Agentic AI Assistant for Enhanced Customer Experience

Introduction

This case study describes the design, development, and deployment of an advanced Agentic AI-powered, multimodal customer support assistant. The assistant operates seamlessly across web chat, web voice, and telephony. It uses a single unified backend with real-time streaming capabilities, intelligent routing, structured output, and personalization for returning or logged-in users.

The solution eliminates customer friction, reduces operational overhead, and delivers human-like conversational capabilities through a sophisticated multi-agent and multimodal architecture.

.

Client Details

Name: Confidential | **Industry:** Ecommerce, Retail, Software | **Location:** USA

Technologies

GPT-4.1-mini, Semantic Kernel, Azure Communication Services, Azure Web Apps, Azure AI Search, Microsoft Teams, FastAPI, Docker

Agentic AI Assistant for Enhanced Customer Experience

Project Description

The Challenge

Traditional customer service often struggles with high call volumes, repetitive inquiries, and the need for 24/7 availability. Customers frequently face long waiting times, and manual processes for tasks like order status checks and appointment bookings can be inefficient and prone to errors. There was a clear need for an intelligent, scalable solution to automate routine tasks, provide instant support, and free up human agents for more complex issues.

Customers frequently experience friction, especially when switching channels or repeating the same information. A scalable, intelligent, multimodal solution was needed to automate routine tasks, provide instant support, and free human agents for complex issues.

The Solution: A Multimodal, Agent-Orchestrated AI Assistant

To overcome these challenges, a unified Agentic AI chatbot was developed, capable of handling all customer interaction modes—chat, voice, and telephony, through a single orchestrated backend.

At its core, the system uses:

- Three Specialized Agents: FAQ Agent, Order Status Agent, Booking Agent
- A Router Agent: Determines user intent dynamically
- A Multimodal Engine: Supports text, real-time voice, and phone calls
- A JSON Output Parser: Converts HTML/Markdown into structured UI elements
- Advanced Personalization: Automatically retrieves user data when logged in
- Barge-in Voice support: Users can interrupt speech output anytime

The result is a unified, intelligent, scalable customer service platform.

Agentic AI Assistant for Enhanced Customer Experience

Key Features and Capabilities

1. Multimodal Interaction (Chat, Voice, and Telephony)

All modes share the same entry point and backend logic.

Web Chat

- Text-based conversation
- Interactive UI through structured JSON responses
- Buttons, actions, links, custom components

Web Voice (via WebSocket Streaming)

- Real-time STT (Azure Speech-to-Text)
- Real-time TTS (Azure Neural Voices)
- Bot produces:
  - A full textual response
  - A compressed spoken summary
- Simultaneous speaking and listening: the bot can continue listening while speaking in certain scenarios
- Barge-in support: User interrupts the bot mid-sentence

Telephony (via Azure Communication Services)

- Full phone-call integration
- Users speak naturally as if talking to a human
- Compatible with Router Agent + downstream agents

All channels share the same intent detection and routing logic, so the user experience is consistent regardless of the interface.

2. Router Agent for Intelligent Intent Routing

Incoming messages (text or STT transcript) are analyzed by the Router Agent to determine user intent:

- FAQ questions → routed to FAQ Agent
- Order lookup → routed to Order Status Agent
- Bookings → routed to Booking Agent
- Ambiguous or multi-intent messages → clarified or split

Router Agent ensures clean transitions and maintains conversational context across steps.

Agentic AI Assistant for Enhanced Customer Experience

3. FAQ Agent with Contextual Hyperlinks

The FAQ Agent delivers accurate and helpful responses based on content sourced from the website's FAQ pages.

Key Enhancements

- Responses include relevant hyperlinks to:
  - Specific FAQ sections
  - Offer pages
  - Policy documents
  - Support articles
- These links open in the background while the chat stays in the foreground
- Rich, navigable answers that go beyond simple text

This encourages self-service and reduces escalation volume.

4. Order Status Agent with Multi-Input Flexibility

The Order Status flow is optimized for both speed and accuracy.

Supported Input Combinations

- Phone Number + Date of Birth
- Order ID + Date of Birth

Both inputs may be provided simultaneously to reduce back-and-forth clarifications.

Logged-In User Personalization

When the customer is authenticated:

- Phone number, email, DOB, and historical order activity are pre-fetched
- No need to re-enter personal details
- Fastest possible path to order verification

Agentic AI Assistant for Enhanced Customer Experience

5. Booking Agent with Multi-Step Orchestration

The Booking Agent performs a full, automated appointment scheduling process:

Workflow

1. Store Selection
2. Exam Selection
3. Fetch Available Time Slots
4. Fetch Available Time Slots
5. Real-Time Availability Check
6. User Profile Handling:
    - New users → onboarding flows
    - Returning/logged-in users → auto-fill personal info
7. Confirmation and Summary Display

Each step uses the JSON parser to generate actionable UI blocks.

