COVAL

Voice AI 2026

What happened in 2025 and what it means for voice agents in 2026

A Strategic Guide for scaling Enterprise Voice AI

Executive Summary

The 2025 Breakthrough

Infrastructure finally reached production-grade quality. We saw a dramatic reduction in latency and a reduction in response times. Speech recognition accuracy improved by 54%. Costs collapsed 60-87% across the entire stack. The market reached $10.3B with 51% year-over-year growth.

You can now build voice agents that actually work.

The 2025 Reality Check

Perfect demos consistently failed when deployed to production. Week one success rates reached 95% in controlled demo environments, but only 62% with real customers in real-world conditions.

This wasn't because the technology wasn't ready but because deployment methodology wasn't ready.

The Critical Insight for 2026

Your competitive advantage won't come from having access to the newest model or achieving the lowest latency. It will come from systematic deployment discipline: comprehensive testing infrastructure, multi-model orchestration capability, and treating voice agents as learning systems that improve with every conversation.

What This Report Covers

We interviewed industry leaders in voice AI and analyzed deployment patterns across thousands of production voice agents.

This report shares what actually works, what doesn't, and what you need to plan for 2026.

Part 1

The 2025 Market Evolution

Understanding What Changed and Why It Matters

How the Conversation Changed

From “How Human Does It Sound?” to “What’s the Resolution Rate?”

In early 2024, enterprise conversations about voice AI centered on demos and surface-level capabilities. Questions like, How human does it sound? Can it really handle interruptions? Is it better than traditional IVR systems?

By late 2025, those conversations fundamentally shifted to business outcomes.

> In 2025 the conversation stopped being about how human the bot sounds and became about resolution rate and handle time.
>
> — Enterprise Voice AI Provider

Enterprises now buy voice AI based on measurable operational impact:

- **Resolution rate and containment metrics**: What percentage of calls does the agent successfully resolve?
- **Average handle time reduction**: How much faster are conversations compared to human agents?
- **Human agent productivity gains**: How much time do we free up for complex issues?
- **Post-escalation outcomes**: What happens when calls transfer to humans?
- **End-to-end customer journey**: Is the complete experience better?

How the Conversation Changed

The User Acceptance Inflection Point

The question every executive asks: Are customers actually willing to talk to bots? The answer from production data is definitively yes.

Enterprises have measured this directly in their deployments:

> Drop-offs have decreased dramatically when the experience is good. People are getting used to voice bots and are pleasantly surprised.
>
> — Enterprise Voice AI Provider

This isn't aspirational — it's happening now in production. Bot recognition drop-off rate (the percentage of users who hang up when they realize they're talking to AI) has become a canonical KPI that enterprises track, and it's declining sharply across the industry.

What this means for You

Stop optimizing for "wow" factor in controlled demos. Start optimizing for the metrics that matter to your C-Suite: cost per resolution, human agent hours saved, customer satisfaction maintained or improved versus baseline.

The market matured in 2025. Your evaluation criteria and deployment strategy should mature with it.

> Voice agents can now be as productive as a human agent. We‘re going to see a lot more shift towards voice.

Dhivya Rajprasad, ,Product for Zoom CXCX

Voice Resurgence—Why Voice Is Winning Over Chat

Why Voice Is Winning Over Chat

For the past decade, contact centers systematically pushed customers toward chat channels. The reasoning was sound: chat was cheaper to operate, more scalable with traditional chatbots, and easier to automate. Voice remained the expensive channel reserved for complex issues requiring human empathy and nuanced problem-solving.

2025 fundamentally flipped this assumption.

The Strategic Shift

The math changed completely. When voice agents achieve 75-85% resolution rates at a fraction of human agent cost, with sub-500ms latency creating natural conversation flow that feels genuinely human, voice becomes the scalable default channel again — not chat.

We’re observing enterprises make counterintuitive strategic bets: organizations with extensive chatbot expertise are now prioritizing voice-first development over expanding their existing chat capabilities.

Voice Resurgence—Why Voice Is Winning Over Chat

Why Voice Is Winning Over Chat

Drop-off rates are declining when users hear voice bots. When quality is high, users are comfortable engaging.

Voice feels more natural for complex multi-turn conversations. Chat requires typing effort and reading comprehension.

Emotion and tone convey critical information faster and more accurately than text alone.

Accessibility advantages are significant: hands-free operation, support for vision-impaired users, multilingual flexibility.

