VPBank Technology Hackathon 2025

General Brief

Challenge Statement	Speak To Input (#20)
Team Name	Group 19 - PipeKat and LodiKat

Team Members

Full Name	Role	Email Address	School Name	Faculty / Area of Study	LinkedIn Profile URL
Pham Nguyen Hai Anh	Solution Architect (Leader)	haithe123123@gmail.com	University of Information Technology	Information Security	https://www.linkedin.com/in/anhpnh/
Bui Ho Ngoc Han	Project Manager	buihongochan.lodi@gmail.com	FPT University	Information Systems	https://www.linkedin.com/in/lodi-bui-han/
Le Minh Nghia	Software Engineer	nghialm2603@gmail.com	Dong Nai Technology University	Information Technology	https://www.linkedin.com/in/minhnghia2k3/

Danh Hoang Hieu Nghi	AI Engineers	hieunghiwork123@gmail.com	Ho Chi Minh City University of Foreign Languages	Information Technology	https://www.linkedin.com/in/hieunghi/
Nguyen Duc Toan	DevOps Engineer	toanndcloud@gmail.com	University of Information Technology	Computer Networks and Communications	http://www.linkedin.com/in/toanndcloud021104

Content Outline

Solutions Introduction

1. Problem Overview – The Hidden Productivity Drain in Manual Data Entry

Vietnamese banking operations face significant challenges in managing the heavy burden of manual data entry across their core systems. Current workflows for vital services like Loan Origination System (LOS), Customer Relationship Management (CRM), Human Resource Management System (HRMS), and Risk and Compliance assessments are still heavily dependent on non-automatic intervention [1]-[4]. These inefficiencies directly impact:

- Turnaround time in loan origination and customer service
- Accuracy of compliance and audit reports
- Overall employee satisfaction and workflow agility

Such dependency on manual input not only strains workforce capacity but also limits the bank's ability to achieve agility and scalability in its operations.

Key Process Bottlenecks:

- Loan Origination System (LOS): Manual re-entry of customer and financial data across multiple modules during credit evaluation leads to processing delays, data inconsistencies, and extended loan approval cycles [4].
- Human Resource Management System (HRMS): Manual updates for employee records, payroll, and attendance create data redundancy, increase error rates, and hinder timely workforce reporting.
- Customer Relationship Management (CRM): Manual entry of customer interactions, service requests, and lead information results in fragmented records, reduced data accuracy, and slower response times, limiting the bank's ability to deliver personalized and timely customer engagement.
- Letter of Credit (LC) Processing: Staff manually sift through countless emails to categorize and process LC applications, amendments, and cancellations. Manual data consistency checks across multiple documents lead to an estimated error rate of 15-20% [21], [22], necessitating costly rework and creating compliance risks.

2. Analysis of Existing Solutions

There are several approaches to solve the manual data entry problem, which are mainly grouped into two groups, listed in the table below:

Approach	Description	Contribution / Benefit	Limitations
Traditional RPA [11], [12]	Script-based bots automate predefined data-entry sequences.	VPBank was able to reduce up to 20 manual steps in processing documents and applications. As a result, approval times were shortened by 30 % to 40 %. This enabled customers to have unsecured loan approved in 2-3 days. [16]	Inflexible to voice input, frequent maintenance when UI or format changes.
Voice AI Assistant [13]-[15]	One conversational AI that interprets and types the inputs.	VIB MyVIB AI Voice Banking: - Helps customer complete a core transaction in under 1 minute via 3 simple steps [18]. - Supports natural Vietnamese (North/Central/South) with NLP and speech-to-text / text-to-speech [19]. TPBank Mobile – VoicePay: - Enables hands-free financial transactions – useful when driving or multitasking [20].	Lacks structured validation, not support with diverse systems.

Table 1 – Comparison of Existing Approaches for Automating Manual Data Entry

To address these limitations, we propose the voice-driven multi AI agents automation solution that turns multilingual speech into structured, validated enterprise data entries and commands, executing directly across existing banking systems without modifying legacy infrastructure.

3. Our Highlighted Solution: The “VPBank Speak-to-Input Multi Agent Automation”

We present the **VPBank Speak-to-Input Multi Agent Automation** solution. This is an intelligent automation solution that enables VPBank staffs and customers to speak multiple languages naturally to fill enterprise forms – from loan application to compliance reporting – instead of typing manually, and can understand, map, and execute the action with banking systems. Through advanced Automatic Speech Recognition (ASR) model (fine-tuned PhoWhisper) and several voice-processing AI models, as well as LLM orchestration (Amazon Bedrock Agents + LangChain/LangGraph), the system converts multilangual voice commands (especially focusing Vietnamese) into validated, structured inputs.

Key Innovations

- **Fine-Tuned Multilingual ASR for Banking Precision**: Optimized English-Vietnamese ASR open-source models (fine-tuned PhoWhisper) built specifically for financial operations and Vietnamese regional accents.
- **Hyper-Localized for Vietnam**: Purpose-built to align with the State Bank of Vietnam (SBV) regulations and local business contexts. The system integrates advanced Vietnamese language processing, automatically corrects common spelling errors with the help of GenAI models and Retrieval-Augmented Generation (RAG).
- **Instant Action Execution:** Uses audio turn detection AI model to respond immediately as the user speaks – processing long or compound command strings in real time to fill enterprise forms, trigger banking functions, or navigate between workflow steps without delay.
- **Interactive Confirmation:** Provides textual and voice-based feedback (using text-to-speech AI model) for confirmation, validation, and updates. Ensuring each transaction or form entry is reviewed for accuracy and completeness before submission, as well as notification for the result.
- **Voice-Driven Input Editing and Control:** Uses Browser AI Agent to empowers users to input, edit, delete, or update content within form fields as well as activate functional buttons on web interfaces – fully replacing manual typing and clicking.
- **Advanced Multi-Agent Architecture**: Our core innovation lies in collaborative intelligence, where multiple specialized agents reason together and share insights. This enables dynamic orchestration of tasks such as intent recognition, data validation, and system interaction through Amazon Bedrock Agents and LangChain/LangGraph frameworks.
- **An Engine for Unrivaled Growth and Prosperity**: Our architecture is designed to match the ambition and pioneering spirit of VPBank. Built on enterprise-grade AWS, it provides the secure, massively

scalable foundation required to not only support VPBank's current leadership position but to propel its digital dominance for the next decade, fostering a more prosperous Vietnam.

Impact of Solution

1. Social Impact & Target Benefit

- Enhancing Accessibility: The solution broadens digital banking access for underserved groups in Vietnam, including 11.6% of the population aged 60+ [5] and 7.06% with disabilities [6]. A context-aware voice interface enables natural spoken transactions, while TTS feedback audibly guides visually impaired users through each step.

- Reducing Customer Friction: Complex forms remain a major barrier to digital adoption, with average completion rates of only 42–58% [7] and 67% of users abandoning them due to length or unclear fields [8]. By allowing users to speak naturally in Vietnamese, the system automatically extracts and validates information, reducing user effort and increasing completion rates and satisfaction.

- **Multilingual and Standardized Vietnamese Support**: The system accurately recognizes diverse Vietnamese accents and informal speech through a hybrid model combining PhoWhisper for phonetic recognition and Claude Sonnet for contextual reasoning. It enforces standardized Vietnamese orthography and entity normalization, ensuring data consistency and compliance across CRM, HRMS, and LOS systems.

2. Strategic Positioning and Differentiation

- Back-Office and Staff-Facing Voice Automation: While most voice banking products in Vietnam (e.g., TPBank VoicePay) focus on retail customer transactions, there is no packaged solution optimized for staff-facing workflows such as L/C, trade finance, enterprise credit, or compliance forms. Our solution directly addresses this gap – enabling internal teams to handle complex business forms hands-free.

- Vietnamese Technical Language and Slot-Filling Accuracy: Following NVIDIA Riva's recommendations, our focus is not limited to reducing WER (Word Error Rate), but to maximizing slot-filling accuracy – mapping each recognized entity precisely to its corresponding form field. By fine-tuning PhoWhisper and expanding its vocabulary with financial terminology, the system ensures precise recognition and field mapping for financial workflows.

3. Addressing Business Pain Points

3.1. Reducing Branch & Call-Center Operational Costs

Pain Point: Banks typically employ 500–1,000 staff at branches and call-centers to handle simple tasks like card registration, or data updates.

Solution: The Voice-AI system automates these repetitive tasks. Estimated cost per transaction:

- **Amazon Bedrock (Claude Sonnet):** Approximately $0.0105 per request (based on ~1,000 input and 500 output tokens, at $3 per million input tokens and $15 per million output tokens according to Anthropic pricing on AWS Bedrock).
- **Browser-Use agent:** Around $0.01–0.02 per task (using Claude Sonnet for automation orchestration).

ROI: This configuration delivers a substantial cost reduction compared to manual staff processing while maintaining accuracy and scalability across thousands of transactions per day.

3.2. Improving Form Completion Rate (Conversion Rate)

Pain Point: In financial-services onboarding, form abandonment remains high. Studies show that when forms are lengthy (more than 10 fields), not mobile-friendly, have unclear validation, or require additional document retrieval, churn is significant.

Solution: A voice interface reduces friction by enabling users to begin anywhere – even if not all information is ready (“My name is Nguyễn Văn A, I’ll tell you the ID later”), the system retains context and supports multi-session continuation (customers can resume after 1-2 days).

Supporting data: Form abandonment statistics show that over 67% of users abandon forms when encountering usability issues [10].

3.3. Seamless Omnichannel Support

Pain Point: Customers may start account opening via web, then switch to mobile app for photo upload, then make a call for verification, and finally visit a branch for signing. Each channel is a separate system, with no shared data.

Solution: The Voice-AI platform works across any WebRTC-enabled endpoint.

3.4. Automated Compliance & Audit Trail

Pain Point: Transactions via call-centers or branches often lack detailed, user-verifiable records. When a complaint arises, banks struggle to show which terms were agreed by the user.

Solution: Every voice interaction is recorded and traced:

- Full transcripts: STT-processed text stored
- Audio recording: Optionally save .wav files
- Action logs: Function calls with timestamps

4. Comparative Analysis with Common Approaches and existing productions

4.1. Method 1 – Traditional Web/App Forms

How It Works: Users manually type text into input fields, select dropdowns, and tick checkboxes.
Comparison: Our Voice AI solution reduces completion time increase the conversion rate, and expands accessibility to the population who cannot effectively use traditional forms – including seniors and users with motor or visual impairments.

4.2. Method 2 – Text-Based Chatbots (e.g., VIB ViePro, ChatGPT-based Assistants)

How It Works: Users type text messages, and the chatbot responds via text. The bot may answer questions or collect information, but users still need to manually fill out forms afterward.
Comparison: Our Voice AI system combines conversation + automation. Users not only converse but also receive completed results – the AI actually fills and submits the form. Voice is also inherently faster than typing: verage speaking speed ≈ 150 wpm vs typing speed ≈ 40 wpm, meaning faster data entry.

4.3. Method 3 – OCR + Document AI (e.g., Ocrolus, ABBYY Vantage)

How It Works: Users upload photos of documents (ID cards, driver's licenses, bank statements). The AI extracts text and auto-fills form fields.
Comparison: Our Voice AI solution is universal – applicable to all forms, not limited by document type. It costs less per transaction and can integrate OCR as a pre-step (extracting ID data) before using voice to fill in missing information.

4.4. Comparative with existing solutions

Besides, the following table provides a holistic comparison of current voice and AI-driven data entry technologies across multiple categories. It highlights where our system – a Vietnamese-first Speak-to-Input AI

platform – holds distinct advantages in enterprise automation, contextual understanding, and compliance-ready deployment.

Solution Category	Representative Products	Strengths	Limitations / Trade-offs	Primary Use Cases
Specialized Voice-to-Form Systems	SayFill, Dataloop Voice Pipeline	Fast voice-to-form interaction; quick go-to-market	Requires detailed field/intent mapping; limited domain vocabulary support	Internal admin workflows, CRM, service forms
ASR / Speech Tech Stack	Google STT, NVIDIA Riva	High accuracy, real-time streaming; Riva supports on-prem deployment and custom vocabulary	Heavy DevOps (Riva) or high cloud cost (Google); limited context retention	Banking, contact centers, internal AI assistants
Vietnam Voice Banking Solutions	TPBank VoicePay, VIB MyVIB Voice Banking	Proven retail use cases; includes biometric voice authentication	End-user focus (transactions, Q&A); lacks back-office or enterprise form automation	Retail banking, customer interaction
IDP / KYC Document Automation	Ocrolus, Hyperscience, ABBYY	>99% accuracy in document parsing and validation	Complex licensing; requires human-in-the-loop validation	Lending, KYC, trade finance

		(e.g., mortgage use cases)		
Audit Report Automation	DataSnipper, Fieldguide, MindBridge	AI-assisted draft generation, risk scoring, and time savings	Quality control and bias concerns; unclear long-term KPI impact	Internal/external audit automation
OS-Level Voice Assistants	Apple Siri, Microsoft Voice Access	On-device security; natural dictation and voice-based input; accessibility for disabled users	Limited to OS ecosystem; poor domain-level context understanding; minimal slot-filling customization	Accessibility, personal assistance, device control

Table 2 – Comparative Analysis of Existing Voice and AI-Driven Data Entry Solutions

Compared to existing solutions, our solution is better because**:**

• **Fine-Tuned Multilingual ASR for Banking Accuracy:** The solution uses optimized English–Vietnamese Automatic Speech Recognition (ASR) model, fine-tuned from PhoWhisper, to deliver exceptional accuracy in financial operations. These models are specifically adapted to understand Vietnamese regional accents and banking terminology, ensuring reliable transcription and command recognition in multilingual environments.

• **Hyper-Localized for the Vietnamese Market:** Developed with a deep understanding of Vietnam's financial landscape, the system is fully aligned with the State Bank of Vietnam (SBV) regulations and local business practices. It integrates advanced Vietnamese language processing and employs Generative AI with Retrieval-Augmented Generation (RAG) to automatically correct common spelling and grammatical errors, enhancing data quality and compliance.

• **Real-Time Action Execution:** Leveraging advanced audio turn detection model, smart-turn, the system processes user speech in real time. This enables immediate execution of complex or multi-step voice commands to populate enterprise forms, trigger banking workflows, or navigate between screens seamlessly – eliminating operational delays.

• **Interactive Voice and Text Confirmation:** The solution delivers instant, interactive feedback through both text and voice channels using state-of-the-art text-to-speech (TTS) technology. Each transaction or form input is validated and confirmed in real time, ensuring accuracy, completeness, and user confidence prior to final submission, as well as notification for the result.

• **Voice-Driven Input Control and Automation:** By integrating a Browser AI Agent, users can efficiently input, modify, delete, or update data within form fields and activate interface functions through natural speech commands. This capability replaces traditional manual data entry, significantly improving productivity and user experience.

• **Advanced Multi-Agent Orchestration:** At the core of the system is a collaborative multi-agent architecture, where specialized AI agents coordinate seamlessly to interpret intent, validate data, and interact with banking systems. Powered by Amazon Bedrock Agents and the LangChain/LangGraph framework, this intelligent orchestration enables adaptive, high-precision automation across banking workflows.

• **A Scalable Foundation for Sustainable Growth:** Built on enterprise-grade AWS infrastructure, the platform provides VPBank with a secure, scalable, and future-ready foundation. It is designed to reinforce VPBank's leadership in digital innovation while supporting the bank's long-term strategy for operational excellence and national economic advancement.

5. Quantifiable Impact

Cost Savings: Automating 60% of monthly transactions cuts cost.
Revenue Growth: Improved onboarding via voice automation raises conversion and account openings.
Time Savings: Average process duration reduced from 6m30s → 1m45s (73% faster).
Customer Experience (CX) Metrics
Beyond quantitative financial impact, the Voice AI platform substantially improves customer experience (CX) – reducing effort, increasing satisfaction, and strengthening brand advocacy.
Customer Effort Score (CES)

Metric	Before	After Voice AI	Improvement
Customer Effort Score (CES)	4.2 / 7 (“High effort”)	2.1 / 7 (“Easy”)	↓ 50% perceived effort

A lower CES indicates that customers complete banking tasks more easily, requiring less cognitive and physical effort – a key driver of digital adoption and repeat usage.

