**ENHANCING CUSTOMER SUPPORT WITH REAL-TIME** **RETRIEVAL-AUGMENTED GENERATION AND HUMAN ESCALATION**

**Abstract**

In today's fast-paced digital economy, customer service plays a critical role in shaping user satisfaction and business performance. However, traditional chatbot systems often fall short when handling complex, multi-step, or domain-specific customer inquiries due to their limited contextual understanding and static response capabilities. This paper presents the design and implementation of a real-time, embeddable customer service solution powered by Retrieval-Augmented Generation (RAG) and supported by human-in-the-loop escalation. The system leverages OpenAI’s large language models for natural language processing, augmented with contextually relevant information retrieved from user-defined intent files. Built with FastAPI, Centrifugo, Celery, and ReactJS, the solution enables low-latency communication, seamless chatbot-human transition, and dynamic analytics dashboards. The architecture supports various integration methods, including iframe embedding, web components, and shareable URLs. Performance evaluation highlights the system’s ability to enhance response accuracy, reduce escalation delays, and deliver actionable insights into customer concerns. By combining AI-driven automation with human oversight, the system offers a scalable, customizable, and intelligent framework for improving customer service efficiency and experience across industries.

**Keyword:** Retrieval-Augmented Generation (RAG), Customer Service Automation, Human-in-the-Loop (HITL), Chatbot Architecture, Large Language Models (LLMs), Real-Time AI Systems

**1. INTRODUCTION**

In the digital era, customer service has emerged as a critical differentiator for business success, especially in highly competitive and consumer-driven markets. Traditional support systems, typically reliant on phone calls, emails, and manual human agents, often fail to meet modern expectations for immediacy and personalization. Long wait times, inconsistent service quality, and the inability to scale during peak demand periods have left customers dissatisfied and businesses under pressure to innovate [5].

To address these limitations, businesses have increasingly turned to artificial intelligence (AI), especially AI-powered chatbots. These systems leverage natural language processing (NLP) to simulate human conversation and automate responses to customer inquiries. However, despite advances in AI, many chatbot systems still suffer from shortcomings such as limited contextual understanding, poor handling of multi-turn dialogue, and inflexible, rule-based logic [1]. These limitations can result in miscommunication, reduced customer trust, and the eventual need for human intervention.

The emergence of large language models (LLMs), such as OpenAI’s GPT-3 and GPT-4, has marked a turning point in conversational AI. These models are capable of generating contextually relevant and coherent responses by learning from vast amounts of structured and unstructured data [3], [9]. When integrated with Retrieval-Augmented Generation (RAG), LLMs can enhance response accuracy by drawing upon external sources of domain-specific knowledge during conversation generation. This makes RAG an effective solution for customer service scenarios that demand both fluency and factual accuracy [7].

Despite the promise of such technologies, research indicates that AI alone cannot fully replace the need for human judgment, especially in handling emotionally sensitive or complex customer queries [15], [8]. Human-in-the-loop (HITL) systems, which allow seamless escalation from chatbot to human agents, have emerged as a practical approach to overcoming these challenges. By combining automation with human oversight, HITL systems ensure that customers receive not only instant replies but also empathetic and accurate resolutions when automation reaches its limits.

This paper presents the design and implementation of a real-time, embeddable customer service system that integrates RAG-based chatbot architecture with dynamic human escalation. The solution is built on a modular web-based framework using FastAPI, ReactJS, and Centrifugo for WebSocket-based communication. It allows businesses to embed a customizable AI assistant into their digital platforms and configure it using domain-specific intent files. When the chatbot encounters queries beyond its capability, the system triggers an escalation mechanism that alerts human agents in real time via NovuHQ notifications.

In addition to its core functionality, the system features an analytics dashboard for performance monitoring, response categorization, and conversational summaries. It supports multiple integration formats, including iframe, web component, and shareable URLs, making it suitable for diverse business contexts, from small enterprises to large-scale service providers.

The remainder of this paper is organized as follows: Section 2 reviews related work in chatbot development, RAG models, and human-in-the-loop systems. Section 3 outlines the system architecture and design. Section 4 discusses implementation specifics. Section 5 evaluates the system's performance, and Section 6 concludes with future research directions.

**2. LITERATURE REVIEW**

The rapid evolution of artificial intelligence (AI) and natural language processing (NLP) has significantly influenced customer service systems, enabling businesses to automate responses, improve scalability, and enhance user satisfaction. This section reviews existing literature on AI-powered chatbots, Retrieval-Augmented Generation (RAG), and human-in-the-loop (HITL) mechanisms, identifying their strengths, limitations, and opportunities for innovation.

### ****2.1 Evolution of Chatbots in Customer Service****

Early chatbot systems, such as ELIZA [13] and PARRY [4], were rule-based and relied on pattern matching to simulate conversation. While pioneering, these systems lacked contextual memory and adaptability. Subsequent developments introduced AI-powered chatbots capable of basic intent recognition using decision trees and keyword extraction [2].

In recent years, the integration of natural language processing (NLP) and machine learning has significantly advanced chatbot capabilities. AI-powered platforms such as Google Dialogflow, IBM Watson Assistant, and Microsoft Bot Framework now support intent classification, entity recognition, and conversational flow design [1]. These platforms offer improved response quality but often fall short in handling dynamic, multi-turn conversations or domain-specific queries without substantial manual training and fine-tuning.

### ****2.2 Large Language Models and RAG in Dialogue Systems****

The emergence of large language models (LLMs) such as GPT-3 and GPT-4 has reshaped conversational AI by enabling more fluent, contextually coherent responses. These models, trained on massive corpora, can generate text, summarize content, and answer questions with high degrees of linguistic accuracy [3], [9]. However, their generative nature can also lead to hallucinations or factually incorrect outputs.

To mitigate this, Retrieval-Augmented Generation (RAG) has been introduced as a hybrid approach that combines document retrieval with generation. Lewis et al. [7] propose RAG as a method that retrieves relevant information from an external knowledge base and conditions the generative model on it, thereby improving factual accuracy and contextual grounding. This is especially valuable in customer service, where responses must be consistent with company policies, FAQs, or service documentation.

Several studies have demonstrated the effectiveness of RAG in dialogue systems. For instance, Wu et al. [14] highlight how retrieval-enhanced dialogue models outperform vanilla transformers in task-specific scenarios. However, few implementations of RAG have been designed to operate in real time or integrate seamlessly with web platforms for live customer interaction.

### ****2.3 Human-in-the-Loop (HITL) Approaches in Customer Support****

Despite advances in LLMs and RAG, studies emphasize that fully autonomous systems often fail to handle emotionally sensitive or high-stakes situations adequately [15], [12]. Human-in-the-loop (HITL) models have thus emerged as a pragmatic solution, where automated systems manage routine interactions and escalate complex issues to human agents.

