Google

Model card published/updated SEP 26, 2025

Gemini 2.5 Flash-Lite

Model Card

Gemini 2.5 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.*

Technical Reports *are similar to academic papers, and describe models' capabilities, limitations and performance benchmarks. The Gemini 2.5 technical report contains additional details about the Gemini 2.5 series of models. We recommend that readers seeking more details and information about these models navigate to the technical report.*

Published: September 26, 2025

Model Information

Description: Gemini 2.5 Flash-Lite is an addition to our hybrid reasoning model family, giving developers the ability to turn a model's thinking on or off. The model is cost-efficient and fast, optimized for high-volume, latency-sensitive tasks like translation and classification. This model offers improved performance compared to 2.0 Flash-Lite, with strong results in coding, math, science, and reasoning benchmarks.

As of September 25, 2025, there is a preview version offered alongside 2.5 Flash-Lite, specified as "Gemini 2.5 Flash-Lite Preview (09-2025)", which is designed to address developer feedback, with key quality improvements, including better instruction following, reduced verbosity, and stronger multimodal and translation capabilities.

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

Outputs: Text, with a 64K token output.

Architecture: The Gemini 2.5 models are sparse mixture-of-experts (MoE) (Clark et al., 2022; Du et al., 2021; Fedus et al., 2021; Jiang et al., 2024, Lepikhin et al., 2020; Riquelme et al., 2021; Roller et al., 2021; Shazeer et al., 2017; transformers (Vaswani et al., 2017) with native multimodal support for text, vision, and audio inputs. Sparse MoE models activate a subset of model parameters per input token by learning to dynamically route tokens to a subset of parameters (experts); this allows them to decouple total model capacity from computation and serving cost per token. Developments to the model architecture contribute to the significantly improved performance of Gemini 2.5 compared to Gemini 1.5 Pro (see Section 3 of the Gemini Technical Report).

Model Data

Training Dataset: The pre-training dataset was a large-scale, diverse collection of data encompassing a wide range of domains and modalities, which included publicly-available web-documents, code (various programming languages), images, audio (including speech and other audio types) and video. The post-training dataset consisted of vetted instruction tuning data and was a collection of multimodal data with paired instructions and responses in addition to human preference and tool-use data.

Training Data Processing: Data filtering and preprocessing included techniques such as deduplication, safety filtering in-line with Google's commitment to advancing AI safely and responsibly and quality filtering to mitigate risks and improve training data reliability.

Implementation and Sustainability

Hardware: Gemini 2.5 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.

Evaluation

Approach: Gemini 2.5 Flash-Lite was evaluated using the methodology below:

- **Gemini results**: All Gemini scores are pass @1."Single attempt" settings allow no majority voting or parallel test-time compute; "multiple attempts" settings allow test-time selection of the candidate answer. They are all run with the AI Studio API with default sampling settings. To reduce variance, we average over multiple trials for smaller benchmarks. Aider Polyglot score is the pass rate average of 3 trials. Vibe-Eval results are reported using Gemini as a judge. Google's scaffolding for "multiple attempts" for SWE-Bench includes drawing multiple trajectories and re-scoring them using model's own judgement. For Aider results differ from the official leaderboard due to a difference in the settings used for evaluation (non-default).

- **Result sources**: Where provider numbers are not available we report numbers from leaderboards reporting results on these benchmarks: Humanity's Last Exam results are sourced from https://agi.safe.ai/ and https://scale.com/leaderboard/humanitys_last_exam, LiveCodeBench results are from https://livecodebench.github.io/leaderboard.html (1/1/2025 - 5/1/2025 in the UI), Aider Polyglot numbers come from https://aider.chat/docs/leaderboards/. FACTS come from https://www.kaggle.com/benchmarks/google/facts-grounding. For MRCR v2 which is not publically 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. The methodology has changed in this table vs previously published results for MRCR v2 as we have decided to focus on a harder, 8-needle version of the benchmark going forward. For Flash-Lite Preview (09-2025) we start measuring SimpleQA Verified instead of SimpleQA for a higher eval signal. See https://www.kaggle.com/benchmarks/deepmind/simpleqa-verified for more details.

* these results are on an earlier HLE dataset, obtained from https://scale.com/leaderboard/humanitys_last_exam_preview
.

