CHALLENGE 2

Milestone 1: Requirements Analysis, Data Preparation, and Model Selection

CS590 Summer 2025

Jun 8, 2025

Full Names	Student ID	email
Alpha Oumar Diallo	002352906	adiallo234@ubishops.ca
Ahmed Issa	002230777	aissa23@UBishops.ca
Amna Khalid	002351415	akhalid@UBishops.ca

1 Project Summary

The project aims to develop an AI-powered chatbot for **Level-0 IT support** for The University. Level-0 (Tier 0) support provides *self-serve* IT solutions such as FAQs, knowledge bases, and automated chatbots. The chatbot will answer routine technical questions (e.g., password resets, network access, software installation) using natural language understanding and a connected knowledge base. Automating these tasks is expected to reduce helpdesk load and improve user satisfaction by providing instant, 24/7 assistance.

1.1 Objectives

- **Automate Routine Queries:** Enable students and staff to get immediate answers to common IT issues (password resets, email setup, Wi-Fi connection, etc.) without human intervention.
- **Reduce Helpdesk Burden:** Decrease the volume of tickets reaching Level-1 support by resolving simple problems at Level-0, allowing IT staff to focus on complex cases.
- **Improve User Experience:** Provide a friendly, conversational interface accessible 24/7, reducing wait times and improving the campus IT support experience.
- **Maintain Data Privacy:** Ensure that all user data remains secure, by choosing appropriate LLM models and deployment strategies (on-premise vs cloud) that comply with privacy requirements.
- **Scalable Solution:** Design the system so it can be extended to cover additional universities or new use cases in the future.

1.2 Scope and Use Cases

Scope: The initial scope is limited to **Level-0 (Tier 0)** IT support tasks at The University. This includes frequently asked questions and straightforward tasks such as:

- Password resets and account unlocks (AD/portal accounts).
- Email configuration and access (campus email, mobile sync).

- Network access troubleshooting (Wi-Fi setup, VPN access).
- Software usage queries (available licensed software, installation steps).
- Campus service guides (printing services, library login).

Non-critical administrative tasks like course registration or financial matters are *out of scope*. Complex issues beyond basic guidance will be escalated to human support.

Use Cases: The project will document specific Level-0 use cases (50 in total; see Section 3) covering categories like *Account Management, Networking, Email, Printing, Software, Hardware* and *Security*. Each use case includes sample user utterances, preconditions, success flow, alternate flow, postconditions, etc., ensuring comprehensive coverage of anticipated queries.

2 Data Collection and Preparation

To ensure the chatbot has accurate and comprehensive knowledge, data was collected from a variety of publicly available IT support resources. These include official FAQs and help articles from major technology providers such as Microsoft and Google, as well as IT services documentation and support pages from universities across Canada, the United States, the United Kingdom, and Australia. These sources provided a broad and diverse set of issues and resolutions relevant to university-level IT support scenarios.

The data collection and preparation process involved several key steps:

- **Knowledge Base Extraction:** Frequently asked questions (FAQs), how-to articles, troubleshooting guides, and help documentation were gathered from the IT websites of various universities and tech companies. This content was curated and reformatted into question-and-answer pairs suitable for ingestion by a large language model (LLM).
- **Support Logs and Transcripts:** Where available, anonymized IT support ticket transcripts and chat logs were considered to better capture the language, phrasing, and flow of real user interactions. These transcripts help in fine-tuning the model to understand natural and sometimes ambiguous user input.
- **General IT Support Corpus:** To supplement the university-specific content, general-purpose IT support datasets were integrated. These included publicly available documentation from sources like Microsoft support forums, Google Workspace help centers, and major antivirus providers. This added depth and broader context to the dataset.
- **Preprocessing:** Collected data underwent thorough cleaning and formatting. Irrelevant or outdated information was removed, formatting was standardized, and conversational dialogues were segmented and tagged where applicable.

The data will be split into training and validation sets if fine-tuning an open model; otherwise, used as context sources for retrieval-augmented generation.

2.1 Functional Requirements

- **Natural Language Interface:** Understand user queries in text (English). Support clarification questions for ambiguous requests.
- **Accurate Responses:** Provide correct, concise answers. For list-style answers (e.g., steps to connect Wi-Fi), respond with clear bullet points or numbered lists.
- **Knowledge Integration:** Access the IT knowledge base (via embeddings or retrieval) to ground responses and avoid hallucinations. Cite relevant policy or help page when possible.
- **Authentication:** Optionally verify user identity for personalized data (e.g., check login status before revealing account info).

- **Escalation Mechanism:** Detect when a query cannot be answered (out of scope or uncertain). In such cases, politely suggest submitting a support ticket to Tier-1 staff.
- **Multi-platform Access:** Deploy on university channels (website chatbot, Microsoft Teams, or Slack).

2.2 Non-Functional Requirements

- **Performance:** Responses should be delivered within 1–2 seconds of user input. The backend should support concurrent conversations.
- **Reliability:** Uptime should be high (target 99%). Gracefully handle downtime or model errors by displaying an apology and forwarding to human support.
- **Security & Privacy:** we will use secure protocols (HTTPS) for data in transit. as we are using cloud APIs, we will definitely ensure compliance with privacy policies; will considering an open-source model on local servers for sensitive data.
- **Maintainability:** Code will be modular (separate NLP pipeline, retrieval, and UI layers). Configuration and thresholds should be easy to update (e.g., adjust confidence cutoffs).
- **Scalability:** The architecture will allow scaling out (e.g., adding more GPU workers or fallback to larger LLM) as usage grows.

