

# Mercury Voice: A Voice AI That Reasons Without the Awkward Pause

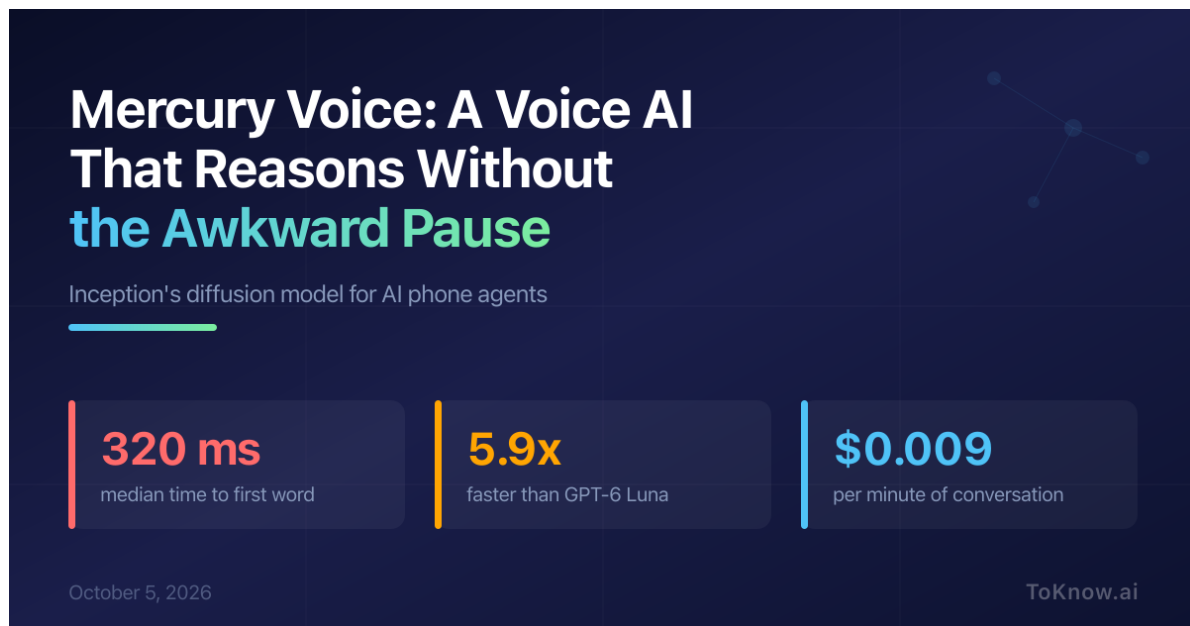
Kabui, Charles

2026-10-05

---

[Read at ToKnow.ai](#)

---



Inception released Mercury Voice, a model built to power AI phone agents and voice assistants. It is a diffusion model, which means it does not write one word at a time the way most chatbots do. Instead it starts with a rough draft and refines the whole answer at once, so the first word is ready almost immediately. On real customer-service prompts, Inception measured a median 320 milliseconds to that first answer word, with the slowest 5% under 750 milliseconds, 5.9

times faster than GPT-6 Luna with reasoning turned off. The company says Mercury Voice also beats Gemma 4 31B, GLM-5.3-Flash, and Qwen3.5-397B across tool-use and conversation tests. It holds 128,000 tokens of context, runs at three reasoning levels, and ships as a drop-in, OpenAI-compatible API.

In a voice agent, the pause after you stop talking is the whole product. Wait past about half a second and the call feels broken, so most voice bots use fast models that skip real reasoning. Mercury Voice reasons and still answers inside that window. Audivi AI runs drive-thru ordering on it, Altur handles bank payment calls, and OpenCall reported a median response near 170 milliseconds in production. Launch pricing is \$0.20 per million input tokens and \$0.75 per million output, about \$0.009 per minute of conversation.

The mechanism is architectural. Standard models write one word at a time, so more thinking means more waiting. Diffusion models generate many words in parallel, which is how Mercury Voice can reason and still stay fast.

Read More: [Gemini 3.8 Live: Voice Models That Think While They Talk](#)

Sources:

- [Introducing Mercury Voice \(Inception\)](#)
- [Introducing Mercury 2.5 \(Inception\)](#)
- [Our Models: Mercury pricing and specs \(Inception\)](#)
- [Inception Launches Mercury 2.5, the Next Tier of Intelligence for Diffusion LLMs \(Business Wire\)](#)

---

**Disclaimer:** For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation. **Read more:** [/terms-of-service](#)