SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Model-Driven Autonomous Voice Agents for Regulated Healthcare Operations

Bhargavi Kalicheti

Independent Researcher, USA

Abstract
Healthcare contact centers represent one of the most operationally complex interfaces between clinical, administrative, and payer systems. Traditional automation rule-based interactive voice response (IVR), deterministic routing, and scripted chatbots have failed to address the inherent variability, ambiguity, and multi-step reasoning required in healthcare interactions. Recent advances in agentic artificial intelligence, powered by large language models (LLMs), tool orchestration frameworks, and structured autonomy protocols, enable a new paradigm: agentic voice AI systems capable of executing healthcare workflows end-to-end with minimal human intervention. This paper presents a comprehensive architectural and systems-level analysis of model-driven autonomous voice agents applied to healthcare contact center operations. We introduce an agent-centric framework integrating perception, reasoning, planning, action, and reflection into voice-driven autonomous systems. The paper examines enabling technologies, including Model Context Protocol (MCP), multi-agent orchestration, dynamic tool invocation, confidence-driven autonomy thresholds, advanced memory architectures, and real-time governance. We analyze how agentic voice systems transform contact centers from reactive routing hubs into intelligent operational engines capable of executing scheduling, authorization, eligibility validation, and benefits navigation workflows at scale. The paper addresses safety, bounded autonomy, self-healing execution patterns, and organizational transformation considerations essential for production deployment in regulated healthcare environments.

Keywords: Agentic AI, Large Language Models, Tool Orchestration, Model Context Protocol, Multi-Agent Systems, Autonomous Workflow Execution, Healthcare Contact Centers, Voice AI, LLM Planning, HIPAA Compliance, React, Chain-Of-Thought Reasoning

1. Introduction

Healthcare contact centers operate at the intersection of clinical care, insurance administration, and regulatory compliance. Every interaction, whether from a patient, provider, pharmacist, or caregiver may require authentication, eligibility reasoning, clinical context awareness, and multi-system coordination. Human agents are traditionally tasked with orchestrating these workflows manually, leading to high operational costs, long handling times, and inconsistent outcomes [1], [2].

Voice-based automation has historically focused on surface-level interaction handling menu navigation, simple intent detection, and call routing. These approaches lack the capacity for goal-driven execution, adaptive reasoning, and context persistence across multi-step workflows. The dominant intent classification paradigm treats interactions as isolated events: models determine what the caller wants but have no capacity to reason about how to accomplish the goal, manage exceptions, or adapt when circumstances change [3], [4].

Agentic voice AI represents a structural evolution beyond conversational AI. Rather than responding to prompts, agentic systems become autonomous actors capable of planning, executing, and reflecting on tasks on behalf of the caller. This evolution is made possible by the convergence of LLMs capable of chain-of-thought reasoning [5], tool-augmented language models [6], emergent planning capabilities [7], and structured interoperability protocols such as Model Context Protocol (MCP) [8]. When these capabilities are integrated into voice-first architectures operating under healthcare regulatory constraints, the result is a new class of autonomous operational system.

This paper presents a systems-level analysis of model-driven autonomous voice agents for regulated healthcare operations, examining architectural design principles, enabling technologies, safety and governance requirements, and operational transformation implications. The contributions include: (1) a layered agentic voice architecture combining perception, reasoning, planning, execution, memory, and governance; (2) a

framework for MCP-based deterministic tool orchestration in healthcare voice systems; (3) confidence-driven autonomy threshold modeling for safe agentic execution; (4) multi-tier memory architecture for cross-call continuity; and (5) an autonomous error recovery framework for self-healing healthcare workflows.

2. From Conversational AI to Agentic Voice Systems: A Structural Evolution

The progression from intent-based conversational AI to agentic systems represents a qualitative rather than quantitative shift in system architecture. Intent classification models treat each interaction as an isolated event: they determine the caller's immediate intent but lack memory, planning depth, or execution awareness. In healthcare, even a seemingly simple request such as "I need to schedule a specialist visit" may require eligibility verification, identity confirmation, clinical constraints validation, plan network lookup, availability negotiation, and documentation updates, a multi-step goal structure that far exceeds the scope of any intent classification model [3].

