zensar

Giving Life to LLM: Generative AI and AI Solutions for Reducing Billing-related Customer Calls

Whitepaper

An RPG Company

Contact centers today are overwhelmed by skyrocketing call volumes driven by the increasing complexity of services, billing issues, and the demand for personalized support. Traditional systems struggle to handle the sheer volume of customer inquiries, resulting in prolonged wait times, higher operational costs, and dissatisfied customers. Advanced artificial intelligence (AI) solutions, particularly generative AI-powered systems, are essential to address these challenges. These systems automate routine inquiries, deliver quick and accurate responses, and reduce the burden on human agents, allowing them to focus on more complex, high-value tasks. This strategic approach not only enhances operational efficiency but also significantly improves the overall customer experience — making it a vital solution for modern contact centers grappling with increased demands.

Billing inquiries alone constitute a significant portion of customer service interactions. Customers often seek clarity on charges, leading to over 55,000 calls each month. This influx creates financial strain and dissatisfaction due to delays in resolution. With a cost of $7 per call, the current system incurs approximately $385,000 in monthly expenses for companies. One of our clients, a leading multinational telecommunications company, faced the same high volume of billing-related calls. We harnessed the power of AI to tackle its challenges effectively. This paper outlines our gen AI and AI approach , designed to reduce call volume, streamline resolution processes, and uphold service quality.

What is the operational challenge we're trying to solve?

The current billing system is burdened by a high volume of customer service interactions, leading to operational inefficiencies and increased costs. Customers are often confused by billing errors, complicated statements, and delayed charges, driving the influx of calls. This presents a significant challenge, as the manual resolution process is time-consuming and costly. Companies aim to reduce call volumes while improving customer service without compromising the quality of support provided.

AI and the leap forward

Our approach helps companies implement generative AI and AI features designed to enhance customer interaction with their billing system, reduce call volume, and streamline the query resolution process.

- **Call volume analysis:** The process begins with analyzing historical call volume data to understand patterns, identify pain points, and set a baseline for measuring improvement.
- **Gen AI solution:** Next, our gen AI solution is applied to address customer billing queries through AI-driven interactions, such as virtual avatars and automated agents, designed to reduce the need for human intervention.
- **Post-monitoring:** After implementing the gen AI solution, ongoing monitoring evaluates its impact on reducing call volume and improving the customer experience, ensuring continuous improvement and adjustment where necessary.