What This Means for You

Challenge the default assumption of chat-first strategies. If you're planning 2026 voice AI deployments, seriously evaluate voice-first architectures.

The TAM expansion opportunity is real — you're not competing in crowded chatbot markets where differentiation is difficult. You're opening entirely new automation channels where the competitive landscape is still forming.

Part 2

The Infrastructure Reality

Multi-Model World and Architecture Decisions

> I am 100% convinced we‘re living in a multi-model world, and figuring out how to use models together is one of the really interesting software engineering questions.

Kwindla Hultman Kramer, Creator of Pipecat

The Multi-Model Reality

The Era of Single Models Is Over

If your voice AI architecture relies on a single LLM doing everything — conversation, reasoning, function calling, safety — you're already at a disadvantage. Production systems in 2025 orchestrate multiple specialized models in parallel.

Why Single Models Can't Win

No single model simultaneously optimizes for speed, reasoning, and cost. Physics and economics prevent it.

Speed-Optimized Models: Sub-200ms response times require smaller parameters and aggressive streaming. These sacrifice reasoning depth for velocity.

Reasoning-Optimized Models: Complex multi-step logic demands larger models with higher token limits. Slower, but dramatically more capable.

Cost-Optimized Models: High-volume tier-1 support needs 10x cheaper inference. Acceptable latency trade-offs enable massive cost reduction.

The constraint: Larger models are slower (physics). More capable models cost more per token (economics).

