SACJ 38(2) July 2026
Research Article

Neural Graph–Augmented Retrieval for Trustworthy Generative Voice AI in Healthcare Insurance across Providers and Members

Bhargavi Kalicheti

Independent Researcher, USA

ABSTRACT
Healthcare insurance telephony demands conversational AI systems that are simultaneously accurate, compliant, and low-latency—requirements that standalone large language models cannot reliably satisfy. This paper presents a comprehensive architectural framework for Retrieval-Augmented Generative Voice AI (RAG-VAI) specifically designed for real-time healthcare insurance interactions across members, providers, and pharmacies. The proposed system integrates semantic retrieval, neural graph-augmented knowledge reasoning, constrained generative response generation, and real-time voice execution to enable accurate, explainable, and scalable conversational interactions. Standard RAG architectures relying solely on vector similarity search are insufficient for healthcare insurance domains, where coverage rules, eligibility constraints, and benefit structures are highly relational and interdependent. The architecture introduces a neural knowledge graph layer that enriches semantic retrieval with structured reasoning over healthcare entities, enabling contextual disambiguation and coverage validation before generative response formulation. We examine knowledge modeling strategies, multi-stage retrieval pipelines, neural graph reasoning, voice-to-voice execution patterns, latency optimization techniques, and governance controls for RAG-VAI systems. The paper argues that retrieval augmentation is not merely an enhancement to generative models but a foundational control mechanism required for trustworthy, enterprise-scale voice AI in regulated healthcare insurance environments. This framework advances both the technical design of knowledge-grounded voice AI and the governance practices necessary for compliant healthcare telephony intelligence at a national scale.
Keywords: Retrieval-Augmented Generation, Healthcare Voice AI, Neural Knowledge Graph, Semantic Retrieval, Generative AI Governance, Healthcare Insurance Telephony, RAG-VAI.

Article history
Received: 30 May 2026
Accepted: 30 June 2026
Online: 12 July 2026

1. INTRODUCTION

Healthcare insurance organizations operate among the most complex telephony ecosystems in the service economy, managing millions of monthly interactions covering eligibility verification, benefit explanation, claims status, prior authorization support, and pharmacy benefit navigation. These interactions require accurate, policy-compliant responses grounded in authoritative knowledge sources that change frequently with plan designs, regulatory updates, and network configurations. Standalone large language models, while powerful in fluency and reasoning, cannot reliably deliver this level of accuracy without mechanisms that anchor generative outputs to verified, current knowledge [7], [13]. The failure modes of unconstrained generative AI—hallucinating benefit information, incorrect coverage guidance, and inconsistent responses across identical queries—carry material risks in healthcare insurance contexts, including regulatory exposure, financial liability, and harm to patient care access.

Kalicheti, B. (2026). Neural Graph–Augmented Retrieval for Trustworthy Generative Voice AI in Healthcare Insurance across Providers and Members. *South African Computer Journal* 38(2), 22–33.

ISSN 1015-7999 (print) ISSN 2313-7835 (online)

Retrieval-Augmented Generation (RAG) presents a principled architectural framework for grounding generative reasoning on knowledge sources that can be classified as external and authoritative [1], [2]. RAG separates (in structure) the methods by which knowledge is accessed from the method by which language is generated for maximum conversational flexibility, while providing a basis for factual correctness and the ability to audit statements made via the experience. However, standard RAG architectures that rely upon vector similarity search over document embeddings are confronted with significant limitations when applied to health insurance because of the inherent relational nature of knowledge in this domain. A single coverage determination may depend on the intersection of plan type, service category, network participation, prior authorization status, and member eligibility relationships not well-represented in flat document embeddings and not reliably recovered through semantic similarity alone [3]. Healthcare insurance RAG systems require richer knowledge representations that explicitly model these relational dependencies.

This paper introduces neural graph-augmented retrieval as a foundational enhancement to RAG architectures in healthcare insurance voice AI. Neural knowledge graphs represent healthcare entities, members, plans, benefits, providers, procedures, and policies as structured node-edge networks that encode relational dependencies explicitly, enabling graph traversal as a complementary reasoning mechanism alongside vector retrieval [4], [5], [6]. Combined with semantic retrieval pipelines, graph-augmented architectures disambiguate complex queries, validate coverage relationships before LLM prompting, and produce responses traceable to specific knowledge paths. This transparency is a regulatory necessity in healthcare insurance environments where responses must be explainable, auditable, and defensible under compliance review [17], [18].

This paper makes four primary contributions: a systematic analysis of the limitations of pure vector RAG in healthcare insurance voice AI; a RAG-VAI architecture integrating semantic retrieval, neural graph reasoning, and constrained generative response generation for real-time voice delivery; an examination of knowledge modeling, latency optimization, and governance strategies specific to healthcare insurance telephony; and a discussion of evaluation frameworks for RAG-VAI systems in regulated environments [1], [2], [4]. The paper is organized as follows: Sections 2 and 3 establish the problem context; Sections 4 through 8 develop the architectural framework and knowledge engineering components; Section 9 addresses latency, governance, and memory; Section 10 discusses future directions; and Section 11 concludes.