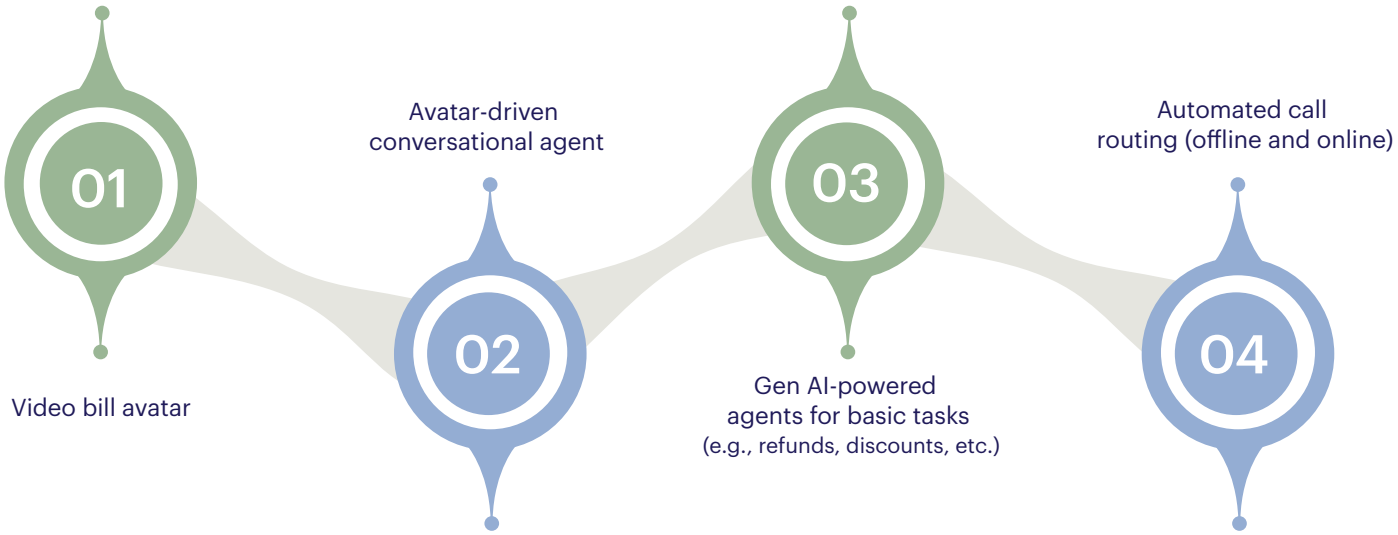


Figure 1: Key features

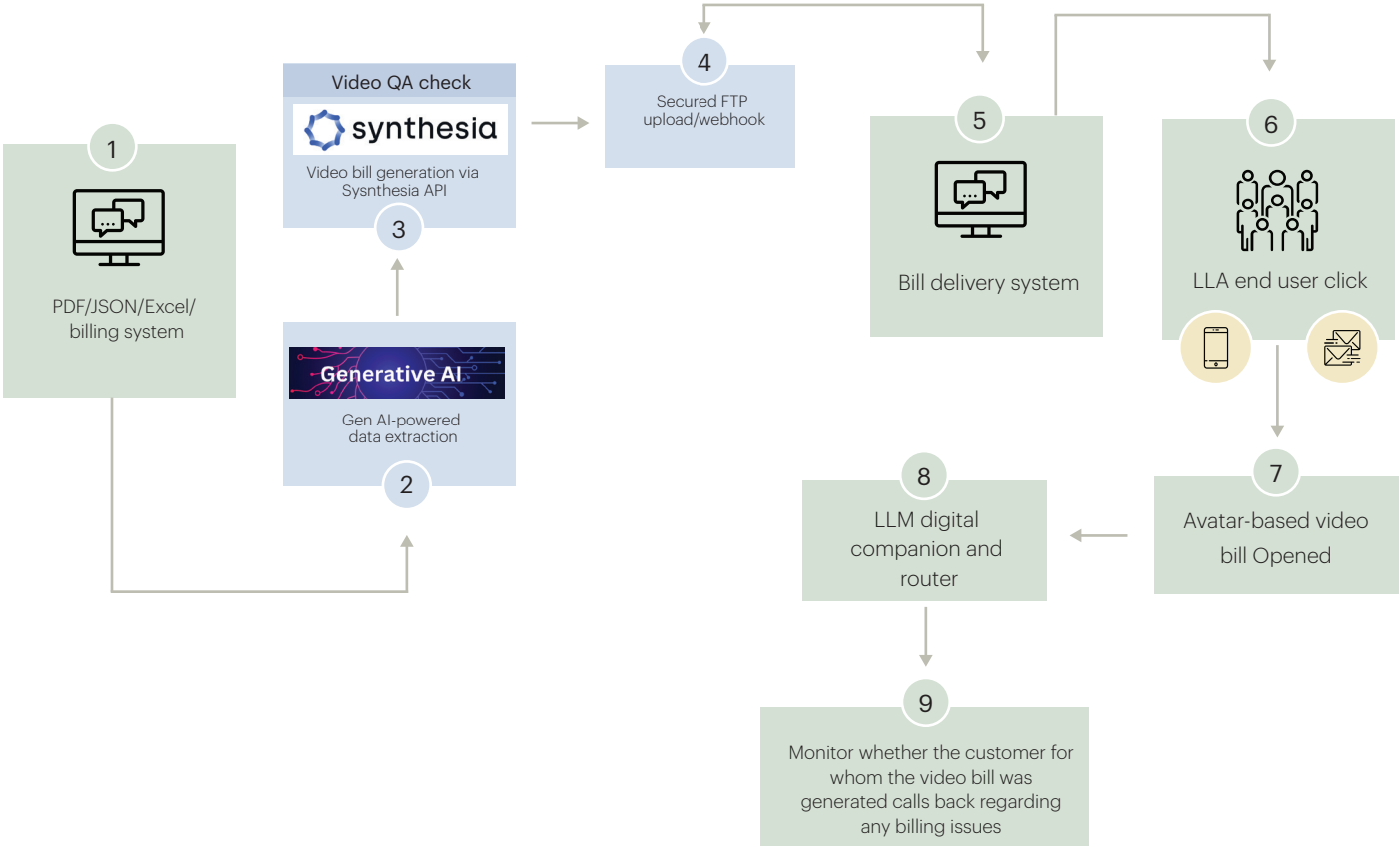


Figure 2: End-to-end flow of the gen AI solution

Feature 1: Video bill avatar

The ‘video bill avatar’ feature provides customers with an interactive way to understand their bills. Instead of just reading the bill, the main components are narrated by a 3D avatar, offering a personalized and clear explanation of the charges, discounts, and other essential details.

