



Gemini 3.8 Audio (Live, Live Extended Thinking, Flash TTS, Flash-Lite TTS)

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, Flash TTS, Flash-Lite TTS) is an addition to the Gemini 3 series of highly-capable models. This model card describes the native capabilities (e.g., audio and video) as additional outputs of Gemini, including optional avatar capabilities (Live Avatar). Information specific to these modalities is specified in-line and referred to as **Gemini 3.8 Live**, **Gemini 3.8 Live Extended Thinking**, **Gemini 3.8 Flash TTS**, and **Gemini 3.8 Flash-Lite TTS** (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.
- Gemini 3.8 Flash TTS and Gemini 3.8 Flash-Lite TTS: Text up to 8K.

Outputs

- Gemini 3.8 Live and Gemini 3.8 Live Extended Thinking: Audio and text, with 64K token output. With Live Avatar: Audio, video, and text, with 24K token output.
- Gemini 3.8 Flash TTS and 3.8 Flash-Lite TTS: Audio, 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 Audio models 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 Audio models, see the Gemini 3 Pro [model card](#).

Implementation and Sustainability

Hardware

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

Gemini 3.8 Live

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

Gemini 3.8 Live Extended Thinking

- [Gemini App](#)
- [Gemini API](#)
- [Google AI Studio](#)
- [Gemini Enterprise Agent Platform](#)
- Google Workspace ([Gmail](#), [Docs](#), and [Keep](#))

Gemini 3.8 Flash TTS

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

Gemini 3.8 Flash-Lite TTS

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

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 models were evaluated across a range of benchmarks. For more details, see:

- [Evals & Methodology for 3.8 for Live and Live Extended Thinking](#)
 - [Evals & Methodology for 3.8 Flash TTS and Flash-Lite TTS](#)
-

Intended Usage and Limitations

Benefit and Intended Usage

Gemini 3.8 Audio models 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.

Gemini 3.8 Live with Live Avatar and Gemini 3.8 Live Extended Thinking with Live Avatar generate expressive videos of natural head movements and speech synchronization based on configured images and audio inputs.

Known Limitations

Gemini 3.8 Audio 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.

Gemini 3.8 Live with Live Avatar and Gemini 3.8 Live Extended Thinking with Live Avatar can support a few minutes of continuous interaction, rather than extended hours.

The knowledge cutoff date is January 2025.

For more information about the known limitations for Gemini 3.8 Audio, see the Gemini 3 Pro [model card](#).

Acceptable Usage

For more information about the acceptable usage for Gemini 3.8 Audio, 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.
- **Child Safety Evaluations** were conducted using techniques developed by expert teams to protect children online and to meet [Google's commitments to child safety](#) across our models and Google products.

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

Safety Policies

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

Frontier Safety Assessment

Gemini 3.8 Audio is part of the Gemini 3 series of models. To assess Gemini 3.8 Audio for frontier safety, we rely on our evaluations of Gemini 3.7 Flash as outlined in our latest [Frontier Safety Framework](#). We found that Gemini 3.7 Flash did not reach any Tracked or Critical Capability Levels (T/CCLs). Our assessments have shown that Gemini 3.8 Live, Gemini 3.8 Live Extended Thinking, Gemini 3.8 Flash TTS, or Gemini 3.8 Flash-Lite TTS do 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, Gemini 3.8 Live Extended Thinking, Gemini 3.8 Flash TTS, or Gemini 3.8 Flash-Lite TTS 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 Audio, see the Gemini 3 Pro [model card](#).