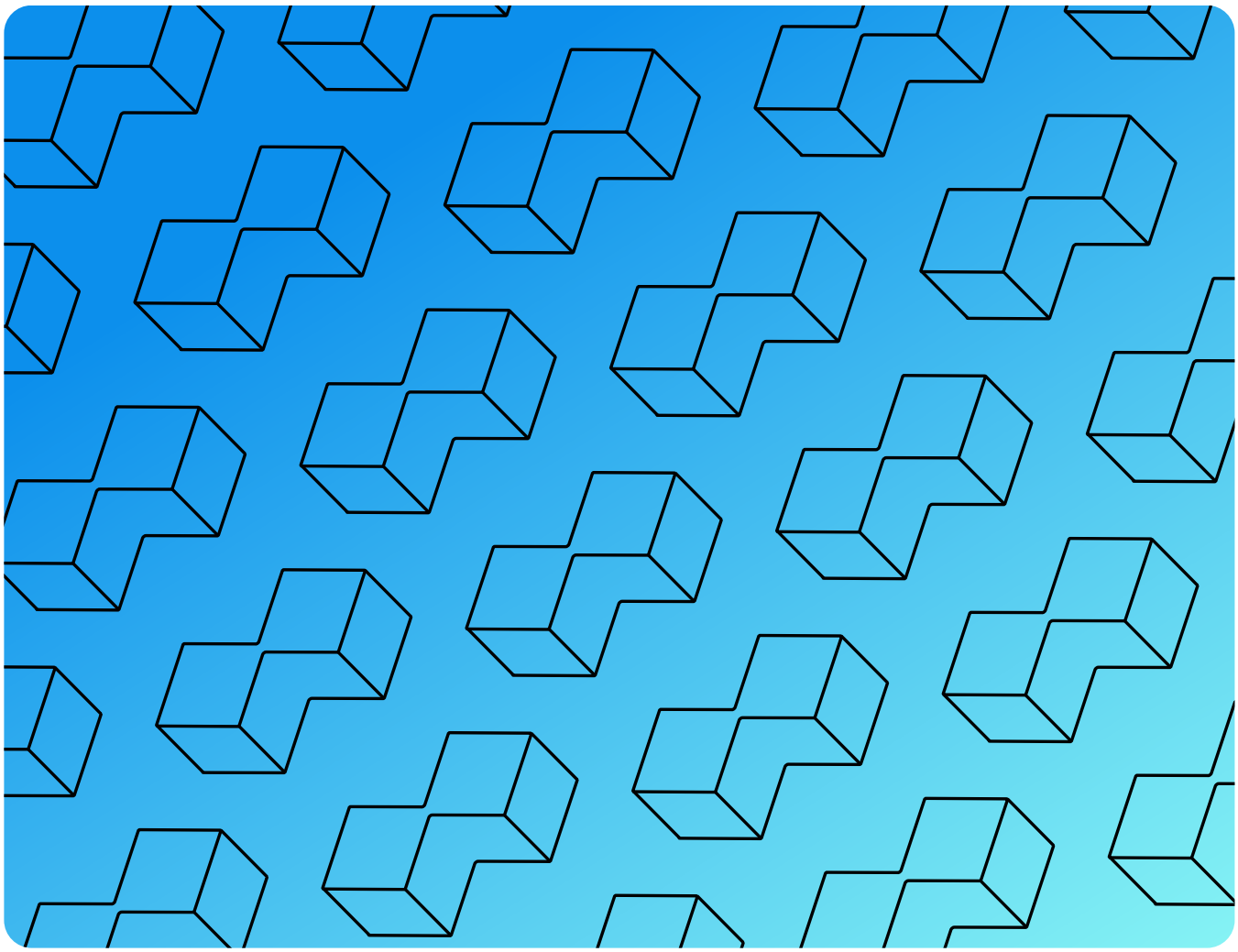
Infinitus

The Infinitus AI machine

How we use humans in concert with our multi-model, multimodal AI system



Executive summary

Automating administrative phone calls in healthcare is no easy feat. The benefit verifications and prior authorization inquiries Infinitus automates can be an hour long, with hundreds of back-and-forth interactions that require our AI to keep track of context. And because these calls are made to humans, our AI agent must respond almost instantaneously, in order to avoid confusing or frustrating the person on the other end of the line.