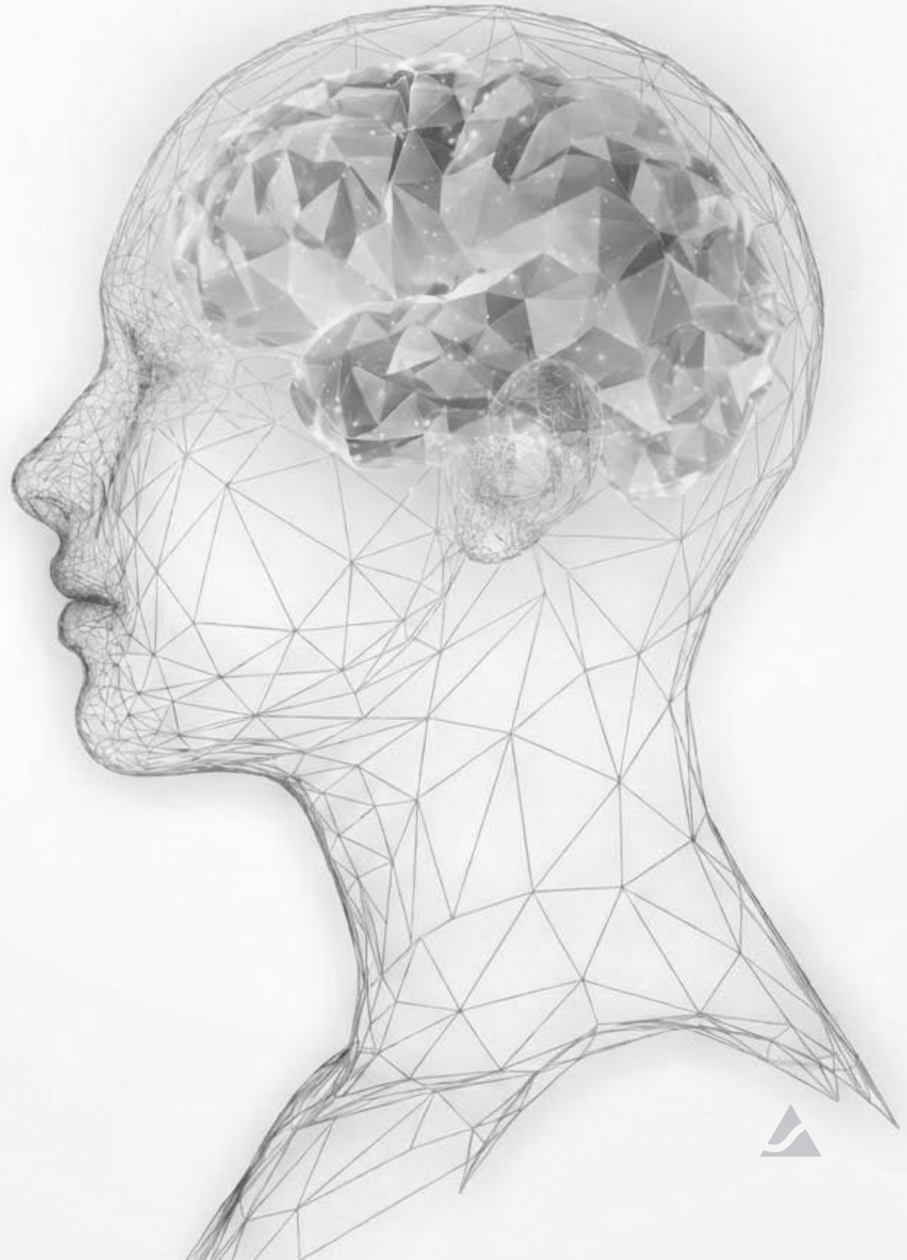
The Multi-Model Reality

Domain Specialization Compounds the Advantage

Function calling models fine-tuned for API interaction outperform general-purpose alternatives.

Emotional intelligence models trained on empathy and tone excel at sensitive interactions.

Compliance-native models for healthcare (HIPAA), legal (privilege), and finance (fiduciary duty) understand regulatory requirements at the model level.



The Multi-Model Reality

Real Production Architecture Example

Customer Service Agent Using Five Models in Parallel:

- **Primary Conversation**: Fast streaming LLM for natural dialogue, sub-300 ms latency
- **Function Calling Specialist:** Optimized for API interactions, database queries, CRM updates with structured output
- **Sentiment Analysis:** Real-time emotional detection triggering escalation or empathy adjustments
- **Guardrails**: Independent safety layer checking responses before delivery, ensuring brand compliance
- **Fallback**: Lightweight backup ensuring service continuity during failures or rate limiting

Hybrid Deployment Adds Complexity

- **On-device inference**: Privacy, zero latency, offline capability
- **Cloud inference:** Computational power, model size flexibility
- **Edge computing**: Reduced latency in specific regions
- **Fallback mechanisms**: Service continuity across infrastructure failures

Coordinating These Models Requires:

- Routing logic determining which model handles which task
- State management maintaining conversation coherence across transitions
- Context sharing enabling models to build on each other's outputs
- Debugging capabilities isolating which model caused failures

The State of the Art Five-Layer Stack of 2025

Layer 1

Listening
Speech-to-Text

2025 pushed streaming accuracy and multilingual capability. Providers achieved 50%+ error reduction while supporting 30+ languages with automatic detection. Domain-specific models for healthcare, finance, and legal emerged as requirements for regulated industries.

Key capability: Real-time transcription accurate enough for compliance-critical deployments.

Market Leaders: Deepgram, AssemblyAI

Layer 2

Speaking
Text-to-Speech

2025 solved the expressiveness-latency trade off through model specialization. Sub-100 ms synthesis became standard for real-time conversations. Voice cloning and steerable tone control enabled brand differentiation.

Key capability: Human-quality voice with emotional depth at conversational speed.

Market Leaders: ElevenLabs, Cartesia, Rime for custom voices

Layer 3

Thinking
Language Models

2025 delivered sub-500 ms response times with sophisticated tool calling. Models now handle complex multi-step workflows, API integrations, and database queries as standard capabilities. Multimodal processing combined vision and audio.

Key capability: Agents that execute actions, not just converse.

Market Leaders: OpenAI, Anthropic, Google

Layer 4

Coordinating
Orchestration

2025 standardized modular architectures enabling provider swaps without code rewrites. Vendor-neutral orchestration became table stakes for cost optimization and avoiding lock-in.

Key capability: Architectural flexibility and quarterly provider optimization.

Market Leaders: Pipecat, LiveKit

Layer 5

Hearing Clearly
Noise Cancellation

2025 achieved robust background voice filtering running on edge devices. Real-world deployment in noisy environments became viable with dramatically reduced false interruptions and improved transcription accuracy.

Key capability: Production-grade performance in uncontrolled acoustic environments.

Market Leaders: Krisp

A note on layers: These five layers represent the foundational infrastructure every voice agent requires. 2025 demonstrated that production deployments increasingly customize additional components including turn detection, interruption handling, audio preprocessing, voice activity detection, and conversation state management based on specific use case requirements.

A Note on Speech-to-Speech

2025: The Year S2S Emerged

Speech-to-speech models burst onto the scene with 85% latency reduction, collapsing pipelines from 2000 ms to 200-300 ms. The technology preserved emotional prosody that text-based systems couldn’t match, showcasing compelling use cases where natural interaction quality created genuine advantages.

