Published: March, 2026

Model Evaluation – Approach, Methodology & Results

Gemini 3.1 Flash Live

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

Capabilities / Benchmarks:

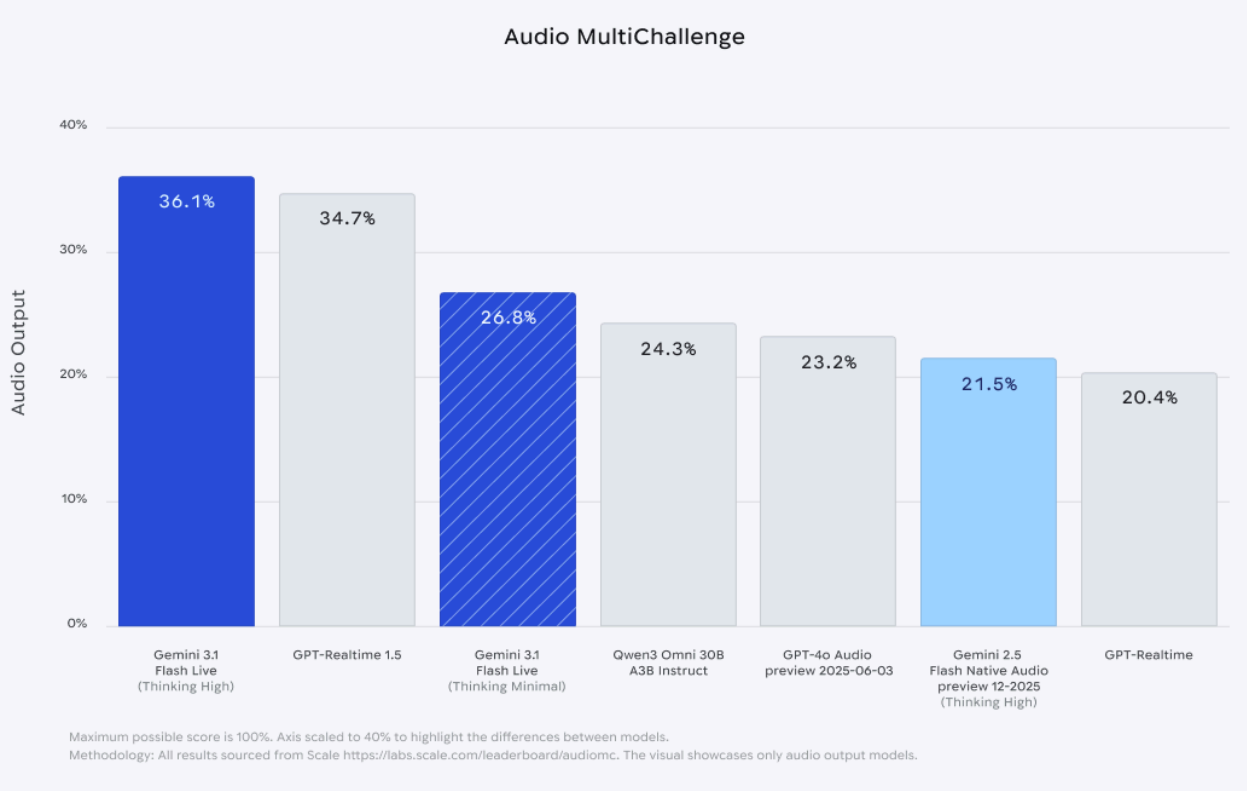
- **Audio Multi Challenge:** This multi-turn benchmark assesses the conversational proficiency of audio-language models and spoken dialogue systems, including speech-to-speech variants. It evaluates their capacity to follow instructions, maintain self-consistency, integrate previous context, and manage 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.

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.