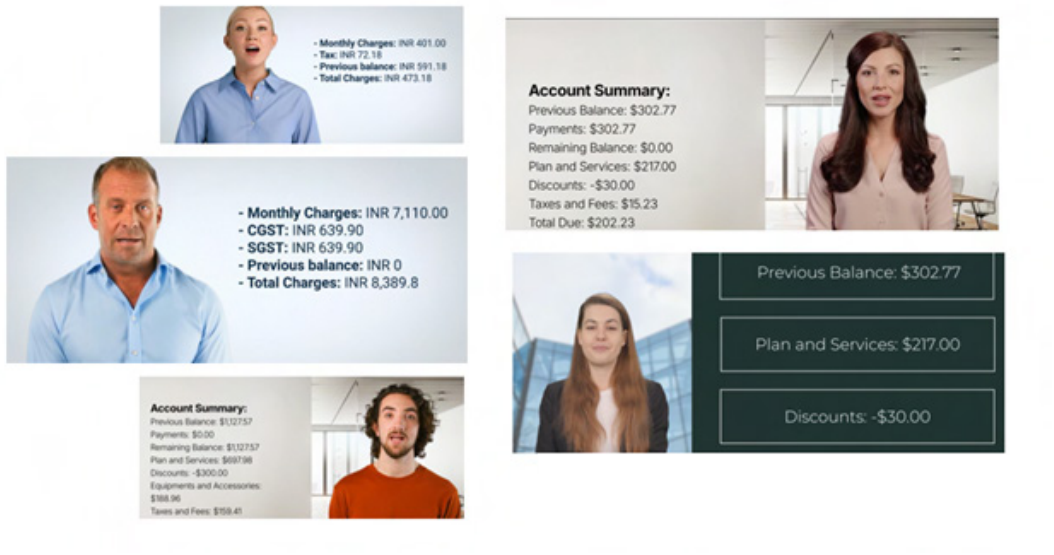


Figure 3: Video Bill Avatars

Technical implementation

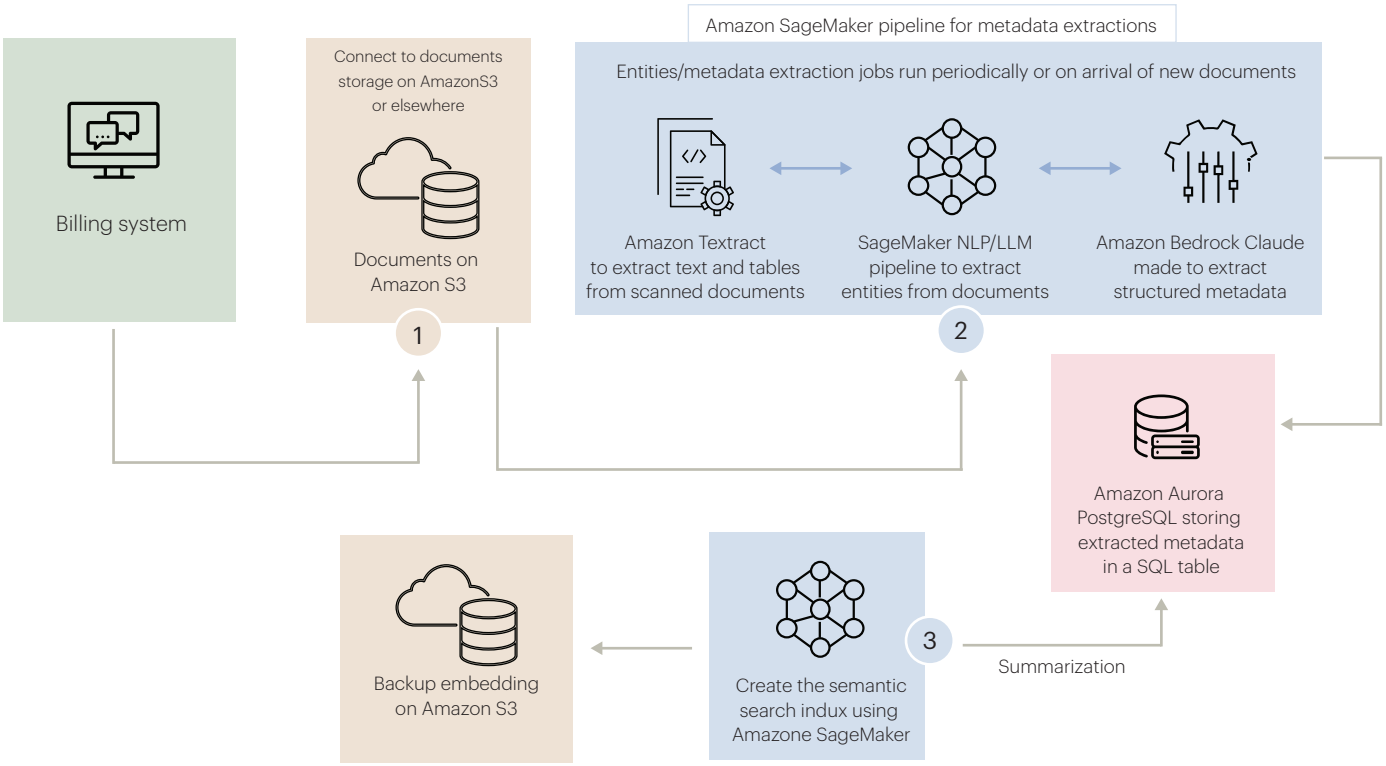


Figure 4: Video bill summarization powered by generative AI and LLM

Technical implementation

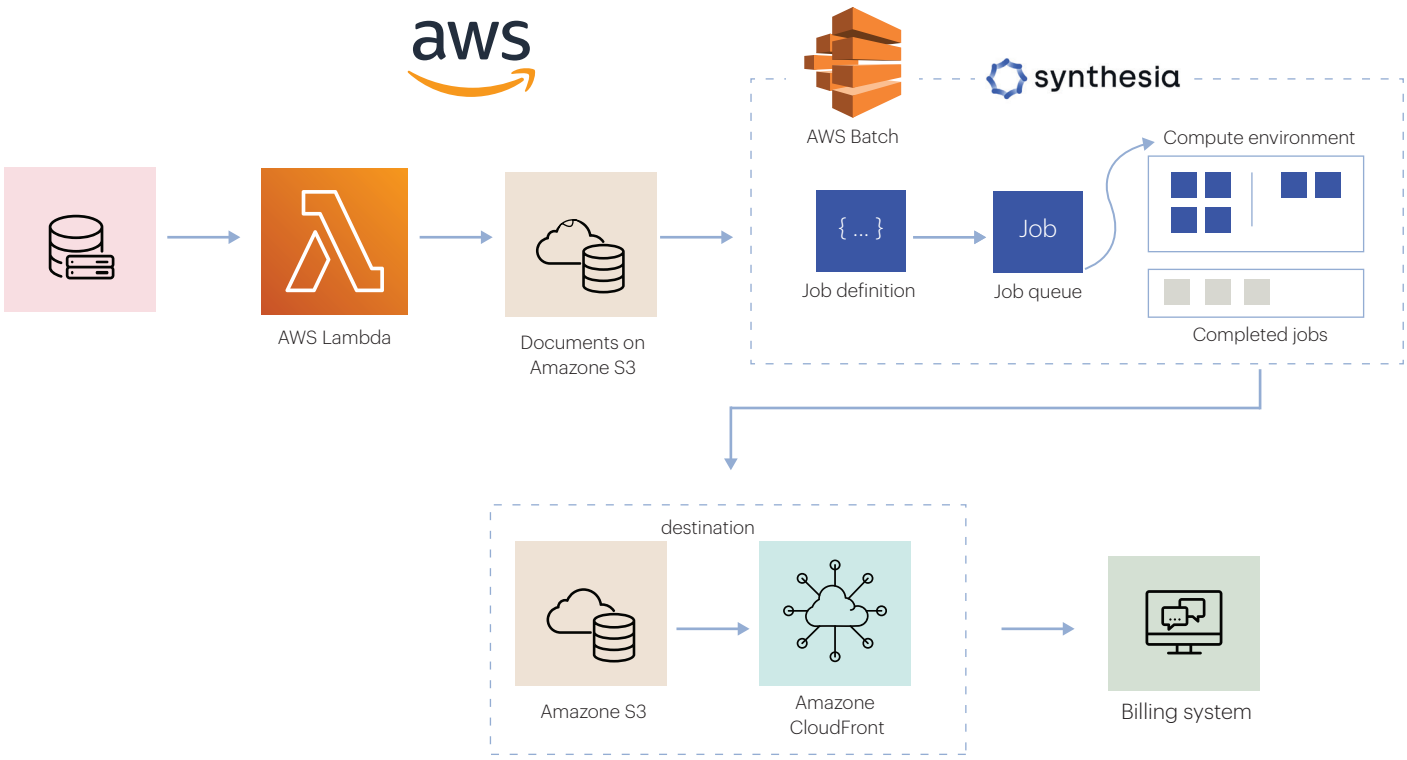


Figure 5: Video bill generation using Synthesia or something similar