Enterprise Reality: Reliability Trumps Naturalism

Enterprise buyers still evaluate on resolution rates, handle time reduction, and containment metrics—not conversation naturalness. Regulated industries require predictable performance with audit trails. The 90%+ success rate threshold remains non-negotiable.

Why Cascaded Architecture Still Dominates in 2026

Traditional cascaded architectures maintain dominance for four reasons:

1. More control through text intermediaries enabling compliance checks
2. More fallbacks ensuring service continuity
3. More debuggability isolating failures to specific layers,
4. More mature evaluation and compliance tooling.

Speech-to-Speech: The Promise & The Reality

Coval's 2026 S2S Prediction

H1 2026: Traditional pipeline dominates with less than 15% S2S adoption. Market focus remains on scaling tool-using agents where traditional architecture offers mature tool calling, proven debugging, and established compliance.

H2 2026: S2S crosses production viability as four enablers unlock adoption:

1. Audio-native evaluation tools mature
2. Debugging capabilities catch up
3. Compliance frameworks adapt
4. Cost parity is achieved.

The Intelligent Strategy

- **Deploy S2S for:** Premium support, basic consultations, therapy/coaching, multilingual scenarios where emotional connection drives value.
- **Maintain traditional for:** High-volume tier-1 support, regulated industries requiring audit trails, complex tool calling, debt collection compliance.
- **Testing Infrastructure matters:** look for audio-native evaluation tools, emotion preservation metrics, and interruption handling measurement.

Perfect traditional fundamentals in H1 2026, experiment with S2S on 5-10% of traffic in H2. Winners will run both architectures optimized independently.

> I think we're going to shift towards thinking about the job we're doing with LLMs as moving everything as much as possible to natural language. There's a tension here because our old software engineering instincts are to wrap these LLMs in deterministic guardrails and APIs and evals. But if you do that, you never get where you want to get.

Kwindla Hultman Kramer, Creator of Pipecat

Key insight: These are natural language machines. We need to figure out patterns for software engineering where it's natural language in and it's mostly natural language out.

Natural Language as Engineering Paradigm

Why Traditional Software Engineering Fails with LLMs

The If-Then-Else Trap

When conversations go off the rails, traditional engineering instinct builds catch statements for every conceivable edge case. But conversations went wrong precisely because you didn't know that particular corner case existed beforehand. You always end up in the finally block thinking: I don't know what to do now.

The problem: You're trying to anticipate every possible conversation path. With infinite variability in human language, this approach is fundamentally doomed to fail.

The Guardrails Paradox

Wrapping non-deterministic systems in deterministic constraints limits their capability to exactly the scenarios you anticipated during development. LLMs excel at handling novel situations you never explicitly programmed for. Rigid if-then-else guardrails prevent this adaptive intelligence from functioning effectively.

You're fighting against the system's core strength: flexible natural language understanding and generation.

Natural Language as Engineering Paradigm

Instead of if-then-else statements catching specific keywords or patterns, build guardrails processes that express everything you want the model to do in natural language.

Guardrails Agents Know

- The full system prompt and all instructions
- Complete conversation context that happened so far
- User sentiment signals and frustration indicators
- Domain knowledge and business rules
- Escalation protocols and recovery strategies

They can intelligently reset conversations using this wealth of contextual information.

Natural Language as Engineering Paradigm

Practical Example Comparison

Traditional Deterministic Approach

- IF user says inappropriate_keyword THEN respond with canned_apology
- IF user repeats question 3 times THEN escalate to human
- IF conversation duration exceeds 10 minutes THEN offer callback
- ELSE fallback to generic error message

Result: Brittle system only handling anticipated scenarios. Poor user experience when novel situations arise. Constant maintenance adding new if-then rules.

Natural Language Approach

Guardrails agent instruction

Ensure conversation stays productive and helpful. If user becomes frustrated, confused, or conversation is not progressing toward resolution, use your understanding of the full context to guide conversation back on track. You have access to the system prompt, all conversation history, user sentiment signals, and domain knowledge. Intelligently determine best path forward.

Result: Adaptive system handling novel situations. Maintains conversational coherence. Better user experience. Fewer hard-coded rules to maintain.

Natural Language as Engineering Paradigm

The Result

Adaptive system handling novel situations. Maintains conversational coherence. Better user experience. Fewer hard-coded rules to maintain.

