International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(4s) 2026 pp. 975-983
https://doi.org/10.70917/ijcisim-2026-2610
Article

A Reference Architecture for Real-Time Conversational Telephony Intelligence in Regulated Healthcare Systems

Bhargavi Kalicheti

Independent Researcher, USA

Abstract: Healthcare insurance organizations face mounting pressure to deliver scalable, accurate, and compliant telephony services to millions of members, providers, and pharmacies. Traditional interactive voice response (IVR) systems, founded on deterministic menu logic and constrained speech grammars, are fundamentally inadequate for the complexity of modern healthcare interactions. This paper presents a cloud-native reference architecture for real-time conversational telephony intelligence designed to transform healthcare insurance contact center operations. The proposed architecture integrates large language models (LLMs), natural language understanding (NLU) pipelines, conversational orchestration layers, and enterprise workflow systems within a compliant, observable, and horizontally scalable infrastructure. Key design principles include stateless service composition, policy-aware generative reasoning, and multi-tier inference routing to balance conversational intelligence with deterministic safety controls. The architecture addresses five interdependent concerns: real-time voice ingestion, intelligent intent processing, workflow-grounded response generation, elastic scalability, and regulatory governance. We analyze system-level trade-offs, deployment strategies, and ethical safeguards necessary for production-grade deployment at national scale. The proposed architecture demonstrates significant potential for improving intent recognition accuracy, first-contact resolution rates, and operational efficiency while maintaining strict adherence to healthcare compliance requirements. This work offers a reusable architectural blueprint for healthcare organizations advancing toward conversational AI-driven telephony intelligence.

Keywords: Conversational Voice AI, Healthcare Telephony, Large Language Models, Cloud-Native Architecture, Natural Language Understanding, IVR Transformation, Generative AI Governance

1. Introduction

Healthcare insurance companies maintain some of the biggest and most intricate telephony systems in the service economy. Members initiate millions of interactions every month to get verified as eligible, receive information about benefits, claims, and are seeking information on prior authorization status. The providers and pharmacies rely on voice channels with time-sensitive administrative inquiries that have a direct connection to clinical decision-making and patient access to care. Although these channels are critically important in operations of healthcare insurance, the majority of healthcare insurance telephony systems remain based on interactivity-based voice response infrastructures developed decades ago, when intelligent voice interaction seemed a far-fetched possibility due to the lack of modern natural language understanding, generative AI, and cloud-scale computing capabilities [1], [20]. The constant disconnect between the requirements of the callers and the capabilities of the legacy systems is a major structural problem that faces healthcare organizations, which want to enhance the quality of services as well as the efficiency of their operations.

The shortcomings of rule-based IVR systems have been long known to operational practice and the literature. These systems limit callers to pre-defined menu hierarchies, use limited speech grammars that cannot handle the variability of natural language, and cannot maintain the context of conversation across interaction turns [4]. The consequence is a chain reaction of operational inefficiencies: high call transfer rates, high average handle time, high abandonment rates, and a long-term overhead on the human agents to fix the breakdowns of automated systems. With the cost of maintaining IVR-based telephony infrastructure growingly challenging to afford within healthcare

organizations to operate and meet regulatory requirements that continue to expand, the cost of the telephony infrastructure has come under a heightened level of scrutiny [5].

New developments in large language models provide previously unheard-of abilities to interpret free-form natural language, preserve multi-turn conversations, and reason about ambiguous and complex inputs [6]. Language models that learn on general text corpora can be fine-tuned to domain-specific scopes and be prompted with structured prompts to support more flexible intent detection and contextualized response generation. The immediate use of LLMs in healthcare telephony, however, presents serious technical and governance issues. The design space of generative AI systems in this field is limited by latency sensitivity, risk of hallucinations, compliance needs, and deterministic performance of regulated workflows [2], [7]. The scale of the conversational intelligence required to roll out the national healthcare insurance business requires an architecturally sound production-tested system that trades generative potential with the dependability of operations [12], [22].