Net Promoter Score (NPS)

Metric	Baseline (VPBank)	Post-Voice AI	Improvement
Net Promoter Score (NPS)	35	52–55	+17–20 points

Summary Table: Quantifiable ROI for VPBank

Category	Baseline	With Voice AI	Impact
Conversion Rate	25%	55%	+30 pp (3–4× more completions)
Customer Effort (CES)	4.2 / 7	2.1 / 7	↓ 50%
Net Promoter Score (NPS)	35	55	+20 points
Transaction Time	6m 30s	1m 45s	↓ 73%
Queue Time (Branch)	25 min	10 min	↓ 60%
Customer Satisfaction	100% baseline	+45% improvement	↑ Significant

Table 8 – Quantifiable ROI and Operational Efficiency Gains from Voice AI Implementation at VPBank

Deep Dive into Solution

1. Data flow diagram

1.1. Level 0 data flow diagram

External Entity:

Entity	Description
Customer	End-users of VPBank's digital services – typically clients applying for loans, checking balances, or performing banking operations via web or app interfaces.
Internal Staff	Bank employees or operations teams using the system to process internal tasks.
Banking Websites	The front-end user interface (UI) – the web portals where form fields, buttons, and menus are located (for both staff and customers).
Banking Systems	Core or back-end systems (mocked in prototype) - includes core banking software, loan management systems, and compliance databases.
External Documents	Documents received from customers or third parties – e.g., loan applications, regulations, or legal forms.
Internal Documents	Institutional policies, regulations, and internal reference documents from the bank.

Data flow diagram description:

The Speak-to-Input System processes voice commands from users, retrieves relevant data for context, executes actions on the banking site, and provides feedback through voice or on-screen text.

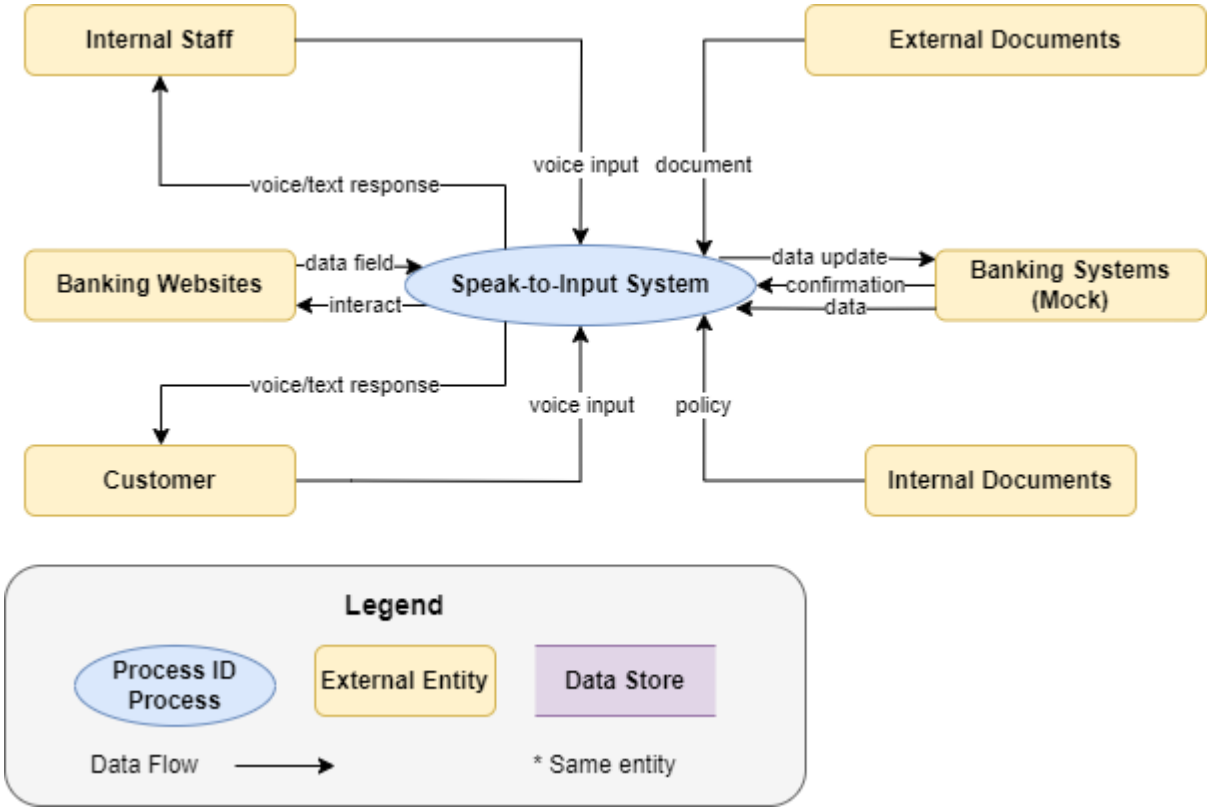

Figure 1: Level 0 Data Flow Diagram

1.2. Level 1 data flow diagram

External entities:

Entity	**Description**
Customer in Public Area	End-user of VPBank's web or mobile platforms who interacts by voice, possibly in noisy environments.

Internal Staff	VPBank employees using the voice-based system for internal workflows such as loan processing, compliance, or customer service tasks.
Banking Systems (Mock)	Represents VPBank's backend systems - core banking, compliance, CRM - simulated for prototype purposes.
Banking Web	The web interface where forms and buttons reside.
Internal Bank Document	Repository of internal policies, form templates, or compliance instructions.
External Regulatory Document	External policies and SBV regulations used for compliance validation.

AI/Automation entities:

AI / Automation Component	**Description**
Noise-Cancel AI Model (Krisp)	Removes background noise from the audio input to ensure clean voice data.
VAD Model (Silero VAD Analyzer)	Detects active speech segments in the audio stream.
Audio Turn Detection Model (Smart-Turn)	Detects when the user starts or finishes speaking to manage conversation turns.
Fine-tuned ASR Model (PhoWhisper)	Converts multilingual speech (especially Vietnamese) into accurate text transcription.

Coordinator GenAI Model (LLM Orchestrator)	Understands user intent from text using LLM frameworks like Amazon Bedrock + LangChain/LangGraph.
Banking Operation Orchestrator (GenAI Model)	Handles domain-specific actions such as editing fields, submitting forms, or fetching data.
Browser AI Agent (Browser-Use Web-UI)	Performs automated interactions on the banking web interface - filling forms, clicking buttons, navigating menus.
Text-to-Speech AI	Converts text-based feedback or confirmations into synthesized voice for user response.

Process "Process Voice": This process transforms raw voice input from customers or staff into clean, structured text. It begins by applying noise cancellation (Krisp) and voice activity detection (VAD) to isolate meaningful speech segments. Using the Smart-Turn model, the system detects conversation turns in real time, enabling smooth multi-speaker interaction. The refined audio is then processed by the fine-tuned PhoWhisper model, which accurately transcribes multilingual (Vietnamese–English) speech into text suitable for banking operations.

Process 2 – Act on Intent: This process interprets the transcribed text to determine the user's intent and execute corresponding actions. The Coordinator GenAI model classifies the intent, while Amazon Bedrock Agents and LangChain/LangGraph orchestrate the appropriate response. The Browser AI Agent then performs real-time actions such as filling, editing, or submitting forms on the banking website. Finally, the Text-to-Speech AI model provides confirmation or feedback to users through voice or text responses.

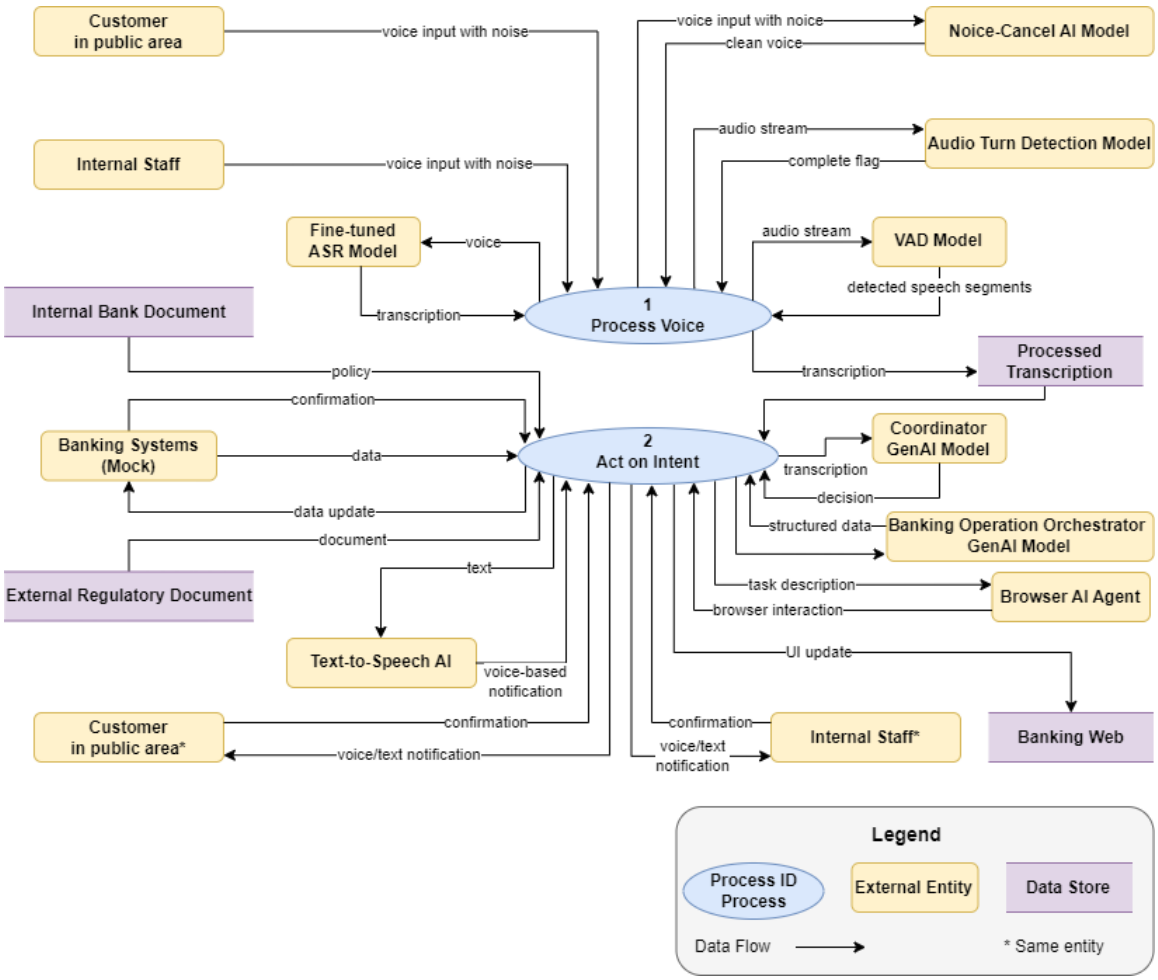

Figure 2: Level 1 Data Flow Diagram

1.3. Level 2 data flow diagram

Process 1.1 – Cancel Noise: The system filters noisy voice input using a Noise-Cancel AI Model to produce clean audio, enhancing clarity for accurate detection and transcription.

Process 1.2 – Detect Voice Activity: The VAD Model detects speech segments within the audio, separating spoken content from silence or noise for smoother downstream processing.

Process 1.3 – Detect Turn: The Audio Turn Detection Model identifies speaker turns between customer and staff, defining clear boundaries to maintain accurate conversation flow before transcription.

Process 1.4 – Transcribe: The Fine-tuned ASR Model (Automatic Speech Recognition) converts processed speech segments into textual transcriptions. This transcription bridges voice interaction and text-based intent recognition.

Process 2.1 – Recognize Intent: The Coordinator GenAI Model extracts intent and key data from transcriptions, converting spoken requests into actionable banking tasks for the next processing stage.

Process 2.2 – Action on Intent Banking Operations: This is the core decision and orchestration process. It takes the recognized intent and interacts with:

- **Banking Systems (Mock)** to validate data and retrieve internal policies or confirmations.
- **External Document sources** for any supporting files.
- **Banking Operation Orchestrator GenAI Model**, which routes and structures the task. The result is structured data and a trigger for the UI to act. This ensures compliance and correct routing for the requested operation.

Process 2.3 – Interact Banking UI: The Browser AI Agent uses structured data to perform actions on the banking website—filling forms, submitting requests, or retrieving data—and returns a success confirmation.

Process 2.4 – Validate & Confirm: The system verifies transaction results and notifies users or staff through voice or text, ensuring both parties are informed of successful or failed actions.

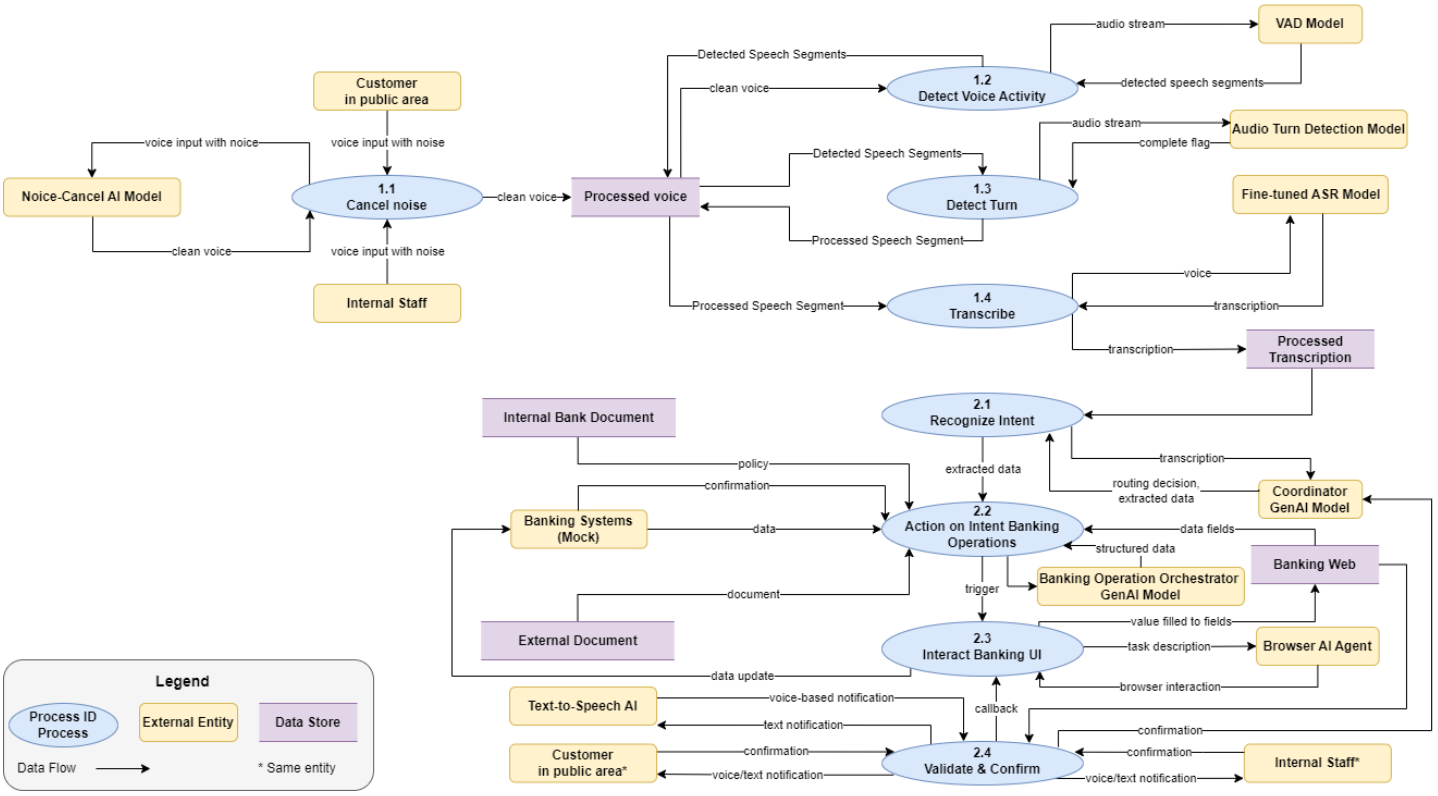

Figure 3: Level 2 Data Flow Diagram

1.4. Level 3 data flow diagram

1.4.1. Level 3 data flow diagram – Loan Submission

User role: Loan applicant

Process 2.1 – Recognize Intent: This process serves as the entry point for understanding the applicant's purpose in the interaction, whether it is to apply for a new loan or update existing loan details. The applicant's voice or text input is captured and processed through Pipecat's **Silero-VAD Analyzer**, which detects speech segments and filters out silence or background noise, ensuring clean transcription. The **Smart-Turn module** then manages multi-turn dialogue, detecting user intent and context from natural conversation. The recognized intent and structured data, including key fields such as "loan type" or "amount," are extracted and forwarded to the **Coordinator GenAI Model**, which validates the intent and determines the next operational step in the workflow. The output of this stage is well-structured and semantically tagged data ready for processing in the subsequent action phase.