An agentic voice AI system is defined by five core capabilities. Perception encompasses understanding spoken language, caller context, and environmental signals, including emotional state and background noise. Reasoning involves interpreting goals, constraints, and dependencies across the conversational context and retrieved domain knowledge. Planning decomposes goals into executable step sequences, adapting dynamically when prerequisites are unmet or conditions change. Action invokes tools, APIs, and enterprise workflows with typed parameters and defined success criteria. Reflection evaluates outcomes against expected results, updates beliefs, and selects corrective actions when execution deviates from plan [9], [10].

The ReAct framework [9] demonstrated that interleaving reasoning and action steps within LLM inference produces substantially more coherent and reliable task execution than either pure reasoning or pure action separately. Applied to voice AI, ReAct-style agents maintain an explicit thought-action-observation loop that produces interpretable, auditable decision traces a property of particular value in regulated healthcare environments where every autonomous decision must be defensible to compliance reviewers. Chain-of-thought prompting [5] further enables agents to decompose complex healthcare queries into verifiable reasoning chains before committing to tool invocations or workflow steps.

3. Layered Architecture for Agentic Voice AI in Healthcare

The agentic voice AI architecture is organized as a layered cognitive pipeline in which each layer operates independently while sharing structured state through defined interfaces. This modularity enables independent scaling, testing, and governance of each architectural component.

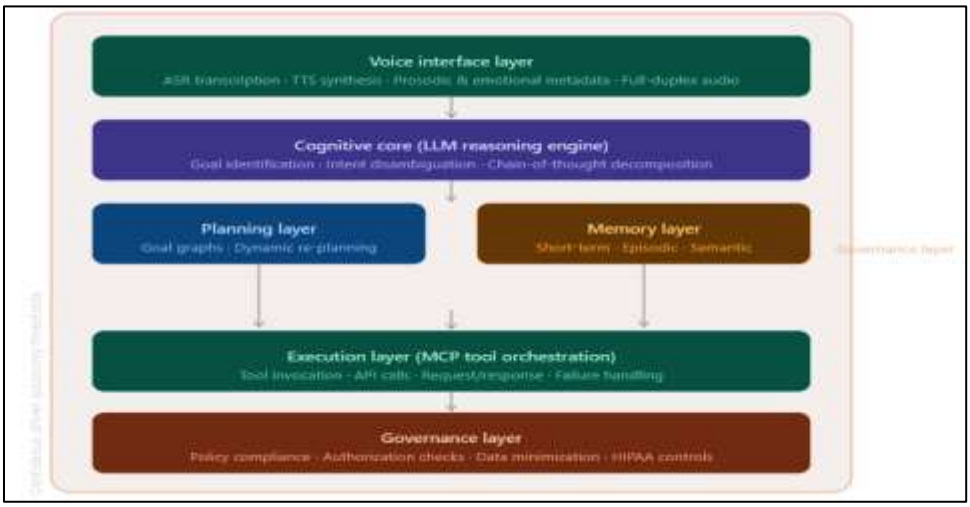

Figure 1 — Layered agentic voice AI architecture.

The voice interface layer manages ASR transcription, TTS synthesis, and full-duplex audio session handling. Raw audio is converted to text and enriched with prosodic and emotional metadata before being passed upward. The cognitive core houses the primary LLM reasoning engine, responsible for goal identification, intent disambiguation, and chain-of-thought decomposition. The planning layer receives decomposed goals

and produces executable step sequences, represented as goal graphs rather than fixed scripts, enabling dynamic re-planning when preconditions fail or caller constraints evolve [7].

The execution layer translates planned steps into concrete tool invocations and API calls, managing request construction, response parsing, and failure handling. The memory layer maintains a three-tier structure: short-term memory preserving in-session conversational state; episodic memory storing structured summaries of prior interactions for cross-call continuity; and long-term semantic memory encoding domain knowledge, procedural rules, and behavioral patterns. The governance layer enforces policy compliance at every execution step, embedding authorization checks, data minimization constraints, and approval checkpoints directly into the execution logic rather than applying them as post-hoc audits [2], [11].

4. Model Context Protocol: Governed Tool Orchestration for Healthcare Agentic Systems

The Model Context Protocol (MCP) formalizes the contract between LLM agents and external tools, data sources, and execution environments [8]. Rather than embedding tool invocation logic within prompts, an approach that is brittle, difficult to audit, and resistant to policy enforcement, MCP defines structured interfaces that specify tool signatures, input schemas, output formats, preconditions, and permission boundaries. This formalization transforms tool invocation from an emergent LLM behavior into a governed, deterministic operation.