2.3 Resources and Tools

- **Models/API:** The project will use Meta's LLaMA 3.1 model, hosted on AWS. Fine-tuning will be conducted using Hugging Face's AutoTrain platform, leveraging the collected dataset to adapt the model to the university IT support domain.
- **Infrastructure:** The backend will be developed using Express.js and Next.js for API routing and frontend integration. The system will be deployed on AWS using Docker containers orchestrated via Kubernetes to ensure high availability, scalability, and modular deployments.
- **Libraries:** Retrieval-augmented generation (RAG) will be implemented using LangChain or Haystack, with model access through the Hugging Face Transformers library. Llama and Anthropic SDKs may be integrated optionally. Rasa will be considered if advanced dialogue state management becomes necessary.
- **Knowledge Base:** The core knowledge base will include university IT documentation (when available), Confluence or SharePoint internal support pages, and curated public resources such as Zendesk and other IT knowledge articles.
- **Collaboration Tools:** Development and project coordination will be managed using GitHub for version control and Microsoft Teams for communication.

2.4 Backend Implementation Plan

The chatbot backend will consist of:

- **API Layer:** Expose endpoints for sending user queries and retrieving answers. Implement chat session management.
- **NLP Engine:** Interface with the chosen LLM (e.g., call llama API or a local Hugging Face model). Process and format the response.
- **Knowledge Retrieval:** If not relying solely on LLM, include an embedding-based search over the IT KB to find relevant articles. Use RAG (Retrieve and Generate) to ground answers.
- **Database:** (Optional) Store conversation logs, user preferences, or any escalations.
- **Integration Layer:** Connect to campus authentication (if checking user identity) and to existing helpdesk systems (for ticket creation if needed).

2.5 Risk Assessment

- **Incorrect Answers (Hallucinations):** LLMs can generate plausible but wrong info. Mitigation: tightly constrain the domain (Level-0 tasks), use knowledge base for factual backup, and mark uncertain cases for human review.
- **Privacy Concerns:** Using cloud-based LLMs risks exposing user queries. Mitigation: use an open-source model (like LLaMA 2 or Mistral) on secure servers, or ensure API contracts prevent data misuse.
- **Data Limitations:** If historical query logs are scarce, the model may lack training data. Mitigation: rely on general IT support knowledge and iteratively refine the bot with real usage feedback.
- **User Adoption:** Users might mistrust or underuse a new bot. Mitigation: emphasize convenience and accuracy, provide easy fallback to human support, and gather user feedback early.
- **Technical Integration:** Chatbot integration (with existing systems or communication tools) may be complex. Mitigation: allocate time for prototyping connectors (e.g., testing Microsoft Teams bot framework).

3 Use Case Documentation

Each use case below follows a standard template with the fields: **ID**, **Category**, **Title**, **Sample Utterances**, **Main Success Flow**, **Alternative Flows**, **Postconditions**, **Priority**, and **Estimated Frequency**. The use cases focus on common Level-0 IT support scenarios.

ID	Category	Title	Sample Utterance(s)
UC001	Authentication & Access	Password Reset	"I forgot my password."
UC002	Authentication & Access	Account Lockout	"My account is locked after multiple failed login attempts."
UC003	Authentication & Access	MFA Setup	"How do I enable two-factor authentication?"
UC004	Authentication & Access	VPN Access Issues	"I canconnect to the campus VPN."
UC005	Connectivity & Network	Wi-Fi Connectivity	"My laptop wonconnect to campus Wi-Fi."
UC006	Connectivity & Network	Network Drive Access	"I canaccess the shared network drive."
UC007	Connectivity & Network	Email Configuration	"How do I set up Outlook with my university email?"
UC008	Connectivity & Network	Remote Desktop Issues	"Remote Desktop Connection is not working."
UC009	Software & Applications	Software Installation	"How do I install approved software like MATLAB?"
UC010	Software & Applications	Application Crashes	"Moodles keeps freezing."
UC011	Software & Applications	License Activation	"My software license expired."
UC012	Software & Applications	Browser Issues	"Chrome is slow or not loading pages."

UC013	Software & Applications	Updates & Patches	"How do I check for OS/software updates?"
UC014	Hardware & Peripherals	Printer Troubleshooting	"The library printer isnresponding."
UC015	Hardware & Peripherals	Peripheral Device Issues	"My external monitor isnworking."
UC016	Security & Compliance	Phishing/Suspicious Emails	"I received a suspicious email—what should I do?"
UC017	Security & Compliance	Antivirus Alerts	"My antivirus flagged a file—is it safe?"
UC018	Collaboration & Tools	Meeting Setup	"How do I schedule a Zoom meeting?"
UC019	Collaboration & Tools	File Recovery	"I accidentally deleted a file—how do I restore it?"
UC020	Collaboration & Tools	Shared Resource Access	"I need access to the SharePoint folder."
UC021	Connectivity & Network	Mobile Wi-Fi Setup	"How do I connect my phone to campus Wi-Fi?"
UC022	Connectivity & Network	Mobile Printing	"How do I print from my phone to the library printer?"
UC023	Software & Applications	Mobile Email Configuration	"How do I set up my university email on my phone?"
UC024	Software & Applications	LMS Mobile App Issues	"The Moodle mobile app won't load my courses."
UC025	Hardware & Peripherals	Laptop Battery Health	"My laptop battery drains too fast—what can I do?"
UC026	Hardware & Peripherals	Audio Device Troubleshooting	"My microphone isn't working in Teams."
UC027	Hardware & Peripherals	Webcam Connectivity	"The webcam isndetected by Zoom."
UC028	Security & Compliance	Password Expiry Notification	"I got a warning my password will expire soon—how to extend?"
UC029	Security & Compliance	Security Patch Information	"How do I install the latest security updates?"
UC030	Authentication & Access	Username Recovery	"I forgot my username—how can I recover it?"
UC031	Collaboration & Tools	Shared Calendar Access	"I can't see the department calendar—can you grant access?"
UC032	Collaboration & Tools	OneDrive Access	"Why can't I access my OneDrive files?"
UC033	Collaboration & Tools	Teams Chat History	"How do I retrieve old Teams messages?"