Process 2.2.1 – Action on Loan Submission Operation: Once the applicant's intent has been recognized, this process translates it into a concrete operational task for loan submission. The **Coordinator GenAI Model** routes extracted intent data to the **LOS Agent GenAI Model**, which interprets the loan request and structures it into fields compatible with the bank's internal Loan Origination System (LOS). It cross-references policies from internal bank documents and applies regulatory checks using external compliance documentation to ensure that the request meets internal and external standards. The resulting structured information – including applicant details, loan type, and requested amount – is prepared for submission to the **Browser AI Agent** and **Loan Submission Form (Web)**. This stage essentially bridges high-level applicant intent with executable instructions for automated form handling and regulatory compliance.

Process 2.3.1 – Interact Loan Submission Form: At this stage, the system automates the physical submission of the loan application using browser automation technology. The **Browser AI Agent**, leveraging the **Browser-use framework (web-ui)**, performs web interactions that mimic human input – typing applicant details, selecting dropdown options, checking boxes, and clicking submission buttons within the official loan application form at vayonline.vpbank.com.vn. The **LOS Agent GenAI Model** provides task descriptions and structured data, while the browser agent executes them through secure and controlled automation. Any feedback from the web form, such as validation errors or confirmation messages, is returned as a callback to the system. This process ensures accurate, efficient, and consistent submission of loan data without requiring direct manual entry by a bank officer.

Process 2.4 – Validate & Confirm: This final process verifies that the loan submission has been completed successfully and that the data is accurate and compliant. The callback response from the user if the value entry is not the same as the user's intent. Any necessary updates are applied to the **LOS (Mock)** database and external regulatory records. Once the validation is complete, the system generates a confirmation for the applicant. Depending on context, this confirmation can be delivered through **Text-to-Speech AI** as a voice notification or as a text-based message for applicants in public or remote environments. The validation ensures that all applicant data aligns with compliance rules and that the transaction is properly logged within the LOS. This process closes the loan submission loop by confirming completion, maintaining data integrity, and enhancing user trust through immediate, AI-driven acknowledgment.

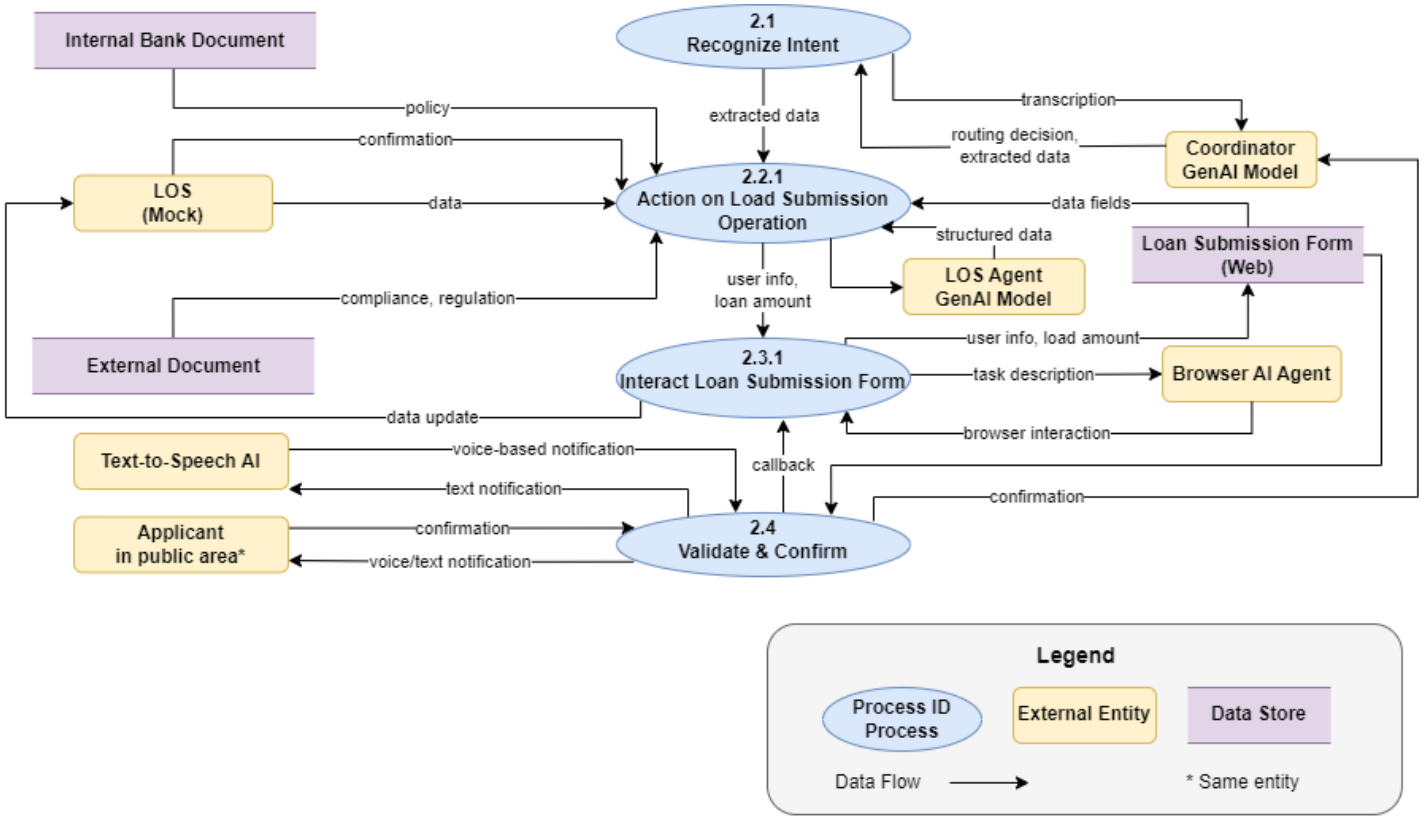

Figure 4: Level 3 Data Flow Diagram – Loan Submission

1.4.2. Level 3 data flow diagram – CRM update customer information

User role: Customer service officer

Process 2.1 – Recognize Intent: This process begins when a customer service officer or internal staff issues a command such as *"Cập nhật địa chỉ khách Nguyễn Văn Bình thành 25A Nguyễn Trãi."* The input is captured and processed through speech analysis modules. Pipecat's Silero-VAD Analyzer detects the active speech segments and the Smart-Turn module interprets conversational context, ensuring that only relevant instructions are passed on. The Coordinator GenAI Model then analyzes the transcription to extract key intent – in this case, a CRM update operation – and identifies relevant entities such as customer name and updated address. The structured data is routed forward for execution, forming the foundation for automated CRM actions.

Process 2.2.2 – Action on CRM Operation: Once the intent to update customer information is recognized, this process translates it into actionable data operations for the Customer Relationship Management (CRM) environment. The Coordinator GenAI Model sends extracted data to the CRM Agent GenAI Model, which organizes it into structured fields (e.g., customer ID, updated address, contact number). The model cross-checks

internal policies and compliance guidelines from internal bank documents and validates the operation against external documents for data protection and record management regulations. The CRM Agent then interacts with the CRM System (Mock) to retrieve relevant customer data and prepares a clean data package for submission to the Browser AI Agent, ensuring that every update complies with both organizational standards and privacy laws.

Process 2.3.2 – Interact Customer Information Management Web: At this stage, the automation of the CRM update task occurs through browser-based actions. The Browser AI Agent simulates a human user interacting with the Customer Information Management Web interface. It inputs updated customer data – such as name, address, or contact details – into designated form fields, selects necessary dropdown options, and submits the information through the web interface. Throughout this process, the CRM Agent GenAI Model provides structured data and task descriptions to guide execution. The Customer Service Officer can monitor or validate the operation in real time. Once the update is submitted, the web system generates a callback confirming the status (success, failure, or data validation error), which is then returned to the central process flow for validation.

Process 2.4 – Validate & Confirm: The final process verifies that the CRM update has been completed successfully and that the data matches compliance and system integrity requirements as well as the Customer Service officer. The callback is sent back if it does not meet the user's intent, and triggered the CRM agent to correct the data from the CRM web interface. Upon successful validation, a confirmation is generated and sent to the Customer Service Officer. Depending on configuration, this can be a voice notification delivered through Text–to–Speech AI or a text–based confirmation for record tracking. This ensures transparency, data accuracy, and immediate feedback to the officer, closing the loop in the automated CRM update workflow.

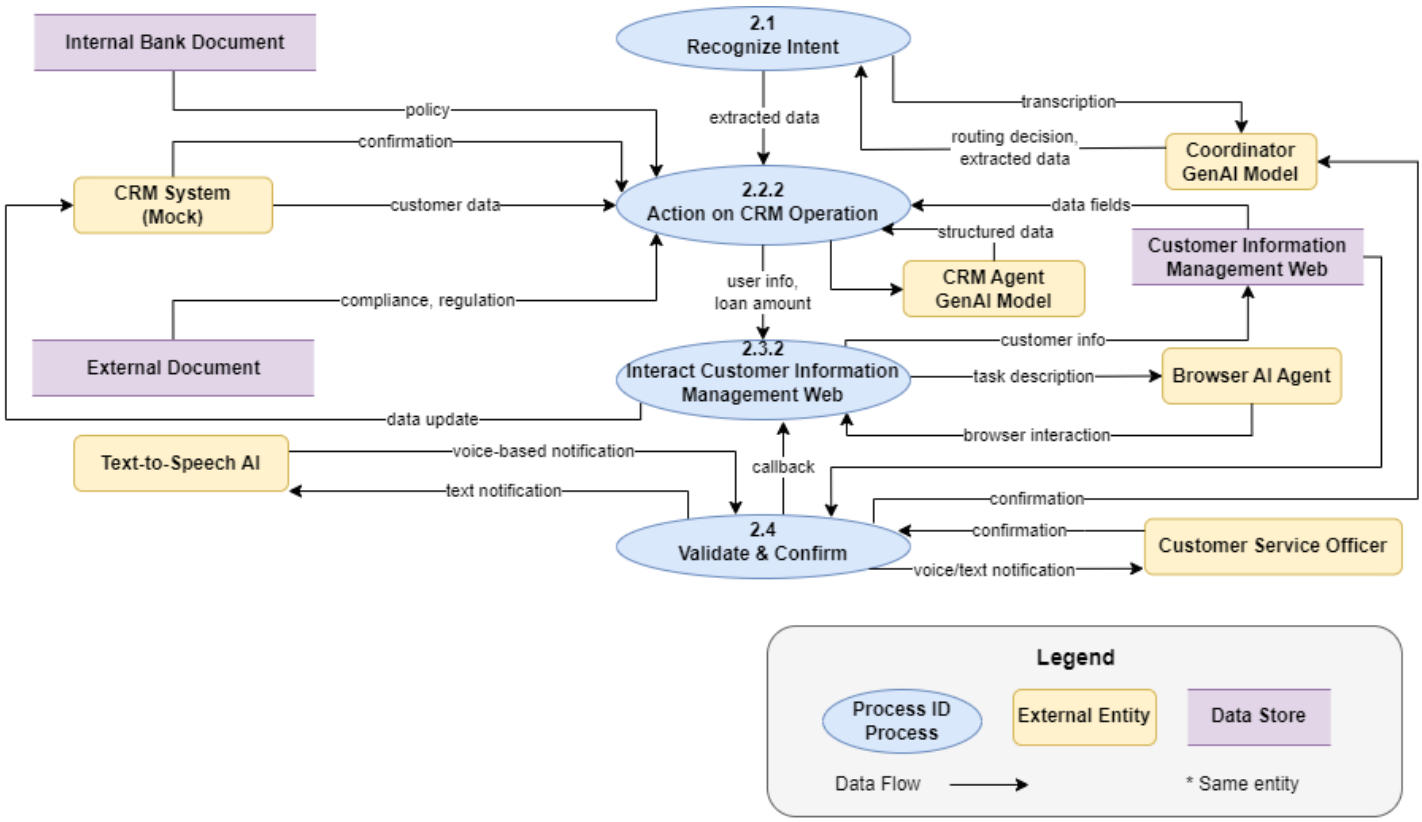

Figure 5: Level 3 Data Flow Diagram – CRM update customer information

1.4.3. Level 3 data flow diagram – HR writes job description

User role: HR officer

Process 2.1 – Recognize Intent: The workflow begins when the HR officer expresses a command such as *"Đăng mô tả công việc vị trí Chuyên viên Tín dụng lên trang tuyển dụng."* The system detects and processes the voice input, identifies intent to create or update the posting, extracts key job details, and forwards them for policy and compliance alignment.

Process 2.2.3 – Action on HR Operation: After identifying intent, the system converts the HR officer's command into structured job data. The HR Agent organizes details, references internal and external regulations for compliance, updates the HR system, and prepares the finalized job description for posting.

Process 2.3.3 – Interact Recruitment Web: The system automates job posting through the Browser AI Agent, which logs in, fills out required fields, uploads documents, and submits the listing. It then receives confirmation or error feedback from the recruitment platform for validation.

Process 2.4 – Validate & Confirm: In this final process, the system validates that the job description has been uploaded successfully and that all required fields meet the recruitment web's standards and correctly filled in based on the HR officer's purpose. The callback progress is triggered to ensure alignment between the recruitment platform and the HR staff, updating internal HR data where necessary. The HR Officer receives a confirmation of successful posting. Depending on the communication mode, this may be delivered as a text notification or a voice alert via Text-to-Speech AI, confirming that the new job posting is live. This process closes the HR operation loop by verifying accuracy, ensuring compliance, and providing timely feedback to the HR officer.

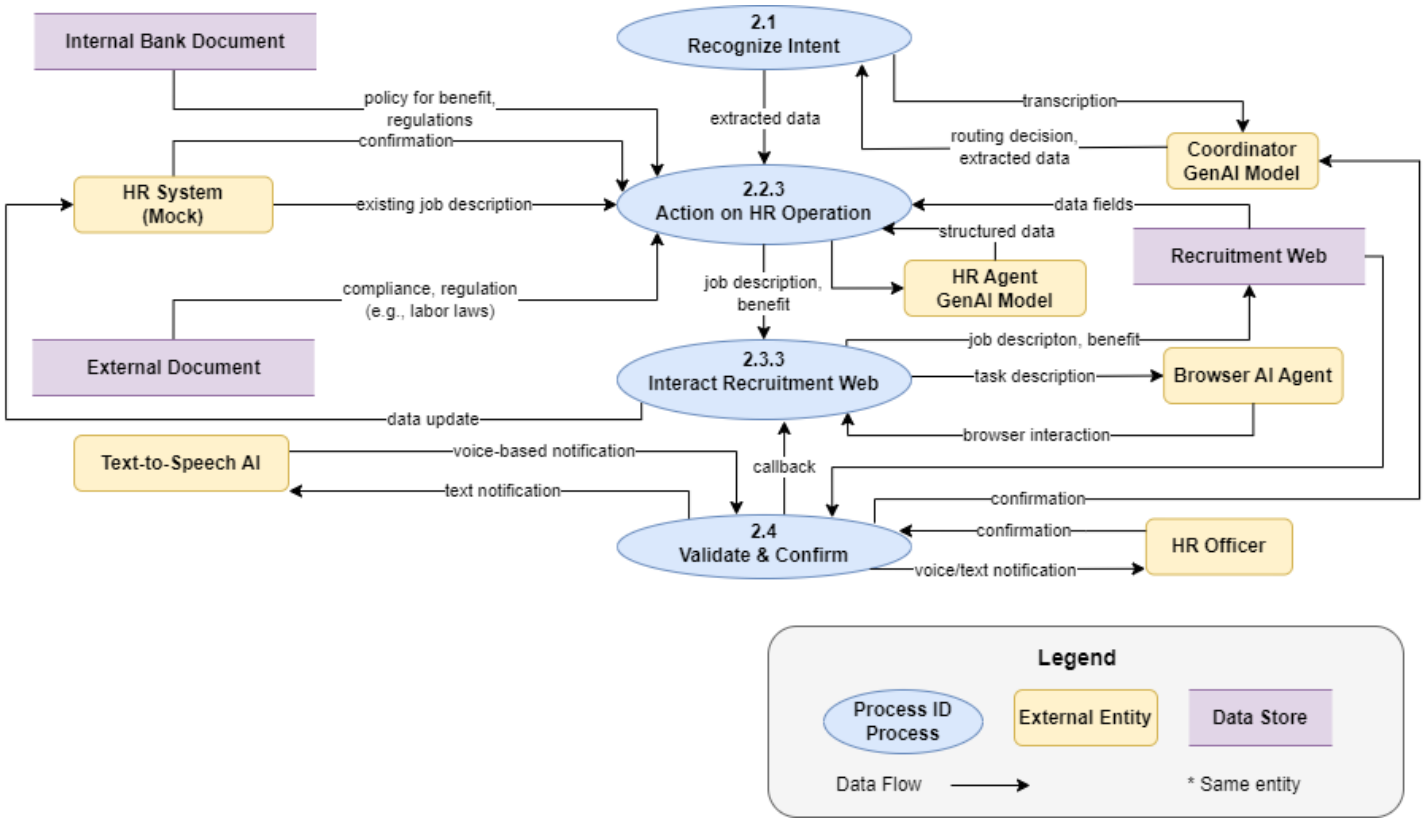

Figure 6: Level 3 Data Flow Diagram – HR writes job description

1.4.4. Level 3 data flow diagram – Compliance Validation

User role: Compliance officer

Process 2.1 – "Recognize Intent": The process begins when the Compliance Officer provides a voice input to the system. The voice data is first captured and processed using **Voice Activity Detection (VAD)** – specifically, something like the *Silero VAD Analyzer* (from pipecat.ai documentation). This VAD module detects when speech starts and stops, helping the system to cleanly segment the spoken input for accurate transcription. The audio is then transcribed into text by the **Coordinator GenAI Model**, which interprets the content and extracts intent, such as identifying the compliance case or policy mentioned. The extracted intent data (e.g., type of report, severity, or compliance policy reference) is sent forward for validation.

