A FIELD GUIDE TO GIVING AI AGENTS A VOICE

TALKING AGENTS

From Audrey in 1952 to Alexa+, Siri AI and real-time phone agents: how voice AI works, what it costs, and how we're wiring a voice and a phone number into hank, our always-on agentic computer.

Get OpenClaw, Hermes and Claude Dispatch talking.

`OpenClaw` `Hermes Agent` `Claude Dispatch` `+ any agent with an API`

One voice layer, one locked relay, any brain behind it.

PART I
The world of talking machines

PART II–III
How it works and what to buy

PART IV–V
The hank build and the rules

CONTENTS

From the world to one machine

I · The world of talking machines

Seventy years in one page 03
Why Siri and Alexa stalled 04
The 2024–2026 turn 05
Follow the money 06

II · How a talking agent works

The 200-millisecond rule 07
Two ways to build a voice 08
The hard parts 09

III · Choosing a stack

Three paths 10
Managed platforms, priced 11
Ears, mouths and brains 12
Open source and phone lines 13
What a minute really costs 14

IV · The hank build

Meet hank 15
The split brain 16
Never open the front door 17
A local voice on Strix Halo 18
The hank voice roadmap 19

V · Rules, gotchas, decisions

Rules of the road 20
Field-tested gotchas 21
Pick your path 22
Glossary 23
Sources 24

PRICES

Checked on vendor pages October 9, 2026. They move monthly; re-check before buying.

STATUS TAGS

LIVE means running on hank today. PLANNED means designed, not yet built.

NOT ADVICE

Part V summarizes laws so you know what to ask. It is not legal advice.

SEVENTY YEARS IN ONE PAGE

Machines have listened since 1952

For sixty years they could only hear commands. Since 2024 they can hold a conversation.

1952 **Audrey.** Bell Labs builds a machine that recognizes spoken digits from one speaker.

1962 **IBM Shoebox.** Sixteen words, demonstrated at the Seattle World's Fair.

1976 **Harpy.** Carnegie Mellon reaches 1,011 words under DARPA's speech program.

1990 **DragonDictate.** Speech recognition becomes a product you can buy.

1997 **NaturallySpeaking.** Continuous dictation on an ordinary PC.

'90s **"How May I Help You?"** AT&T Labs routes phone calls from callers' open-ended answers. The IVR era begins.

2007 **Siri Inc.** Spins out of SRI's DARPA-funded CALO assistant project.

2011 **Siri on iPhone 4S.** Apple bought Siri in 2010. The voice assistant goes mass-market.

2014 **Echo and Alexa.** Amazon puts a far-field microphone in the kitchen.

2016 **Google Assistant.** Pitched at I/O as "an ongoing two-way dialogue."

2018 **Google Duplex.** An AI books a hair salon by phone. Backlash forces built-in disclosure.

2022 **Whisper and Nuance.** OpenAI releases Whisper. Microsoft closes its $19.7B buy of Nuance.

2024 **Speech-to-speech.** GPT-4o voice, then OpenAI's Realtime API (Oct 1). Kyutai's Moshi listens and talks at once.

2025 **Alexa+ and Sesame.** Generative Alexa is announced. Sesame's demo voice goes viral for sounding human.

2026 **The rebuild year.** Alexa+ reaches all US users (Feb). Gemini replaces Assistant on Android (Sep 4). iOS 27 ships Siri AI (Sep 14).

The pattern: recognition kept getting better, but until 2024 the brain behind it was a rulebook. The rest of this guide is about what replaces it.

WHY SIRI AND ALEXA STALLED

Half a billion speakers, mostly setting timers

COMMAND-AND-CONTROL ASSISTANT · 2011–2023

Wake word › Speech to text › Intent classifier › Slot filling › Hand-coded handler › Canned reply

AGENTIC VOICE · 2024 ONWARD

Listen continuously › Language model understands the request › Calls tools: calendar, CRM, email › Answers in its own words

500M+
Alexa devices sold, per Amazon.

~$10B
Projected 2022 loss for the Amazon unit that includes Alexa (Business Insider).

48%
of smart-speaker owners had ever used a third-party voice app (Voicebot, 2020).

Legacy Alexa ran on "a collection of small machine learning models and thousands of hand-crafted and hard-coded rules," insiders told Fortune. Every request had to match an intent somebody had built ahead of time. Music, Home and Shopping each lived in a separate domain. Improving one could break another.

The result was a billion interactions a week that were hard to make money from: music, weather, timers. Siri hit the same wall. Its interim iOS 18 build was a "grafted-together" system with two brains. Both companies have now rebuilt their assistants on large language models.

LESSON FOR YOUR AGENT

Open conversation and reliable commands are different jobs. Google's Gemini switchover was reportedly held up partly because it fumbled basic commands like timers. A phone agent needs both: free-form talk, plus a short list of actions that never fail.

THE 2024–2026 TURN

Three breakthroughs in two years

Rulebook assistants could only map speech to a fixed menu. Three advances turned voice into something an agent can actually talk with.

01

Speech-to-speech

One model hears audio and speaks audio, with no text in between. OpenAI's Realtime API reached general release on Aug 28, 2025, with phone (SIP) calling built in. Google and Amazon followed.

02

Full duplex

Kyutai's open-source Moshi (2024) listens and talks at the same time, the way people do. It answers in about 200–240 ms. Interruptions stop being a crash and become part of the conversation.

03

Voices that pass

Sesame's 2025 demo, "Crossing the uncanny valley of voice," went viral. ElevenLabs' newest voices start speaking in about 75–100 ms. A synthetic voice no longer gives itself away.

THE GIANTS REBUILD · STATUS OCTOBER 2026

Alexa+	All US customers since Feb 4, 2026. Free with Prime, $19.99/mo without. Runs on Amazon Nova and Anthropic models. Books restaurants, rides and home services.
Siri AI	Shipped in iOS 27 on Sep 14, 2026. The release is a beta, in English only, on "an entirely new architecture." Apple and Google said in January that Apple's next models are built on Gemini.
Gemini	Replaced Google Assistant on Android phones, tablets and watches starting Sep 4, 2026. Speakers, TVs and cars come later.
Developer APIs	OpenAI gpt-realtime-2.1 (Jul 2026), Google Gemini Live API, and Amazon Nova 2.5 Sonic (Oct 5, 2026). You can rent all three by the minute.