Modern HITL systems include escalation triggers based on sentiment analysis, confidence thresholds, or explicit customer requests [11]. Such systems are employed by platforms like Zendesk and LivePerson to strike a balance between automation efficiency and human empathy. However, many of these implementations lack the modularity and embeddability necessary for widespread adoption by small and medium-sized enterprises (SMEs).

### ****2.4 Real-Time Systems and Embeddable Architectures****

Real-time chatbot systems require robust backend architectures capable of handling concurrent users and asynchronous message processing. Technologies such as WebSockets (via Centrifugo) and task queues (e.g., Celery) are often employed to maintain low-latency performance [6]. Additionally, the rise of API-first development paradigms has facilitated the creation of embeddable components that can be seamlessly integrated into third-party platforms.

While some literature has focused on the technical architecture of scalable chatbot platforms [10], relatively few studies explore systems that are both embeddable and configurable through domain-specific intent files provided by businesses. The lack of such flexible, plug-and-play AI solutions remains a significant research gap, particularly for SMEs lacking the resources to build custom AI infrastructure.

### ****2.5 Research Gap****

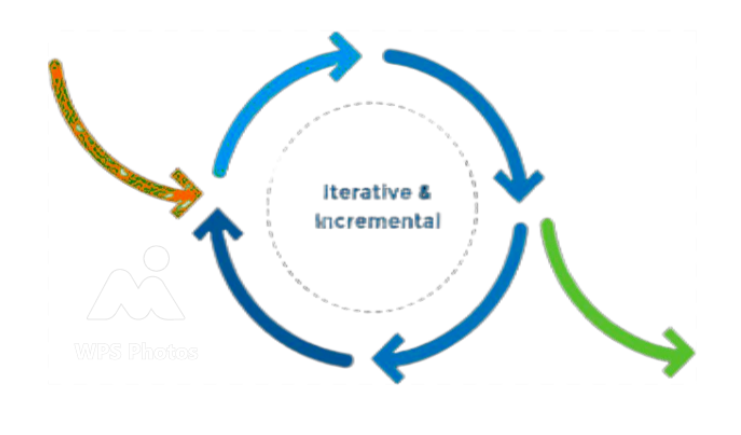
Although numerous works have explored AI chatbots, RAG models, and HITL strategies independently, few have addressed their integration into a unified, embeddable customer service framework. Existing platforms often require complex setups, lack domain adaptability, or do not offer real-time escalation. This paper contributes to filling that gap by presenting a real-time, web-based, RAG-powered customer service system with configurable human escalation and analytics support.

**3. METHODOLOGY**

This section outlines the methodological approach adopted in designing and implementing the real-time, embeddable Retrieval-Augmented Generation (RAG) system for customer service with human-in-the-loop (HITL) escalation. The methodology spans the selection of development models, system components, and architectural decisions aimed at achieving the project objectives—namely, real-time processing, intelligent query handling, and seamless human agent escalation.

### ****3.1 Development Approach****

The **Iterative Software Development Model** was adopted due to its flexibility and emphasis on incremental refinement. Each iteration involved the design, implementation, and evaluation of specific components, such as the chatbot engine, the intent-based retrieval mechanism, and the escalation framework. Feedback was continuously incorporated to enhance system usability and performance.

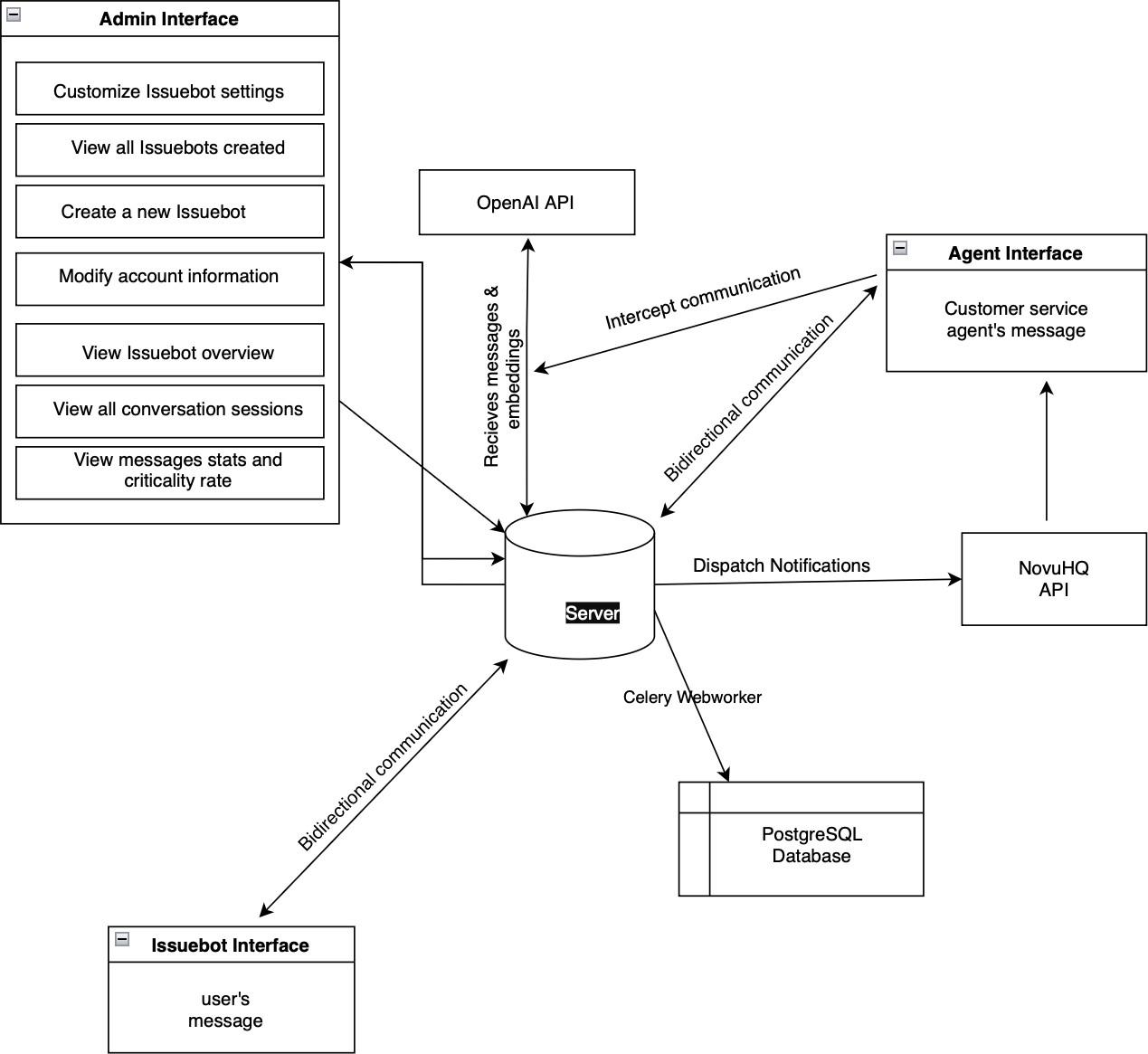


**Figure 1**. Iterative Software Development Cycle

### ****3.2 System Workflow Overview****

The system consists of several integrated components that handle user input, process retrieval and generation tasks, and escalate unresolved issues to human agents. The core flow involves:

