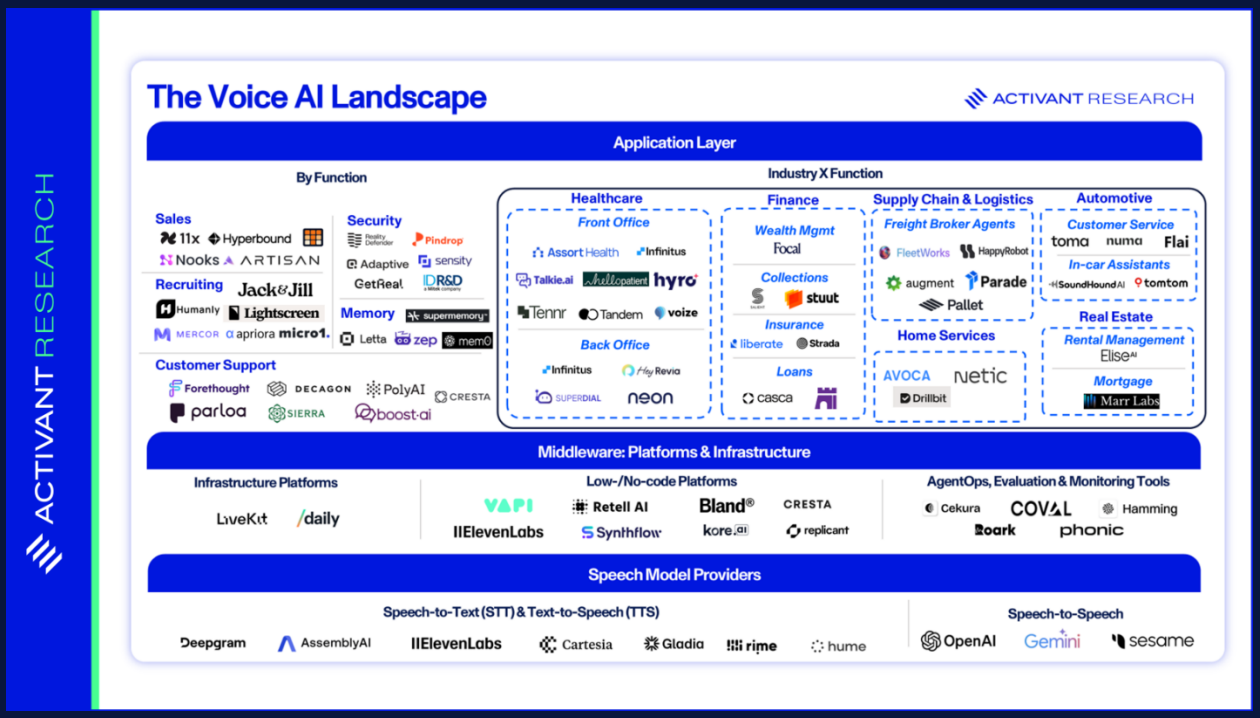
ACTIVANT RESEARCH

Voice Agents 2.0

From “Sounds Human” to “Thinks Human”



Emma Rowand, Scott Watson

Q4 2025

When we first published *Conversational AI and the Future of Customer Service* in early 2024, the market thesis was centered on realism: Could generative AI replicate natural, multi-turn dialogue well enough to move beyond clumsy IVR systems and high-turnover human call centers? At the time, we focused on how AI would disrupt the $314 billion call center market by achieving human parity in sound and flow.[1] Almost two years later, that feels outdated. Infrastructure improvements have pushed response latencies below 300 milliseconds, aligning AI interaction with natural human speech. The standard for realism itself has elevated significantly and what seemed ambitious is now table stakes.

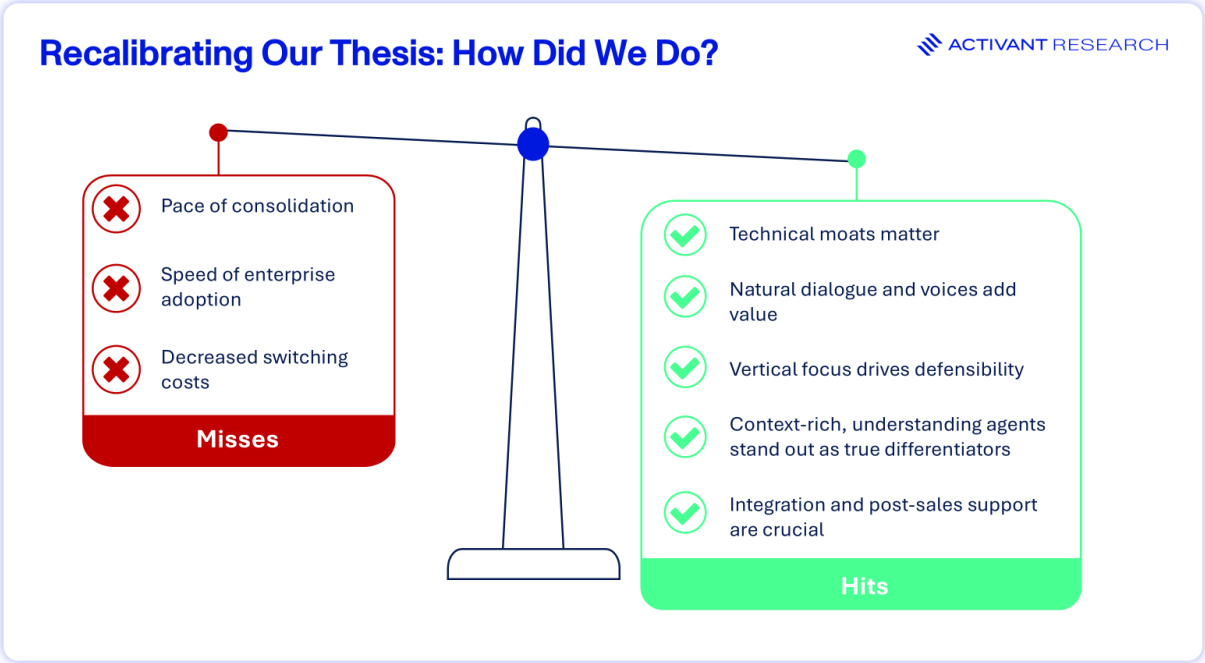
What began as a customer service tool is developing into a cross-industry infrastructure shift. Industries including healthcare, finance, retail, gaming, translation, education, and automotive are integrating voice capabilities. Even Wimbledon has adopted electronic line-calling technology that uses voice systems to call “out” or “fault” on the courts.[2] However, a critical tension has surfaced: while consumer-facing AI applications proliferate, enterprise adoption for core processes remains measured and cautious. The reason for this gap will define the next phase as the competitive focus migrates from “sounds human” to “thinks human.”

