Google Gemini 3.8 Live Models Add Background Tasks to Voice Conversations

TechRounder PDF Edition

Live article:
https://www.techrounder.com/ai/google-gemini-3-8-live-models-add-background-tasks-to-voice-conversations/

By Vipin PG | Published September 16, 2026 | Updated September 16, 2026 | Format: Analysis | 4 min read

In brief

Google has introduced Gemini 3.8 Live and Gemini 3.8 Live Extended Thinking, two audio models designed to reason, call tools and keep talking during a live voice session.

Google has introduced Gemini 3.8 Live and Gemini 3.8 Live Extended Thinking, two audio models designed to reason, call tools and keep talking during a live voice session. The announcement was made on September 15, 2026, and the models are rolling out through the Gemini API, Google AI Studio, Gemini Live and selected Google products.

The release targets a problem that has limited many voice assistants: a conversation often has to stop while the system completes a search, API request or multi-step task. Google's new models can continue streaming a spoken response while work happens in the background.

What Gemini 3.8 Live adds

Google describes Gemini 3.8 Live as the model for scale and cost efficiency. It supports audio, images, video and text as inputs, and can return audio and text. The model can use visual context during a conversation, so a voice agent can respond to what a user says and what its camera or other visual input shows.

It can also make asynchronous function calls. A developer can let the model call an external service while the conversation continues, rather than blocking the user until the service responds. That could help with voice interfaces for customer support, scheduling, troubleshooting and other workflows that need both dialogue and actions.

Google says Gemini 3.8 Live can detect and switch between 97 supported languages during a conversation. The developer announcement also points to better handling of alphanumeric information such as confirmation codes, claim numbers and technical data.

Extended Thinking lets the model reason while speaking

Gemini 3.8 Live Extended Thinking is aimed at more complicated requests. Google says it can perform background reasoning and asynchronous tool calls while streaming continuous audio. It can also narrate progress with short verbal cues while it works through a task.

The Gemini API documentation describes a different integration model for developers. Receiving 'turnComplete: true' does not necessarily mean the server is finished. The client must continue listening for later audio, reasoning updates or tool calls until the server reports an idle state through 'interaction_status'.

Function calling is asynchronous only. Synchronous blocking calls are not supported, and developers can configure the thinking level as low, medium or high. Google's documentation also says caching, code execution, file search, structured outputs, URL context and Google Maps grounding are not supported for the Extended Thinking model.

Where users and developers can try it

Google says both models are available through the Gemini API and Google AI Studio. Gemini 3.8 Live is also rolling out in Search Live. The Extended Thinking model is beginning to appear in Gemini Live and in some Workspace experiences, including Docs, Gmail and Keep, depending on the user's plan and region.

For enterprises, the models are in private preview in Gemini Enterprise, with additional customer experience and Workspace availability planned. Google also lists integrations with platforms such as LiveKit, LangChain, Pipecat, Vercel, Agora and Fishjam, which handle parts of the real-time media infrastructure.

Google's developer announcement lists pricing of $0.005 per minute for audio input and $0.018 per minute for audio output. Actual application costs will also depend on how long users speak, how much audio the system returns and how frequently the agent invokes external tools.

What the early benchmark claims mean

Google says Gemini 3.8 Live Extended Thinking recorded an 82.6 score on Artificial Analysis' Speech to Speech Quality Index. It also reports scores of 68.6% on the ?-Voice agentic task benchmark, 35.1% on Sierra's ?-Voice-banking benchmark and 97.7% on Big Bench Audio. Gemini 3.8 Live is described as second in the Speech Agent Arena.

These figures come from Google's launch material and apply to specific test conditions. They show the areas Google chose to measure, while application performance will also depend on latency, interruption handling, microphone quality, tool reliability and the design of the surrounding application.

What developers still need to handle

Google's model card says Gemini 3.8 Live and Gemini 3.8 Live Extended Thinking can still produce hallucinations. It also warns about occasional slowness or timeouts. The model card lists January 2025 as the knowledge cutoff, so a live conversation does not automatically mean the model has current knowledge unless the application supplies search or other external data.

The same model card says Google did not find meaningful new capabilities or material performance increases over Gemini 3.7 Flash for these models. Based on that assessment, Google says the models are not likely to reach its Tracked or Critical Capability Levels. The assessment covers those capability thresholds; voice agents can still behave incorrectly in production.

For developers, the main change is architectural. A voice agent built around the new model must treat a conversation as an ongoing stream of audio, state updates and background work. An app that closes a session as soon as it receives 'turnComplete: true' could miss a later tool result or response.

Gemini 3.8 Live combines speech, visual context and background actions in one service. Developers still need to set permissions, handle tool failures, provide fallbacks and show users when the system is thinking or acting.