This article introduces an end-to-end reference architecture of real time conversational telephony intelligence in regulated health care systems. The proposed design is based on the principles of cloud-native design and combines the use of LLMs, NLU components, and conversational orchestration, enterprise workflow systems, and multi-layer governance controls into a single, production-deployable ecosystem. Three key contributions of the paper are as follows: it discusses the architectural concepts that separate production-grade conversational voice AI and research-oriented deployments; it proposes a layered ecosystem model that explicitly considers integration of generative reasoning with deterministic policy enforcement; and it discusses scalability and governance mechanisms needed to support sustained operation at a national scale. The rest of the paper will be structured in the following way: Section 2 describes drawbacks of legacy IVR systems; Sections 3-8 elaborate on the architectural framework; Section 9 talks about directions of the future; and Section 10 provides a conclusion.

2. Weaknesses of Traditional Healthcare IVR Systems

Traditional IVR systems are designed based on deterministic call flows where all interactions paths are defined by the system designers. Such philosophy of design puts an asymmetric load on callers, who have to adjust their behavior to the system navigation forms, instead of asking what they want in their own words. The intent recognition is achieved via touch-tone inputs or limited scope speech grammar, which do not support the natural diversity of healthcare questions [4]. Callers with non-standard terms, changing the subject during a call or indicating complex requests often misrouted or sent back to beginning. This structural inflexibility generates systematically high intent misclassification rates compared to human-assisted channels, leading to increased transfer rates and poorer overall caller experience.

A second category of restrictions has to do with conversational memory and state persistence. IVR systems are made to handle interactions in an isolated, stateless manner. Every call starts with no prior context and callers are consistently asked to re-authenticate and re-enter information that was previously given during steps of interaction or after agent transfers. In healthcare insurance, where transactions frequently require overlaying eligibility requirements, benefit regulations specific to the plan, and multi-level service processes, this lack of contextual continuity is a basic incompatibility with the requirements of the callers [5]. Its operational implications are high average handle time, high rate of repeat calls, and quantifiable decrease in scores of customer satisfaction, which directly translate into high operational cost and organizational ability to support increasing membership and provider bases.

In addition to the quality of interaction, the traditional IVR systems have great maintenance and adaption fees restricting the agility of the organization. Any change in call flow logic, new intent grammar library, or new benefit structure takes weeks to months of dedicated development cycles. This adaptation latency in healthcare insurance, where benefit designs, network designs, and regulatory requirements are changing constantly, produces periods where system behavior and the real world of operations are out of step with each other [1]. Organizations can either use inaccurate automated replies or direct all the associated calls to human operators, neither of which is operationally viable at scale. The aggregate impact of these structural constraints has created a resurgence of interest in conversational AI engines with the ability to adjust dynamically without there being a complete redevelopment cycle.

Performance Metric	Traditional IVR (Baseline)	Conversational Voice AI (Target)	Improvement Direction	Notes

Intent Recognition Accuracy (%)	52	88	Higher is better	Based on free-speech input scenarios
First-Contact Resolution Rate (%)	34	71	Higher is better	Single-call resolution without transfer
Call Transfer Rate (%)	58	18	Lower is better	Unnecessary agent transfers per 100 calls
Average Handle Time (minutes)	7.4	3.2	Lower is better	End-to-end call duration including self-service
Caller Satisfaction Score (1–10)	4.1	7.8	Higher is better	Post-call survey composite
System Adaptation Time (weeks)	12	1	Lower is better	Time to deploy benefit or policy update
Natural Language Variability Support	Narrow grammars only	Unrestricted free-form	Qualitative	Accent, dialect, multi-intent
Multi-Turn Context Retention	None (stateless)	Full session context	Qualitative	Persistent across dialog turns

Table 1: Performance Metric Comparison – Traditional IVR vs. Conversational Voice AI (Conceptual)

3. Paradigm Shift: Conversational Voice AI

Conversational voice AI is a paradigm shift in the model of telephony interaction to navigate the menu to intent-based dialogue. This paradigm allows the caller to state his or her requirements in a natural language and the system understands the meanings, maintains the context and dynamically adjusts the conversation towards the aim of the caller [4]. Instead of forcing callers to standardized routes, conversational systems embrace natural variations in human speech, such as incomplete expressions, switching to subsequent topics, corrections, and follow-up clarifications. This change is facilitated by progress in automatic speech recognition, natural language understanding, and generative language modeling, which together allow systems to perceive, reason, and respond in a conversational level of sophistication previously accomplished only by human agents [13], [16], [17].

