Google

Gemini 3.1 Flash Lite

Model Card

Gemini 3.1 Flash-Lite - 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.*

Updated: May 2026

Model Information

Description: Gemini 3.1 Flash-Lite is an addition to the Gemini 3 series of highly-capable, natively multimodal, reasoning models. The model is cost-efficient and fast, optimized for high-volume, latency-sensitive tasks like translation and classification.

Model dependencies: Gemini 3.1 Flash-Lite 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 Flash-Lite is based on Gemini 3 Pro. For more information about the model architecture for Gemini 3.1 Flash-Lite, see the Gemini 3 Pro model card.

Model Data

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

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

Implementation and Sustainability

Hardware: Gemini 3.1 Flash-Lite 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 3.1 Flash-Lite is distributed in the following channels; respective documentation shared in line:

- Google Cloud / Vertex AI
- Google AI Studio
- Gemini API
- Gemini App
- Google Search AI Overviews

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 Flash-Lite was evaluated across a range of benchmarks, including speed, reasoning, multimodal capabilities, factuality, agentic tool use, multi-lingual performance, coding, and long-context. Benchmark details on approach, results, and their methodologies can be found at: https://deepmind.google/models/evals-methodology/gemini-3-1-flash-lite

Results: Gemini 3.1 Flash-Lite results as of March, 2026 are below:

Benchmark		Gemini 3.1 Flash-Lite High	Gemini 2.5 Flash Dynamic	Gemini 2.5 Flash-Lite Dynamic	GPT-5 mini High	Claude 4.5 Haiku Extended Thinking	Grok 4.1 Fast Reasoning
Input price $/1M tokens, no caching	Lower is better	$0.25	$0.30	$0.10	$0.25	$1.00	$0.20
Output price $/1M tokens	Lower is better	$1.50	$2.50	$0.40	$2.00	$5.00	$0.50
Output speed Tokens/s		363	249	366	71	108	145
Humanity's Last Exam Academic reasoning (full set, text + MM)	No tools	16.0%	11.0%	6.9%	16.7%	9.7%	**17.6%**
GPQA Diamond Scientific knowledge	No tools	**86.9%**	82.8%	66.7%	82.3%	73.0%	84.3%
MMMU-Pro Multimodal understanding and reasoning	No tools	**76.8%**	66.7%	51.0%	74.1%	58.0%	63.0%
CharXiv Reasoning Information synthesis from complex charts		73.2%	63.7%	55.5%	**75.5%** (+ python)	61.7%	31.6%
Video-MMMU Knowledge acquisition from videos		**84.8%**	79.2%	60.7%	82.5%	—	74.6%
SimpleQA Verified Parametric knowledge		**43.3%**	28.1%	11.5%	9.5%	5.5%	19.5%
FACTS Benchmark Suite Factuality benchmark across grounding, parametric, search, and MM.		40.6%	**50.4%**	17.9%	33.7%	18.6%	42.1%
MMMLU Multilingual Q&A		**88.9%**	86.6%	84.5%	84.9%	83.0%	86.8%
LiveCodeBench Code generation (UI: 1/1/2025-5/1/2025)		72.0%	62.6%	34.3%	**80.4%**	53.2%	76.5%
MRCR v2 (8-needle) Long context performance	128k (average)	**60.1%**	54.3%	30.6%	52.5%	35.3%	54.6%
	1M (pointwise)	12.3%	**21.0%**	5.4%	Not Supported	Not Supported	6.1%

Intended Usage and Limitations

Benefit and Intended Usage: Gemini 3.1 Flash-Lite is well suited for applications that require high volume, cost-efficient and low latency tasks.

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

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

Ethics and Content Safety

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

Safety Policies: For more information about the safety policies for Gemini 3.1 Flash-Lite, 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 Flash-Lite outperforms Gemini 2.5 Flash-Lite across both safety and tone, while keeping unjustified refusals low. We mark improvements in green and regressions in red.

Evaluation[1]	Description	Gemini 3.1 Flash-Lite vs. Gemini 2.5 Flash-Lite
Text to Text Safety	Automated content safety evaluation measuring safety policies	-1.18%
Multilingual Safety	Automated safety policy evaluation across multiple languages	-1.84%
Image to Text Safety	Automated content safety evaluation measuring safety policies	-21.7%
Tone[2]	Automated evaluation measuring objective tone of model refusal	+14.59%
Unjustified-refusals	Automated evaluation measuring model's ability to respond to borderline prompts while remaining safe	-14.41%

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 3.1 Flash-Lite 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.

[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.

Frontier Safety Assessment: Gemini 3.1 Flash-Lite is part of the Gemini 3 family of models. We rely on our evaluation of Gemini 3.1 Pro with Deep Think mode for Frontier Safety as it is the most generally capable model as of publication of this model card, and it did not reach any Critical Capability Levels (CCLs) outlined in our Frontier Safety Framework. Our assessments have shown that Gemini 3.1 Flash-Lite is less capable than Gemini 3.1 Pro, therefore based on Gemini 3.1 Pro, we are confident that Gemini 3.1 Flash-Lite is also unlikely to reach any CCLs. For more information, read the Gemini 3.1 Pro Model Card.

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