NetWebMedia

2026

CONVERSATIONAL AI GUIDE

OpenAI Realtime API: Building AI Customer Conversations That Convert

A practitioner's guide to deploying voice AI for B2B customer conversations — from architecture to compliance to performance measurement

Marcus Chen
SEO Director, NetWebMedia

13 pages
netwebmedia.com

Contents

Performance Measurement: The Metrics That Matter for Voice AI

The performance framework for voice AI — from call mechanics to revenue contribution — and how to build the reporting that makes ongoing optimization possible.

EXECUTIVE SUMMARY

OpenAI Realtime API: Building AI Customer Conversations That Convert

Voice AI has crossed a threshold in 2026 that changes what's possible in B2B customer conversations. Sub-300ms response latency, dramatically improved natural language understanding, and reliable context management across multi-turn conversations mean that AI-handled calls are now indistinguishable from human calls for large categories of interactions. The OpenAI Realtime API is the infrastructure layer enabling this shift — providing the low-latency speech-to-speech capabilities that make voice AI conversations feel natural rather than robotic. But technical capability and business deployment readiness are different things. This guide addresses the full deployment picture: where voice AI belongs in your customer conversation stack, how to build the architecture correctly, how to comply with disclosure requirements, how to design conversations that convert, and how to measure whether the investment is working.

IN THIS GUIDE

- ✓ Why sub-300ms latency is not a technical benchmark but a user experience threshold that determines whether voice AI feels like a conversation or a phone tree
- ✓ The use case qualification framework for deciding where voice AI delivers ROI and where it destroys customer experience
- ✓ A practical architecture overview covering telephony integration, CRM connection, and session management
- ✓ Conversation design principles specific to AI — why human call scripts fail for AI deployment and what to build instead
- ✓ The performance measurement framework for voice AI that goes beyond call volume to revenue contribution

Who this is for: B2B marketing and sales technology leaders evaluating voice AI for customer-facing applications — qualification calls, appointment setting, customer onboarding, or support escalation.

SECTION 1

The Latency Breakthrough: Why Sub-300ms Changes Everything

Human conversation has a physical rhythm. Turn-taking in natural dialogue involves response latencies between 200-500 milliseconds — the gap between one speaker finishing and the next beginning. When response latency exceeds 700-800 milliseconds, human speakers begin to feel that something is wrong with the connection or that the listener isn't engaged. At 1.5+ seconds of latency — which characterized most voice AI systems before 2024 — the interaction doesn't feel like a conversation at all. It feels like navigating a phone tree with speech recognition. The consequence wasn't just annoyance: high-latency voice AI produced measurably worse outcomes because callers changed their behavior in response to the latency, using shorter, less natural utterances and losing the conversational context that makes qualification and persuasion possible. The OpenAI Realtime API's sub-300ms speech-to-speech latency changes this dynamic fundamentally. At 250-300ms, the response feels conversational — not quite as instantaneous as face-to-face dialogue, but within the natural variation of a phone call with a human. Users do not modify their speech patterns to accommodate the AI. They speak naturally, provide richer context, and engage in the multi-turn exchanges that qualification, objection handling, and appointment setting require.

The technical mechanism behind the latency breakthrough: the Realtime API processes audio directly rather than converting speech to text, sending text to a language model, and converting the response back to speech (the pipeline that created high latency in prior-generation systems). The direct audio processing eliminates two conversion steps and their associated delays. It also preserves prosodic signals — the tone, emphasis, and pacing in human speech — that carry meaning lost in text transcription. This means the AI can respond to how something is said, not just what is said: a question phrased with uncertainty is processed differently from the same question phrased with confidence. For B2B applications where qualification depends on reading buyer intent signals, this prosodic awareness is a meaningful capability upgrade.

