



PROJECT FINAL REPORT

BCIS 5140 Term Project – FALL 2025

Playstaytion Pet Resort & Training – Agentic AI Voice Assistant



GROUP 2

**Rajdeep Kaur Sandhu
Sheetal Kantilal
Shoaib Khan
Uma Maheshwari Kannan**

TABLE OF CONTENTS

<i>EXECUTIVE SUMMARY</i>	3
<i>PROJECT GOALS AND OUTCOMES</i>	5
<i>SCOPE OF THE PROJECT</i>	6
<i>PROJECT APPROACH</i>	8
<i>SOLUTION REQUIREMENTS</i>	10
<i>SOLUTION DESIGN</i>	12
<i>SOLUTION IMPLEMENTATION</i>	22
<i>APPENDIX: PROTOTYPE</i>	26
<i>REFERENCES</i>	27

EXECUTIVE SUMMARY

PlayStayTion Pet Resort and Training is a locally owned pet care facility in North Texas that offers boarding, daycare, grooming, and training services for dogs. Known for its five-star care and welcoming environment, the resort provides a safe, fun, and comfortable experience for every pet. Over the years, PlayStayTion has built a loyal community of customers who appreciate its dedication and personal service. However, like many growing small businesses, it faces challenges in keeping up with rising customer demand. Missed calls, delayed responses, and limited after-hours availability often lead to customer frustration, lost reservations, and reduced efficiency. These issues affect both the company's revenue and its reputation, motivating the management team to find a smarter, cost-effective way to stay connected with customers around the clock.

The purpose of this project is to explore how agentic AI can help address these challenges and improve customer service at PlayStayTion. Our team is designing an AI Voice Assistant that can manage customer calls, answer frequently asked questions, assist with booking inquiries, and provide accurate, real-time information about services. The project serves as a proof of concept to show how conversational AI can create smoother, more consistent customer interactions while freeing staff to focus on in-person care. The system integrates Twilio for call handling, Voiceflow for conversation design, ProPet CRM for customer and booking data, and a Vector Database that supports Retrieval-Augmented Generation for more complex or policy-related questions.

The AI assistant functions as a virtual front-desk helper that understands what a caller needs and responds naturally using verified information. It can identify customer intent, extract key details from speech, and deliver personalized answers. When the assistant encounters a question that requires human attention, it transfers the call to a staff member, ensuring that service remains accurate and personal. For this prototype, all interactions will be logged in a Google Sheets database to measure accuracy, response time, and escalation patterns. These insights will help evaluate the assistant's performance and identify opportunities for refinement.

Our project's scope is limited to developing and testing a voice-based prototype that handles inbound calls, FAQs, and booking inquiries through read-only integration with ProPet CRM. It will not process payments, modify customer data, or complete reservations. The focus remains on evaluating communication efficiency, accuracy, and feasibility within a small-business setting.

By automating routine inquiries, PlayStayTion can reduce missed calls, shorten response times, and provide faster and more reliable service. The AI system will enable the business to remain available twenty-four hours a day, giving customers the convenience and confidence they expect without adding additional staff. More importantly, the project examines how small businesses can adopt AI responsibly, using technology to support human performance rather than replace it.

This project demonstrates why AI adoption is important for small, service-driven businesses like PlayStayTion. It provides a forward-looking model for combining efficiency, reliability, and personal service in one system. The interim report presents the vision, goals, and structure of the prototype, all centered on one outcome: building a responsive and customer-focused communication experience for PlayStayTion Pet Resort and Training in North Texas.

OUR UNDERSTANDING OF YOUR BUSINESS NEEDS

Client Name: PlayStayTion Pet Resort & Training

Industry: Pet Boarding, Grooming, and Training

Location: Sadler, Texas

Website: www.playstaytionpetresort.com



ABOUT THE CLIENT

PlayStayTion Pet Resort and Training is a locally owned pet care facility in North Texas that offers boarding, daycare, grooming, and training services for dogs. The business is recognized for its five-star service standards, attentive staff, and a welcoming environment that provides pets with comfort, safety, and engaging daily routines. Its operations run throughout the year, and the facility serves a diverse mix of new and returning customers who rely on timely communication for service inquiries, reservations, and general information. Strong community ties and a commitment to personalized care have contributed to a loyal customer base, making effective communication an essential component of the overall service experience at PlayStayTion.

BUSINESS NEEDS

As customer demand has grown, the volume of incoming calls and routine inquiries has increased significantly. Staff members often balance customer phone calls with in-person responsibilities such as check-ins, pet supervision, and grooming coordination. This creates periods when calls may go unanswered or responses are delayed. These communication gaps affect customer satisfaction by limiting access to basic information, boarding availability, and appointment inquiries. They also influence operational efficiency since the staff must frequently interrupt ongoing tasks to respond to repetitive questions. As a result, the current communication model places strain on both employees and the business workflow.

The primary business need is to establish a more consistent and reliable communication process that ensures customers receive timely and accurate information. PlayStayTion requires a method to manage high-volume or repetitive inquiries without overburdening staff, while still maintaining the personalized service that customers expect. A dependable communication system is necessary to reduce missed opportunities, support staff productivity, and reinforce the overall customer experience that the brand is known for.

From a business perspective, the scope of this project focuses specifically on improving the communication workflow. This includes understanding the types of customer inquiries the business receives, identifying where bottlenecks occur, and determining how consistent information delivery can be achieved. The project does not modify pricing, service policies, staffing models, or facility operations. Instead, it concentrates on designing a communication framework that enhances responsiveness and supports PlayStayTion's long-term growth and service quality.

PROJECT GOALS AND OUTCOMES

GOAL OF THE PROJECT

The goal of this project is to design and evaluate a feasibility prototype of an Agentic AI Voice Assistant that supports customer communication at PlayStayTion Pet Resort and Training. The project aims to understand how artificial intelligence can assist with managing routine inbound calls, answering common service-related questions, and guiding customers through basic booking inquiries. A primary objective is to determine whether AI can improve responsiveness and reduce communication bottlenecks while maintaining the business's existing service standards and operational procedures.

The project also seeks to explore how conversational AI can complement staff efforts by handling repetitive inquiries, improving consistency in information delivery, and reducing missed-call occurrences during busy or after-hours periods. By studying interaction patterns and system performance, the project will identify how an AI-assisted communication model could enhance customer access to information, support staff productivity, and contribute to a more predictable communication workflow. The prototype further aims to generate insights into the conditions and safeguards required for responsible AI adoption within a small-business environment.