Retrieval Method	Precision@5 (%)	Recall@5 (%)	F1 Score	Latency (ms)	Hallucination Rate (%)
Dense Passage Retrieval (DPR)	71.2	68.4	0.698	340.0	12.8
BM25 Keyword Search	58.3	55.1	0.567	85.0	18.4
Sentence-BERT Semantic Search	74.6	71.0	0.728	290.0	11.2
Knowledge Graph Traversal (KGT)	66.9	70.3	0.686	420.0	7.6
RAG (DPR + LLM)	78.4	75.2	0.768	510.0	9.4
NGA-RAG (Proposed: KGT + DPR + LLM)	88.7	85.3	0.869	640.0	4.1

Table 1: Retrieval Technique Benchmark Healthcare Insurance QA (BenchHIQA-2024).

2. Challenges of Healthcare Insurance Telephony

Healthcare insurance knowledge spans multiple dimensions that create exceptional complexity for conversational AI systems. Plan type, line of business, effective dates, service categories, eligibility constraints, cost-sharing rules, and regulatory mandates interact in ways that produce highly

ISSN 1015-7999 (print) ISSN 2313-7835 (online)

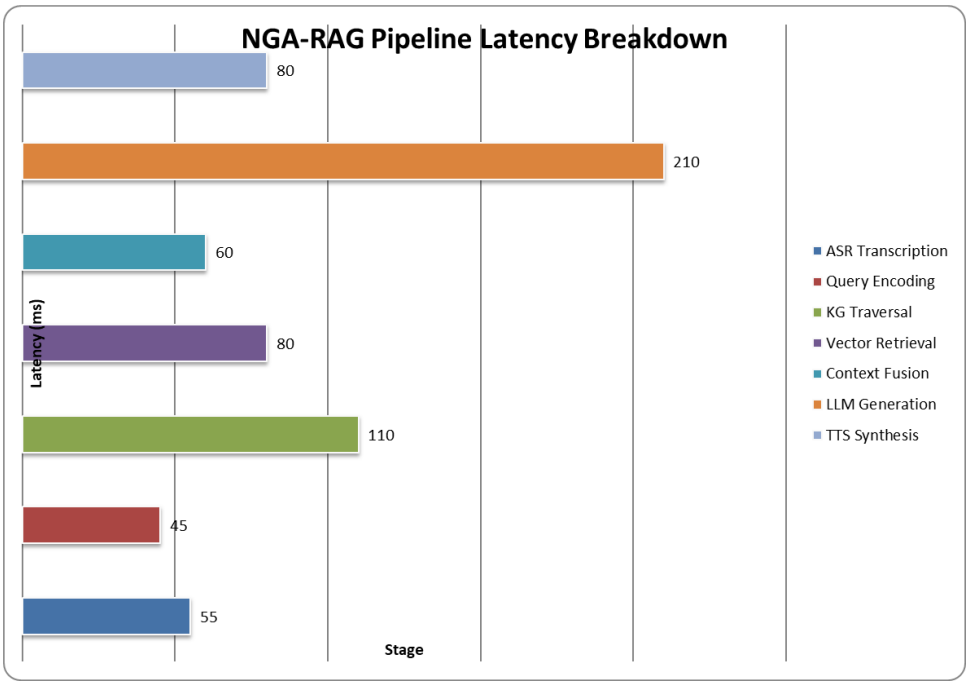
context-dependent answers to seemingly simple questions. A query about coverage for a specific procedure may require evaluating network participation, prior authorization requirements, frequency limits, deductible accumulation, and coordination of benefits rules simultaneously [8]. Static IVR systems cannot model this complexity without exponential growth in call flow logic, while standalone generative models cannot reliably produce responses accurate across all relevant dimensions without grounding in verified sources. The knowledge complexity of healthcare insurance constitutes a fundamental design constraint that shapes every aspect of voice AI architecture.

Callers rarely express requests in standardized language. The same coverage inquiry may be articulated using clinical terminology, colloquial phrasing, insurance jargon, or incomplete information. A member asking about “eye surgery” may intend cataract extraction, LASIK refractive surgery, or glaucoma treatment, each carrying different coverage rules and prior authorization requirements. As such, traditional speech grammars and keyword-based intent classifiers cannot generalize well for the complete range of communication that could take place between a health insurance company and its insureds. Furthermore, pure generative model architectures may misinterpret ambiguous queries, resulting in the creation of confident but incorrect responses [8], [16]. Effective use of voice AI within the healthcare insurance space requires the ability to disambiguate caller intent through knowledge-based reasoning instead of probabilistic inference. Therefore, the coverage response generated by the system must reflect the specific conditions associated with the caller's health insurance benefit plan and their call inquiry.