UC034	Collaboration & Tools	File Sharing Permissions	"I don't have edit rights on a shared document."
UC035	Software & Applications	PDF Reader Problems	"Acrobat Reader keeps crashing on PDFs."
UC036	Software & Applications	Virtual Lab Access	"I can't access the virtual lab environment."
UC037	Hardware & Peripherals	Projector Connectivity	"The classroom projector won't connect to my laptop."
UC038	Hardware & Peripherals	Smart Classroom Controls	"The smart board touchscreen isn't responding."
UC039	Connectivity & Network	IP Address Assignment	"My device isn't getting an IP address on the network."
UC040	Connectivity & Network	DNS Resolution Issues	"I can't reach the university website—DNS error."
UC041	Hardware & Peripherals	Headset Configuration	"How do I set up my headset for Teams calls?"
UC042	Software & Applications	VPN Client Update	"How do I update the VPN client on my laptop?"
UC043	Connectivity & Network	Network Printer Driver	"Why is my computer not finding the printer driver?"
UC044	Collaboration & Tools	Canvas Access	"I can't access my course on Canvas."
UC045	Security & Compliance	Security Policy Clarification	"What are the password complexity requirements?"
UC046	Authentication & Access	SSO Logout Issues	"Why am I still logged in after signing out?"
UC047	Software & Applications	Microsoft Teams Plugin	"How do I install the Teams integration in Outlook?"
UC048	Collaboration & Tools	Video Conferencing Setup	"How do I join a Zoom meeting with the conference room system?"
UC049	Hardware & Peripherals	Docking Station Connectivity	"My laptop won't recognize the docking station."
UC050	Software & Applications	Antivirus Definition Update	"How do I manually update antivirus definitions?"
UC051	Authentication & Access	CAS Login Loop	"Why do I keep getting redirected back to the login page?"
UC052	Authentication & Access	NetID Activation	"How do I activate my NetID for the first time?"
UC053	Authentication & Access	Session Timeout	"Why was I logged out suddenly?"
UC054	Connectivity & Network	Eduroam Configuration	"How do I connect to Eduroam Wi-Fi?"

UC055	Connectivity & Network	Bandwidth Monitoring	"Can I see how much data I'm using on campus Wi-Fi?"
UC056	Software & Applications	SPSS License Help	"SPSS is asking for a license key—what should I do?"
UC057	Software & Applications	Zoom Recording Access	"Where can I find my Zoom recordings?"
UC058	Hardware & Peripherals	Computer Lab Issues	"The keyboard in the lab isn't working."
UC059	Hardware & Peripherals	USB Device Not Recognized	"My flash drive isn't showing up."
UC060	Collaboration & Tools	Shared Drive Quota	"My Google Drive says it's full—what can I do?"
UC061	Security & Compliance	Login Attempt Alerts	"I received a login alert I don't recognize."
UC062	Security & Compliance	Data Encryption Policy	"Do I need to encrypt my files before sharing?"
UC063	Security & Compliance	Public Wi-Fi Safety	"Is it safe to use public Wi-Fi at the library?"
UC064	Software & Applications	RDP Shortcut Setup	"How do I create a shortcut to my remote desktop?"
UC065	Collaboration & Tools	Meeting Room Booking	"How do I reserve a conference room?"
UC066	Software & Applications	SSH Access Setup	"How do I SSH into the university servers?"
UC067	Connectivity & Network	Port Activation Request	"How can I request a network port in my office?"
UC068	Hardware & Peripherals	Screen Resolution Issues	"My display resolution is wrong—how do I fix it?"
UC069	Collaboration & Tools	Google Workspace Access	"How do I access Google Docs using my university email?"
UC070	Software & Applications	Email Signature Configuration	"How do I set up an email signature in Outlook?"