The project also aims to establish a clear foundation for future system development by identifying which elements of customer communication are most suitable for automation and which require human judgment. By defining these boundaries, the project will help PlayStayTion plan for long-term integration of AI tools in a way that aligns with its customer service philosophy and operational workflow. This phase will focus on feasibility and design validation, ensuring that any recommended next steps are grounded in evidence, practical considerations, and the unique needs of the business.

INTENDED OUTCOMES

The project is expected to produce outcomes that demonstrate whether automated voice interaction can improve communication efficiency and customer experience at PlayStayTion. The prototype will help determine if AI can reduce call wait times, lower missed-call rates, and provide reliable answers to common questions and booking inquiries. It will also show how well the system recognizes when human support is needed and how smoothly it transitions calls to staff when required.

Another intended outcome is the creation of structured interaction data that captures details such as intent patterns, confidence levels, and escalation triggers. This dataset will support performance evaluation and provide direction for future refinements. The project also aims to assess whether an AI-driven assistant can maintain clarity, accuracy, and consistency in customer communication. Overall, the outcomes will help PlayStayTion understand the benefits and limitations of AI-assisted communication and determine whether the approach can be expanded to broader use in the future.

PROJECT SCOPE, CONSTRAINTS, AND ASSUMPTIONS

SCOPE OF THE PROJECT

The scope of this project is centered on creating and evaluating a feasibility prototype of an Agentic AI Voice Assistant designed to support customer communication at PlayStayTion Pet Resort and Training. The work focuses specifically on the early inquiry stage of customer interaction, where timely access to information and consistent responses matter most. This project explores whether a conversational AI system can reliably recognize caller intent, provide accurate service information, and assist customers with basic inquiry needs in a controlled, low-risk environment. The prototype is intended as a learning and exploration tool rather than a production system. It allows the team to study interaction patterns, test natural-language capabilities, and assess how AI may complement staff efforts without altering existing business processes. By concentrating on feasibility and user experience, the project aims to generate meaningful evidence that will help PlayStayTion consider whether expanded AI integration would be valuable in future phases.

IN SCOPE

- 🐾 Developing a feasibility prototype of the AI voice assistant
- 🐾 Handling inbound voice calls for general information and basic booking inquiries
- 🐾 Identifying caller intent and retrieving information from knowledge bases
- 🐾 Accessing availability information through read-only ProPet CRM integration
- 🐾 Logging interactions for performance evaluation
- 🐾 Escalating low-confidence or complex queries to staff

OUT OF SCOPE

- 🐾 Completing or modifying bookings
- 🐾 Writing data to the ProPet CRM system
- 🐾 Processing payments or transactions
- 🐾 Adding communication channels such as chat, SMS, or mobile apps
- 🐾 Deploying the solution as a production-ready system
- 🐾 Automating grooming or training assessments
- 🐾 Conducting outbound communications

PROJECT DELIVERABLES

Deliverable Name	Deliverable Description	Due Date
AI Voice Agent Prototype	A feasibility prototype demonstrating intent detection, FAQ responses, availability checks, and staff escalation.	October 24, 2025

Interaction Log Dataset	Structured logs of customer interactions including intent, response patterns, and confidence scores for evaluation.	November 4, 2025
Communication Workflow Review	A documented assessment of inquiry types, communication gaps, and prototype-supported improvements.	November 15, 2025
Analytics and Performance Summary	A performance evaluation report based on interaction logs, focusing on accuracy, latency, and escalation outcomes.	November 21, 2025
Architecture Diagram and Technical Report	Complete documentation of the prototype architecture, data flow, agent design, and implementation logic.	December 9, 2025

Table 1. Project Deliverables and Due Dates

LIMITATIONS, CONSTRAINTS, AND ASSUMPTIONS

LIMITATIONS

- 🦋 The system is a feasibility prototype and not a production-level solution, so performance, scalability, and long-term reliability are not fully representative of a deployed system.
- 🦋 The prototype can only read data from ProPet CRM and cannot modify customer or booking records.
- 🦋 Test accuracy depends on the quality and variety of sample utterances used during evaluation.
- 🦋 RAG responses are constrained by the completeness and clarity of the documents provided by PlayStayTion.

CONSTRAINTS

- 🦋 ProPet API restrictions prevent write operations and limit booking-related functionality.
- 🦋 The project must be completed within the timeframe and resources available for the academic course.
- 🦋 The team is limited to documents, policies, and information supplied by PlayStayTion.
- 🦋 Staff availability influences the ability to test escalation scenarios and human handoff flows.

ASSUMPTIONS

- 🦋 Test users will speak naturally and clearly when interacting with the voice assistant.
- 🦋 Knowledge base materials will be maintained and kept accurate by PlayStayTion staff.
- 🦋 The ProPet CRM system will provide reliable availability information through read-only API access.
- 🦋 Project evaluation will focus on communication feasibility rather than full operational automation.

PROJECT APPROACH

WHY AGENTIC AI

Agentic AI is used in this project because it provides a structured way for an autonomous system to understand customer intent, make context-aware decisions, and take appropriate actions during live voice interactions. PlayStayTion receives many routine questions that follow predictable patterns, such as inquiries about hours, pricing, boarding requirements, and availability. Agentic AI offers a flexible framework for managing these interactions by combining natural language understanding, decision logic, and controlled actions within defined boundaries. This approach allows the assistant to respond consistently, escalate when needed, and maintain alignment with PlayStayTion's service expectations. The agentic model is especially suitable for a feasibility prototype because it enables the team to test how well AI can support communication tasks without altering existing workflows or introducing risk to business operations.

AGENTIC AI ARCHITECTURES

The project uses a modular agent architecture that divides the system into specialized components responsible for intent detection, FAQ responses, complex query handling, booking information retrieval, and customer response generation. Each agent operates within a defined role, allowing the system to route interactions to the most appropriate component. Voiceflow provides the agent-building environment, while Twilio enables speech-to-text and text-to-speech processing for voice interaction. Information retrieval is supported by two knowledge bases: a short-answer FAQ set and a long-form document set designed for contextual responses. The architecture incorporates read-only integration with the ProPet CRM system to retrieve real-time availability information when customers ask about bookings. All agents are orchestrated within a controlled workflow that ensures accuracy, transparency, and human oversight through staff escalation whenever AI confidence is low or the question requires human judgment.

