Proposal — Vantage Premium Voice on Cloudflare, with Claude as the Brain

For Bob, Niven, and Paul · 2026-07-18 · Proof-of-Concept proposal (approval requested — no build yet)

The idea in one paragraph

Move the Vantage AI Receptionist and Ask Vantage Copilot to a **single voice stack that runs on Cloudflare**: WebRTC audio over Cloudflare's own Realtime network, **Claude as the reasoning brain** (via Cloudflare AI Gateway), and **premium ElevenLabs voices** — a warm, confident male in the spirit of the "Ash" voice we use for narration, plus **an ElevenLabs equivalent of "Coral"** — the gracious female Coordinator voice Hidden Creek uses today. This replaces the OpenAI realtime-voice path we're moving away from, keeps everything inside Cloudflare for per-tenant isolation and flat cost, upgrades the brain from a Cloudflare open model to Claude, and — because Claude follows instructions and calls tools far more reliably — unlocks more real automation (booking, quoting, lead capture) than we can do today.

What "using Claude" changes vs. Cloudflare's AI today

- **Same "binding," stronger result.** You ground Claude exactly the way you ground Cloudflare's AI today — a system prompt plus tenant knowledge (retrieval) plus tool definitions. This is *not* fine-tuning; it's the same instruction-and-knowledge approach, and Claude simply follows it more faithfully. (True fine-tuning exists on Anthropic's platform if we ever want it, but the PoC doesn't need it.)
- **More automation, done reliably.** Claude's tool/function-calling is materially better than an open 8B model's. That's the difference between a receptionist that *reads a script* and one that actually **books a tour, quotes the rate card (and only the rate card), captures a lead, checks a calendar, or takes a message** — the automation you expect from this route.
- **Consistency with the rest of our stack.** We already run Claude for WebDesigner/taste; one vendor, one prompting discipline, plus cost levers (prompt caching –90% on the fixed tenant prompt, Batch API –50%).

Proposed architecture (all on Cloudflare)

Layer	Choice	Where it runs
WebRTC transport	Cloudflare **Realtime SFU/TURN** (anycast, $0.05/realtime-GB)	Cloudflare edge
Orchestration + memory + tools	**Durable Object** voice agent (Agents SDK `withVoice`), SQLite conversation history	Cloudflare edge
Speech-to-text	**Deepgram** (Nova-3 / Flux) via Workers AI — in-network	Cloudflare edge
Brain (LLM)	**Claude** (Haiku 4.5 default; Sonnet 5 for the Copilot) via **AI Gateway**	Anthropic, proxied through Cloudflare

Layer	Choice	Where it runs
Text-to-speech (premium)	**ElevenLabs Flash v2.5** — an "Ash-style" male + a "Coral-style" female (~75 ms)	ElevenLabs
Fallback brain	Workers AI open model via AI Gateway auto-fallback	Cloudflare edge

The pipeline is provider-agnostic by design, so we can A/B a cheaper in-network TTS (Deepgram Aura-2, ~90 ms) against the premium ElevenLabs voice per tenant tier.

Voice/IP note: we will **not** clone OpenAI's "Ash" (their asset). We'll select or design ElevenLabs equivalents for BOTH of today's OpenAI presets — the warm male "Ash" and the gracious female "Coral" (the venue Coordinator preset). Neither is an ElevenLabs voice today.

Value proposition

- **Clients (venues):** a receptionist that *sounds elite* and matches the brand (Coordinator-caliber), answers 24/7, books and quotes accurately, captures every lead, and handles Spanish for DFW. Category-first — our venue-competitive scan found an AI receptionist in **0 of 9** competitors.
- **EdgePointe:** removes the OpenAI realtime dependency (and the "realtime-not-connecting → robot fallback" bug), consolidates voice onto Cloudflare (per-tenant isolation, no in-network egress, flat cost), differentiates every tenant with premium voice, and adds automation that raises the tier's value — all while holding or lowering cost.
- **Owners:** an Ask Vantage Copilot they can *talk to* in the console, that actually performs actions.