Healthcare agentic systems have four critical requirements that MCP directly addresses: deterministic tool invocation with typed parameters and validated responses; auditable decision paths where every tool call is logged with its context, parameters, and outcome; explicit permission boundaries enforcing least-privilege access to sensitive healthcare data; and versioned context schemas enabling reproducible audits of past agent decisions. Together, these properties transform LLMs from probabilistic responders into governed decision engines that satisfy healthcare compliance requirements.

The agent execution cycle on MCP in healthcare voice AI can be described as follows: voice input is transcribed and enriched; the agent identifies the goal of the caller and breaks it down into a plan of steps. The agent queries the MCP tool registry to find the authorized tools to execute each step of the plan. tools are invoked with validated parameters and their responses are added to the agent state. the agent re-plans as needed based on the tools. The agent does not generate responses based on unconstrained generative inference on healthcare data at any point--all factual assertions are based on tool-returned structured data, and the major hallucination risk in healthcare AI deployments is eliminated [12].

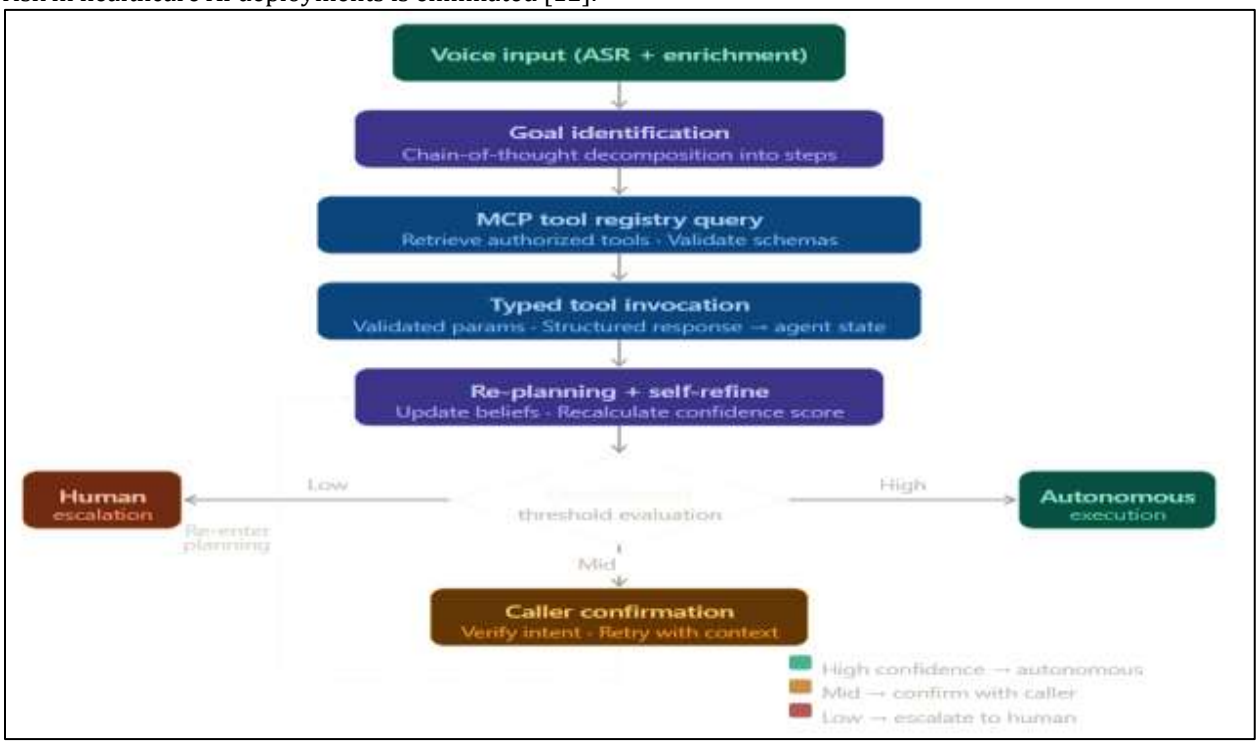

Figure 2 — MCP agent execution cycle and confidence-driven autonomy routing.

5. Multi-Agent Coordination of Complex Medical Workflow

Multifaceted healthcare call center communication habitually goes beyond the sequence of actions of a single agent. Temporary authorization processes, such as those needed to verify concurrent eligibility, clinical criteria evaluation, formulary search, and documentation, are better suited to specialized agents running in parallel under supervisory coordination. Multi-agent architectures delegate these tasks to specialized agents having clear functions, communication interfaces and context-sharing protocols [13], [14].

