Google

February 2026

Gemini 3.1 Pro

Model Card

***Model Cards** are intended to provide essential information on Gemini models, including known limitations, mitigation approaches, and safety performance. Model cards may be updated from time-to-time; for example, to include updated evaluations as the model is improved or revised. See the Google DeepMind site for a comprehensive list of model cards.*

Published: February 2026

Model Information

Description: Gemini 3.1 Pro is the next iteration in the Gemini 3 series of models, a suite of highly capable, natively multimodal reasoning models. As of this model card's date of publication, Gemini 3.1 Pro is Google's most advanced model for complex tasks. Gemini 3.1 Pro can comprehend vast datasets and challenging problems from massively multimodal information sources, including text, audio, images, video, and entire code repositories.

Model dependencies: Gemini 3.1 Pro is based on Gemini 3 Pro.

Inputs: Text strings (e.g., a question, a prompt, document(s) to be summarized), images, audio, and video files, with a token context window of up to 1M.

Outputs: Text, with a 64K token output.

Architecture: Gemini 3.1 Pro is based on Gemini 3 Pro. For more information about the model architecture for Gemini 3.1 Pro, see the Gemini 3 Pro model card.

Model Data

Training Dataset: Gemini 3.1 Pro is based on Gemini 3 Pro. For more information about the training dataset for Gemini 3.1 Pro, see the Gemini 3 Pro model card.

Training Data Processing: For more information about the training data processing for Gemini 3.1 Pro, see the Gemini 3 Pro model card.

Implementation and Sustainability

Hardware: Gemini 3.1 Pro is based on Gemini 3 Pro. For more information about the hardware for Gemini 3.1 Pro and our continued commitment to operate sustainably, see the Gemini 3 Pro model card.

Software: Gemini 3.1 Pro is based on Gemini 3 Pro. For more information about the software for Gemini 3.1 Pro, see the Gemini 3 Pro model card.

Distribution

Gemini 3.1 Pro is distributed in the following channels; respective documentation shared in line:

- Gemini App
- Google Cloud / Vertex AI
- Google AI Studio
- Gemini API
- Google Antigravity
- NotebookLM

Our models are available to downstream providers via an application program interface (API) and subject to relevant terms of use. There is no required hardware or software to use the model. For AI Studio and Gemini API, see the Gemini API Additional Terms of Service; for Vertex AI, see Google Cloud Platform Terms of Service. For more information, see Gemini Model API instructions and Gemini API in Vertex AI quickstart.

Evaluation

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. Additional benchmarks and details on approach, results and their methodologies can be found at: deepmind.google/models/evals-methodology/gemini-3-1-pro.

Results: Gemini 3.1 Pro significantly outperforms Gemini 3 Pro across a range of benchmarks requiring enhanced reasoning and multimodal capabilities. Results as of February 2026 are listed 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	— —

Intended Usage and Limitations

Benefit and Intended Usage: Gemini 3.1 Pro is the next iteration in the Gemini 3.0 series of models, a suite of highly intelligent and adaptive models, capable of helping with real-world complexity, solving problems that require enhanced reasoning and intelligence, creativity, strategic planning and making improvements step-by-step. It is particularly well-suited for applications that require:

- agentic performance
- advanced coding
- long context and/or multimodal understanding
- algorithmic development

Known Limitations: For more information about the known limitations for Gemini 3.1 Pro, see the Gemini 3 Pro model card.

Acceptable Usage: For more information about the acceptable usage for Gemini 3.1 Pro, see the Gemini 3 Pro model card.

Ethics and Content Safety

Evaluation Approach: For more information about the evaluation approach for Gemini 3.1 Pro, see the Gemini 3 Pro model card.

Safety Policies: For more information about the safety policies for Gemini 3.1 Pro, see the Gemini 3 Pro model card.

Training and Development Evaluation Results: Results for some of the internal safety evaluations conducted during the development phase are listed below. The evaluation results are for automated evaluations and not human evaluation or red teaming. Scores are provided as an absolute percentage increase or decrease in performance compared to the indicated model, as described below. Overall, Gemini 3.1 Pro outperforms Gemini 3.0 Pro across both safety and tone, while keeping unjustified refusals low. We mark improvements in green and regressions in red. Safety evaluations of Gemini 3.1 Pro produced results consistent with the original Gemini 3.0 Pro safety assessment.

Evaluation[1]	**Description**	Gemini 3.1 Pro vs. Gemini 3.0 Pro
Text to Text Safety	Automated content safety evaluation measuring safety policies	+0.10% (non-egregious)
Multilingual Safety	Automated safety policy evaluation across multiple languages	+0.11% (non-egregious)
Image to Text Safety	Automated content safety evaluation measuring safety policies	-0.33%
Tone[2]	Automated evaluation measuring objective tone of model refusal	+0.02%
Unjustified-refusals	Automated evaluation measuring model's ability to respond to borderline prompts while remaining safe	-0.08%

We continue to improve our internal evaluations, including refining automated evaluations to reduce false positives and negatives, as well as update query sets to ensure balance and maintain a high standard of results. The performance results reported below are computed with improved evaluations and thus are not directly comparable with performance results found in previous Gemini model cards.