Differentiation no longer comes from voice quality alone; it comes from memory, context-awareness, and verifiable governance. In this report, we revisit our initial assumptions and explore where true, defensible value is accruing.

A note on terminology: While often called "chatbots," that term fails to capture the complexity of systems integrated into enterprise workflows. We will use "voice agents" to describe AI that handles complex tasks, memory, and actions.

Revisiting Our 2024 Thesis: Hype vs. Enterprise Reality

Our initial analysis predicted consolidation and easier technological adoption. The market, however, proved more nuanced. Enterprise buyers, wary of brand risk and integration complexity, have prioritized control over speed, reshaping go-to-market strategies and technology moats.



Thesis Revisited: Adoption Pacing and the Wedge Strategy

Initial Expectation: Full automation would rapidly displace human agents in customer-facing roles.

Market Reality: Augmentation has taken precedence over full automation. The vision of fully autonomous agents ran into the high stakes of brand reputation. McDonald's, for example, faced a public relations headache when viral videos exposed incorrect orders from its automated system.[3]

Enterprises have learned to separate ambition from deployment reality. A 2024 Gartner survey revealed significant consumer apprehension, with many fearing AI would make it harder to reach a human.[4] Consequently, leaders are implementing Voice AI through a **wedge strategy**: targeting high-volume, lower-risk use cases first. This typically includes after-hours support, overflow call routing, and internal agent assistance (e.g. summaries and recommendations). This strategy works because it delivers immediate ROI without jeopardizing core operations. For instance, startups like Toma target automotive dealerships where 56% of leads arrive after hours.[5] By capturing and qualifying these leads instead of letting them go to voicemail, the voice agent demonstrates clear value, building internal trust for future expansion. Since most enterprise adoption paths are multi-modal, the wedge strategy provides an entry point that expands into omni-channel automation and ultimately positions the voice agent as the system of record.

Yet despite these gains, board-level concerns persist. In July 2025, the World Economic Forum reported that as AI becomes more agentic in enterprise operations, issues of trust, explainability, data control, and regulatory compliance continue to delay adoption.[6]

Thesis Revisited: Market Consolidation and Switching Costs

Initial Expectation: The market would consolidate rapidly, with category leaders emerging and as many as half the players on our initial market map disappearing within a year.

Market Reality: Consolidation has been slower than anticipated. According to CB Insights, most Voice AI companies are still iterating on product rather than achieving true scale.[7] The field remains fragmented, buoyed by broad definitions of the category and steady VC funding (Voice AI companies raised $371M in equity funding by July, matching the total for all of 2024 within just the first seven months of the year).[8]

The primary barrier to consolidation is not a lack of M&A interest, but high switching costs. Our initial thesis assumed stack modularity would simplify migration. However, **enterprise adoption continues to move at the speed of IT, not the speed of innovation.** Activant's expert network notes that while an SMB might switch vendors in weeks, large enterprise migrations stretch from eight to twelve months. In regulated sectors like healthcare, compliance verification can add another two to four months.

These barriers are less contractual than technical and operational. Systems are deeply coupled with provider-specific IDs and workflow logic. Re-implementation requires re-training staff and sacrificing operational knowledge built around a specific vendor's quirks. As one product head noted in an expert call, "The longer you stay with one vendor, the more cost-effective it becomes to continue."

The bottom line: enterprises are moving cautiously, redefining defensibility and pointing us toward the next phase of value creation in Voice AI.

The New Moats: Architecture, Memory, and Pricing

The market's cautious pace and fragmentation point to a new set of requirements. Value has adapted from off-the-shelf model performance to full-stack control and contextual intelligence. In enterprise contexts, voice quality is the customer experience, and small improvements translate into measurable business value.

From Commodity to Technical Moat: The Return of Architecture

In 2024, many Voice AI vendors appeared interchangeable, leveraging similar underlying foundation models from hyperscalers. From the outside looking in, our internal feature analysis revealed limited visible differentiation, making it difficult to identify clear technical advantages.

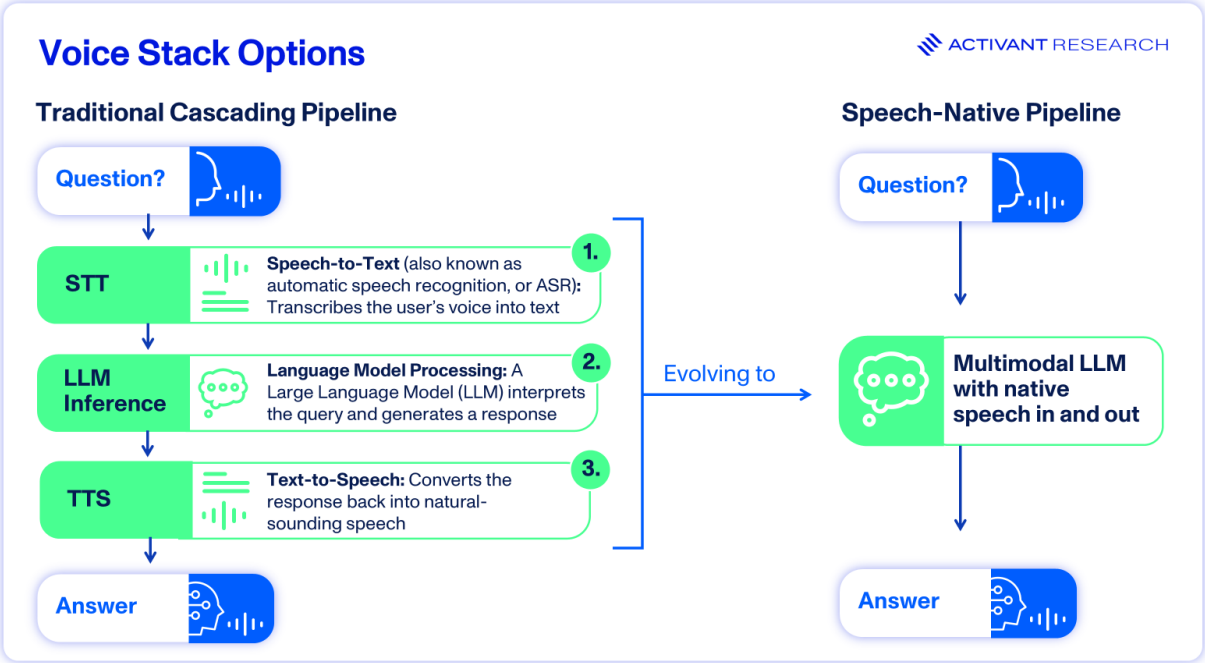
Today, however, the market is rewarding companies that own more of their stack to control performance and cost.