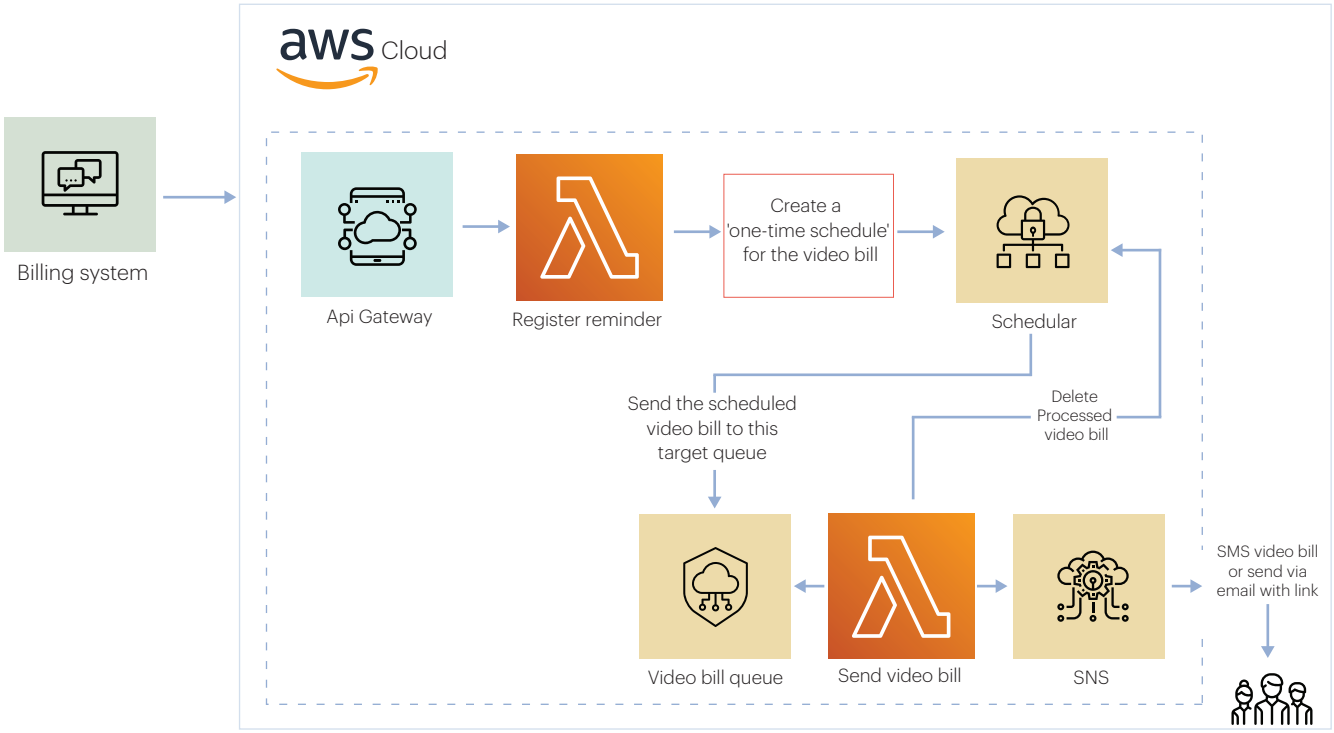


Figure 6: Video bill delivery

Avatar integration

We've developed a 3D avatar driven by advanced gen AI-powered bill-to-speech technology to provide personalized explanations of the bill's components. This avatar enhances customer interaction by delivering clear, real-time explanations.

