Model card published April 16, 2025

Gemini 2.5 Pro 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. Additional reports focused on dangerous capability evaluations will be published at regular cadences, including one coming soon.

Model Information

Description: Gemini 2.5 Pro 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 Pro Preview is Google's most advanced model for complex tasks and can comprehend vast datasets and complex problems from different information sources, including text, audio, images, video, and even entire code repositories.

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 Pro 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. Refinements in architectural design and optimization methods led to substantial improvements in training stability and computational efficiency. Gemini 2.5 Pro Preview was carefully designed and calibrated to balance quality and performance for complex tasks, improving over previous generations.

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 Pro 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. 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 Pro Preview was evaluated against performance benchmarks detailed below:

- **Gemini 2.5 Pro Preview Results:** All Gemini 2.5 Pro scores were 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-pro-exp-03-25 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. All SWE-bench Verified numbers followed official provider reports, using different scaffolding 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).

- **Single attempt vs multiple attempts**: When two numbers were reported for the same eval 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, AIME 2025 numbers, LiveCodeBench results 10/1/2024 - 2/1/2025 in the UI) and Aider Polyglot numbers. For MRCR 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 Pro Preview demonstrated strong performance across a range of benchmarks requiring enhanced reasoning. Detailed results as of March 2025 are listed below:

Capability **Benchmark**		**Gemini 2.5 Pro** (Experimental 03-25)	**OpenAI** **03-mini** High	**OpenAI** **GPT-4.5**	**Claude 3.7** **Sonnet** 64K Extended Thinking	**Grok 3 Beta** Extended Thinking	**DeepSeek** **R1**
Reasoning & Knowledge **Humanity's Last Exam (no tools)**		**18.8%**	14.0%*	6.4%	8.9%	—	8.6%*
Science **GPQA diamond**	single attempt (pass@1)	**84.0%**	79.7%	71.4%	78.2%	80.2%	71.5%
	multiple attempts	—	—	—	**84.8%**	84.6%	—
Mathematics **AIME 2025**	single attempt (pass@1)	**86.7%**	86.5%	—	49.5%	77.3%	70.0%
	multiple attempts	—	—	—	—	**93.3%**	—
Mathematics **AIME 2024**	single attempt (pass@1)	**92.0%**	87.3%	36.7%	61.3%	83.9%	79.8%
	multiple attempts	—	—	—	80.0%	**93.3%**	—
Code generation **LiveCodeBench V5**	single attempt (pass@1)	70.4%	**74.1%**	—	—	70.6%	64.3%
	multiple attempts	—	—	—	—	**79.4%**	—
Code editing **Aider Polyglot**		**74.0% / 68.6%** whole/diff	60.4% diff	44.9% diff	64.9% diff	—	56.9% diff
Agentic coding **SWE-bench verified**		63.8%	49.3%	38.0%	**70.3%**	—	49.2%
Factuality **SimpleQA**		52.9%	13.8%	**62.5%**	—	43.6%	30.1%
Visual reasoning **MMMU**	single attempt (pass@1)	**81.7%**	no MM support	74.4%	75.0%	76.0%	no MM support
	multiple attempts	—	no MM support	—	—	78.0%	no MM support
Image understanding **Vibe-Eval (Reka)**		69.4%	no MM support	—	—	—	no MM support
Long Context **MRCR**	128k (average)	**94.5%**	61.4%	64.0%	—	—	—
	1M (pointwise)	83.1%	—	—	—	—	—
Multilingual performance **Global MMLU (Lite)**		89.8%	—	—	—	—	—

* indicates evaluated on text problems only (without images)

Intended Usage and Limitations

Benefit and Intended Usage: Gemini 2.5 Pro Preview is a thinking model, capable of reasoning before responding, resulting in enhanced performance and improved accuracy. Gemini 2.5 Pro Preview is well-suited for applications that require:

- enhanced reasoning;
- advanced coding;
- multimodal understanding;
- long context.

Known Limitations: Gemini 2.5 Pro 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. The knowledge cutoff date for Gemini 2.5 Pro Preview was January 2025. See the Ethics and Safety Section for additional information on known limitations.

Ethics and Safety

Evaluation Approach: The development of Gemini 2.5 Pro Preview 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 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:** automated and human 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 for safety and security considerations at scale, complementing human efforts and static 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 compared to Gemini 1.5 Pro 002.

Evaluation	Description	**Gemini 2.5 Pro Preview** (in comparison to Gemini 1.5 Pro 002)
Text to Text Safety	Automated content safety evaluation measuring safety policies	-3.6%
Multilingual Safety	Automated safety policy evaluation across multiple languages	-2.1%
Tone	Automated evaluation measuring objective tone of model refusal	+5.70%
Instruction Following	Automated evaluation measuring model's ability to follow instructions while remaining safe	+12.60%
Image to Text Safety	Automated content safety evaluation measuring safety policies	-4.2%

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 saw Gemini 2.5 Pro Preview displaying low violation rates across modalities and improvement in safety profile compared to Gemini 1.5.

Known Safety Limitations: The main safety limitations for Gemini 2.5 Pro Preview are over-refusals and tone. The model will sometimes refuse to answer on prompts where an answer would not violate policies. Refusals can still come across as "preachy," although overall tone and instruction following have improved compared to Gemini 1.5.

Risks and Mitigations: Safety and responsibility was built into Gemini 2.5 Pro 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.