DATA SOURCES AND INTEGRATION REQUIREMENTS

The project relies on several structured and unstructured data sources. Customer service information, such as frequently asked questions, pricing details, policies, and general service descriptions, is uploaded into the FAQ knowledge base. Longer documents, including training instructions, boarding policies, grooming guidelines, and operational notes, are stored in the long-form knowledge base used for retrieval-augmented responses. Real-time availability data is accessed through the ProPet CRM system using read-only API calls that provide information on service openings and booking capacity. Twilio provides transcript data through speech-to-text processing, which is then passed into Voiceflow's intent classifier. All interactions are logged in a Google Sheets database through API calls for evaluation. The integration requirements include secure communication between Voiceflow and the data sources, accurate handling of user input, and reliable API responses from ProPet to ensure that availability information is consistent and up to date.

SUCCESS METRICS AND EVALUATION PLAN

The evaluation of the feasibility prototype will focus on how effectively the AI Voice Assistant supports PlayStayTion's customer communication process. The goal is to determine whether the system can deliver accurate information, maintain a natural and clear interaction flow, and escalate calls appropriately when human assistance is required. The prototype will be tested using sample utterances that represent typical customer questions, and all interactions will be logged for analysis. The evaluation plan will use a combination of technical performance measures and qualitative indicators to assess whether the system is reliable, usable, and aligned with PlayStayTion's communication needs.

METRICS

- 👤 Intent accuracy across all major inquiry categories
- 👤 Response relevance and correctness based on predefined service information
- 👤 Latency, measured as the time between user input and system response
- 👤 Escalation accuracy, including whether the system routes calls correctly when confidence is low
- 👤 Consistency of information delivery across repeated and similar test cases
- 👤 System stability during multiple test sessions
- 👤 User experience indicators, such as clarity of responses and ease of interaction
- 👤 Completeness and structure of interaction logs for monitoring, review, and improvement

SOLUTION REQUIREMENTS

PROJECT STAKEHOLDERS

The PlayStayTion AI Voice Assistant involves several stakeholder groups who either interact directly with the prototype or support its operation. Each stakeholder contributes to the system's functionality, maintenance, and evaluation during the feasibility phase. Pet owners represent the primary users, as they initiate calls for information about boarding, daycare, grooming, and training services. The AI Voice Assistant, built using Twilio and Voiceflow, acts as the automated interface that processes speech, identifies intent, retrieves information from the knowledge bases, and provides responses. Front desk staff serve as human support for escalations when questions require personal judgment or when the AI's confidence is low. Administrative staff maintain service documents and knowledge base materials to ensure accurate responses. The ProPet CRM system functions as the central data provider for service availability and customer information. The project team manages the design, development, and testing of the prototype. Together, these stakeholders shape the functionality, relevance, and evaluation of the AI system.

Stakeholder	Role	Responsibilities
Pet Owners (Customers)	End users	Call PlayStayTion for questions, service information, and booking-related inquiries.
AI Voice Assistant (Twilio + Voiceflow)	Automated system	Handle incoming calls, convert speech to text, detect intent, retrieve information, and respond.
Front Desk Staff	Human support	Manage escalations, handle special cases, and provide assistance for complex questions.
Admin or Manager	System operator	Maintain the FAQ and long-form knowledge bases, review interaction logs, and oversee accuracy.
ProPet CRM System	Data provider	Supply service availability, customer information, and booking data through read-only API access.
Project Team	Designers and evaluators	Develop the prototype, manage integrations, evaluate performance, and document findings.

Table 2. Stakeholder Roles and Responsibilities

BUSINESS REQUIREMENTS – STAKEHOLDER ANALYSIS

Understanding stakeholder needs is essential for shaping the business requirements of the AI Voice Assistant prototype. Pet owners require timely responses, clear information, and predictable communication, especially during peak hours or after business hours when staff may be unavailable. They expect accurate details about boarding rules, grooming requirements, service hours, pricing, and availability. Front desk staff need support in managing repetitive inquiries so they can focus on in-person care, customer check-ins, and operational tasks. They require a system that escalates calls appropriately and does not disrupt existing workflows.

Administrative staff depend on an AI system that reflects current service policies, meaning the knowledge bases must be accurate and easy to update. They also need transparent logs that allow them to monitor communication quality. PlayStayTion management requires evidence that AI can improve service efficiency, reduce missed calls, and enhance consistency without compromising the business's personal service philosophy. The project team needs access to service information, policies, and operational details to build a prototype that accurately mirrors customer interactions. The ProPet CRM system must provide reliable availability data to support booking-related inquiries. This analysis ensures that business requirements reflect the needs of all parties and supports an AI solution that is accurate, responsible, and aligned with PlayStayTion's operational goals.

SOLUTION DESIGN

AGENTIC SOLUTION ENVIRONMENT

For the business solution, the overall environment for working is studied to outline system requirement and architecture plan. It is noted that there are three major divisions to work to enable the prototype: Workflow dashboard, Platform settings, and External Telephony Integration. For Telephony integration, Twilio account is created, and a number is bought to test the working of integration. It is not assigned at early development stage. The Voiceflow platform settings are thoroughly explored such as modality, behavior, testing and evaluation, etc. The workflow is the main ground for creating voice agents. After creating a new project, it is here that agent architecture is designed and iteratively changed after testing agents in development stage. After finalizing the workflow, the workflow in production environment is published to be assigned to already bought Twilio phone number.