1. A customer initiates a query through the web-based chatbot interface.
2. The backend receives the message and generates an embedding of the user input.
3. Using vector similarity, the system retrieves relevant content from a user-uploaded **intent file.**
4. A RAG-enabled prompt is formed using the retrieved content and passed to the **OpenAI GPT API**, which returns a response.
5. If the chatbot cannot confidently respond or the user explicitly requests human assistance, the message is escalated to a human agent using **Centrifugo** (WebSocket-based real-time communication) and **NovuHQ** notifications.
6. The agent seamlessly joins the session, and both interaction data and system metrics are logged in the PostgreSQL database.

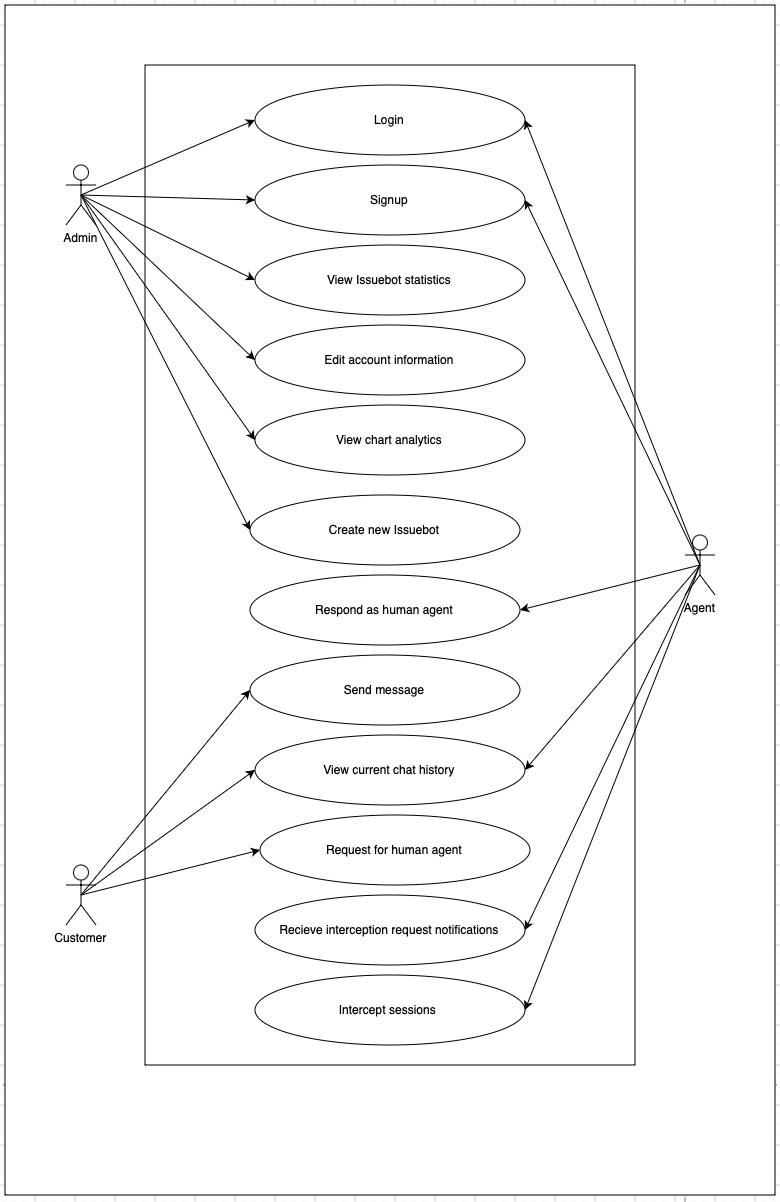


**Figure 2.** High-Level Architecture of the System

### ****3.3 Use Case and Functional Design****

The system supports three primary actors:

1. **End Users** (Customers) interact with the chatbot, request assistance, and receive responses.
2. **Human Agents** receive escalation alerts, join ongoing chats, and intervene when necessary.
3. **Administrators** upload and manage intent files, configure escalation rules, and monitor system performance.



**Figure 3.** Actor-Based Use Case Representation

### ****3.4 Sequence of Operations****

The communication and task flow are captured in a **sequence diagram**, showing the interactions among the frontend, backend, database, RAG engine, and escalation services. It emphasizes the decision points for automated vs. human response routing.

