Model Evaluation – Approach, Methodology & Results

Gemini 3 Pro

Approach: Gemini 3 Pro was evaluated across a range of benchmarks, including reasoning, multimodal capabilities, agentic tool use, multi-lingual performance, and long-context.

Methodology: All Gemini scores are pass @1. "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-pro-preview with default sampling settings unless indicated otherwise. 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 below. For Claude Sonnet 4.5, and GPT-5.1 we default to reporting high reasoning results, but when reported results are not available we use best available reasoning results. Google DeepMind calculated the following scores using official provider APIs, as self-reported or official leaderboard numbers were unavailable: MMMU-Pro, ScreenSpot-Pro, CharXiv Reasoning, OmniDocBench 1.5, Video-MMMU, MMMLU, and Global PIQA unless indicated otherwise.

Additional Details: Our benchmarks span several capabilities as of November, 2025:

- **Reasoning and Academic Knowledge:**
  - *Humanity's Last Exam* results for Gemini 2.5 Pro and Claude Sonnet 4.5 are from ScaleAI leaderboard & GPT-5.1 from Artificial Analysis due to result availability. Gemini 3 Pro results are self-computed. For search and code on results we run the Gemini model using Gemini API with a blocklist implemented to avoid results that could include benchmark numbers like huggingface.com and others.
  - *ARC-AGI-2* results are sourced from the ARC Prize website and are ARC Prize Verified. The set reported is semi-private.
  - *MathArena Apex* results are reported by matharena.ai.

- **Image**
  - *MMMU-Pro* scores are averaged across the Standard (10 options) and Vision settings. GPT-5.1 results are sourced from Artificial Analysis.
  - *ScreenSpotPro* results for Gemini 3 use function calling with a "capture screenshot" tool that passes the captured screenshot back to the model, and the media_resolution parameter to "extra_high". This setting is coming soon to our API. With media_resoulution set to "high", Gemini 3.0 scores 60.5.
  - *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).

- **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 below table. Scores for existing models are from the public leaderboard.
  - *Terminal-Bench 2.0* results are reported from the public leaderboard and follow the default agent harness (Terminus 2).
  - *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 10x 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 with a system instruction. All scores reported above are the average of scores on the three individual categories: Retail, Airline and Telecom. Gemini 3.0 Pro scores 85.3%, 73.0% and 98.0% on these categories respectively.
  - *Vending-bench 2* results are reported from https://andonlabs.com/evals/vending-bench.

- **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 we will be releasing soon.
  - SimpleQA Verified results are reported from the official Kaggle leaderboard.

- **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

Results: Gemini 3 Pro significantly outperforms Gemini 2.5 Pro across our range of aforementioned benchmarks. Results as of November, 2025 are listed below:

Benchmark	Description		Gemini 3 Pro	Gemini 2.5 Pro	Claude Sonnet 4.5	GPT-5.1
Humanity's Last Exam	Academic reasoning	No tools With search and code execution	**37.5%** **45.8%**	21.6% —	13.7% —	26.5% —
ARC-AGI-2	Visual reasoning puzzles	ARC Prize Verified	**31.1%**	4.9%	13.6%	17.6%
GPQA Diamond	Scientific knowledge	No tools	**91.9%**	86.4%	83.4%	88.1%
AIME 2025	Mathematics	No tools With code execution	**95.0%** **100%**	88.0% —	87.0% **100%**	94.0% —
MathArena Apex	Challenging Math Contest problems		**23.4%**	0.5%	1.6%	1.0%
MMMU-Pro	Multimodal understanding and reasoning		**81.0%**	68.0%	68.0%	76.0%
ScreenSpot-Pro	Screen understanding		**72.7%**	11.4%	36.2%	3.5%
CharXiv Reasoning	Information synthesis from complex charts		**81.4%**	69.6%	68.5%	69.5%
OmniDocBench 1.5	OCR	Overall Edit Distance, lower is better	**0.115**	0.145	0.145	0.147
Video-MMMU	Knowledge acquisition from videos		**87.6%**	83.6%	77.8%	80.4%
LiveCodeBench Pro	Competitive coding problems from Codeforces, ICPC, and IOI	Elo Rating, higher is better	**2,439**	1,775	1,418	2,243
Terminal-Bench 2.0	Agentic terminal coding	Terminus-2 agent	**54.2%**	32.6%	42.8%	47.6%
SWE-Bench Verified	Agentic coding	Single attempt	76.2%	59.6%	**77.2%**	76.3%
τ2-bench	Agentic tool use		**85.4%**	54.9%	84.7%	80.2%
Vending-Bench 2	Long-horizon agentic tasks	Net worth (mean), higher is better	**$5,478.16**	$573.64	$3,838.74	$1,473.43
FACTS Benchmark Suite	Held out internal grounding, parametric, MM, and search retrieval benchmarks		**70.5%**	63.4%	50.4%	50.8%
SimpleQA Verified	Parametric knowledge		**72.1%**	54.5%	29.3%	34.9%
MMMLU	Multilingual Q&A		**91.8%**	89.5%	89.1%	91.0%
Global PIQA	Commonsense reasoning across 100 Languages and Cultures		**93.4%**	91.5%	90.1%	90.9%
MRCR v2 (8-needle)	Long context performance	128k (average) 1M (pointwise)	**77.0%** **26.3%**	58.0% 16.4%	47.1% not supported	61.6% not supported

For details on our evaluation methodology please see deepmind.google/models/evals-methodology/gemini-3-pro