Voice-related interactions have little tolerance for delay, which can put stress on callers who place calls to the health insurance industry in order to obtain information that they need in an urgent or time-sensitive manner. Consequently, the end-to-end time required to process a call within the healthcare insurance sector must be held below time limits that allow for the flow of natural and human conversational pacing. The retrieval and reasoning processes that take place within standard RAG architectures add additional layers to the overall responsiveness of RAG versus stand-alone generative model architectures [16]. Healthcare insurance systems must meet specific regulatory compliance requirements regarding their members and providers. They are required to produce responses that comply with all applicable policy directives, regulatory guidance, and organizational guidelines, and they must also be able to provide a reconstruction of the reasoning that took place in creating that answer [17], [18]. Within the regulatory compliance requirements and the performance requirements for real-time systems, engineering boundary conditions exist, and all design decisions for RAG-VAI will adhere to these engineering boundaries.

Kalicheti, B. (2026). Neural Graph–Augmented Retrieval for Trustworthy Generative Voice AI in Healthcare Insurance across Providers and Members. *South African Computer Journal* 38(2), 22–33.

ISSN 1015-7999 (print) ISSN 2313-7835 (online)



Graph 2: NGA-RAG Latency Breakdown by Pipeline Stage (ms)

3. Why Pure Generative AI Is Not Enough

Large language models (LLMs) have shown exceptional capability regarding natural language fluency, understanding of context, and the ability to make open-domain inferences. However, these strengths do not translate reliably to the specific requirements of healthcare insurance telephony [13]. LLMs encode knowledge in model parameters learned during pre-training on broad corpora, making it fundamentally difficult to guarantee the accuracy of responses about specific, frequently updated insurance policies. Hallucination the generation of confident, fluent, but factually incorrect statements, is a well-documented failure mode of LLMs [7]. In healthcare insurance contexts, a hallucinated benefit limit, an incorrect deductible figure, or a fabricated prior authorization requirement can result in material harm to the caller and regulatory exposure for the organization.

Beyond accuracy concerns, standalone generative models exhibit response inconsistency across identical queries, a characteristic that is unacceptable in regulated administrative contexts where callers and downstream systems rely on consistent information. The parametric knowledge encoded in LLMs also degrades in relevance over time as plan designs, regulatory requirements, and network configurations change, while model retraining cycles are too slow to keep pace with these updates [1], [14]. Furthermore, LLM inference latency is variable and can spike under load, degrading conversational pacing in ways that erode caller trust. These limitations collectively establish that pure generative AI is not a viable architecture for healthcare insurance telephony, regardless of model capability advances.

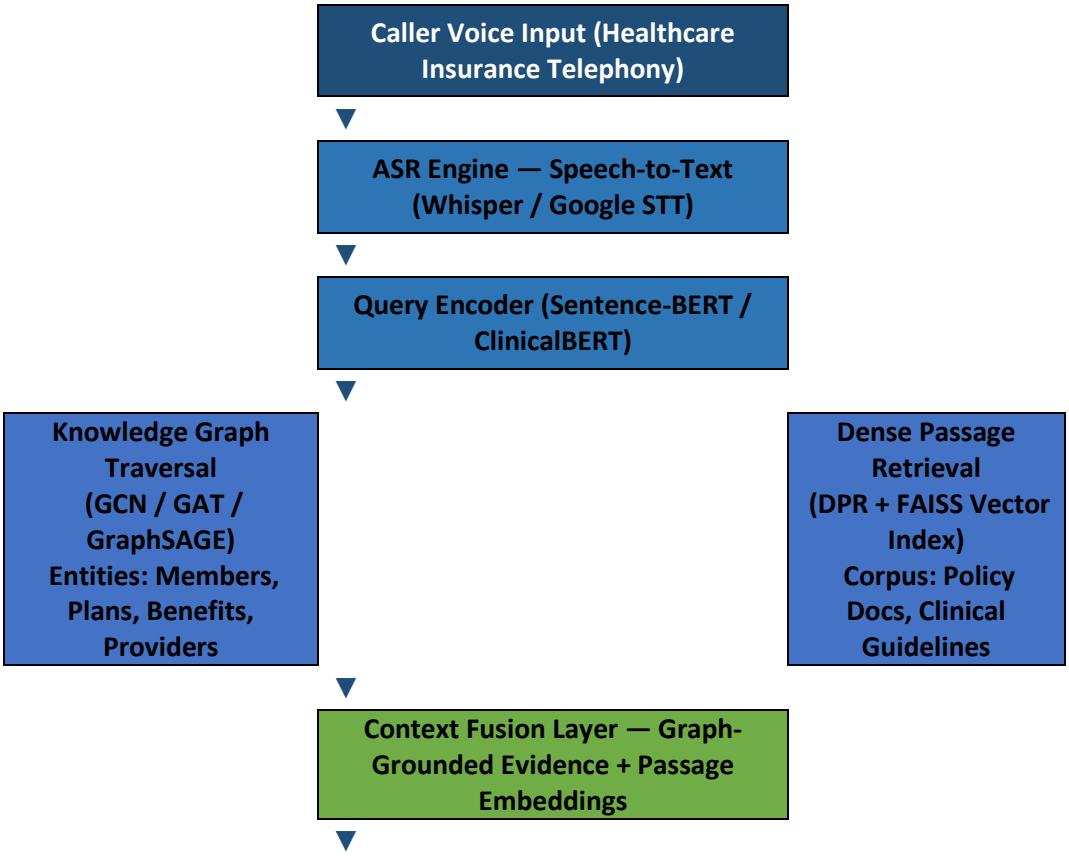
The regulatory dimension of healthcare insurance amplifies these limitations. Responses provided through telephony channels may be cited by callers in appeals, grievances, and regulatory complaints. Systems must support full auditability, enabling reviewers to trace any response to the knowledge sources and reasoning steps that produced it. Generative models, which produce responses through opaque parameter activations without explicit source references, cannot provide this traceability [17], [18]. Compliance obligations in healthcare insurance telephony therefore, require architectural mechanisms that externalize knowledge access, make retrieval sources explicit, and constrain generative behavior to verifiable content requirements that define the foundational rationale for the RAG-VAI architecture presented in this paper.