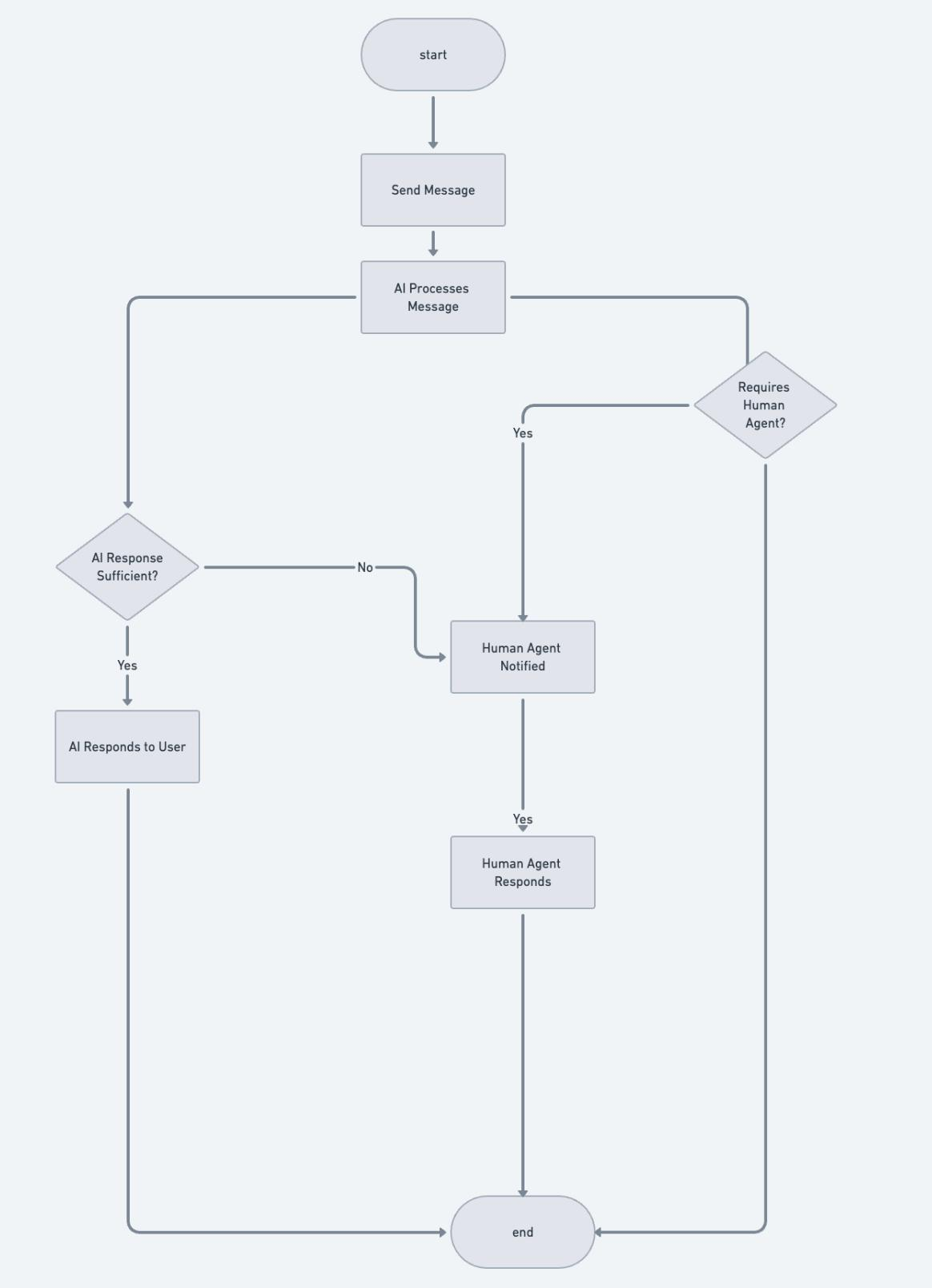
### 

**Figure 4.** System Message and Escalation Sequence

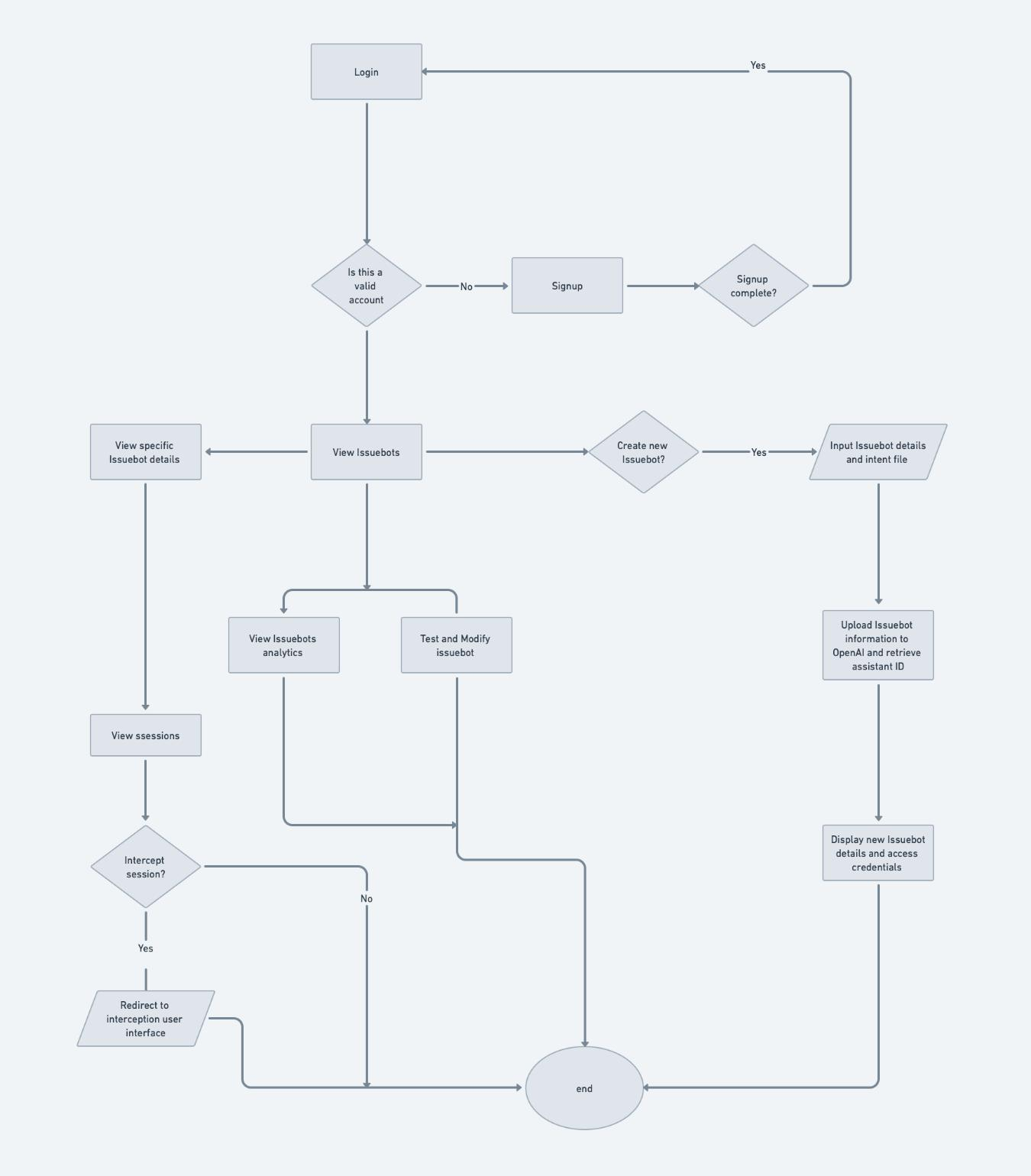
### ****3.5 User Experience and Backend Logic Flow****

To capture both user and system behaviours, two diagrams are proposed:

1. **User Flow Diagram:** Captures typical customer interactions, from initiating a query to receiving assistance.
2. **System Workflow Diagram:** Represents how backend components like FastAPI, Celery, and Centrifugo handle requests asynchronously and communicate with external APIs.



