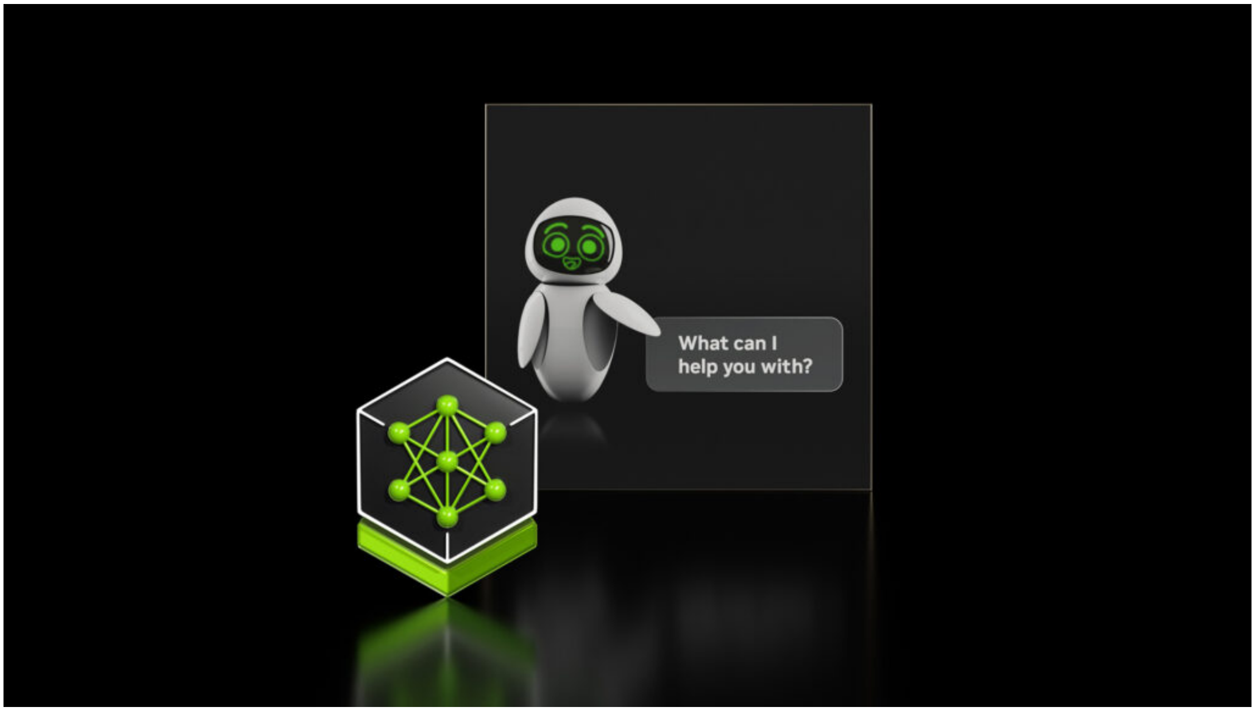
Three Building Blocks for Creating AI Virtual Assistants for Customer Service with an NVIDIA AI Blueprint



Dec 11, 2024
By Isabel Hulseman and Ruchika Kharwar
+27
Like
Discuss (0)

AI-Generated Summary

- Companies are using generative AI to improve customer service by enhancing operational efficiency, reducing costs, and maximizing ROI with the help of NVIDIA's AI Blueprint for AI virtual assistants.

- NVIDIA's AI Blueprint enables developers to create AI virtual assistants that use retrieval-augmented generation to provide personalized and accurate responses to customer queries in real-time.
- The NVIDIA AI Blueprint is designed to integrate with existing customer service applications, ensuring data integrity and governance, and can be customized for specific industry use cases with the help of NVIDIA consulting partners like Accenture and Deloitte.

AI-generated content may summarize information incompletely. Verify important information. Learn more

In today's fast-paced business environment, providing exceptional customer service is no longer just a nice-to-have—it's a necessity. Whether addressing technical issues, resolving billing questions, or providing service updates, customers expect quick, accurate, and personalized responses at their convenience. However, achieving this level of service comes with significant challenges.

Legacy approaches, such as static scripts or manual processes, often fall short when it comes to delivering personalized and real-time support. Additionally, many customer service operations rely on sensitive and fragmented data, which is subject to strict data governance and privacy regulations. With the rise of generative AI, companies aim to revolutionize customer service by enhancing operational efficiency, cutting costs, and maximizing ROI.

Integrating AI into existing systems presents challenges related to transparency, accuracy, and security, which can impede adoption and disrupt workflows. To overcome these hurdles, companies are leveraging generative AI-powered virtual assistants to manage a wide range of tasks, ultimately improving response times and freeing up resources.

