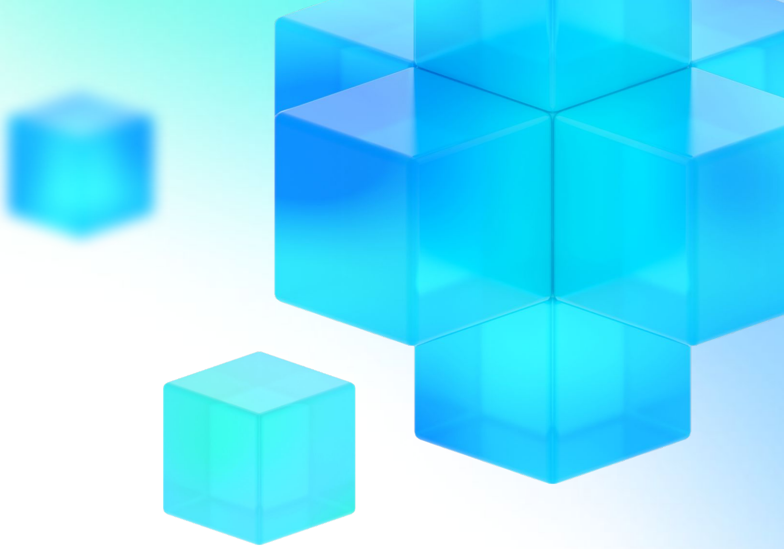


Using AI to establish and maintain trust

A technical exploration of how Infinitus' AI architecture is designed to achieve patient, payor, and provider trust

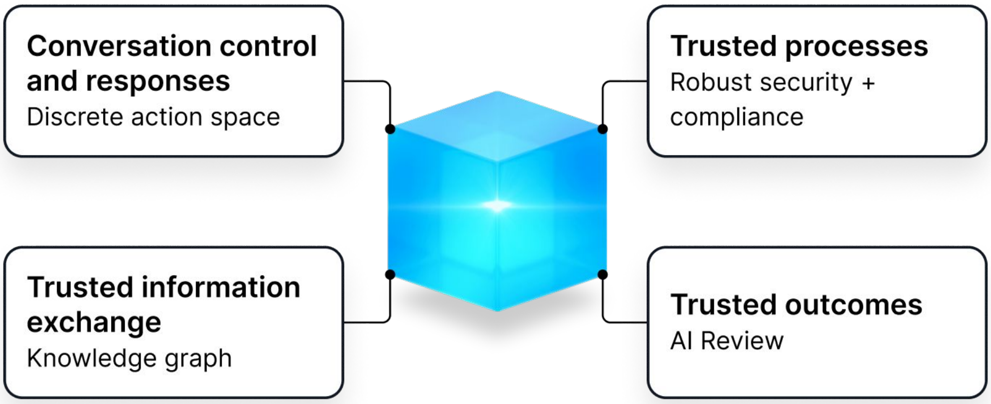
By Shyam Rajagopalan
CTO & co-founder, Infinitus

Infinitus

Introduction: The Infinitus AI system

As AI systems become increasingly central to healthcare communication, achieving both high accuracy and security remains a significant challenge, particularly when handling sensitive data such as patient information and support cases. At the same time, building and maintaining trust is critical; an AI platform that checks every box isn't worth much if it's not trusted by the entire healthcare ecosystem.

Infinitus Systems addresses these challenges by deploying a uniquely structured AI platform that integrates four components specifically built to foster trust among patients, providers, and payors:



Conversation control and responses
Discrete action space

Trusted processes
Robust security + compliance

Trusted information exchange
Knowledge graph

Trusted outcomes
AI Review

Discrete action space (trusted conversation control and responses): Limits AI responses to a set list, ensuring accuracy and compliance, preventing errors like incorrect patient data.

Knowledge graph (trusted information exchange): Learns from millions of calls, verifying info, understanding complex plans, resulting in higher accuracy than human agents.