Agent	Primary Function	Key Data Sources	Escalation Trigger
Intake agent	Caller authentication, identity verification, initial context capture	IVR session, member ID systems, call history	Identity mismatch, authentication failure
Eligibility agent	Member benefit and coverage record queries	Payer eligibility APIs, enrollment databases	Coverage gap, inactive plan, coordination of benefits conflict
Clinical rules agent	Authorization criteria evaluation against clinical guidelines	Clinical policy engines, formulary databases, ICD/CPT code tables	Criteria not met, missing clinical documentation
Scheduling agent	Provider availability lookup, appointment creation	EHR scheduling systems, provider networks, calendar APIs	No available providers, out-of-network scenario
Compliance agent	Real-time policy enforcement across all agent actions	HIPAA policy objects, organizational rules, scope restriction tables	Policy violation detected, data minimization breach, irreversible action pending approval
Supervisory agent	Conflict detection, output reconciliation, human handoff coordination	All agent outputs, confidence score logs	Agent output disagreement, confidence below threshold, policy limit reached

Table 1 — Specialized Agent Roles in Healthcare Multi-Agent Architecture.

A healthcare contact center multi-agent system of specialized agent functions would consist of an intake agent to identify, authenticate, and capture an initial context of a caller, an eligibility agent to query member benefit and coverage records, a clinical rules agent to verify an authorization requirement against clinical guidelines, a scheduling agent to access provider availability systems, and a compliance agent to oversee all agent activity in relation to regulatory and policy limits. These agents communicate via shared context objects and event-driven signals, enabling asynchronous coordination without tight coupling.

Toolformer [6] showed that language models can be trained to call external tools without much supervision, which means that they learn when and how to invoke APIs based on context. This has been used to provide dynamic delegation to multi-agent healthcare systems, where a primary reasoning agent is able to recognize subtasks that a specialized agent is authorized to perform and delegates them instead of using hardcoded orchestration logic. A supervisory agent is in charge of execution, detects conflicts or inconsistencies in agent outputs, and solves disagreements via structured deliberation and escalates to human agents when policy limits or confidence limits are not satisfied.

Reflexion [10] proposed an agent way of learning by executing failures, based on verbal self-reflection, where experiential summaries are stored in memory to direct later decisions. In medical contexts, this mechanism allows agents to achieve accuracy in authorization requests with time, minimizing denials that can be due to a partial record or a misaligned clinical code, without retraining the underlying language model.

6. Multi-Tier Memory Architecture of Cross-Call Healthcare Agent Continuity

The memory systems that are necessary in the context of effective autonomous operation in healthcare contact centers extend far beyond the scope of a single interaction. Multiple calls, in many instances, are made by members to healthcare organizations to update them on the changing circumstances, preexisting authorization moves, claims in adjudication, and continuing care coordination occurrences. Callers can only repeat information to agents, as the agents do not have access to cross-call memory and cannot reason about prior interactions.

The proposed multi-tier healthcare voice agent memory architecture is based on the idea of MemGPT [15], which showed that LLM agents can operate in hierarchical memory systems similar to the virtual memory of operating systems, dynamically paging in the active attention windows contexts of the retrieval relevance. The three layer design incorporates short-term memory that retains within-session state such as confirmed identifiers, workflow steps in progress and clarifications pending, episodic memory with the structured interaction summaries per call such as plan-specific coverage rule, formulary tier, clinical criteria and pattern of behavior through historical interaction analysis, and long-term semantic memory coding plan-specific coverage rule, formulary tier, clinical criteria, and behavioral pattern.

The access to memory with telephony-imposed latency restrictions necessitates predictive retrieval: early turns of conversation have agents pre-load the probable necessary episodic and semantic context depending on both the identity and initial intent cues of the caller, covering the latency of memory access by the natural dialogue processing. Policy-sensitive memory filters provide minimum-necessary disclosure at the time of retrieval so that only data needed by the current step of interaction is revealed to the reasoning layer- a direct implementation of the minimum requirements of HIPAA in the agent architecture.