The voice technology Apple, Amazon and Google rebuilt their assistants on is now available to a one-person business for a few cents a minute.

FOLLOW THE MONEY

Why voice is a big deal

17M+

contact-center agents worldwide (Gartner, 2022). Other counts run as high as 29M seats.

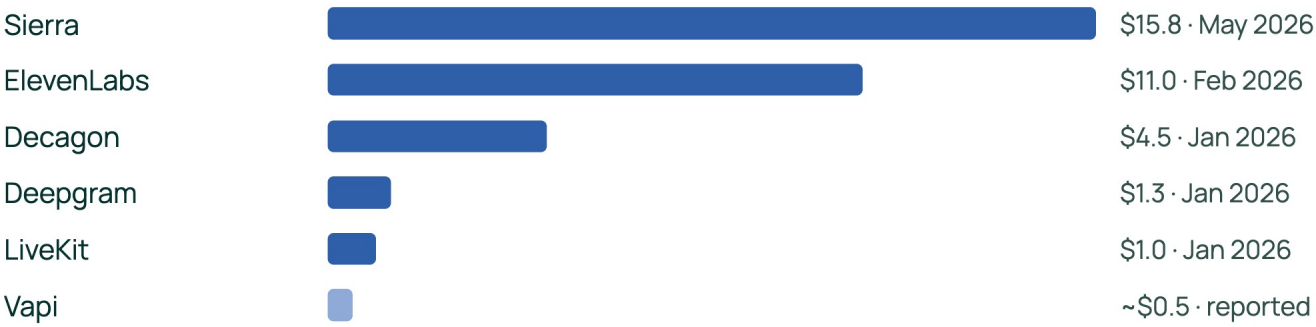
$8.01

average cost of one live-agent contact by phone, chat or email. Self-service averages $0.10 (Gartner, 2019).

5–15¢

per minute for a working AI phone agent in October 2026, all-in. Page 14 shows the math.

LATEST VALUATIONS, VOICE AND CUSTOMER-SERVICE AI · $ BILLIONS



ElevenLabs reported over $330M in annual recurring revenue at the end of 2025. Bland handles over 3.5M calls a week.

The forecasts

One firm puts voice AI agents at $2.4B in 2024, rising to $47.5B by 2034. Another puts conversational AI at $49.9B by 2030. These are paid research reports with different definitions. Read them as direction, not data. The investment above is real money.

The builders

Y Combinator funded 90 voice-agent companies from 2020 to early 2025. By early 2025, 22% of a new YC batch was building with voice. About 69% of YC's voice companies sell to businesses, not consumers.

WHERE THE OPPORTUNITY IS

The platforms sell minutes. Most small and mid-size businesses can't connect that voice to their own calendar, CRM and phone system safely. That integration is the gap, and filling it is what this guide teaches.

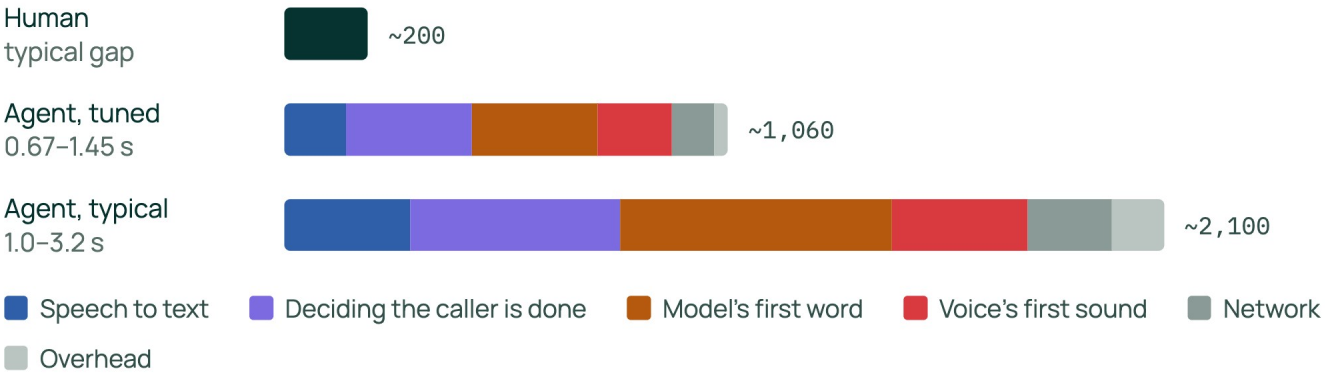
THE 200-MILLISECOND RULE

People answer in a fifth of a second

A 2009 study of ten languages found that people leave very short gaps between turns. The most common gap is 0–200 ms. Averages run from 7 ms in Japanese to 468 ms in Danish. Planning even one word takes about 600 ms, so we start building our answer while the other person is still talking.

Machines can't match that yet, and callers forgive some delay. Industry targets aim for a reply 600–1,200 ms after the caller stops. Past roughly 600–800 ms, a pause starts to sound like reluctance. Steady timing matters more than a fast average.

WHERE A VOICE AGENT'S REPLY TIME GOES · MIDPOINTS IN MILLISECONDS



Stage ranges from Hamming AI's 2026 latency study. A phone-network hop typically adds another 100–200 ms.

Knowing when they're done

The biggest hidden cost. Wait too long and the agent feels slow. Jump in too early and it interrupts.

The model's first word

Big reasoning models are slow to start. Keep a fast model on the call and send hard work elsewhere (page 16).

Streaming everything

Start speaking the first sentence while the rest is still being written. Never wait for the full answer.