Over the last five years, Infinitus has automated over 3 million such calls, giving back nearly 75 million minutes of healthcare back office time while delivering significant savings. These phone calls have been made on behalf of over 80,000 providers across the US, acquiring data that is at least 10% more accurate than equivalent human-made calls. Leveraging the Infinitus platform has helped providers avoid wrongly denied insurance claims and other delays in medication access.

Today, we're excited to unveil the details of many of these advancements. We hope this additional level of transparency shines some light on the complexity of the problem, the ingenuity of our methods, and why we are confident that we can dramatically improve the status quo.

Contents

01 A multi-model, multimodal AI system

02 Humans operating in concert with AI

03 Delivering on the quality promise

04 Creating time for healthcare

01

A multi-model, multimodal AI system

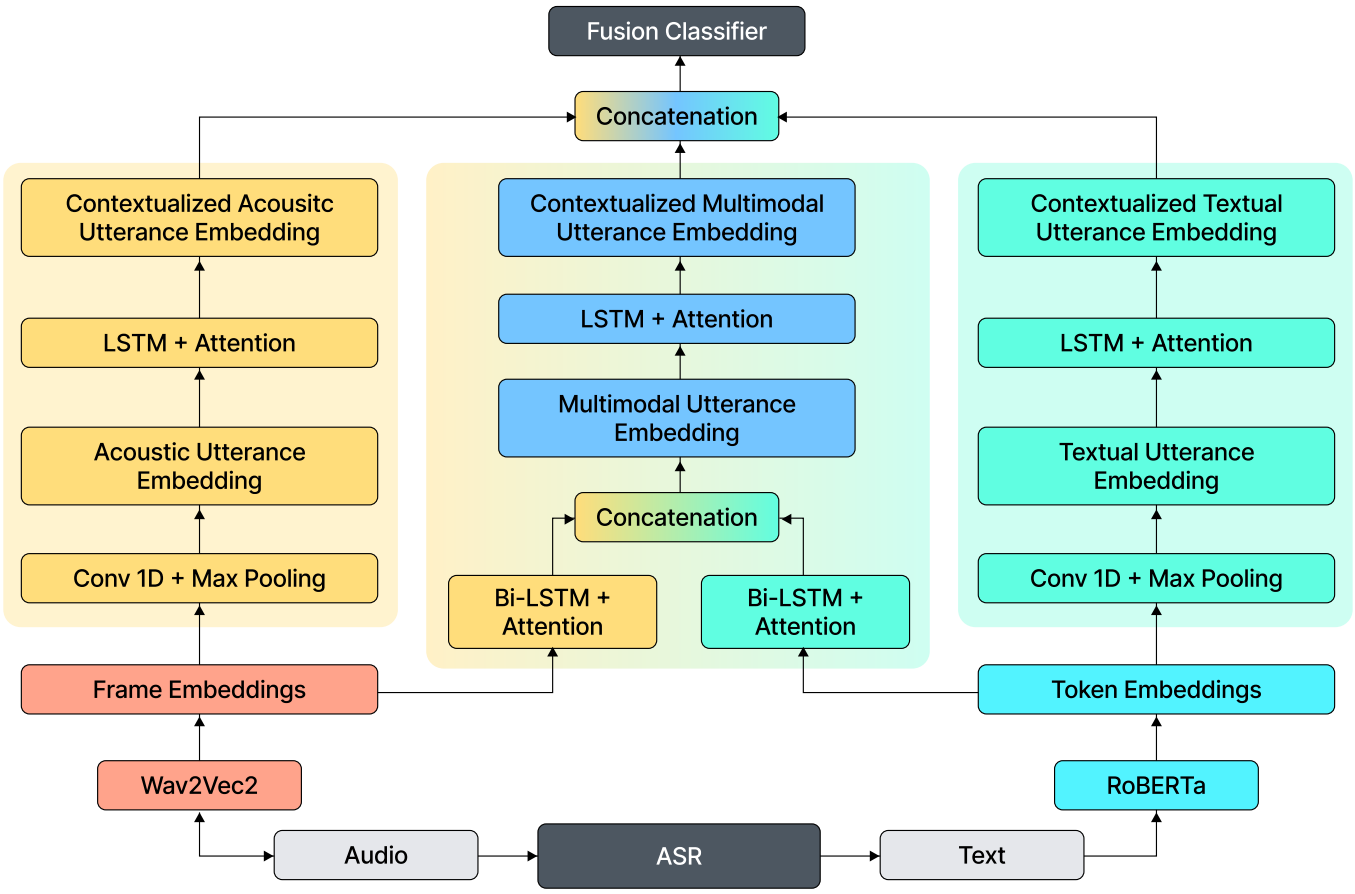
When we started Infinitus in 2019 to automate complicated healthcare back-office phone calls that often are an hour or more long, with hundreds of back-and-forth interactions, many thought we were crazy for attempting to take on this science problem in one of the hardest industries to sell into. Cutting-edge natural language processing systems didn't have the context windows nor the knowledge base to deal with such domain-specific jargon, complexity and SOPs.