7. Confidence-Based Autonomy Levels and Human in the Loop Control

Confidence modeling is not an option in autonomous healthcare systems; it is the dominant safety mechanism that decides where agentic performance and human supervision intersect. An agent choice is supported by a clear confidence score based on the uncertainty of the reasoning of the LLM, the completeness of the input, the reliability of the tool responses, and the success rates of similar tasks in the past [11]. Confidence levels are correlated with tiers of autonomy: high-confidence decisions are processed autonomously, mid-confidence decisions are processed with confirmation of a caller; low-confidence decisions are processed to human agents with full-context handoff.

Self-Refine [16] proved that without further training, LLMs can use self-generated feedback to refine the results of their outputs, generating even higher quality responses by critiquing and refining the results of their intermediate output. Self-Refine mechanisms applied to healthcare agent confidence modeling allow healthcare agents to carefully inspect their own reasoning chains prior to making commitments to tool invocations or generating caller-facing output, detecting inconsistencies or other information gaps and asking about clarification instead of generating outputs with high confidence but erroneous.

Dynamic confidence adjustment makes sure that confidence does not remain constant throughout the course of an interaction. Confidence is then recalculated at every planning step as more context is accumulated by the responses of the tools, or by confirmation of the caller, or episodic memory recall. This allows gradual automation: an agent can start an interaction with mid-confidence and reach high-confidence autonomous execution when identity and intent are completely known. Another layer of governance is offered by constitutional AI principles [17], which incorporate value-congruent constraints directly into agent reasoning to enforce high-confidence autonomous decisions to be within ethical and regulatory limits even in unforeseen conditions not explicitly encoded into the policy engines of rule-based policies.

Autonomy tier	Confidence range	System behavior	Example healthcare scenario	Human involvement
Tier 1 — Full autonomy	High (≥ 0.85)	Agent executes end-to-end without interruption; outcome logged for audit	Routine eligibility verification with confirmed identity and complete data	None — post-hoc audit only

Tier 2 — Supervised autonomy	High-mid (0.70–0.84)	Agent executes but surfaces summary to caller for passive confirmation	Scheduling appointment after eligibility confirmed; caller confirms provider preference	Passive caller confirmation
Tier 3 — Active confirmation	Mid (0.50–0.69)	Agent pauses at key decision point; presents options; proceeds on caller input	Prior authorization where clinical criteria are partially met; agent requests missing information	Active caller input required
Tier 4 — Assisted escalation	Low-mid (0.30–0.49)	Agent completes resolvable subtasks; escalates outstanding steps with full context	Complex multi-payer coordination; agent resolves eligibility, escalates authorization	Human agent takes over partial workflow
Tier 5 — Full escalation	Low (< 0.30)	Immediate human handoff with complete context packet: reasoning chain, tool call history, confidence scores	Disputed claim, clinical exception, identity verification failure	Human agent owns entire interaction

Table 2 — Confidence-Based Autonomy Tiers and System Behavior.

8. Self-healing and autonomous error recovery workflow execution

The nature of healthcare settings is imperfect systems: timeouts in the back-end, incomplete data, discrepancies in records of members, and disagreements in policy interpretation are all normal modes of operation and not failure modes. AI systems using agentic voice should not assume that errors are a terminal condition of the execution model but rather a recoverable state.

The self-healing execution model enacts four recovery patterns. Combining adaptive backoff and state preservation in retry strategies enables agents to re-invoke failed tool invocations with altered parameters or different service endpoints, maintaining conversational continuity. Selection of alternative execution paths allows agents to accomplish the same workflow objective using a different sequence of tools in case the first one is blocked: e.g., by accessing eligibility information in a second system when the primary eligibility API is not available. Partial completion with follow-up enables the agents to complete resolvable sections of a multi-step workflow and postpone the execution of outstanding steps, informing the caller that the work is partially complete and not giving up on the whole interaction.

Chain-of-thought decomposition makes explicit failure reasoning possible, enabling agents to diagnose the cause of a failure and choose contextually appropriate recovery actions instead of using the generic retry logic. An agent that recognizes a benefits eligibility time out as due to a known downstream system degradation, e.g., can not only alert the caller to the anticipated delay but may also provide alternative paths to resolution instead of possibly retrying or escalating futilely. This interpretive failure management significantly minimizes all the unnecessary escalations, a critical operational efficiency measure in large-scale healthcare contact center implementations [1].