- Benchmark your current IVR or chatbot solution's response latency before evaluating Realtime API — document the baseline
- Test with your target audience: have 5-10 prospects or customers interact with a prototype and rate conversational naturalness
- Measure speech modification rate: do users shorten or simplify their responses to accommodate the AI? High modification = bad UX
- Document the qualification variables your current human calls capture that would need to be preserved in AI conversations

- Set a minimum acceptable latency threshold before deployment: 300ms or below for conversational use cases

> *The 300ms threshold is not arbitrary — it's grounded in human conversation research. Below it, users speak naturally and engage. Above 700ms, users modify behavior in ways that reduce conversation quality and conversion rates.*

<300ms the OpenAI Realtime API's target response latency — enabling voice AI conversations that feel natural rather than robotic for the first time at production scale

SECTION 2

Use Case Qualification: Where Voice AI Works vs. Where It Shouldn't

Not every customer conversation is appropriate for AI deployment, and deploying voice AI in the wrong context produces customer experience damage that's difficult to repair. The use case qualification framework evaluates each conversation type across four dimensions. Dimension one: conversation structure. Highly structured conversations — those that follow predictable flows with defined information exchange and a clear endpoint — are strong AI candidates. Qualification calls, appointment setting, survey administration, status check calls, and initial intake follow structured patterns that AI handles reliably. Unstructured conversations — escalated complaints, sensitive situations, complex negotiations — require the judgment, empathy, and authority that human agents provide. Dimension two: stakes and sensitivity. Low-stakes, low-sensitivity interactions are appropriate for AI. High-stakes interactions (large enterprise deals, customer at risk of churn, post-incident resolution) or sensitive situations (billing disputes, medical or legal contexts) should route to humans. Dimension three: information availability. Voice AI is only as useful as the context it has. If the AI has access to the caller's account history, prior interactions, and relevant product/service data, it can conduct informed conversations. If it's operating without context, it can't personalize and can't avoid asking questions the caller has already answered.

Dimension four: conversion importance. For conversations where conversion (booking, qualification advancement, purchase) is the primary objective, test AI performance against human benchmark before full deployment. Voice AI typically achieves 70-85% of human agent conversion rates in well-designed deployments for appropriate use cases — a genuine ROI gain given the volume and cost advantages. For high-conversion-value conversations where even a 10% conversion rate difference represents significant revenue, the threshold for deployment should be higher and the testing more rigorous. The qualification scoring: a conversation type that scores well on all four dimensions is a strong AI candidate. A conversation type that fails on any one dimension requires either redesign (can it be restructured to qualify?) or exclusion from AI deployment.

- Dimension 1 — Structure: score each conversation type as structured (AI candidate), semi-structured (design required), or unstructured (human only)
- Dimension 2 — Stakes/sensitivity: low stakes + low sensitivity = AI appropriate; either elevated = route to human
- Dimension 3 — Information availability: map what CRM and account data the AI would have access to for each conversation type
- Dimension 4 — Conversion importance: set a minimum acceptable AI conversion rate benchmark before deployment for high-stakes use cases

- Run the four-dimension framework across your top 5-10 customer conversation types and produce a qualification scorecard
- Start AI deployment with 2-3 highest-qualifying use cases — do not deploy across all use cases simultaneously

> *The most common voice AI deployment failure is applying it to conversation types that fail the structure or sensitivity dimensions — producing customer experience damage that takes months to undo with the affected customer segment.*

70–85% of human agent conversion rates achievable by voice AI in well-designed deployments for appropriately qualified use cases — a compelling ROI when applied at volume

SECTION 3

Architecture Overview: API, Telephony, CRM Integration

A production voice AI system using the OpenAI Realtime API has five architectural components. Component one: the OpenAI Realtime API session layer. This handles the speech-to-speech conversation processing — receiving audio input, processing it through the model, and returning audio output within the sub-300ms latency target. Sessions are stateless by default: each call is a fresh context unless context is injected at session start. Session configuration includes the system prompt (the AI's persona, instructions, and constraints), the voice selection, and the turn detection settings that determine when the AI recognizes a speaker has finished their turn. Component two: the telephony integration layer. The Realtime API processes audio streams, not phone calls. Connecting it to a telephone infrastructure requires a telephony provider with WebSocket audio streaming support — Twilio, SignalWire, and Vonage all offer this integration path. The telephony layer handles PSTN connectivity, call routing, DTMF detection, and recording. It passes audio streams to and from the Realtime API in real time.