ISSN 1015-7999 (print) ISSN 2313-7835 (online)

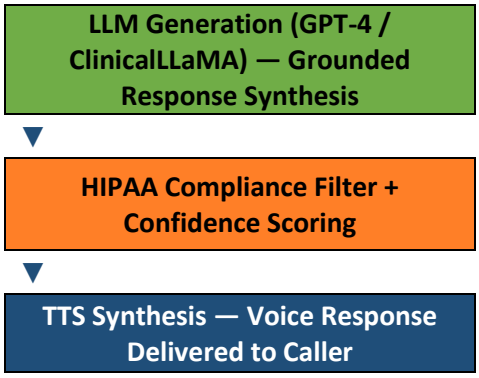
4. RAG as a Control Structure

RAG places a structural divide between how knowledge is accessed and how language is generated, thus addressing the limitations of using an LLM by itself in a regulated healthcare setting [1], [2]. Rather than relying on model parameters to encode domain knowledge, the system retrieves relevant content from curated, versioned knowledge sources at inference time and constrains generation to that context. This architecture transforms the LLM from a knowledge repository into a language synthesis engine: it is responsible for formulating coherent, natural-language responses from provided evidence, not for independently recalling healthcare insurance specifics. The retrieval component ensures that the factual basis of every response is explicitly grounded in verified sources, enabling source attribution that satisfies both operational accuracy requirements and regulatory auditability demands.

RAG offers voice systems numerous advantages beyond just accuracy. Based on the query complexity, conversation context, and how much latency can be tolerated, the rate and depth of retrieval can be dynamically adjusted, giving RAG systems far more performance options than conventional standalone models [3]. Shallow retrieval with template-based responses is enough for basic fact resourcing (such as a deductible or tiers of a network) since the response can be completed before the applicable latency threshold. Substantially higher retrieval depth using LLM-mediated synthesis is required to address complex queries (which often contain multiple criteria). Consequently, the lengthier retrieval will substantially exceed the allowed latency time per response due to the complexity of the interaction [16]. This ability to systematically adapt to the varying levels of latency in telephonic interactions for health insurance will allow RAG to be uniquely suited to healthcare insurance telephonic interactions from simple status inquiries to complicated illustrations of the comparison of benefits.



ISSN 1015-7999 (print) ISSN 2313-7835 (online)



Flow Diagram 1: Neural Graph-Augmented Retrieval (NGA-RAG) Architecture for Healthcare Voice AI.

From a governance perspective, RAG introduces a new trust model for generative AI that takes traditional opaque recollection of parameters and replaces it with auditable access to knowledge. Each response generated by RAG can be traced back to the specific documents retrieved, including chunk identifiers and retrieval scores, which will provide a verifiable audit trail and support for post-hoc audits and dispute resolution [17], [18]. Knowledge stores are versioned so that you can see all the available data when you access it. This allows for retrospective review of call answers in comparison to what was available when you made your call. While RAG optimizes performance, it also provides the foundation for compliance and governance of voice AI technologies in healthcare insurance. Therefore, it is an integral part of the proper implementation of voice AI technologies at scale.

5 - RAG-VAI Architecture

The new RAG-VAI architecture has six independent levels of interdependent scalability based on common observability. The Voice Input Layer is responsible for receiving calls and processing them in real time using ASR with a partial hypothesis that allows for turn-taking detection. The Intent and Context Processing Layer integrates a lightweight NLU classification process with multi-turn conversational memory management to produce structured representations of the query that will be used for retrieving relevant data. The Knowledge Retrieval Layer performs multi-stage searches on vector index and neural knowledge graph databases to produce the most relevant data for each detected intent. The Generative Reasoning Layer synthesizes the retrieved data into coherent responses in an audio format that will meet token limit restrictions. The Voice Response Orchestration Layer will generate text-to-speech audio response(s) for the caller. The Governance and Observability Layer captures decision traces, source attributions, and model behavior metrics in compliance-grade storage [1], [3].

