Google

Model card published: December, 2025

Gemini 3 Flash

Model Card

Gemini 3 Flash - 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: December 2025

Model Information

Description: Gemini 3 Flash is the next iteration in the Gemini 3 series of highly-capable, natively multimodal, reasoning models. Gemini 3 Flash is built off of the Gemini 3 Pro reasoning foundation with thinking levels to control the mix of quality, cost and latency.

Model dependencies: Gemini 3 Flash 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 Flash is based on Gemini 3 Pro. For more information about the model architecture for Gemini 3 Pro, see the Gemini 3 Pro model card.

Model Data

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

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

Implementation and Sustainability

Hardware: Gemini 3 Flash was trained using Google's Tensor Processing Units (TPUs). TPUs are specifically designed to handle the massive computations involved in training LLMs and can speed up training considerably compared to CPUs. TPUs often come with large amounts of high-bandwidth memory, allowing for the handling of large models and batch sizes during training, which can lead to better model quality. TPU Pods (large clusters of TPUs) also provide a scalable solution for handling the growing complexity of large foundation models. Training can be distributed across multiple TPU devices for faster and more efficient processing.

The efficiencies gained through the use of TPUs are aligned with Google's commitment to operate sustainably.

Software: Training was done using JAX and ML Pathways.

Distribution

Gemini 3 Flash is distributed similarly to Gemini 3 Pro. For more information about the distribution of Gemini 3 Flash, see the Gemini 3 Pro model card.

Evaluation

Approach: Gemini 3 Flash 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.com/models/evals-methodology/gemini-3-flash.

Results: Gemini 3 Flash significantly outperforms Gemini 2.5 Pro across a range of benchmarks requiring enhanced reasoning and multimodal capabilities. Results as of December, 2025 are listed 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%

Intended Usage and Limitations

Benefit and Intended Usage: Gemini 3 Flash is well-suited for users and developers, specific use cases include: agentic workflows, every day coding, reasoning and planning, and multimodal analysis.

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

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

Ethics and Content Safety

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

Safety Policies: For more information about the safety policies for Gemini 3 Flash, 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 Flash outperforms Gemini 2.5 Flash across both safety and tone, while keeping unjustified refusals low. We mark improvements in green and regressions in red.

Evaluation Description	**Gemini 3 Flash** vs. Gemini 2.5 Flash
Text to Text Safety Automated content safety evaluation measuring safety policies	-3.1%
Multilingual Safety Automated safety policy evaluation across multiple languages	+0.1% non-egregious
Image to Text Safety Automated content safety evaluation measuring safety policies	-2.3%
Tone Automated evaluation measuring objective tone of model refusal	+3.8%
Unjustified-refusals Automated evaluation measuring model's ability to respond to borderline prompts while remaining safe	-10.4%

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 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 Flash 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 or improved safety performance compared to Gemini 2.5 Flash. Like 3 Pro, the scope of red teaming covered potential issues outside of our strict policies, and found no egregious concerns.

Frontier Safety Assessment: We evaluated Gemini 3 Pro Preview for Frontier Safety and reported the results in the Gemini 3 Pro Frontier Safety Framework Report, finding that it did not reach any critical capability levels (CCLs) outlined in our Frontier Safety Framework. As Gemini 3 Flash is less capable than Gemini 3 Pro, and the Gemini 3 Pro model results give us confidence that Gemini 3 Flash is unlikely to reach any CCLs, we can rely on results reported for Gemini 3 Pro. Therefore, in line with the risk acceptance criteria outlined in our FSF (and our general responsibility and safety practices), we deemed Gemini 3 Flash was acceptable for deployment.

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