Component three: the context injection layer. This is the most important and most commonly underbuilt component. At session start, before the first utterance, the system needs to inject the caller's relevant context into the session — account status, prior call history, open tickets, stage in the sales cycle, and any data that enables personalized, informed conversation. This requires a lookup function: when a call comes in, identify the caller (by phone number, or by prompted authentication), query the CRM or relevant data source, and include the retrieved context in the system prompt before the session starts. Component four: the action execution layer. Voice AI conversations often trigger downstream actions: booking an appointment, logging a call summary to the CRM, routing a follow-up task, or sending a confirmation email. These actions require API integrations from the AI session to the relevant downstream systems. Component five: session recording and logging — essential for quality assurance, compliance, and conversation design iteration.

- Realtime API session: configure system prompt, voice, turn detection, and response constraints before any caller interaction
- Telephony integration: evaluate Twilio, SignalWire, or Vonage for WebSocket audio streaming support and PSTN capability
- Context injection: build the caller lookup function that loads CRM context before the first AI utterance
- Action execution: map every downstream action the AI should trigger (CRM logging, scheduling, follow-up) and build the integrations

- Session recording and logging: required for QA, compliance, and conversation design iteration — configure from day one
- Error handling: build fallback routing to human agents for session failures, low-confidence responses, and escalation requests

> *Context injection is the most underbuilt component in voice AI deployments and the most impactful for conversation quality. An AI that knows who it's talking to and why they're calling converts at dramatically higher rates than an AI starting every conversation from scratch.*

5 architectural components required for a production voice AI deployment: session layer, telephony, context injection, action execution, and recording/logging

SECTION 4

Disclosure and Compliance: FTC Guidance and Best Practices

Deploying voice AI for customer conversations without proper disclosure is a significant legal and reputational risk. The regulatory landscape is evolving rapidly, with the FTC issuing guidance on AI disclosure and multiple states enacting specific requirements for AI-handled calls. The FTC's current position on AI voice disclosure: organizations using AI in customer-facing voice interactions must disclose that the interaction is AI-handled if the customer would materially change their behavior or the information they share based on knowing this. Practically, this means disclosure is required for any AI-handled call that collects personally identifiable information, involves sales or financial decisions, or handles sensitive topics. The FTC has signaled that 'deceptive AI' — AI that is specifically designed to pass as human — will face enforcement action. Several states have enacted explicit AI disclosure requirements that go beyond FTC guidance. California's law requires disclosure at the start of any AI-handled call if the caller asks whether they are speaking with a human. Illinois, Texas, and New York have similar or stricter requirements. The regulatory picture is changing fast enough that legal review before deployment is non-negotiable.

Best practice disclosure approach: disclose proactively at the start of every AI-handled call, not reactively when asked. Something as simple as 'Hi, I'm an AI assistant from [Company Name] — I'm here to help with [purpose of call]. You can ask to speak with a person at any time.' This disclosure is short, clear, and positions the AI capability as a feature rather than a liability. Research consistently shows that proactive disclosure has minimal negative impact on call completion rates for appropriate use cases — and that customers who discover non-disclosed AI after the fact generate disproportionate negative sentiment. Beyond disclosure, compliance for voice AI requires: TCPA compliance for outbound AI-initiated calls (written consent requirements), recording disclosure compliance by state (many states require all-party consent for call recording), and data retention and deletion policies for call transcripts and audio recordings consistent with your privacy policy.