Process 2.2.3 – "Action on Compliance Validation Operation": Once the intent and extracted data are identified, the system executes the **Compliance Validation Operation**. This process uses structured data and compliance regulations retrieved from both internal bank documents and external documents. The Compliance Management System (Mock) supplies internal policy references and confirmation requirements, while external sources provide updated regulatory standards. The Coordinator GenAI Model routes data fields and ensures consistency before passing them to the Compliance Validation Agent. This agent structures the data into standardized formats suitable for submission to the Compliance Portal Web.

Process 2.3.3 – "Interact Compliance Portal Web": The Browser AI Agent acts as an intermediary, simulating or automating browser interactions such as filling forms, uploading attachments, and submitting the compliance report. It receives structured task descriptions from the previous step (e.g., field mappings, form elements like those seen in VPBank's online loan submission form). The agent interacts with the web interface to upload the compliance report and sends confirmation or any task exceptions back to the validation process.

Process 2.4 – "Validate & Confirm": After the submission, the system executes Validate & Confirm to ensure the compliance report was first filled correctly, then successfully registered on the portal. The Compliance Officer returns confirmation signals – text-based and voice-based. If he/she does not confirm, then the process "Interact Compliance Portal Web" is triggered again to edit the fields. On the other hand, The Text-to-Speech AI component converts textual confirmation into a voice notification, which is delivered back to the Compliance Officer. This allows the officer to receive real-time audible updates confirming successful submission, error notices, or validation outcomes. The Compliance Management System (Mock) and external documents sources may also be updated if new compliance information or corrections are identified.

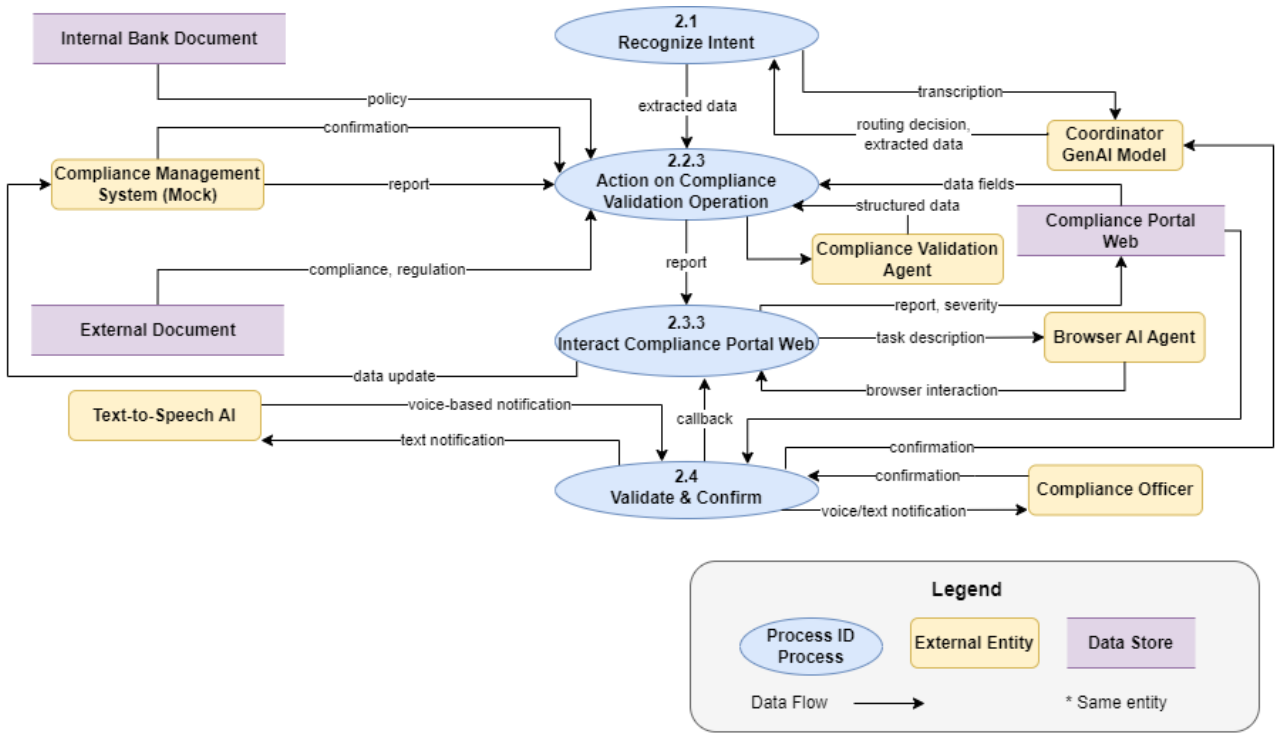

Figure 7: Level 3 Data Flow Diagram – Compliance Validation

2. Tech-stack

Category / Layer	**Technology / Service**
Frontend Framework	Typescript, ReactJS 19
Backend Framework	Python, FastAPI
Conversational Orchestration	PipeCat
Multi-Agent Framework	LangGraph

Speech-to-Text	Whisper (PhoWhisper)
Text-to-Speech	ElevenLabs
Generative AI & NLU	Amazon Bedrock
Voice Processing	Voice Activity Detection (VAD)
State Management	PipeCat In-memory State
Task Execution	Custom Execution Agent (Browser Automation)

3 Evaluation Metrics

To objectively evaluate the performance and accuracy of the voice automation workflow, multiple quantitative and qualitative metrics are defined. Each metric is directly tied to system reliability, user experience, and business value

Metric	Target	Validation Method
Workflow Correction	> 90 %	Manual / Auto validation
Word Correction	95%	Text Comparison
Voice Latency	< 4 s	Pipecat Latency Metric

User Satisfaction	> 4.5 / 5	Pilot survey

Table 9 – Evaluation Metrics for Voice AI System Performance

3.1 Workflow Correction Rate (WCR)

Definition: Measures the accuracy of the AI agent in correctly filling out form fields or workflow objects according to predefined templates.

Formula:

$$WCR = Total\ number\ of\ required\ fields/Number\ of\ correctly\ filled\ fields \times 100\%$$

Example:

- If a loan application form has 10 fields, and the bot fills 9 fields correctly, then WCR = 90%.
- If only 5 fields are correct, then WCR = 50%.

This metric evaluates the structural correctness of the workflow – whether the agent follows the intended process logic (e.g., correctly identifying “Customer Name”, “Loan Amount”, “Date”, etc.).

3.2 Word Correction Rate (Word Accuracy Score)

Definition: Assesses the textual accuracy of what the AI fills into each field, by comparing the generated content with the ground–truth data in the test case.

Formula:

$$Word\ Accuracy = 1 - Edit\ Distance\ (Levenshtein)/Length\ of\ Ground\ Truth$$

Example:
If the correct name is **“Hiếu Nghị”** and the bot writes **“Hiếu Nghĩa”**, one character mismatch is detected → ~85–90% accuracy.

Purpose: Evaluates **semantic fidelity** and **linguistic correctness**, especially important for Vietnamese tone–sensitive fields such as names, addresses, and financial terms.

Application:

- Automated form–filling verification
- Named entity accuracy for banking customers
- Context validation for text generation from voice

3.3. End–to–End Voice Latency (E2E–VL)

Definition: Measures the total response time between the end of the user's spoken input and the start of the bot's voice or textual reply.

Formula:

$$E2E\_VL = t\ (bot\ response) - t\ (user\ speak)$$

Ideal Range: < **4.0 seconds** for natural conversational flow. Quantifies system responsiveness and ensures **real–time voice interaction** feels human–like.

Measurement:

- Collected automatically from Pipecat logs
- Time includes VAD → STT → LLM→ TTS Pipeline

Application Used during performance testing of:

- Real–time customer service scenarios
- On–premise vs cloud latency comparison
- Network optimization (WebRTC)

3.4. User Satisfaction Score (USS)

Definition: A qualitative metric representing how satisfied end–users are with the overall voice interaction experience.

Measurement Methods:

- Post–interaction surveys (1–5 scale or Net Promoter Score).

- Sentiment analysis of voice or text responses.
- Session completion rate vs dropout rate.

Formula (Example):

$$\text{USS} = \text{Number of positive ratings/Total Responses} \times 100\%$$

Purpose: Tracks user experience and adoption, providing early indicators of business value and system usability.

Application:

- Business KPI dashboard for VPBank AI adoption.
- Continuous improvement cycle (data feedback → fine–tuning).

4. Security and Governance

Encryption: All data – from user audio to responses – is encrypted using **AES-256 at rest (via AWS KMS)** and **TLS in transit**. This secures stored transcripts, agent states, and real-time WebRTC or API, ensuring sensitive content cannot be intercepted or altered.

Authentication: Amazon Cognito enforces **Role-Based Access Control (RBAC)**, defining which agents or workflows users can access to ensure to log in securely using authenticated credentials.

PII Protection: Sensitive data such as names or account numbers undergo **pre-inference masking**, ensures personally identifiable information is never stored in database unmasked.

Logging: All agent actions, API calls, and voice interactions are recorded in **AWS S3 and CloudTrail** for a complete, compliant **audit trail** supporting financial sector governance and accountability.

Isolation: AI agents and Bedrock services operate within VPC Endpoints, keeping all traffic off the public internet.

5. Enhance voice recognition system

Fine-Tuned Model: Built on the advanced PhoWhisper foundation, our smart voice detection module is carefully fine-tuned using a high-quality dataset of real Vietnamese voices, including numerous regional accents and both male and female speakers.

Bilingual Capability: Training incorporates both pure Vietnamese and bilingual Vietnamese-English speech samples, enabling accurate recognition for code-mixed interactions commonly found in banking and customer service.

Continuous Improvement: The system is retrained regularly with updated voice data, allowing it to adapt to new accents, language patterns, and user requirements over time.

Superior Accuracy: Superior Accuracy: Achieves Word Error Rate (WER) below 5% for Vietnamese and bilingual (Vietnamese-English) conversations, much better than the regular PhoWhisper model, which shows WER of 7–9% for Vietnamese and 27–30% for English speech.

Business Impact: Enhanced recognition speed and reliability enable compliance operations to be completed faster and with fewer errors, supporting efficient workflows and a smoother experience for both staff and customers.

6. Continuous Improvement

Data Feedback Loop: All anonymized user voice commands and transcriptions from the Pipecat conversational flows are securely stored in Amazon S3. These datasets are automatically tagged and versioned for auditability. At the end of each month, the S3 dataset is used to retrain and fine–tune speech recognition and dialog models through Amazon SageMaker pipelines. This enables the system to continuously adapt to new conversational patterns, business workflows (Loan/HR/CRM), and user intents without exposing any personal or financial data – thanks to pre–inference masking and anonymization mechanisms applied during ingestion.

Accent Adaptation: To improve voice recognition accuracy, regional and local accent samples are periodically collected through PhoWhisper (Speech–to–Text) logs. These anonymized samples feed into a fine–tuning process for the speech recognition sub–model, enabling the pipeline to adapt to Vietnamese–English mixed speech or other regional language variations common among users. This continuous accent adaptation ensures better transcription accuracy and smoother downstream interaction for the Workflow Router Agent and all connected domain agents.

Model Observability: The AI pipeline integrates Prometheus and Grafana for real–time observability and model health tracking. Metrics such as model drift, intent classification accuracy, and latency of Pipecat Flows are continuously monitored. Whenever performance degradation or drift is detected, alerts are automatically triggered to data scientists or MLOps engineers, prompting evaluation and retraining cycles in SageMaker. This monitoring loop ensures that both the Generative AI (Bedrock) and Workflow Automation (LangGraph + Pipecat) layers remain accurate, stable, and compliant over time.

ARCHITECTURE OF SOLUTION

1. High–Level Architecture

1.1. System Overview

The solution overview is designed in the figure below:

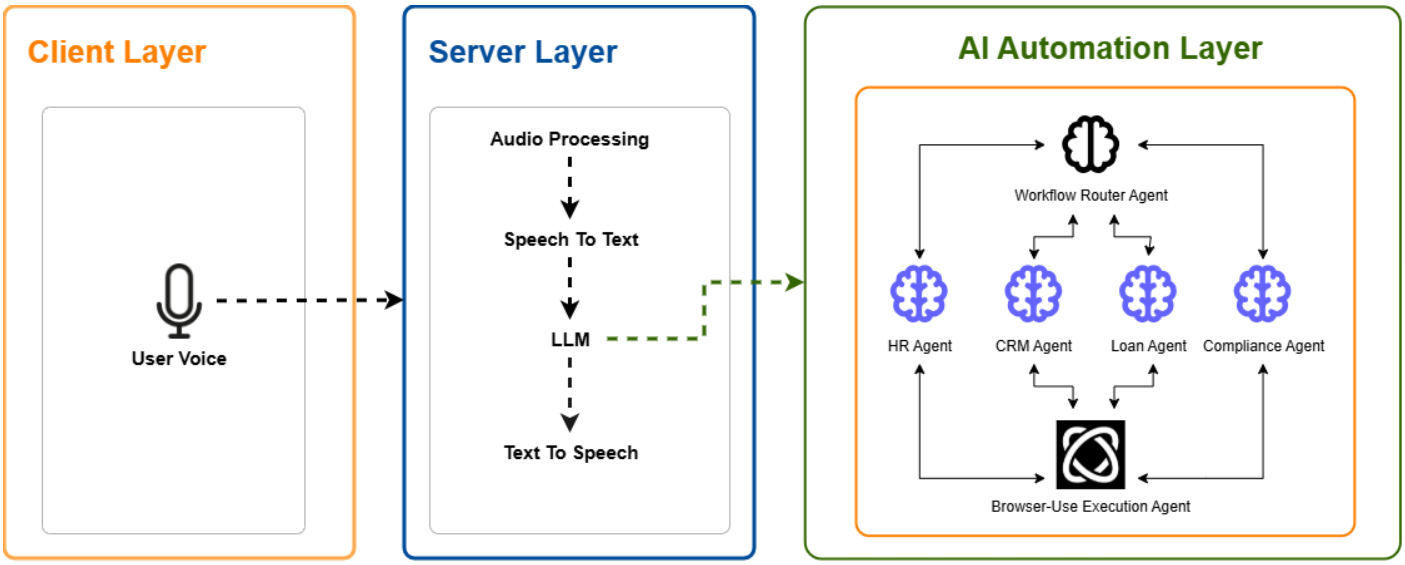

Figure 8: Abstract System Overview Architecture Diagram

Architecture Principles:

1. Simplicity

The overall system follows a layered architecture consisting of three main components:

Client Layer – User Interaction Interface: This layer includes all components that directly interact with end users, such as web or mobile applications. The user's voice is captured through a microphone and transmitted securely via **WebRTC Transport** for real–time audio streaming. The goal of this layer is to deliver a natural, fast, and frictionless voice–based interaction experience.

Server Layer – Processing and Intelligence Core: *This layer handles all voice understanding, intent recognition, and workflow execution logic. It consists of two sub–layers:*

Voice Layer: Processes incoming audio and converts it into text through a Speech–to–Text (STT) module. The resulting text is passed to a **Large Language Model (LLM)** that interprets the business context and determines the user's intent. The system can then generate a voice response using **Text–to–Speech (TTS)** and deliver it back to the client via **RTVI Events**, enabling a natural two–way conversation. This is the layer where **Pipecat** and **WebRTC** integration enables real–time, bidirectional conversation.