Table 1: Level 0 IT Support Use Cases for Bishop's University

4 LLM Comparison Matrix

Criteria	**GPT-4 (OpenAI)**	**Claude (Anthropic)**	**LLaMA 2 (Meta/HF)**	**Mistral (Mistral AI)**

Provider	OpenAI (USA, commercial)	Anthropic (USA, commercial)	Meta / Hugging Face (open-source via license)	Mistral AI (France, open-source)
Model Types	GPT-4 family (proprietary)	Claude 3 family (Proprietary)	LLaMA 2 family (7B, 13B, 70B, open)	Mistral family (7B, 8x7B mixture, open)
Release	2023 (GPT-4o in 2023)	2023 (Claude 3, etc.)	2023 (LLaMA 2 series)	2023 (Mistral 7B, Mixtral)
Key Strengths	*Top-tier generalist*: highest accuracy and reasoning on broad tasks (creative writing, coding, multi-modal)index=0. Integrated in many tools (Copilot, ChatGPT).	*Long-context & safety*: excels at summarization, instruction-following and safe outputs (trained with Constitutional AI). Good with large documents.	*Open and customizable*: strong baseline NLP performance given parameters. Free for re-search/commercial use. Fine-tunable by developers.	*Efficiency & open*: highly efficient inference (reportedly 1,000 words/sec on open-source engine). Good performance for its size. Suitable for on-prem deployment.
Performance	Very high (often leads benchmarks, excels in coding and reasoning).	High (comparable to GPT-4 in many tasks, especially summarization).	Moderate to high (70B close to GPT-4 performance on many tasks, smaller sizes less so).	Moderate to high (Mistral 7B often outperforms larger LLaMAs on some benchmarks).
Latency / Speed	*Slower inference*: large model size and cloud overhead can cause higher latency. Occasional delays reported.	*Moderate*: large model with long context is slower.	*Variable*: smaller (7B/13B) models can run quickly on suitable hardware; larger (70B) is slower unless optimized.	*Fast inference*: optimized for efficiency on modern hardware, supports fast response (report claims very low latency).
Cost	High (proprietary API pricing; e.g. GPT-4 costs \$0.03–\$0.06 per 1K tokens). Requires paid subscription or API usage.	High (proprietary API, similar cost profile, Anthropic Claude pricing is comparable).	Low (models are free; only cost is compute for hosting/fine-tuning).	Low (open and free; cost is com-pute/infrastructure only).
Integration Ease	Very easy (well-documented API and SDKs; integrated in many platforms). Broad tool support.	Easy (API available; some integration similar to GPT).	Moderate (no official cloud API; requires model hosting or use via Hugging Face or private infrastructure).	Moderate to easy (open models available via Hugging Face/Ollama; can be self-hosted or via community APIs).

Customization	Limited (no direct fine-tuning by users; can only prompt-engineer). Fine-tune options limited compared to open models.	Limited (no user fine-tuning; customize via prompts only).	High (full fine-tuning and prompt-tuning allowed under license; can adapt to domain data).	High (open-source; supports fine-tuning or adapter methods on private data).
Deployment	Cloud only (OpenAI API, Azure). No on-premise solution.	Cloud only (Anthropic API). No on-premise.	Flexible (can run on local servers, cloud, or edge; supports embedded runtimes).	Flexible (open, can run locally or on cloud; designed for edge and on-prem use).
Data Privacy	Low (data is sent to OpenAI; sensitive info depends on trust, but OpenAI does not use data for training by default). No complete control.	Low (data sent to Anthropic; similar privacy caveats as other cloud APIs).	High (can be hosted locally; data never leaves organization).	High (fully open-source; can deploy on-premise to keep data private).
Context Window	Up to 128K tokens (GPT-4o), enabling very long inputs.	Very large (200K tokens, excellent for lengthy documents).	Standard (4096 tokens for base models; some Llama-2 derivatives now support 32K).	Moderate (depends on variant; e.g. Mixtral 8x7B supports long context 32K).
Support / Community	Very extensive (large user base, many examples, continuous updates).	Growing (enterprise focus, but fewer community tools than OpenAI).	Large (open-source community, many forks and tools available).	Growing (new but active; open models available on Hugging Face; French-based ecosystem support).

4.1 LLM Section

One of the key decisions in the system design was the selection of Meta's LLaMA as the foundational large language model (LLM). This choice was driven by several compelling factors. First, LLaMA is one of the most advanced open-weight models available, offering strong performance across both general and domain-specific tasks without the licensing constraints of proprietary alternatives like OpenAI's GPT-4 or Anthropic's Claude. Its architecture is optimized for inference efficiency, making it well-suited for on-premises or cloud deployments with manageable resource demands, especially when fine-tuned on a domain-specific dataset. LLaMA supports continued pretraining and instruction tuning, which aligns with our objective of customizing the model to handle the nuances of university IT support inquiries. Furthermore, the availability of LLaMA on Hugging Face through the AutoTrain interface allows for streamlined fine-tuning and integration into modern retrieval-augmented generation (RAG) pipelines. Its strong support from the open-source community, including compatibility with LangChain and Hugging Face Transformers, makes it a practical and forward-compatible foundation for building a scalable, privacy-conscious, and adaptable chatbot.

5 Conclusion

This proposal outlines a comprehensive approach to developing an intelligent IT support chatbot destined for university environments. By leveraging publicly available knowledge bases, anonymized support logs, and curated documentation from leading technology providers and academic institutions, we ensure that the model is trained on relevant, high-quality data. The choice of Meta's LLaMA provides a powerful and open-source large language model foundation, offering flexibility, cost-efficiency, and strong performance for domain-specific tasks.

Our architecture combines modern backend frameworks, cloud-native infrastructure, and retrieval-augmented generation (RAG) techniques to deliver scalable and contextually accurate responses. With fine-tuning conducted through Hugging Face's AutoTrain and deployment on AWS using Docker and Kubernetes, the system is designed for both robustness and future scalability.