What This Means for Evaluation

You cannot evaluate natural language systems with only deterministic checks. Regex matching and keyword detection catch some problems but miss most interesting failures requiring contextual understanding.

Need hybrid measurement: Machines (latency, compliance checks) + AI judges (conversation quality, tone appropriateness) + Humans (ground truth validation, discovering new patterns).

This is why comprehensive simulation and systematic testing are non-negotiable for production deployment.

Part 3

The Deployment Gap

Why Perfect Demos Fail in Production

Why Perfect Demos Fail in Production

The Controlled Demo Environment

- ✓ Quiet rooms with professional studio-quality microphones
- ✓ Native English speakers with clear articulation and standard pronunciation
- ✓ Linear conversations following expected scripted flows
- ✓ Single test persona, no interruptions or topic changes
- ✓ Controlled acoustic conditions with no background noise

Result: 95%+ success rate that looks production-ready

The Production Reality

- ✗ Speaker phones in cars with road noise, restaurants with ambient chatter, busy streets with traffic
- ✗ 100+ different accents, dialects, and speech patterns reflecting global user diversity
- ✗ Users who interrupt mid-sentence, change their minds, multitask while talking
- ✗ Background TV audio, other conversations, crying children, barking dogs
- ✗ Uncontrolled acoustic environments creating unpredictable audio quality

Result: Week 1 production success drops to 62%

Why Perfect Demos Fail in Production

Root Cause Analysis

Infrastructure is ready. 85% latency reduction achieved. 54% accuracy improvement delivered. Costs collapsed 60-87%. The technology works.

The barrier is not technological capability. The barrier is deployment methodology.

Real users are fundamentally messy and unpredictable. Real acoustic environments have uncontrolled characteristics. Real conversations include topic changes, interruptions, ambiguous requests, contextual references. If you only test in controlled conditions, you're building for a world that doesn't exist in production.

The Solution: Systematic Simulation Testing

- Simulate millions of conversations covering comprehensive edge cases BEFORE production deployment. Just as autonomous vehicles require millions of simulated miles before road deployment, voice agents need thousands of simulated conversations before customer deployment.
- Evaluate with hybrid metrics combining deterministic checks, AI-based judges, and strategic human review.
- Monitor continuously in production with automated evaluation of every conversation.
- Iterate based on real user failure data feeding back into test library for continuous improvement.

Three-Layer Testing Framework

Layer 1

Regression Testing

50-100 core scenarios covering happy path conversations and expected user flows.

Layer 2

Adversarial Testing

20-30 edge cases specifically targeting background noise, diverse accents, user interruptions, complexity.

Layer 3

Production-Derived Testing

Continuous learning system converting real production failures into new test cases.

Outcome

What This Means for You

Companies that invest in systematic testing achieve 90%+ production success within 3 months, with predictable performance and rapid iteration cycles.

Those that skip testing start at 62% success, take 6-9 months to reach 85% reliability, and spend significantly more on production firefighting.

The Evaluation Adoption Gap

The Harsh Reality in Late 2025

Many production voice AI deployments processing millions of conversations monthly have zero to minimal evaluation infrastructure. The reason cited by teams: It's hard.

The Recommended Sequence Most Teams Skip: Building evals by hand, then integrate them into DevOps tooling, and then layer simulation-based testing.

Most teams aren't even on step one. They cannot improve their voice agents in production because they have no systematic way to measure what's working versus what's broken. Without measurement, optimization is impossible.

Why Evaluation Adoption Lags Deployment

Building effective evaluation infrastructure requires deep expertise in:

- Natural language processing and semantic understanding
- Audio quality assessment and acoustic analysis
- Conversation flow dynamics and dialogue coherence
- Business metrics alignment with technical measurements
- Statistical sampling methodology and test design

Most teams deploying voice agents don't have this multidisciplinary expertise in-house and significantly underestimate the complexity involved. At the same time, there's a perceived time-to-market pressure ("ship now, optimize later").

Result: Ship without evaluation capability, discover serious problems with real paying customers, spend 10x more time firefighting production incidents than proper evaluation infrastructure would have cost upfront.