LLM-based advanced RAG for billing solutions

A large language model (LLM)-based advanced retrieval-augmented generation (RAG) system retrieves sequential monthly bill details and tracks temporal changes in the billing history for each customer. Utilizing the LLM to handle natural language processing (NLP) tasks allows us to analyze bill content, identify discrepancies, and generate tailored responses to explain charges and billing changes. This personalized communication will significantly reduce customer confusion and call volume. The system will also feature an interactive virtual assistant that provides real-time bill explanations, ensuring a seamless customer experience while optimizing operational efficiency.

Multi-channel video bill

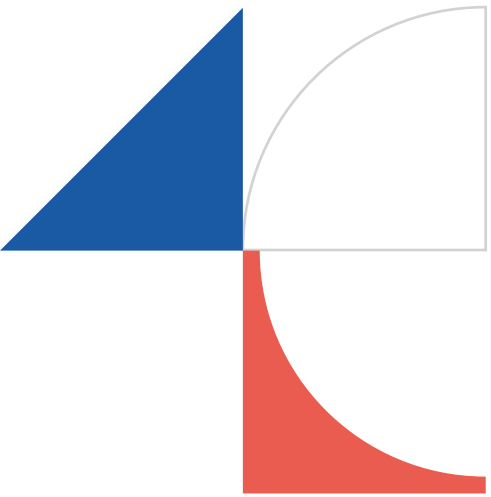
We ensure customers can access their video bills across multiple channels, including website, email, and mobile app. These video bills have a user-friendly interface, offering intuitive options to pause, replay, and skip sections as needed. This multi-channel approach allows customers to view their bills conveniently, ensuring an engaging and seamless user experience regardless of the platform.

Avatar rendering

The 3D avatar we develop is equipped with advanced gen AI-powered text-to-speech technology and enhances the overall customer experience with its realistic and human-like interaction capabilities. The avatar's facial expressions, body language, and lip-syncing accuracy ensure natural, life-like communication. It responds dynamically to each bill's content, adjusting its tone, emphasis, and expression based on the complexity of the information. This level of detail in avatar rendering creates a more immersive experience, making customers feel like they are interacting with an actual representative while still benefiting from automated, scalable AI-driven support.

LLM summarization

We utilize the LLM's capabilities to summarize bills using an RAG approach, allowing us to retrieve and condense customer bill details dynamically. This process will involve analyzing a customer's billing history, identifying important changes or charges, and creating concise explanations that are easy to understand. The LLM will break down complex billing structures into clear, plain-language summaries, ensuring that each customer has a comprehensive view of their bills. Additionally, the summarization process will be adaptive, meaning the LLM will learn from past interactions and continuously improve its ability to generate tailored summaries, enhancing the personalization of bill explanations over time.



Benefits

- **Improved understanding:** Customers can understand their bills better through interactive video explanations, significantly reducing confusion and the need for direct support.
- **Cost reduction:** We can achieve significant cost savings by reducing the number of customers who need to contact customer service for bill explanations.

Feature 2: Avatar-driven conversational agent, digital companion, and routing

This feature introduces a conversational AI agent, represented by an avatar that customers can interact with. Customers can ask questions about their bills, and the avatar will respond in natural language, providing a seamless and human-like interaction.