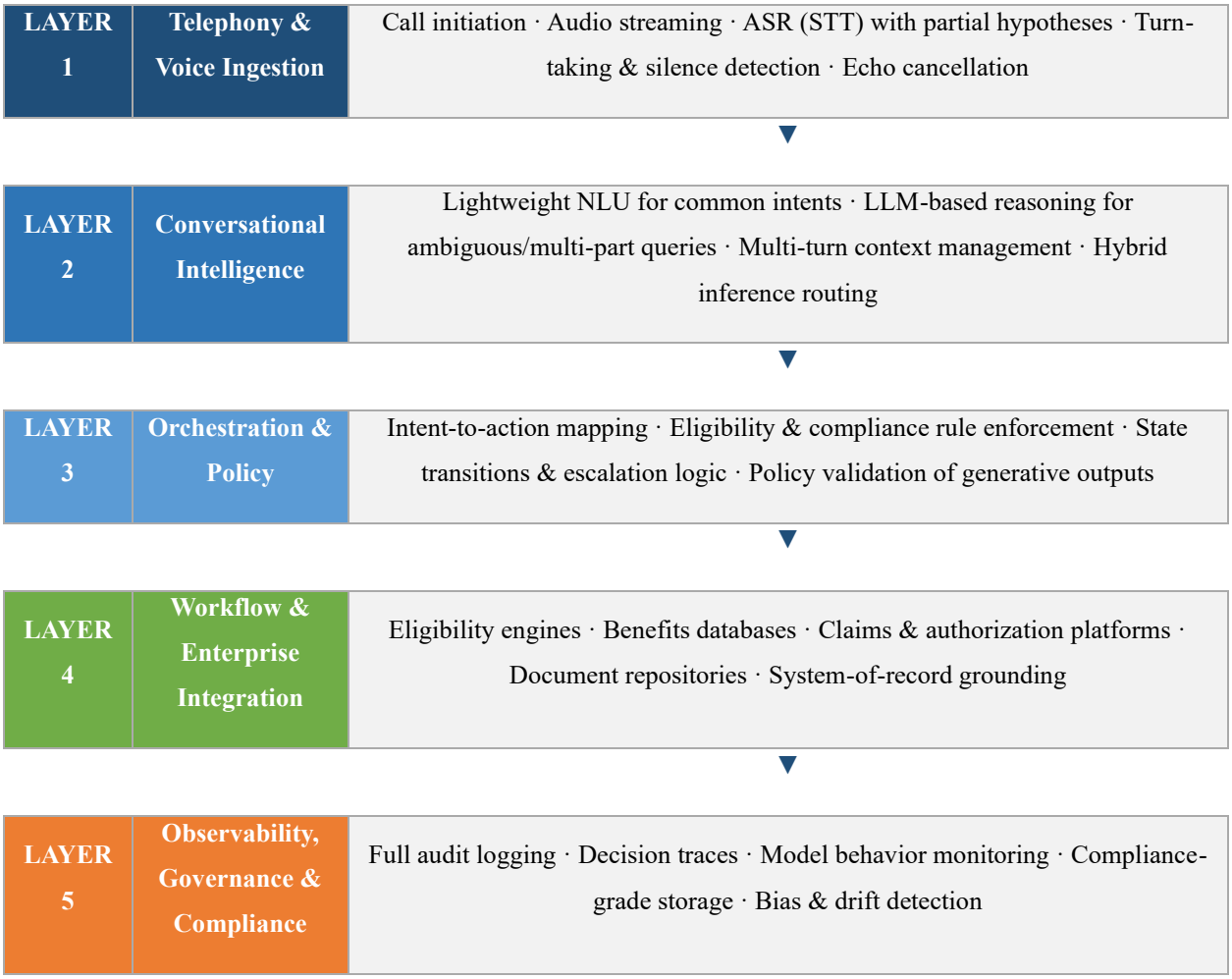
The capabilities of modern voice AI systems to achieve conversational intelligence relies on the large language models. Such models have impressive emergent language comprehension, multi-step reasoning and coherent response generation skills in a variety of domains [10]. Utilizing LLM in the context of organized conversation management systems, it is possible to decode complex multi-part queries, infer implicit intent based on partial information and produce contextually based responses in agreement with domain specific information [2], [6]. In telephony in healthcare insurance, this is a system that can process layered eligibility questions, benefit comparisons and authorization status questions in a single coherent conversational turn - interaction that would otherwise need multiple menuing or agent-to-agent transfers in a traditional IVR setting [3].