AI review (trusted outcomes): Automated system checks calls for accuracy and standards, with "human-in-the-loop" for complex cases, ensuring data reliability.

Security and compliance program (trusted processes): Adheres to SOC2 and HIPAA, and has an internal GRC process that includes bias testing, PHI redaction, and data retention controls for secure, unbiased data handling.

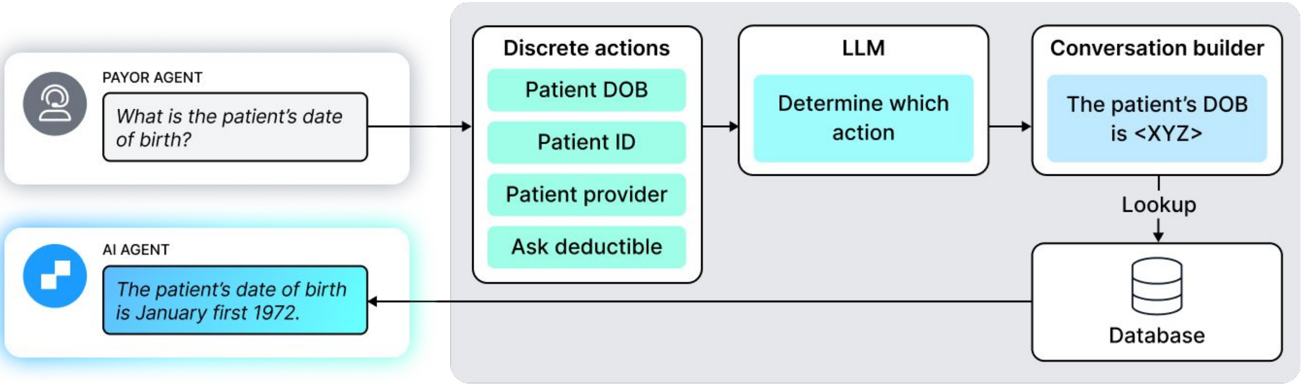
Discrete action space: Trusted conversation control and hallucination free

While large language models (LLMs) bring powerful understanding and conversational capabilities, they may occasionally introduce unwanted variability, especially in healthcare scenarios requiring exact language. Infinitus has designed its system to overcome this by implementing a discrete action space – a limited set of actions the AI can take or speak informed by our knowledge graph and customer SOPs while still using the latest LLMs to understand what is being said by the other party. This approach enables several critical benefits over traditional LLMs.

Precision in key details

A traditional LLM might inadvertently generate incorrect information in a high-stakes conversation, such as an inaccurate birthdate or member ID, especially in longer, complex conversations. By separating the action space from language generation, Infinitus ensures that critical details like patient information are treated deterministically, preventing costly mistakes.

For example, the AI might choose to confirm the patient's date of birth during a conversation but instead of creating a response, the LLM is forced to use a fixed rule-based response where it cannot generate a response (and thus, cannot hallucinate). This approach helps avoid costly mistakes especially and ensures sensitive information is always handled safely and appropriately When a human agent requests a patient's date of birth (DOB), the system ensures the question aligns with predefined prompts that the LLM is authorized to handle. The input is first validated against a set of rules, then passed to the LLM. The LLM generates a response without including any personally identifiable information (PII). Instead, it uses a placeholder for the DOB, which is later securely retrieved from the database and presented to the agent with the actual value.



Compliance-focused language

Many customers need the AI to follow specific, compliant language for actions like greeting a payor or asking for benefits. Infinitus’ discrete action space enables the platform to use compliant language and to ensure consistent adherence to compliance requirements. This capability is especially valuable for organizations where regulatory compliance or internal protocol mandates exact phrasing, offering customers a solution that is both reliable and adaptable.