This post outlines how developers can use the NVIDIA AI Blueprint for AI virtual assistants to scale operations with generative AI. By leveraging this information, including sample code, businesses can meet the growing demands for exceptional customer service while ensuring data integrity and governance. Whether improving existing systems or creating new ones, this blueprint empowers teams to meet customer needs with efficient and meaningful interactions.

Smarter AI virtual assistants with an AI query engine using retrieval-augmented generation

When building an AI virtual assistant, it’s important to align with the unique use case requirements, institutional knowledge, and needs of the organization. Traditional bots, however, often rely on rigid frameworks and outdated methods that struggle to meet the evolving demands of today’s customer service landscape.

Across every industry, AI-based assistants can be transformational. For example, telecommunications companies, and the majority of retail and service providers, can use AI virtual assistants to enhance customer experience by offering support 24 hours a day, 7 days a week while handling a wide range of customer queries in multiple languages and providing dynamic, personalized interactions that streamline troubleshooting and account management. This helps reduce wait times and ensures consistent service across diverse customer needs.

Another example is within the healthcare insurance payor industry, where ensuring a positive member experience is critical. Virtual assistants enhance this experience by providing personalized support to members, addressing their claims, coverage inquiries, benefits, and payment issues, all while ensuring compliance with healthcare regulations. This also helps reduce the administrative burden on healthcare workers.

With the NVIDIA AI platform, organizations can create an AI query engine that uses retrieval-augmented generation (RAG) to connect AI applications to enterprise data. The AI virtual assistant blueprint enables developers to quickly get started building solutions that provide enhanced customer experiences. It is built using the following NVIDIA NIM microservices:

- **NVIDIA NIM for LLM:** Brings the power of state-of-the-art large language models (LLMs) to applications, providing unmatched natural language processing with remarkable efficiency.
    - Llama 3.1 70B Instruct NIM**:** Powers complex conversations with superior contextual understanding, reasoning, and text generation.
- NVIDIA NeMo **Retriever NIM:** This collection provides easy access to state-of-the-art models that serve as foundational building blocks for RAG pipelines. These pipelines, when integrated into virtual assistant solutions, enable seamless access to enterprise data, unlocking institutional knowledge via fast, accurate, and scalable answers.
    - **NeMo** Retriever Embedding NIM**:** Boosts text question-answering retrieval performance, providing high-quality embeddings for the downstream virtual assistant.
    - **NeMo** Retriever Reranking NIM**:** Enhances the retrieval performance further with a fine-tuned reranker, finding the most relevant passages to provide as context when querying an LLM.

The blueprint is designed to integrate seamlessly with existing customer service applications without breaking information security mandates. Thanks to the portability of NVIDIA NIM, organizations can integrate data wherever it resides. By bringing generative AI to the data, this architecture enables AI virtual assistants to provide more personalized experiences tailored to each customer by leveraging their unique profiles, user interaction histories, and other relevant data.

A blueprint is a starting point that can be customized for an enterprise's unique use case. For example, integrate other NIM microservices, such as the Nemotron 4 Hindi 4B Instruct, to enable an AI virtual assistant to communicate in the local language. Other microservices can enable additional capabilities such as synthetic data generation and model fine-tuning to better align with your specific use case requirements. Give the AI virtual assistant a humanlike interface when connected to the digital human AI Blueprint.

With the implementation of a RAG backend with proprietary data (both company and user profile and their specific data), the AI virtual assistant can engage in highly contextual conversations, addressing the specifics of each customer's needs in real-time. Additionally, the solution operates securely within your existing governance frameworks, ensuring compliance with privacy and security protocols especially when working with sensitive data.

Three building blocks for creating your own AI virtual assistant

As a developer, you can build your own AI virtual assistant that retrieves the most relevant and up-to-date information, in real time, with ever-improving humanlike responses. Figure 1 shows the AI virtual assistant architecture diagram which includes three functional components.