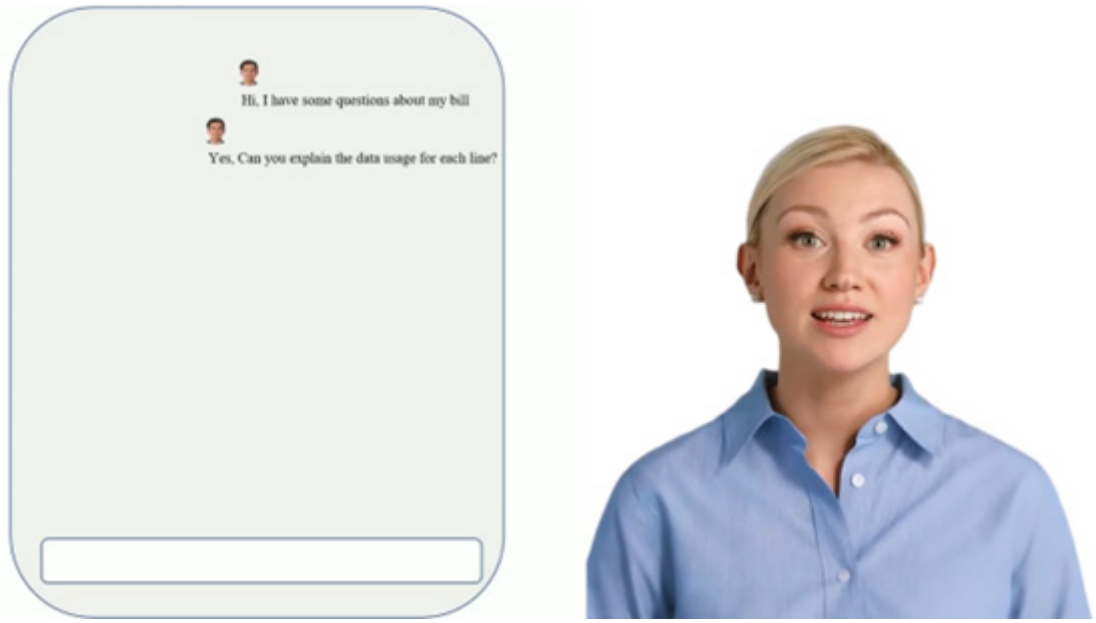


Figure 7: Interactive conversation BOT with avatar

- **User interaction with chat avatar**

  Customers who are unsatisfied with their video bill can interact with a chat-based avatar agent via various devices, including desktop and mobile. The avatar uses speech-to-speech technology to enable real-time, dynamic conversations, allowing users to ask additional questions about their bills and receive explanations directly from the system.

- **Authentication and authorization**

  Before interacting with the avatar, users are authenticated and authorized through a secure identity management system, ensuring only verified customers can access their billing details, thus maintaining privacy and security.

- **Conversation handling and moderation**

  Customer queries are processed by a conversation history service, which logs user interactions in real time. A moderation model filters and analyzes incoming queries to ensure the system only processes appropriate and bill-related questions. This helps streamline the interaction and keeps the conversation focused on billing inquiries.

- **Query processing**

  Once a query is validated, it is sent to the core system, where an agent determines which tools or models should be invoked based on the customer's question. Depending on the type of question, the system may access information from various data sources, including customer-specific billing data.

- **Customer billing data storage (Vector DB)**

  Customer billing details are stored in a vector database, which indexes all past bills for each user. This allows the system to perform fast, semantic searches and retrieve relevant bill details based on customer inquiries. The vector database efficiently handles large amounts of historical billing data and supports advanced search capabilities.

- **RAG-driven speech response**

  The system uses a RAG approach powered by a speech-to-speech engine. When a customer asks a question, the system retrieves the relevant information from the vector database and uses a language model to generate a clear, speech-based response. The avatar delivers this response in real time, offering a personalized and conversational user experience.

- **Explanation generation**

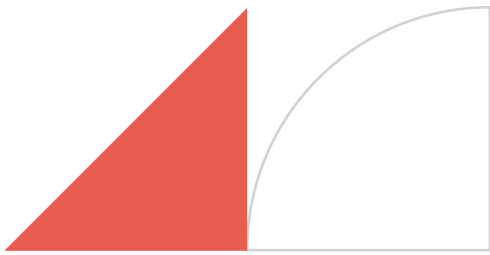
  If required, the system can generate complex billing details. The summarization feature allows the avatar to provide concise explanations of the most critical aspects of a customer's bill, such as charges, credits, and changes over time. These explanations are powered by the same language model, ensuring consistency in the responses provided.

- **Backend data management and extension**

  The system architecture is designed to allow for easy scalability and extensions. New tools and modules can be added to the backend as necessary, allowing for future enhancements to the billing explanation process or additional use cases.

- **Cloud and model execution**

  The entire interaction, from query processing to response generation, occurs in the cloud, where language generation and semantic search models are executed. This ensures scalability, reliability, and real-time responsiveness in the system's interactions with users.