Dependency management for streamlined interactions

In addition to a discrete action space, Infinitus manages dependencies within conversations to ensure optimal flow, skipping irrelevant questions when certain criteria are met. If prior authorization, for instance, is unnecessary for a medication, the system skips related questions like what dates the prior authorization is approved for, saving time and enhancing efficiency.

This ensures conversations remain focused and prevents unnecessary, repetitive exchanges, ultimately increasing productivity and reducing call duration. This also makes Infinitus calls much quicker for payors and others receiving Infinitus calls. In internal testing, we’ve found dependency management in longer conversations is very hard for a traditional LLM to manage with a purely prompt-based system.

Comparison to traditional LLMs

Traditional LLMs, though capable of impressive language understanding and generation, present certain limitations in settings requiring structured accuracy and compliance. Without a controlled action space, an LLM might deviate from the intended flow, potentially overlooking critical details or introducing inaccurate information. Infinitus’ system leverages LLMs for natural language understanding only within specific contexts and goals, integrating them into a graph-based, deterministic conversational builder that ensures strict adherence to a structured conversational path. This blend of structured determinism and LLM-based understanding harnesses the strengths of both approaches, creating a system that balances flexibility with accuracy.

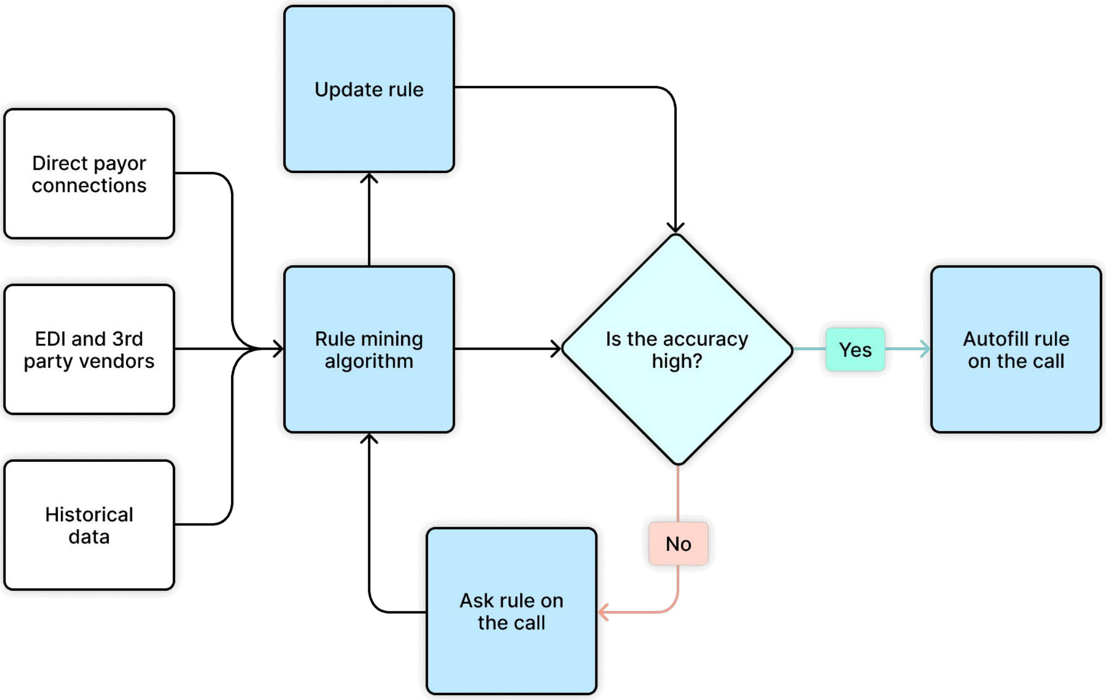
Knowledge graph: Trusted information exchange

Infinitus leverages a specialized knowledge graph for trusted information exchange, enabling its AI agent to actively verify information, understand complex payor plans, and integrate diverse data sources, delivering outputs validated by customers as consistently more accurate (> 10%) than human agents.

