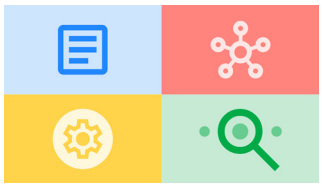

Gemini 2.5 Flash Preview Model Card

Model Cards are intended to provide essential information on Gemini models, including known limitations, mitigation approaches, and safety performance. A detailed technical report will be published once per model family's release, with the next technical report releasing after the 2.5 series is made generally available. Model cards may be updated from time-to-time; for example, to include updated evaluations as the model is improved or revised.

Update: June 6, 2025

Model Information

Description: Gemini 2.5 Flash Preview is the next iteration in the Gemini 2.0 series of models, a suite of highly-capable, natively multimodal, reasoning models. Gemini 2.5 Flash Preview is Google's first fully hybrid reasoning model, giving developers the ability to turn a model's thinking on or off. The model also allows developers to set thinking budgets to find the right tradeoff between quality, cost, and latency. This model card has been updated to contain information for Gemini 2.5 Flash Preview (04-17) and Gemini 2.5 Flash Preview (05-20)[1].

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: Gemini 2.5 Flash Preview 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 2.0 and 1.5.

[1] We've updated the naming convention throughout this model card to clearly differentiate two Gemini 2.5 Flash Preview versions. The model previously identified as "Gemini 2.5 Flash Preview" is referred to as "Gemini 2.5 Flash Preview (04-17)". The latest version is referred to as "Gemini 2.5 Flash Preview (05-20)". Where no distinction is needed, both models are collectively referred to as Gemini 2.5 Flash Preview.

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 Preview 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 Preview was evaluated using the methodology below:

- **Gemini results:** All Gemini 2.5 Flash scores are pass @1 (no majority voting or parallel test time compute unless indicated otherwise). They were all run with the AI Studio API for the model-id gemini-2.5-flash-preview-05-20, model-id gemini-2.5-flash-preview-04-17 and gemini-2.0-flash with default sampling settings. To reduce variance, we averaged over multiple trials for smaller benchmarks. Vibe-Eval results were reported using Gemini as a judge.

- **Non-Gemini results:** All the results for non-Gemini models were sourced from providers' self-reported numbers unless mentioned otherwise below. All SWE-bench Verified numbers follow official provider reports, using different scaffoldings and infrastructure. Google's scaffolding includes drawing multiple trajectories and re-scoring them using the model's own judgement.

- **Thinking vs not-thinking**: For Claude 3.7 Sonnet: GPQA, AIME 2024, MMMU came with 64k extended thinking, Aider with 32k, and HLE with 16k. Remaining results came from the non thinking model due to result availability. For Grok-3, all results came with extended reasoning except for SimpleQA (based on xAI reports) and Aider.

- **Single attempt vs multiple attempts**: When two numbers were reported for the same evaluation, the higher number used majority voting with n=64 for Grok models and internal scoring with parallel test time compute for Anthropic models.

- **Results sources:** Where provider numbers were not available we reported numbers from leaderboards reporting results on these benchmarks: Humanity's Last Exam results were sourced from here and here, AIME 2025 numbers, LiveCodeBench results (10/1/2024 - 2/1/2025 in the UI), Aider Polyglot numbers and FACTS. For MRCR v2 we included 128k results as a cumulative score to ensure they can be comparable with previous results and a pointwise value for 1M context window to show the capability of the model at full length.

Results: Gemini 2.5 Flash Preview demonstrated strong performance across a range of benchmarks. Detailed results as of May 2025 are listed below:

Capability **Benchmark**[2]		**Gemini 2.5 Flash Preview** (05-20)	**Gemini 2.5 Flash Preview** (04-17) Thinking	**Gemini 2.0 Flash** Non-thinking	**OpenAI o4-mini**	**Claude 3.7 Sonnet** 64k Ext. thinking	**Grok 3 Beta** Ext. thinking	**DeepSeek R1**
Reasoning & knowledge **Humanity's Last Exam (no tools)**		11.0%	12.1%	5.1%	14.3%	8.9%	—	8.6%*
Science **GPQA diamond**	single attempt (pass@1)	82.8%	78.3%	60.1%	81.4%	78.2%	80.2%	71.5%
	multiple attempts		—	—	—	84.8%	84.6%	—
Mathematics **AIME 2025**	single attempt (pass@1)	72.0%	78.0%	27.5%	92.7%	49.5%	77.3%	70.0%
	multiple attempts	—	—	—	—	—	93.3%	—
Code generation **LiveCodeBench v5**	single attempt (pass@1)	63.9%	63.5%	34.5%	—	—	70.6%	64.3%
	multiple attempts	—	—	—	—	—	79.4%	—
Code editing **Aider Polyglot**		61.9% / 56.7% whole / diff	51.1% / 44.2% whole / diff	22.2% whole	68.9% / 58.2% whole / diff	64.9% diff	53.3% diff	56.9% diff
Agentic Coding **SWE-Bench Verified**		60.4%	—	—	68.1%	70.3%	—	49.2%
Factuality **SimpleQA**		26.9%	29.7%	29.9%	—	—	43.6%	30.1%
Factuality **FACTS Grounding**		85.3%	—	84.6%	62.1%	78.8%	74.8%	56.8%
Visual reasoning **MMMU**	single attempt (pass@1)	79.7%	76.7%	71.7%	81.6%	75.0%	76.0%	no MM support
	multiple attempts	—	—	—	—	—	78.0%	no MM support
Image understanding **Vibe-Eval (Reka)**		65.4%	62.0%	56.4%	—	—	—	no MM support
Long context **MRCR v2**	128k (average)	74%	—	36%	49%	—	54%	45%
	1M (pointwise)	32%	—	6%	—	—	—	—
Multilingual performance **Global MMLU (Lite)**		88.4%	88.4%	83.4%	—	—	—	—

* indicates evaluated on text problems only (without images)

[2]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 Preview is well suited for applications that require:

- cost-efficient thinking;
- well-rounded capabilities.

Known Limitations: Gemini 2.5 Flash Preview 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 Preview 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 Preview 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 evaluators who sit outside of the model development team, used to independently assess responsibility and safety governance decisions;
- **Google DeepMind Responsibility and Safety Council (RSC)**, Google DeepMind's internal 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.

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 compared to Gemini 2.0 Flash.

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, and this nuance can impact our automated safety scores.

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. Furthermore, this review confirmed losses are narrowly concentrated around **explicit** requests to produce sexually suggestive content or hateful content, mostly in the context of creative use-cases (e.g. historical fiction). We have not observed increased violations outside these specific contexts.

While we prioritize sharing this information promptly, our approach to capturing key insights is an ongoing process. 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 run Assurance Evaluations which are independent evaluations 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 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. We mark improvements in green and regressions in red.

Evaluation[3]	Description	**Gemini 2.5 Flash Preview (05-20)** (in comparison to Gemini 2.0 Flash)	**Gemini 2.5 Flash Preview(04-17)** (in comparison to Gemini 2.0 Flash)
Text to Text Safety	Automated content safety evaluation measuring safety policies	+8.7% (non egregious)	+4.1% (non egregious)
Multilingual Safety	Automated safety policy evaluation across multiple languages	+1.55%	-0.49%
Image to Text Safety	Automated content safety evaluation measuring safety policies	+12.9% (non egregious)	+9.60% (non egregious)
Tone	Automated evaluation measuring objective tone of model refusal	+12.7%	+10.10%
Instruction Following	Automated evaluation measuring model's ability to follow instructions while remaining safe	+33.5%	+30.10%

Assurance Evaluations Results: We conduct baseline assurance evaluations to guide decisions on model releases. These standard safety tests 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, but prompt sets are held out to prevent overfitting and preserve the results' ability to inform decision making. For content safety policies, including child safety, we saw similar safety performance compared to Gemini 2.0 Flash.

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 Preview 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 is unlikely to reach critical capability levels, we can rely on Frontier Safety evaluations reported for Gemini 2.5 Pro Preview.

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

[3]The ordering of evaluations in this table has changed from previous iterations of the 2.5 Flash 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 Preview 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.