Although such is possible, the architectural needs of the implementation of LLMs in regulated healthcare telephony are such that such phenomena are absent in the general-purpose conversational AI systems. Healthcare communication is a field where the health information is secured, must be completely audited, and should be in

accordance with the regulatory frameworks such as the Health Insurance Portability and Accountability Act. Generative models generate probabilistic outputs that can differ on the same inputs - a property that cannot exist in regulated administrative processes [12]. Moreover, large model inference latency profiles do not match with sub-second voice telephony response time requirements. The architecture discussed in this paper directly provides solutions to these tensions by incorporating generative intelligence into a deterministic, policy-enforcing orchestration layer that maintains compliance without compromising the quality of conversations.

4. End-to-End Ecosystem Architecture

The ecosystem architecture proposed is structured into five mutually dependent functional layers that can be scaled independently and managed by a common observability infrastructure. The Telephony and Voice Ingestion Layer process call establishment, audio streaming, speech-to-text processing and turn taking. With the history of low-latency automatic speech recognition, based on the development of end-to-end deep learning in speech [23], partial hypothesis generation is possible, allowing downstream intent processing to be started before utterance completion, minimizing the effective response latency. The Conversational Intelligence Layer is a lightweight NLU classifiers-based layer that uses LLM-based reasoning selectively when dealing with ambiguous or multi-part queries [15]. This hybrid inference model maintains performance efficiency to high frequency interaction pattern but still has the generative flexibility to support a complex or exception-driven conversation.



Flow Diagram 1: End-to-End Conversational Voice AI Ecosystem Architecture (Five-Layer Model)

The Orchestration and Policy Layer manages the conversation by mapping identified intent to permitted actions, implementing eligibility and compliance policies, and state transitions and escalation logic. This layer prevents the generation of model outputs directly performing external actions without the policy is verified, maintaining a

deterministic control over consequential operations like eligibility checks, authorization requests or document access requests [12]. The Workflow and Enterprise Integration Layer links the results of conversations to authoritative healthcare information systems such as eligibility engines, benefits databases, claims systems, and clinical authorization platforms. The reactions provided to callers are based on validated, system-of-record information as opposed to generated content, which makes them factual and defensible under regulations. This divide between conversational intelligence and data access is an architectural principle that has been observed to set the difference between production-grade and experimental systems [2], [9].

Observability, Governance, and Compliance Layer offers nonstop tracking, auditing, and governing capabilities of the whole ecosystem. Telephony systems in healthcare are also regulated so that the decisions of the system should be fully audited, the interaction logic should be controlled by a version and the routing behavior has to be explained [18]. This tier records transcripts of interactions, decision records, model invocation records and latency telemetry to compliance grade storage, allowing post-hoc audit and continuous quality improvement. Automated evaluation pipelines detect drift, bias signs, and policy violations in model behavior by observing anomalies that are sent to humans [11], [14]. The incorporation of governance as a structural layer, and not a post-deployment consideration, is indicative of a privacy-by-design practice and in line with regulatory requirements of AI deployments of healthcare6. Cloud-Native Design Principles

