Google

Gemini 3.1 Flash with Native Audio Capabilities (Flash Live) Model Card

Gemini 3.1 Flash with Native Audio Capabilities (Flash Live) - 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: March 2026

Model Information

Description: Gemini 3.1 Flash with Native Audio Capabilities (Flash Live) (hereafter, "Gemini 3.1 Flash Live"), is a member of the Gemini series of models, a suite of highly-capable, natively multimodal reasoning models.

Model dependencies: Gemini 3.1 Flash Live is based on Gemini 3 Pro.

Inputs: Audio, images, video, and text with a token context window of up to 128K.

Outputs: Audio and text, with 64K token output.

Architecture: Gemini 3.1 Flash Live is based on Gemini 3 Pro. For more information about the model architecture for Gemini 3.1 Flash Live, see the Gemini 3 Pro model card.

Model Data

Training Dataset: Gemini 3.1 Flash Live is based on Gemini 3 Pro. For more information about the training dataset for Gemini 3.1 Flash Live, see the Gemini 3 Pro model card.

Training Data Processing: For more information about the training data processing for Gemini 3.1 Flash Live, see the Gemini 3 Pro model card.

Implementation and Sustainability

Hardware: Gemini 3.1 Flash Live is based on Gemini 3 Pro. For more information about the hardware for Gemini 3.1 Flash Live and our continued commitment to operate sustainably, see the Gemini 3 Pro model card.

Software: Gemini 3.1 Flash Live is based on Gemini 3 Pro. For more information about the software for Gemini 3.1 Flash Live, see the Gemini 3 Pro model card.

Distribution

Gemini 3.1 Flash Live is distributed in the following channels; respective documentation shared in line:

- Gemini App
- Google AI Studio
- Gemini API
- Google Antigravity
- NotebookLM

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 Vertex AI, see Google Cloud Platform Terms of Service. For more information, see Gemini Model API instructions and Gemini API in Vertex AI quickstart.

Evaluation

The following Evaluation approach and results are for Gemini 3.1 Flash Live.

Approach: Gemini 3.1 Flash Live was evaluated using the methodology below:

Capabilities / Benchmarks:

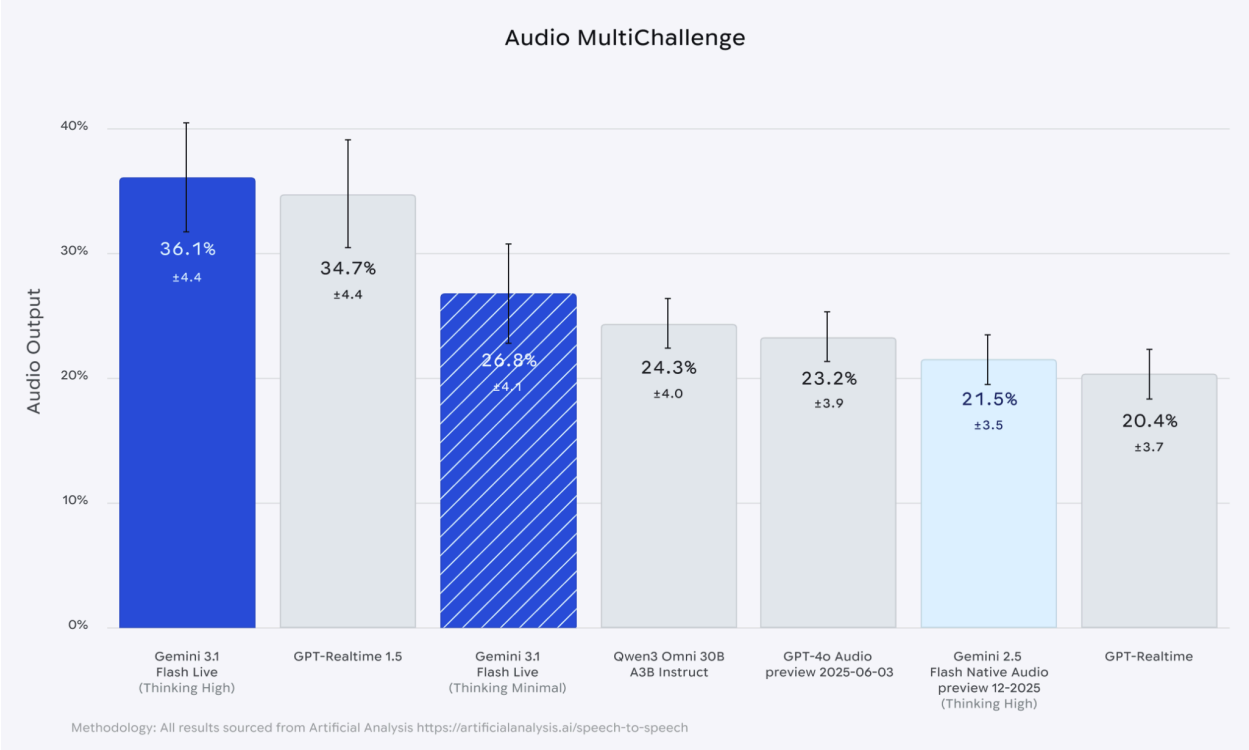
- **Audio Multi Challenge:** This multi-turn benchmark assesses the conversational proficiency of audio-language models and spoken dialogue systems, including speech-to-speech variants. It evaluates their capacity to follow instructions, maintain self-consistency, integrate previous context, and manage natural speech corrections throughout long-form dialogues.
- **Big Bench Audio:** This single turn benchmark consists of 1,000 audio recordings that pair an audio clip (ranging from speech to natural sounds) with a text question. It measures five diverse audio comprehension skills: audio captioning, speech understanding, audio scene understanding, accent/language identification, and sound recognition.
- **ComplexFuncBench**: This static context multi-turn benchmark measures the model's ability to perform a sequence of interdependent function calls related to travel booking. Since this was originally a text-to-text evaluation, we synthesized audio for each prompt and used the published scoring apparatus to evaluate the performance of the Gemini realtime API. More details on ComplexFuncBench can be found here.

