Published: February, 2026

Model Evaluation – Approach, Methodology & Results

Gemini 3.1 Pro

Approach: Gemini 3.1 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 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-3.1-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 Opus 4.6, Sonnet 4.6, and GPT-5.2 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 February, 2026:

- **Reasoning and Academic Knowledge:**
  - *Humanity's Last Exam* results for Gemini 3 Pro are from the ScaleAI leaderboard. Gemini 3.1 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 similar sites.
  - *ARC-AGI-2* results are sourced from the ARC Prize website and are ARC Prize Verified. The set reported is semi-private.
  - *GPQA Diamond* results for Gemini 3.1 Pro are self computed.
- **Code**
  - *Terminal-Bench 2.0* results are self computed for Gemini 3.1 Pro and for other models are reported from the public leaderboard. Results are reported for the default agent harness (Terminus 2) and for other best self-reported harnesses where applicable.
  - *SWE-Bench Pro (Public)* and *SWE-bench Verified* numbers follow official provider methodology, 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 for SWE-Bench Verified and 5x runs for SWE-Bench Pro.
    - For SWE-Bench Verified we discovered bugs with 3 items on the official test harness which make them impossible for any solutions to pass:
      - astropy__astropy-7606: the official dataset contains a nonexistent PASS_TO_PASS test called "astropy/units/tests/test_units.py::test_compose_roundtrip[]. The issues was also discussed at https://github.com/SWE-bench/SWE-bench/issues/223

        - sphinx-doc__sphinx-8595 & sphinx-doc__sphinx-9711 : There is a bug in the official harness which causes the pytest -rA change in tox.ini to get reverted when using the latest v2 official docker images.
          The specific commands that caused the issues are + *git checkout b19bce971e82f2497d67fdacdeca8db08ae0ba56* in sphinx-doc__sphinx-8595's and + *git checkout 81a4fd973d4cfcb25d01a7b0be62cdb28f82406d* in sphinx-doc__sphinx-9711.
      - Gemini-3.1 Pro-Preview passed all three items with fixes for these bugs in our internal implementation, so we adjusted the score for this model to reflect the improved pass rate (an increase of 0.6%).
    - *LiveCodeBench Pro*: We report ELO scores from the public leaderboard for all models.
    - *SciCode* results are sourced from Artificial Analysis.
- **Expert tasks** – *GDPval-AA Elo* results are sourced from the Artificial Analysis public leaderboard.
- **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. The airline version was excluded due to lower quality grading.
    - *MCP Atlas* results are based on the public set and sourced from Turing.
    - *BrowseComp* results for Gemini 3.1 and 3 Pro utilize Deep Research with access to search, python, and browsing. Other model results are sourced from providers' self reported numbers.
- **Image** – *MMMU-Pro* scores are averaged across the Standard (10 options) and Vision settings.
- **Multilinguality** – Multilingual MMLU results for Gemini 3.1 Pro and 3 Pro are self computed.
- **Long Context** – For MRCR v2 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 are also releasing the full dataset for reproducibility in our repository: https://github.com/google-deepmind/eval_hub/tree/master/eval_hub/mrcr_v2

Results: Gemini 3.1 Pro results as of February, 2026 are below:

Benchmark		Gemini 3.1 Pro Thinking (High)	Gemini 3 Pro Thinking (High)	Sonnet 4.6 Thinking (Max)	Opus 4.6 Thinking (Max)	GPT-5.2 Thinking (xhigh)	GPT-5.3-Codex Thinking (xhigh)
Humanity's Last Exam Academic reasoning (full set, text + MM)	No tools Search (blocklist) + Code	**44.4%** 51.4%	37.5% 45.8%	33.2% 49.0%	40.0% **53.1%**	34.5% 45.5%	— —
ARC-AGI-2 Abstract reasoning puzzles	ARC Prize Verified	**77.1%**	31.1%	58.3%	68.8%	52.9%	—
GPQA Diamond Scientific knowledge	No tools	**94.3%**	91.9%	89.9%	91.3%	92.4%	—
Terminal-Bench 2.0 Agentic terminal coding	Terminus-2 harness Other best self-reported harness	**68.5%** —	56.9% —	59.1% —	65.4% —	54.0% 62.2% (Codex)	64.7% **77.3%** (Codex)
SWE-Bench Verified Agentic coding	Single attempt	80.6%	76.2%	79.6%	**80.8%**	80.0%	—
SWE-Bench Pro (Public) Diverse agentic coding tasks	Single attempt	54.2%	43.3%	—	—	55.6%	**56.8%**
LiveCodeBench Pro Competitive coding problems from Codeforces, ICPC, and IOI	Elo	**2887**	2439	—	—	2393	—
SciCode Scientific research coding		**59%**	56%	47%	52%	52%	—
APEX-Agents Long horizon professional tasks		**33.5%**	18.4%	—	29.8%	23.0%	—
GDPval-AA Elo Expert tasks		1317	1195	**1633**	1606	1462	—
τ2-bench Agentic tool use	Retail Telecom	90.8% **99.3%**	85.3% 98.0%	91.7% 97.9%	**91.9%** **99.3%**	82.0% 98.7%	— —
MCP Atlas Multi-step workflows using MCP		**69.2%**	54.1%	61.3%	59.5%	60.6%	—
BrowseComp Agentic search	Search + Python + Browse	**85.9%**	59.2%	74.7%	84.0%	65.8%	—
MMMU Pro Multimodal understanding and reasoning	No tools	80.5%	**81.0%**	74.5%	73.9%	79.5%	—
MMMLU Multilingual Q&A		**92.6%**	91.8%	89.3%	91.1%	89.6%	—
MRCR v2 (8-needle) Long context performance	128k (average) 1M (pointwise)	**84.9%** 26.3%	77.0% 26.3%	**84.9%** Not supported	84.0% Not supported	83.8% Not supported	— —