**Figure 5.** User flow diagram



**Figure 6.** System Processing Flowchart

### ****3.6 Data Modelling and Storage****

The backend of the proposed system is supported by a relational data model implemented in **PostgreSQL**, designed to manage real-time chatbot sessions, track escalation events, and provide analytical insights. The core entities include users, chat sessions, messages, intent files, escalation logs, and analytics.

Each user, whether a customer, human agent, or administrator, is stored with a role-based profile. When a user initiates a conversation, a session is created and linked to all messages exchanged within that session. These messages are recorded with sender type and timestamps for traceability. If a conversation requires human intervention, an escalation log is created to capture details such as the responding agent, response time, and escalation type (manual or automatic). Meanwhile, businesses upload intent files, which are indexed and used during the retrieval step of the RAG workflow. To enable performance monitoring, interaction data is aggregated in the analytics table, recording metrics like average response time, resolution duration, and escalation frequency. The model is normalized for integrity and indexed for speed, supporting high-performance query operations in real-time environments.

### ****3.7 Tools and Technologies****

### ****Table 1.** Tools, Technologies, and Purpose**

|  |  |  |
| --- | --- | --- |
| Component | Technology | Purpose |
| Frontend | ReactJS, Vite | Lightweight, interactive user interface |
| Backend | FastAPI | Core API logic and business rule processing |
| Chat Engine | OpenAI GPT (via API) | Natural language understanding and generation |
| Retrieval Mechanism | Vector similarity search (e.g., FAISS or in-memory) | Matches user input with relevant intent data |
| WebSocket Layer | Centrifugo | Enables real-time messaging |
| Notifications | NovuHQ | Triggers alerts for escalation |
| Task Queue | Celery | Handles background jobs and long-running tasks |
| Database | PostgreSQL | Stores structured data and chat histories |

This methodological framework ensured that the system was both technically robust and adaptable to varying customer service contexts. The next section presents the actual implementation and graphical user interface of the completed system.

## **4. IMPLEMENTATION**

The implementation phase focused on translating the system design into a functional, web-based application that supports real-time, intelligent customer interaction. The system was developed using a modular, API-first architecture to ensure flexibility, embeddability, and integration with external platforms.

### ****4.1 Technology Stack and Tooling****

The backend was implemented using **FastAPI**, a modern Python framework well-suited for asynchronous APIs. It handles request routing, session management, and integration with external services. **Celery** was used for executing background jobs such as logging, session analysis, and periodic updates. The real-time communication layer was built using **Centrifugo**, which enables WebSocket connections between the frontend and backend. For notifications during escalation, **NovuHQ** was integrated to push alerts to available agents.

The frontend was built using **ReactJS** and bundled with **Vite**, offering a fast, lightweight interface that supports chat interactions, bot testing, analytics dashboards, and administration tools. User authentication was managed using Google OAuth and password-based login, with session handling secured by token-based authorization.

### ****4.2 Chatbot and Retrieval-Augmented Generation Integration****

The chatbot functionality was powered by **OpenAI's GPT-3.5-turbo model**, with RAG (Retrieval-Augmented Generation) used to enhance accuracy. When a user submits a query, the system:

1. Generates an embedding of the query using OpenAI’s embedding API,
2. Searches a pre-computed vector space derived from a user-uploaded intent file (using FAISS or in-memory search),
3. Combines the retrieved context with the query,
4. Constructs a prompt and sends it to the OpenAI chat API for response generation.

The response is then returned to the frontend in real-time. If the chatbot determines, based on predefined rules or confidence levels, that the query is outside its scope, it flags the message for escalation.

### ****4.3 Human Escalation Mechanism****

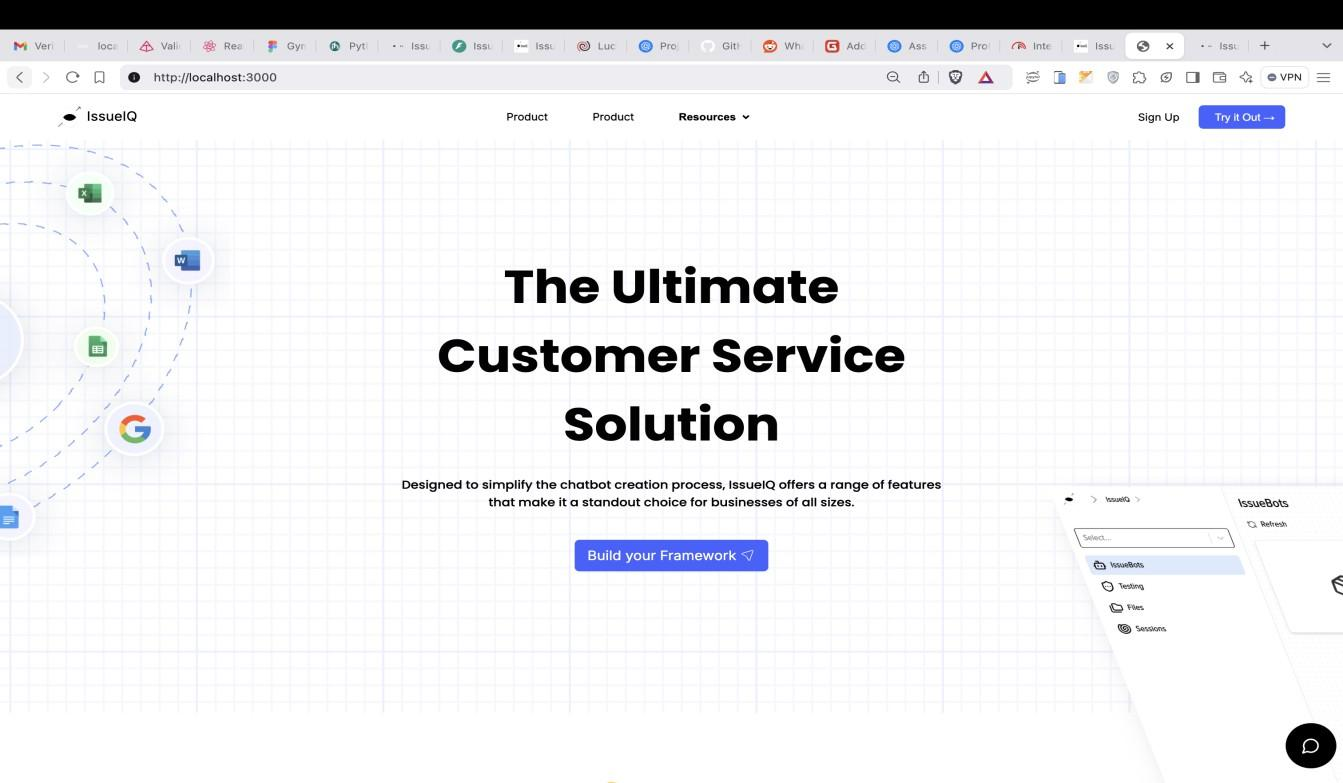
Upon escalation, the system uses **NovuHQ** to send real-time notifications to available agents. The agent receives the session ID and full message context and can immediately join the conversation through the agent interface. A seamless transition is achieved via Centrifugo’s WebSocket layer, which synchronizes the chat window without reloading or disrupting the user experience.