TARGETS · SILENCE AFTER THE CALLER STOPS, 0 TO 2.4 SECONDS



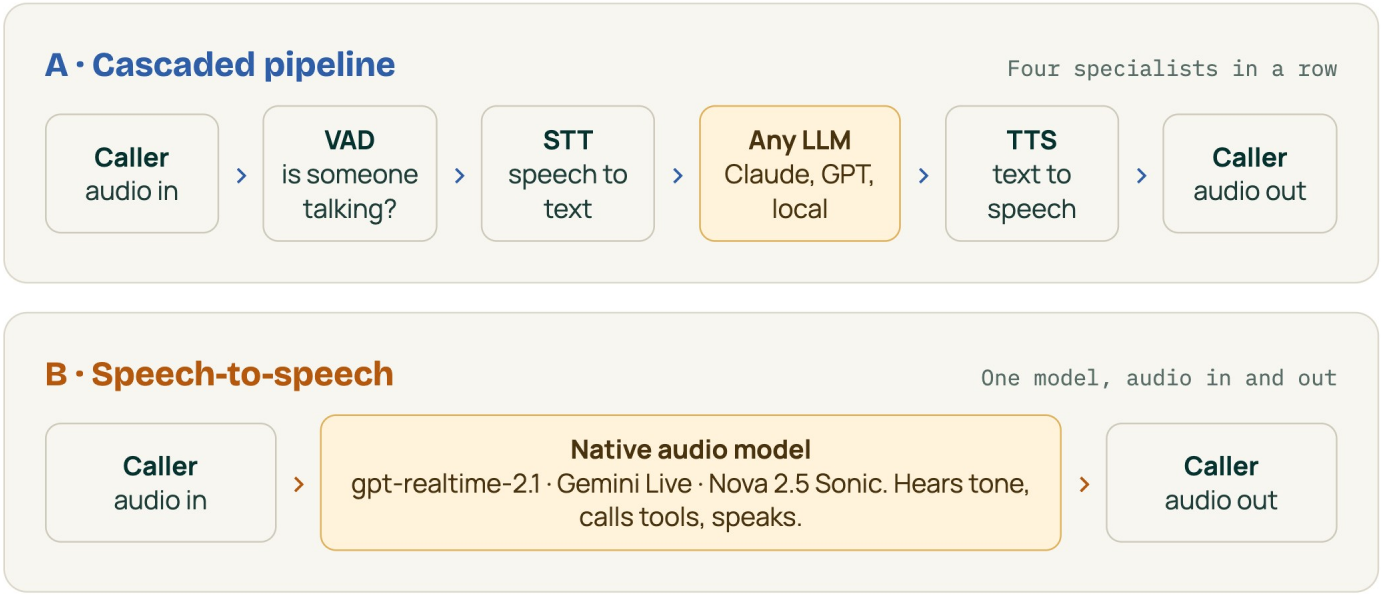
0–600 ms
Close to human

600–1,200 ms
Industry target band

Over 1,200 ms
Beyond target: the pause starts to sound like hesitation

TWO WAYS TO BUILD A VOICE

A relay team, or one model that does it all



	A · CASCADED	B · SPEECH-TO-SPEECH
Speed	Each handoff adds delay. About 0.7–1.5 s when tuned.	Fastest. No text handoffs in the loop.
Choice of brain	Any model, including your own local one.	Whatever model the vendor ships.
Tone and emotion	Lost when speech becomes text.	Hears sighs, hurry and hesitation.
Debugging	Every step leaves a text log.	Harder to control exact wording.
Cost	Shop each part. Usually cheaper.	Long calls grow costly as the transcript piles up.
Best for	Agents with a brain of their own, like hank.	Fast, natural front desks and companions.

Our pick for hank: cascaded. hank's value is its memory, tools and local models. Speech-to-speech would hand the conversation to someone else's brain. A cascade keeps hank thinking while a rented voice does the talking.

THE HARD PARTS

The voice is easy. Taking turns is hard.

Good voices are a commodity now. What separates a demo from a phone agent people tolerate is these four problems.

1 Knowing when to talk

A voice activity detector (VAD) only hears sound or silence, and a pause is not the end of a thought. Turn-detection models listen to *how* a sentence ends.

Silero VAD: free, about 2 MB, under 1 ms per audio chunk on a CPU. Pipecat Smart Turn v3.1: open, 23 languages, about 12 ms on a CPU. Deepgram Flux builds turn detection into its speech recognition.

2 Being interrupted

When the caller talks over the agent ("barge-in"), it must stop at once and drop any audio already queued. Then it must remember only what the caller actually heard.

On Twilio Media Streams, a "clear" message flushes queued audio and "mark" messages track playback. Twilio ConversationRelay has a setting to ignore backchannel words like "yeah" and "uh-huh."

3 Phone-quality audio

The phone network carries 8 kHz mono audio in μ-law encoding (G.711). Most voice models expect 16–24 kHz. Every call needs resampling in both directions.

Twilio sends base64 μ-law frames over a WebSocket. Audio you send back must have no file header. A stray WAV header plays as a click or as silence.

4 Never dead air

Silence on a phone line sounds like a dropped call. Any step that might take a while needs a spoken bridge ("Let me check that") and a timeout with a graceful line.

If the backend dies, the agent should say so and offer a callback or a person. It should never just hang up.

AND ECHO

On phone lines the carrier mostly handles echo. In browsers and apps you rely on WebRTC's built-in echo cancellation. When it fails, the agent hears its own voice and interrupts itself. Many "false barge-ins" are really echo.

THREE PATHS

Rent it, assemble it, or own it

PATH 1 · RENT

Managed voice platform

The vendor runs listening, turn-taking, voice and phone number. You write a prompt and point it at your own model or tools.

ElevenLabs Agents, Vapi, Retell, Bland, Synthflow, Telnyx Voice AI

PATH 2 · ASSEMBLE

Open framework, rented parts

You run the conversation loop yourself with open-source code. You choose each ear, brain and mouth, and connect a phone carrier.

Pipecat or LiveKit Agents + Deepgram, Cartesia or ElevenLabs + Twilio or Telnyx

PATH 3 · OWN

Fully local on your hardware

Speech recognition, model and voice all run on your own machine. Only the phone call itself leaves the building.

whisper.cpp or Parakeet + Ollama + Kokoro or Piper, wired with Pipecat

HOW THEY COMPARE · OUR ASSESSMENT, FIVE DOTS = BEST

	RENT	ASSEMBLE	OWN
Speed to first call	●●●●●	●●●○○	●●○○○
Cost per minute at volume	●●○○○	●●●●○	●●●●●
Voice quality today	●●●●●	●●●●●	●●●○○
Control and customizing	●●○○○	●●●●○	●●●●●
Privacy of conversations	●●○○○	●●●○○	●●●●●
Low upkeep	●●●●●	●●○○○	●○○○○

EVERY PATH RENTS A PHONE LINE

"Self-hosted" means your own software, not your own phone wires. You still need Twilio, Telnyx or another carrier to reach the phone network.