The knowledge graph in action: Powering accurate and validated AI

The Infinitus knowledge graph actively learns by applying rule-mining algorithms to data from over three million historical calls, deriving precise, verified statements about payor rules and plan specifics. This dynamic system runs daily, adding validated rules (meeting >95% accuracy) and removing outdated ones, ensuring real-time intelligence.

This mined knowledge empowers the AI agent to operate smoothly and even push back on incorrect payor agent information. Furthermore, Infinitus ensures quality through continuous monitoring and daily cross-verification by human reviewers, leading to customer-validated results consistently reported as approximately 10% more accurate on average than human agents. This demonstrated reliability builds significant trust and powers all Infinitus AI offerings.



The need for specialized healthcare knowledge

The need for such a robust knowledge base stems from the immense complexity of the US healthcare system. Infinitus AI agents require deep, accurate understanding to navigate thousands of unique insurers, including variations even within large groups like Blue Cross Blue Shield, and a multitude of plan designs (commercial, Medicare, ACA, supplemental, etc.). Since plan details dictate critical patient costs like co-pays and deductibles, precision is paramount for providers and patients seeking to avoid unexpected bills. The knowledge graph enables the AI agent to correctly interpret this complex landscape, categorize policies accurately, and handle interactions effectively based on verified information.

Overcoming technical and data challenges

Acquiring this essential benefit and authorization data faces significant technical hurdles, primarily because this information isn't centrally available and often exists only in non-digital formats like PDFs or spreadsheets. The sheer volume of plans (e.g., over 43 Medicare options per beneficiary in recent years) and variability based on factors like state regulations add layers of difficulty.

Furthermore, healthcare rules are constantly evolving due to legislation like the ACA and IRA, demanding continuous updates. The Infinitus knowledge graph is designed to overcome these issues with constantly refreshed data, managing updates and maintaining accuracy far more effectively than manual processes.

AI review: Trusted outcomes

Infinitus tackles the challenge of ensuring data accuracy from its AI-driven healthcare calls using an automated AI review system. This system acts as a quality "judge," meticulously evaluating every call for consistency and adherence to standards, overcoming the scalability and subjectivity limitations of manual review to build customer trust through reliable data delivery. Customers find Infinitus AI Agents consistently > 10% more accurate than human agents.

The need for AI review

Infinitus' AI voice agents navigate complex, lengthy calls with healthcare payors, patients, and providers. During payor calls, for example, they can gather over 200 data points regarding patient benefits and prior authorizations.

