

Introducing Gemini 3.8 Live and 3.8 Live Extended Thinking

15.9.2026 - Tom Ouyang, Malini Jaganathan | Google Italia

Gemini 3.8 Live and Gemini 3.8 Live Extended Thinking are our most advanced live dialogue models yet. Major upgrades in intelligence and parallel reasoning make them more intuitive to collaborate with and use to execute complex tasks using your voice.

Today, we're introducing two new models that bring advancements in near real-time reasoning to more effectively enable voice agents and make conversing with AI feel more intuitive and intelligent.

- Gemini 3.8 Live: Built for scale and cost efficiency, combining conversational intelligence with fluid dialogue and visual grounding.
- Gemini 3.8 Live Extended Thinking: Built for high-complexity tasks, with increased intelligence and multi-step reasoning.

For developers and enterprises, these models deliver the building blocks for reliable, production-ready voice agents. They also make speaking with Gemini across the Gemini app, Google Workspace, and Search more fluid and collaborative — helping you tackle complex tasks using just your voice.

Experience more fluid, intelligent conversations

Gemini 3.8 Live Extended Thinking provides enterprise-grade task completion and intelligence, capturing the #1 overall spot on Artificial Analysis' Speech to Speech Quality Index (82.6), and leads in agentic task completion with 68.6% on τ -Voice and 35.1% on Sierra's τ -Voice-banking benchmark. It also provides strong reasoning capabilities, scoring 97.7% on Big Bench Audio, while maintaining a highly competitive price point compared to other frontier models.

Gemini 3.8 Live has shown a high preference among users, securing a second place in the Speech Agent Arena. In addition to this performance, it remains highly cost-effective — providing developers and enterprises with a capable and efficient model built for scale.

On ServiceNow's EVA-Bench, a benchmark for evaluating voice agents, our models push the Pareto Frontier for complex workflows by successfully balancing accuracy with conversational quality.

Note: This was run on the Live API on Gemini Enterprise Agent Platform.

Gemini 3.8 Live processes visual inputs in near real-time, enriching conversations with context for more helpful responses. It automatically detects and transitions between 97 supported languages mid-conversation. It executes tools and API calls in the background while continuing the conversation, so the model can acknowledge requests and keep chatting while tasks finish in the background.

Gemini 3.8 Live guides employee onboarding in real time, using visual context to answer live questions.

Watch Gemini 3.8 Live play chess in near real-time using visual context, reasoning, and natural conversational flow.

For tasks that require deeper reasoning, 3.8 Live Extended Thinking reasons and speaks

simultaneously. It delivers increased intelligence for complex workflows while maintaining an uninterrupted conversational flow — using early verbal cues like “*Let me check that...*” to acknowledge prompts naturally, and live progress narration to walk users through multi-step background tasks as they progress.

Watch Gemini 3.8 Live Extended Thinking transform raw sketches and near real-time voice feedback into functional React components.

See Gemini 3.8 Live Extended Thinking coordinate multi-step bookings and asynchronous function calls — all without interrupting natural live conversation.

Watch Gemini 3.8 Live build complete business plans and custom marketing toolkits on the fly through natural speech.

Across Google Workspace and Search, our Live models deliver more intuitive, collaborative experiences — especially when tackling your most complex tasks.

Try Gemini 3.8 Live Extended Thinking in Google Workspace with Docs Live, Gmail Live, and Keep Live.

Empowering the developer and enterprise voice ecosystem

By using the Gemini Live API, developer platforms such as Agora, Fishjam, LiveKit, Pipecat, Vercel, and Vision Agents enable developers to build and deploy high-performance voice-driven interfaces with ease. These platforms manage complex real-time media streaming infrastructure behind the scenes, allowing developers to focus entirely on crafting the user experience.

We’re also partnering with companies like Salesforce, Genspark, and Lumeris who are excited about 3.8 Live and 3.8 Live Extended Thinking, highlighting its impressive latency, fluidity, and tool-calling capabilities.

Ensure transparency with SynthID watermarking

All audio generated by our AI products is watermarked with SynthID. This imperceptible watermark is woven directly into the audio output, ensuring AI-generated content remains detectable to help prevent misinformation. For details on our approach to safety and responsibility, review the model card.

Start using our latest Gemini Audio models:

3.8 Live is rolling out starting today:

- For developers: In the Gemini API and Google AI Studio
- For enterprises: In private preview in Gemini Enterprise and coming soon to Gemini Enterprise for Customer Experience
- For everyone: In Search Live

3.8 Live Extended Thinking is rolling out starting today:

- For developers: In the Gemini API and Google AI Studio
- For enterprises: In private preview in Gemini Enterprise and coming soon to Gemini Enterprise for Customer Experience and Google Workspace business customers

- For everyone: In Gemini Live and for Google AI Pro and Ultra subscribers in Workspace in Docs, and all Google AI subscribers in Gmail and Keep

<https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-8-live-gemini-3-8-live-extended-thinking>