



# Gemini 3.5 Flash-Lite Model card

## Gemini 3.5 Flash-Lite — 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 3.5 Flash-Lite is an addition to the Gemini 3 series of highly-capable, natively multimodal, reasoning models. The model is cost-efficient and fast, optimized for high-volume, latency-sensitive tasks like translation and classification as well as supporting agentic workflows.

### Model dependencies

Gemini 3.5 Flash-Lite is based on Gemini 3.1 Flash-Lite.

### 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 3.5 Flash-Lite is based on Gemini 3.1 Flash-Lite. For more information about the model architecture for Gemini 3.5 Flash-Lite, see the Gemini 3.1 Flash-Lite [model card](#).

---

# Model Data

## Training Dataset

Gemini 3.5 Flash-Lite is based on Gemini 3.1 Flash-Lite. For more information about the training dataset for 3.5 Flash-Lite, see the Gemini 3.1 Flash-Lite [model card](#).

## Training Data Processing

For more information about the training data processing for 3.5 Flash-Lite, see the Gemini 3.1 Flash-Lite [model card](#).

---

# Implementation and Sustainability

## Hardware

Gemini 3.5 Flash-Lite is based on Gemini 3.1 Flash-Lite. For more information about the hardware for Gemini 3.5 Flash-Lite and our continued [commitment to operate sustainably](#), see the Gemini 3.1 Flash-Lite [model card](#).

## Software

Gemini 3.5 Flash-Lite is based on Gemini 3.1 Flash-Lite. For more information about the software for Gemini 3.5 Flash-Lite, see the Gemini 3.1 Flash-Lite [model card](#).

---

# Distribution

Gemini 3.5 Flash-Lite is distributed in the following channels; respective documentation shared in line:

- [Gemini App](#)
- [Google AI Studio](#)
- [Gemini Enterprise Agent Platform](#)
- [Gemini API](#)

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 3.5 Flash-Lite 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-3-5-flash-lite](#).

## Results

Results as of July, 2026 are listed below:

| Benchmark                                 |                              | Gemini 3.5<br>Flash-Lite | Gemini 3.1<br>Flash-Lite | GPT-5.4<br>mini | Claude<br>Haiku 4.5 |
|-------------------------------------------|------------------------------|--------------------------|--------------------------|-----------------|---------------------|
| Input price                               | \$/M tokens                  | \$0.30                   | \$0.25                   | \$0.75          | \$1.00              |
| Output price                              | \$/M tokens                  | \$2.50                   | \$1.50                   | \$4.50          | \$5.00              |
| SWE-Bench Pro (Public)                    | Diverse agentic coding tasks | 54.2%                    | 38.3%                    | 54.4%           | 39.5%               |
| Terminal-bench 2.1                        | Agentic terminal coding      | 54.0%                    | 31.0%                    | 59.2%           | 44.2%               |
| Terminus-2 harness                        |                              |                          |                          |                 |                     |
| MLE-Bench                                 | Machine Learning Engineering | 39.2%                    | 22.0%                    | -               | -                   |
| GDPVal-AA v2                              | Knowledge work               | 1140                     | 642                      | 1171            | 907                 |
| Elo                                       |                              |                          |                          |                 |                     |
| OSWorld-Verified                          | Computer use                 | 74.0%                    | 54.3%                    | 72.1%           | 50.7%               |
| CharXiv Reasoning                         | no tools                     | 74.5%                    | 73.2%                    | 80.3%           | 61.7%               |
| Information synthesis from complex charts | with tools                   | 76.5%                    | 75.6%                    | -               | -                   |
| GDM-MRCR v2 (8-needle)                    | 128k (average)               | 72.2%                    | 60.1%                    | 42.7%           | 35.3%               |
| Long context performance                  | 1M (pointwise)               | 21.3%                    | 12.3%                    | -               | -                   |

Methodology: [deepmind.google/models/evals-methodology/gemini-3-5-flash-lite](#)

---

# Intended Usage and Limitations

## Benefit and Intended Usage

Gemini 3.5 Flash-Lite is well-suited for users, developers, and enterprises, some use cases include: agentic workflows, coding tasks, as well as high throughput tasks like document processing or search.

## Known Limitations

Gemini 3.5 Flash-Lite 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 3.5 Flash-Lite 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 3.1 Flash-Lite](#) model card.

## Acceptable Usage

For more information about the acceptable usage for Gemini 3.5 Flash-Lite, see the [Gemini 3.1 Flash-Lite](#) model card.

---

## Ethics and Content Safety

### Evaluation Approach

For more information about the evaluation approach for Gemini 3.5 Flash-Lite, see the [Gemini 3.1 Flash-Lite](#) model card.

### Safety Policies

For more information about the safety policies for Gemini 3.5 Flash-Lite, see the [Gemini 3.1 Flash-Lite](#) 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 3.5 Flash-Lite outperforms Gemini 3.1 Flash-Lite across both safety and tone. There was a slight regression in unjustified refusals, however we see this because the model is more likely to refuse to answer on sensitive topics to remain safe. We mark improvements in bolded green and regressions in red.

| Evaluation           | Description                                                                                          | Gemini 3.5 Flash-Lite<br>vs. Gemini 3.1 Flash-Lite<br>(percentage point<br>increase/decrease) |
|----------------------|------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|
| Text to Text Safety  | Automated content safety evaluation measuring safety policies                                        | <b>-8.14%</b><br><i>Lower is better</i>                                                       |
| Multilingual Safety  | Automated safety policy evaluation across multiple languages                                         | <b>-0.92%</b><br><i>Lower is better</i>                                                       |
| Image to Text Safety | Automated content safety evaluation measuring safety policies                                        | 0%<br><i>Lower is better</i>                                                                  |
| Tone <sup>1</sup>    | Automated evaluation measuring objective tone of model refusal                                       | <b>+3.04%</b><br><i>Higher is better</i>                                                      |
| Unjustified-refusals | Automated evaluation measuring model's ability to respond to borderline prompts while remaining safe | <b>+5.32%</b><br><i>Lower is better</i>                                                       |

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.

## 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 3.5

<sup>1</sup> 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 3.1 Flash-Lite.

Flash-Lite 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 3.1 Flash-Lite. 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 3.5 Flash-Lite 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 3.5 Flash-Lite 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 3.5 Flash-Lite is also unlikely to reach any CCLs.

For more information on our Frontier Safety Assessment, read the [Gemini 3.1 Pro Model Card](#).

## Risks and Mitigations

To see additional information about the risks and mitigations for Gemini 3.5 Flash-Lite, see the Gemini 3.1 Pro [model card](#).