The Knowledge Retrieval Layer is the architectural centerpiece that distinguishes RAG-VAI from conventional conversational AI. Voice queries are first processed by the ASR and NLU components to produce structured intent representations with extracted entities such as procedure codes, benefit categories, and member identifiers. These representations drive a multi-stage retrieval process that combines dense vector search over semantically chunked policy documents with structured graph queries over the neural knowledge graph [3], [21]. Using Dense Passage Retrieval (DPR) in the Vector Retrieval phase, returns document segments for the purpose of identifying broadly relevant segments of text. Using a Graph Query, validates the coverage conditions of the entity dependencies in the retrieved document segments. The conclusion of the two phases will yield an assembled, ranked, and deduplicated retrieved context, which gets injected into the Generative Reasoning Prompt. It is important to note that the constraints imposed on the Generative Reasoning Layer prevent unconstrained generation from overriding the evidence produced by the retrieval phase.

Kalicheti, B. (2026). Neural Graph–Augmented Retrieval for Trustworthy Generative Voice AI in Healthcare Insurance across Providers and Members. *South African Computer Journal* 38(2), 22–33.

ISSN 1015-7999 (print) ISSN 2313-7835 (online)

When the Language Model (LLM) receives the context that was retrieved based on the caller's Intent Representation and the system prompt (structured) that specifically indicates the scope of the response to the supporting evidence, there are constraints placed on each response to verify that the response is complete per the necessary schema validation rules (for example, checking for essential elements such as coverage, limits, and next steps). Orchestration will have visibility into all responses generated by the LLM for the purpose of verifying compliance with the policy before the TTS Voice synthesizes the audio, and the audio is delivered to the caller; therefore, all unverified responses would be prevented from being sent to the caller. This architecture positions the LLM as a retrieval-conditioned synthesizer rather than an autonomous knowledge source, fundamentally changing the failure mode from hallucination to retrieval failure, a mode that is observable, measurable, and addressable through knowledge base maintenance [17], [18].

6. Knowledge Modeling for Healthcare Insurance Domains

Effective RAG performance in healthcare insurance depends critically on the quality of knowledge representation in the retrieval store. Insurance documents, including benefit summaries, coverage policies, clinical criteria, regulatory guidance, and formularies, are segmented into semantically meaningful units aligned with coverage rules, procedural requirements, and benefit descriptions. Chunk boundaries are selected to preserve logical completeness while minimizing noise: a chunk covering a prior authorization criterion should contain all conditions that must be met together, rather than splitting related conditions across multiple retrievable units [21], [22]. Embedding models fine-tuned on healthcare insurance language, including clinical terminology and plan-specific nomenclature, produce vector representations that are more aligned with insurance domain semantics than general-purpose embeddings [22].

Metadata-driven indexing enables hybrid forms of retrieval that combine semantic similarity and deterministic filtering. Each knowledge chunk has been enhanced with structured metadata associated with each knowledge chunk (e.g., type of plan, where the person will be getting service, type of service, what is the condition for eligibility, what tier of the provider network they are in, and the effective date of the service). This metadata enables pre-retrieval filtering that narrows the candidate set to knowledge that is relevant to the caller's plan context before semantic similarity ranking is applied [18]. For example, an eligibility verification query from a commercial plan caller will only retrieve knowledge chunks tagged for the relevant line of business, preventing plan-specific rules from contaminating responses. Metadata filtering also supports compliance use cases: retroactive review can reconstruct the knowledge base state applicable to a given interaction date by filtering on effective period metadata.

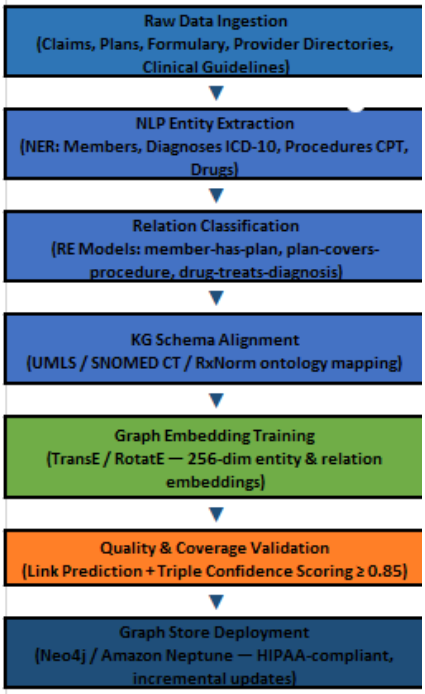
Knowledge currency is a persistent operational challenge in healthcare insurance environments where plan configurations, benefit structures, and regulatory requirements change on quarterly or annual cycles. The knowledge base will keep explicit version history for all knowledge artefacts, allowing for temporal access, which prevents the effect of using out-of-date or future effective policies on response outputs. Update pipelines will provide regular build cycles and automatically embed/re-index all affected knowledge chunks due to ingestion of new information documents, policy changes, and formula changes [17]. Detection of changes at the chunk level will reduce the scope of re-indexing by limiting re-indexing to affected items and reducing the cost associated with maintaining the knowledge base. This versioned knowledge architecture ensures that retrieval remains aligned with current operational reality throughout the insurance plan year, reducing the risk of benefit misinformation caused by stale knowledge.

