Model card published April 10, 2025

Gemini 2.0 Flash-Lite
Model Card

Model Cards are intended to provide essential information on Gemini models, including known limitations, mitigation approaches, and safety performance. These cards are accompanied by a detailed technical report that will be published once per model family's release, and additional reports focused on dangerous capability evaluations that will be published at regular cadences.

Model Information

Description: Gemini 2.0 Flash-Lite is a member of the Gemini 2.0 series of models, a suite of highly-capable, natively multimodal models designed to power a new era of agentic systems. Gemini 2.0 Flash-Lite is Google's most cost-efficient model, striking a balance between efficiency and quality targeting low-cost workflows.

Inputs: Text strings (e.g., a question, a prompt, a document(s) to be summarized), images, audio, and video files, with a 1,048,576 token context window.

Outputs: Text, with an 8,192 token output.

Architecture: The Gemini 2.0 series builds upon the sparse Mixture-of-Experts (MoE) Transformer architecture (Clark et al., 2020; Fedus et al., 2021; Lepikhin et al., 2020; Riquelme et al., 2021; Shazeer et al., 2017; Zoph et al., 2022) used in Gemini 1.5. Key enhancements in Gemini 2.0 include refined architectural design and novel optimization methods, leading to substantial improvements in training stability and computational efficiency. Each model within the 2.0 family, including Gemini 2.0 Flash-Lite, is carefully designed and calibrated to achieve an optimal balance between quality and performance for their specific downstream applications.

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 2.0 Flash-Lite leverages Trillium, the sixth-generation of Google's Tensor Processing Units, for both training and inference. Trillium provides significant improvements in training performance, inference throughput, and energy efficiency, and is capable of scaling a single distributed training job to hundreds of thousands of accelerators across multiple datacenters.

More efficient TPU hardware design led to an improvement in the carbon-efficiency of AI workloads that was three times greater over two generations — from TPUv4 to Trillium. The efficiencies gained through the use of Trillium TPUs are aligned with Google's commitment to operate sustainably.

Software: Training was done using JAX and ML Pathways.

Evaluation

Approach: Gemini 2.0 Flash-Lite was evaluated against performance benchmarks below.

Results: Gemini 2.0 Flash-Lite performs better than Gemini 1.5 Flash on the majority of benchmarks, at the same speed and cost. Detailed results are listed below:

Capability	Benchmark	Description	**Gemini 1.5 Flash**	**Gemini 1.5 Pro**	**Gemini 2.0 Flash-Lite** (Public Preview)	**Gemini 2.0 Flash** (GA)
General	MMLU-Pro	Enhanced version of popular MMLU dataset with questions across multiple subjects with higher difficulty tasks	67.3%	75.8%	71.6%	**77.6%**
Code	LiveCodeBench (v5)	Code generation in Python. Subset covering more recent examples (in the UI: 10/01/2024 -02/01/2025)	30.7%	34.2%	28.9%	**34.5%**
	Bird-SQL (Dev)	Benchmark evaluating converting natural language questions into executable SQL	45.6%	54.4%	57.4%	**58.7%**
Reasoning	GPQA (diamond)	Challenging dataset of questions written by domain experts in biology, physics, and chemistry	51.0%	59.1%	51.5%	**60.1%**
Factuality	SimpleQA	World knowledge factuality with no search enabled	8.6%	24.9%	21.7%	**29.9%**
	FACTS Grounding	Ability to provide factually correct responses given documents and diverse user requests	82.9%	80.0%	83.6%	**84.6%**
Multilingual	Global MMLU (Lite)	MMLU translated by human translators into 15 languages. The Lite version includes 200 Culturally Sensitive and 200 Culturally Agnostic samples per language	73.7%	80.8%	78.2%	**83.4%**
Math	MATH	Challenging math problems (incl. algebra, geometry, pre-calculus, and others)	77.9%	86.5%	86.8%	**90.9%**
	HiddenMath	Competition-level math problems, held out dataset AIME/AMC-like, crafted by experts and not leaked on the web	47.2%	52.0%	55.3%	**63.5%**
Long-context	MRCR (1M)	Novel, diagnostic long-context understanding evaluation	71.9%	**82.6%**	58.0%	70.5%
Image	MMMU	Multi-discipline college-level multimodal understanding and reasoning problems	62.3%	65.9%	68.0%	**71.7%**
Audio	CoVoST2 (21 lang)	Automatic speech translation (BLEU score)	37.4	**40.1**	38.4	39.0
Video	EgoSchema (test)	Video analysis across multiple domains	66.8%	**71.2%**	67.2%	71.1%