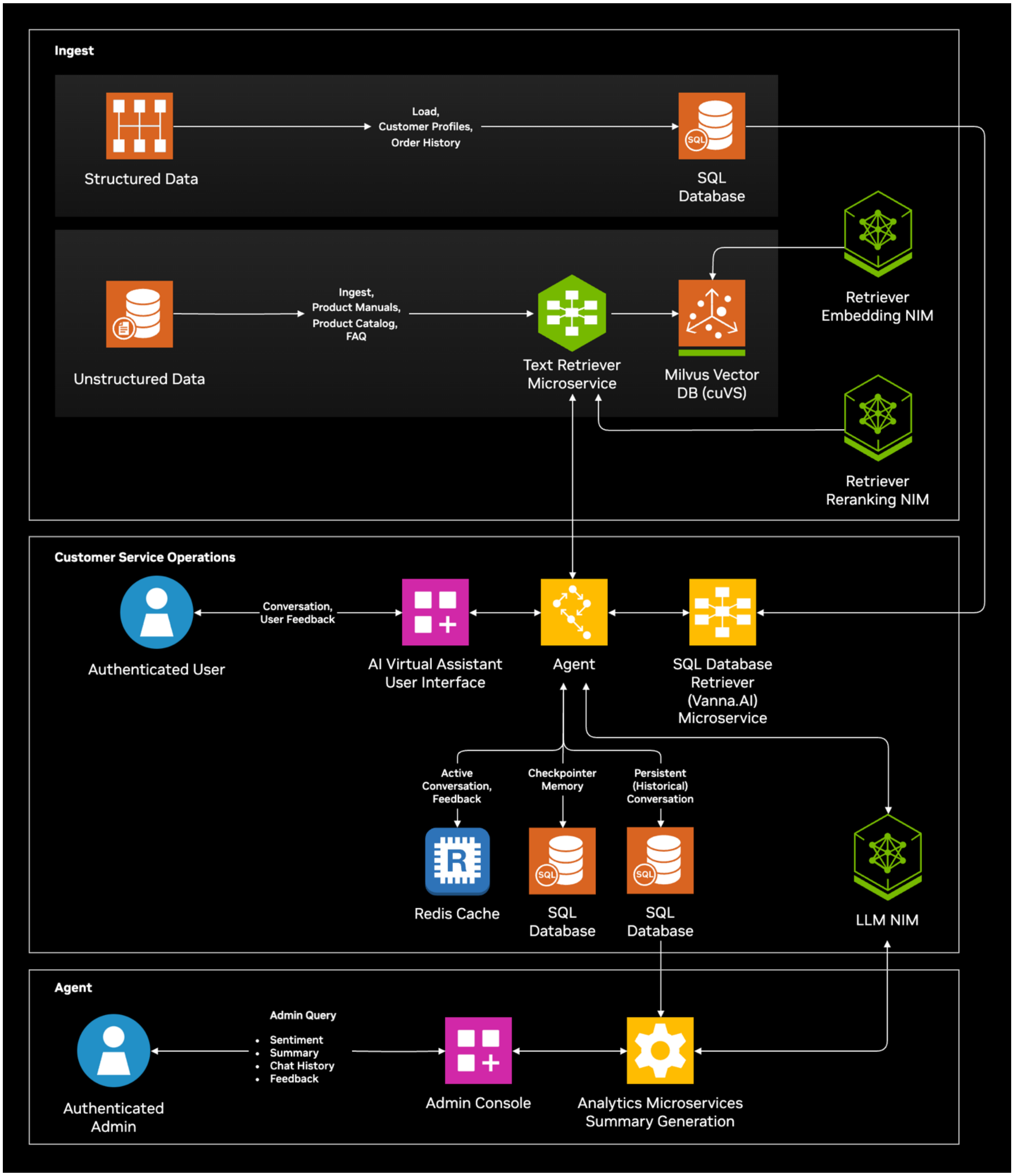


Figure 1. The NVIDIA AI Blueprint for AI virtual assistants

1. Data ingestion and retrieval pipeline

Pipeline administrators use the ingestion pipeline to load structured and unstructured data into the databases. Examples of structured data include customer profiles, order history, and order status. Unstructured data includes product manuals, the product catalog, and supporting material such as FAQ documents.

2. AI agent

The AI virtual assistant is the second functional component. Users interact with the virtual assistant through a user interface. An AI agent, implemented in the LangGraph agentic LLM programming framework, plans how to handle complex customer queries and solves recursively. The LangGraph agent uses the tool calling feature of the Llama 3.1 70B Instruct NIM to retrieve information from both the unstructured and structured data sources, then generates an accurate response.

The AI agent also uses short-term and long-term memory functions to enable multi-turn conversation history. The active conversation queries and responses are embedded so they can be retrieved later in the conversation as additional context. This allows more human-like interactions and eliminates the need for customers to repeat information they've already shared with the agent.

Finally, at the end of the conversation, the AI agent summarizes the discussion along with a sentiment determination and stores the conversation history in the structured database. Subsequent interactions from the same user can be retrieved as additional context in future conversations. Call summarization and conversation history retrieval can reduce call time and improve customer experience. Sentiment determination can provide valuable insights to the customer service administrator regarding the agent's effectiveness.

3. Operations pipeline

The customer operations pipeline is the third functional component of the overall solution. This pipeline provides important information and insight to the customer service operators. Administrators can use the operations pipeline to review chat history, user feedback, sentiment analysis data, and call summaries. The analytics microservice, which leverages the Llama 3.1 70B Instruct NIM, can be used to generate analytics such as average call time, time to resolution, and customer satisfaction. The analytics are also leveraged as user feedback to retrain the LLM models to improve accuracy.