7. Retrieval Strategies for Real-Time Voice Systems

Multi-stage retrieval balances recall completeness with latency efficiency by decomposing the search process into sequential refinement stages. Stage one applies coarse metadata filtering and

ISSN 1015-7999 (print) ISSN 2313-7835 (online)

approximate nearest neighbor search to identify a candidate set of potentially relevant chunks from the full knowledge store, using efficient indexing structures that support sub-100-millisecond lookup at production scale [3], [20]. Stage two applies semantic similarity ranking over the candidate set using higher-quality, computationally intensive embedding comparisons to reorder candidates by relevance. Stage three applies rule-based refinement that checks candidate chunks against eligibility constraints, effective date validity, and plan-specific applicability filters. By adopting a staged methodology, the number of costly operations performed against the entire knowledge base is lessened, since only validated and contextually appropriate knowledge is supplied to the generative reasoned phase.



Flow Diagram 2: Healthcare Insurance Knowledge Graph Construction and Update Pipeline

Retrieval depth and number of retrieved chunks are dynamically adjusted based upon intent confidence signals provided by the NLU module, such that high-confidence intents requiring well-established information requirements will invoke shallow retrieval with a narrow search scope, thus minimizing latency while meeting users' informational needs in routine transactions like checking deductible status or confirming network tier [20]. Conversely, ambiguous intents with numerous potential meanings will cause deeper retrieval through multiple domains of knowledge, resulting in a richer context that facilitates LLM-driven disambiguation. This adaptive depth strategy prevents over-retrieval, which increases prompt length, LLM processing time, and noise in the reasoning context while ensuring that complex queries receive the knowledge depth they require.

Retrieval latency optimization relies on multiple complementary strategies. Session caching allows for the caching of previously retrieved knowledge representations for each turn as an in-memory representation, enabling an immediate retrieval of any knowledge that has been previously accessed without needing to perform another vector search [16]. Caches located on regional edge sites (next to telephony gateways) contain highly utilized knowledge paths related to the highest volume of inquiries per geographic market, which greatly reduces round-trip latency for the most frequently asked coverage questions. Additionally, the presence of embedding-based in-memory caches for the most frequently retrieved knowledge segments eliminates the need to compute knowledge locations each time knowledge is requested. Parallel execution of ASR final-pass processing and retrieval query formulation allows retrieval to begin concurrently with utterance completion, hiding

ISSN 1015-7999 (print) ISSN 2313-7835 (online)

retrieval latency within the tail of the speech recognition window and contributing meaningfully to the overall latency budget.

8. Neural Graph–Augmented Retrieval for Healthcare Knowledge Reasoning

While vector retrieval over document embeddings identifies semantically relevant content, it is inherently limited in its ability to reason over structured relational knowledge. Explicit dependencies between entities characterize healthcare insurance knowledge: a coverage determination for a procedure depends not only on the procedure code but on its relationship to a benefit category, its inclusion in a specific network tier, its prior authorization requirements, its frequency limits, and its interaction with the member's deductible accumulation state [4], [5], [11], [23]. These multi-hop relational dependencies require traversal of structured knowledge representations rather than similarity matching over flat embeddings. Knowledge graph embedding methods that encode entities and relationships in continuous vector spaces enable neural graph reasoning to be integrated with retrieval pipelines without abandoning the efficiency advantages of dense vector computation [11].

The graph-augmented retrieval architecture introduces a neural knowledge graph alongside the vector index as a complementary knowledge representation layer. The knowledge graph encodes healthcare entities as nodes and coverage relationships as typed, directional edges, with edge types representing relationships such as “covered-under,” “requires-authorization,” “network-tier,” and “subject-to-limit.” Graph construction and maintenance are automated through information extraction pipelines that parse benefit documents and update the graph alongside the vector index [6], [12]. During retrieval, vector candidates from the initial similarity search are enriched through graph traversal: relevant subgraphs are activated based on detected entities in the query, and graph paths connecting the query entities to coverage conditions are retrieved as structured evidence. This hybrid vector-plus-graph retrieval produces a richer, more structured context for the generative reasoning stage, improving both accuracy and explainability [22].

Neural graph augmentation provides three specific capabilities that vector-only RAG cannot deliver. First, contextual disambiguation: when a caller query contains ambiguous terms, graph traversal over entity relationships narrows the interpretation space by evaluating which entities have valid coverage relationships in the caller's plan context, reducing unnecessary clarification turns. Second, coverage validation: before injecting retrieved content into the generative prompt, the system verifies that the retrieved coverage rules are valid for the specific member context by traversing eligibility and plan membership edges, preventing the generation of coverage responses that are factually accurate in the general case but inapplicable to the specific caller [4]. Third, explainability: graph paths traversed during reasoning are logged as structured evidence chains that provide an auditable, human-interpretable record of the relational reasoning underlying each response, directly satisfying regulatory auditability requirements in healthcare insurance telephony environments [17].