A crucial distinction has since emerged. For basic applications, "good enough" voice realism is quickly becoming commoditized. But for enterprise-grade use-cases, voice quality remains a moat. The difference between a neutral but robotic voice and one that conveys emotional nuance can drive a 20% increase in call success.[9] At scale, that improvement reduces call transfers, improves Customer Satisfaction (CSAT) and Net Promoter Scores (NPS), and unlocks millions in cost savings.[10] A premium tier of proprietary models offers demonstrably superior emotional nuance and prosody. This elite quality tier is far from a commodity and serves as a differentiator for market leaders capable of developing their own core speech synthesis technology.

Meta's acquisition of PlayAI illustrates this shift. It wasn't just buying a library of realistic voices; it was securing proprietary technology that offered full control over the speech generation process and guaranteed sub-100ms latency.[11] Voice synthesis is evolving from a feature add-on into core infrastructure, becoming as valuable as foundation models themselves.[12] PlayAI provides Meta with scalable text-to-speech, while WaveForms brings advances in emotional intelligence and its pursuit of the "Speech Turing Test," which measures whether listeners can tell human and synthetic speech apart.[13] NiCE's acquisition of Cognigy adds enterprise agents that think, adapt, and act independently to deliver human-like service to their Contact Center as a Service (CCaaS) platform.[14] These deals signal a race toward deeper technical capability, where full-stack ownership delivers reliable and scalable enterprise products, differentiating them from vendors who are merely "wrappers" on third-party APIs.

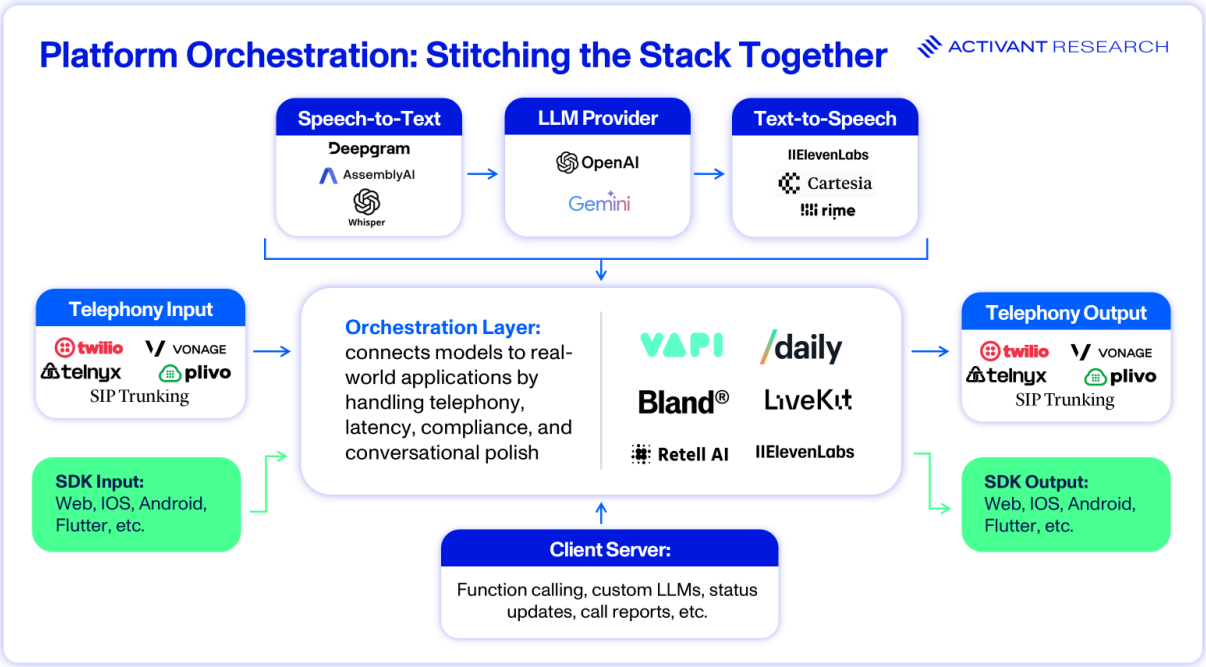
The Technical Fork in the Road:

This leads to the fundamental architectural choice facing developers today: cascading pipelines versus speech-native models.



Traditionally, Voice AI has followed a cascading three-step process, with each stage often powered by different vendors. Speech-to-text (STT) engines like Deepgram's Nova 3, AssemblyAI's Universal Streaming or OpenAI's Whisper first transcribe audio into text. Large language models such as OpenAI or Gemini then generate a response, which is finally converted back into voice through text-to-speech (TTS) models like ElevenLabs' Flash v2, Cartesia's Sonic, or Rime's Arcana . The voice stack is typically stitched together by orchestration platforms such as Vapi, Retell AI, or Bland, combining STT, LLM, and TTS components with telephony integration, observability, compliance, and security, while also adding conversational polish through features like[15]:

- **Endpointing:** Detects precisely when a speaker has finished talking
- **Interruption handling**: Lets users cut in without breaking the conversation flow
- **Noise and voice filtering:** Blocks background speech from TVs, echoes, or other voices
- **Backchanneling:** Inserts natural affirmations like "yeah" or "got it" at the right moments
- **Emotion detection:** Reads user tone and passes emotional context to the LLM
- **Filler injection:** Adds natural pauses and fillers ("um," "like," "so") to make responses more human



Source: Vapi, How Vapi Works, 2025

Cascading pipelines remain enterprise standard because each step creates a control surface for validation and logging. However, they come with three major obstacles:

1. **Latency:** The cumulative delay often exceeds 500 milliseconds, breaking the flow of natural conversation (where pauses over 200ms feel unnatural) and frustrating users.[16]
2. **Information Loss:** Converting speech to text strips away crucial paralinguistic signals such as tone, hesitation, and emphasis. These signals carry up to 38% of emotional meaning according to foundational communication research.[17] Losing this context prevents the agent from detecting user frustration or urgency.
3. **Cost:** Stacking models from different vendors multiplies infrastructure expenses.