9. Governance, Safety Controls, and Observability in Real-Time

In regulated healthcare settings, agentic voice AI systems must be governed in ways that can be executed at agent speed and not as a post-processing audit but as constraints that are implemented as part of the execution pipeline. The governance layer imposes policy adherence on all tool invocations, steps of response generation, and all state transitions, a perimeter of unbroken safety surrounding agent autonomy.

Policy-conscious implementation provides regulatory constraints within the planning model of the agent. Scope restrictions indicate the number of things that an agent may do on behalf of a particular caller in a particular situation. The minimization of data restricts the amount of information that is brought up in the queries of tools and the response of a caller to only the amount that is necessary due to the step of workflow being undertaken. The approval checkpoints need express authorization prior to the agent taking any irreversible action like filing claims, authorizing, or making schedules. These constraints are modeled as formal policy objects instead of natural language policies and can be enforced deterministically regardless of variability in LLM behavior.

Observability infrastructure: Infrastructure offers decision traceability, with each agent decision being recorded as a formatted record of input context, reasoning chain, tool calls with parameters and responses, confidence scores, and outcomes. This logging system can be used to detect anomalies in real-time, conduct post-hoc regulatory audits, perform benchmarking, and constantly better agent policies. Operational visibility to agent behavior at a national contact center scale is through key performance metrics, including task completion rate, autonomous resolution rate, frequency of escalation, latency per planning step, and rate of policy violation.

10. Agent Scheduling and Optimization of LLM Resources (Load aware)

The systems of agentic voice work under exact real-time requirements, and at the same time, they produce significantly higher compute loads than the traditional intent-based systems. Multi-step reasoning chain inference, parallel tools execution, and memory retrieval. Multi-step reasoning chains, parallel tool execution, and memory retrieval should execute within the conversational latency budget and support thousands of calls simultaneously at the scale of national contact centers.

Model tiering tasks and reasoning tasks to the models that are calibrated to their complexity. Lightweight fine-tuned models (sub-100 ms inference latency) are used to do simple identifier confirmation and eligibility lookup. Complex prior authorization reasoning, clinical matching, and multi-step workflow planning call on more complex models with greater reasoning capacity at the cost of increased latency, used in interactions where their ability has been proven to be required. The tiering approach can cut down the average cost of inference by 40-60 percent in production deployments without compromising the quality of response to high-complexity tasks [18].

Parallel reasoning processes run parallel branches of planning, taking advantage of the timing properties of conversational pauses to push forward a multiplicity of workflow steps at a time. Predictive prefetching proactively loads the tools it predicts are required and pre-invokes the predicted context of the memory during idle conversational periods, which is the analogy of speculative execution in computer architecture. Structured memory summarization as context compression lowers the token usage in reasoning chains, preventing quadratic increases in the cost of attention as agent context grows, as with raw conversational transcripts. These optimizations taken collectively allow agentic voice AI to work within the 400 ms conversational latency budget and conduct planning cycles that require multiple invocations of tools and memory lookups.

11. Conclusion

Self-directed autonomous voice agents as models are a paradigm shift in the work of the healthcare contact centers. agentic voice systems can overcome the shortcomings of conversational AI, and become real agents in the provision of healthcare services by combining LLM-based chain-of-thought reasoning, MCP-controlled tool coordination, multi-agent coordination, confidence-based autonomy limits, multi-level memory, and self-healing workflow execution.

The architectural design in this paper supports the entire range of needs to be deployed in a regulated healthcare setup: the planning and execution capacity to deploy end-to-end automation of workflows; the safety and governance facilities to enable autonomous operation under HIPAA; the memory architecture to support cross-call continuity; and the load management policy to support national-scale real-time operation. With the ongoing evolution of agentic AI systems, healthcare organizations that incorporate strong

architectural frameworks of regulated autonomous performance will be in a position to achieve radical enhancements in operational efficiency, the experience of callers, and regulatory responsibility.
Future research directions are agent-to-agent negotiation across organizational groups to support the workflow of multi-payer authorization; federated agent learning that does not involve centralized data aggregation; regulatory co-design of autonomy governance structures, and cross-modal agents that cut across voice, text, and clinical data modalities in single healthcare operating systems.

