



Gemini 3.8 Audio (Live, Live Extended Thinking) Model card

Gemini 3.8 Audio — 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 Audio (Live, Live Extended Thinking) is an addition to the Gemini 3 series of highly-capable, natively multimodal, reasoning models. The models are cost-efficient and fast, optimized for high-volume, latency-sensitive tasks like real-time dialogue. This model card describes the native audio capabilities as additional outputs of Gemini. Information specific to these modalities is specified in-line and referred to as **Gemini 3.8 Live**, and **Gemini 3.8 Live Extended Thinking** referred to collectively as Gemini 3.8 Audio.

Model dependencies

Gemini 3.8 Audio is based on Gemini 3 Pro.

Inputs

Gemini 3.8 Live and Gemini 3.8 Live Extended Thinking: Audio, images, video, and text with a token context window of up to 128K.

Outputs

Gemini 3.8 Live and 3.8 Live Extended Thinking: Audio and text, with 64K token output.

Architecture

Gemini 3.8 Audio models are based on Gemini 3 Pro. For more information about the model architecture, see the Gemini 3 Pro [model card](#).

Model Data

Training Dataset

Gemini 3.8 Live and Gemini 3.8 Live Extended Thinking are based on Gemini 3 Pro. For more information about the training dataset, see the Gemini 3 Pro [model card](#).

Training Data Processing

For more information about the training data processing for Gemini 3.8 Live and Gemini 3.8 Live Extended Thinking, see the Gemini 3 Pro [model card](#).

Implementation and Sustainability

Hardware

Gemini 3.8 Live and Gemini 3.8 Live Extended Thinking are based on Gemini 3 Pro. For more information about the hardware for Gemini 3 Pro and our continued [commitment to operate sustainably](#), see the Gemini 3 Pro [model card](#).

Software

Gemini 3.8 Live and Gemini 3.8 Live Extended Thinking are based on Gemini 3 Pro. For more information about the software for Gemini 3 Pro, see the Gemini 3 Pro [model card](#).

Distribution

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

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

- [Google Cloud / Vertex AI](#)
- [Google Search Live](#)

Gemini 3.8 Live Extended Thinking is distributed in the following channels; respective documentation shared in line:

- [Gemini App](#)
- [Gemini API](#)
- [Google AI Studio](#)
- [Google Cloud / Vertex AI](#)
- Google Workspace ([Gmail](#), [Docs](#), and [Keep](#))

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 Audio (Gemini 3.8 Live and Gemini 3.8 Live Extended Thinking) was evaluated across a range of benchmarks. See [Evals & Methodology for 3.8 Live Extended Thinking](#) for more details.

Intended Usage and Limitations

Benefit and Intended Usage

Gemini 3.8 Live and Gemini 3.8 Live Extended Thinking process continuous streams of audio, video, and text to deliver spoken responses in real-time to create natural conversational experiences well-suited for users, developers, and enterprises.

Known Limitations

Gemini 3.8 Live and Gemini 3.8 Live Extended Thinking 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 is January 2025.

For more information about the known limitations for Gemini 3 Pro, see the Gemini 3 Pro [model card](#).

Acceptable Usage

For more information about the acceptable usage for Gemini 3.8 Live and Gemini 3.8 Live Extended Thinking, see the Gemini 3 Pro [model card](#).

Ethics and Content Safety

Evaluation Approach

Gemini 3.8 Audio was developed in partnership with internal safety 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](#), as well as Google's Generative AI policies (e.g., [Generative AI Prohibited Use Policy](#) and the [Gemini API Additional Terms of Service](#)).

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 Evaluations** conducted by specialist teams across the policies and desiderata to ensure the model adheres to safety policies and desired outcomes.

Additional information: For more information about the evaluation approach for Gemini 3.8 Live and Gemini 3.8 Live Extended Thinking, see the Gemini 3 Pro [model card](#).

Safety Policies

For more information about the safety policies for Gemini 3.8 Live and Gemini 3.8 Live Extended Thinking, see the Gemini 3 Pro [model card](#).

Frontier Safety Assessment

Gemini 3.8 Audio is part of the Gemini 3 series of models. We evaluated Gemini 3.7 Flash as outlined in our latest [Frontier Safety Framework](#), and we found that it did not reach any Tracked or Critical Capability Levels (T/CCLs). Our assessments have shown that Gemini 3.8 Live or Gemini 3.8 Live Extended Thinking does not have meaningful new capabilities or material increases in performance compared to Gemini 3.7 Flash; therefore, based on Gemini 3.7 Flash results, we are confident that Gemini 3.8 Live and Gemini 3.8 Live Extended Thinking are not likely to reach any T/CCLs.

For more information on our Frontier Safety Assessment, read the [Gemini 3.7 Flash Model Card](#).

Risks and Mitigations

For more information about the risks and mitigations for Gemini 3.8 Live and Gemini 3.8 Live Extended Thinking, see the Gemini 3 Pro [model card](#).