Agents can respond directly, resolve the issue, or escalate further as needed. The conversation can also be returned to the chatbot after the issue is clarified.

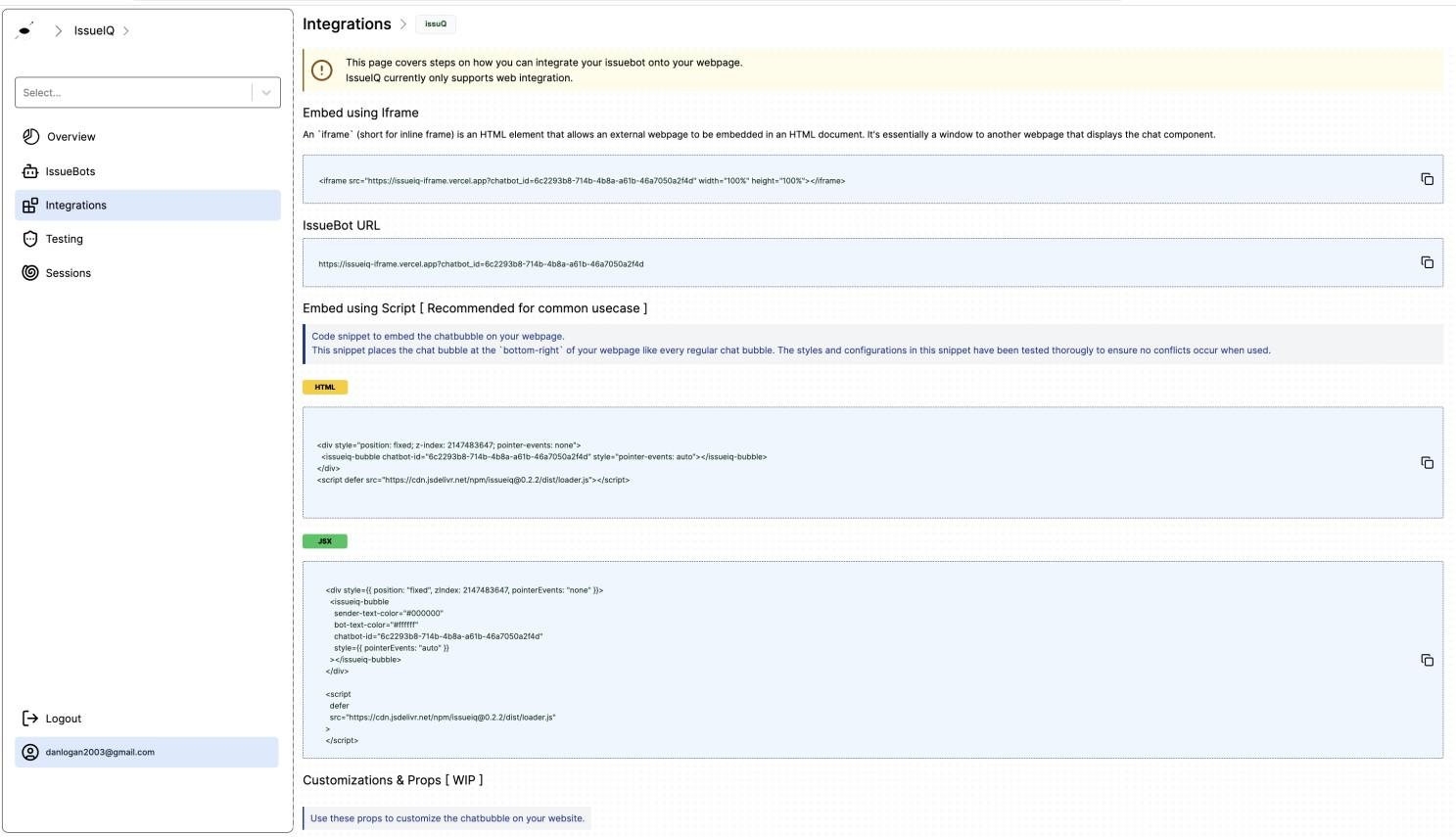
### ****4.4 Key Functional Modules****

The system was structured around several key user-facing modules:

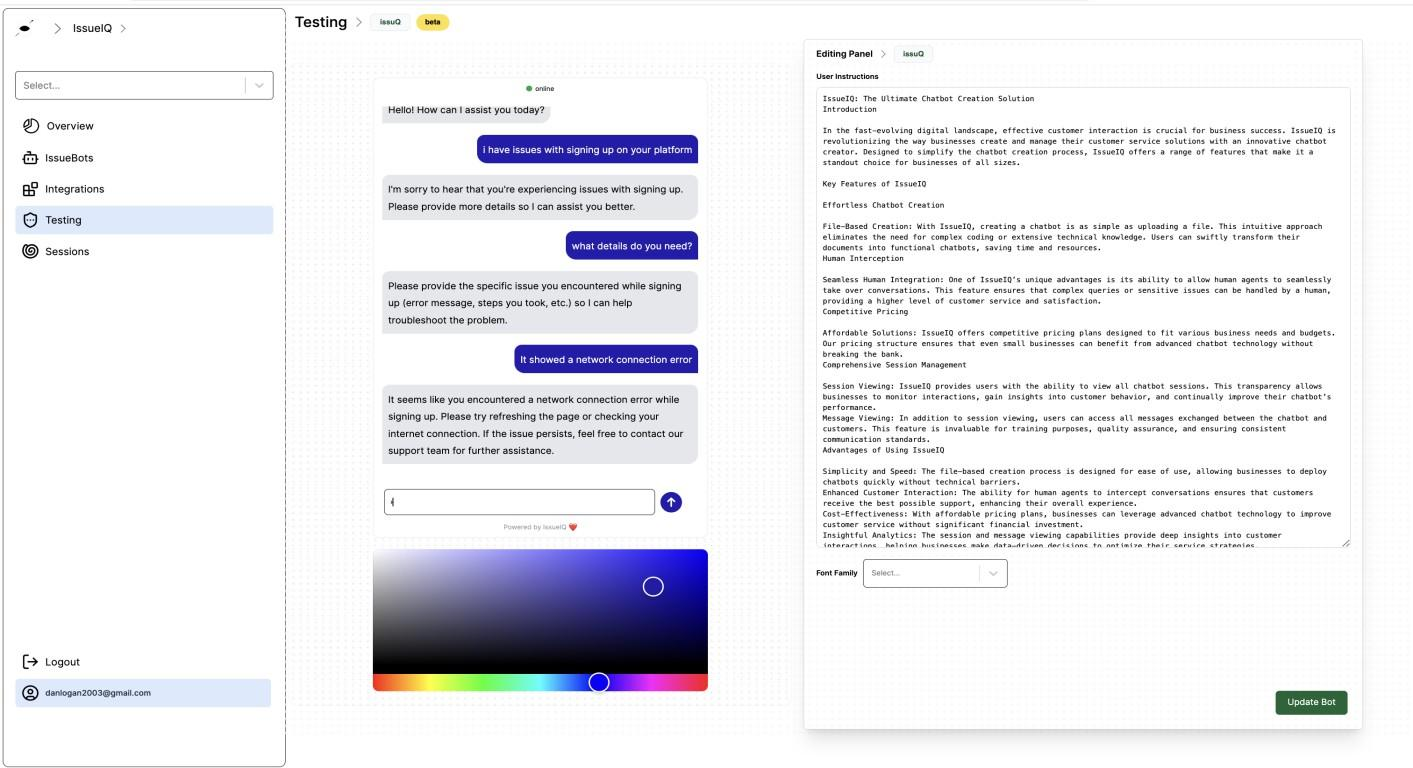
1. **Landing Page:** Presents the platform and guides users to log in or register.
2. **Authentication Module:** Supports secure login with password or Google OAuth.
3. **IssueBot Management:** Allows users to create and manage their chatbot instances, upload intent files, and configure escalation behaviour.
4. **Testing Interface:** Let users simulate conversations before deployment.
5. **Integrations Page:** Provides embeddable iframe code, shareable URLs, and script-based web components for third-party websites.
6. **Sessions Dashboard:** Displays active and historical chat sessions with escalation logs.
7. **Analytics Module:** Visualizes key performance indicators such as resolution time, response latency, and escalation rate.