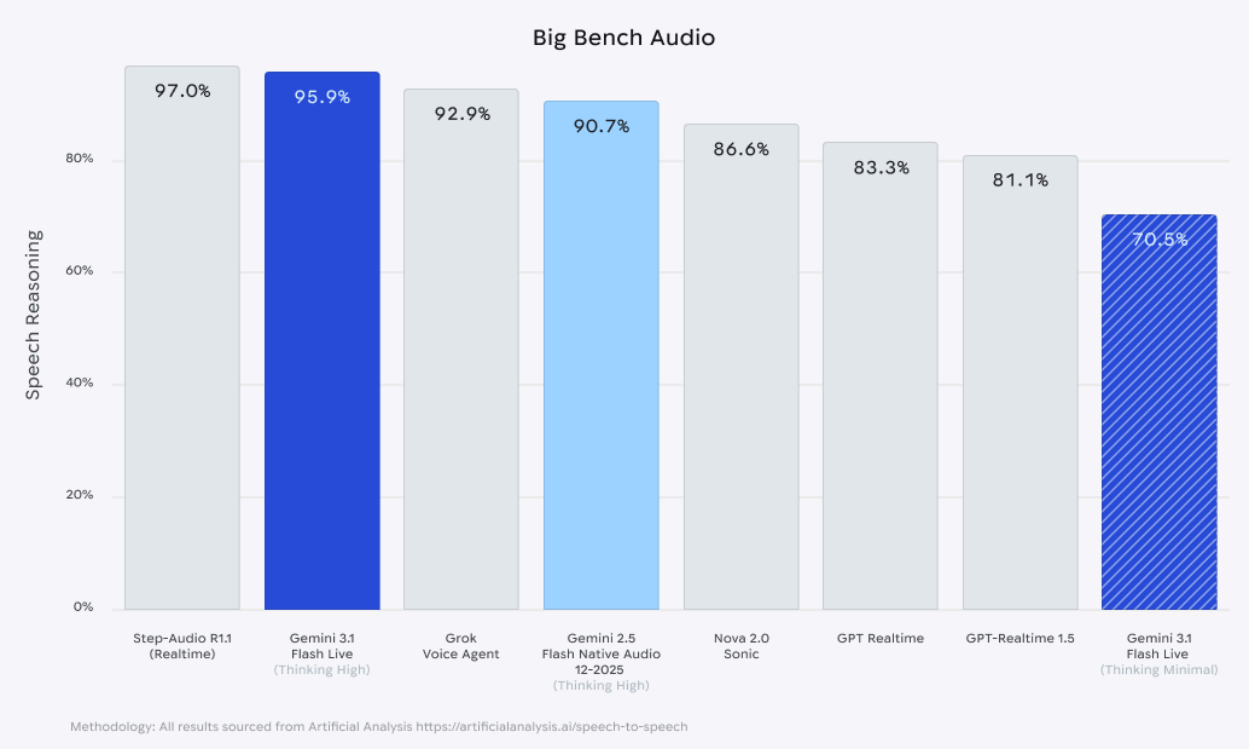
Live Scores available:

Scale AI Multi Challenge Leaderboard



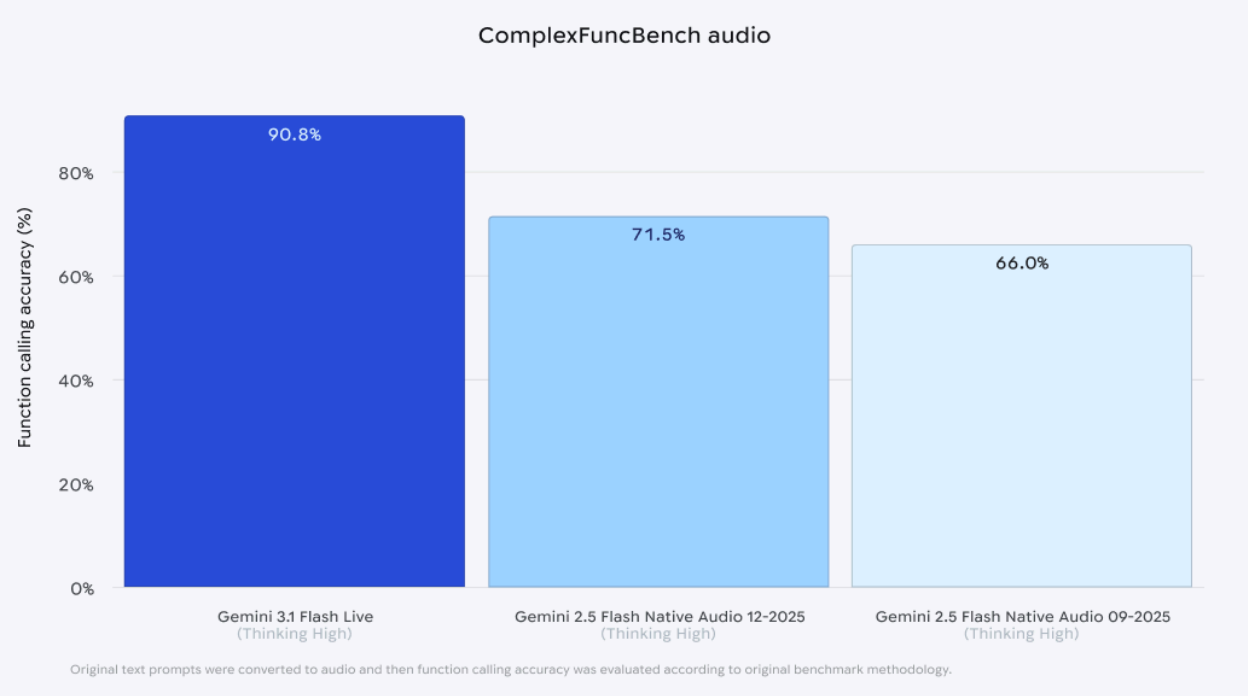
Model	Overall Score	Gemini 2.5 Flash Native Audio Preview 12-2025
Gemini 3.1 Flash Live (Thinking = High)	36.1%	21.5%
Gemini 3.1 Flash Live (Thinking = Minimal)	26.8%	13.9%

[Artificial Analysis Speech-to-Speech Leaderboard]



Model	Overall Score	Gemini 2.5 Flash Native Audio Preview 12-2025
Gemini 3.1 Flash Live (Thinking = High)	95.9%	90.7%
Gemini 3.1 Flash Live (Thinking = Minimal)	70.5%	69%

ComplexFunBench Audio



Model	Overall Score	Gemini 2.5 Flash Native Audio Preview 12-2025
Gemini 3.1 Flash Live (Thinking = High)	90.8%	71.5%