Google

Gemini 3.8 Flash
Model card

Google

Gemini 3.8 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: September, 2026

Model Information

Description

Gemini 3.8 Flash is the next iteration in the Gemini 3 model family, building on Gemini 3.7 Flash, delivering performance advancements across software engineering and agentic knowledge workflows. It continues to support customizable effort levels to control the mix of quality, cost and latency.

Model dependencies

Gemini 3.8 Flash is based on Gemini 3.7 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 3.8 Flash is based on Gemini 3.7 Flash. For more information about the model architecture for Gemini 3.8 Flash, see the Gemini 3.7 Flash model card.

Google

Model Data

Training Dataset

Gemini 3.8 Flash is based on Gemini 3.7 Flash. For more information about the training dataset for Gemini 3.8 Flash, see the Gemini 3.7 Flash model card.

Training Data Processing

For more information about the training data processing for Gemini 3.8 Flash, see the Gemini 3.7 Flash model card.

Implementation and Sustainability

Hardware

Gemini 3.8 Flash is based on Gemini 3.7 Flash. For more information about the hardware for Gemini 3.8 Flash and our continued commitment to operate sustainably, see the Gemini 3.7 Flash model card.

Software

Gemini 3.8 Flash is based on Gemini 3.7 Flash. For more information about the software for Gemini 3.8 Flash, Gemini 3.7 Flash model card.

Distribution

Gemini 3.8 Flash is distributed in the following channels; respective documentation shared in line:

- Gemini app
- Gemini Enterprise Agent Platform
- Google AI Studio
- Gemini API
- Google AI Mode
- Google Antigravity

Google

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 3.8 Flash was evaluated across a range of benchmarks, including coding, knowledge work, multimodal capabilities, long-context, computer use, scientific reasoning. Additional benchmarks and details on approach, results and their methodologies can be found at: deepmind.com/models/evals-methodology/gemini-3-8-flash.

Google

Results

Results as of September, 2026 are listed below:

	Gemini 3.8 Flash	Gemini 3.7 Flash	Claude Opus 5	Claude Sonnet 5	GPT-5.6 Sol	GPT-5.6 Terra
Input price $/1M tokens, no caching	**$0.75** ($1.50 regular)	**$0.75** ($1.50 regular)	$5.00	$2.00	$4.00	$2.00
Output price $/1M tokens	**$3.75** ($7.50 regular)	**$3.75** ($7.50 regular)	$25.00	$10.00	$20.00	$12.00
DeepSWE v1.1 Long-horizon software engineering	73.7%	65.3%	**74.0%**	53.8%	72.7%	69.6%
GDPVal-AA v2 Knowledge work Elo	1545	1482	**1824**	1584	1710	1528
Vals Finance Agent v2 Financial analyst tasks	**61.4%**	59.0%	58.6%	53.9%	53.8%	54.4%
Harvey's Legal Agent Benchmark Complex legal workflows All pass rate	**10.0%**	8.8%	6.7%	5.0%	2.5%	0.8%
Terminal-bench 2.1 Agentic terminal coding	**89.4%**	85.8%	89.1%	80.4%	88.8%	87.4%
Terminal-bench 4.0 General agent capabilities	19.1%	11.2%	**51.8%**	12.4%	37.3%	23.6%
GDP.PDF Expert PDF document comprehension All pass rate	35.0%	34.0%	37.0%	28.0%	**40.0%**	29.0%
CharXiv Reasoning Information synthesis from complex charts No tools	**86.2%**	84.5%	83.7%	70.1%	85.8%	85.9%
LVBench Long video understanding	**87.8%** (agentic) **87.1%** (static)	85.4%	75.4%	68.5%	82.1%	78.9%
HLE-Verified Multidisciplinary expert reasoning	**54.9%**	53.6%	54.4%	31.0%	54.5%	51.1%
OSWorld-2.0 Agentic computer use Partial score batch tool enabled	59.0%	50.6%	**75.4%**	42.6%	62.6%	50.2%
BioMysteryBench Bioinformatics research workflows Human Solvable	88.8%	87.1%	**90.1%**	87.5%	79.5%	83.8%
BioMysteryBench Bioinformatics research workflows Human Difficult	**56.5%**	43.5%	49.4%	34.1%	44.7%	49.4%
LABBench2 Biology real-world research tasks	**86.2%**	82.1%	84.2%	80.1%	82.1%	81.2%

Methodology: deepmind.google/models/evals-methodology/gemini-3-8-flash

* For 3.7 and 3.8 Flash, introductory price expires on December 31, 2026. Starting January 1, 2027, $1.50/1M input tokens and $7.50/1M output tokens will apply.

For details on our evaluation methodology please see:
deepmind.com/models/evals-methodology/gemini-3-8-flash

Google

Intended Usage and Limitations

Benefit and Intended Usage

Gemini 3.8 Flash is well-suited for users, developers, and enterprises, designed for cost-effective scaling of general-purpose, production-ready agents. Some use cases include: software engineering, agent tasks, and complex knowledge workflows.

Known Limitations

Gemini 3.8 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. At times, the model might use more tokens to maximize performance, especially at higher effort levels.

The knowledge cutoff date for Gemini 3.8 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 3.7 Flash model card.

Acceptable Usage

For more information about the acceptable usage for Gemini 3.7 Flash, see the Gemini 3.7 Flash model card.

Ethics and Content Safety

Evaluation Approach

For more information about the evaluation approach for Gemini 3.8 Flash, see the Gemini 3.7 Flash model card.

Safety Policies

For more information about the safety policies for Gemini 3.8 Flash, see the Gemini 3.7 Flash model card.

Google

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 3.8 Flash performs similarly to Gemini 3.7 Flash across both safety and tone, with low unjustified refusals. Safety performance across non-English languages regressed slightly relative to 3.7 Flash.

Evaluation	**Description**	**Gemini 3.8 Flash** vs. Gemini 3.7 Flash
Text to Text Safety	Automated content safety evaluation measuring safety policies	-0.4pp *Lower is better*
Multilingual Safety	Automated safety policy evaluation across multiple languages	+5.4pp *Lower is better*
Image to Text Safety	Automated content safety evaluation measuring safety policies	0.0pp *Lower is better*
Tone[1]	Automated evaluation measuring the objective tone of model responses	+0.2pp *Higher is better*
Unjustified-refusals	Automated evaluation measuring model's ability to respond to borderline prompts while remaining safe	+1.1pp *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 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 3 Flash. We mark improvements in green and regressions in red.

Google

Human Red Teaming Results

We conduct manual red teaming by specialist teams who sit outside of the model development team. High-level findings are fed back to the model team. For child safety evaluations, Gemini 3.8 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 3.7 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 3.8 Flash is part of the Gemini 3 series of models. We evaluated Gemini 3.7 Flash as outlined in our latest Frontier Safety Framework (April-2026), and found that it did not reach any Tracked or Critical Capability Levels (T/CCLs). Our assessments have shown that Gemini 3.8 Flash 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 3.7 Flash; therefore, based on Gemini 3.7 Flash results, we are confident that Gemini 3.8 Flash is also unlikely to reach any T/CCLs.

For more information on our Frontier Safety assessment, read the Gemini 3.7 Flash model card.