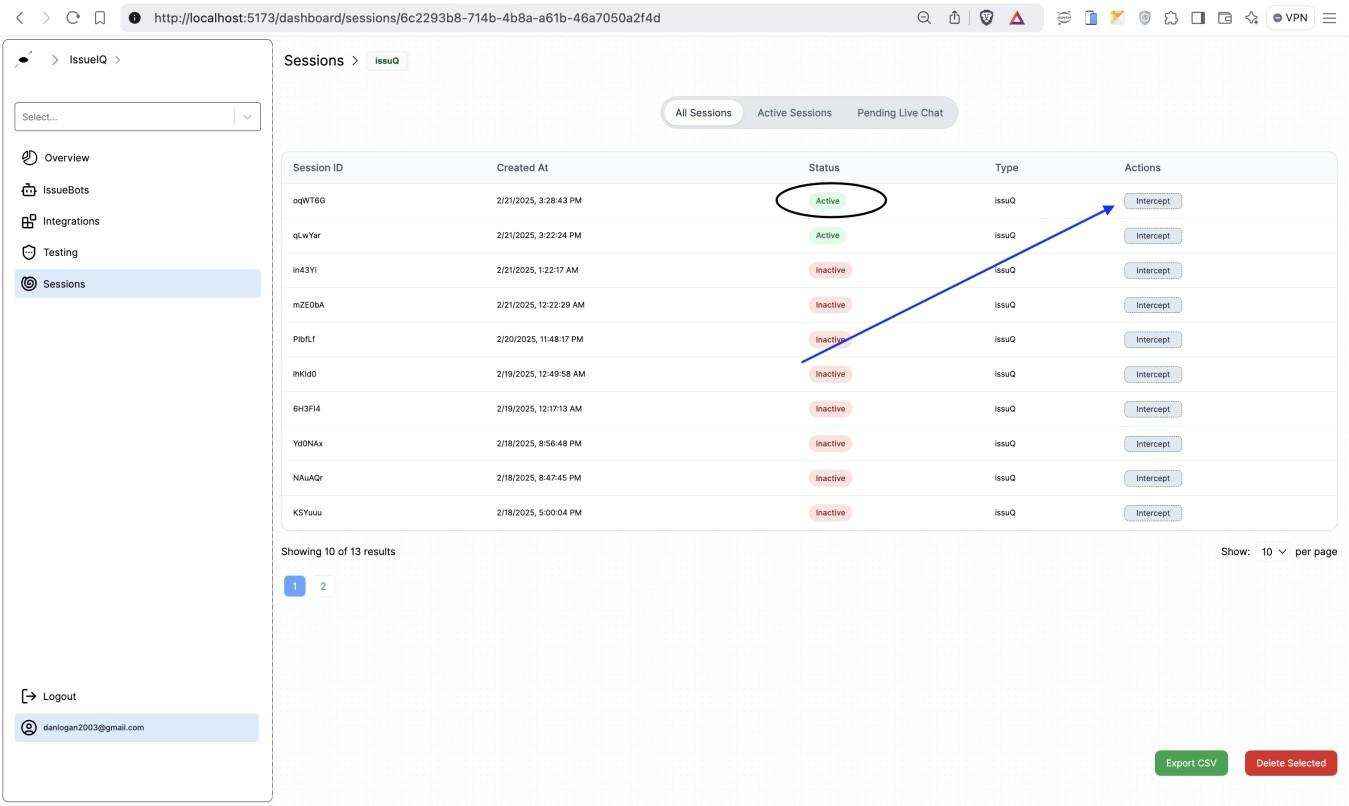
**Figure 7.** Landing Page



**Figure 8.** Integration page



**Figure 9.** Testing page



**Figure 10.** Testing page

### 

**Figure 11.** Overview and Analytics Page

### ****4.5 Deployment and Hosting****

The system was containerized using Docker and deployed to a cloud-based environment to enable scalability. PostgreSQL was hosted separately to handle persistent storage, while the WebSocket server and backend logic were deployed as independent microservices. The frontend was hosted on a static CDN for optimal load speed.

This implementation provides a production-ready foundation for businesses seeking to integrate intelligent, real-time customer support into their platforms. The next section evaluates the system’s performance and discusses its limitations and future directions.

**5. EVALUATION AND RESULTS**

This section evaluates the performance, usability, and reliability of the implemented real-time, embeddable RAG-based customer service system. The goal was to determine how well the system fulfilled its objectives in real-world or simulated scenarios, particularly in terms of response accuracy, escalation efficiency, and user satisfaction.