Evaluation Methodology

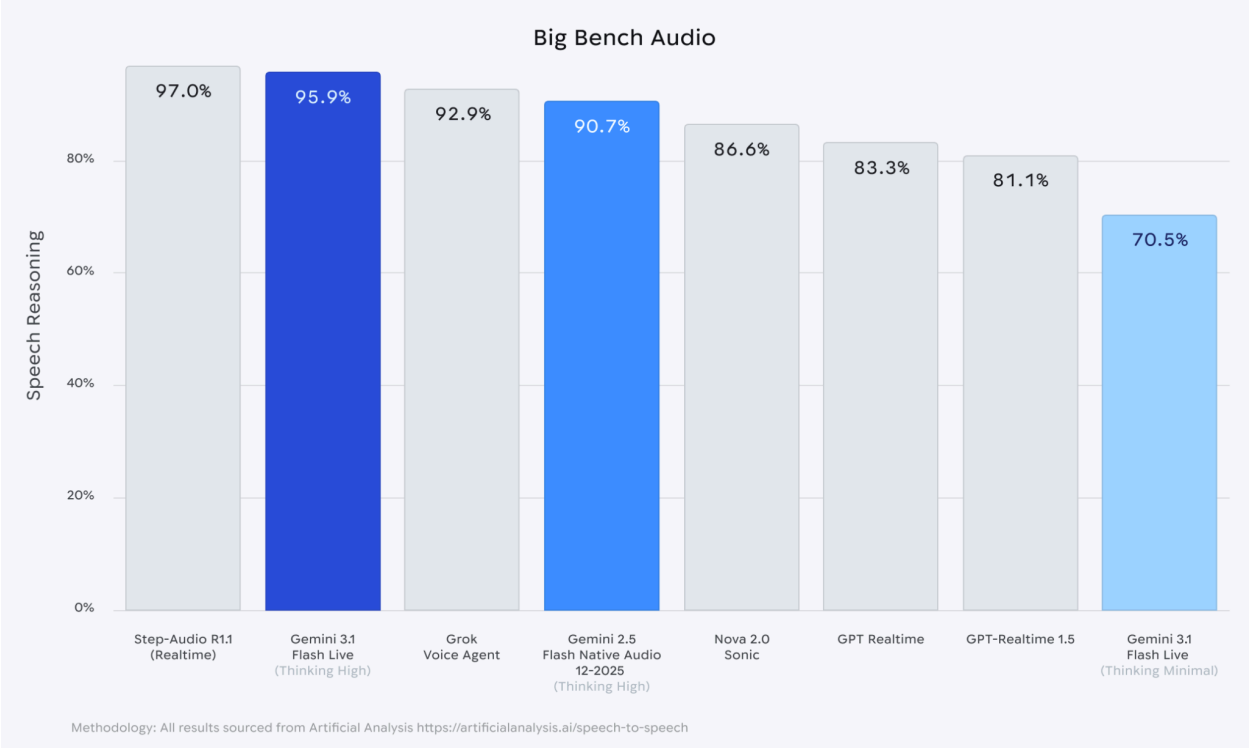
- We ran ComplexFunBench using the Gemini Live API. Big Bench Audio and Audio Multi Challenge were run externally by Artificial Analysis and Scale AI respectively.