AI Automation Layer: The AI Automation Layer consists of **domain-specific AI Agents** such as the HR Agent, CRM Agent, and Loan Agent. A Workflow Router Agent identifies the user's intent and directs requests to the appropriate agent. Each agent can autonomously perform actions on business systems through the Browser–Use Execution Agent, including editing, deleting, or updating form data and triggering functional buttons on web interfaces - effectively simulating human interactions similar to robotic process automation (RPA). This modular, agent-based architecture makes the system highly extensible, allowing new agents or workflows to be added independently without affecting existing operations.

2. Scalability

The system is designed using a microservice-based architecture that enables independent scaling of each functional layer. The Server Layer, which hosts the Voice Processing (STT/TTS) and Automation Workflow components, leverages **Amazon ECS with AWS Fargate** and **ECS Service Auto Scaling** to dynamically adjust compute capacity according to workload demand. This elasticity ensures high responsiveness during conversation peaks – for instance, when multiple users interact concurrently through WebRTC channels – while maintaining optimal cost efficiency during off-peak periods.

3. Security

All incoming traffic is routed through a centralized API Gateway, which acts as the single secure entry point to backend services. The gateway enforces strict authentication and authorization policies using Amazon Cognito (RBAC) integrated with VPBank's SSO identity provider.

Voice streams transmitted over WebRTC are end–to–end encrypted, ensuring privacy for real–time audio sessions. Sensitive data is protected through multiple layers of governance:

- AES–256 encryption at rest managed by AWS KMS
- TLS 1.2+ encryption in transit
- Pre–inference masking of Personally Identifiable Information (PII) before any text or voice data reaches inference endpoints (e.g., Whisper, Bedrock)

- Comprehensive audit logging via S3 audit trails and CloudTrail with 7–year retention
- Network isolation through VPC Endpoints and PrivateLink for hybrid or on–premises connections

Additionally, AWS Bedrock Guardrails are applied to constrain LLM behavior, ensuring compliance with corporate data policies and regulatory requirements.

4. Cost Efficiency

The architecture follows a **serverless–first** and **managed–service–first** design philosophy to minimize operational complexity and cost.

Core AI capabilities are delegated to fully managed AWS services, eliminating the need to maintain GPU clusters or audio–processing servers. Key components include:

- **PhoWhisper (managed via Pipecat)** for **Speech–to–Text (STT)**
- **ElevenLabs** for **Text–to–Speech (TTS)**
- **Amazon Bedrock Claude** for **LLM reasoning, workflow orchestration, and dialog management**

These services are billed only when invoked, ensuring that costs scale linearly with actual system usage. By combining **on–demand execution**, **auto–scaling**, and **Pipecat's managed voice/agent orchestration**, the solution achieves high availability and low latency while remaining **cost–efficient and operationally lean**.

1.2. Component Details – Client Layer

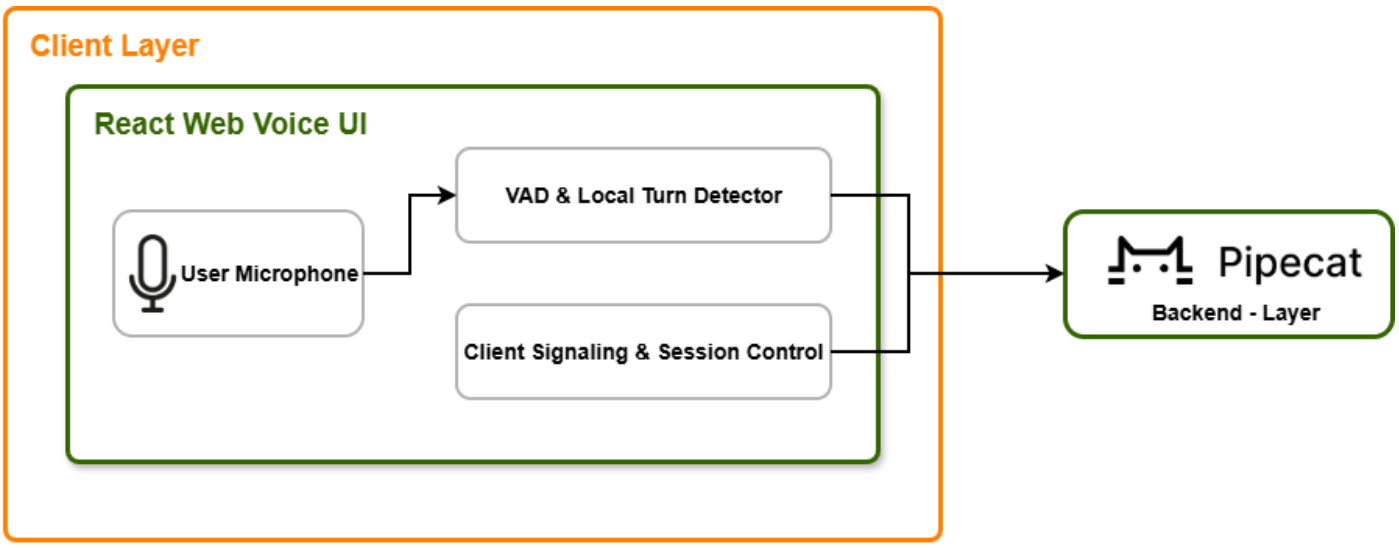

Figure 9: Client Layer Diagram

The client architecture orchestrates multiple layers – audio capture, local enhancement, activity detection, and real–time communication – before handing off to backend AI services. At the heart of this architecture lies a WebRTC–based audio pipeline, powered by **Pipecat's client SDK**. The browser initializes audio capture through the WebRTC getUserMedia() API, which opens a live audio stream from the user's microphone. This stream is encoded in a lightweight codec such as Opus or raw PCM and processed locally before being transmitted. WebRTC's built–in noise reduction and echo cancellation provide a first level of cleanup, but to achieve production–grade clarity, the system integrates **Pipecat's Krisp feature** for AI–powered noise suppression.

Krisp runs as an **on–device WebAssembly audio processor**, which means that the audio never leaves the browser for preprocessing. Instead, Krisp intercepts the raw microphone stream, identifies and removes environmental noise – such as keyboard clicks, fan hum, or background chatter – while preserving the natural tone and intelligibility of the speaker's voice. This real–time enhancement step dramatically improves both the transcription accuracy of downstream Speech–to–Text models and the overall responsiveness of the system. Because Krisp operates locally, it also strengthens privacy and reduces cloud processing costs.

Once the audio is cleaned, it flows into the **Voice Activity Detection (VAD)** module. The VAD continuously analyzes short segments of the audio stream to determine whether the user is actively speaking or silent. When speech is detected, the system begins streaming audio frames to the backend; when silence is sustained for a certain threshold, it automatically pauses transmission. This intelligent gating mechanism avoids sending unnecessary silent data to the backend, which optimizes both cost and bandwidth usage. It also enables a more natural user experience – the user simply speaks, pauses, and continues without needing to manually trigger recording.

When speech is active, audio frames are packaged and transmitted through a **secure WebRTC** connection to the Pipecat gateway. WebRTC are essential for real–time, bidirectional communication: the client sends binary audio chunks upstream while simultaneously receiving transcription updates, intent parsing, and AI–generated responses downstream. The connection is persistent and event–driven, ensuring millisecond–level latency between user speech and the corresponding AI feedback.

On the user interface, this architecture supports live transcription visualization – partial text updates appear on the screen as the model processes each segment. The client also renders playback of synthesized speech responses (from AWS Polly or other TTS providers) and can animate "listening" or "thinking" indicators based on the VAD state. The entire flow is orchestrated through React hooks or equivalent frontend logic, using the Pipecat client SDK to handle Krisp activation, VAD thresholds, and WebRTCsession management.

Security is embedded throughout the client pipeline. All data is transmitted over encrypted **WebRTC** channels, and each session includes authentication metadata (such as JWTs or signed tokens). Importantly, because Krisp operates entirely within the user's device, no raw voice data is stored or transmitted externally during preprocessing.

1.3. Component Details – Voice Processing

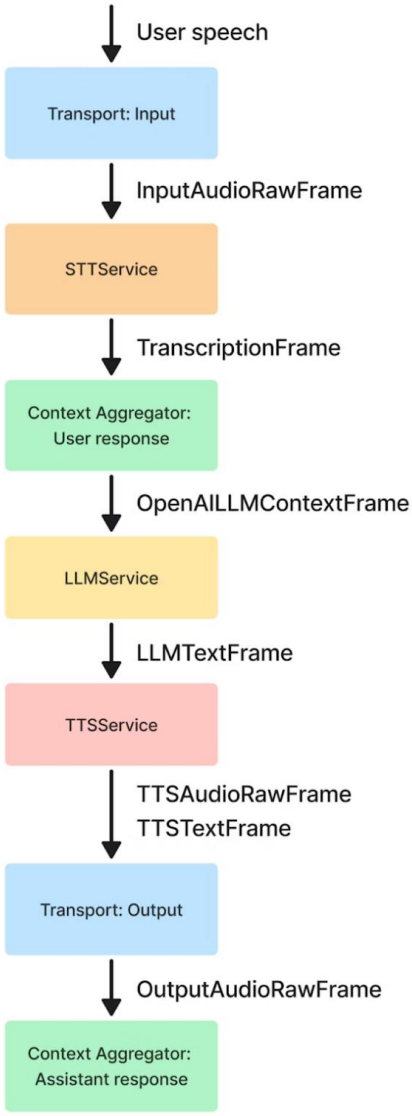

Figure 10: Detailed Voice Processing Layer Diagram

The **Voice Pipeline** serves as the entry point of the conversational system, enabling real–time, low–latency voice interaction between users and backend AI agents. Built on Pipecat, an open–source audio orchestration framework, this layer integrates multiple subcomponents to handle streaming, detection, transcription, and preprocessing efficiently.

Pipecat Framework: Pipecat acts as the core orchestrator for audio data flow. It manages:

- Audio streaming and buffering across the network
- Bidirectional WebRTC connections
- Real–time synchronization with AI inference layers

Through its modular pipeline design, Pipecat ensures that each processing unit (VAD, STT, TTS, and Agent Orchestration) can be independently optimized or replaced without affecting overall stability.

WebRTC Connection: The client (web or mobile) communicates with the server using WebRTC, which Pipecat handles natively. This ensures:

- Low–latency, full–duplex voice transmission
- End–to–end encryption (SRTP/TLS)
- Adaptive bitrate streaming, optimizing audio quality under fluctuating network conditions

PhoWhisper Speech–to–Text (STT): After VAD filtering, the audio stream is routed to fine-tuned PhoWhisper, a Vietnamese–optimized speech–to–text model derived from **OpenAI Whisper.** The fine-tuned progress is described in section 1.5 – Fine-Tune Automatic Speech Recognition Model Detail in “ARCHITECTURE OF SOLUTION”.

The model provides**:**

- High accuracy for tonal languages, preserving diacritics and phonetic nuances in Vietnamese.
- Dialect robustness, handling **Northern, Central, and Southern** variations fluently.
- Banking domain adaptation, recognizing terminology specific to loan origination, CRM updates, and HR workflows.

Deployment Flexibility and Data Governance

Unlike most commercial transcription APIs, PhoWhisper supports on–premise or private–cloud deployment, ensuring:

- Full data sovereignty – no customer audio leaves the organization’s controlled environment.
- Compliance with internal and regional banking regulations.

- Customizable fine–tuning on proprietary datasets using AWS SageMaker for continuous improvement.

Domain and Accent Adaptation, PhoWhisper is continually improved through:

- Domain–specific fine–tuning, using banking, compliance, or HR datasets for contextual understanding.
- Accent adaptation, trained on Vietnamese regional speech samples to improve recognition accuracy across user demographics.
- Feedback loop integration, where anonymized user commands are periodically aggregated and used for monthly model refinement.

These adaptation mechanisms are critical to sustaining long–term accuracy and user trust in voice–driven enterprise automation.

Open–Source Advantage

PhoWhisper and Pipecat, being open–source frameworks, provide flexibility beyond commercial lock–ins.

- Extend pipeline logic (e.g., integrate Krisp for noise suppression).
- Adjust VAD sensitivity or transcription thresholds.
- Customize downstream routing to proprietary workflow agents.

1.4. Component Details – AI Agent System

The Voice–AI automation framework is designed as a modular multi–agent system, where each agent specializes in a distinct enterprise function. The **Workflow Router Agent** orchestrates the overall process, routing requests to task–specific agents – including the Loan, CRM, HR, Compliance, and Browser Execution agents – ensuring seamless task automation, security, and compliance.

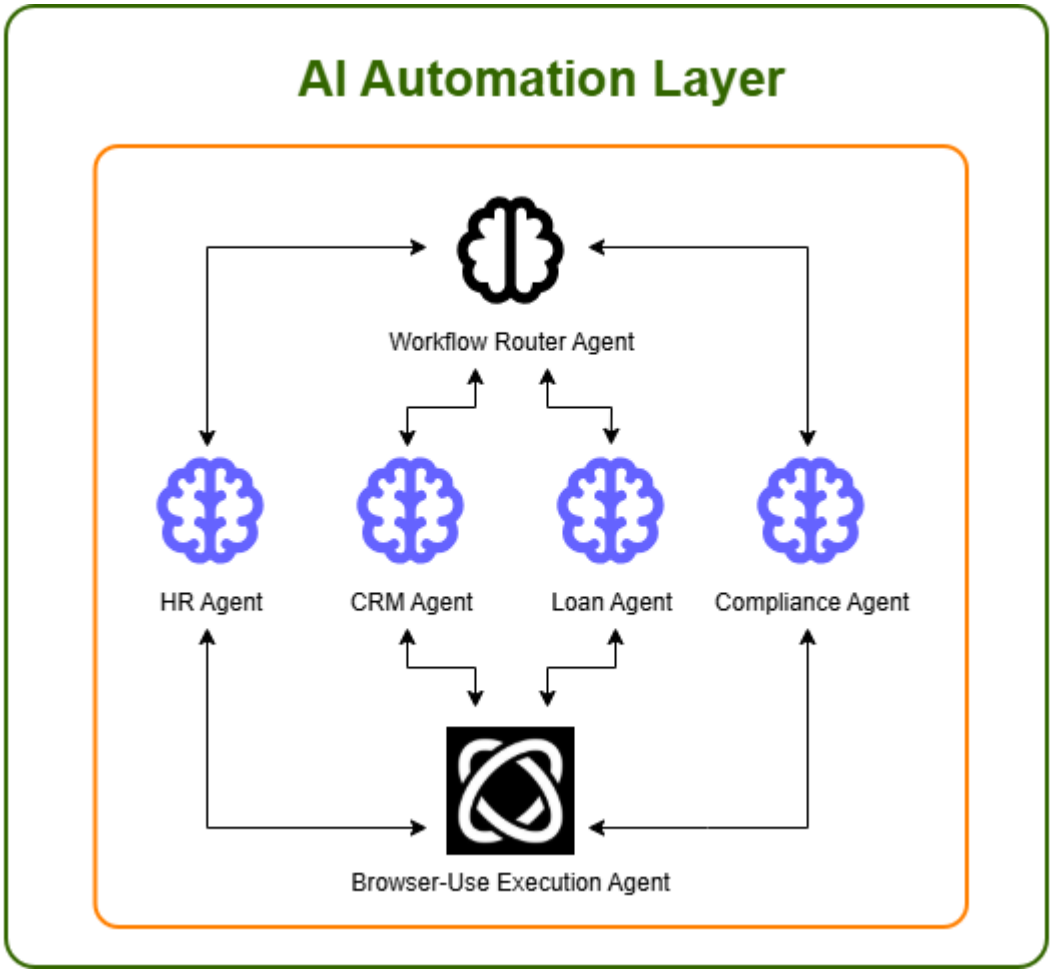

Figure 11: Multi–Agent Layer Diagram

Workflow Router Agent

This is the **core orchestration agent**, responsible for **intent recognition**, **context management**, and **inter–agent coordination**. It receives natural language inputs (voice or text), identifies the corresponding business domain, and delegates the task to the appropriate specialized agent.

Example workflow: A user says: “Fill in customer Nguyễn Văn An, loan amount 500 million, term 24 months.” → The Router Agent detects a *Retail Lending* intent. → It then dispatches the request to the **Loan Agent** for structured processing.

Key functions:

- Intent classification and routing
- Context and session management
- Agent–to–agent communication via API, HTTP, or WebRTC
- Logging, audit trail generation, and compliance reporting hooks

Compliance Agent

Handles **Use Case 4 – Risk, Compliance & Audit Automation**. This agent ensures all automated workflows comply with regulatory and institutional policies, particularly in AML/KYC reporting and data governance.