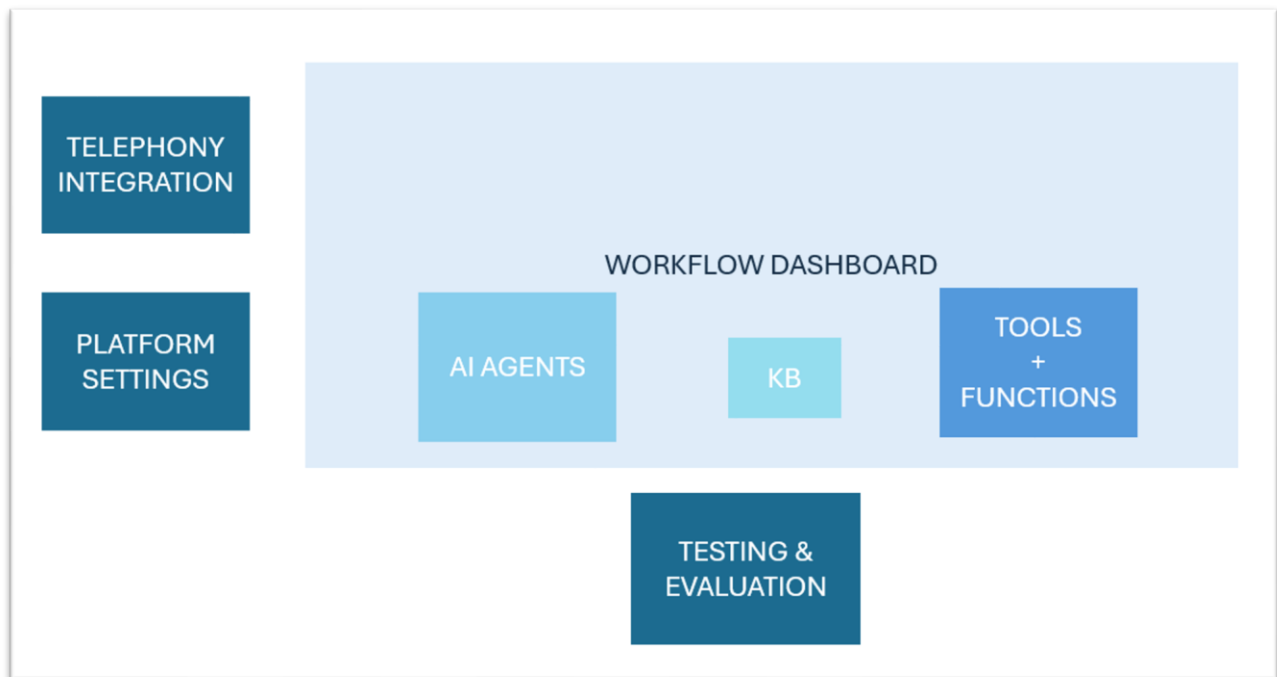


Figure 1. Core Environment Components for the PlayStayTion AI Voice Assistant







ARCHITECTURE

The workflow architecture for the conversation design consists of combination of routing and specialized agents. To keep the architecture simple, a greeting agent receives a call gathering initial user details and transferring them to customer logs using an integration step to an external database (Google sheet used for prototype). After which, the user details such as reason for calling is fed to

a Supervisor Agent where the intent classification occurs for routing the query to the appropriate specialized agent like FAQ agent, Booking Info agent, Complex Info agent, and Human Escalation agent. Each specialized agent is sent to a path to return to supervisor agent for addressing any newer queries. In case of Human escalation agent, call forward and update customer log actions are triggered depending upon customer enquiry.

AGENT ROLES

Every agent performs a distinctive task to optimize tasks.

-  **Greeting Agent:** Gathers information from the user about reason for calling and basic details like name and pet name. This agent only ensures that user is politely handled to understand the context of query.
-  **Supervisor Agent:** This agent understands query further to classify the intent correctly and route to the appropriate specialized agent.
-  **FAQ Agent:** Responds to short queries from the knowledge base in KB_FAQ folder.
-  **Complex Info Agent:** Responds to queries that do not have direct answers from KB_FAQ and require referring to more detailed documents from KB_LONG folder. This agent can provide information in depth.
-  **Booking Info Agent:** Responds to queries which need real time integration with the Propet CRM regarding booking related queries.
-  **Human Escalation Agent:** this agent is routed when the supervisor agent cannot classify the intent to be routed to any of the above-mentioned agents, or when user explicitly mentions to talk to a human employee. This agent helps to retain customers and prevents AI from hallucinating and giving incorrect responses. It guides the route to Call forwarding or storing the user query depending upon time of the call and Office hours.

INTEGRATION

The integrations conducted for making the voice agent prototype includes Twilio and External API. Both integrations are possible within Voiceflow platform. For Twilio Integration, a Twilio account is created and under the Interface dashboard, telephony is selected to fill the necessary details. For the external API database, Google sheet within the agent block of Voiceflow is used that asks for spreadsheet url, range, values, and capture response along with LLM description.

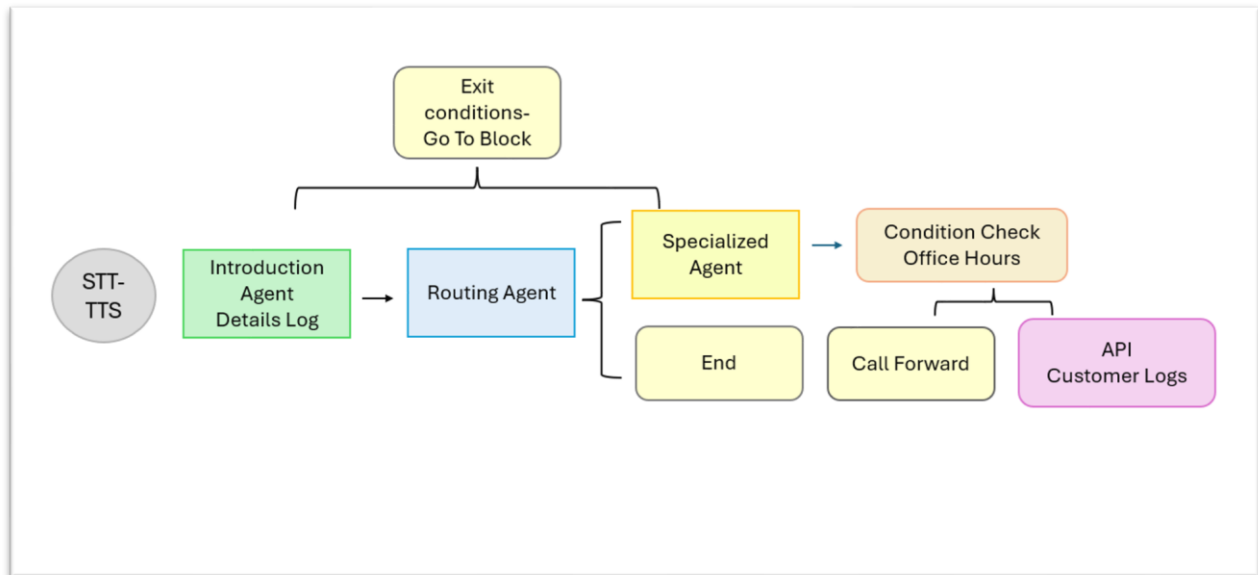


Figure 2. Agent Workflow and Integration Flow for the AI Voice Assistant

DATA ACCESS

KNOWLEDGE BASE STRUCTURE

Knowledge base is compiled with two folders names KB_FAQ and KB_LONG. The first folder contains the frequently asked questions to answer user queries quickly. For the other folder, more detailed documents are added such as all extra information from the website. Keeping two separate folders help the specialised agents to gather information depending upon the complexity of the query.