MIXING IS NORMAL

hank's plan rents the voice (Path 1) but keeps the brain local (Path 3). Most good builds are hybrids. Page 19 shows how.

MANAGED PLATFORMS, PRICED · OCTOBER 9, 2026

The sticker price is only part of the bill

PLATFORM	PLATFORM FEE	YOU ALSO PAY · NOTES
ElevenLabs Agents	$0.08/min	LLM and telephony at cost. Plans $6–$990/mo. 4 to 40 calls at once depending on plan. Best-known voices.
Vapi	$0.05/min	Speech, model, voice and carrier passed through at cost. Vapi's own estimate for 1,000 minutes: $82–$129 total.
Retell AI	$0.055/min	Voice $0.015–0.10, LLM, telephony $0.015. Typically about $0.11/min all-in. 20 calls at once free.
Telnyx Voice AI	$0.05/min	Includes speech and voice. Runs on its own carrier network. About $0.06/min all-in by Telnyx's math. Numbers $1/mo.
Twilio Conversation Relay	$0.07/min	Plus normal Twilio voice minutes. Twilio handles speech and voice. Your server only sends and receives text.
Deepgram Voice Agent	$0.075/min	Bundled ears, brain and voice. $0.05/min if you bring your own LLM and voice.
Bland *	$0.11–0.14	Lower rate on $299–$499/mo plans. Built for high-volume outbound. $0.015 minimum per outbound attempt.
Synthflow *	$0.09/min	Plus LLM $0.02–0.05 and telephony $0.02. Five calls at once included.
LiveKit Cloud	$0.01/min	Beyond plan minutes (1,000 free). Hosts your own open-source agent. Models and phone line extra.

Platform fees from each vendor's pricing page, October 9, 2026. * Figures from a competitor's published comparison; confirm with the vendor.

Count the add-ons

The model, the voice and the carrier are usually billed separately. All-in, most stacks land between $0.06 and $0.15 a minute.

Check call limits

Every plan caps how many calls run at once. Extra lines cost $8–$20 each per month on several platforms.

Vendors disappear

Hume shuts down its voice APIs on Nov 13, 2026. Build so you can swap any one piece without starting over.

EARS, MOUTHS AND BRAINS · OCTOBER 9, 2026

The parts, if you assemble your own

Ears · streaming speech-to-text

Deepgram Nova-3	$0.0048/min	Promotional rate; list price $0.0077.
Deepgram Flux	$0.0065/min	Detects when the caller is done.
AssemblyAI Universal-Streaming	$0.15/hr	About $0.0025/min. Six languages.
ElevenLabs Scribe v2 Realtime	$0.39/hr	About 150 ms, per ElevenLabs.
OpenAI gpt-4o-transcribe	Retiring	Ends Feb 26, 2027. Successor $0.017/min.

Mouths · text-to-speech

ElevenLabs Flash v2.5	$0.04 / 1k chars	About 75 ms to first sound.
Cartesia Sonic 3.5	~$0.04–0.05 / 1k	351 ms (p90) in an Aug 2026 benchmark.
Rime Mist v3	$0.03 / 1k chars	Claims 37 ms median to first audio.
Deepgram Aura-2	$0.030 / 1k chars	Same vendor as Nova-3 and Flux.
Inworld TTS-2	$0.025 / 1k chars	$25 per million characters.

All-in-one brains · speech-to-speech

OpenAI gpt-realtime-2.1	~$0.05/min	Has SIP calling. Mini about $0.016/min.
Google Gemini Live API	~$0.01–0.02	$0.005 listening, $0.018 speaking.
Amazon Nova 2.5 Sonic	Bedrock rates	Oct 5, 2026. Plugs into Twilio, Connect.

Rule of thumb: an agent that talks 40% of a call speaks about 360 characters per call-minute. At $0.04 per 1,000 characters, that is about $0.015 a minute. Per-minute figures for the all-in-one models are our estimates.

OPEN SOURCE AND PHONE LINES

The free code, and the line you still rent

Frameworks that run the conversation loop

Pipecat	16.3k ★	From Daily. BSD license. Built-in phone audio for Twilio, Telnyx, Plivo and Vonage. Ships the Smart Turn model.
LiveKit Agents	13.0k ★	Apache 2.0. Inbound and outbound calls over SIP. The company is valued at $1B and its customers include OpenAI.
TEN Framework	11.1k ★	From Agora. Includes its own voice-activity and turn-detection models.
Asterisk · FreeSWITCH	Classic	Long-standing open-source phone systems. Connect here when the agent must transfer callers to real office extensions.
Vocode	Dormant	No updates since November 2024. Older tutorials still recommend it. Don't start new work on it.

GitHub stars as of October 9, 2026.

Renting the line · US

Twilio local number	$1.15/mo
Twilio toll-free number	$2.15/mo
Twilio inbound, local	$0.0085/min
Twilio outbound	$0.014/min
Twilio Media Streams	$0.0044/min
Telnyx number · SIP inbound	$1/mo · $0.0032

Three ways to hand a call to your agent

01 **Audio stream.** Twilio Media Streams sends raw 8 kHz audio to your server over a WebSocket. Full control, and the most work.

02 **Text relay.** Twilio ConversationRelay does the listening and speaking. Your server only exchanges text.