Technical implementation

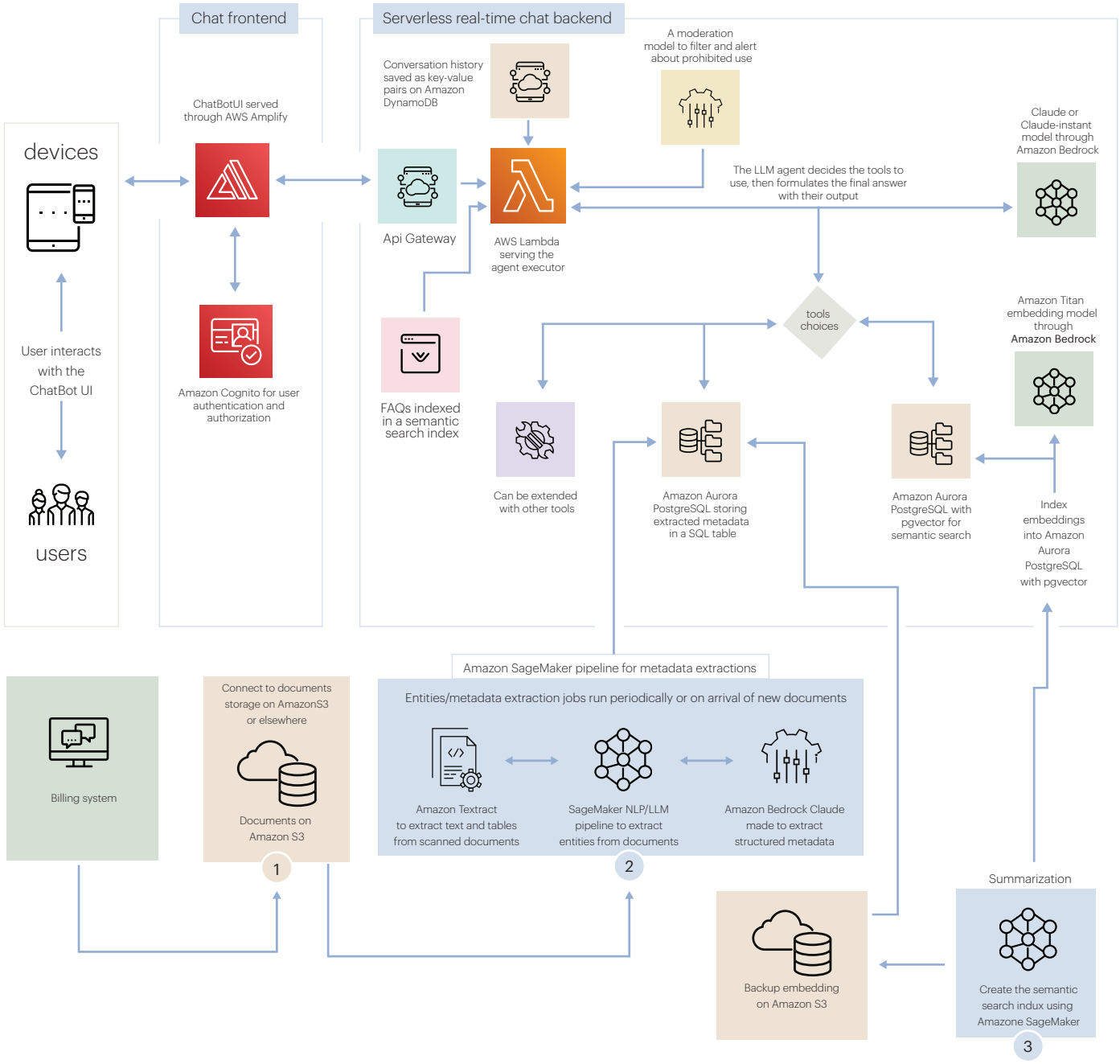


Figure 8: Technical implementation of avatar-driven conversational agent, digital companion, and routing

Intelligent routing to human agent

If a customer is still not satisfied with the avatar's responses or requests further escalation, the system intelligently routes the query to a human agent. This escalation is triggered based on predefined thresholds such as repeated questions or explicit requests for human intervention. The system transfers the complete conversation history to the human agent, including the avatar's responses and the customer's specific queries. This ensures that the agent is fully informed of the context, reducing the need for the customer to repeat information and enabling the agent to provide a more personalized and efficient resolution.

Benefits

- **Enhanced customer experience:** Customers will have access to a virtual assistant that provides immediate answers, improving satisfaction and reducing frustration.
- **Scalability:** The system can handle an unlimited number of queries simultaneously, further reducing the need for human intervention.

Feature 3: Automated call routing (offline and online)

Technical implementation

- **Call analysis:** An AI-powered system will analyze incoming calls to identify the nature of the issue. It will use speech recognition and natural language understanding to categorize the query.
- **Routing logic:** Based on the analysis, the system will route the call accordingly:
  - **Online resolution:** If the issue is straightforward and can be handled by the avatar-driven agent, the customer will be directed to the online platform.
  - **Human support:** If the issue is complex or sensitive, the call will be routed to a human agent for personalized support.
- **Feedback loop:** The system will continuously learn from call outcomes to improve its routing accuracy.

Benefits

- **Efficiency:** Ensures that customers are directed to the most appropriate resource, reducing wait times and improving the likelihood of a first-call resolution.
- **Resource optimization:** Human agents will only handle calls requiring expertise, allowing companies to optimize resources and reduce operational costs.

Feature 4: Gen AI-powered agents for basic tasks (e.g., refunds, etc.)

Description

This feature introduces gen AI-powered agents capable of handling basic, repetitive tasks such as processing refunds. These tasks often consume significant human resources but can be efficiently managed by AI, reducing operational costs and improving response times.

Technical implementation

- **Gen AI integration:** The gen AI agents are designed to execute specific tasks like processing refunds by interacting with companies' backend systems. These agents will follow predefined workflows to ensure compliance with company policies and transaction accuracy.
- **Task automation:** The agents will automate the end-to-end process of handling refunds, including validating refund requests, updating customer accounts, and issuing payments. They will have decision-making capabilities to handle straightforward cases independently, while more complex cases can be escalated to human agents.
- **User interaction:** Customers can request a refund through the companies' website, mobile app, and customer service portal. The gen AI agent will guide them through the process, providing real-time updates and confirmation once the refund is processed.
- **Audit and compliance**: Transactions that gen AI agents process are logged and audited to ensure compliance with regulatory requirements and company standards.

Benefits

- **Efficiency:** Automating basic tasks like refunds reduces the workload on human agents, allowing them to focus on more complex customer issues.
- **Cost savings:** By handling routine tasks autonomously, companies can significantly reduce operational costs.
- **Improved customer experience:** Customers receive faster responses and resolutions for basic tasks, leading to higher satisfaction and trust in the service.