9. Governance, Conversation Memory, and Latency Optimization

Healthcare-related conversations tend to occur in incremental segments over multiple query attempts, whereby each successive query is based on the response to the most recent query. A caller who opens with an eligibility inquiry may follow with questions about cost-sharing, prior authorization, and in-network provider options that all depend on the eligibility context established earlier. RAG-VAI systems maintain structured conversational memory, capturing confirmed constraints such as verified plan type and service category retrieved sources, and prior clarifications [16]. This memory informs subsequent retrieval queries, narrowing the search scope to knowledge applicable to the established context and reducing redundant retrieval computation. Conversational memory also enables coherence checking: new queries are evaluated against prior context to detect

Kalicheti, B. (2026). Neural Graph–Augmented Retrieval for Trustworthy Generative Voice AI in Healthcare Insurance across Providers and Members. *South African Computer Journal* 38(2), 22–33.

ISSN 1015-7999 (print) ISSN 2313-7835 (online)

topic shifts or apparent contradictions that may indicate caller confusion or system misunderstanding.

Latency reduction in RAG-VAI systems requires parallel execution of all independent processing stages. Intent extraction, entity detection, retrieval query formulation, and metadata filter construction execute concurrently rather than sequentially during the intent processing stage [3], [20]. Hierarchical caching reduces retrieval latency across three levels: session-level in-memory caches for multi-turn coherence, regional edge caches for high-frequency knowledge paths, and in-memory embedding caches for the most commonly retrieved document segments. Incremental and streaming generation, beginning text-to-speech synthesis as partial responses become available rather than waiting for complete response generation, reduces the caller-perceived silence period by pipelining synthesis with LLM output. These combined optimizations allow RAG-VAI systems to satisfy voice telephony latency constraints despite the additional computational stages that retrieval introduces relative to standalone generative architectures.

Governance in RAG-VAI systems is architecturally embedded rather than procedurally imposed. Source attribution logging captures the specific retrieved chunks and graph paths that informed each response, creating immutable evidence chains stored in compliance-grade infrastructure [17], [18]. Versioned knowledge access ensures that the retrieval system state at the time of any interaction can be reconstructed for retrospective audit, enabling organizations to demonstrate that responses were consistent with the knowledge available at the time of the call. Deterministic response templates for high-confidence, rule-governed interactions reduce LLM involvement in compliance-critical responses, further strengthening auditability. Continuous monitoring for retrieval quality degradation, response schema violations, and anomalous knowledge access patterns enables early detection of governance failures before they manifest as caller-visible errors [19].

10. Future Directions

Emerging research directions for RAG-VAI in healthcare insurance include adaptive retrieval learning, where the retrieval strategy is updated based on observed response quality signals such as caller confirmation, clarification requests, and agent escalation rates to progressively improve retrieval precision for the specific knowledge patterns of a given deployment [1], [14]. Agentic orchestration of insurance workflows, where the RAG-VAI system not only retrieves and synthesizes information but actively executes multi-step administrative tasks using retrieved knowledge to guide each action, represents a significant expansion of the capability frontier. Federated knowledge architectures that enable retrieval across distributed, organization-specific knowledge stores without centralizing sensitive plan data may address data residency requirements that currently constrain multi-cloud healthcare insurance deployments.

Multilingual RAG-VAI systems capable of retrieving, reasoning, and generating responses across diverse language communities without duplicating knowledge engineering or governance infrastructure will expand equitable access to intelligent healthcare telephony services [8]. Formal verification of retrieval and reasoning pipelines through methods such as retrieval correctness certification and coverage validity assertion testing may enable stronger governance assurances than current monitoring-based approaches [18]. Standardized benchmarks for RAG-VAI evaluation in regulated domains, covering retrieval precision under noisy voice input, grounded response accuracy, containment appropriateness, and latency percentiles, will support systematic progress measurement and enable cross-system comparisons that advance the state of practice in trustworthy healthcare voice AI.

11. Conclusion

This paper has presented a neural graph-augmented Retrieval-Augmented Generation architecture for trustworthy generative voice AI in healthcare insurance telephony. The architecture addresses

Kalicheti, B. (2026). Neural Graph–Augmented Retrieval for Trustworthy Generative Voice AI in Healthcare Insurance across Providers and Members. *South African Computer Journal* 38(2), 22–33.

ISSN 1015-7999 (print) ISSN 2313-7835 (online)

the fundamental limitations of standalone LLMs and vector-only RAG through a hybrid knowledge layer that combines semantic retrieval with structured graph reasoning over healthcare domain knowledge. By grounding generative responses in verified, source-attributed evidence chains and validating coverage relationships through neural graph traversal before LLM prompting, the proposed system delivers a level of factual accuracy, explainability, and regulatory traceability that is not achievable with conventional generative AI approaches [1], [17]. The neural graph augmentation component specifically advances the state of RAG architecture for healthcare domains by enabling the relational reasoning that insurance knowledge requires.