Human in the loop and human collaboration

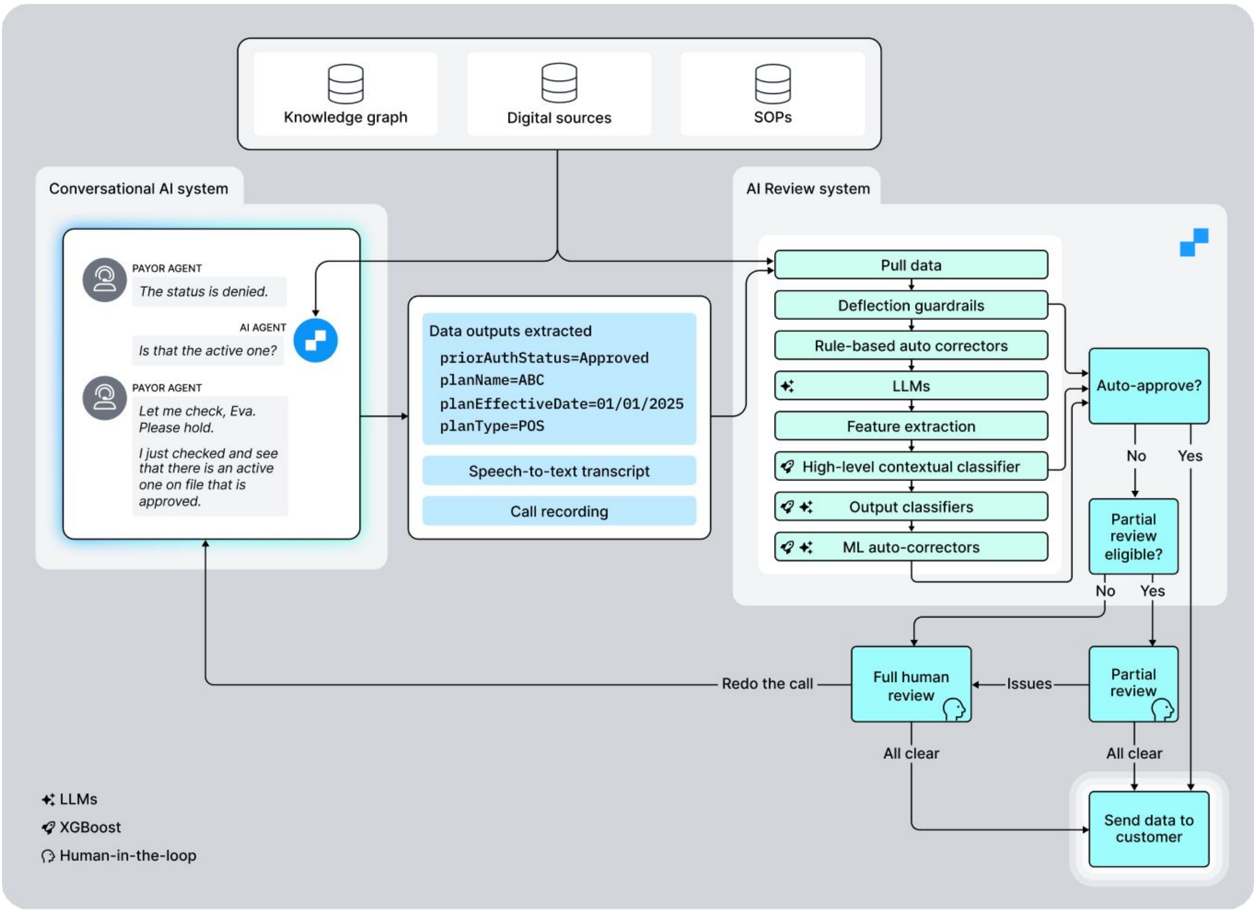
While LLMs are utilized, Infinitus determined that relying solely on them for review resulted in lower accuracy due to potential inconsistencies, hallucinations, and difficulties in maintaining brittle, domain-specific prompts. Instead, the outputs from LLMs are used as powerful features within statistical XGBoost classifiers. These classifiers excel at learning complex patterns from historical data, including real-world human corrections, leading to more reliable, scalable, and explainable quality predictions.

The system firmly embraces a human-in-the-loop philosophy where AI handles the volume, automatically approving calls or flagging specific issues for focused human review. This collaboration optimizes human expertise for complex cases, ensures unbiased evaluation, and provides feedback for continuous AI improvement, augmenting rather than replacing human intelligence.

The accuracy of this data is paramount, as inconsistencies can lead to repeated calls, delays in patient care, and significant back-office inefficiencies. Traditional manual review by human experts, while capable of nuanced understanding, suffers from inconsistency, fatigue, subjectivity, and simply cannot scale to the sheer volume of calls. To overcome these limitations, Infinitus developed an AI review system. This automated process acts as a proactive quality check, evaluating every call for contradictions, adherence to customer SOPs, and data integrity before any information is sent to the customer, flagging calls that need correction or human review.