6 High-level system architecture

Appendix:

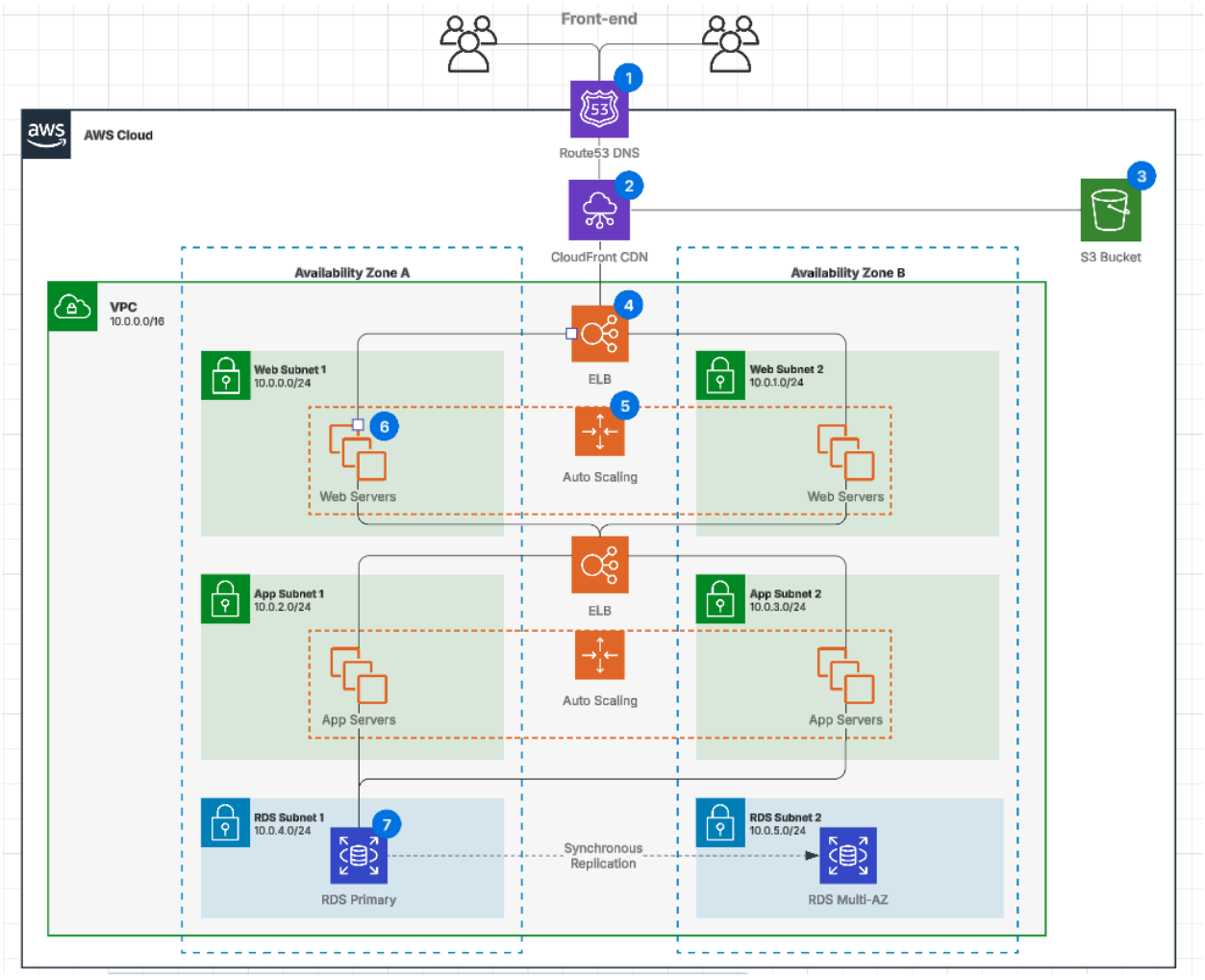

Figure 1: AWS Mori Hosting