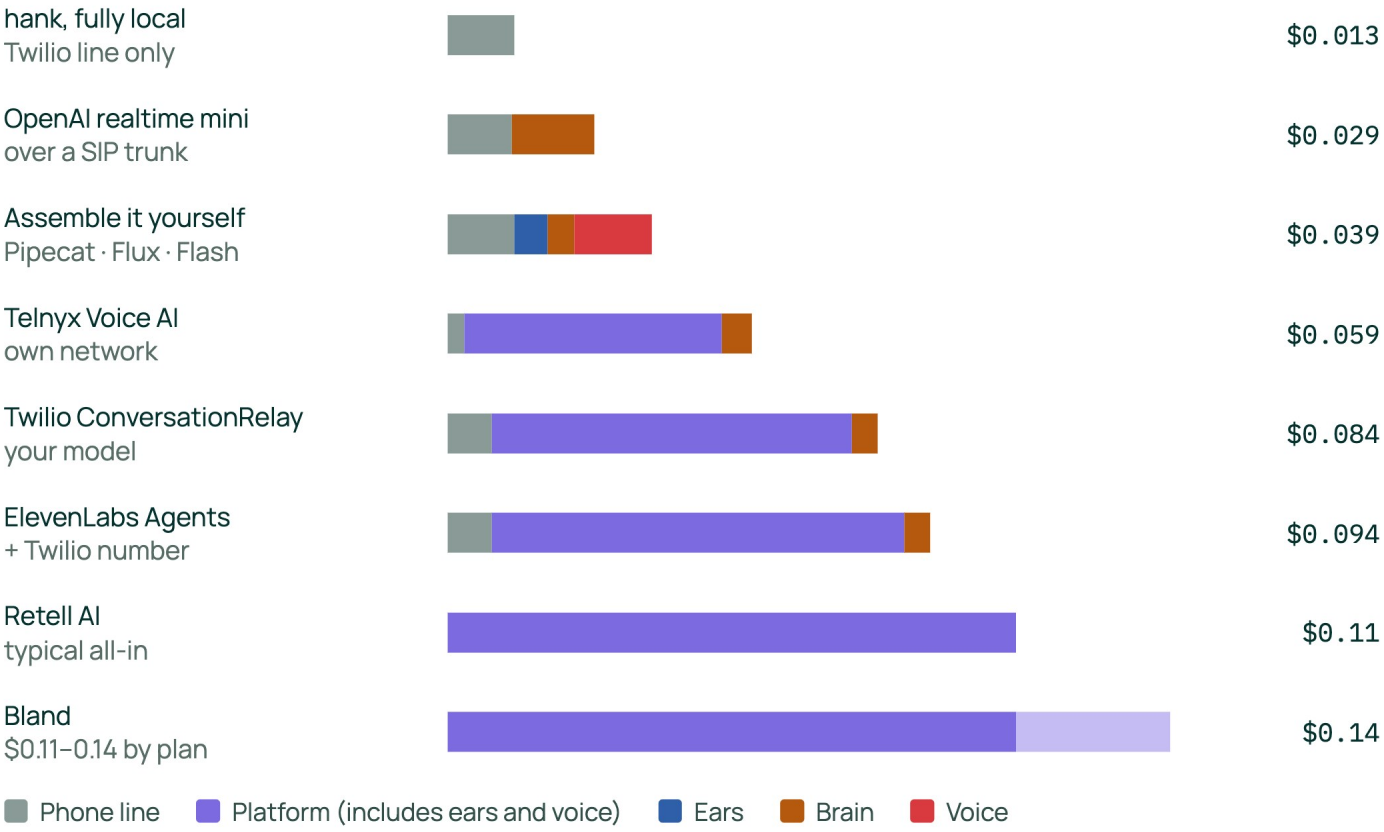
03 **SIP trunk.** Point the carrier straight at LiveKit, OpenAI Realtime, Vapi or Retell. This skips the per-minute stream fee.

Get past the spam filters. Carriers sign each call with a STIR/SHAKEN rating. "A" means verified and is least likely to show "Spam Likely." Since Feb 1, 2025, US carriers block texts from local numbers that aren't registered for A2P 10DLC. Register before your agent texts anyone.

WHAT A MINUTE REALLY COSTS

One inbound US call-minute, eight ways

Built from the October 2026 list prices on pages 11–13. Where a small cloud model does the thinking, we assume about $0.005 a minute.



1,000 MIN / MONTH
Local **$13** · DIY **$39**
ElevenLabs **~$94**

10,000 MIN / MONTH
Local **$130** · DIY **$390**
ElevenLabs **~$940**

FOR COMPARISON
A live-agent contact averages **$8.01** (Gartner).

Local is not free. It costs hardware you already paid for, electricity and your own time. It also handles only as many calls at once as the box can carry. Monthly figures leave out plan fees, and the platform rows above don't include their monthly plan either.

MEET HANK

A mini-PC that thinks around the clock

hank is Winfield's agentic computer. It runs its own language models, keeps its own memory, and handles work while we sleep. Part IV is the honest story of giving it a voice.

PROCESSOR	MEMORY	GRAPHICS	SYSTEM
AMD Ryzen AI Max+ 395 "Strix Halo"	**128 GB unified** Shared by CPU and GPU	**Radeon 8060S** Runs models via Vulkan	**Ubuntu Linux** Headless, on 24/7

What runs on hank

Ollama serving gpt-oss:120b and qwen3:30b-a3b, at no cost per token	LIVE
Hermes Agent, Nous Research's open-source agent, with tools and memory	LIVE
Telegram, for texting hank from a phone	LIVE
Eight in-house tools, including dashboard, reports, contacts, notifications and Cost Guard	LIVE
A private Tailscale network joining hank to our other machines	LIVE
A voice and phone line, behind a locked relay	PLANNED

House rules hank lives by

- ✓ **Fail closed.** A missing key means refuse, never guess.
- ✓ **Scoped keys.** Each service gets only the access it needs.
- ✓ **No public doors.** Nothing on hank is reachable from the open internet.
- ✓ **Spending cap.** An alert at $2 a day in paid API calls, a hard stop at $5.
- ✓ **Local first.** Paid cloud models are the exception, not the default.
- ✓ **A human runs admin.** hank prepares admin (sudo) commands. A person runs them.