Intended Usage and Limitations

Benefit and Intended Usage: Gemini 2.0 Flash-Lite models offer enhanced multimodal understanding, enabling reasoning across images, video, audio, and text. Gemini 2.0 Flash-Lite can be used for applications that require operational efficiency on devices with limited computational power, including applications such as wide-ranging language tasks such as text generation, summarization, translation, and question answering. Gemini 2.0 Flash-Lite is the fastest and most cost-efficient Gemini Flash model, and is well suited for low cost workflows.

Known Limitations: Gemini 2.0 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. Gemini 2.0 Flash-Lite does not include all of the same features as Gemini 2.0 Flash, including multimodal API support and others. More information is available on AI Studio Gemini API documentation and Vertex AI documentation. The knowledge cutoff date for Gemini 2.0 Flash-Lite was June 2024. See the Ethics and Safety Section for additional information on known limitations.

Ethics and Safety

Evaluation Approach: The development of Gemini 2.0 models was driven in partnership with internal safety, security, and responsibility teams. A range of evaluations and red teaming activities were held prior to release to improve models 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:** automated evaluations completed throughout and after model training;
- **Human red teaming** conducted by specialist teams across the policies and desiderata;
- **Automated red teaming** to dynamically evaluate Gemini at scale, complementing human efforts and static evaluations for both security and safety-focused evaluations;
- **Assurance Evaluations** conducted by evaluators who sit outside of the model development team, used to assess responsibility and safety governance decisions;
- **Frontier Safety Framework** evaluations according to Google DeepMind's Frontier Safety Framework (FSF);
- **Google DeepMind Responsibility and Safety Council (RSC),** Google DeepMind's governance body, reviewed the initial ethics and safety assessments on novel model capabilities in order to provide feedback and guidance during model development. The

RSC also reviewed data on the model's performance via assurance evaluations and made release decisions.

Training and Development Evaluation Results: Results for some of the internal safety evaluations conducted during the training and 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 Gemini 1.5 Pro 002. For safety evaluations, a decrease in percentage represents a reduction in violation rates compared to Gemini 1.5 Pro 002, while for tone a positive percentage increase is representative of an improvement in the tone of model refusal, for instruction following a decrease in percentage is representative of a slight decline in performance compared to Gemini 1.5 Pro 002.

Evaluation	**Description**	**Gemini 2.0 Flash-Lite** (in comparison to Gemini 1.5 Pro 002)
Text to Text Safety	Automated content safety evaluation measuring safety policies	-1.40%
Multilingual Safety	Automated safety policy evaluation across multiple languages and safety policies	-2.0%
Tone	Automated evaluation measuring objective tone of model refusal	+4.40%
Instruction Following	Automated evaluation measuring model's ability to follow instructions while remaining safe	-1.10%
Image to Text Safety	Automated content safety evaluation measuring safety policies	+2.30%

Assurance Evaluations Results: Our baseline assurance evaluations are conducted for model release decision-making for all models. They look at model behavior, including within the context of Google's content policies and modality-specific risk areas. High level findings are fed back to the model team, but prompt sets are held-out to prevent overfitting and preserve the results' ability to inform decision making.

For content policies, we see the Gemini 2.0 family of models displaying lower violation rates in most modalities than Gemini 1.5 Pro, which in turn was a significant improvement on Gemini 1.0. They tended to demonstrate a small regression on our content policy evaluation for image-to-text, though the overall violation rates remained low.

Known Safety Limitations: The main safety limitations for Gemini 2.0 Flash-Lite are over-refusals and tone. The model will sometimes refuse to answer on prompts where an answer would not violate policies (e.g. “Do I sound Italian?”). Refusals can still come across as “preachy,” although tone has improved compared to Gemini 1.5.

Risks and Mitigations: Safety and responsibility was built into Gemini 2.0 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.