KNOWLEDGE BASE SETTINGS

It is significant to maintain an optimal setting for the various factors that impact the retrieval augmentation for the agents. Therefore, the Model chosen for retrieval chunks from Knowledge base is GPT 4, since it is intelligent and had reduced latency and doesn't consume as many token credits as newer models. Temperature is maintained at 0.16, chunks limit to 3, max tokens to around 2000 to retrieve important information.

Import file [X]

File(s)

KB_LONG.docx

0.03mb file successfully uploaded.

LLM chunking strategy

Select strategy (optional) [v]

- ☐ **Smart chunking**
Breaks up your data into logical sections grouped by topic. Best for complex documents with varied topics.
- ☐ **FAQ optimization**
Creates sample questions that each section could answer. Best for building FAQs.
- ☐ **Remove HTML and noise**
Cleans up messy website formatting to make text easier to group. Best for website or markdown formatting.
- ☐ **Add topic headers**
Adds brief summaries at the start of each section. Best for long documents needing context.
- ☐ **Summarize**
Keeps only the key points and removes filler content. Best for dense, lengthy content.

Figure 3. Chunking strategy of Knowledge base

GROUNDING

CONDITION LOGIC & API INTEGRATION

To maintain the Agent's response to accurate data, a JavaScript ensures that the office hours are checked and based on the condition, user is either forwarded the call if within office hours or their query is logged into database to be checked by staff the next working day. Although outside scope of the prototype, integration to the Propet CRM provides real world access to the reservation, boarding & training bookings for answering users booking related queries.

MEMORY USAGE

To maintain context awareness, conversation history {vf_memory} is added as an asset in the prompt of the agents so that the AI agent refers to prior utterance to respond like humans. It helps to add personalization and avoid repetitive conversations.

ACTIONS

Built in actions within the platform are used to perform certain operations.

Go To Block: It is an inbuilt action within voiceflow platform that enables the agent to return to an agent. All specialized agents are sent back to supervised agent using this action.

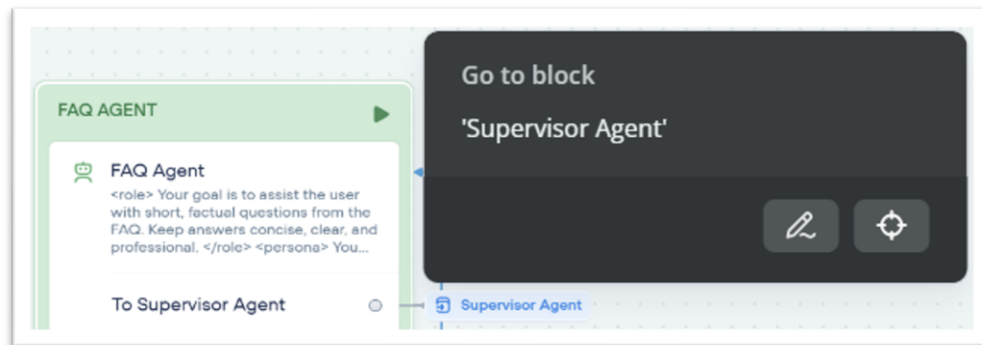


Figure 4. Shows Go to block for a specialized agent

End: The end action helps the agent to directly end the active state. It is added as an exit condition on the supervisor agent as well as to the human escalation agent.

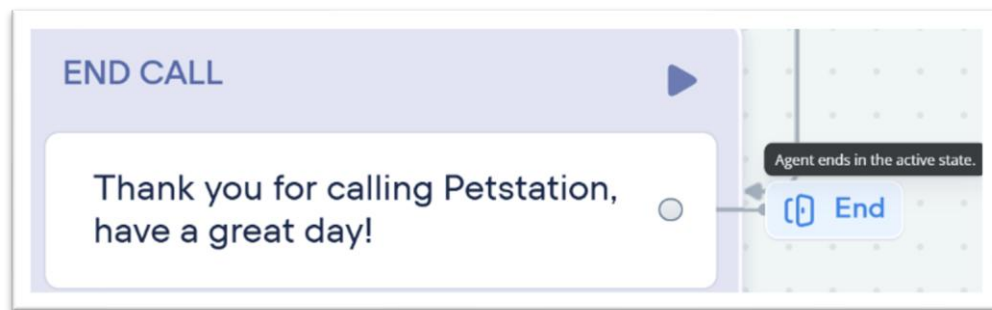


Figure 5. Shows End action after identifying intent

Forward Call: This action enables the call forwarding to the concerned person. In this case, the call is forwarded to the human staff member on request of the customer, after checking the condition of office hours.

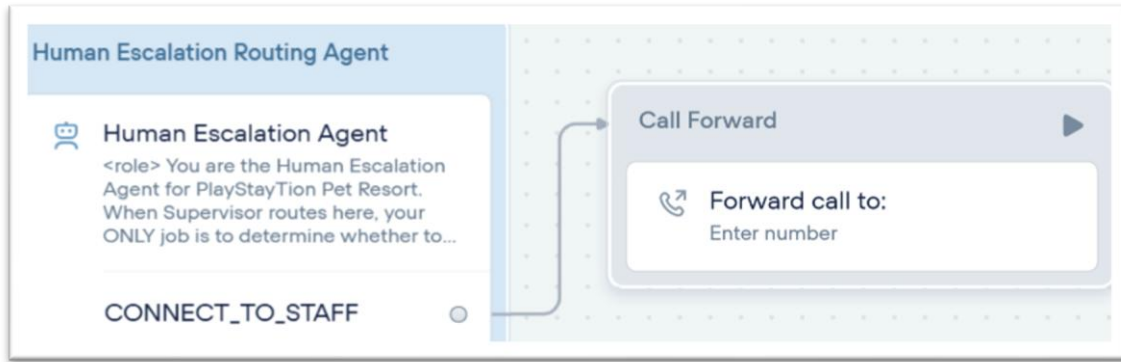


Figure 6. Shows Call Forward action to connect to a number

TOOLS

Tools are required to access external datastores and add logical conditions for various agents.