Scores available:

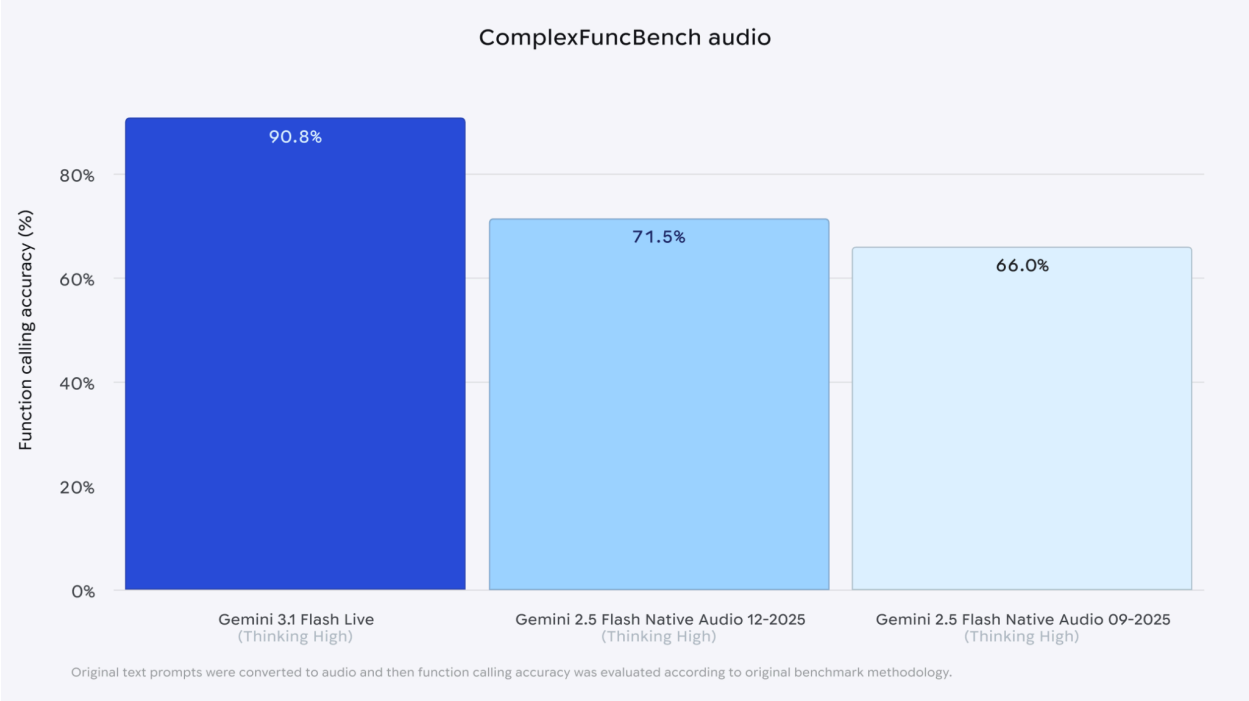
[Scale AI Multi Challenge Leaderboard]



[Artificial Analysis Speech-to-Speech Leaderboard]



ComplexFunBench Audio



Intended Usage and Limitations

Benefit and Intended Usage: Gemini 3.1 Flash Live enables low-latency, real-time voice and video interactions. It processes continuous streams of audio, video, or text to deliver immediate, human-like spoken responses, creating a natural conversational experience for your users.

Known Limitations: For more information about the known limitations for Gemini 3.1 Flash Live, see the Gemini 3 Pro model card.

Acceptable Usage: For more information about the acceptable usage for Gemini 3.1 Flash Live, see the Gemini 3 Pro model card.

Ethics and Content Safety

Evaluation Approach: Gemini 3.1 Flash Live 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. Gen 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;
- **Ethics & Safety Reviews** were conducted ahead of the model's release.

Safety Policies: Gemini's safety policies are based on Google's standard framework, which aim to prevent our Generative AI models from generating harmful content, including:

1. Content related to child sexual abuse material and exploitation
2. Hate speech (e.g. dehumanizing members of protected groups)
3. Dangerous content (e.g. promoting suicide, or instructing in activities that could cause real-world harm)
4. Harassment (e.g. encouraging violence against people)
5. Sexually explicit content
6. Medical advice that runs contrary to scientific or medical consensus

Frontier Safety Assessment: Gemini 3.1 Flash Live is part of the Gemini 3 family of models. For frontier safety, we rely on our evaluation of Gemini 3.1 Pro with Deep Think mode as it is the most generally capable model as of publication of this model card, and it did not reach the Critical Capability Levels (CCLs) outlined in our Frontier Safety Framework. Our assessments have shown that Gemini 3.1 Flash Live is less capable than Gemini 3.1 Pro, therefore based on Gemini 3.1 Pro, we are confident that Gemini 3.1 Flash is also unlikely to reach any CCLs. For more information, read the Gemini 3.1 Pro Model Card.

Risks and Mitigations: For more information about the risks and mitigations for Gemini 3.1 Flash Live, see the Gemini 3 Pro model card.