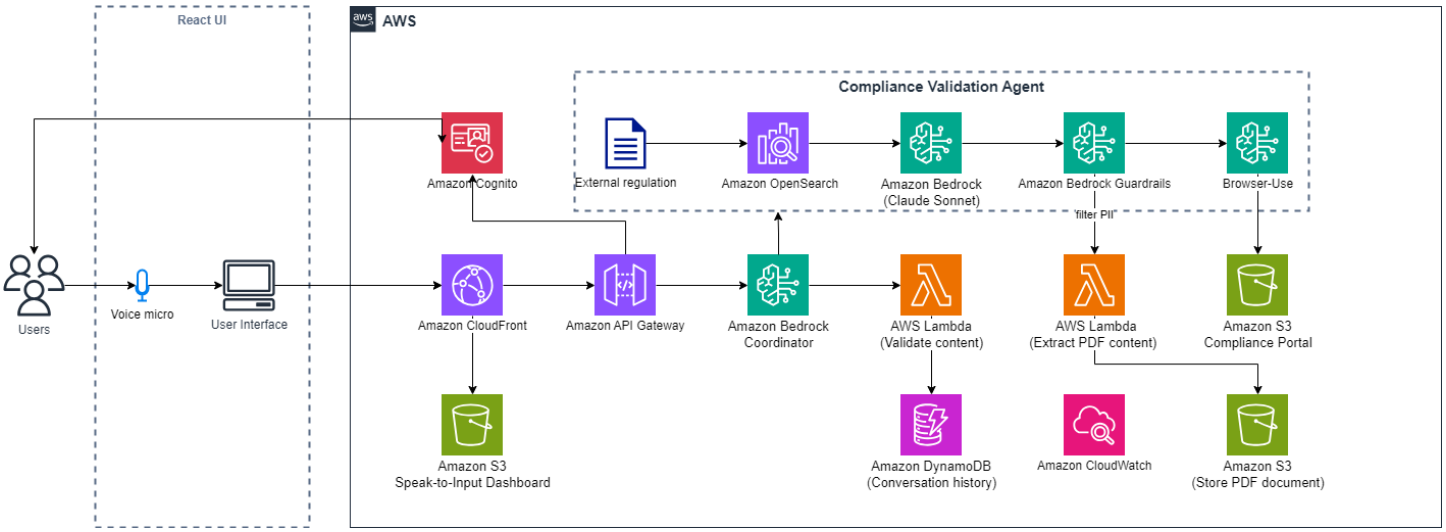

Figure 12 – Compliance Agent Architecture

(LOS agent, CRM agent and HR agent has the same architecture)

Functions:

- Automates report generation for AML, KYC, and regulatory submissions
- Cross–validates extracted data against internal systems for consistency
- Logs every action with a timestamp and user signature for auditability
- Enforces compliance with GDPR, PDPA, and SBV (State Bank of Vietnam) data retention regulations

Example:

A compliance officer states: “Generate September AML report, check completion status, and verify no violations.”

→ The Compliance Agent retrieves data, fills the report, validates entries, and archives all logs for audit trail compliance.

The Compliance Validation Agent implements a comprehensive AWS serverless architecture designed for automated document compliance filling and checking against regulatory standards and organizational policies.

Workflow Process

Frontend Layer

- React UI provides the user interface for reading the compliance report.
- Amazon CloudFront delivers the application with global distribution.
- Amazon S3 hosts static web assets.

API and Authentication Layer

- Amazon API Gateway manages API requests and routing to backend services.
- Amazon Cognito handles user authentication and role-based access control.
- Ensures secure document submission and compliance report access.

Document Processing Pipeline

- AWS Lambda (Extract PDF content) processes uploaded documents for compliance analysis.
- Amazon S3 (Store PDF document) securely stores original documents with audit trails.

AI-Powered Compliance Validation

- AWS Lambda (Validate content) orchestrates the compliance checking process.
- Knowledge Base contains regulatory frameworks and compliance rules:
- Amazon Bedrock Guardrails enforce content safety and compliance boundaries
- Amazon Bedrock (Claude 4.5 Sonnet) performs intelligent document analysis
- Browser-Use to interact with the compliance portal website.
- Amazon OpenSearch enables semantic search across compliance databases

Compliance Analysis Engine

The validation engine performs:

- Regulatory Compliance Checking: Validates documents against industry standards (UCP 600, ISBP 821,...).
- Policy Adherence Verification: Ensures alignment with organizational policies.
- Risk Assessment: Identifies potential compliance violations and risk levels.
- Gap Analysis: Highlights missing required elements or documentation.

Data Storage and Audit Layer

- Amazon S3 (Output) stores compliance reports and validation results.
- Amazon DynamoDB maintains audit trails, conversation history, and compliance status.
- Full traceability for regulatory reporting requirements.

Monitoring and Observability

- Amazon CloudWatch provides comprehensive monitoring, alerting, and audit logging.
- Real-time compliance dashboard and automated notifications for violations.

Loan Agent

Handles **Use Case 1 – Loan Origination & KYC Automation**. It converts spoken instructions into structured loan application data and ensures pre–submission validation.

Functions:

- Parses natural language commands from Relationship Managers (RMs)
- Extracts entities such as *customer name*, *loan amount*, and *term length*
- Interfaces with the **Browser–Use Execution Agent** to fill loan origination forms in the LOS system

Requests confirmation and correction before final submission

CRM Agent

Handles **Use Case 2 – CRM Update & Customer Interaction Logging**. It keeps customer data synchronized across systems without manual entry.

Functions:

- Interprets natural commands, e.g., “Update customer Nguyễn Văn Bình’s address to 25A Nguyễn Trãi.”
- Executes updates via CRM APIs or browser automation
- Logs all interactions and updates automatically to ensure traceability

HR Agent

Handles **Use Case 3 – HR & Internal Workflow Automation**. It automates employee self–service actions and HR form processing.

Functions:

- Parses input like “Create a leave request from Oct 22 to 24 for personal reasons.”
- Identifies the relevant HR form, fills and validates data
- Submits through **Browser–Use Execution Agent**, with confirmation sent to the employee
- Ensures role–based access control (RBAC) for internal compliance

Browser–Use Execution Agent

This is the **execution layer** of the automation framework – effectively the system’s “hands.” It performs browser–level actions securely through sandboxed environments

Functions:

- Automates form–filling, button clicks, and file uploads
- Interacts with enterprise portals (LOS, CRM, HRIS, Compliance dashboards)
- Captures screenshots and action logs for traceability
- Operates within a secure, isolated environment with TLS encryption

Context Management and Inter-Agent Communication

In this multi-agent system, The Workflow Router Agent serves as the central orchestrator, managing a shared context that stores all session information (history, user intent, collected data like customer name or loan amount). When a user makes a request, the Router Agent analyzes it, attaches the relevant context, and delegates the task to a **specialized agent** (e.g., Loan Agent, CRM Agent). That agent performs its function, updates the context with new data (e.g., “KYC status: verified”), and returns the result to the Router Agent. This ensures all agents operate from a **single, consistent source of truth**, enabling smooth handling of complex, multi-step workflows without losing information.

1.5. Fine-Tune Automatic Speech Recognition Model Detail

The PhoWhisper model is fine-tuned through a structured pipeline, designed to adapt a pre-trained speech-to-text model for high accuracy in Vietnamese and bilingual enterprise environments

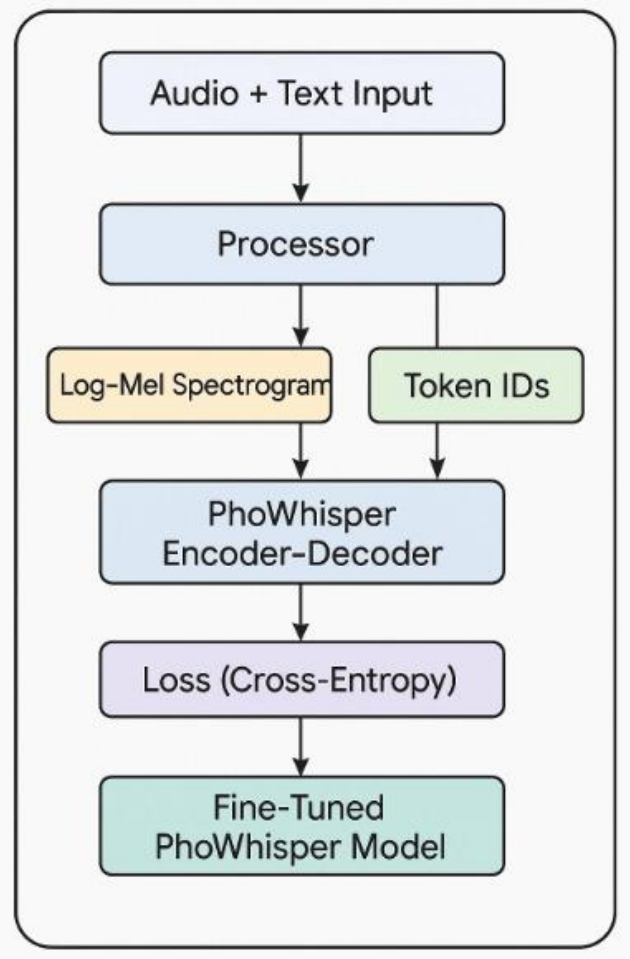

Figure 13 : PhoWhisper Model Fine-Tuning Pipeline

Audio + Text Input: The training process starts by pairing audio recordings with their corresponding text transcripts. This provides the model with both what was said (audio) and the correct answer (text).

Processor: The processor prepares each pair for model ingestion. It transforms:

- Audio into log-Mel spectrograms, a visual format representing sound frequencies over time that helps the model "see" the sounds.
- Text into token IDs, turning words and sentences into sequences of numbers that the model can understand.

Log-Mel Spectrogram & Token IDs: These outputs represent the two main information streams going into the model:

- The audio stream (log-Mel spectrogram) captures all the acoustic and phonetic detail in the voice recording.
- The text stream (token IDs) encodes the linguistic content for training.

PhoWhisper Encoder-Decoder: Both the spectrogram and token IDs feed into the PhoWhisper encoder-decoder neural network:

- The encoder analyzes the audio features, searching for patterns and meaning.
- The decoder predicts the text transcription, learning the mapping between the spoken sound and written language.

Loss (Cross-Entropy): During training, the model's output is compared with the correct transcript using cross-entropy loss:

- This loss function measures how close the predicted transcription is to the real answer.
- It guides the model to improve its weights and decision-making with each training step.

Fine-Tuned PhoWhisper Model: After many rounds of learning, the process produces a fine-tuned PhoWhisper model:

- This model is specifically adapted for Vietnamese (and bilingual) speech, recognizing accents, tones, and domain-specific terminology.
- The result is a high-accuracy speech-to-text engine ready for deployment in banking, compliance, or other enterprise environments.

Architecture:

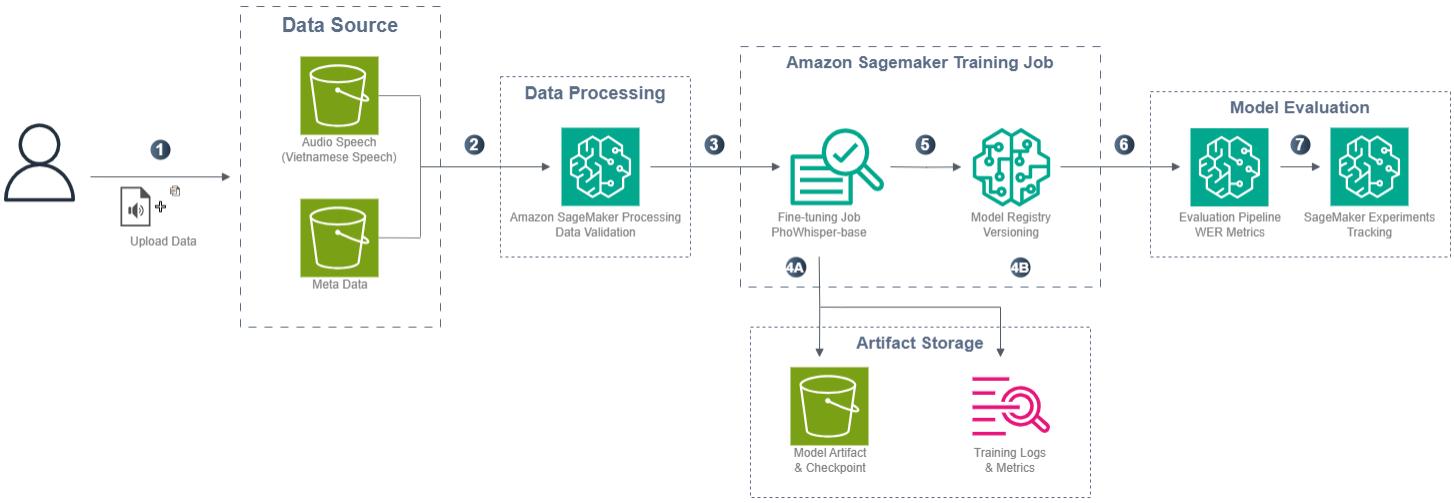

Figure 14: Fine-tuning PhoWhisper model architecture

Description:

Step 1: Data Ingestion

- **User uploads raw data** to **Amazon S3**.
- The data is divided into two parts:
    - **Audio Speech (S3):** Contains all Vietnamese audio files (e.g., .wav, .mp3).
    - **Metadata (S3):** A file (e.g., .csv, .json) describing each audio file, including **transcripts**, **dialect**, and **gender**.

Steps 2 & 3: Data Processing

- **Service:** Amazon SageMaker Processing Job.
- **Purpose:** Automatically clean and prepare data before training.
- **Key tasks:**
    - **Data Merging:** Link each audio file with its corresponding metadata.

- **Text Processing:** Add special conditioning tags (e.g., [North] [Male] chào bạn).
- **Audio Processing:** Resample all audio to a standard format (e.g., 16kHz mono).
- **Data Validation:** Check data integrity (e.g., missing or corrupted files).
- **Data Splitting:** Divide data into **train/validation/test** sets.

- **Output:** Clean and validated data ready for training.

Steps 4 & 5: Model Training & Version Management

- **Service:** Amazon SageMaker Training Job.
- **Purpose:** The core step - fine-tuning the model.
- **Process flow:**
  - **(4A) Fine-Tuning Job:**
    SageMaker automatically spins up GPU instances, loads the **PhoWhisper-base** model, and trains it on the processed data.
  - **(4B) Artifact Storage:**
    - **Model Artifacts & Checkpoints** are saved to S3 (ensuring progress recovery).
    - **Training Logs & Metrics** (loss, accuracy, etc.) are streamed to **Amazon CloudWatch** for real-time monitoring.
  - **(5) Model Registry Versioning:**
    Once training completes, the model is automatically **registered in the SageMaker Model Registry**.
    Each training run produces a **new version** (v1, v2, v3...) for organized lifecycle management.

Steps 6 & 7: Model Evaluation & Experiment Tracking

- **Services:** SageMaker Processing Job (for evaluation) + SageMaker Experiments (for tracking).

- **Purpose:** Automatically assess the trained model's performance.
- **Process flow:**
    - The **evaluation pipeline** fetches the latest model (from Step 5) and the **test dataset** (from Step 2).
    - **(6) Evaluation Pipeline:** Runs inference on test data and computes key metrics - especially **WER (Word Error Rate)**, both overall and by subgroup (dialect, gender).
    - **(7) Experiments Tracking:**
      All evaluation results (WER scores) and training parameters (e.g., learning rate, epochs) are logged in **SageMaker Experiments**, functioning like an **experiment notebook** to compare runs and identify the best-performing model.

1.6. Component Details – Security & Monitoring

The **Security & Monitoring** layer serves as the foundation for ensuring data protection, system reliability, and compliance across the entire VPBank AI Workflow Automation ecosystem. This layer operates transparently across all components – from the **client–facing voice interface (Pipecat & WebRTC)** to backend orchestration services – providing continuous enforcement of security controls and visibility into operational health.

Security Stack

1. Cognito for User Authentication

Amazon Cognito is used as the centralized identity management service, integrated seamlessly with **VPBank's Single Sign–On (SSO)** platform. This integration ensures that both internal users (relationship managers, customer service representatives, HR staff) and authorized external users can securely access AI–driven workflows using existing corporate credentials. Cognito enables:

- **Multi–Factor Authentication (MFA)** for sensitive transactions.
- **Role–Based Access Control (RBAC)** to enforce least–privilege access at the service and API level.
- **Session token expiration policies**, mitigating the risk of unauthorized reuse.