Cost–benefit (estimates — to be confirmed by the PoC)

Rough per-minute of live conversation:

Component	~Cost / min	Notes
Cloudflare Realtime SFU/TURN	~$0.0001	voice is low-bandwidth; effectively negligible
STT — Deepgram (Workers AI)	~$0.005	in-network
Brain — Claude Haiku (cached prompt)	~$0.005–0.015	tenant prompt cached at –90%
TTS — **ElevenLabs Flash (premium)**	~$0.040	$0.05 / 1k chars; the largest slice
Total (premium voice)	**~$0.05–0.06 / min**	**≈ $0.18–0.22 per 3.5-min call**
Alt: TTS — Deepgram Aura-2 (in-network)	*~$0.025*	*total ~$0.035/min for non-premium tenants*

For comparison, the OpenAI realtime speech-to-speech path we're leaving is commonly ~**$0.15–0.30/min** (premium audio-token pricing) — so the proposed premium stack is **comparable-to-cheaper with a better brain, our choice of voice, and everything on Cloudflare.** The LLM is a small slice of every call, so paying for a frontier brain barely moves total cost while sharply improving reliability.

Pricing sources: Cloudflare Realtime, Workers AI + Deepgram, Claude API pricing, ElevenLabs pricing, Deepgram pricing, Cloudflare voice agents.

How we would test it (PoC flow)

1. **Sandbox:** one tenant (`hiddencreek-demo`), two new ElevenLabs voices — a Coral-style female and an Ash-style male — behind a feature flag (`VANTAGE_VOICE_POC`, default off). Nothing touches live tenants.
2. **Stand up** a Worker + Durable-Object voice agent: WebRTC via Realtime SFU, Deepgram STT, **Claude via AI Gateway** grounded on the tenant's existing content + a small tool set (book-tour, quote-rate-card, capture-lead — stubs OK), ElevenLabs TTS.
3. **Scripted test calls** (in-browser widget — no phone number needed): book a tour, ask hours, ask price → *quote the rate card only, never invent a number*, capture a lead, one Spanish turn, a "take a message" fallback, and a barge-in/interrupt.
4. **Measure:** transcription accuracy, tool-call correctness, on-script adherence (zero fabricated prices), subjective voice quality, first-audio latency, voice-to-voice latency, and real cost per call — **side-by-side against the current path.**
5. **Go** / **no-go** against the success criteria below.

Success criteria: voice-to-voice < 800 ms; tool-call accuracy ≥ 95%; **zero** fabricated prices across the script; voice quality rated ≥ current; cost/min ≤ current; ≥ 99% connection success over the test-call set.

Risks & mitigations

- **Cloudflare's Agents voice pipeline is still experimental** → PoC only, flag-gated, with a small evals harness before anything goes near a live call.
- **External hops add latency** (Claude, ElevenLabs) → Flash TTS (75 ms) + Haiku + sentence-streaming; option to drop to in-network Deepgram voice where latency is critical.
- **Premium voice cost at high volume** → tier it (ElevenLabs for premium/Concierge tenants, Deepgram Aura-2 for lower tiers); use bundle plans.
- **Phone/telephony still needs Twilio + EIN/A2P** → PoC uses the in-browser WebRTC widget, so it proceeds now, ahead of the EIN.
- **Vendor dependency** → AI Gateway auto-fallback to a Workers AI brain; provider-agnostic TTS interface keeps voices swappable.

Recommendation & the ask

Approve a **time-boxed, flag-gated PoC** on `hiddencreek-demo` — one tenant, two premium voices, Claude brain — measured head-to-head against the current stack, shipped through our normal GitHub CI/CD, touching no live tenant. It directly tests three things at once: moving off OpenAI voice, upgrading the brain to Claude, and proving the added automation.

On your approval, I'll deliver the detailed PoC build plan (tasks, flags, test harness, evals, timeline, and exact voice selections) — and only build once you approve *that*.