Cloud-native architectures offer the underlying capability that is needed to support conversational voice AI on the scale and reliability that is needed in national healthcare insurance telephony. The ecosystem embraces stateless service composition as a fundamental design concept: conversational processing elements do not hold local session state, and all context is externalized to low-latency distributed data stores. The design allows horizontal scaling of separate service parts in reaction to changes in call volume with no coordination overhead or session affinity limitations [19], [21]. Stateless design has been used to help the platform absorb surge traffic during peak times such as open enrollment cycles, benefit renewal windows, or even events related to regulatory deadlines by quickly scaling the busiest components of the system to a state of peak performance and ensuring the same latency profiles across all busy caller sessions.

Microservice decomposition isolates faults, avoiding cascading failures in isolated components of the system. The speech-to-text service, intent classification engine, LLM inference cluster, workflow integration adapter, and text-to-speech synthesis service are all independently deployed services with explicit API contracts and circuit-breaker patterns that gracefully degrade the service in case of unavailability of upstream dependencies [19]. Canary and blue-green deployment strategies allow safe deployment of conversational logic updates, such as NLU model versions, LLM prompt execution, and workflow rule updates, without affecting active calls. Practices in infrastructure-as-code maintain versioned, reproducible, auditable deployment settings, which can be used to support regulatory compliance and operational resilience in the event of an infrastructure change or disaster recovery operation.

The multi-region deployment architectures also enhance the performance and resilience properties of conversational voice AI on the national level. Key inference components can be deployed regionally to minimize round-trip latency by serving telephony traffic using geographically close compute resources [21]. The multi-zone redundancy in each cloud region serves to defend against local infrastructure outages, which would otherwise impair the availability to the geographically concentrated populations of callers. Active-active deployment models, in which several regions are used to serve live traffic, can provide a zero-downtime maintenance window and can facilitate regulatory needs of business continuity and disaster recovery in healthcare operations. The horizontal scalability, fault isolation and geographic distribution together create a resilience posture that is suitable to mission-critical healthcare telephony infrastructure.

6. Generative Reasoning Grounding into Regulated Workflows

The fundamental engineering challenge to integrating generative AI in regulated healthcare telephony lies in the fact that while generative models exhibit probabilistic behaviors, the telephony operation is grounded within deterministic workflows. Unguided outputs from LLMs are unsuitable in the healthcare insurance context, which necessitates responses that are both factually accurate, policy-aligned and defensible in regulatory compliance. The proposed architecture achieves the desired control by positioning generative reasoning as an inferential and synthesising layer rather than as an automated conversational agent [2], [9]. The output of generative models is constrained using prescribed response schemas which define what kinds of responses are allowed, what data sources can be referenced and what kinds of compliance assurances are mandatory. The generative model operates in a precisely defined prompt context that specifies information regarding the caller profile, interaction circumstances and operational constraints.

A series of policy validation layers review all generative outputs prior to their dissemination to callers; it assesses conformance to operational policies, regulatory constraints and conversational state variables. These validators deterministically evaluate the generated content against rules specified in structured rule bases, which are compiled from coverage rules, regulatory laws and operational policies. Response items that fail validation are then either rejected (regenerated under stronger controls) or relayed to a human operator path, ensuring the full conversational context is transmitted to them [12], [14]. Such validation infrastructure ensures that irrespective of generative model outputs, the caller ultimately receives factually correct and policy-compliant information and maintains trust in the system. Structured execution plans control interaction execution by constraining generative reasoning and forcing execution along a defined set of valid actions rather than independent model-guided steps.