This ensures every interaction with the AI workflow – whether via web dashboard, voice client, or automation API – is authenticated, authorized, and logged under a verifiable identity.

2. KMS for Data Encryption

All data within the AI platform, including voice recordings, transcribed text, and workflow metadata, is encrypted using **AWS Key Management Service (KMS)** with **AES–256** encryption both **at rest** and **in transit**.

- **At Rest:** Data stored in S3, DynamoDB, and EBS volumes is encrypted using customer–managed KMS keys.
- **In Transit:** All network communication (WebRTC, API Gateway, internal Lambda calls) is encrypted via **TLS 1.2+**.

KMS provides centralized lifecycle management of encryption keys – including key generation, automatic rotation, and fine–grained access logging – ensuring that only authorized services and personnel can access sensitive data.

This design is particularly critical for compliance with **Vietnamese banking data regulations** and internal **VPBank Information Security (InfoSec) policiesAudit Trail and Data Retention.**

VPBank maintains a **comprehensive, immutable audit trail** for all security– and system–related events. All logs are stored in **Amazon S3** with **seven years of retention**, ensuring durability (99.999999999%), version control, and accessibility for regulatory or investigative purposes.

The audit trail includes:

- **Authentication logs** (from Cognito and SSO).
- **API call history** (via AWS CloudTrail).
- **Voice transaction metadata** (timestamp, latency, user ID, session outcome).
- **System monitoring events** (CloudWatch and Pipecat telemetry).

This long–term log retention is critical for:

1. **Regulatory Compliance:** Satisfying national and internal audit requirements for data traceability.
2. **Forensic Analysis:** Enabling retrospective analysis in case of anomalies or breaches.
3. **Security Monitoring:** Allowing continuous pattern analysis to identify early warning indicators of potential threats.

All audit logs are periodically verified for integrity and backed by **cross–region replication**, ensuring business continuity even under disaster recovery scenarios.

Security Monitoring and Governance Integration

To maintain continuous observability and proactive defense, the platform integrates multiple monitoring layers:

- **Amazon CloudWatch** for real–time metric collection (latency, error rate, model drift).

2. Detailed Flow Diagrams

2.1. Delivery Demo

Basic Usage Guide

Activate the Assistant: Open the demo interface and press the microphone button to begin. Give a Command: Clearly and naturally state your request. You can issue a command for the entire process or break it down into smaller steps. Observe: Watch as the AI assistant automatically navigates the website, locates the correct input fields, and enters the data exactly as you commanded.

Detailed Demo Scenarios

Scenario: Loan Submission

This scenario simulates the multi-step process of applying for a loan submission online. You can begin by saying, "Open the VPBank loan submission page." Follow up with commands to fill in the required information, such as, "Start filling in the personal details. Full name: Nguyễn Thị B. Date of birth: May 15, 1990. Phone number: 0987654321." You can continue with, "Next, fill in the address information. Permanent address: 123 Lang Street, Dong Da, Hanoi." Then, you may say “Pass the Captcha”. To complete the process, you might say, "Review all the information and press the confirm button."

Expected Output: The browser will automatically open the VPBank loan submission page. The "Full name," "Date of birth," "Phone number," and "Permanent address" fields on the webpage will be accurately populated with the data you provided. After all information is entered, the assistant will automatically click the the “Captcha” field and finally click "Confirm" button (or equivalent), and the screen will navigate to a success page or display a corresponding confirmation message.

Scenario: Submitting a Detailed Customer Support Ticket

This scenario demonstrates the assistant's ability to handle long text fields and select options from dropdown lists. You could start with, "Go to the customer support page and create a new ticket." Then, specify the details: "In the 'Request Type' section, select 'Online Transaction Issue'." Continue by providing specific data: "Enter the transaction ID as VPB-98765. In the issue description box, enter: 'I made a transfer at 10 AM this morning, but the recipient has not yet received the funds. Please investigate this for me'." Finally, command the assistant to "Submit the ticket."

Expected Output: The browser will navigate to the customer support page. The assistant will select "Online Transaction Issue" from the list of request types. The "Transaction ID" field will be filled with "VPB-98765," and the entire descriptive sentence will be accurately entered into the corresponding text box. Finally, the assistant will click the "Submit" button, and the screen will display a "Ticket submitted successfully" message along with a tracking ID.

2.2. Sequence Diagram

This sequence diagram illustrates a VPBank customer or employee’s use of voice commands to complete a loan application:

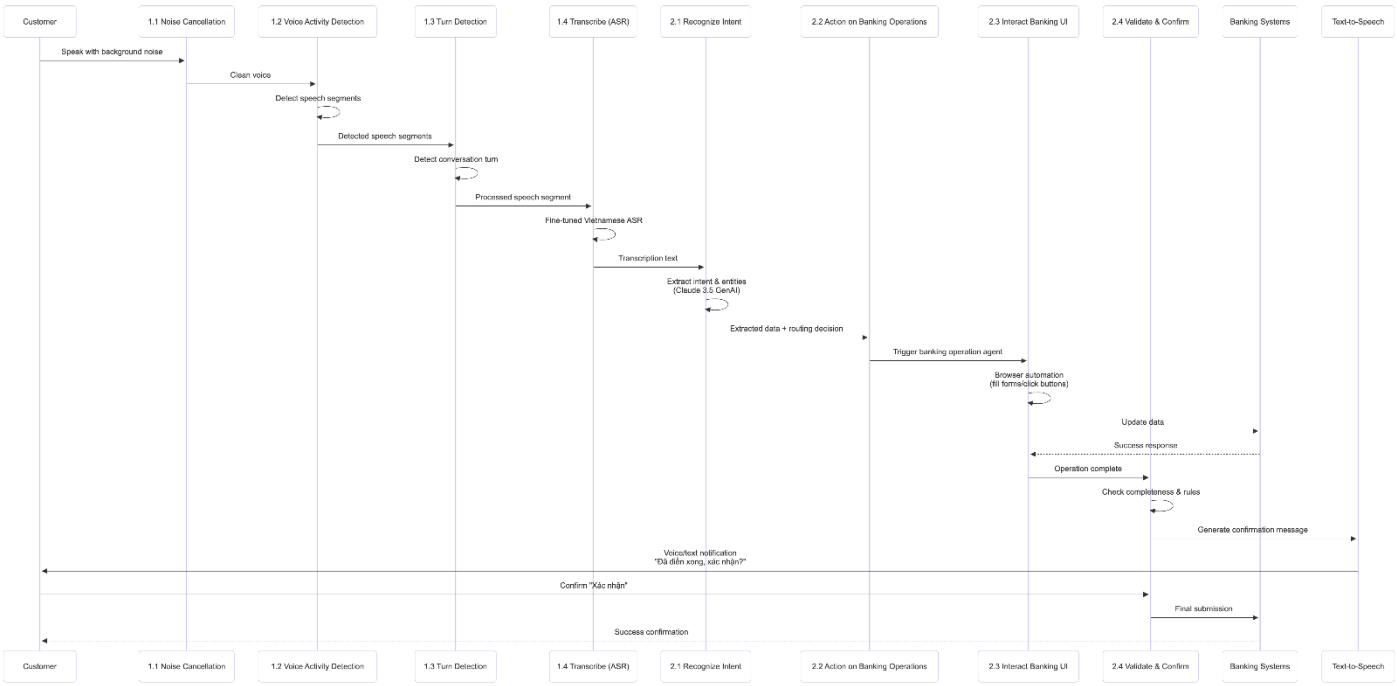

Figure 15 : Sequence Diagram of Voice–Based Banking Flow

2.1.1. Customer Speaks

- The customer interacts naturally, even in a **noisy environment**.
- The system begins capturing the **audio input**.

2.1.2. Preprocessing & Speech Segmentation

(1.1) Noise Cancellation

- Filters out ambient noise (background chatter, traffic, etc.)
- Produces a **clean, clear voice signal** for accurate processing.

(1.2) Voice Activity Detection (VAD)

- Detects **when speech starts and stops**.
- Separates voice segments from silence, ensuring efficient processing.

(1.3) Turn Detection

- Identifies **when the customer finishes speaking**.
- Signals the system to move forward to speech recognition.

2.1.3. Speech–to–Text Conversion

(1.4) Automatic Speech Recognition (ASR)

- Converts spoken Vietnamese into text format.
- Uses a fine–tuned Vietnamese ASR model for local language accuracy.
- Output: Transcribed text of the customer's request.

2.1.4. Understanding & Decision–Making

(2.1) Intent Recognition

- A Generative AI model (e.g., Claude 3.5) analyzes the transcribed text.
- Extracts:
    - Intent → What the user wants (e.g., apply for a loan, transfer money).
    - Entities → Key data (e.g., amount, recipient, account number).
- Produces structured data and a **routing decision**.

(2.2) Action Handling

- Determines which banking operation to perform.
- Prepares an action plan for the automation layer (e.g., open loan form, fill data).

2.1.5. Automation & System Interaction

(2.3) Browser Automation

- Automates interaction with the banking user interface (UI).
- Executes actions such as filling forms, selecting options, and clicking buttons.
- Sends data to the bank's backend system.

Bank System Interaction

- The core banking system validates and processes the submitted data.
- Returns a success or error response to the automation layer.

2.1.6. Validation & Confirmation

(2.4) Validate & Confirm

- Reviews results to ensure data completeness and compliance.
- Generates a summary or confirmation prompt.

Example Voice Output:

"Đã điền xong, xác nhận?" ("All filled in, confirm?")

2.1.7. Customer Confirmation & Finalization

Customer Confirms

- The user responds with a simple "Xác nhận" (Confirm).

Final Submission

- The system executes the final submission to the bank.

Success Notification

- The banking system confirms success, and the customer receives a spoken or text acknowledgment.

Summary

This end–to–end voice flow combines speech recognition, GenAI intent understanding, and browser automation to deliver a hands-free, secure banking experience for VPBank users.

2.3. Voice Pipeline Detail (Pipecat)

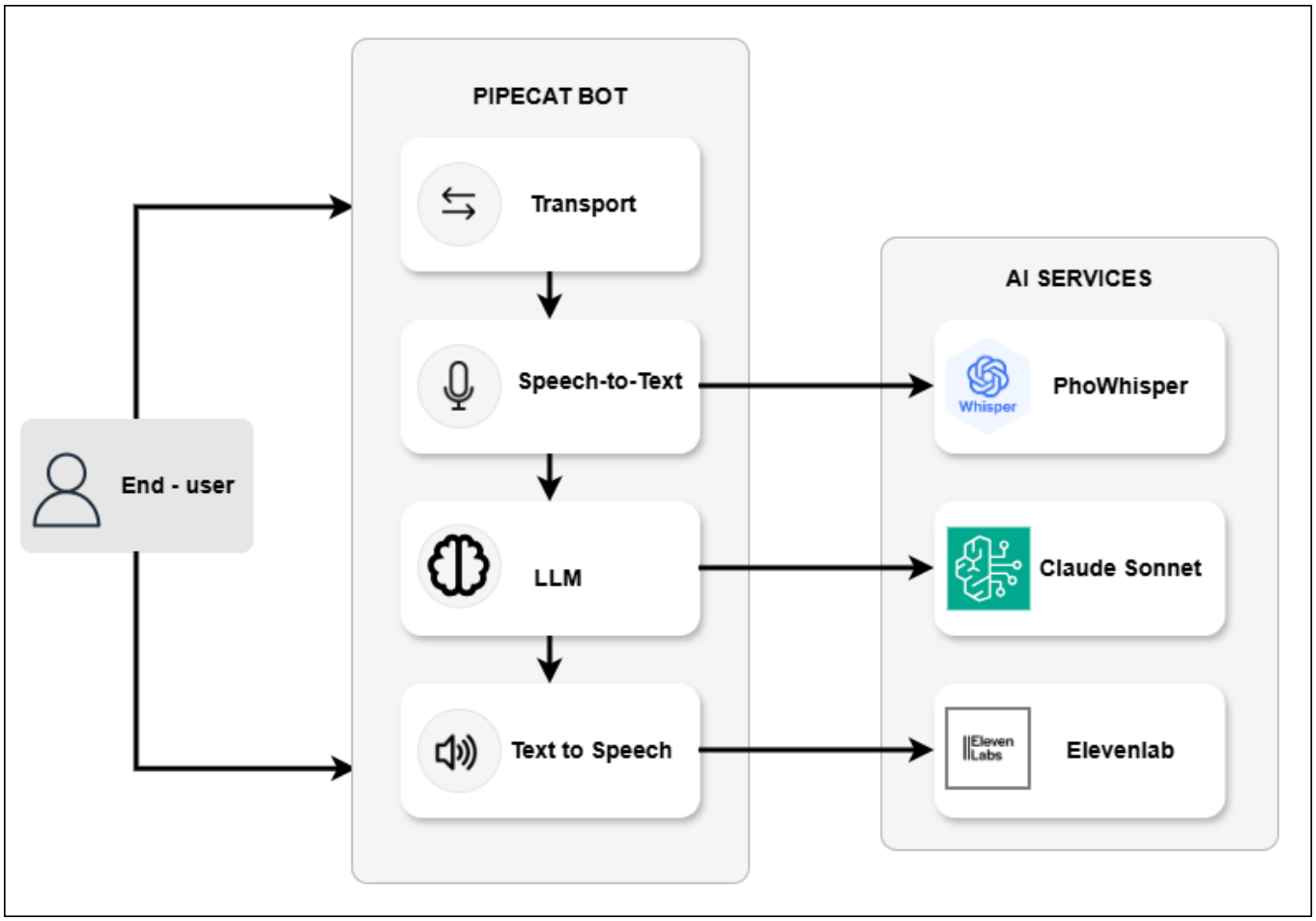

Figure 16: Detailed Diagram about Pipecat Component

The **Voice Pipeline** is the core mechanism enabling real–time, low–latency conversational interaction between the user and the AI workflow system. It is built upon Pipecat, an open–source framework designed for managing audio streaming, buffering, and WebRTC–based communication.

When a user speaks a command (e.g., “Điền khách hàng Nguyễn Văn An vay 500 triệu, kỳ hạn 24 tháng”), Pipecat captures and streams the audio input through the following sequential process:

1. **Audio Capture and Buffering:** The user’s microphone input is captured in real time by Pipecat’s WebRTC connection, which provides a secure, low–latency audio channel. The audio is temporarily buffered to maintain a continuous, stable stream for downstream processing.

2. **Voice Activity Detection (VAD):** Pipecat's integrated VAD module analyzes the incoming stream to detect when the user starts and finishes speaking. This prevents unnecessary processing of silence or background noise, optimizing both performance and resource utilization.
3. **Speech–to–Text Conversion (PhoWhisper):** Once speech is detected, the audio is sent PhoWhisper for Vietnamese language optimization). The transcription service converts the speech into text and provides a confidence score for each segment.
4. **Confidence Validation and Routing:** If the transcription confidence score is ≥ 80%, the resulting text is forwarded to the Agent Orchestrator, which interprets the intent and routes it to the appropriate domain–specific AI agent (e.g., Loan, CRM, or HR). If confidence is < 80%, the system automatically requests the user to repeat the command for clarification, ensuring that only high–quality inputs are processed.

This design guarantees accuracy, privacy, and responsiveness while maintaining conversational fluidity. Pipecat's event–driven pipeline enables flexible integration with multiple speech models and voice analysis tools such as Krisp for real–time noise suppression, further enhancing user experience and STT accuracy.

2.4. Text–to–Speech Response

Once the Agent Orchestrator done and produces a text–based response the Text–to–Speech (TTS) pipeline transforms this response into natural–sounding Vietnamese speech.

Text–to–Speech Response Flow:

1. The AI Agent generates a Vietnamese text response.
2. The synthesized audio stream is sent through Pipecat Sender, which handles buffering and transport management.
3. Audio is delivered to the user's browser over a WebRTC connection, maintaining sub–second latency for real–time conversational flow.
4. The user hears the AI's spoken reply as if communicating with a live assistant.

This TTS flow enables a fully interactive voice experience, bridging natural human speech with AI reasoning results. By leveraging Elevenlabs and Pipecat's streaming architecture, the system achieves both low latency and high audio quality, essential for enterprise–grade customer interactions such as loan application assistance, CRM updates, or internal HR workflows.

3. Deployment Architecture

AWS Architecture Implementation for the Voice–Driven Automation Platform implement the AWS architecture that powers our voice–driven automation platform. The design is engineered for scalability, security, and high availability, providing a robust foundation for the core features detailed in the project's implementation strategy documents. It translates the conceptual framework of adaptive conversations and structured data extraction into a tangible, high–performance cloud infrastructure.