Five years in, our multi-model, multimodal platform is increasingly accomplishing harder and harder tasks. The longest call we processed in the last quarter was 3 hours and 15 minutes long and contained 306 conversational turns. Our platform employs over 100 models, from highly-performant shallow models to custom fine-tuned and in-house audio and text models to large-language models (LLMs), which enable us to optimize the user experience during each phase of a call. The parameters of our models range from dozens in shallow models to minimize latency, 100 million for our larger in-house models, to as much as 10 trillion in the GenAI models that we employ.

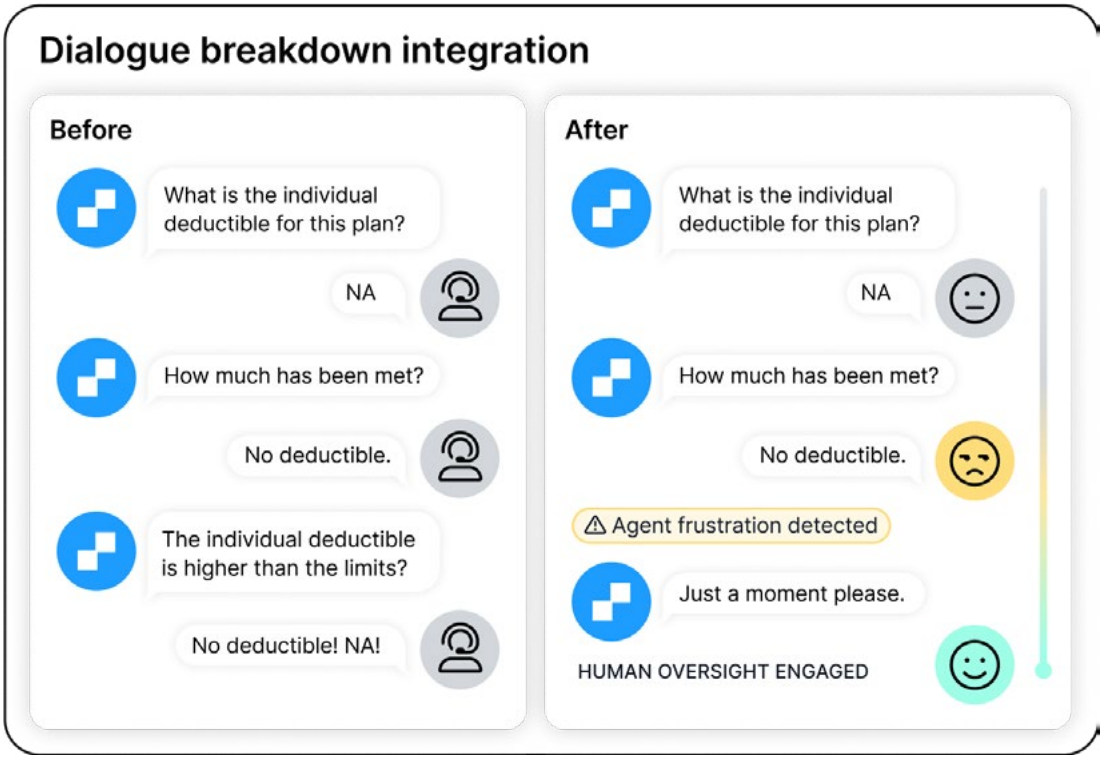
As we've delivered value to over a dozen Fortune 500 companies, we have learned that doing so while having the right guardrails to be secure and compliant isn't going to be delivered with a simple wrapper around the latest transformer (GPT, Gemini, etc.) models.