API Integration: Google sheet is added as a tool API for adding new rows for every caller that makes a call. It is connected to the local file using spreadsheet url, range, values, and capture response. In certain cases, other than url nothing needs to be filled since the LLM instructs to capture details from user reply.

The screenshot shows the 'Integration tool' configuration interface. At the top, a dropdown menu is set to 'Add to sheet'. Below this, the 'LLM description' field contains the text: 'Append a new customer log after the Greeting Agent collects the caller's reason, name, and pet name.' To the right of this field is a 'Reset' button. Below the description, the 'Input variables' section lists three variables: 'spreadsheetURL' (Value provided), 'range' (Agent collect), and 'values' (Value provided). Each variable has a corresponding toggle switch that is currently turned on. At the bottom, the 'Capture response' section has a play button and a plus sign. A green status bar at the very bottom indicates 'Google Sheets is connected.' with a 'Manage' button to its right.

Figure 7. Shows Integration API Tool for adding new sheet

Function-Javascript: A JavaScript function from tools is added to create a conditional logic to capture time of the user's call and to estimate whether it is within or outside office hours. The respective code is added in the appendix for more information.

PROTOTYPE DESIGN AND IMPLEMENTATION ENVIRONMENT

For the initial stage of the Project, all settings are adjusted before start designing the conversation design in the development Environment.

Platform Setup: For setting the platform, a voiceflow account is created.

Modality: Under interface settings, modality chosen is voice.

Voice Settings: Under behaviour settings, it is ensured that the chosen voice is natural and appropriate for answering questions by an AI pet resort receptionist. The voice provider company (Eleven Labs), voice type, and model (Eleven Turbi v2.5) are selected. Background voice and audio cue is added to make sure there is no awkward silence when the user calls. Other features such as speed (0.93), stability (0.46) and similarity boost (0.75) are adjusted and kept at a balanced score on the range provided.

Workflow Design: After completing the settings, conversation design is initiated in workflow dashboard.

- 👤 **Agent Block Design:** Major architecture plan is outlined first with all agent blocks lined up. Three major agents are planned in a hierarchy; Greeting Agent-Supervisor Agent-Specialized Agents, and later exit conditions are added to each agent to connect them cohesively for smooth interaction.
- 👤 **Prompt Instructions:** The prompt instructions within agents are detailed to include role, persona, behaviour, actions, exit conditions, conversation flow, conversation memory {vf_memory}, and certain rules specific to specialized agents.
- 👤 **Model Settings:** For maintaining optimum intelligence and latency while keeping token credit economical, model selection is planned based on agent use. For example, for greeting agent and supervisor agent since it doesn't need to retrieve information from knowledge base but majorly needs to understand customer query correctly, therefore in such cases temperature of model is more deterministic (0.2) and less creative. While for other specialized models, more creativity is added so temperature is increased (0.3).
- 👤 **Toggle on buttons:** Other than the Instructions, Knowledge base is toggled on for three agents- FAQ, Complex Info and Booking Info agent.

- 👤 **Tools:** As mentioned above, API integration for adding and updating rows in Google sheet is enabled for logging customer data.
- 👤 **Functions:** For enabling conditional logic, JavaScript function is coded to store variables like isOfficeHours. It guides the human escalation agent to decide whether to forward a call to human staff or log user query in external API.

TESTING DURING DEVELOPMENT ENVIRONMENT:

Multiple tests are conducted to evaluate if the voice agent is answering calls as planned. Priority metrics were as follows:



Figure 8. Satisfaction and deflection testing scores of the workflow in development stage

- 👤 **Routing Accuracy:** Major concern is whether the supervisor agent can route the call appropriately to the right agent.
- 👤 **Voice Behaviour:** The speech speed and tone behaviour of the agent is also closely monitored, and observations are noted for improvement. The expectation is to create an agent which speaks in a natural and polite tone at a speed that everybody can understand without any discomfort.
- 👤 **Performance of tools & Functions:** Another important feature that is evaluated is performance of the tools and functions.
- 👤 **Observation & Fixes:** It is noted that handled variables and logging the variables related to user details to external API presented challenges. Some unsatisfying attempts included capturing one variable per message and capturing all variables in a single entity and later extracting each using JavaScript. Connecting JavaScript to agent poses another challenge since objects cannot be directly linked to ai agents on voiceflow platform. The voice behaviour of the agent is evaluated to cause many issues such as unnecessary repetitions, or unpredictable call endings or memorizing previous caller's history.