As healthcare organizations advance toward AI-driven telephony, the quality of the knowledge architecture underlying conversational systems will increasingly determine the difference between responsible deployment and regulatory liability. The architectural principles outlined in this paper provide a rigorous design foundation for RAG-VAI systems that balance the conversational capability of generative AI with the accuracy and governance requirements of regulated healthcare insurance operations [9], [10]. Future advances in knowledge graph construction, retrieval efficiency, and federated knowledge management will further mature this architecture, progressively expanding the scope of healthcare insurance interactions that can be handled with verifiable, trustworthy conversational AI at a national scale.

REFERENCES

1. Patrick Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2020. [Online]. Available: https://arxiv.org/abs/2005.11401
2. Kelvin Guu et al., "REALM: Retrieval-augmented language model pre-training," arXiv:2002.08909 [cs.CL], 2020. [Online]. Available: https://arxiv.org/abs/2002.08909
3. V. Karpukhin et al., "Dense passage retrieval for open-domain question answering," in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 6769–6781. [Online]. Available: https://arxiv.org/abs/2004.04906
4. Vladimir Karpukhin et al., "Semi-supervised classification with graph convolutional networks," arXiv:2004.04906 [cs.CL], 2017. [Online]. Available: https://arxiv.org/abs/2004.04906
5. Petar Veličković et al., "Graph attention networks," arXiv:1710.10903 [stat.ML], 2018. [Online]. Available: https://arxiv.org/abs/1710.10903
6. William L. Hamilton et al., "Inductive representation learning on large graphs," arXiv:1706.02216 [cs.SI], 2017. [Online]. Available: https://arxiv.org/abs/1706.02216
7. Karan Singhal et al., "Large language models encode clinical knowledge," Nature, 2023. [Online]. Available: https://doi.org/10.1038/s41586-023-06291-2
8. Liliana Laranjo et al., "Conversational agents in healthcare: A systematic review," JAMIA, 2018. [Online]. Available: https://doi.org/10.1093/jamia/ocy072
9. E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," Nature Medicine, 2019. [Online]. Available: https://doi.org/10.1038/s41591-018-0300-7
10. Pranav Rajpurkar et al., "AI in health and medicine," Nature Medicine, 2022. [Online]. Available: https://doi.org/10.1038/s41591-021-01614-0
11. Antoine Bordes et al., "Translating embeddings for modeling multi-relational data," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2013. [Online]. Available: https://papers.nips.cc/paper_files/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html
12. Zhiqing Sun et al., "RotatE: Knowledge graph embedding by relational rotation in complex space," in Proc. International Conference on Learning Representations (ICLR), 2019. [Online]. Available: https://arxiv.org/abs/1902.10197

Kalicheti, B. (2026). Neural Graph–Augmented Retrieval for Trustworthy Generative Voice AI in Healthcare Insurance across Providers and Members. *South African Computer Journal* 38(2), 22–33.

ISSN 1015-7999 (print) ISSN 2313-7835 (online)

13. Tom B. Brown et al., "Language models are few-shot learners," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2020. [Online]. Available: https://arxiv.org/abs/2005.14165
14. Rishi Bommasani et al., "On the opportunities and risks of foundation models," arXiv preprint arXiv:2108.07258, 2021. [Online]. Available: https://arxiv.org/abs/2108.07258
15. Jacob Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv:1810.04805 [cs.CL], 2019. [Online]. Available: https://arxiv.org/abs/1810.04805
16. Steve Young et al., "Pomdp-based statistical spoken dialogue systems: A review," Proceedings of the IEEE, 2013. [Online]. Available: https://doi.org/10.1109/JPROC.2012.2225812
17. Danton S. Char et al., "Implementing machine learning in health care addressing ethical challenges," New England Journal of Medicine, 2018. [Online]. Available: https://doi.org/10.1056/NEJMp1714229
18. W. Nicholson Price II, I. Glenn Cohen, "Privacy in the age of medical big data," Nature Medicine, 2019. [Online]. Available: https://www.nature.com/articles/s41591-018-0272-7
19. Ziad Obermeyer et al., "Dissecting racial bias in an algorithm used to manage the health of populations," Science, 2019. [Online]. Available: https://doi.org/10.1126/science.aax2342
20. Jeff Johnson et al., "Billion-scale similarity search with GPUs," IEEE Transactions on Big Data, 2021. [Online]. Available: https://doi.org/10.1109/TBDATA.2019.2921572
21. Nils Reimers, Iryna Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," arXiv:1908.10084 [cs.CL], 2019. [Online]. Available: https://arxiv.org/abs/1908.10084
22. Emily Alsentzer et al., "Publicly available clinical BERT embeddings," arXiv:1904.03323 [cs.CL], 2019. [Online]. Available: https://arxiv.org/abs/1904.03323
23. Maximilian Nickel et al., "A review of relational machine learning for knowledge graphs," Proceedings of the IEEE, 2016. [Online]. Available: https://doi.org/10.1109/JPROC.2015.2483592

Kalicheti, B. (2026). Neural Graph–Augmented Retrieval for Trustworthy Generative Voice AI in Healthcare Insurance across Providers and Members. *South African Computer Journal* 38(2), 22–33.

ISSN 1015-7999 (print) ISSN 2313-7835 (online)