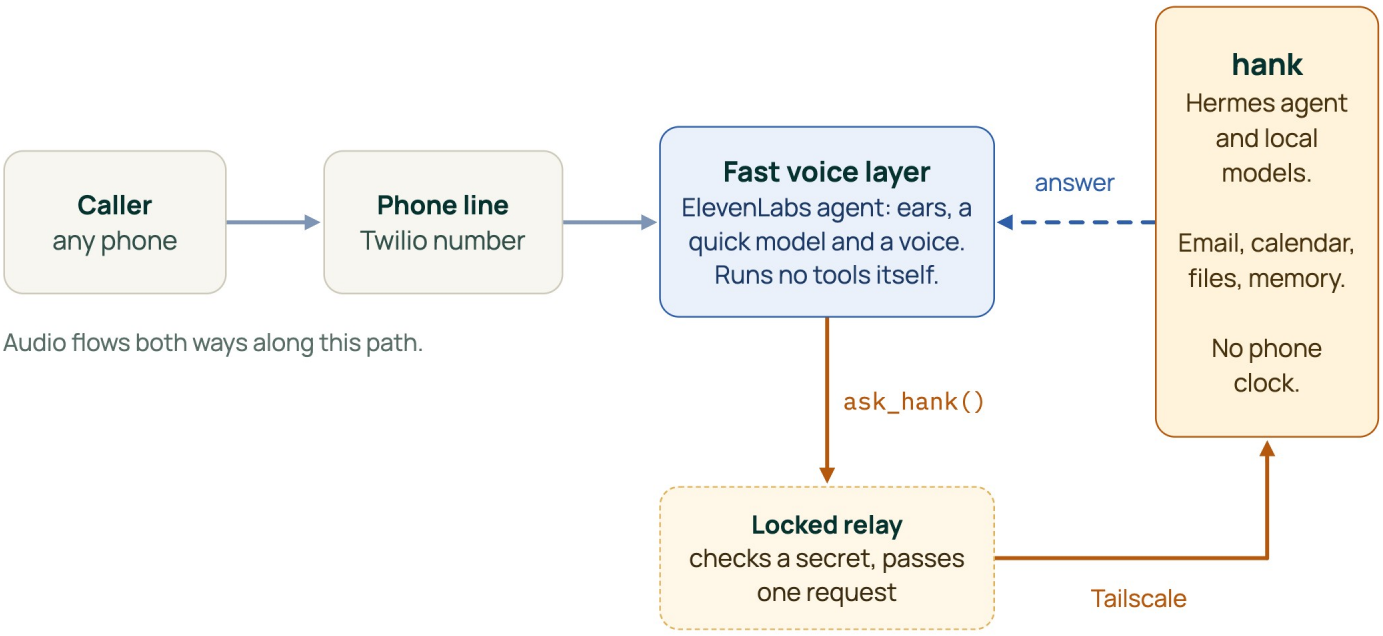
WHERE THE VOICE STANDS, OCTOBER 2026

Our previous agent box had a working phone line using ElevenLabs and Twilio. We shut it off in August because it was exposed to the public internet. hank's daily phone-agent digest is built and waiting on a fresh ElevenLabs key. The next three pages show how we rebuild the voice properly.

THE SPLIT BRAIN

Fast mouth, slow brain, never on the same clock

Our first phone build pointed calls straight at the full agent. Real tasks take the agent seconds to half a minute. The voice platform gave up after roughly eight seconds, and the calls dropped. The fix is to split the job in two.



Audio flows both ways along this path.

The mouth stays fast

A small, quick model handles greetings, small talk and "let me check on that." It answers in well under a second.

One tool, one door

The voice layer gets exactly one tool, ask_hank. It can't read email or touch files. It can only ask.

The brain takes its time

hank works on its own schedule and returns a short spoken-style answer. Long jobs become a text or email follow-up.

WHAT THE CALLER HEARS · ILLUSTRATIVE

Caller	"Did the quote for the Ohio job go out?"
Voice layer	"Let me check with hank. One moment."
hank	*searches sent mail, finds it, replies in one sentence*
Voice layer	"Yes, it went out Tuesday afternoon. Want me to text you a copy?"

NEVER OPEN THE FRONT DOOR

The voice reaches in. The internet doesn't.

A cloud voice platform has to call your agent over the public internet. Most tutorials say to open a tunnel straight to your machine. We did that once. Here is why we stopped, and what replaced it.

BEFORE · RETIRED AUGUST 2026

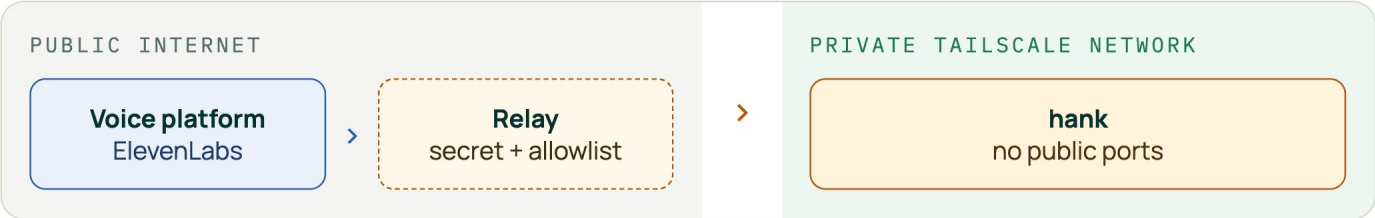
A tunnel straight into the agent box

A Tailscale Funnel gave the agent's own server a public web address. An audit then found API keys that defaulted to empty, which meant no key at all. Anyone who found the address could have knocked.

NOW · HANK'S DESIGN

A small locked relay in the cloud

The only public piece is a small relay on a $5-a-month server or a Cloudflare endpoint. It checks a secret and forwards one kind of request over Tailscale. hank itself opens nothing to the internet.



The relay checklist

01 **Fail closed.** No secret, or a wrong one, gets a refusal.

02 **One door, one verb.** A single endpoint that accepts only "ask hank."

03 **Known callers only.** An allowlist of phone numbers that may reach hank.

04 **Rate limits.** Cap requests per minute and per day.

05 **Log everything.** Every request, every answer, never the secrets.

06 **Keys out of chat.** Keys go straight into the secrets file, never into a chat window or a log.

07 **Rotate.** New keys on a schedule, old ones revoked.

08 **Kill switch.** One command turns the phone line off.

A LOCAL VOICE ON STRIX HALO

Everything but the phone line, at home

This is the stack we plan to run on hank's AMD hardware. Most voice tutorials assume an NVIDIA card, so we picked parts that run on AMD. We haven't benchmarked it on hank yet. Measure before you trust our numbers.

Turn-taking	Silero VAD plus Pipecat Smart Turn. Both are tiny and run on the CPU.	CPU · MIT / open
Ears	NVIDIA's Parakeet model via ONNX on the CPU (about 36× real time on a desktop Ryzen), or whisper.cpp on the GPU through Vulkan.	CPU or Vulkan
Brain	qwen3:30b-a3b answers the live turn: it activates only about 3B parameters, so it's quick. gpt-oss:120b handles deep work through hank.	Ollama · LIVE today
Mouth	Kokoro-82M has 54 voices and an Apache license. Piper is very fast on a CPU but is GPL-licensed. Chatterbox Turbo adds voice cloning and watermarks its audio.	CPU · ROCm
Conductor	Pipecat runs the loop and speaks Twilio's and Telnyx's phone audio natively.	Python · BSD

WON'T RUN, OR CAN'T SELL

Kyutai Unmute and Sesame CSM are documented for NVIDIA CUDA only. F5-TTS and XTTS v2 weights are licensed for non-commercial use only. Coqui, XTTS's maker, shut down in January 2024.