Conclusion

Organizations can anticipate a significant reduction in customer inquiries by implementing advanced AI and generative AI features, potentially deflecting up to 60 percent of routine calls. This reduction leads to substantial cost savings and more efficient resource allocation while improving overall customer satisfaction. Using interactive video explanations, conversational AI-driven agents, intelligent call routing, and AI-powered task automation, we can create a comprehensive solution that addresses the volume and complexity of customer interactions across various service areas, particularly for recurring inquiries.

References

- https://aws.amazon.com/blogs/machine-learning/boosting-rag-based-intelligent-document-assistants-using-entity-extraction-sql-querying-and-agents-with-amazon-bedrock/
- https://aws.amazon.com/startups/learn/serverless-retrieval-augmented-generation-on-aws#overview
- https://levelup.gitconnected.com/using-aws-eventbridge-scheduler-to-build-a-serverless-reminder-application-ba3086cf8e
- https://aws.amazon.com/startups/learn/serverless-retrieval-augmented-generation-on-aws#overview
- https://aws.amazon.com/blogs/machine-learning/unlock-the-potential-of-generative-ai-in-industrial-operations/?advocacy_source=everyonesocial&trk=global_employee_advocacy&sc_channel=sm&es_id=f1a21df807

Authors:

Indrajit Kar, Head - AI

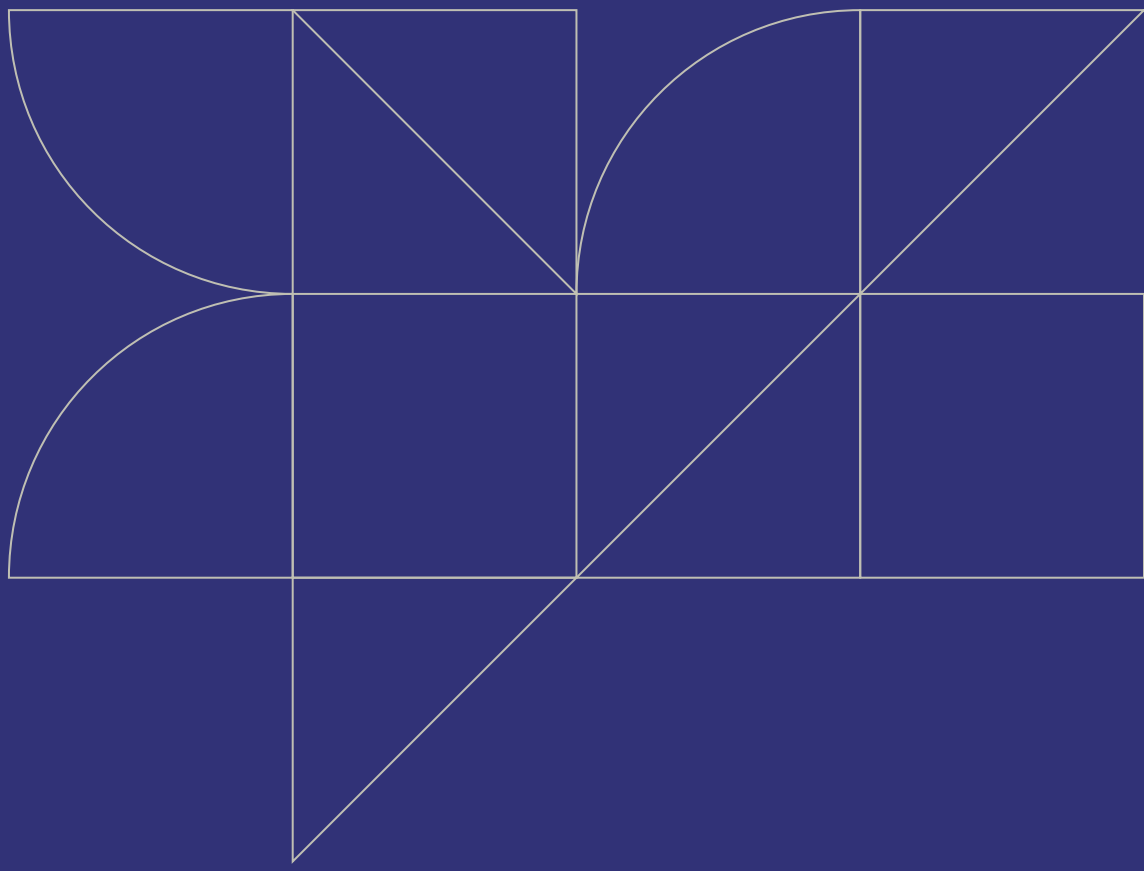
Vikram Parihar, Head - Sales

Zonunfeli Ralte, AI/ML Manager

Vidhi Rai, 3D and Avatar Lead

Souvik Majumdar, Generative AI Specialist

Shrestha Mitra, Assistant Manager



zensar

An RPG Company

At Zensar, we're 'experience-led everything.' We are committed to conceptualizing, designing, engineering, marketing, and managing digital solutions and experiences for over 145 leading enterprises. Using our 3Es of experience, engineering, and engagement, we harness the power of technology, creativity, and insight to deliver impact.

Part of the $4.8 billion RPG Group, we are headquartered in Pune, India. Our 10,000+ employees work across 30+ locations worldwide, including Milpitas, Seattle, Princeton, Cape Town, London, Zurich, Singapore, and Mexico City.

For more information, please contact: **info@zensar.com | www.zensar.com**