In Voice AI, every second counts. A single extra second of delay can cut user satisfaction by 16% and drive abandonment up 23%, with two-thirds of users bailing for a human agent if pauses drag on.[18] Add to that the 91% of consumers who report frustrating digital experiences and the 70% ready to switch brands after a single bad AI interaction and the stakes become clear.[19]

OpenAI's launch of Advanced Voice mode in September 2024 followed by Realtime API in October marked a potential turning point: a single-step speech-to-speech model that preserves nuance, reduces latency, and produces more natural, emotionally aware responses, building conversational polish directly into the stack. Since then, the ecosystem has expanded with Moshi, an open source alternative, and Microsoft's Azure Voice Live API, released in July as the enterprise platform built on top of GPT-realtime, with added compliance and other enterprise features. By August 2025, GPT-realtime achieved 82.8% accuracy on the Big Bench Audio reasoning benchmark and reached enterprise deployment at Zillow, T-Mobile, and Lemonade, powering

applications from voice-driven home search to customer support and insurance claims.[20] These speech-native agents shift interaction from turn-based exchanges to fluid, real-time dialogue, creating a technical moat for providers and enabling use cases where speed, accuracy, and empathy are non-negotiable.

Cascading pipelines remain the enterprise default because they are proven and governable. But cracks are showing as latency, redundancy, and compute inefficiency degrade customer experience and drive up costs. Real-time speech offers immediacy but leaves little room for compliance and safety guardrails, slowing adoption. **These models will not replace cascading pipelines but will coexist, each serving distinct use cases.** In heavily regulated industries, compliance and auditability demand speech-to-text for observability, recall, and transcription (such as healthcare records or legal documentation), while premium text-to-speech will remain indispensable in creative sectors like gaming, narration, and advertising, where fidelity and emotional nuance drive value. The inflection point for mass enterprise adoption will come when speech-native models integrate robust validation and policy controls to combine efficiency with enterprise-grade reliability.

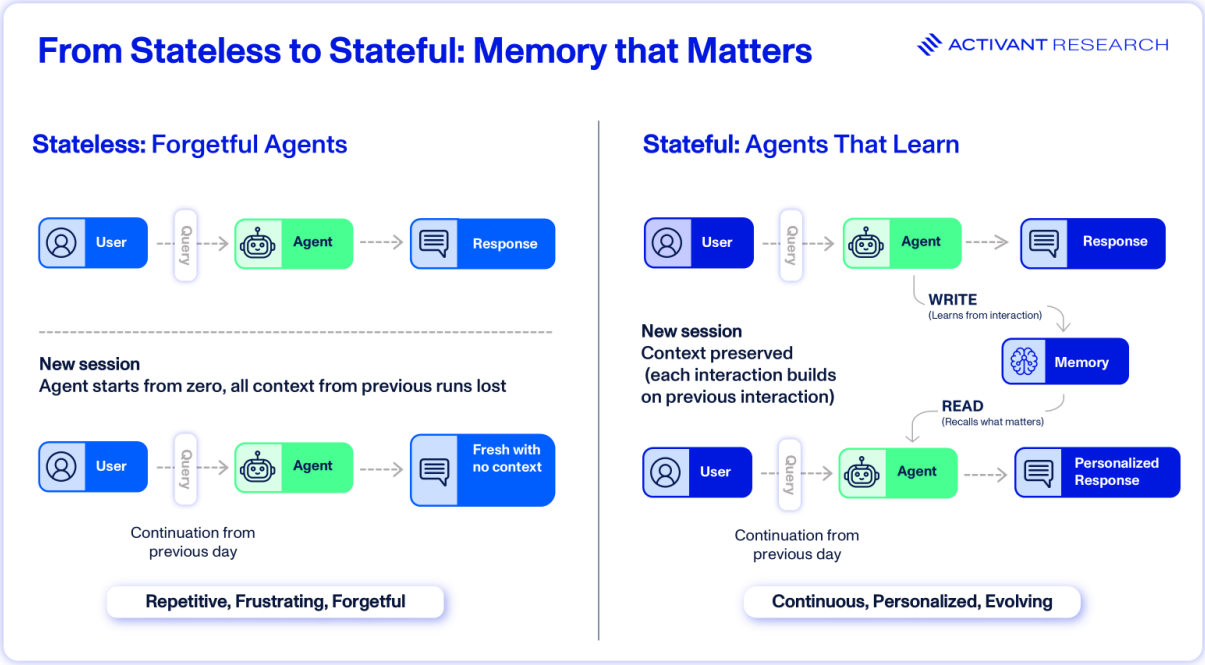
The Emerging Differentiator: Stateful Agents

Most AI systems today behave as if they have amnesia: they're smart but lack continuity. Without memory, every interaction starts fresh, forcing users to repeat themselves. Advances in model reasoning, falling costs for retrieval augmented generation (RAG), and new standards for connecting to external applications like Modal Context Protocol (MCP) now make it possible for voice agents to carry context across sessions.[21]

Rather than treating every interaction as a one-off exchange, memory creates a persistent internal state that evolves with use. This allows agents to capture user preferences, past decisions, and long-term context, sustaining continuity across weeks or months. Unlike RAG, which grounds responses in external data at inference but remains stateless, memory provides persistent awareness of identity and history. This transforms AI from a transactional tool into a relationship-based system, positioning it as the backbone for next-generation customer engagement, productivity, and enterprise applications.

To achieve true intelligence, next-generation agents must integrate multiple types of memory:

- **Working memory (short-term):** keeps track of immediate context
- **Factual Memory (long-term):** stores user preferences and domain knowledge
- **Episodic Memory (long-term):** recalls specific past events or outcomes
- **Semantic Memory (long-term):** builds generalized knowledge over time
- **Contextual handoff:** transfers the relevant information to a human upon escalation



Source: mem0, Memory in Agents: What, Why and How, 2025

Enterprises using Google Cloud can leverage Memory Bank in Vertex AI Agent Engine to store session history and deliver personalization, continuity, richer context, and smoother user experiences.[22] For those hesitant about vendor lock-in or who need cross-cloud and on-prem deployment, emerging open-source platforms offer greater flexibility and control. Mem0 is the simplest entry point, a lightweight memory layer for persistent storage and embedding-based retrieval that enables personalization without complex schema design. Zep builds on this with temporal knowledge graphs, fact invalidation, and engineered context blocks to capture evolving user and system states, reducing hallucinations and keeping agents current. At the most comprehensive level, Letta, developed from the MemGPT framework, is tool-agnostic and enables stateful agents with full memory hierarchies, self-editing blocks, and dynamic context management, allowing agents to not only recall but also adapt and learn over time.