The Build vs. Buy Calculation for Evaluation Infrastructure

Building comprehensive evaluation in-house requires

- 6-12 months of dedicated engineering time
- Specialized expertise across NLP, audio, conversation design
- Ongoing maintenance as models and use cases evolve
- Integration with existing DevOps and monitoring infrastructure

Using third-party evaluation platforms requires

- 2-4 weeks to full production deployment
- Immediate access to specialized expertise and proven methodologies
- Continuous platform improvements without internal engineering
- Integration support and best practices from production deployments

Most teams try to build in-house, underestimate scope, fail to deliver comprehensive solution after 6+ months, then adopt platform after expensive production incidents force the decision.

The Build vs. Buy Calculation for Evaluation Infrastructure

What Evaluation-First Teams Do Differently

- **Start with Hand-Built Evals**: Manual tests of 10-20 core flows catch 30-40% of issues and build evaluation culture.
- **Progress to DevOps Integration**: Automated CI/CD evaluation blocks catastrophic failures before production.
- **Scale with Simulation**: Comprehensive testing across thousands of scenarios makes reliability a competitive advantage.

The 2026 Opportunity

Evaluation and observability tools becoming more mainstream throughout 2026 as adoption diffuses across the industry.

Companies adopting evaluation-first approach in Q1 2026 will have 12-18 month advantage over competitors still operating without systematic testing.

The gap between leaders with systematic testing and laggards without evaluation will widen dramatically as S2S and multi-model complexity increases.

Part 4

Enterprise Implementation Reality

What Actually Works in Production

> People thought the voice agent is so cheap, it would just work from day one. The realization comes that it involves a lot of things.

Enterprise Voice AI Provider

The Professional Services Reality Check

Why Voice Agents Aren't Plug-and-Play

2025 showed that Voice agents require a similar level of professional services intensity as full contact center deployments.

And it will take time to reduce friction, especially with regards to integrating with contact centers.

> For contact centers, PSO involvement will remain the norm for at least the next couple of years.
>
> — Enterprise Voice AI Provider

The 2026 Opportunity

Voice Quality Management: Accent robustness, noise handling in uncontrolled environments, audio consistency across devices and networks.

Brand Customization: voice cloning and tone guidelines, conversational style matching brand personality, personalization.

Domain Expertise: Knowledge base creation and maintenance, use case-specific prompt engineering, integration with contact center platforms and CRM.

The Professional Services Reality Check

Why Voice Agents Aren't Plug-and-Play

Voice deployment requires expertise in THREE domains simultaneously:

- **Conversation Design**: Flow architecture, turn-taking, error recovery, escalation
- **Brand Design**: Voice selection, tone guidelines, personality definition
- **ML/LLM Tuning**: Prompts, tools, knowledge bases, guardrails

Most organizations have expertise in one, maybe two — almost none have all three.

Technology vs Skill Gap

Voice AI infrastructure advances monthly, but building internal expertise takes months — creating a persistent deployment bottleneck that technology alone cannot solve.

Organizations need professional services for expert guidance, forward-deployed engineers bridging platform and customer expertise, evaluation platforms enabling testing without deep ML knowledge, and strategic hiring of conversation designers and AI ops engineers.

Even platform leaders require PSO for contact center deployments despite building self-serve tools, proving the skills gap is structural, not temporary.

Voice Agents as Learning Systems

The Paradigm Shift & Learning Loop

Winners in 2026 don't treat voice agents as static endpoints. They build continuous improvement loops where every conversation makes the system smarter.

VA → Human Handoff

- VA attempts resolution, captures full conversation data
- Escalations include complete context (no customer repetition)
- Reduces human agent handle time dramatically

Analysis → Recommendations

- ML identifies patterns in what humans handle that VA couldn't
- System recommends new automation: "500 password reset calls—add this skill"
- Knowledge base auto-updates based on human agent responses

Continuous Improvement

- Routing intelligence improves (better escalation decisions)
- Resolution rates increase without manual intervention
- System compounds improvement with every conversation

Voice Agents as Learning Systems

What This Means for You

Architect for the learning loop from day one

Data flow: Capture every VA conversation with full fidelity, ensure seamless context handoff to human agents, track human agent patterns post-escalation, feed learnings back to VA improvement pipeline.

Reporting infrastructure: End-to-end visibility combining VA and human metrics, track week-over-week improvement trajectories, identify automation expansion opportunities based on data, measure ROI continuously.

Continuous improvement process: Weekly VA performance reviews, monthly knowledge base updates driven by pattern analysis, quarterly strategy reviews evaluating new skills to add, automated ML-driven recommendations.