One example of this is our real-time multimodal dialogue breakdown model. Many points of conversational AI breakdown can only be detected in audio-only signals or at the intersection of audio and text signals so using a text-only model will yield bad recipient experience and lower data quality.

Multimodal architecture to solve dialogue breakdowns:



An example set of conversational turns with and without the multimodal dialogue breakdown model:



This has been made possible by our world-class team of AI scientists, researchers, and engineers. Here are some of the papers published by our research group:

- **2024:** Multimodal Contextual Dialogue Breakdown Detection for Conversational AI Models
- **2024:** Graph Integrated Language Transformers for Next Action Prediction in Complex Phone Calls
- **2023:** Leveraging Explicit Procedural Instructions for Data-Efficient Action Prediction

02

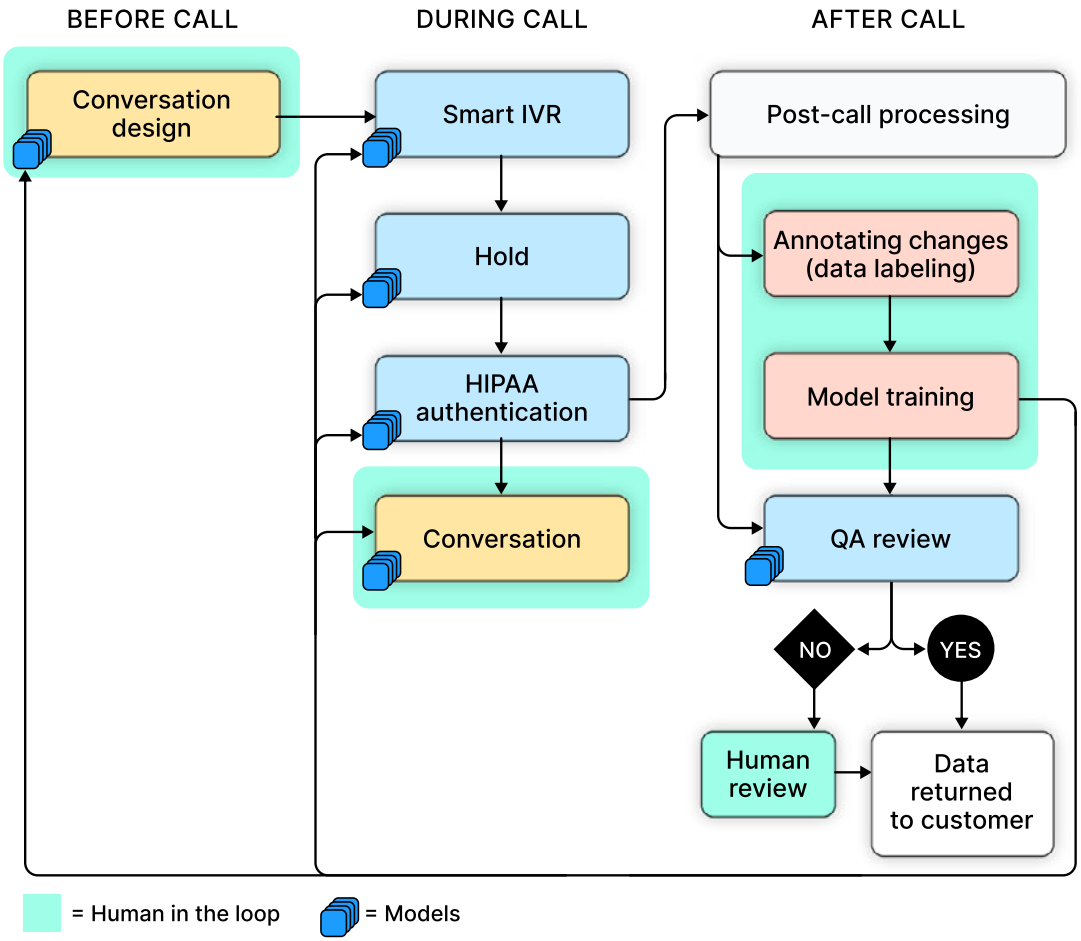
Humans operating in concert with AI

Like most modern AI systems, we also learn a lot from human feedback. And our corpus of human feedback in healthcare back-office calls is gargantuan: it spans over 300 million utterances that have fed into our foundational platform that now automates the vast majority of time spent on tasks.

Infinitus employs human-in-the-loop machine learning before, during, and after each call. Not only does this enable us to deliver higher accuracy than equivalent human callers, but it allows us to continuously improve the abilities and autonomous abilities of our system.

Our goal is always to deliver a seamless experience to those receiving calls from our system while delivering an unparalleled level of quality for those using our services. Humans play a number of roles in accomplishing this, whether it is our conversational designers who configure conversations while employing best practices and SOPs from our experience, our AI training team that can be routed in during “break-glass” moments of a call to help our AI agent navigate through an edge case that hasn’t been encountered before, or our quality assurance teams who can listen to call fragments and correct outputs if our system deems that is the necessary path.

