Google

Gemini 3.6 Flash
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.

Published: July 2026

Model Information

Description

Gemini 3.6 Flash is an addition to the Gemini 3 series of highly-capable, natively multimodal, reasoning models. Gemini 3.6 Flash is our workhorse model that delivers better coding, knowledge work, and multimodal performance, while providing better token-efficiency than Gemini 3.5 Flash.

Model dependencies

Gemini 3.6 Flash is based on Gemini 3.5 Flash.

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 3.6 Flash is based on Gemini 3.5 Flash. For more information about the model architecture for Gemini 3.6 Flash, see the Gemini 3.5 Flash model card.

Model Data

Training Dataset

Gemini 3.6 Flash is based on Gemini 3.5 Flash. For more information about the training dataset for Gemini 3.6 Flash, see the Gemini 3.5 Flash model card.

Training Data Processing

For more information about the training data processing for Gemini 3.6 Flash, see the Gemini 3.5 Flash model card.

Implementation and Sustainability

Hardware

Gemini 3.6 Flash is based on Gemini 3.5 Flash. For more information about the hardware for Gemini 3.6 Flash and our continued commitment to operate sustainably, see the Gemini 3.5 Flash model card.

Software

Gemini 3.6 Flash is based on Gemini 3.5 Flash. For more information about the software for Gemini 3.6 Flash, see the Gemini 3.5 Flash model card.

Distribution

Gemini 3.6 Flash is distributed in the following channels; respective documentation shared in line:

- Gemini App
- Gemini Enterprise App
- Gemini Enterprise Agent Platform
- Google AI Studio
- Gemini API
- Google Antigravity

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 Gemini Enterprise Agent Platform, see Google Cloud Platform Terms of Service. For more information, see Gemini Model API instructions and Gemini API quickstart.

Evaluation

Approach

Gemini 3.6 Flash was evaluated across a range of benchmarks, including reasoning, coding, agentic, multimodal capabilities, and long-context. Additional benchmarks and details on approach, results and their methodologies can be found at: deepmind.com/models/evals-methodology/gemini-3-6-flash.

Results

Results as of July, 2026 are listed below:

Benchmark		Gemini 3.6 Flash	Gemini 3.5 Flash	Gemini 3.1 Pro	GPT-5.6 Luna	Grok 4.5	Claude Sonnet 5
Input price $/1M tokens		$1.50	$1.50	$2.00	$1.00	$2.00	$3.00 (full-price) $2.00 (temp discount)
Output price $/1M tokens		$7.50	$9.00	$12.00	$6.00	$6.00	$15.00 (full-price) $10.00 (temp discount)
SWE-Bench Pro (Public) Diverse agentic coding tasks		58.7%	55.1%	54.2%	62.7%	**64.7%**	63.2%
DeepSWE v1.1 Long-horizon software engineering		49%	37%	12%	**67%**	54%	54%
Terminal-bench 2.1 Agentic terminal coding	Terminus-2 harness	78.0%	76.2%	73.8%	**84.7%**	83.3%	80.4%
MLE-Bench Machine Learning Engineering		63.9%	49.7%	42.6%	47.6%	43.2%	**66.9%**
GDPVal-AA v2 Knowledge work	Elo	1421	1349	965	1584	1535	**1607**
OSWorld-Verified Computer use		**83.0%**	78.4%	76.2%	72.6%	-	81.2%
CharXiv Reasoning Information synthesis from complex charts	no tools	**85.2%**	84.2%	83.3%	82.7%	81.6%	77.0%
	with tools	**89.4%**	84.9%	83.2%	-	-	88.3%
GDM-MRCR v2 (8-needle) Long context performance	128k (average)	**91.8%**	77.3%	84.9%	74.8%	81.4%	71.6%
	1M (pointwise)	**54.0%**	26.6%	26.3%	-	-	-

Methodology: deepmind.google/models/evals-methodology/gemini-3-6-flash

---

Intended Usage and Limitations

Benefit and Intended Usage

Gemini 3.6 Flash is well-suited for users, developers, and enterprises, some use cases include: agentic workflows, coding tasks, and multi-week enterprise processes.

Known Limitations

Gemini 3.6 Flash may exhibit some of the general limitations of foundation models, such as hallucinations. In addition to this, we are continually working to improve jailbreak resistance and have recently strengthened the mitigations across Frontier Safety. There may also be occasional slowness or timeout issues. The knowledge cutoff date for Gemini 3.6 Flash is March 2026 – users can expect updated information for some domains while in others they may experience the model's knowledge is limited to January 2025 (in line with the Gemini 3 Model Family). For more information about known limitations, see the Gemini 3.5 Flash model card.

Acceptable Usage

For more information about the acceptable usage for Gemini 3.6 Flash, see the Gemini 3.5 Flash model card.

Ethics and Content Safety

Evaluation Approach

For more information about the evaluation approach for Gemini 3.6 Flash, see the Gemini 3.5 Flash model card.

Safety Policies

For more information about the safety policies for Gemini 3.6 Flash, see the Gemini 3.5 Flash 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 3.6 Flash outperforms Gemini 3.5 Flash across both multi-lingual and content safety, while continuing to keep unjustified refusals low. Flash 3.6 saw slight regressions in tone, however this does not impact the overall safety of the model. We mark improvements in bolded green and regressions in red.

Evaluation	**Description**	**Gemini 3.6 Flash** vs. Gemini 3.5 Flash *(percentage point increase/decrease)*
Text to Text Safety	Automated content safety evaluation measuring safety policies	**-1.35%** *Lower is better*
Multilingual Safety	Automated safety policy evaluation across multiple languages	**-5.45%** *Lower is better*
Image to Text Safety	Automated content safety evaluation measuring safety policies	0% *Lower is better*
Tone[1]	Automated evaluation measuring objective tone of model refusal	-3.31% *Higher is better*
Unjustified-refusals	Automated evaluation measuring model's ability to respond to borderline prompts while remaining safe	+0.25% *Lower is better*

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.

[1] For tone, 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 3.5 Flash.

Human Red Teaming Results

We conduct manual red teaming by specialist teams who sit outside of the model development team in Google's Trust & Safety organization. High-level findings are fed back to the model team. For child safety evaluations, Gemini 3.6 Flash 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 3.5 Flash. Additionally, the scope of red teaming covered potential issues outside of our strict policies, compared performance to Gemini 3.1 Pro, and found no egregious concerns.

Frontier Safety Assessment

Gemini 3.6 Flash is part of the Gemini 3 series of models. We evaluated Gemini 3.1 Pro for Frontier Safety as it was 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, while Gemini 3.6 Flash excels at agents and coding, it does not have meaningful new capabilities or material increases in performance with respect to the domains outlined in our Frontier Safety Framework compared to Gemini 3.1 Pro; therefore, based on Gemini 3.1 Pro results, we are confident that Gemini 3.6 Flash is also unlikely to reach any CCLs.

As previous models in the Gemini 3 series reached the alert threshold for cyber, we performed additional testing in this domain and found that Gemini 3.6 Flash remains below the cyber CCL. For more information on our Frontier Safety Assessment, read the Gemini 3.1 Pro model card.

Risks and Mitigations

To make Gemini 3.6 Flash more resistant to jailbreaks, we enhanced Frontier Safety safeguards in the domains of Chemical, Biological, Radiological, and Nuclear (CBRN) and cyber offense misuses. We also trained the model to minimize refusals for beneficial uses. To see additional information about risks and mitigations for the Gemini 3 series, see the Gemini 3.1 Pro model card.