Results: 2.5 Flash-Lite has all-round, significantly higher performance than 2.0 Flash-Lite on coding, math, science, reasoning and multimodal benchmarks. Detailed results as of June 2025 are listed below, and updated as of September 2025 to include 2.5 Flash-Lite Preview (09-2025) benchmarks.

Capability **Benchmark**[1]		**Gemini 2.5 Flash-Lite Preview** Non-thinking (09-2025)	**Gemini 2.5 Flash-Lite Preview** Thinking (09-2025)	**Gemini 2.5 Flash-Lite** Non-thinking (06-17)	**Gemini 2.5 Flash-Lite** Thinking (06-17)	**Gemini 2.0 Flash**
Reasoning & knowledge **Humanity's Last Exam (no tools)**		6.4%	7.3%	5.1%	6.9%	5.1%*
Science **GPQA diamond**		70.2%	71.7%	64.6%	66.7%	65.2%
Mathematics **AIME 2025**		50.1%	48.2%	49.8%	63.1%	29.7%
Code generation **LiveCodeBench v5** (UI: 1/1/2025-5/1/2025)		52.1%	58.4%	33.7%	34.3%	29.1%
Code editing **Aider Polyglot**		—	—	26.7% whole	27.1% whole	21.3% whole
Agentic Coding **SWE-Bench Verified**	single attempt	41.3%	38.9%	31.6%	27.6%	21.4%
	multiple attempts	—	—	42.6%	44.9%	34.2%
Factuality **SimpleQA**		—	—	10.7%	13.0%	29.9%
Factuality **SimpleQA Verified**		11.3%	9.6%	—	—	—
Factuality **FACTS Grounding**		86.9%	87.5%	84.1%	86.8%	84.6%
Visual reasoning **MMMU**	single attempt	74.0%	72.0%	72.9%	72.9%	69.3%
Image understanding **Vibe-Eval (Reka)**		58.4%	59.8%	51.3%	57.5%	56.4%
Long context **MRCR v2 (8-needle)**	128k (average)	12.0%	25.6%	16.6%	30.6^	19%
	1M (pointwise)	6.5%	7.7%	4.1%	5.40%	5.3%
Multilingual performance **Global MMLU (Lite)**		82.9%	84.9%	81.1%	84.5%	83.4%

* indicates evaluated on text problems only (without images)

[1]We regularly update evaluation processes to include new and emerging quality evaluations and benchmarks. The results reported above include additional or updated benchmarks which may not have been included in previous Gemini model cards. Results are thus not directly comparable with performance results found in previous Gemini model cards.

Intended Usage and Limitations

Benefit and Intended Usage: Gemini 2.5 Flash-Lite is well suited for applications that require high volume, low-cost and low latency tasks. Gemini 2.5 Flash-Lite Preview (09-2025) provides additional developer-focused updates, with quality improvements in complex instruction following, producing more concise answers, and better translation quality, image understanding, and audio transcription.

Known Limitations: Gemini 2.5 Flash-Lite may exhibit some of the general limitations of foundation models, such as hallucinations, and limitations around causal understanding, complex logical deduction, and counterfactual reasoning. Adherence to thinking budgets may not be consistent. The knowledge cutoff date for Gemini 2.5 Flash-Lite was January 2025. See the Ethics and Safety section below for additional information on known limitations.

Ethics and Safety

Evaluation Approach: Gemini 2.5 Flash-Lite was developed in partnership with internal safety, security, and responsibility teams. A range of evaluations and red teaming activities were conducted to help improve the model and inform decision-making. These evaluations and activities align with Google's AI Principles and responsible AI approach.

Evaluation types included but were not limited to:

- **Training/Development Evaluations** including automated and human evaluations carried out continuously throughout and after the model's training, to monitor its progress and performance;
- **Human Red Teaming** conducted by specialist teams across the policies and desiderata, deliberately trying to spot weaknesses and ensure the model adheres to safety policies and desired outcomes;
- **Automated Red Teaming** to dynamically evaluate Gemini for safety and security considerations at scale, complementing human red teaming and static evaluations;
- **Assurance Evaluations** conducted by human evaluators independent of the model development team, and assess responsibility and safety governance decisions
- **Ethics & Safety Reviews** were conducted ahead of the model's release

In addition, we perform testing following the guidelines in Google DeepMind's Frontier Safety Framework (FSF).