### ****5.1 Evaluation Strategy****

A hybrid evaluation approach was adopted, consisting of both **quantitative** and **qualitative** assessments. Key metrics included:

1. **Average Response Time:** Time between user input and bot/agent reply.
2. **Resolution Rate:** Percentage of sessions that concluded without requiring human intervention.
3. **Escalation Latency:** Time between escalation trigger and human agent response.
4. **Accuracy of Responses**: Measured by comparing bot answers to known intent file information.
5. **User Satisfaction**: Captured via feedback prompts after session completion.

The system was deployed in a controlled environment for testing, with simulated users engaging with the chatbot across various scenarios.

### ****5.2 Quantitative Results****

The system demonstrated strong performance across key metrics during evaluation. Chatbot responses were consistently delivered in under three seconds, ensuring real-time interaction. Out of 150 test sessions, approximately 71% were fully resolved by the chatbot, while the remaining 29% required escalation to human agents.

Escalated cases were handled efficiently, with agents typically responding within 2.8 minutes of receiving a notification. The accuracy of chatbot responses was high, over 90%, when supported by well-structured intent files. However, performance dropped slightly with less organized or incomplete input data.

Overall, the results confirmed the system’s ability to deliver fast, accurate, and scalable customer service with effective human support when needed.

### ****5.4 System Monitoring and Logging****

The system includes a built-in monitoring and analytics module that tracks chatbot interactions, escalation events, and user behaviour in real time. Each chat session is logged from start to finish, capturing key data such as response times, session length, and escalation triggers. This information is processed asynchronously and visualized through an admin-facing dashboard.

During evaluation, the dashboard proved effective for identifying peak usage periods, frequent escalation triggers, and underperforming intent matches. Administrators were able to view trends such as the types of questions that led to escalations and the average response time from human agents. These insights allowed for the refinement of intent files and better agent allocation.

In addition to performance metrics, audit logs record system changes and user activities, ensuring transparency and accountability. Overall, the monitoring tools supported both system optimization and operational oversight.

### ****5.5 Summary of Findings****

The evaluation demonstrated that the system:

1. Provides low-latency, accurate, and domain-aware chatbot responses,
2. Seamlessly escalates queries to human agents in real time,
3. Supports efficient administration through logging and analytics,
4. Is adaptable for deployment across industries requiring customer service automation.

The hybrid AI-human approach shows clear advantages over traditional standalone chatbot systems, especially in handling complex, multi-step queries where context and nuance are critical.

## **6. CONCLUSION AND FUTURE WORK**

This study presented the design and implementation of a real-time, embeddable customer service system that integrates Retrieval-Augmented Generation (RAG) with human-in-the-loop (HITL) escalation. By combining large language model capabilities with contextual retrieval and agent intervention, the system offers a scalable and intelligent approach to modern customer support.

The system performed strongly during evaluation, demonstrating fast response times, high resolution rates, and seamless escalation processes. Users appreciated the intuitive interface, real-time interaction, and smooth transition between bot and human agents. Human agents also highlighted the value of session history access, which enabled them to respond with full conversational context. These features contributed to a user-friendly and effective hybrid support experience.

Nonetheless, some limitations emerged that point toward areas for future development. The system's performance was closely tied to the quality of the uploaded intent files, poorly structured or overly technical data reduced response accuracy. The chatbot also struggled occasionally with ambiguous or sarcastic input, leading to misinterpretations or unnecessary escalations. Additionally, the system currently supports only English, which limits its reach in multilingual environments.

Looking ahead, several enhancements are planned. Future iterations will explore:

1. **Multilingual support**, enabling broader deployment across diverse user bases;
2. **Advanced sentiment and tone detection**, improving the chatbot’s handling of emotionally nuanced queries;
3. **Voice-based interaction**, expanding accessibility;
4. **Adaptive learning**, allowing the system to improve over time based on user feedback and past interactions.

In conclusion, the system lays a strong foundation for intelligent customer support that blends AI automation with human empathy. With continued refinement, it holds promise as a flexible, enterprise-ready solution for organizations seeking to elevate their customer service capabilities.

**REFERENCES**

[1] E. Adamopoulou and L. Moussiades, “Chatbots: History, technology, and applications,” Machine Learning with Applications, vol. 2, p. 100006, 2020.

[2] B. AbuShawar and E. Atwell, “ALICE chatbot: Trials and outputs,” Computación y Sistemas, vol. 19, no. 4, pp. 625–632, 2015.

[3] T. B. Brown et al., “Language models are few-shot learners,” in Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901, 2020.

[4] K. M. Colby, Artificial Paranoia: A Computer Simulation of Paranoid Processes. Oxford, UK: Pergamon Press, 1975.

[5] M. Dixon, K. Freeman, and N. Toman, “Stop trying to delight your customers,” Harvard Business Review, vol. 88, no. 7/8, pp. 116–122, 2010.

[6] S. Kashyap and M. Salunke, “Real-time chatbot systems using WebSockets and asynchronous queues,” International Journal of Advanced Computer Science and Applications, vol. 14, no. 2, pp. 31–39, 2023.

[7] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474, 2020.

[8] R. Nirala, V. Singh, and K. Purani, “Machine learning-powered chatbots in customer service: Capabilities and limitations,” Journal of Retailing and Consumer Services, vol. 67, p. 102966, 2022.

[9] OpenAI, “GPT-4 Technical Report,” OpenAI, 2023. [Online]. Available: https://openai.com/research/gpt-4

[10] T. Polzehl, Y. Cao, V. I. Sanchez Carmona, X. Liu, C. Hu, and S. Möller, “Personalizing customer service chatbots: The role of information savviness,” ACM Transactions on Interactive Intelligent Systems, vol. 12, no. 4, pp. 1–26, 2022.

[11] A. Shah, R. Jain, and D. Patel, “AI-human collaboration in conversational interfaces: A survey,” Journal of Human-Computer Interaction, vol. 38, no. 1, pp. 45–67, 2022.

[12] I. Ullah, S. Ji, and C. Kim, “What (de)motivates customers to use AI-powered conversational agents: A behavioural reasoning approach,” Technological Forecasting and Social Change, vol. 188, p. 122318, 2023.

[13] J. Weizenbaum, “ELIZA—a computer program for the study of natural language communication between man and machine,” Communications of the ACM, vol. 9, no. 1, pp. 36–45, 1966.

[14] Y. Wu, M. Galley, C. Brockett, and B. Dolan, “Improving knowledge-grounded dialogue generation with retrieval-enhanced adversarial training,” in Proc. of the 59th Annual Meeting of the Association for Computational Linguistics, pp. 3101–3110, 2021.

[15] Y. Xu, C.-H. Shieh, P. van Esch, and I.-L. Ling, “AI customer service: Task complexity, problem-solving ability, and usage intention,” Journal of Service Research, vol. 23, no. 4, pp. 458–473, 2020.