servicenow

ServiceNow Voice AI Model Cards

OpenAI Whisper (Cartesia Ink-Whisper Variant)

Cartesia (Sonic)

Model Name

Ink-Whisper (based on Whisper large-v3)

Model Developer

OpenAI (base model)

Enhanced by

Cartesia

Version

Ink-Whisper (Cartesia's optimized variant of Whisper large-v3)

License

- Base Model: MIT License (OpenAI)
- Implementation & Support: Please refer to your agreement with ServiceNow for all license information

Usage in Our System

Ink-Whisper, a variant of OpenAI's Whisper, specifically optimized for low-latency transcription in conversational settings, is utilized within ServiceNow AI Voice Agents as the Speech-to-Text (STT) engine. This implementation processes spoken user input and converts it to text for analysis by our AI orchestration layer.

Ink-Whisper is the fastest, most affordable STT model—designed for enterprise-grade voice agents.

Key Capabilities

Base Whisper Capabilities

- Transcribes audio into the same language or translates into English
- Supports 57+ languages with high accuracy (trained on 98 languages)
- 1.55B parameter model size for comprehensive language understanding
- Zero-shot capabilities for handling diverse speech patterns

Cartesia Enhancements:

- Dynamic chunking for variable-length audio segments (vs. standard 30-second chunks)
- Optimized for real-time conversational AI with ultra-low latency
- Enhanced accuracy in challenging conditions
- Telephony artifacts: Low-bandwidth, compressed audio adds distortion
- Proper nouns and domain terms: Names of products, drugs, or financial instruments
- Background noise: Traffic, restaurant chatter, crying babies, and static
- Disfluencies and silence: Fillers like "um" and pauses
- Accents and variation: Voices come in all kinds
- Fewer errors and less hallucination, especially during silence or audio gaps

Model Name
Sonic-2

Model Developer
Cartesia

Version
Sonic-2-2025-03-07

Source
Cartesia Sonic Documentation

License
Please refer to your agreement with ServiceNow for all license information.

Usage in Our System
Sonic-2 is leveraged within ServiceNow AI Voice Agents as the Text-to-Speech (TTS) engine, converting text responses from our AI orchestration layer into natural-sounding speech output.

The model provides ultra-realistic, expressive voice generation with extremely low latency (as low as 40ms), enabling fluid, human-like conversation experiences across enterprise workflows.

Key Capabilities
- Supports 15 languages: English, French, German, Spanish, Portuguese, Chinese, Japanese, Hindi, Italian, Korean, Dutch, Polish, Russian, Swedish, and Turkish
- High-fidelity voice cloning for customized voice experiences
- Controllable speech parameters (speed, tone, emotion)
- Timestamps for precise audio-text alignment
- Infill support for seamless speech editing