A requirement in terms of the core architecture is decision traceability; it should be natively supported. The output decision trace log, recorded for every interaction, contains signals indicating the caller intent, data retrieved and contextual rules that guided the generation. These traces are stored on a compliance-compliant infrastructure under suitable access controls and storage durations, and can be accessed for quality assurance and auditing processes [11], [18]. Decision traces must also be human-readable to support auditing processes, where compliance and quality analysts should be able to understand system behavior with a minimum level of technical expertise regarding the model internals. This process also establishes confidence with organizational members, and makes them comfortable with using generative AI in high-stakes operations such as healthcare insurance.

7. Scalability Architecture for National-Scale Healthcare Telephony Intelligence

Supporting conversational voice AI at a national scale mandates a scalability infrastructure that can handle both regular volumes of call traffic, and sporadic surges. It proposes decomposing the voice AI pipeline into a set of individually scalable microservices that range from ASR ingestion, to NLU and intent analysis, knowledge retrieval and grounding, generative reasoning and response planning, to text-to-speech generation [21]. Each of these service layers are horizontally scaled based on workload-specific indicators which allow the system to cope with varied volumes of call traffic without propagating the bottleneck up the pipeline [19]. Computationally intensive generative inference workloads are addressed through tiered model routing; requests can be directed to more lightweight inference models to fulfill straightforward intents and reserved for advanced, high complexity models that may be required for a variety of reasoning steps or for error resolution.

STAGE 1	**Caller Interaction Initiation**	Telephony ingress → Predictive capacity provisioning based on historical traffic patterns
▼		
STAGE 2	**Horizontal Microservice Scaling**	ASR / NLU / LLM / TTS services scale independently · Workload-specific signals · No cascading bottlenecks
▼		
STAGE 3	**Stateless Conversational Execution**	Context externalized to distributed state stores · Stateless instances replaced without disruption · Multi-region replication
▼		
STAGE 4	**Multi-Tier LLM Inference Routing**	Low-complexity → lightweight model · High-complexity → advanced LLM · Geographic routing for latency optimization
▼		

STAGE 5	**Observability-Driven Autoscaling**	Closed-loop metrics (concurrency, latency p95, containment rate) → automated scaling policies → continuous refinement

Flow Diagram 2: Scalability Architecture for National-Scale Healthcare Telephony Intelligence

Externalized conversational state support allows seamless horizontal scaling without any risk of losing session context in the distributed system. Every context-related aspect of the interaction-including validated caller attributes, confirmed intent, history of conversational turns, and dialogue state-is maintained in low-latency, geographically replicated distributed stores, ensuring continuous operation even in the event of failure at the instance level or at a given data center [19], [21]. The stateless design allows for instances of the voice AI system to be spun up as needed to handle increased workload, thus maintaining conversational integrity at the time of failover. Scaling strategies based on history can be employed to preemptively provision computing resources for predictable periods of elevated traffic, such as holidays or benefit enrollment periods. This provisioning reduces start-up time for new instances and maintains caller experience quality at these times.

An observability driven autoscale infrastructure will trigger scale-up or scale-down events across the voice AI pipeline in response to system performance metrics such as number of active users, end-to-end latency, ASR throughput, frequency of LLM calls, and percentage of issues resolved by the system without human intervention [14]. Scaling plans are precisely crafted to respond optimally to each performance indicator while balancing system resource utilization. This feedback-controlled scaling system can also provide detailed post-mortem information by correlates scaling events with measures of interaction quality, allowing for subsequent tuning of autoscale policies and of compute resources. Combining anticipatory and responsive scaling allows the voice AI system to reliably serve dozens of millions of healthcare insurance telephone calls annually.

8. Ethical and Responsible AI Considerations

Calls within a healthcare insurance telephony context can originate from some of the most vulnerable populations in healthcare system, including elderly callers, those managing chronic diseases or callers disputing their insurance claim. Deploying AI-driven telephony systems in these high-stakes scenarios requires ethical responsibility that extends beyond accuracy to fairness, transparency and human dignity [11], [14]. Bias detection is a crucial initial step; ASR systems must be rigorously tested for fairness of accuracy and error rate across demographics, accents and speech patterns found in the caller population. NLU models must be evaluated for accuracy of intent classification across disparate linguistic populations to avoid systematically poorer self-service rates for certain demographics, which can indirectly lead to poorer health outcomes [13].