As we have built this team, we have represented all of the customer personas (providers, payors, pharmacies, pharma) we serve and hope to serve. While improving our AI systems is the primary responsibility of this team, providing a **trusted set of guardrails** to novel AI systems is something our customers deeply value.



03

Delivering on the quality promise

In an internal study, we performed two duplicative benefit verification tasks for each patient for a set of patients across a representative sampling of major medical and PBM payors over a four-day period. The goal was to see the consistency of results that we received across the duplicative calls. We found that 30% of these parallel calls had at least one critical data element that had differing values reported to us by payor agents.

To counter this, our AI system draws from a robust in-house knowledge graph that has up-to-date intelligence on payor rules and guidelines that we have gathered from our millions of calls and access to EDI, payor APIs, and policy documents. The impact of this part of our platform is valued by our customers because it reduces downstream denials, but also something that our payor partners appreciate because it provides a level of real-time QA which reduces correction call-backs.

Here is a **video with two examples** of what our knowledge graph enables for our system.

04

Creating time for healthcare

Our vision at Infinitus has always been to make healthcare “wait-less”. We believe in the “infinite” possibilities of AI to solve complex problems and help us fulfill our mission of making time for healthcare. Our system has enabled us to create several standardized call flows across dozens of therapeutic areas and hundreds of specialty medications, procedures, and tests. We are increasing the accuracy and transparency of the data collected and decreasing time to therapy all while remaining flexible, a benefit regularly highlighted by our customers.

For more information about Infinitus or to learn more about our purpose-built AI system, watch a demo or reach out to a member of our team today.

About Infinitus

Infinitus Systems is an AI healthcare company. By automating tedious and time-consuming administrative processes across the ecosystem, we’re creating time for healthcare to improve access, adherence, and affordability.

infinitus.ai