- Review FTC AI disclosure guidance and applicable state-level laws before writing a single line of deployment code
- Engage legal counsel with AI and consumer protection experience for the compliance review — this is not an area for internal interpretation
- Implement proactive disclosure at the call start: identify as AI, state the call purpose, offer human transfer option
- TCPA review for outbound AI calls: written consent requirements apply to AI-initiated calls to mobile numbers

- Recording disclosure: implement state-appropriate all-party or one-party consent notifications for call recording
- Data retention policy: document how long call transcripts and recordings are retained and what deletion rights apply

> ***Proactive AI disclosure at call start has minimal impact on call completion rates for appropriate use cases — and avoids the disproportionate backlash that comes when customers discover non-disclosed AI after the fact.***

12+ US states with enacted or pending specific AI disclosure requirements for customer-facing conversations as of early 2026 — making legal review non-negotiable before deployment

SECTION 5

Conversation Design: Scripting for AI (Different from Human Scripts)

Human call scripts are written for humans with human intuitions. They assume the agent can read the room, improvise when the call goes off-script, and exercise judgment about when to push and when to back off. AI conversation design requires fundamentally different principles. The most common mistake in voice AI deployment is handing developers a human call script and asking them to build the AI to follow it. Human scripts create brittle AI conversations that fail the moment a caller says something unexpected. AI conversation design is built around three principles. Principle one: intent coverage, not script following. Instead of designing a linear script, design a conversation that can achieve its objective from multiple entry points and response sequences. Map the universe of intents a caller might express — agreement, objection, confusion, redirection, escalation request — and design responses for each. The conversation should feel natural regardless of which path the caller takes. Principle two: graceful degradation. Design explicit fallback paths for low-confidence responses. When the AI isn't sure what the caller said or means, the response should acknowledge naturally and ask a clarifying question — not loop on misunderstood utterances or deliver a response that ignores the caller's actual input.

Principle three: the handoff as a design element, not a failure mode. Human transfer should be built as a seamless feature of the AI conversation, not as an error state. Design the handoff trigger points (customer requests a human, AI confidence falls below threshold, conversation type exceeds AI scope), the handoff script (what the AI says to the caller before transferring), and the context package (what summary information gets passed to the human agent). Callers who experience a smooth AI-to-human handoff with context preservation rate the overall experience significantly higher than those who have to re-explain their situation to the human agent after transfer. Conversation design is an iterative process. The first deployment version will have gaps — intents that weren't anticipated, phrasing that confuses the model, objection paths that weren't scripted. Build the design for iteration: session logging, QA review process, and a structured conversation design update cadence are as important as the initial design.

- Principle 1 — Intent coverage: map all caller intents (agreement, objection, confusion, escalation) before writing a single response
- Principle 2 — Graceful degradation: design explicit fallback paths for low-confidence responses — no loops, no ignored utterances
- Principle 3 — Handoff as a feature: design the transfer trigger points, the handoff script, and the context package passed to human agents
- Build objection handling paths for the top 5 objections your human agents face — AI encounters the same objections

- Write response variations for the same intent expressed in different ways — callers phrase things unpredictably
- Test with real callers before go-live: at least 50 test calls with your target audience to surface design gaps

> *The handoff to a human agent should be a designed feature, not an error state. Callers who experience a smooth AI-to-human transfer with context preservation rate the experience as highly as fully human calls — the failure is in transfers that lose context.*

50+ test calls with target audience callers recommended before go-live — fewer than this and the conversation design gaps that surface in production will be preventable and costly

SECTION 6

CRM Integration: Giving the AI Context It Needs to Be Useful

An AI-handled call that starts with 'Can you tell me a bit about what brings you in today?' when the caller has been a customer for three years and called twice last month about the same issue isn't just a bad experience — it actively damages the relationship. CRM integration is the difference between a voice AI that knows who it's talking to and one that treats every call as if it's the first contact. The integration has two components: pre-call context injection and post-call data writing. Pre-call context injection: when a call comes in, the caller's phone number triggers a CRM lookup. The lookup returns the relevant account data — company name, contact name, account tier, open opportunities, recent tickets, prior call summaries, and any CRM notes the team has tagged as 'AI-brief-relevant.' This data is injected into the session system prompt in a structured format: not as a data dump but as a contextual brief formatted for the AI to reference naturally. The AI should know this information without making the caller feel surveilled — 'I see you've been looking at our enterprise plan' lands differently than reading back a list of their data.