Transparency must also be provided. Callers must be clearly notified of the fact they are speaking to a system that uses AI and understandable explanations for automated decisions should be readily available. In the healthcare insurance context, where interactions often involve disputed claims and crucial coverage information, this level of clarity is expected by regulations [12]. The proposed architecture supports transparency through clearly defined explanation mechanisms, human-readable decision traces and readily available escalation paths to human agents that pass down the relevant conversational history [8]. These features are essential in order to ensure the cost savings from AI are not attained at the expense of the caller's ability to influence or appeal decisions made on their behalf.

Continuous post-deployment monitoring is key for detecting unanticipated emergent issues such as system drifts in the models, negative side-effects and fairness challenges that may not manifest at initial deployment. Built-in monitoring infrastructure within the observability layer tracks indicators such as bias and hallucination, and surfacing anomalies through automated alerting workflows that are routed to human operators for review [14], [18]. External audits should be routinely performed and findings must be reported across departments of quality assurance, compliance and engineering. An ethical design, supported by strong governance in the AI pipeline, should be the norm.

9. Future Directions

Future avenues within the realm of conversational voice AI in healthcare will involve more intelligent, agentic conversational systems. Such systems will be capable of complex multi-step task planning, performing actions, and initiating outreaches, and are in contrast to the current response-based conversational AI systems. These agentic systems will be able to schedule follow-ups, perform member and provider assistance, execute various backend administrative processes, and ultimately complete entire care pathways without constant user initiation [3]. Advanced multilingual and accent robust voice models is another important area of future work, which will be especially crucial

for organizations that serve linguistically diverse communities within nationally mandated programs such as Medicaid and Medicare Advantage. Future work in language and speaker representation learning is likely to achieve single models that can effectively serve various linguistic groups and avoid the overhead of multiple, individually deployed systems [16].

In-depth caller data integration can facilitate anticipatory intent recognition and advanced service design. Systems will be able to determine user needs by considering factors such as previous call history, utilization of benefits and global demographic data [9]. In the healthcare domain where additional privacy risks are introduced by prolonged behavioral profiling, managing these risks is an absolute prerequisite to realizing these capabilities. Privacy-preserving AI technologies such as federated learning will permit the personalisation of services without the risks of large-scale sensitive data collection, making advances in the domain align with regulatory requirements. Regulatory definitions of allowable AI system behaviors in healthcare will continue to be a dominant driver for the evolution of conversational AI architectures [18].

10. Conclusion

This paper has presented an end-to-end reference architecture for real-time conversational telephony intelligence in regulated healthcare systems. The architecture addresses the structural limitations of legacy IVR platforms by integrating generative language models, natural language understanding pipelines, policy-aware orchestration, and cloud-native scalability within a unified, governance-embedded ecosystem. The proposed design demonstrates that the capabilities of large language models can be responsibly harnessed within healthcare telephony by constraining generative behavior within deterministic policy frameworks, grounding responses in authoritative enterprise data, and embedding compliance controls as structural system components rather than post-deployment additions [2], [18]. The result is a platform capable of delivering intelligent, contextually aware conversational interactions while maintaining the reliability, auditability, and regulatory compliance that healthcare insurance operations demand.

As healthcare organizations confront growing call volumes, escalating operational costs, and increasing member expectations for service quality through voice channels, conversational voice AI architectures represent a structural transformation of healthcare telephony rather than an incremental improvement. The architectural principles outlined in this paper provide a production-validated blueprint for organizations embarking on this transformation, offering design guidance that balances the promise of generative intelligence with the imperatives of safety, compliance, and operational resilience [1], [14]. Future progress in foundation model capabilities, privacy-preserving computation, and evolving regulatory frameworks will further expand the boundaries of what is achievable in healthcare telephony intelligence, making principled architectural design an enduring requirement for responsible innovation in this critical domain.