A collaboration of many models

The Infinitus AI Review system employs a sophisticated multi-layered pipeline that integrates various models and data sources. It analyzes call recordings, text transcripts, internal knowledge graphs, digital sources, and specific standard operating procedures (SOPs). Early layers use rules to deflect overly complex scenarios to human experts and perform basic auto-corrections like formatting. Large Language Models (LLMs) then process call audio and text to extract structured information and reasoning, which serve as inputs (features) for downstream machine learning classifiers. Subsequent layers utilize statistical models, specifically XGBoost classifiers, to assess overall call quality and evaluate the likelihood of required corrections for individual data outputs based on historical human reviewer actions, leading to a final aggregated decision.



The status is denied.

Is that the active one?

Let me check, Eva. Please hold.

I just checked and see that there is an active one on file that is approved.

Data outputs extracted

```
priorAuthStatus=Approved
planName=ABC
planEffectiveDate=01/01/2025
planType=POS
```

Speech-to-text transcript

Call recording

Security, privacy, and compliance: Trusted processes

Bias testing methodology and 99% accuracy

Infinitus sets a high standard for accuracy across diverse demographic and therapeutic areas, achieving over 99% accuracy across all sectors, from geography, age, and economic indicators to various healthcare specialties.
This ensures Infinitus AI agents remain unbiased and equally effective for all patient groups.

Infinitus' bias testing methodology compares AI-collected data with human-reviewers, monitoring for any variances that might hint at demographic bias. When any category falls below the 99% accuracy threshold, it is flagged and quickly addressed by the machine learning team.

PHI and machine learning

In addition to its focus on accuracy and bias, Infinitus prioritizes the security of healthcare information with SOC2 and HIPAA compliance as foundational principles. The platform employs robust data management policies, including automated and human-verified redaction of personally identifiable information (PHI). PHI is not stored in training datasets, and data retention policies ensure that PHI is deleted within a fixed period as per customer agreements, guaranteeing data is handled with the highest standards of care. Infinitus' full list of security principles is available at https://www.infinitus.ai/security/.

Governance and risk

The Infinitus Governance, Risk, and Compliance (GRC) council meets regularly to ensure that security, privacy, and ethical standards remain robust and up-to-date. This council includes cross-functional leadership from security, machine learning, and engineering, ensuring that security considerations are prioritized at every stage of system development. Furthermore, Infinitus has a formal vendor management process that ensures all third-party vendors adhere to the same standards of security and compliance.

Bringing it all together: A secure, reliable platform for healthcare

Infinitus has created a uniquely structured conversational AI platform that integrates accuracy and security safeguards essential to healthcare. By combining a highly accurate, bias-monitored AI system with a controlled, action-based framework and SOC2- and HIPAA-compliant security standards, Infinitus delivers a dependable and ethical solution for healthcare communication.

Through a blend of cutting-edge AI techniques and robust operational guardrails, Infinitus demonstrates a commitment to delivering conversational AI that is both powerful and secure, meeting the exacting needs of the healthcare industry and establishing new standards in conversational AI integrity and reliability.

To learn more about Infinitus and our voice AI agents that are built for use cases where trust is critical, visit **infinitus.ai** or **contact us directly**.

Infinitus

About Infinitus

Infinitus is the leader in trusted AI agents and co-pilots for healthcare that improve the patient, clinician, and staff experience. Infinitus' AI voice agents and copilots help automate, expedite, and simplify time-consuming but critical conversational touchpoints that drive patient, clinician, and staff frustration and dissatisfaction. Infinitus' proven AI native architecture combines multimodal, multi-model AI with a robust knowledge graph, a proprietary discrete action space, and human-in-the-loop guardrails to ensure data accuracy, safety, and support for healthcare organizations at scale. Infinitus was recently named to Fast Company's World's Most Innovative Companies of 2025 and is a trusted solutions partner to 44% of Fortune 50 healthcare companies. Learn more at **https://www.infinitus.ai/**

HIPAA COMPLIANT

AICPA SOC aicpa.org/soc4so SOC for Service Organizations | Service Organizations

Infinitus