Post-call data writing: after every AI-handled call, the session summary, call outcome, key information exchanged, and any action items should be written back to the CRM as a call log entry. This serves two purposes: it keeps the CRM current for any human agent who interacts with the caller next, and it builds the data trail that enables AI performance analysis and model improvement. The post-call write should include: a structured summary (2-3 sentences), the call outcome (qualified, scheduled, referred to human, unresolved), key information captured (qualification data, expressed objections, stated timeline), and any action items triggered (follow-up send, appointment booked, escalation routed). Configure the post-call write to happen automatically at session end — not as a manual step that depends on a human reviewing the transcript. CRM integration requirements checklist: does your CRM have an API that supports real-time lookup? Does the lookup performance meet the session start latency budget (typically <500ms)? Does your CRM API support the write-back fields you need?

- Build the caller lookup function: phone number → CRM query → structured context package returned in <500ms
- Format the context package as a brief, not a data dump: relevant account status, prior interaction summary, open items
- Test context injection with edge cases: new contacts (no CRM record), inbound from company main line, multi-contact accounts
- Configure automatic post-call write-back: structured summary, outcome, key data captured, action items — no manual step required

- Define 'AI-brief-relevant' CRM fields with the sales and CS teams — not everything in the CRM belongs in the call brief
- Build the post-call write-back schema collaboratively with the teams who read the CRM — they need to find AI call logs useful

> *Pre-call CRM context injection is the single configuration change that most dramatically improves voice AI conversation quality and conversion rates — it's the difference between a knowledgeable advisor and an impersonal intake form.*

34% average improvement in voice AI conversion rates when pre-call CRM context injection is implemented vs. calls handled with no account context

SECTION 7

Handoff Design: When and How to Transfer to Humans

The handoff from AI to human agent is the most consequential design decision in a voice AI deployment. A well-designed handoff is invisible — the caller experiences a seamless transition with no information loss, and the human agent picks up exactly where the AI left off. A poorly designed handoff requires the caller to re-explain their situation, creates obvious gaps in the agent's knowledge that undermine trust, and produces the category of negative voice AI feedback that is most commonly cited in surveys: 'I had to repeat myself three times to three different people.' Handoff trigger design has three categories. Caller-initiated triggers: any time the caller explicitly requests a human — 'I want to talk to a person,' 'Can I speak to a manager,' 'This isn't helping.' These should immediately initiate handoff with no friction. The AI should never attempt to dissuade a caller from requesting human transfer. System-initiated triggers: when the AI confidence falls below a defined threshold, when the conversation has exceeded a defined duration without reaching the intended outcome, or when the conversation has entered a topic category outside the AI's designated scope. Escalation-triggered handoffs: when the caller signals high frustration, urgent need, or a sensitive situation that the AI detects through tone or content — even if the caller hasn't explicitly requested a human.

The context package at handoff is as important as the transfer trigger. When the AI transfers a call, the human agent should receive before the caller comes on the line: a 2-3 sentence call summary, the caller's account context (from the CRM brief), the key points covered in the AI conversation, any information the caller provided, and the reason for transfer. This context delivery should be automated — spoken as a brief pre-briefing to the agent before the transfer completes, or pushed as a screen pop in the agent's interface. The handoff script the AI uses with the caller: warm transfer language, not cold. 'I'm going to connect you with one of our team members right now — I've shared everything we've discussed so you won't need to repeat yourself' sets the right expectation and reduces the negative experience of the transfer itself.

- Caller-initiated triggers: any explicit human request — no friction, no AI attempt to redirect, immediate transfer
- System-initiated triggers: define confidence threshold, maximum conversation duration, and out-of-scope topic categories
- Escalation triggers: configure sentiment detection for high frustration or urgency signals even before explicit request
- Context package: automate pre-briefing delivery to agent before transfer completes — spoken briefing or screen pop

- Handoff script: warm transfer language that sets expectation of context continuity ('I've shared everything we've discussed')
- Test every handoff path before go-live: caller-initiated, system-initiated, and escalation-triggered transfers

> *The handoff script that tells the caller 'I've shared everything we discussed so you won't need to repeat yourself' sets an expectation you must deliver on — the human agent needs that context package before the caller speaks, not after.*