Startups are also embedding memory at the heart of their products. Lemni allows customer-facing agents to draw from a single source of truth, ensuring that every interaction begins where the last one ended. Abridge is evolving from simple transcription to a living medical record, integrating prior notes and preferences, and informing healthcare professionals with better context and history.[23] And London-based Convergence.ai was acquired by Salesforce just 12 months after launch, bringing its AI agents, designed to develop skills through long-term memory and continuous learning, into the Agentforce platform.[24,25]

Parloa is putting its recent $350M raise into building a "multi-model, contextual experience" where AI agents maintain customer context across channels and adapt tone to emotional cues in real time. In a payment reminder deployment with Waterfield Tech for a global e-commerce retailer, 66% of customers promised payment after speaking with Parloa's voice agent versus 51% with a

human, and 62% followed through compared to 57%, suggesting that an agent fine-tuned for empathy can outperform a human working from a script.[26]

With 96% of consumers more likely to purchase when messages are personalized, and 66% reporting stronger affinity for brands that remember their preferences, memory is no longer a nice-to-have, it is a growth driver.[27] Persistent memory ensures that agents remember not just what was said but how it was said. When tone, hesitation, or frustration from past interactions informs the next one, the result is continuity that feels recognizably human. Reducing repetition lowers abandonment, personalization boosts conversion, and continuity across sessions turns an agent from a cost-saving tool into a relationship driver. Persistent memory also poses regulatory challenges like the EU's General Data Protection Regulation's (GDPR) "right to be forgotten," where customers can request deletion of their history. Agents that can remember what matters but also forget on demand will be crucial. **The leap from "sounds human" to "thinks human" through persistent, policy-aware memory is where we think lasting business value will be created.**

Realigning Incentives in Voice AI Pricing

How a Simple Call Adds Up

ACTIVANT RESEARCH

Component	Rate	Average Cost per Call (6 min)[1]
Base Hosting (Vapi)	$0.05/min	$0.30
STT (Deepgram)	$0.01/min	$0.06
LLM (Gemini 2.0 Flash)	~$0.09/1M tokens	$0.00011*
TTS (ElevenLabs)	$0.036/min	$0.22
Telephony (Vonage)	$0.00814/min	$0.05
Total Cost per Call using a Voice Agent		**$0.62**
Average Cost per Customer Service Call observed across industries[2]		**$4.15**

*Calculations based on the assumption of 150 words per minute -200 tokens per min
[1]Average Handle Time (AHT): 6 minutes; ; Sprinklr, Call Center Statistics, 2025
[2]Average Cost per Customer Service Call: $2.70 - $5.60; Sprinklr, Call Center Statistics, 2025

As Activant has maintained, usage-based billing (UBB) is becoming a critical component of modern pricing, aligning cost with consumption and supporting product-led growth.[28] But UBB is not one-size-fits-all. In Voice AI, cost-per-minute models misalign incentives by rewarding duration over efficiency – a long but unproductive conversation drives cost without creating value.

Additionally, as leading vendors like OpenAI cut rates – with GPT-realtime now priced 20% below GPT-4o-realtime-preview – pricing pressure starts to create a race to the bottom.[29] Early on,

usage-based pricing lowers barriers to adoption, but as value is proven, pricing should shift toward outcomes or results.

For agents that rival or surpass humans, value is not in minutes but in outcomes like appointments booked, claims resolved, and revenue generated. Much like SDRs are compensated on sales, outcome-based pricing for voice agents will capture more margin, align incentives, and build stickier enterprise relationships. UBB still has a role to play in pilots and early adoption, but at scale, outcome-based models are where value will be realized.

The Road Ahead

Cost efficiency is irrelevant if latency frustrates customers and robotic tone erodes trust, which is why the next leap is likely to come from speech-native models built for speed, fidelity, and natural conversation at scale. For enterprises, voice quality now shapes brand identity: what works for a global bank doesn't work for a gaming company, and a health insurer can't sound like a Gen Z influencer. Paralinguistic nuance and contextual adaptation are becoming critical.

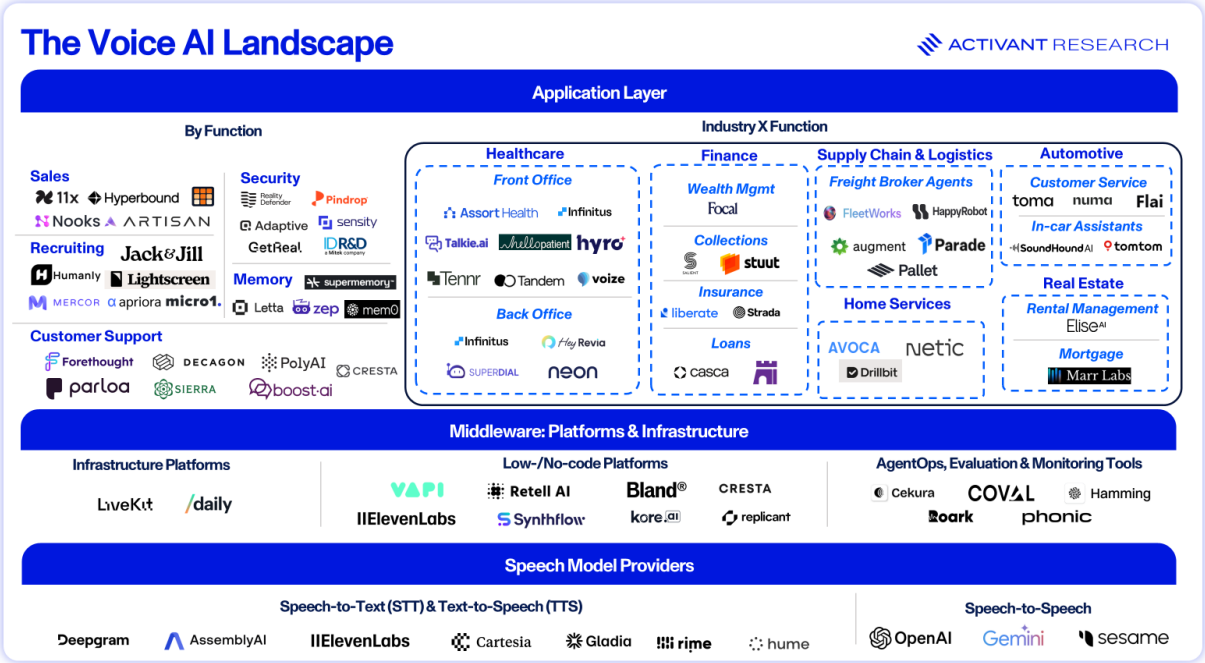
This is not a zero-sum game. Enterprises will still rely on cascaded architectures where compliance, auditability, and deterministic control are non-negotiable, while deploying speech-native stacks for real-time, emotionally aware interactions. The likely outcome is hybrid, with a single exchange drawing on multiple models optimized for distinct functions. The winners will be those who orchestrate both seamlessly, grounding voice systems in vertical workflows to deliver trust, efficiency, and differentiated customer experiences.