We expect variation in our automated safety evaluations results, which is why we review flagged content to check for egregious or dangerous material. Our manual review confirmed losses were

[1]The ordering of evaluations in this table has changed from previous iterations of the 2.5 Flash-Lite model card in order to list safety evaluations together and improve readability. The type of evaluations listed have remained the same.

[2] For tone and instruction following, a positive percentage increase represents an improvement in the tone of the model on sensitive topics and the model's ability to follow instructions while remaining safe compared to Gemini 2.5 Pro. We mark improvements in green and regressions in red.

overwhelmingly either a) false positives or b) not egregious.

Human Red Teaming Results: We conduct manual red teaming by specialist teams who sit outside of the model development team. High-level findings are fed back to the model team. For child safety evaluations, Gemini 3.1 Pro satisfied required launch thresholds, which were developed by expert teams to protect children online and meet Google's commitments to child safety across our models and Google products. For content safety policies generally, including child safety, we saw similar safety performance compared to Gemini 3.0 Pro.

Risks and Mitigations: For more information about the risks and mitigations for Gemini 3.1 Pro, see the Gemini 3 Pro model card.

Frontier Safety

Our Frontier Safety Framework (FSF) includes rigorous evaluations that address risks of severe harm from frontier models, covering five risk domains: CBRN (chemical, biological, radiological and nuclear information risks), cyber, harmful manipulation, machine learning R&D and misalignment.

Our frontier safety strategy is based on a "safety buffer" to prevent models from reaching critical capability levels (CCLs), i.e. if a frontier model does not reach the alert threshold for a CCL, we can assume models developed before the next regular testing interval will not reach that CCL. We conduct continuous testing, evaluating models at a fixed cadence and when a significant capability jump is detected. (Read more about this in our approach to technical AGI safety.)

Following FSF protocols, we conducted a full evaluation of Gemini 3.1 Pro (focusing on Deep Think mode). We found that the model remains below alert thresholds for the CBRN, harmful manipulation, machine learning R&D, and misalignment CCLs. As previous models passed the alert threshold for cyber, we performed more additional testing in this domain on Gemini 3.1 Pro with and without Deep Think mode, and found that the model remains below the cyber CCL.

More details on our evaluations and the mitigations we deploy can be found in the Gemini 3 Pro Frontier Safety Framework Report.

FSF Results as of February, 2026:

Domain	Key Results for Gemini 3.1 Pro	CCL	CCL reached?
CBRN	(Deep Think mode) The model can provide highly accurate and actionable information but still fails to offer novel or sufficiently complete and detailed instructions for critical stages, to significantly enhance the capabilities of low to medium resourced threat actors required for the CCL. We continue to deploy mitigations in this domain.	Uplift Level 1	CCL not reached
Cyber	(3.1 Pro) We conducted additional testing on the model in this domain as Gemini 3 Pro had previously reached the alert threshold. The model shows an increase in cyber capabilities compared to Gemini 3 Pro. As with Gemini 3 Pro, the model has reached the alert threshold, but still does not reach the levels of uplift required for the CCL. (Deep Think mode) Accounting for inference costs, the model with Deep Think mode performs considerably worse than without Deep Think mode. Even at high levels of inference, results for the model with Deep Think mode do not suggest higher capability than without Deep Think mode. We continue to deploy mitigations in this domain.	Uplift Level 1	CCL not reached
Harmful Manipulation	(Deep Think mode) Evaluations indicated that the model showed higher manipulative efficacy for belief change metrics compared to a non-AI baseline, with the maximum odds ratio of 3.6x, which is the same as Gemini 3 Pro, and did not reach the alert threshold.	Level 1 (exploratory)	CCL not reached

Machine Learning R&D	(Deep Think mode) The model shows gains on RE-Bench compared to Gemini 3 Pro, with a human-normalised average score of 1.27 compared to Gemini 3 Pro's score of 1.04. On one particular challenge, *Optimise LLM Foundry*, it scores double the human-normalised baseline score (reducing the runtime of a fine-tuning script from 300 seconds to 47 seconds, compared to the human reference solution of 94 seconds). However, the model's average performance across all challenges remains beneath the alert threshold for the CCLs.	Acceleration level 1 Automation level 1	CCLs not reached
Misalignment (Exploratory)	(Deep Think mode) On stealth evaluations, the model performs similarly to Gemini 3 Pro. On situational awareness, the model is stronger than Gemini 3 Pro: on three challenges which no other model has been able to consistently solve, *max tokens, context size mod,* and *oversight frequency,* the model achieves a success rate of almost 100%. However, its performance on other challenges is inconsistent, and thus the model does not reach the alert threshold.	Instrumental Reasoning Levels 1 + 2 (exploratory)	CCLs not reached