References

[1] Sridhar Rangu et al., "Analyzing the Impact of AI-Powered Call Center Automation on Operational Efficiency in Healthcare," ResearchGate, 2023. https://www.researchgate.net/publication/391717271_Analyzing_the_Impact_of_AI-Powered_Call_Center_Automation_on_Operational_Efficiency_in_Healthcare

[2] Md Faiyazuddin et al., "The Impact of Artificial Intelligence on Healthcare: A Comprehensive Review of Advancements in Diagnostics, Treatment, and Operational Efficiency," McKinsey Global Institute, Tech. Rep., Dec. 2021.https://pmc.ncbi.nlm.nih.gov/articles/PMC11702416/

[3] Tom Young et al., "Recent Trends in Deep Learning Based Natural Language Processing," IEEE Comput. Intell. Mag., 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8416973

[4] Ronan Collobert, Jason Weston, "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning," ACM Digital Library, 2018. [Online]. Available: https://dl.acm.org/doi/10.1145/1390156.1390177

[5] Jason Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," arXiv:2201.11903 [cs.CL], 2022. [Online]. Available: https://arxiv.org/abs/2201.11903

[6] Timo Schick et al., "Toolformer: Language Models Can Teach Themselves to Use Tools," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2023. [Online]. Available: https://arxiv.org/abs/2302.04761

[7] Wenlong Huang et al., "Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents," arXiv:2201.07207 [cs.LG], 2022. [Online]. Available: https://arxiv.org/abs/2201.07207

[8] Anthropic, "Introducing the Model Context Protocol," Anthropic Technical Documentation, 2024. [Online]. Available: https://www.anthropic.com/news/model-context-protocol

[9] Shunyu Yao et al., "ReAct: Synergizing Reasoning and Acting in Language Models," arXiv:2210.03629 [cs.CL], 2023. [Online]. Available: https://arxiv.org/abs/2210.03629

[10] Noah Shinn, et al., "Reflexion: Language Agents with Verbal Reinforcement Learning," arXiv:2303.11366 [cs.AI], 2023. [Online]. Available: https://arxiv.org/abs/2303.11366

[11] Richard S. Sutton and Andrew G. Barto, “Reinforcement Learning: An Introduction,” A Bradford Book, 2018. [Online]. Available: https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf

[12] Patrick Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," arXiv:2005.11401 [cs.CL], 2020. [Online]. Available: https://arxiv.org/abs/2005.11401

[13] MICHAE L WOOLDRIDG E, “An Introduction to MultiAgent Systems,” 2nd ed. Hoboken, NJ, USA: Wiley, 2009. [Online]. Available: https://uranos.ch/research/references/Wooldridge_2001/TLTK.pdf

[14] Yongliang Shen et al., "HuggingGPT: Solving AI Tasks with ChatGPT and Its Friends in Hugging Face," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2023. [Online]. Available: https://arxiv.org/abs/2303.17580

[15] Charles Packer et al., "MemGPT: Towards LLMs as Operating Systems," arXiv:2310.08560 [cs.AI], 2024. [Online]. Available: https://arxiv.org/abs/2310.08560

[16] Aman Madaan et al., "Self-Refine: Iterative Refinement with Self-Feedback," arXiv:2303.17651 [cs.CL]. 2023. [Online]. Available: https://arxiv.org/abs/2303.17651

[17] Yuntao Bai et al., "Constitutional AI: Harmlessness from AI Feedback," arXiv:2212.08073 [cs.CL], 2022. [Online]. Available: https://arxiv.org/abs/2212.08073

[18] Tim Dettmers et al., "QLoRA: Efficient Finetuning of Quantized LLMs," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2023. [Online]. Available: https://arxiv.org/abs/2305.14314

[19] Long Ouyang et al., "Training Language Models to Follow Instructions with Human Feedback,"arXiv:2203.02155 [cs.CL], 2022. [Online]. Available: https://arxiv.org/abs/2203.02155

[20] Aohan Zeng et al., "AgentTuning: Enabling Generalized Agent Abilities for LLMs," arXiv preprint arXiv:2310.12823 [cs.CL], 2023. [Online]. Available: https://arxiv.org/abs/2310.12823

[21] Qingxiu Dong et al., "A Survey on In-Context Learning," arXiv preprint arXiv:2301.00234, Jan. 2023. [Online]. Available: https://arxiv.org/abs/2301.00234
[22] Tom B. Brown et al., "Language Models are Few-Shot Learners," arXiv:2005.14165 [cs.CL] 2020, [Online]. Available: https://arxiv.org/abs/2005.14165
[23] Eric J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," Nat. Med., 2019. [Online]. Available: https://www.nature.com/articles/s41591-018-0300-7