3× increase in negative voice AI experience ratings when callers must re-explain their situation to the human agent after transfer vs. transfers with full context continuity

SECTION 8

Testing and Quality Assurance Before Go-Live

Voice AI deployments that skip structured testing produce preventable production failures: conversation paths that loop on misunderstood utterances, CRM lookups that return wrong data or time out, handoff transfers that lose context, and disclosure language that doesn't satisfy compliance requirements. The testing protocol has three phases. Phase one: conversation design testing. Before any technical integration, test the conversation design with human role-players taking the caller role. Walk every designed intent path, every objection path, every escalation trigger. Identify gaps in the intent coverage and responses that don't feel natural. This phase can be run with simple voice or text interaction — the goal is validating the conversation logic, not the technical implementation. Phase two: integration testing. With the conversation design validated, test the full technical stack: CRM lookup performance and accuracy, context injection fidelity, action execution (appointment booking, CRM write-back, follow-up trigger), handoff transfer quality, and session recording. Run test calls with synthetic caller profiles representing your target segments. Measure latency at every step. Test edge cases: unknown callers (no CRM record), callers with complex histories, calls that hit escalation triggers.

Phase three: live pilot testing. Before full production release, run a limited pilot with real callers on a defined subset of your call volume — typically 5-10% routed to the AI system, with all others going to human agents. Monitor every AI call in the pilot. Review session recordings daily. Track conversion rate, handoff rate, caller sentiment, and QA scores. Run the pilot for a minimum of two to three weeks before making go/no-go decisions. The go-live threshold: the AI should be achieving at minimum 80% of human agent performance on your primary success metric (conversion rate, qualification rate, appointment set rate) in the pilot before full production deployment. Below that threshold, identify the specific conversation design or technical issues causing underperformance and iterate before expanding. A staged rollout — 10% to 25% to 50% to 100% of volume — gives you data and course-correction opportunities at each stage.

- Phase 1 — Conversation design testing: human role-players walk every intent path before technical integration
- Phase 2 — Integration testing: test CRM lookup, context injection, action execution, handoff, and recording with synthetic caller profiles
- Phase 2 edge cases: test unknown callers, complex histories, escalation triggers, out-of-scope topics
- Phase 3 — Live pilot: 5-10% of call volume, minimum 2-3 weeks, daily review of session recordings

- Go-live threshold: minimum 80% of human agent performance on primary success metric before full rollout
- Staged rollout: 10% → 25% → 50% → 100% with performance validation at each stage

> *The 80% of human performance threshold is the minimum go-live standard, not the target. Most well-designed voice AI deployments reach 80-90% of human performance within 60 days of go-live through conversation design iteration.*

80% of human agent performance on the primary success metric is the minimum threshold for voice AI go-live — below this, specific conversation design gaps need to be addressed before full deployment

SECTION 9

Performance Measurement: The Metrics That Matter for Voice AI

Voice AI performance measurement needs to cover three levels: technical performance (is the system working?), conversation performance (is the AI having effective conversations?), and business performance (is it contributing to revenue outcomes?). Technical performance metrics: session success rate (percentage of initiated calls that complete without system error), average response latency (should consistently stay below 300ms), CRM lookup success rate, and handoff transfer completion rate. These metrics should be monitored daily with automated alerts for degradation. A response latency spike or CRM lookup failure rate above 2% should trigger immediate investigation — these are the technical issues that produce the worst caller experiences. Conversation performance metrics: call completion rate (callers who stay through the conversation to a defined endpoint), primary intent success rate (percentage of calls that achieve the AI's designated objective — qualification, appointment set, etc.), handoff rate (percentage of calls transferred to human agents), and caller sentiment score (derived from session transcript analysis). The handoff rate is a particularly useful diagnostic: if it's trending up, the conversation design is failing to serve more caller intents than intended.