DRIVER NOTES

AMD lists the Ryzen AI Max+ 395 under ROCm 7.2.1 with official support on Ubuntu 24.04. For llama.cpp and whisper.cpp, Vulkan is often as fast as ROCm. Stay on the 24.04 LTS release for now.

The honest trade-off: local voices are good, not yet ElevenLabs-good. That's why hank's first phone line rents its voice, and why the local stack starts with private uses like voice notes and an intercom.

THE HANK VOICE ROADMAP

Five steps, each safe to stop at

Each phase works on its own and adds exactly one new kind of exposure. If a phase misbehaves, step back one.

0 UNDER WAY

Unblock

Create a fresh, scoped ElevenLabs key with a spending cap. Put it straight into hank's secrets file. Resume the paused daily phone-agent digest, then revoke the old key. **Done when** the digest arrives on its own for a week.

1 PLANNED

Voice notes, no phone

Speak to hank through Telegram voice notes. Local ears transcribe them and hank answers with audio. Nothing new faces the internet. **Done when** a spoken request round-trips reliably.

2 PLANNED

A private phone line

One Twilio number, an ElevenLabs agent with a fast model, and a single ask_hank tool through the locked relay. Only the owner's numbers may call. Cost Guard counts voice minutes. **Done when** a call can check email and the calendar without a dropped call.

3 PLANNED

A business line

hank answers inbound calls for intake and scheduling. Every call opens by saying it is an AI and asks before recording. Any caller can reach a person. **Done when** callers' questions resolve or reach a person, never a dead end.

4 PLANNED

Bring it home

Swap rented ears and voice for the Strix Halo stack wherever quality holds, with the rented voice kept as a fallback. **Done when** blind listeners can't reliably pick the local voice.

Measure every phase. Log reply time, drops and cost per call from day one.

One change at a time. Never swap the voice and the brain in the same week.

Rollback ready. Each phase can be switched off with one command.

RULES OF THE ROAD · AS OF OCTOBER 2026

An AI voice is a regulated voice

US · FCC, FEB 2024

AI voices count as "artificial"

Under the TCPA, robocalls using AI voices need the recipient's prior consent. The FCC, state attorneys general and private plaintiffs can all enforce it.

US · CALIFORNIA

Don't pretend to be human

The 2019 BOT Act bars undisclosed bots used to sell or to sway votes. SB 243 (signed Oct 2025) requires companion chatbots to disclose they are AI.

EU · AI ACT ART. 50

Disclose from the first word

Since Aug 2, 2026, AI that talks to people must say so at the first interaction, and synthetic audio must be machine-marked. Fines reach €15M or 3% of turnover.

US · 11 STATES

Everyone must agree to recording

CA, DE, FL, IL, MD, MA, MT, NV, NH, PA and WA require consent from every party. You can't tell where a caller is, so ask every time.

US · TENNESSEE, FEDERAL

A voice belongs to its owner

Tennessee's ELVIS Act (2024) protects a person's voice against unauthorized cloning. The federal NO FAKES Act cleared committee in June 2026 but stalled on Sep 30.

CARRIERS

Prove who's calling and texting

STIR/SHAKEN caller authentication has been mandatory since June 2021. Texts from unregistered local numbers have been blocked since Feb 1, 2025 (A2P 10DLC).

HANK'S HOUSE RULES FOR VOICE

1. Say it's an AI in the first sentence.
2. Ask before recording, every call.
3. No AI cold calls without written consent.
4. Only clone a voice you own.
5. Always offer a human.
6. Honor "stop calling" immediately.

A DISCLOSURE-FIRST OPENING LINE · ADAPT IT TO YOUR BUSINESS

"Hi, you've reached hank, the AI assistant for [YOUR COMPANY]. I can record this call so we can follow up. Is that okay? And if you'd rather talk with a person at any point, just say so."

Discloses AI, asks about recording and offers a human in under fifteen seconds. If the caller says no, don't record.

This page is a summary to help you ask the right questions. It is not legal advice.

FIELD-TESTED GOTCHAS

Twelve ways a talking agent breaks

Items marked OURS happened on our own machines.

01 **Smart models drop calls.** OURS Keep the full agent out of the live loop. Use the split brain on page 16.

02 **Public tunnels leak.** OURS A quick tunnel to your agent becomes permanent exposure. Use a relay.

03 **Empty keys mean open doors.** OURS A key that defaults to "" turns off security. Missing must mean refuse.

04 **Phone keys aren't account keys.** OURS Some platforms use separate credentials for telephony. Mix them up and the errors are baffling.

05 **Keys pasted in chat are burned.** If a key passes through a chat, a Telegram thread or a log, rotate it.

06 **Cold AI calls backfire.** People spot an AI voice fast, and the TCPA requires consent. Use email or text for cold outreach.

07 **Wrong audio format.** Phone audio is 8 kHz μ-law. A mismatch sounds like static, chipmunks or nothing at all.

08 **Dead air.** A stalled backend must say something and offer a callback. It must never just go quiet.

09 **The agent interrupts itself.** That's usually echo, or backchannel words like "uh-huh" counted as interruptions.

10 **Long calls get pricey.** Speech-to-speech models re-read the growing conversation every turn. Cap call length or summarize as you go.

11 **Vendors vanish.** Hume's voice APIs close Nov 13, 2026. OpenAI's gpt-4o-transcribe ends in Feb 2027. Keep every piece swappable.

12 **Texting takes paperwork.** Local numbers need A2P 10DLC registration before carriers pass your texts. Start early.

THE ONE THAT WORKED AS DESIGNED

When hank's daily phone-agent digest found no valid ElevenLabs key, the job paused itself and flagged the problem instead of running blind. That's fail-closed doing its job. Build every voice job to stop loudly.