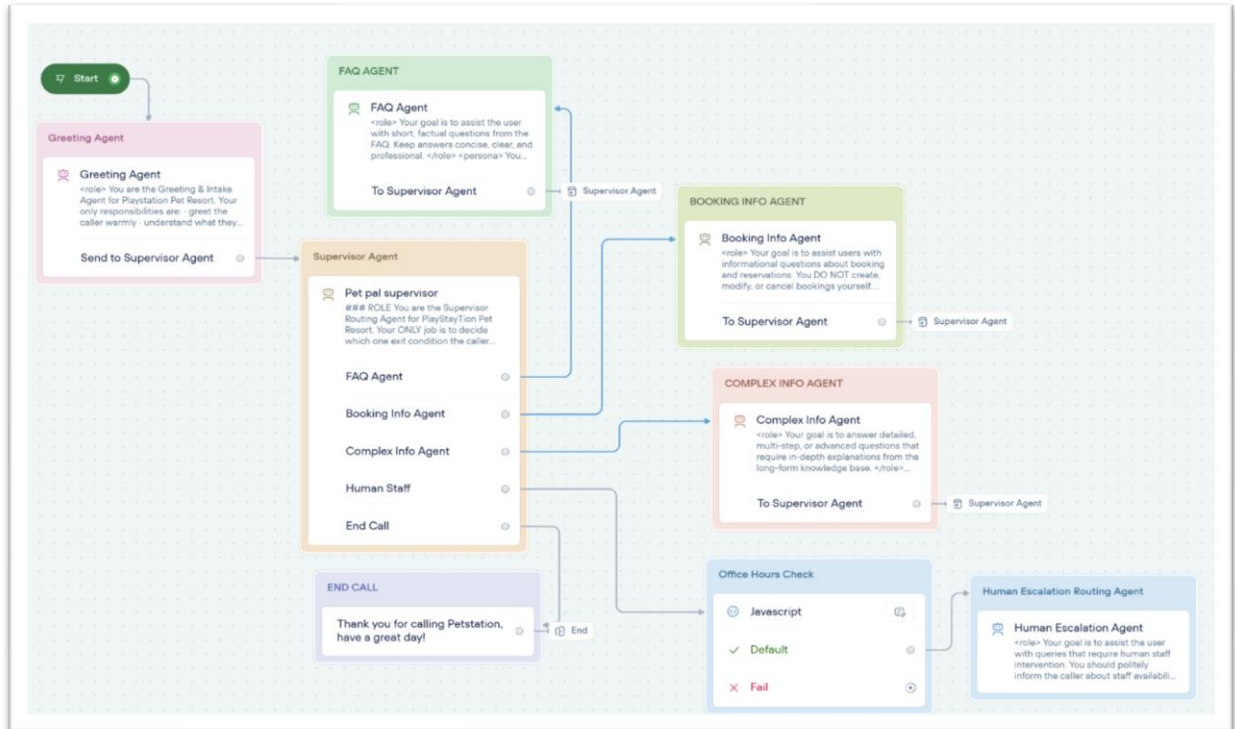


Figure 9. Agent Architecture for the Voice Assistant Prototype

PUBLISHING WORKFLOW FOR PRODUCTION ENVIRONMENT:

After completing the testing and observing the evaluation, the workflow is finally published in production environment to be integrated to Twilio telephony to make the prototype working.

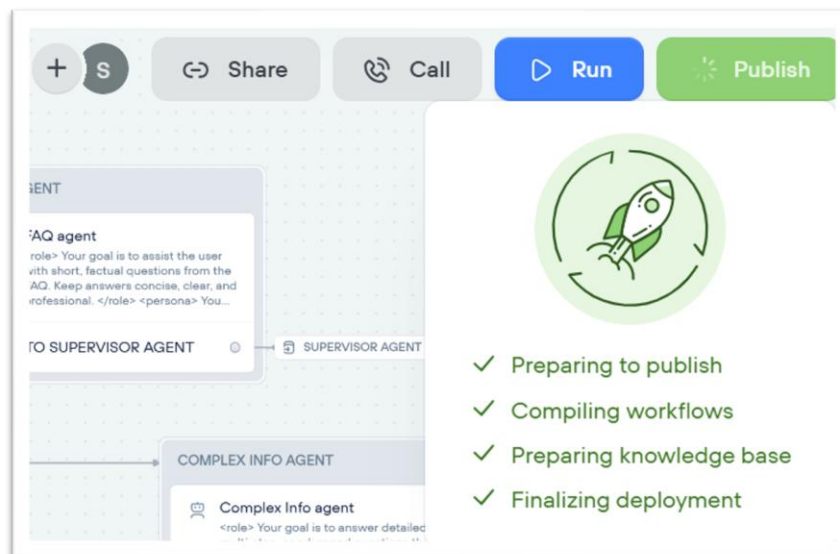


Figure 10. Publishing the workflow for Production environment

SOLUTION IMPLEMENTATION

IMPLEMENTATION ROLLOUT PLAN

Only 30 percent of the planned system has been completed as an initial prototype. This early-stage prototype includes the basic conversational flow, the preliminary agent structure, and read-only data integrations. The remaining 70 percent of the solution will be completed during the full implementation phase, which will include advanced automation, improved reliability, enhanced integration, and expanded feature coverage. The rollout will follow a phased, systematic approach to ensure that the final system is accurate, stable, and production-ready. After implementation, structured support will be provided to maintain performance and update system content as PlayStayTion's requirements evolve.

PHASE DESCRIPTION

Phase 1: Requirements Finalization.

The team will finalize all service details, clarify remaining FAQ gaps, refine booking-related workflows, and document escalation requirements. This ensures that the full implementation reflects PlayStayTion's actual operational needs.

Phase 2: Architecture Expansion.

The prototype will be expanded into a complete agent architecture with improved routing, more advanced intent handling, and full flow coverage.

Phase 3: Knowledge Base Completion

All relevant business documents, grooming guidelines, policies, price lists, and operational instructions will be added to the knowledge base to improve retrieval accuracy.

Phase 4: Full Integrations

The system will be expanded beyond read-only mode to support deeper CRM interactions, booking automations, and optional future integrations.

Phase 5: Logic and Safety Automation

Office-hour logic, error-handling behaviors, escalation rules, and fallback strategies will be fully implemented to ensure predictable and safe system responses.

Phase 6: End-to-End Testing and Optimization

The completed system (100 percent) will undergo comprehensive testing, including performance checks, accuracy validation, user-testing scenarios, and optimization for production deployment.

Phase / Step	What Will Be Completed	Purpose
Requirement Understanding	Finalization of all service details, FAQ gaps, and booking flows	Ensures complete accuracy in the final build
Architecture Expansion	Enhancement of the current 30% prototype into a full agent system	Supports full routing and interaction capabilities
Knowledge Base Completion	Expansion of KB_FAQ and KB_LONG with all service documents	Provides richer and more accurate responses
Full Integrations	Addition of booking automation and CRM-level integration (future scope)	Improves end-to-end customer experience
Logic and Automation	Implementation of office-hour logic, escalation rules, and error handling	Ensures reliability and response safety
System Testing and Optimization	Testing the remaining 70% build and validating full system behavior	Achieves production-level quality

Table 3. Implementation Plan

POST-IMPLEMENTATION SUPPORT PLAN

After deployment, the system will require continuous monitoring and maintenance to ensure that performance remains stable and aligned with PlayStayTion’s ongoing operational needs. Support activities will focus on system performance, content updates, and continuous optimization to adapt to new customer patterns and updated business policies.

SUPPORT ACTIVITIES

Performance Monitoring: The team will review system logs, evaluate intent accuracy, analyze escalation frequency, and track response latency to identify areas for improvement.

Knowledge Base Maintenance: All service information, grooming and boarding rules, operational guidelines, and pricing updates will be periodically reviewed and updated to ensure accuracy.

Monthly Optimization: Regular refinements will be applied to improve prompt quality, reduce repetition, enhance grounding, and strengthen fallback logic based on real-world usage.

Future Enhancements: Optional expansions may include two-way CRM updates, SMS and chat-based support, payment processing, reservation automation, and multi-language capabilities.

ECONOMIC FEASIBILITY

The AI system is extremely cost-effective because it uses low-cost platforms and replaces repetitive manual tasks. Compared to hiring an additional staff member, the AI assistant costs much less while providing 24/7 availability.