Business performance metrics: the metrics that connect voice AI activity to revenue outcomes. For B2B qualification calls: qualified lead rate (percentage of AI-handled calls that produce a SQL), pipeline contribution (revenue value of opportunities originating from AI-qualified calls), cost per qualified conversation (fully loaded AI infrastructure cost divided by qualified conversations), and conversion rate comparison (AI-handled vs. human-handled calls for the same conversation type). Report business performance on a 30-day rolling basis and compare AI-handled cohorts to human-handled cohorts with as many confounding variables controlled as possible — same time period, same lead source, same product line. The ROI summary metric: compare cost per qualified conversation AI vs. human. Most mature B2B voice AI deployments achieve 40-65% cost reduction per qualified conversation while maintaining 80-90% of human conversion rates — the ROI calculation for this ratio is typically straightforward at any reasonable volume.

- Technical metrics: session success rate, response latency, CRM lookup success rate, handoff transfer completion — daily monitoring with automated alerts
- Conversation metrics: completion rate, primary intent success rate, handoff rate, caller sentiment score
- Business metrics: qualified lead rate, pipeline contribution, cost per qualified conversation, AI vs. human conversion rate comparison
- Alert threshold: >2% CRM lookup failure rate or >300ms average latency triggers immediate investigation

- Report business performance on 30-day rolling basis with AI vs. human cohort comparison for the same conversation type
- ROI target: 40-65% cost reduction per qualified conversation at 80-90% of human conversion rate is achievable benchmark

> *Handoff rate trending upward is the earliest indicator of conversation design gaps — it means callers are requesting humans or triggering system escalations at higher rates than intended, which points to specific conversation paths that need redesign.*

40–65%

cost reduction per qualified conversation achievable with mature voice AI deployment vs. fully human-handled calls — the ROI benchmark for B2B voice AI go/no-go decisions

Voice AI Deployment Implementation Checklist

Phase 1 — Foundation

- [ ] Run use case qualification framework across top 5-10 customer conversation types
- [ ] Select 2-3 highest-qualifying use cases for initial deployment
- [ ] Engage legal counsel for FTC guidance review and applicable state-level disclosure compliance
- [ ] Draft disclosure language and human transfer offer for all AI-handled call openings
- [ ] Select telephony provider with WebSocket audio streaming support for Realtime API integration
- [ ] Map CRM lookup requirements: which fields needed, API performance specifications, edge cases
- [ ] Define the post-call write-back schema with sales and CS team input

Phase 2 — Launch

- [ ] Complete conversation design: intent coverage map, objection paths, escalation triggers, handoff design
- [ ] Phase 1 testing: human role-players walk all conversation paths, identify design gaps
- [ ] Build and test CRM integration: lookup performance, context injection, post-call write-back
- [ ] Phase 2 integration testing: full stack with synthetic caller profiles, all edge cases
- [ ] Configure session recording, logging, and QA review workflow
- [ ] Deploy disclosure compliance: proactive disclosure script, all-party consent recording notification where required
- [ ] Launch 5-10% pilot on qualified call volume, monitor daily for 2-3 weeks

Phase 3 — Optimize

- [ ] Evaluate pilot against 80% of human performance go-live threshold
- [ ] Document conversation design gaps from pilot and iterate before expanding volume
- [ ] Set up automated performance monitoring: latency, session success, CRM lookup rates
- [ ] Build 30-day rolling business performance report with AI vs. human cohort comparison
- [ ] Execute staged rollout: 25% → 50% → 100% with performance validation at each stage
- [ ] Schedule quarterly conversation design review based on accumulated session data

NetWebMedia

Build Voice AI That Actually Converts — Not Just a Talking Phone Tree

NetWebMedia designs and implements OpenAI Realtime API voice AI deployments for B2B organizations — from use case qualification and conversation design through CRM integration, compliance review, and performance measurement. We've built voice AI systems for qualification, appointment setting, and onboarding use cases, and we know where the design mistakes happen that produce poor caller experiences and below-threshold conversion rates. If you're evaluating voice AI for your customer conversation stack, we can tell you whether your use case qualifies and what a successful deployment looks like.

- AI Marketing Automation
- AEO & AI-First SEO
- Autonomous AI Agents
- Paid Media + AI Creative
- CRM + AI Workflows

netwebmedia.com/contact