6. Structured Output with JSON Parser

The AI often produces Markdown/HTML. A custom JSON Parser converts this into a standardized schema that the frontend can render:

Supports

- Buttons (actions, payloads)
- Links
- UI prompts
- Mixed-mode guidance (UI + manual entry)

This creates a rich, app-like experience from natural language output.

7. Real-Time Voice Interaction Enhancements

Barge-In Capability

- Users can talk while the bot is speaking
- Bot immediately stops speaking and listens
- Works on both web and telephony

Agentic AI Assistant for Enhanced Customer Experience

Simultaneous Listening/Speaking

- The bot can preemptively detect user speech
- Delivers a natural, human-like interaction pattern

These features dramatically improve usability in voice-first scenarios.

8. Human Handoff through Microsoft Teams

When the AI detects complex issues, emotional tone, or user request for escalation:

- The session can be transferred to a live human agent
- Transfer uses Microsoft Teams integration, ensuring enterprise-grade reliability
- Conversation context is preserved

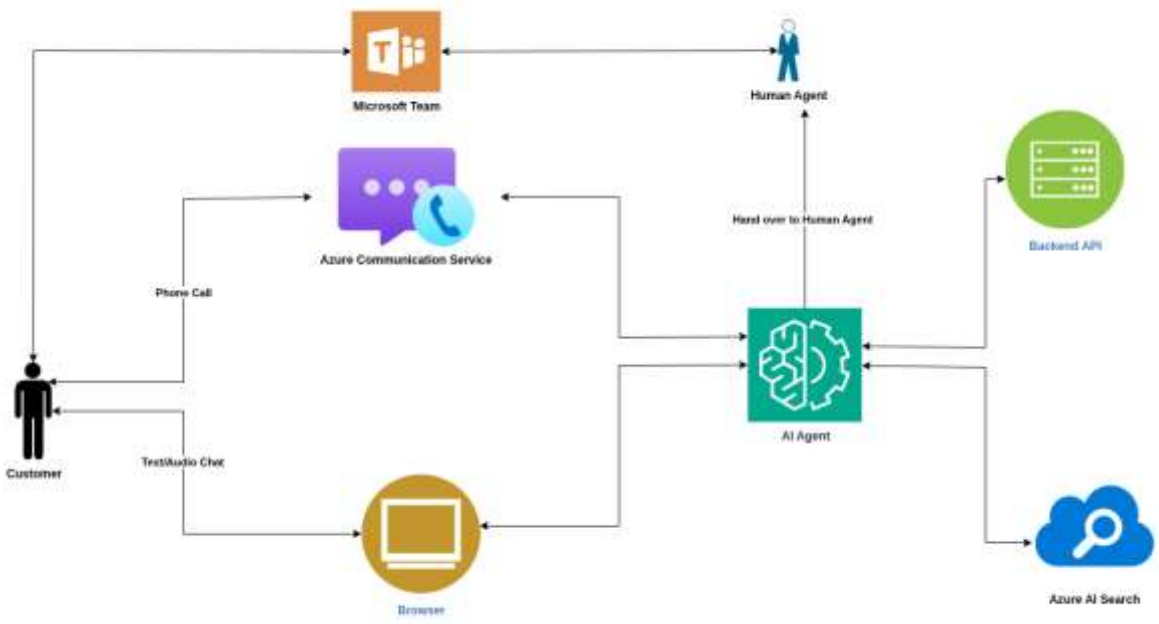
Technology Stack:

The solution is built on a modern and scalable technology stack:

- Text Version and Audio Mode: Utilizes the GPT-4.1-mini model for its efficiency and capabilities in text-based conversations. Semantic Kernel is employed for creating and orchestrating the multiple AI agents, allowing for complex multi-step interactions.
- Phone: Azure Communication Service is integrated to handle the underlying phone call functionalities.
- Deployment: The application is developed using Quart for its high performance and ease of use. It is deployed as a Docker container on Azure Web Apps, ensuring scalability, reliability, and easy management.
- Knowledge Base: For efficient FAQ handling, Azure AI Search is used to provide the underlying knowledge base, empowering the Large Language Model (LLM) with accurate and relevant information.

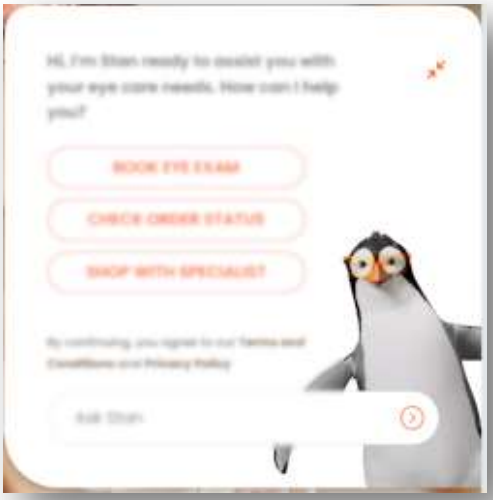
Agentic AI Assistant for Enhanced Customer Experience