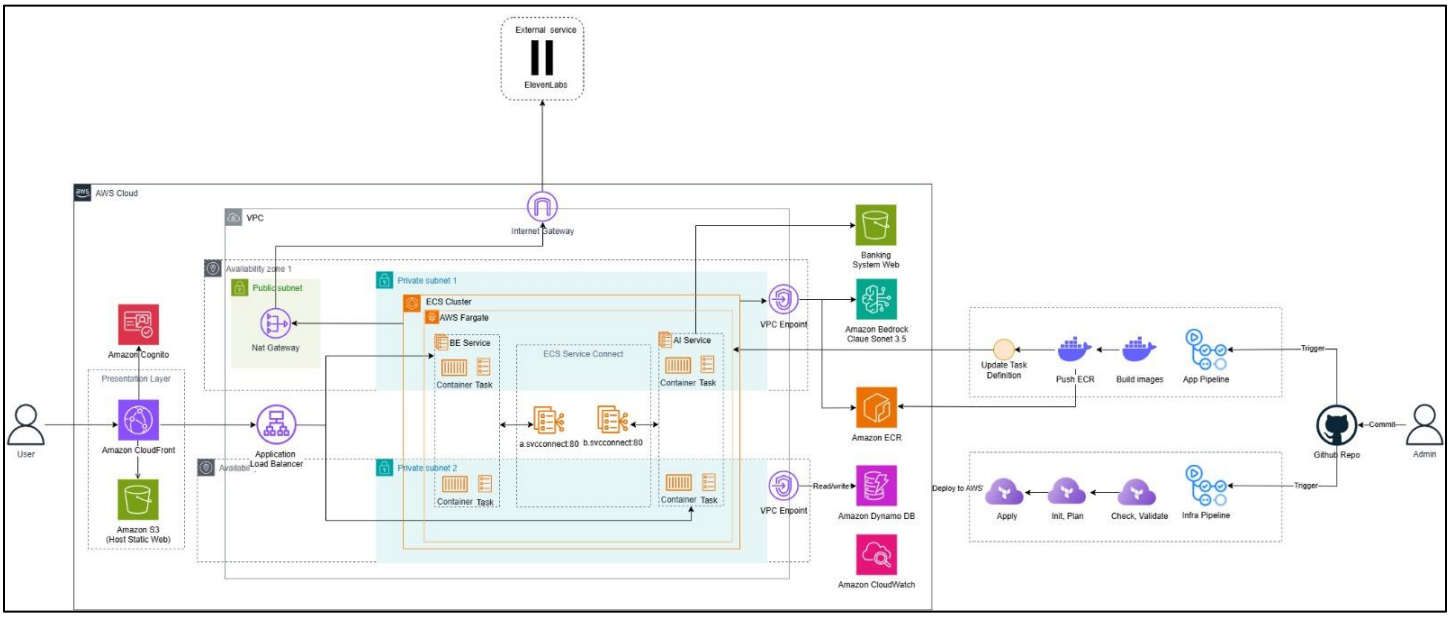

Figure 17 : AWS Deployment Diagram

Infrastructure as Code Deployment: The cloud infrastructure is managed with Terraform using the Infrastructure as Code (IaC) approach for consistent, automated deployment. Terraform Cloud handles execution, state storage, locking, and centralized policy and access control across all environments.

Interface Delivery and Session Initiation: The user experience starts with a responsive web interface delivered via Amazon CloudFront and Amazon S3 for low-latency access. The interface captures voice input, manages the local session, and establishes a secure backend connection for real-time interaction.

User Authentication: User identity and access management are handled through Amazon Cognito, which provides a fully managed authentication and authorization solution for the application.

The Conversational Engine on AWS Fargate: The heart of the platform is the PipeCat conversational engine, which runs as a containerized application on Amazon ECS with the AWS Fargate launch type. This serverless compute model eliminates the need to manage underlying EC2 instances, enabling fully automated scaling, improved resource isolation, and consistent performance under varying workloads.

High Availability: The services are deployed across private subnets in two separate Availability Zones, ensuring high availability and resilience against single-zone failures.

Container Management: The Docker images for the engine are built and stored in Amazon ECR (Elastic Container Registry) and are securely pulled into the Fargate environment through dedicated VPC Endpoints, ensuring private network connectivity and enhanced security.

Service Communication: Internal communication between microservices (e.g., AI Service and Backend Service) is managed by ECS Service Connect, providing a resilient service discovery mechanism within the VPC.

AI-Powered Cognition, Data Persistence, and Voice Synthesis The conversational engine's ability to understand, remember, and respond is powered by managed AWS and external services.

Cognitive Processing: Within a given conversational flow, the engine constructs specific prompts which are sent to Amazon Bedrock through a secure VPC Endpoint. We leverage Bedrock's powerful Claude models to perform the cognitive work: interpreting user intent, extracting structured data, and deciding the next best action.

State Management: The agent uses Amazon DynamoDB for persistent and low-latency data storage, accessed securely through a VPC Endpoint. This component is essential for maintaining conversation state, storing user session data, and preserving audit logs.

Voice Synthesis: To complete the interaction loop, the engine converts its text-based responses into natural speech using the ElevenLabs API. Outbound requests to this external service are routed securely from the private subnets through a NAT Gateway.

System Observability and Audit Trails: The entire system is monitored using Amazon CloudWatch, which collects logs and metrics from all running Fargate tasks. Each instance of the PipeCat engine streams its logs directly to CloudWatch, providing real-time visibility into application performance and behavior. Combined with the state data stored in Amazon DynamoDB, this setup creates a detailed and immutable audit trail of every action taken by the agent. CloudWatch metrics also enable automatic scaling of Fargate services based on system demand.

Administration and CI/CD Pipeline: The system uses an automated CI/CD pipeline built with GitHub Actions and Terraform Cloud. All source code is maintained in a GitHub repository, which is organized into two separate sections - application and infrastructure. Each section is managed through its own pipeline: the *application pipeline builds and pushes Docker images to* Amazon ECR and updates ECS Fargate tasks, while the *infrastructure pipeline* triggers Terraform Cloud to automatically validate, plan, and apply infrastructure changes. This structure ensures consistent, secure, and fully automated deployments across all environments.

4. Cost Estimation

Service Name	Description	Region	Monthly Cost	Properties
Amazon CloudFront	Delivery static web content	US East (N. Virginia)	$11.00	Data transfer out to internet: 100 GB per month Data transfer out to origin: 100 GB per month Number of requests (HTTPS): 500,000 per month
AWS Fargate	Self–host PhoWhisper	US East (N. Virginia)	$27.26	Operating system: Linux CPU Architecture: x86 Average duration: 45 minutes Number of tasks or pods: 3 per day Amount of ephemeral storage: 50 GB
Workload 1	Core LLM	US East (N. Virginia)	$86.40	Average requests per minute: 2 Hours per day at this rate: 1 Average input tokens per request: 500 Average output tokens per request: 1,500
Application Load Balancer		US East (N. Virginia)	$22.27	Number of Application Load Balancers: 1
Amazon API Gateway		US East (N. Virginia)	$0.07	HTTP API requests units: millions Average size of each request: 34 KB Requests: 20,000 per month

S3 Standard		US East (N. Virginia)	$0.23	S3 Standard storage: 10 GB per month
VPN Connection		US East (N. Virginia)	$0.00	Working days per month: 22
NAT Gateway		US East (N. Virginia)	$32.85	Number of NAT Gateways: 1
Amazon CloudWatch		US East (N. Virginia)	$25.23	Standard Logs: Data Ingested: 50 GB
Amazon Elastic Container Registry		US East (N. Virginia)	$20.00	Amount of data stored: 200 GB per month
TOTAL			**$225.31**	

5. Roadmap

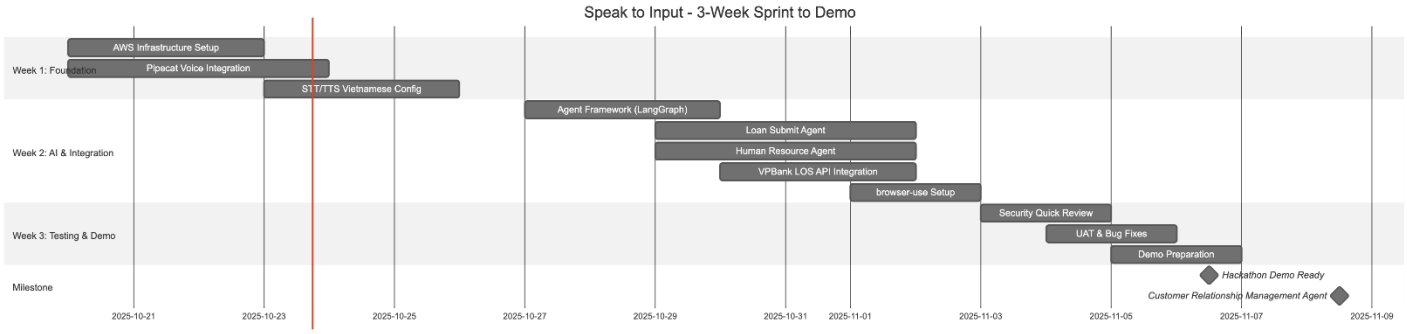

Figure 18 : Timeline Roadmap Project

Total Duration: 18 days (October 20 – November 6, 2025)

Week 1: Foundation & Core Setup (Oct 20–26)

- AWS infrastructure + Pipecat integration (parallel)
- Basic STT/TTS with Vietnamese support

Week 2: AI Agents & Integration (Oct 27 – Nov 2)

- Loan Submit Agent
- Human Resource (HR) Agent
- VPBank LOS API integration + browser–use setup

Week 3: Testing & Demo (Nov 3–6)

- Security review + UAT testing
- **Demo–ready prototype by Nov 6**

Key Overlaps: Testing begins while integration completes; Security review runs parallel to final integration work to save time.

6. Team Structure (5 Members)

Role/Function	Description	Team Size
Solution Architect	Technical Leadership	Hải Anh
AI/ML Engineer	LLM Integration, Agent Development, Pipecat Voice Agent	Hiếu Nghị
Fullstack Engineer	React Voice Widget, Backend API, System Integration	Minh Nghĩa
DevOps Engineer	AWS Infrastructure Implement	Đức Toàn

Project Manager	Tracking Progress, Quality Control	Lodi Bùi

7. Daily Task Distribution

Day	**Milestone**	**Owner**	**Deliverable**
1–2	AWS Infrastructure	DevOps	VPC, ECS cluster setup
3–4	Pipecat Integration	AI/ML	Voice gateway running
5–6	STT/TTS Working	AI/ML	Vietnamese transcription demo
7–8	Agent Framework	AI/ML	LangGraph orchestrator
9–10	Multi–Agent System	AI/ML + Backend	6 agents working
11–12	browser–use	Backend	Browser automation live
13–14	VPBank Integration	Backend	LOS/CRM connected
15–16	Security & Testing	All Team	Penetration test passed
17–18	UAT	QA + BA	User acceptance
19–20	Pilot Launch	All Team	50 users onboarded

Reference

[1] VPBank. (2025). Vay tín chấp theo lương – Điều kiện, quy trình, thủ tục vay. https://www.vpbank.com.vn/bi-kip-va-chia-se/retail-story-and-tips/loans-category/vay-tin-chap-theo-luong

[2] VPBank. (2024). Hướng dẫn mở tài khoản thanh toán VPBank và kích hoạt VPBank Neo. https://www.vpbank.com.vn/-/media/vpbank-latest/1retail/promotion/other/2024/appota/huong-dan-mo-tkvpbank-va-kich-hoat-vpbank-neo.pdf

[3] VPBank. (2024). Chính sách phòng chống rửa tiền, chống tài trợ khủng bố, chống tài trợ phổ biến vũ khí hủy diệt hàng loạt và tuân thủ cấm vận, trừng phạt tại Ngân hàng TMCP Việt Nam Thịnh Vượng. https://www.vpbank.com.vn/-/media/vpbank-latest/8aboutvpbank/ve-vpb_he-thong-kiem-soat-tuan-thu/2024/chinh-sach-phong-chong-rua-tien-vpbank-2024.pdf

[4] VIB. (2025). Thời gian xử lý hồ sơ vay ngân hàng mất bao lâu? https://www.vib.com.vn/vn/cam-nang/vay/vay-the-chap/thoi-gian-xu-ly-ho-so-vay-ngan-hang

[5] General Statistics Office of Viet Nam, & UNFPA. (2021). Population ageing and older persons in Viet Nam: Key findings from the 2019 Census. Hanoi, Vietnam: UNFPA. https://vietnam.unfpa.org/sites/default/files/pub-pdf/ageing_report_from_census_2019_eng_final27082021.pdf

[6] UNICEF Viet Nam, & General Statistics Office of Viet Nam. (2016). Results of the survey on people with disabilities in Viet Nam. Hanoi, Vietnam: UNICEF. https://www.unicef.org/vietnam/media/2786/file/Main%20report%20people%20with%20disabilities%20survey.pdf

[7] Zuko Analytics. (2024). Form benchmark report 2024. Zuko Analytics Ltd. https://www.zuko.io/benchmarking/industry-benchmarking

[8] Baymard Institute. (2023). E-commerce checkout usability: Quantitative study. Copenhagen, Denmark: Baymard Institute. https://baymard.com/lists/cart-abandonment-rate

[9] Le, T.-T., Nguyen, L. T., & Nguyen, D. Q. (2024). PhoWhisper: Automatic speech recognition for Vietnamese. ICLR 2024 Tiny Papers Track. https://arxiv.org/pdf/2406.02555

[10] Hakob, M. (2024, September 3). Form abandonment: How to recover lost leads and improve conversions. FormStory Blog. https://formstory.io/learn/form-abandonment-tracking/

[11] Gopalakrishnan, K. (2023). Robotic process automation: Streamlining operations and revolutionizing traditional banking processes. Journal of Artificial Intelligence & Cloud Computing, 2(4), 1–4. https://doi.org/10.47363/JAICC/2023(2)332

[12] FPT IS. (2024). RPA là gì? https://fpt-is.com/goc-nhin-so/rpa-la-gi/

[13] TPBank. (2024). Chuyển tiền bằng giọng nói – VoicePay: Đẳng cấp Siri tài chính trên app TPBank. https://tpb.vn/tin-tuc/tin-tpbank/chuyen-tien-bang-giong-noi-voicepay-dang-dap-siri-tai-chinh-tren-app-tpbank

[14] VIB. (2024). Giao dịch bằng giọng nói – Voice banking. https://www.vib.com.vn/vn/cam-nang/ngan-hang-so/tien-ich-va-trai-nghiem/giao-dich-bang-giong-noi-voice-banking

[15] FPT AI. (n.d.). Voice-based transactions: Inevitable trend in digital banking. https://fpt.ai/blogs/voice-based-transactions-inevitable-trend-digital-banking/

[16] VPBank. (2020, November 27). VPBank năm thứ 3 liên tiếp nhận giải thưởng "Ngân hàng chuyển đổi số tiêu biểu". https://www.vpbank.com.vn/tin-tuc/thong-cao-bao-chi/2020/vpbank-earned-outstanding-digital-transformation-bank-award-third-year-in-a-row

[17] Amazon Web Services. (2025). AWS cost calculator estimate. https://calculator.aws/#/estimate?id=24af81e1675e2d145576b008e6f441b3e71103c3

[18] Cong Thuong. (2022). Đã đến thời của giao dịch ngân hàng bằng giọng nói?

https://cand.com.vn/doanh-nghiep/da-den-thoi-cua-giao-dich-ngan-hang-bang-giong-noi-i658651/

[19] Cong Thuong. (2022). Ngân hàng số phiên bản MyVIB 2.0 với công nghệ AR và cloud native. https://congthuong.vn/ngan-hang-so-phien-ban-myvib-20-voi-cong-nghe-ar-va-cloud-native-179068.html

[20] TPBank. (2024). Chuyển tiền bằng giọng nói – VoicePay: Đẳng cấp Siri tài chính trên app TPBank. https://tpb.vn/tin-tuc/tin-tpbank/chuyen-tien-bang-giong-noi-voicepay-dang-dap-siri-tai-chinh-tren-app-tpbank

[21] Forrester Research. (2023). The state of customer experience 2023: Personalization and voice interface trends. Forrester.

[22] Bain & Company. (2024). How the Net Promoter Score® relates to growth. https://www.netpromotersystem.com/about/how-net-promoter-score-relates-to-growth/

[23] Smith, P. (2020). Transforming contact center WFM through artificial intelligence: Implementation and impact. International Journal of Research in Computer Applications and Information Technology, 7(2), 161–168. https://iaeme.com/MasterAdmin/Journal_Uploads/IJRCAIT/VOLUME_7_ISSUE_2/IJRCAIT_07_02_161.pdf