PICK YOUR PATH

Find your situation, start there

IF YOU…	START WITH	≈ PER MIN
want a demo this weekend	ElevenLabs Agents and a Twilio number. Your agent sits behind a relay as the custom brain.	$0.09–0.10
want managed, but cheaper	Telnyx Voice AI (carrier and AI in one) or Vapi with pass-through pricing.	$0.06–0.13
need a natural, fast front desk	Speech-to-speech: gpt-realtime-2.1-mini or Gemini Live over a SIP trunk, with a few well-defined tools.	$0.03
have a developer and real volume	Pipecat or LiveKit Agents, Deepgram Flux, ElevenLabs Flash or Cartesia, and a Twilio or Telnyx SIP trunk.	$0.04
can't let audio leave the building	The hank-style local stack (page 18). Rent only the number. Budget real hardware and testing time.	$0.013 + box
must transfer to office phones	Any path above, plus a SIP connection to your Asterisk, FreeSWITCH or hosted phone system.	varies
plan to cold-call consumers	Stop. Get prior written consent first (page 20). For cold outreach, email or text usually works better anyway.	risk

Still unsure? Answer three questions.

1

How many minutes a month?

Under a few thousand: renting wins on time saved. Tens of thousands: assembling pays for a developer.

2

Does it need your private data?

If it reads email, files or the CRM, use the split brain and the relay. The voice layer never holds those keys.

3

Who may hear the audio?

If the answer is "no one outside," go local. Otherwise read each vendor's data-retention terms.

GLOSSARY

The words vendors assume you know

A2P 10DLC. US registration for business texting from ordinary local numbers. Unregistered texts are blocked.

Backchannel. "Uh-huh," "yeah," "okay": sounds that mean "keep going," not "stop."

Barge-in. The caller talks over the agent. A good agent stops at once.

Cascaded pipeline. Speech-to-text, then a language model, then text-to-speech, in a row.

Endpointing. Deciding the caller has finished their turn. Also called turn detection.

Full duplex. Listening and speaking at the same time, as people do.

G.711 μ-law. The 8 kHz audio encoding used on ordinary phone calls.

IVR. The "press 1 for sales" phone menus that came before AI agents.

Media Streams. Twilio's feature that sends a call's raw audio to your server over a WebSocket.

Relay. A small public server that checks a secret and forwards one kind of request to a private machine.

SIP trunk. A carrier connection that delivers phone calls to software over the internet.

Speech-to-speech. One model that takes audio in and gives audio out, with no text step in between.

STIR/SHAKEN. Carriers' signed caller ID. A higher rating means fewer "Spam Likely" labels.

STT / ASR. Speech-to-text, or automatic speech recognition: the agent's ears.

TCPA. The US law governing automated and artificial-voice calls. It covers AI voices since 2024.

TTFT / TTFB. Time to first token, or first byte: how long until the model, or the voice, starts.

TTS. Text-to-speech: the agent's mouth.

VAD. Voice activity detection: is anyone talking right now?

Vulkan / ROCm. Two ways to run AI models on AMD graphics. Vulkan is broader; ROCm is AMD's own.

WebSocket. A connection that stays open so audio or text can stream both ways.

SOURCES · CHECKED OCTOBER 9, 2026

Where every number came from

HISTORY AND THE GIANTS

Wikipedia, "Speech recognition" and "Dragon NaturallySpeaking"

SRI International, history of Siri · sri.com

Google blog, Assistant at I/O, May 18, 2016

Engadget, Duplex disclosure, May 10, 2018

OpenAI, "Introducing the Realtime API," Oct 2024

The Decoder, Kyutai Moshi; Sesame CSM-1B

Fortune, Alexa insiders, Jun 12, 2024

AppleInsider, Alexa $10B (2022); Alexa+ (Feb 2026)

Voicebot.ai, smart speaker usage, May 2020

Google–Apple joint statement, Jan 12, 2026

Macworld, iOS 27 Siri AI; AI Weekly, Gemini cutoff

MONEY AND MARKETS

Gartner press releases, Sep 2019 and Aug 2022

a16z, "AI Voice Agents: 2025 Update"

eWeek (ElevenLabs); Quartz (Sierra); Bloomberg Law (Decagon)

Built In SF (Deepgram); Fortune (LiveKit; Bland, Jun 2026)

Market.us; MarketsandMarkets via Speech Technology

HOW PEOPLE TAKE TURNS

Stivers et al., PNAS 106(26), 2009

Levinson & Torreira, Frontiers in Psychology, 2015

Parloa Labs, "The latency paradox"

Hamming AI, voice latency study, Jan 2026

PRICING PAGES

elevenlabs.io/pricing · vapi.ai/pricing · retellai.com/pricing

telnyx.com/pricing · twilio.com voice and AI pricing

deepgram.com/pricing · livekit.com/pricing

assemblyai.com · rime.ai · cartesia.ai · inworld.ai

OpenAI and Google Gemini API pricing; AWS Nova

Ringly.io comparisons (Bland, Synthflow)

TECHNICAL DOCS AND CODE

GitHub: pipecat, livekit/agents, TEN, vocode, silero-vad

GitHub: whisper.cpp, kyutai unmute, sesame csm

Daily, Smart Turn v3.1; Deepgram, Flux launch

Twilio docs: Media Streams, ConversationRelay

AMD ROCm Radeon/Ryzen compatibility matrix

Hugging Face: Kokoro-82M; Resemble: Chatterbox

Hume developer docs, API shutdown notice

LAW AND POLICY

Akin Gump, FCC AI-voice ruling, Feb 2024

Rev, US call-recording laws by state

Greenberg Traurig (BOT Act); Skadden (SB 243)

Manatt (ELVIS Act); Daily Journal (NO FAKES), Oct 2026

Jones Walker, EU AI Act Article 50

ElevenLabs voice-cloning docs and help center

The hank material comes from Winfield's own build logs. Market forecasts are vendor research reports, quoted as direction only. Full links for every source are available on request at winfieldtech.com.

If an agent can learn to answer the phone, it can help your business do the same.

Winfield Technologies researched and wrote this guide with hank, the company's always-on AI agent. Every failure in it happened to us first. Every price was checked on the vendor's own page. Every claim about hank is marked as live or planned.

Want a voice agent connected to your own calendar, CRM and phone system, without opening your network to the internet? That's the work we do.

Works with OpenClaw, Hermes, Claude Dispatch and any agent with an API.



Winfield Technologies Inc.
winfieldtech.com

TALKING AGENTS · 2026 EDITION
SET IN BRICOLAGE, MANROPE AND PLEX MONO
© 2026 WINFIELD TECHNOLOGIES INC.