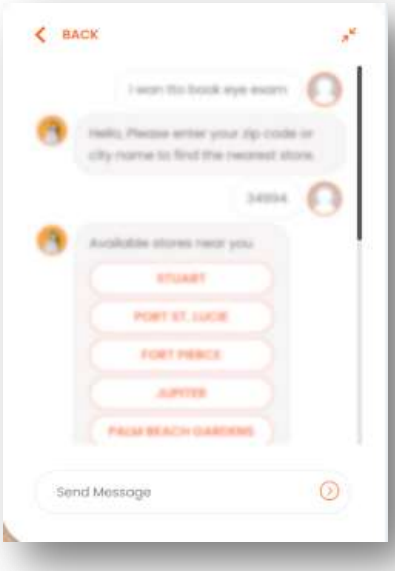
Architecture

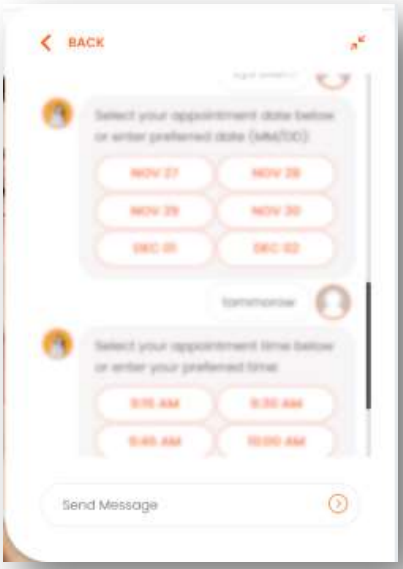


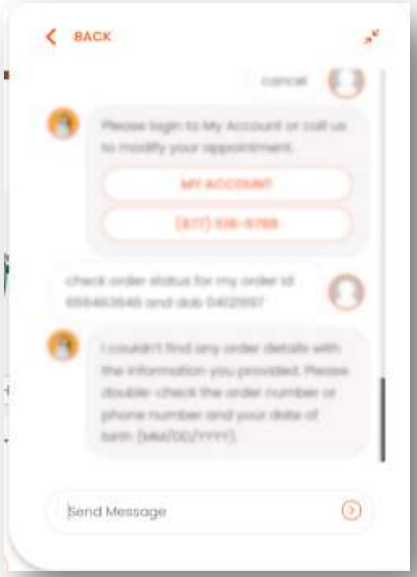
Agentic AI Assistant for Enhanced Customer Experience

Screenshots

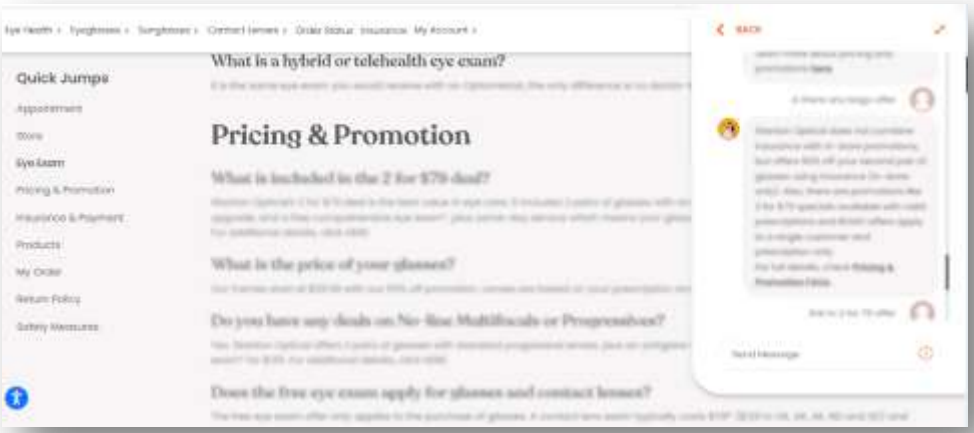








Agentic AI Assistant for Enhanced Customer Experience



Agentic AI Assistant for Enhanced Customer Experience

Benefits and Outcomes

This Agentic AI chatbot solution delivers significant benefits, transforming customer service operations:

- **Enhanced Customer Satisfaction**: Provides instant, 24/7 support across preferred channels, reducing wait times and offering a more personalized experience.
- **Operational Efficiency**: Automates repetitive tasks, allowing human agents to focus on high-value interactions and complex issues, leading to reduced operational costs.
- **Scalability**: The cloud-native deployment on Azure ensures the solution can easily scale to meet fluctuating demand, maintaining performance even during peak periods.
- **Improved Accuracy**: Leveraging advanced AI models and a dedicated knowledge base ensures consistent and accurate information delivery.
- **Seamless Integration**: The ability to transfer calls to human agents via Microsoft Teams ensures a smooth transition for users requiring further assistance, maintaining a high level of service.