> Every enterprise will have its own distinct agent and voice model - tuned to its customers and workflows. Value shifts from minutes to outcomes, and the edge comes from focusing on full vertical workflows and the orchestration powering them - blending control with emotionally intelligent speech in real time."
>
> **Mati Staniszewski**
> Co-Founder & CEO, ElevenLabs

Verticalization: Where Value Accrues

As identified in our previous research on vertical software, we observed that value emerges when solutions go deep into industry workflows rather than spreading horizontally.[30] Vertical SaaS is outpacing traditional enterprise software by roughly three percentage points of CAGR, while also delivering 10%-20% higher year-over-year revenue growth and nearly double the sales efficiency.[31]

The same dynamic is now emerging in Voice AI: agents that integrate with systems of record, leverage domain-specific data, and align with regulatory requirements deliver greater stickiness, higher switching costs, and more durable moats than general-purpose platforms can achieve.



Verticalization is no longer optional – it's being driven by structural forces that enterprises can't ignore.

Driver 1: Growing Customer Expectations

Voice remains the dominant channel for customer engagement, driving up to 70% of interactions, particularly in urgent or complex scenarios.[32] Expectations are rising fast: 71% of customers now demand personalized experiences, and a single bad experience would prompt 78% of U.S. consumers to consider switching brands.[33,34] As Forbes notes, customers expect not only faster responses, but also hyper-personalized engagement.[35]

Yet expectations differ both by industry and by task. In finance, for example, most Americans are comfortable with AI for fraud detection (70%) or credit scoring (64%), but only 13%-15% trust it with high-stakes decisions like retirement planning or asset allocation.[36,37] The wedge strategy, built on high-volume, low-risk use cases, fits well in industries like automotive, retail, and home services. In regulated sectors such as finance and healthcare, go-to-market and adoption take a different route. Trust comes first, forcing leading voice AI players to emphasize safety and credibility before scale. Verticalization then becomes critical, tuning models and workflows for industry-specific accuracy and aligning with varying levels of consumer comfort. But delivering speed, empathy and trust together remains a longer-term challenge.

Driver 2: Regulation Forces Fragmentation

The regulatory landscape has grown more demanding, turning compliance from a checkbox item into a core architectural constraint. Key developments include:

- **TCPA and FCC Rulings:** The FCC affirmed that AI-generated voices fall under the Telephone Consumer Protection Act (TCPA), mandating explicit consent.[38] Proposed legislation like the "Keep Call Centers in America Act" further demands disclosure and an immediate right to escalate to a U.S.-based human agent.[39]
- **Global Frameworks:** The EU AI Act imposes high-risk classifications on emotion recognition systems and mandates data transparency, carrying penalties up to 7% of global turnover. [40, 41,42]

Voice AI platforms are being forced to re-architect around auditability, traceability, and identity verification, giving rise to a new ecosystem of compliance-first tooling. A wave of early-stage startups (including Phonic, Cekura, Coval, and Hamming) are building infrastructure to observe and evaluate voice agents in production, while Adaptive Security provides AI-driven phishing simulations and compliance training. Their focus on real-time compliance monitoring, continuous QA, and conversational reliability ensures that speed and naturalism is matched with enterprise-grade governance.

For enterprises in highly regulated industries, compliance is not optional—it is a prerequisite for adoption. Compliance requirements are not transferable across industries: a HIPAA-compliant agent in healthcare operates under different rules than a PCI-compliant agent in finance. Memory-rich systems introduce new liabilities such as data privacy and residency requirements. This regulatory fragmentation inherently breaks apart horizontal platforms, shaping vertical stacks and making it harder for generalist providers to scale. Compliance-first startups are emerging as essential enablers of adoption, positioning themselves as foundational to the Voice AI stack. Those able to solve for secure storage, policy enforcement, and explainability are best positioned to define the next wave of defensible platforms.

Driver 3: Security Becomes a Functional Vertical

As voice becomes an identity layer, security threats have become more sophisticated. Synthetic voice fraud attacks have surged, with Pindrop reporting significant increases in insurance (+475%), banking (+149%) and retail (+107%).[43] High-profile incidents, like the $25 million deepfake wire fraud in Hong Kong, illustrate that legacy defenses are insufficient.[44]

As our identity security research highlights, enterprises have spent decades building Identity and Access Management (IAM) and Multi-Factor authentication (MFA), yet breaches persist because attackers exploit the last mile of security – the point where identities are authenticated and authorized.[45] In fact, two-thirds of all data breaches involve a “human element”.[46] Voice inherits these same vulnerabilities and once it becomes a credential, synthetic speech can bypass traditional defenses just as stolen passwords or hijacked cookies do.

AI’s expanding memory and recall will make voice agents more proactive and context-aware but will also raise the bar for privacy and security. To address this, security is verticalizing by function,

requiring real-time inference, multimodal analysis (audio, metadata, behavior), and partnerships between model vendors and security firms. Companies like Reality Defender, Pindrop and ID R&D are building this new layer of infrastructure dedicated to voice authentication.

Industry pain points are rarely standalone – they are defined as much by compliance and security as by customer expectations, workflow complexity and integration. We believe that the winners will be those that verticalize, capturing defensible value where industry and function meet.

Going Deep: Voice as the Operating Layer

This tension between efficiency and empathy plays out differently across industries. Each sector has pain points that generalist systems struggle to solve, from compliance-heavy workflows to revenue leakage from missed calls. Value accrues where voice agents embed deeply into the operating system of an industry, addressing needs that are too complex or specialized for horizontal platforms. Consider the following interesting players who have addressed industry-specific requirements and challenges:

Finance and insurance are defined by high-stakes interactions that demand strict identity verification and compliance. Salient provides an AI loan servicing platform, Pindrop manages fraud alerts, and Casca accelerates loan qualification, all integrating with systems of record under PCI-DSS and KYC standards. In insurance, Liberate automates policyholder interactions like claims intake, servicing, and multichannel communication, while Strada focuses on voice-first sales, renewals, and distributor-side workflows to help agencies and brokers capture more revenue.