Platform	Monthly Cost	Notes
Twilio	\$1.50	Phone number + usage
Voiceflow	\$60	AI agent conversation builder
Google Sheets	\$0	Used for call logs
Total Monthly Cost	\$61.50	\$738 per year

Table 4. Monthly Cost Breakdown of AI System Components

Note: The final cost may change depending on the GPT model selected for the full implementation. More advanced models may provide better accuracy but will also increase monthly token usage, leading to cost variations. This cost variation is expected and will be calculated during the final deployment stage

Hiring a front-desk staff member is significantly more expensive. The AI assistant saves thousands each year and gives a strong return on investment.

Option	Monthly Cost	Yearly Cost
Human Staff	~\$3,120	~\$37,440
AI Voice Assistant	~\$61.50	~\$738
Yearly Savings	-	\$36,702

Table 5. Cost Comparison Between Human Staff and AI Voice Assistant

Note: The cost of a human front-desk employee was estimated using a standard Texas wage of \$15/hour, which equals \$120 per day for an 8-hour shift. With approximately 26 working days per month, this results in \$3,120 per month, or \$37,440 per year. This calculation illustrates the high ongoing cost of staffing compared to the AI system.

Using these numbers, the return on investment is extremely high.

Metric	Value
Annual Savings	\$36,702

Annual AI Cost	\$738
ROI	4,874%

Table 6. Return on Investment (ROI) Summary

PROJECT LIMITATIONS

Even though the prototype works well, it has some limitations. It cannot complete bookings, and accuracy depends on the documents provided. Real customer calls may also vary based on noise, accents, and unclear speech.

Limitation	Description
Prototype only	Not built for full business deployment
CRM Credentials	We did not check the CRM process as we did not request from client
Depends on document quality	Poor documents reduce answer quality
Limited speech testing	Not tested with all accents/noisy calls
Human escalation depends on staff	Staff must be available to take calls

Table 7. Prototype Limitations and Descriptions

PLATFORM SETTINGS FEASIBILITY CONSIDERATIONS

Settings on the Platform are adjusted with the following Considerations- Intelligence, Latency, and Tokens Used. For Intent Classification in supervisor agent Instructions, the model chosen is optimally selected as G.P.T 4.1 mini. As compared to all earlier and later models (including G.P.T. 5.1 mini), the chosen model proves to be effectively intelligent to classify intent and yet produce faster results due to reduced latency. Also, it reflects higher intelligence than GPT 4.1 nano and GPT 4.0 mini but less latency than GPT 4.1, GPT 4.0. Therefore, feasibility wise, G.P.T 4.1 mini is a very appropriate model for our use case.

Other setting features such as priority processing and LLM fallback are enabled for the prototype to test the Voice agent's ability to interact effectively, though it is known that these can cause higher cost but it's a tradeoff that is being intentionally taken into consideration since these adjustments ensure better quality and continuity during outages.

APPENDIX: PROTOTYPE

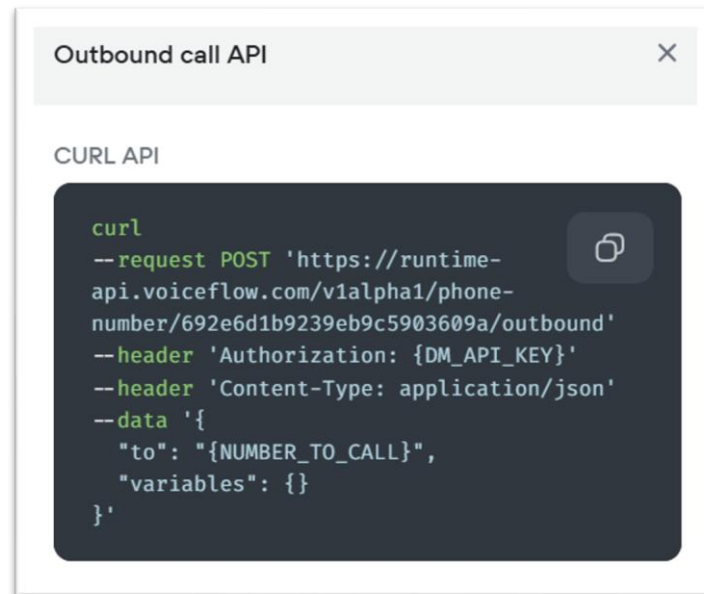


Figure 12. Outbound Call API Integration for Twilio

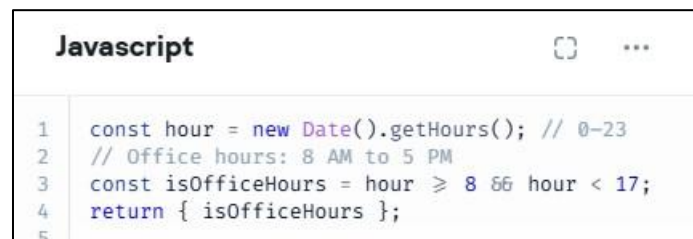


Figure 13. Conditional Logic code for determining Office Hours

REFERENCES

1. Bottaro, J. P., & Ramgopal, K. (2024, April 25). *Musings on building a generative AI product*. LinkedIn. <https://www.linkedin.com/blog/engineering/generative-ai/musings-on-building-a-generativeai-product>
2. Umbral. (2025, August 4). *How to evaluate production AI agents (NEW UPDATE) -- Voiceflow* [Video]. YouTube. https://www.youtube.com/watch?v=10dzi_Hdih4
3. Voiceflow. (2023, May 16). *Introducing Voiceflow's GPT-powered AI Builder*. Voiceflow. <https://www.voiceflow.com/pathways/introducing-voiceflows-gpt-powered-ai-builder>
4. Eddie Chen | Chat + Voice AI. (2025, March) *Master Voiceflow in 2 hours: Complete beginner's guide for 2025* [Video]. YouTube. <https://www.youtube.com/watch?v=sq2uqeyqL3A>
5. Blackader, B., Buesing, E., Amar, J., Raabe, J., Mehndiratta, M., & Gupta, V. (2025, March 19). *The contact center crossroads: Finding the right mix of humans and AI*. McKinsey & Company. <https://www.mckinsey.com/capabilities/operations/our-insights/the-contact-center-crossroads-finding>
6. Voiceflow. (n.d.). *GDPR compliance*. Voiceflow. <https://www.voiceflow.com/legal/gdpr#Learn-more>