

Build real-time voice applications with Gemini 3.8 Live and 3.5 Transcribe

15.9.2026 - Alisa Fortin, Thor Schaeff | Google Italia

New Gemini Audio models are available for developers to build more intelligent conversational experiences via the Gemini API and Google AI Studio.

Today, we released new Gemini Live models in the Gemini API and Google AI Studio, expanding our developer suite for building real-time, voice-first product experiences:

Gemini 3.8 Live and 3.8 Live Extended Thinking: Gemini 3.8 Live brings a step change to our native speech-to-speech models, capable of performing tasks while maintaining dialogue. For complex requests, 3.8 Live Extended Thinking delivers deeper reasoning, ranking #1 on Artificial Analysis' leaderboard.

Gemini 3.5 Transcribe: Our dedicated speech-to-text model brings highly precise transcription across 85+ languages. Released last month, it achieved an average Word Error Rate (WER) of 4.0% (streaming) and 2.6% (non-streaming).

Gemini 3.8 Live & 3.8 Live Extended Thinking: Build more intelligent conversational agents

Our new models, Gemini 3.8 Live and 3.8 Live Extended Thinking enable developers to build voice agents that reason and execute tasks while maintaining the flow of conversations. Key capabilities include:

- Asynchronous function calling: Execute API and tool calls in the background while continuing to stream audio responses to the user
- Visual context: Ground dialogue in live visual inputs to enable agents that understand what users say *and* see
- Alphanumeric precision: Accurately parse confirmation codes, claim numbers, and technical data
- Multilingual support: Reach global audiences with coverage for 97+ languages and accent consistency
- Incremental content updates: Seamlessly merge real-time audio with structured data to return context-aware responses

3.8 Live Extended Thinking also supports configurable thinking to handle complex, multi-step reasoning in the background, while responding or narrating its progress in the main conversation. These models represent a step-change from our previous live models and provide a more streamlined alternative to cascaded architectures.

Gemini 3.8 Live and Gemini 3.8 Live Extended Thinking are available via the Live API. Competitively priced at \$0.005/min for audio input and \$0.018/min¹ for audio output, they allow developers to scale voice applications with industry-leading performance.

Developers can also access the models through Agora, Fishjam, LiveKit, Pipecat, Vercel, and Vision Agents, Live API integration partners that handle media streaming infrastructure for real-world deployment:

Gemini 3.5 Transcribe: Convert streamed speech to text

Real-time speech understanding is critical for voice-first interfaces. Last month, we released Gemini 3.5 Transcribe for low-latency transcription with high precision, achieving a 4.0% WER, and useful features:

- Automatic code-switching: Handle intra-sentence and inter-sentential code- and language-switching without manual configuration
- Custom vocabulary biasing: Steer speech recognition toward domain-specific terms, uncommon jargon, company names, and proper nouns by passing a `custom_vocabulary` list of up to 1,000 terms
- Smart transcription mode: Deliver polished, reader-ready transcripts with structured formatting, self-corrections, and disfluency removal that eliminates filler words

3.5 Transcribe supports 85+ languages and provides a strong listening engine for voice experiences and stateless tasks like sub-second captioning, call center agents, and real-time audio analytics. You can also access the model via the Interactions API to transcribe audio files up to 1 hour long with structured timestamps and speaker diarization. Read our developer guide to learn more.

Our complete audio suite for developers

To get started, try out the models in [ai.studio/live](https://ai.google.dev/studio/live), clone example apps from GitHub, or equip your agent with our live api skill.

You can also create audio experiences with our speech and music generation models, all available in the Gemini API:

- Gemini 3.5 Live Translate: Speech-to-speech translation across more than 70 languages
- Gemini 3.1 Flash TTS: Highly configurable speech generation (with more updates coming soon)
- Lyria 3.5: Production-grade music generation

The mic is yours, and we can't wait to hear what you build!

<https://blog.google/innovation-and-ai/technology/developers-tools/build-real-time-voice-applications-gemini-audio>