Nearly a quarter of the **healthcare** industry's $5 trillion in annual spending is lost to administrative overhead.[47] On the front end of care delivery, Assort Health streamlines intake, scheduling, and administrative coordination while adapting to the nuances of specialty clinics, while Tennr focuses specifically on solving pre-visit patient processing. Other voice agents are automating revenue cycle management by handling payer calls and prior authorizations (Infinitus Health, SuperDial) and navigating the complex specialty drug process with AI-powered outreach and patient support (Neon Health). Meanwhile, engagement platforms like Hello Patient and Hyro manage real patient conversations around the clock, booking appointments, answering questions, and re-engaging patients. Collectively, these solutions automate routine but costly tasks, allowing practices to redirect staff capacity toward direct patient care, all while adhering strictly to HIPAA.

In regulated industries like healthcare, 'trust' isn't a marketing word. A single utterance can change care or move money, so voice AI must be trusted to do the right thing, every time. This is more than just compliance; it's predictable, auditable behavior under real-world constraints. That's why we built Infinitus the way we did: trust-by-design with agents constrained by clear rules and anchored to healthcare knowledge, so automation can be both safe and scalable."

Ankit Jain
Co-Founder & CEO, Infinitus Systems

Value isn't limited to regulated industries – supply chain, home services, real estate, automotive, hospitality, and media all face similar pain points and are ripe for AI-driven automation. In **home services,** a $211.7 billion market projected to quadruple by 2032, Avoca handles lead intake, estimates and invoicing for trades like plumbing and electrical work, while Netic.ai proactively drives sales through maintenance and upgrade nudges.[48] In **supply chain and logistics,** where fragmented communication still dominates, HappyRobot books freight and resolves load issues, FleetWorks vets carriers, and Augie from Augment automates the entire order-to-cash cycle.

In **real estate**, EliseAI streamlines rent collection, guest communication, and leasing, while Marr Labs powers mortgage agents that capture and qualify leads. In **automotive**, where missed calls cost dealers nearly $50,000 a year, Toma and Numa convert inbound and after-hours inquiries into sales. In **hospitality,** Slang.ai serves as a 24/7 phone concierge, booking reservations and integrating with platforms like OpenTable, while quick-service chains like Domino's already rely on Rime for order taking. ElevenLabs has even partnered with book publisher, HarperCollins to turn backlist titles into audiobooks, and DJ Sam Feldt uses its voice cloning to create broadcast-quality voiceovers for social media, radio, and festival announcements.

The ability to leverage voice while pulling and using data across multiple data silo's is a massive unlock for businesses. The benefits of the space is the rapid pace at adoption, allowing us to switch in and switch out providers to deliver the most process automation for our clients. For example, we can use one model or provider for a IVR tree while handling inbound calls differently."

Tarek Alaruri
Co-Founder & CEO, Stuut

In the next phase of enterprise Voice AI, durable value should be created when technology solves industry-specific pain points, embeds into daily workflows, and becomes the operating layer that drives revenue and efficiency.

Investment View and Forward Outlook

The prevailing narrative, echoed by Sam Altman and others, is that AI will soon replace customer service entirely.[49] We once shared that belief, but it overlooks the nuance of human needs. Customer service remains one of the most fraught consumer experiences. Frustrations are well documented: long queues, dropped calls, irrelevant automated replies, and the need to repeat the same issue to multiple agents. Above all, customers resent not being able to reach a real person.[50]

At its core, enterprise communication is about acknowledgment—being heard, understood, and respected. Whether in customer service, sales engagement or internal collaboration, voice interactions shape how organizations build trust, resolve issues and strengthen relationships. The challenge for enterprises is achieving efficiency without sacrificing loyalty. As LLMs advance, speed and accuracy will become table stakes. What remains is friendliness, cultural nuance, empathy, and the qualities that make humans human. When combined with multilingual capabilities, voice agents can help customers meet in their preferred language, making interactions feel more inclusive, personal, and in some cases, democratizing access to services that were previously inaccessible.[51]

What's missing today are voice agents capable of managing nuance, maintaining continuity across interactions, and escalating seamlessly to humans when necessary. Lasting value will come from agents that demonstrate empathy, cultural awareness, and tone—qualities that make interactions feel human – across every touchpoint. Whether these agents should go as far as building genuine rapport remains an open question.

Our Investment Thesis for Voice AI 2.0

Voice AI is entering its second act. Beyond proving speech realism, the next phase is about defensibility and enterprise scale. Four drivers shape where we foresee lasting value being created:

1. **Realism is Table Stakes; Intelligence is the Moat:** The competitive edge has shifted from voice quality to memory, contextual awareness, and governance. A premium tier of high-fidelity, proprietary models provides a distinct advantage over commoditized offerings.
2. **Architecture Controls Destiny:** Durable value accrues to companies that own the key components of the technology stack. Full-stack control allows for superior management of latency (with benchmarks now pushing sub-100ms), cost, and feature development, creating a defensible barrier.
3. **Verticalization Wins:** General-purpose platforms will struggle against specialized agents designed for specific industry workflows and compliance requirements. Defensibility lies at the intersection of industry depth and functional specialization (e.g., healthcare + RCM automation).
4. **Price Outcomes, Not Minutes:** The economic model must evolve from usage-based pricing driven by consumption to value-based pricing driven by results delivered.

All these elements matter, but the most compelling opportunities lie with founders who truly understand the person on the other end of the call, building voice agents that reflect their wants and needs. They see voice not just as a channel replacement but as an operational unlock, creating systems that embed into enterprise workflows while balancing efficiency with the governance and intelligence required to earn lasting trust.

At Activant, we believe Voice AI will permeate daily life, from ordering takeout and scheduling appointments to job interviews and even companionship, but success hinges on clear market strategy. Vendors that remain point solutions risk being commoditized. Durable players will either own the stack (controlling performance and cost end-to-end), move up the stack (embedding memory, orchestration, compliance), or verticalize (integrating directly into industry workflows).