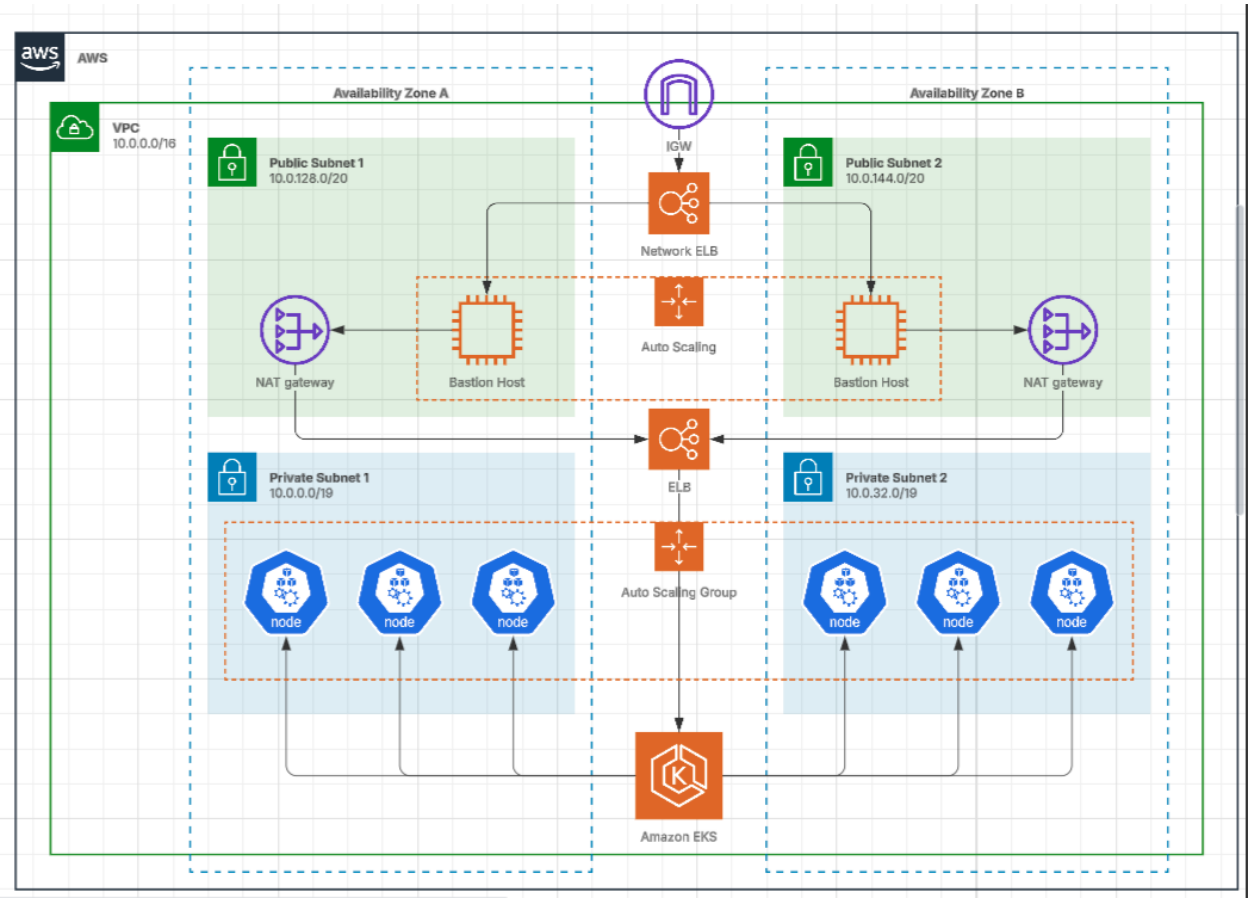

Figure 2: AWS Kubernetes nodes with EKS

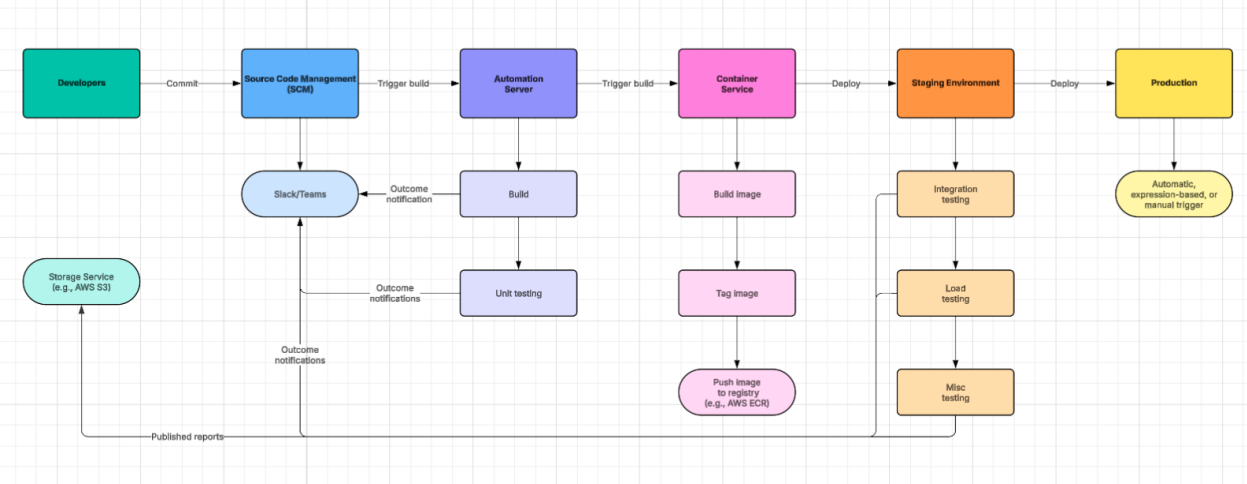