Safety Policies: Gemini safety policies align with Google's standard framework for the types of harmful content that we make best efforts to prevent our Generative AI models from generating, including the following types of harmful content:

1. Child sexual abuse and exploitation
2. Hate speech (e.g., dehumanizing members of protected groups)
3. Dangerous content (e.g., promoting suicide, or instructing in activities that could cause real-world harm)
4. Harassment (e.g., encouraging violence against people)
5. Sexually explicit content
6. Medical advice that runs contrary to scientific or medical consensus

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, and scores are provided as an absolute percentage increase or decrease in performance in comparison to the indicated model, as described below.

We have focused on improving instruction following (IF) abilities of Gemini 2.5. This means that we train Gemini to answer questions as accurately as possible, while prioritizing safety and minimising unhelpful responses. New models are more willing to engage with prompts that previous models may have incorrectly refused.

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. 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. In addition to continuing to improve our evaluations, we also leverage expert red teamers to assess the safety profile of our models (see below section).

For safety evaluations, a decrease in percentage represents a reduction in violation rates compared to Gemini 2.0 Flash-Lite and an increase in percentage represents an increase in violation rates. 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 2.0 Flash-Lite. We mark improvements in green and regressions in red.

Evaluation[2]	Description	**Gemini 2.5 Flash-Lite Preview** Non-thinking (09-2025) vs. Gemini 2.0 Flash-Lite	**Gemini 2.5 Flash-Lite Preview** Thinking (09-2025) vs. Gemini 2.0 Flash-Lite	**Gemini 2.5 Flash-Lite** Non-thinking (06-17) vs. Gemini 2.0 Flash-Lite	**Gemini 2.5 Flash-Lite** Thinking (06-17) vs. Gemini 2.0 Flash-Lite
Text to Text Safety	Automated content safety evaluation measuring safety policies	+3.7% (non egregious)	+0.1% (non egregious)	+5.7% (non egregious)	+4.9% (non egregious)
Multilingual Safety	Automated safety policy evaluation across multiple languages	+4.9% (non egregious)	+2.0% (non egregious)	+3.5% (non egregious)	+2.9% (non egregious)
Image to Text Safety	Automated content safety evaluation measuring safety policies	+1.2% (non egregious)	-1.3%	-4.4%	-0.5%
Tone	Automated evaluation measuring objective tone of model refusal	+2.4%	-1.4%	-9.9%	-13.2%
Instruction Following	Automated evaluation measuring model's ability to follow instructions while remaining safe	+37.7%	+34.3%	+31.5%	+29.8%

Assurance Evaluations Results: We conduct baseline assurance evaluations to guide decisions on model releases. These evaluations look at model behavior, including within the context of the safety policies and modality-specific risk areas. High-level findings are fed back to the model team. For content safety policies, including child safety, we saw similar or improved safety performance compared to Gemini 2.0 Flash-Lite.

Frontier Safety Assessment: We evaluated Gemini 2.5 Pro Preview for Frontier Safety and reported the results in the 2.5 Pro Preview model card, finding that it did not reach any critical capability levels outlined in our Frontier Safety Framework. As Gemini 2.5 Flash-Lite is less capable than Gemini 2.5 Pro Preview, and the Gemini 2.5 Pro model results give us confidence that Gemini 2.5 Flash-Lite is unlikely to reach critical capability levels, we are using Frontier Safety evaluations reported for Gemini 2.5 Pro Preview.

Known Safety Limitations: The main safety limitations for Gemini 2.5 Flash-Lite are related to tone. The model will sometimes respond in a way which can come across as "preachy". However, Gemini 2.5 Flash-Lite still has measurable improvements in tone over previous Flash-Lite models.

[2]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.

Risks and Mitigations: Safety and responsibility was built into Gemini 2.5 Flash-Lite throughout the training and deployment lifecycle, including pre-training, post-training, and product-level mitigations. Mitigations include, but are not limited to:

- dataset filtering;
- conditional pre-training;
- supervised fine-tuning;
- reinforcement learning from human and critic feedback;
- safety policies and desiderata;
- product-level mitigations such as safety filtering.