And we're already seeing this play out. ElevenLabs has evolved from a pure TTS model provider into a creative and enterprise voice platform, unifying speech-to-text, text-to-speech, voice cloning, and orchestration within a fully owned stack. Its Agents Platform integrates with tools like Salesforce and Stripe, with custom agents and workflow templates for telecom, retail, financial services, and technology, and upcoming multi-agent workflows for more complex enterprise use cases. Deepgram's Voice Agent API follows a similar path, giving enterprises the choice to use its native models (Nova-3 and Aura-2) or bring their own while maintaining orchestration and control. On the vertical front, Sierra has evolved from a generalist customer service AI into an industry-specific platform, launching tailored solutions for financial services, healthcare, telecom, retail, and more. Assort Health, meanwhile, is deeply embedding itself within healthcare, expanding from voice AI for patient scheduling into a full omnichannel patient-experience platform that now spans billing, refills, lab results, and multi-specialty workflows illustrating a clear land-and-expand strategy beyond its initial scheduling wedge.

The strongest vertical opportunities lie in industries where dialogue drives value and high-frequency interactions often depend on costly or limited human talent. True defensibility will come from embedding directly into regulated, compliance-heavy workflows, where integration creates both stickiness and barriers to entry. The next wave of Voice AI will separate leaders from followers with the winners pairing domain expertise with relentless focus on product–market fit, delivering solutions that are not just faster but fundamentally better—empowering humans where it matters most.

We would like to hear from you if our work resonates or if you have a different perspective. If you're building in this space, we'd love to connect!

---

[1] Businesswire, Global Call Centers Strategic Business Report 2023: Market to Reach $494.7 Billion by 2030 from $314.5 Billion in 2022 - U.S. is Estimated at $110 Billion, While China is Forecast to Grow at 6.6% CAGR - ResearchAndMarkets.com, 2023

[2] Wimbledon, The precision operation: Introducing Electronic Line Calling, 2025

[3] CB Insights, Voice AI's sweet spot: ordering fries with that, 2025

[4] Gartner, Gartner Survey Finds 64% of Customers Would Prefer That Companies Didn't Use AI For Customer Service, 2024

[5] Juice Digital, Missed Call Text Back: Maximising Sales [Updated], 2023

[6] World Economic Forum, Enterprise AI is at a tipping Point, here's what comes next, 2025

[7] CB Insights, Voice AI's sweet spot: ordering fries with that, 2025

[8] CB Insights, Voice AI is having a moment: Here are the startups that could get acquired next, 2025

[9] Activant Expert Network

[10] Activant Expert Network

[11] CB Insights, Voice AI is having a moment: Here are the startups that could get acquired next, 2025

[12] CB Insights, Voice AI is having a moment: Here are the startups that could get acquired next, 2025

[13] TechCrunch, Meta acquires AI audio startup WaveForms, 2025

[14] NiCE, Acquisition combines NiCE's purpose-built CX AI platform, CXone Mpower, with the enterprise leader in conversational and agentic AI, enabling organizations to accelerate AI adoption in customer experience across the front and back office, 2025

[15] Vapi, Orchestration Models, 2025

[16] Gnani.ai, Latency is the Silent Killer of Voice AI—Here's How We Solved It, 2025

[17] Forbes, Strong Nonverbal Skills Matter Now More Than Ever In This "New Normal", 2020

[18] Gnani.ai, Latency is the Silent Killer of Voice AI—Here's How We Solved It, 2025

[19] Techradar, The trust recession: why customers don't trust AI (and how to fix it), 2025

[20] OpenAI, Introducing gpt-realtime and Realtime API updates for production voice agents, 2025

[21] The Wall Street Journal, The AI Experience Is Going From '50 First Dates' to 'Cheers', 2025

[22] Google Cloud, Announcing Vertex AI Agent Engine Memory Bank available for everyone in preview, 2025

[23] The Wall Street Journal, The AI Experience Is Going From '50 First Dates' to 'Cheers', 2025

[24] The Next Web, New UK startup raises $12M for personal AI agents with long-term memory, 2024

[25] Salesforce, Salesforce Signs Definitive Agreement to Acquire Convergence.ai, 2025

[26] Parloa, Why multilingual agentic AI is key to global customer experience, 2026

[27] Businesswire, New Global Study Reveals Consumers Demand More Personalization in Marketing; 81% Ignore Irrelevant Messages, While Personalized Experiences Drive Loyalty and Sales, 2025

[28] Activant Capital, Usage-Based Billing, 2024

[29] OpenAI, Introducing gpt-realtime and Realtime API updates for production voice agents, 2025

[30] Activant Capital, Vertical Software Is Having A Moment, 2025

[31] Activant Capital, Vertical Software Is Having A Moment, 2025

[32] Techradar, Why voice still rules in the AI-powered contact center, 2025

[33] McKinsey & Company, The value of getting personalization right—or wrong—is multiplying, 2021

[34] Verint, New Study Reveals 2025 as the Year AI-Powered CX Delivers Real-World Value, 2025

[35] Forbes, From Emotion To Empathy: Bringing Human Experience To Voice AI, 2025

[36] TD Stories, TD Bank Survey Finds Americans are Ready to Embrace AI, but Have Yet to Unlock Its Full Potential, 2025

[37] Northwestern Mutual, Human Connection Over Machines: Americans Trust Advisors More Than AI for Financial Advice, Finds Northwestern Mutual's 2025 Planning & Progress Study, 2025

[38] Federal Communications Commission, FCC Confirms that TCPA Applies to AI Technologies that Generate Human Voices, 2024

[39] CBS News, New bill aims to protect American call center jobs and consumers from AI, 2025

[40] EU Artificial Intelligence Act, Chapter XII: Penalties, Article 99: Penalties, 2025

[41] EU Artificial Intelligence Act, Annex III: High-Risk AI Systems Referred to in Article 6(2), 2025

[42] EU Artificial Intelligence Act, High-level summary of the AI Act, 2024

[43] Pindrop, 2025 Voice Intelligence and Security Report, 2025

[44] CNN, Finance worker pays out $25 million after video call with deepfake 'chief financial officer', 2024

[45] Activant Capital, Taking Identity Security Forward, 2025

[46] Verizon, 2025 Data Breach Investigations Report, 2025

[47] StatRanker, The Impact of Administrative Costs on U.S. Healthcare Spending, 2025

[48] Verified Market Research, U.S. Home Services Market Size And Forecast, 2025

[49] Futurism, Sam Altman Says OpenAI Is Poised to Wipe Out Entire Categories of Human Jobs, 2025

[50] The Sun, ON HOLD Biggest customer service bugbears revealed – including rubbish hold music and not speaking to a real person, 2025

[51] Zendesk, CX Trends 2025: CX Trendsetters surge ahead of peers with human-centric AI as their advantage, 2025