Figure 3: Deployment Automation

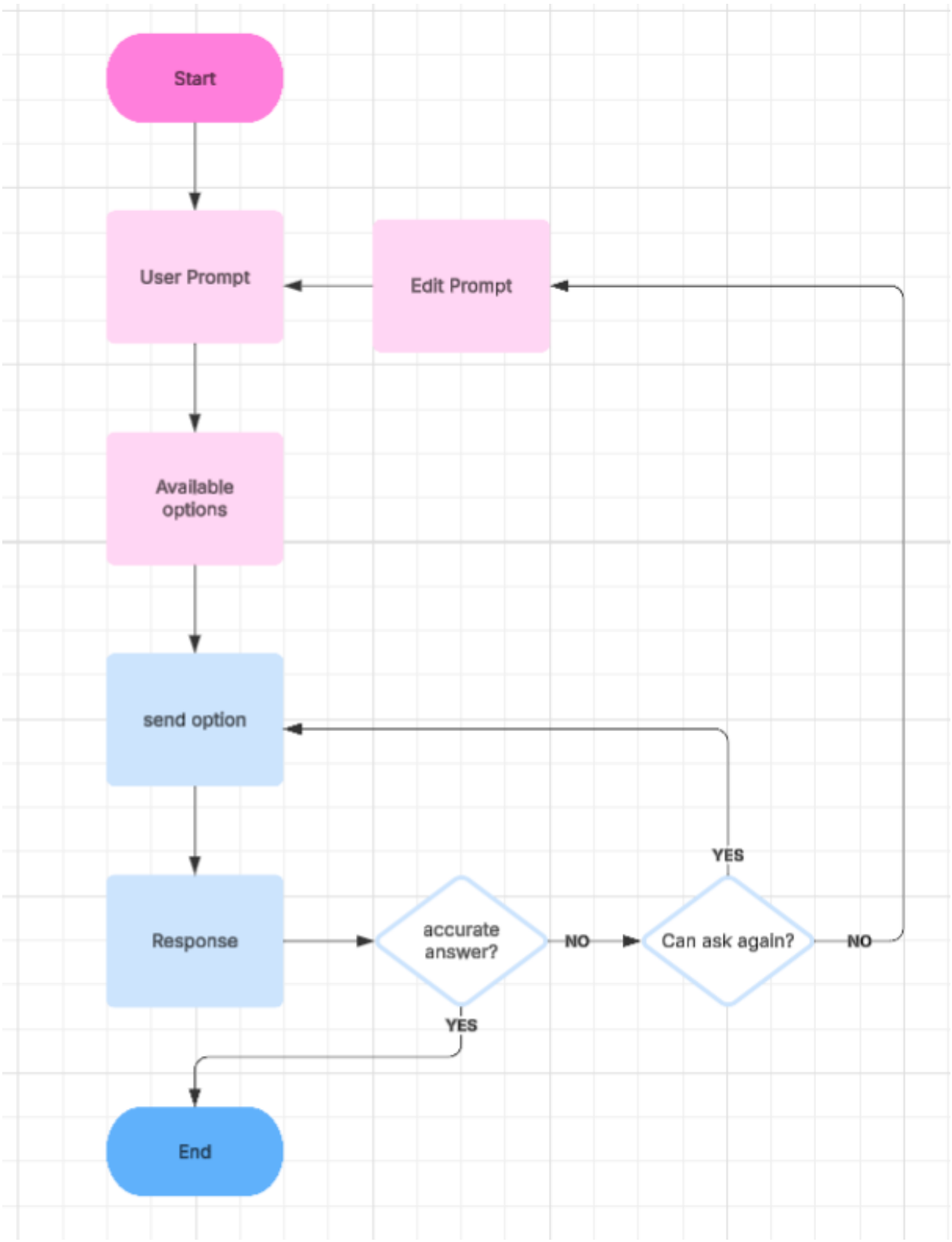

Figure 4: Flow Process of Mori