References

1. E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," Nature Medicine, 2019. https://www.nature.com/articles/s41591-018-0300-7
2. K. Singhal et al., " Large language models encode clinical knowledge," Nature, 2023. https://www.nature.com/articles/s41586-023-06291-2
3. M. Moor et al., "Foundation models for generalist medical artificial intelligence," Nature, 2023. https://www.nature.com/articles/s41586-023-05881-4
4. L. Laranjo et al., "Conversational agents in healthcare: A systematic review," Journal of the American Medical Informatics Association, 2018. https://pubmed.ncbi.nlm.nih.gov/30010941/
5. A. Palanica et al., "Physicians' perceptions of chatbots in health care: Cross-sectional web-based survey," Journal of Medical Internet Research, 2019. https://pmc.ncbi.nlm.nih.gov/articles/PMC6473203/
6. T. B. Brown et al., "Language models are few-shot learners," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2020. https://arxiv.org/abs/2005.14165
7. J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), 2019. https://arxiv.org/abs/1810.04805
8. A. Vaswani et al., "Attention is all you need," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017. https://arxiv.org/abs/1706.03762
9. R. Bommasani et al., "On the opportunities and risks of foundation models," arXiv preprint arXiv:2108.07258, 2021. https://arxiv.org/abs/2108.07258
10. J. Wei et al., "Emergent abilities of large language models," Transactions on Machine Learning Research, 2022. https://arxiv.org/abs/2206.07682

11. Z. Obermeyer et al., "Dissecting racial bias in an algorithm used to manage the health of populations," Science, 2019. https://pubmed.ncbi.nlm.nih.gov/31649194/
12. D. S. Char et al., "Implementing machine learning in health care — addressing ethical challenges," New England Journal of Medicine, 2018. https://pubmed.ncbi.nlm.nih.gov/29539284/
13. W. N. Price and I. G. Cohen, "Privacy in the age of medical big data," Nature Medicine, 2019. https://pubmed.ncbi.nlm.nih.gov/29539284/
14. N. Schwalbe and B. Wahl, "Artificial intelligence and the future of global health: Insights from a digital intelligence platform," npj Digital Medicine, 2020. https://pmc.ncbi.nlm.nih.gov/articles/PMC7255280/
15. J. Gao et al., "Neural approaches to conversational AI: Question answering, task-oriented dialogues and social chatbots," Foundations and Trends in Information Retrieval, 2019. https://arxiv.org/abs/1809.08267
16. A. Y. Hannun et al., "Deep speech: Scaling end-to-end speech recognition," arXiv preprint arXiv:1412.5567, 2014. https://arxiv.org/abs/1412.5567
17. S. Young et al., "Pomdp-based statistical spoken dialogue systems: A review," Proceedings of the IEEE, 2013. https://www.semanticscholar.org/paper/POMDP-Based-Statistical-Spoken-Dialog-Systems%3A-A-Young-Gasic/84b520a8d6de79f62bb095b565d077e95bfb6f5b
18. P. Rajpurkar et al., "AI in health and medicine," Nature Medicine, 2022. https://www.nature.com/articles/s41591-021-01614-0
19. N. Dragoni et al., "Microservices: Yesterday, today, and tomorrow," in Present and Ulterior Software Engineering, M. Mazzara and B. Meyer, Eds. Cham, Switzerland: Springer, 2017. https://link.springer.com/chapter/10.1007/978-3-319-67425-4_12
20. A. L. Beam and I. S. Kohane, "Big data and machine learning in health care," JAMA, 2018. https://pubmed.ncbi.nlm.nih.gov/29532063/
21. B. Burns et al., "Borg, omega, and kubernetes: Lessons learned from three container-management systems over a decade," ACM Queue, 2016. https://dl.acm.org/doi/10.1145/2890784
22. A. B. Rajkomar et al., "Machine learning in medicine," New England Journal of Medicine, 2019. https://dl.acm.org/doi/10.1145/2890784
23. G. Hinton et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Processing Magazine, 2012. https://ieeexplore.ieee.org/document/6296526