and Cultures		92.8%	**93.4%**	90.2%	91.5%	90.1%	91.2%	85.6%
MRCR v2 (8-needle)	Long context performance	128k (average)	67.2%	77.0%	54.3%	58.0%	47.1% not supported	**81.9%** not supported	54.6%
		1M (pointwise)	22.1%	**26.3%**	21.0%	16.4%			6.1%