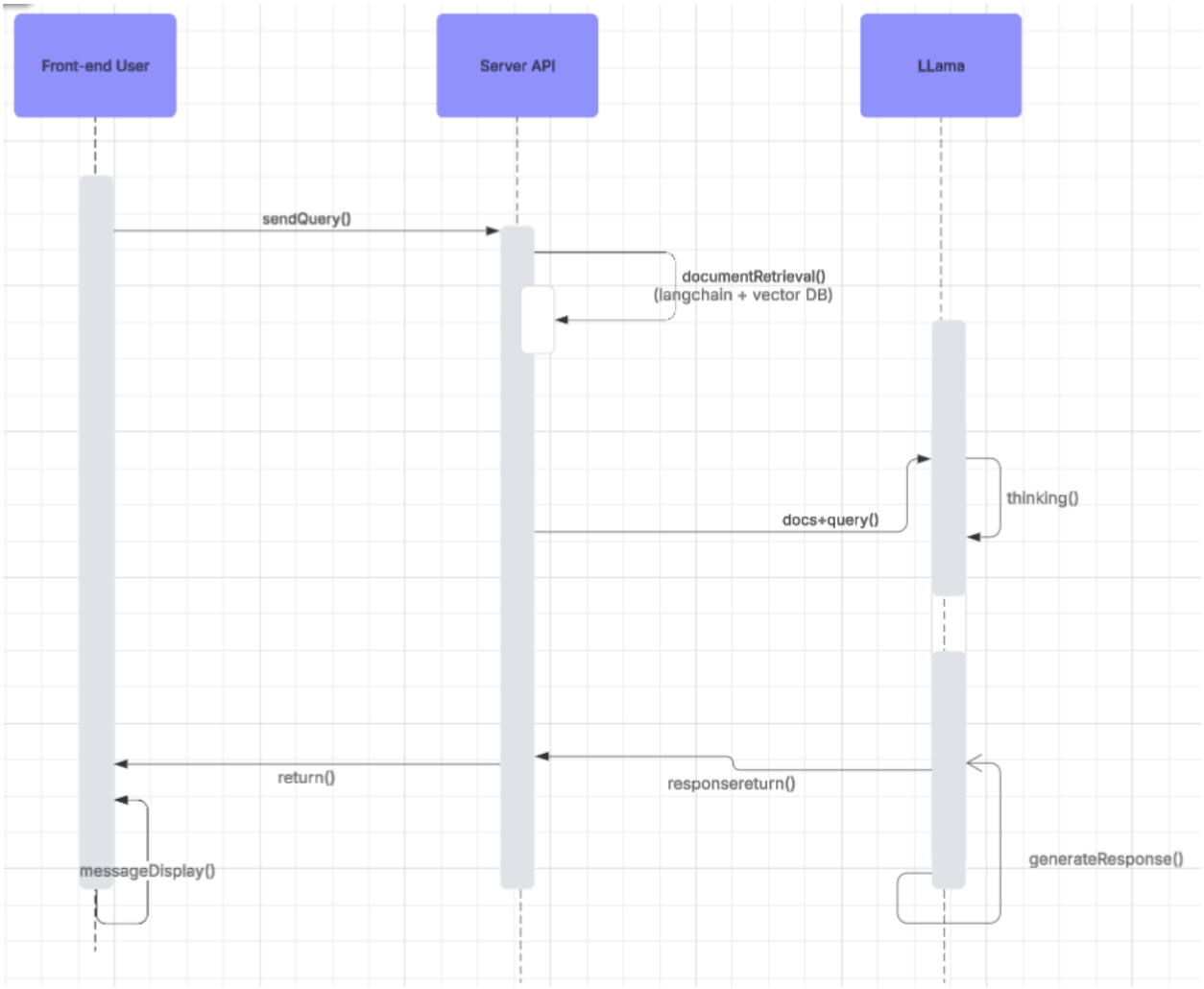

Figure 5: Sequence Flow Diagram

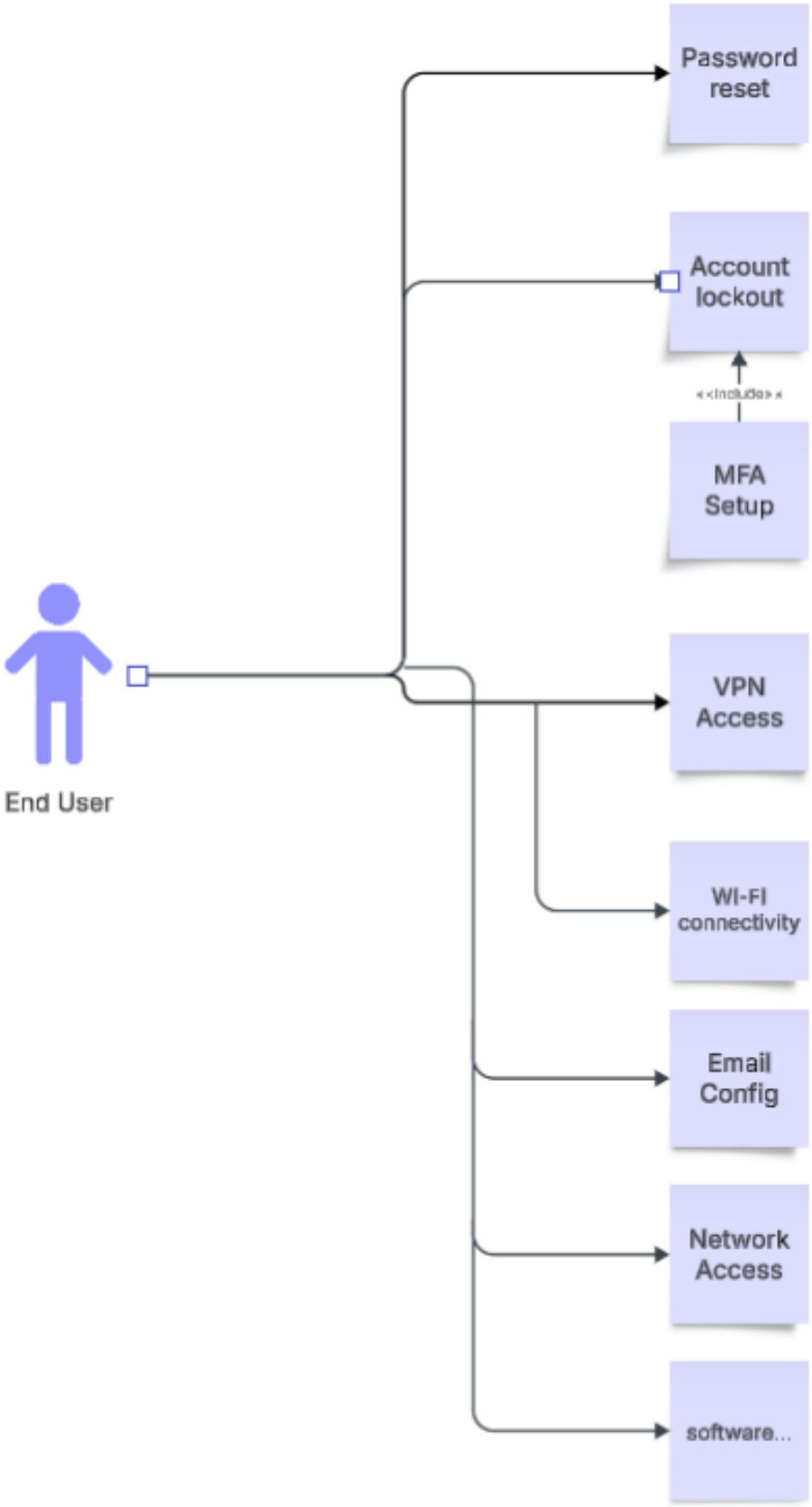

Figure 6: Mori Use Cases