The competitive advantage in 2026 isn't deploying the best voice agent on day one — it's building the best learning system that improves faster than competitors.

Part 5

Key Takeaways

What You Need to Remember for 2026

Five Critical Takeaways for 2026

Takeaway 1

Infrastructure Matured, but Demo to Production Gap Emerged

Technology reached production-grade in 2025: 85% latency reduction, 54% accuracy improvement, 60-87% cost collapse.

However, Week 1 production success dropped to 62% despite 95% demo success.

The barrier is no longer technological capability — it's systematic deployment methodology.

Takeaway 2

Systematic Testing Infrastructure Is Non-Negotiable

Dedicate 20-30% of total investment to evaluation. $50K spent on evaluation prevents $500K in production firefighting.

Three-layer framework required: regression testing (50-100 scenarios), adversarial testing (20-30 edge cases), production-derived testing (continuous learning).

Takeaway 3

The Multi-Model Reality

Production systems orchestrate multiple specialized models: speed-optimized, reasoning-optimized, cost-optimized, domain-specific.

No single model can simultaneously optimize for all requirements. Build natural language guardrails that understand context. Requires hybrid evaluation: machines + AI judges + humans.

Takeaway 4

Voice Agents as Learning Systems

Winners build continuous improvement loops: VA conversations → human agent patterns → ML-driven recommendations → knowledge base updates → improved routing → repeat.

Not static deployments, but living infrastructure that gets smarter with every conversation.

Takeaway 5

PSO-Heavy Reality and Voice Resurgence

Voice still requires services intensity: voice quality tuning, brand representation, personalization, three-dimensional expertise. Technology improving faster than skills.

Voice-first strategies open new TAM versus crowded chatbot markets as quality improves and drop-off rates decline.

Coval Is Your Partner in Systematic Deployment

What Coval Provides

Simulation Infrastructure at Production Scale

Generate thousands of diverse conversations before production. Catch 80% of issues in simulation versus expensive customer discovery. Enable 10-20x faster prompt iteration cycles based on data, not guesswork.

Evaluation Methodology That Actually Works

Hybrid metrics: technical (latency, audio-quality) + AI judges (conversation quality, tone) + humans (ground truth). Customizable for domain-specific requirements (HIPAA, finance, legal). 5x faster false positive refinement for AI-judge-metrics

Production Monitoring with Continuous Learning

Automated evaluation of every production call. Real-time anomaly detection identifying edge cases. Test-case library grows continuously. Metrics improve through human annotations.

Why Coval Exists

Most teams try building evaluation in-house, underestimate complexity, fail after 6-12 months. Coval gets you operational in 2-4 weeks with proven methodologies.

About This Report

Built from Real-World Deployment Experience

This report synthesizes insights from thousands of voice AI deployments observed throughout 2025 — from successful production rollouts to costly failures that revealed critical deployment gaps. We analyzed patterns across enterprise contact centers, healthcare systems, financial services, and consumer applications to identify what actually works versus what sounds good in demos.

16 Industry Leaders Interviewed

We conducted in-depth interviews with 16 leaders spanning the voice AI ecosystem: vertical players building specialized voice solutions, architecture providers enabling modular infrastructure, and enterprise voice AI platforms deploying at scale. These conversations revealed the systematic deployment methodologies separating companies achieving 90%+ production success from those struggling at ~60%.

Coval's Unique Vantage Point

Coval builds evaluation infrastructure that sits on top of voice AI deployments, providing visibility into production performance across hundreds of implementations. This positioning gives us unparalleled insight into what breaks in production, which testing strategies prevent failures, and how the evaluation adoption gap creates competitive advantages for systematic deployers.

The Result

A strategic guide for 2026 based on what actually works in production — not vendor marketing or theoretical best practices. The companies winning in voice AI aren't deploying the newest technology first. They're building the most reliable systems through systematic testing, multi-model orchestration, and treating voice agents as learning systems that improve with every conversation.

Coval Case Studies

Financial Services

Validate AI agents used for account management, payment and billing support, and loan applications so responses are accurate, and compliant.

Healthcare

Validate agents used for appointment scheduling, patient check-ins, and new member onboarding to ensure accuracy, consistency, and safe.

HR and Recruiting

Validate AI agents used for candidate screening calls so evaluations are accurate, unbiased, and reliable at scale.

COVAL

www.coval.dev

@covaldev

LinkedIn

sales@coval.dev