Topic: Internal Bank Customer Support Chatbot

Team 09: Sambisha Godi, Mokhinur Talibzhanova, Drishti Chulani, Sarah Dsouza

Overview

This project focuses on an internal decision-support chatbot designed for banking customer support and fraud/compliance teams. The system assists human agents by categorizing incoming customer support tickets into predefined, policy-relevant classes (e.g., fraud-related issues, account access problems, or cases requiring escalation). In addition, it retrieves relevant policy and knowledge-base documents and suggests compliant draft responses. The chatbot functions as an internal "copilot" rather than an autonomous system; all final decisions and customer communications remain the responsibility of human agents.

System Purpose

The chatbot is intended to support consistent, policy-aligned handling of customer inquiries in a highly regulated environment. Its scope is intentionally limited to three core functions:

- Categorizing customer messages by issue type and risk class, rather than prioritizing or ranking them.
- Surfacing relevant internal-style policy and regulatory documents to inform agent judgment.
- Drafting preliminary responses that follow approved language and compliance constraints.

The system does not execute actions, resolve disputes, or communicate with customers without human review.

Ethical Analysis

Banking customer support involves significant ethical and legal obligations. Errors in categorization, unsupported guidance, or overly confident responses may result in financial harm to customers and regulatory or legal exposure for institutions. Ethical considerations therefore play a central role in system design, particularly with respect to safety, compliance, accountability, transparency, and the preservation of human oversight.

Knowledge Base and Corpus

- **Regulatory Guidance:** Includes **CFPB Regulation E** for error resolution and **FDIC** consumer compliance.
- **Risk Management:** Documentation from the **FFIEC (BSA/AML Manual)** and **Federal Reserve** publications regarding fraud prevention.
- **Internal Industry Standards:** Policy materials such as **JPMorgan's complaint handling procedures**

Stakeholders

The system affects multiple stakeholder groups:

- **Direct users:** Bank employees and customers
- **Affected individuals:** Bank and credit card companies
- **The organization:** Banking institutions responsible for consumer protection, compliance, and reputational risk.
- **Regulators:** Financial and banking authorities overseeing fraud prevention and customer protection like the FDIC, The Federal Reserve System, OCC
- **Vulnerable groups:** Elderly customers, individuals with low financial literacy, financially stressed users, etc

Metric–Value Tension

A central ethical tension arises between operational efficiency and risk-sensitive accuracy. A tempting metric to optimize is response speed or automation coverage, which reduces handling time and costs. However, over-optimizing for speed risks undermines:

- Accuracy of categorization
- Customer safety
- Regulatory compliance

In fraud and compliance contexts, slower but cautious categorization and escalation are ethically preferable to rapid but potentially misleading responses.

Automation Risks and Control Measures

Automated decision-support systems in banking introduce risks related to misclassification of high-risk inquiries, ungrounded or overly confident guidance, and overreliance on automation in sensitive contexts. These failures may delay escalation to fraud or compliance teams, imply outcomes not supported by policy, and reduce human judgment in complex or adversarial cases, increasing regulatory, financial, and ethical exposure.

To mitigate these risks, the system enforces conservative classification thresholds and mandatory human review for high-risk or ambiguous cases . We apply explicit "Abstain & Escalate" rules whenever inputs lack clarity or indicate potential harm. All outputs are strictly grounded in policy-specific document chunks using approved compliance language . Evaluation prioritizes escalation accuracy and policy alignment over response speed to reinforce human accountability and risk-sensitive handling .

Abstain and Escalate Rules

To reduce the risk of harm, the chatbot incorporates explicit abstention and escalation criteria. The system must refuse to answer or route cases to a human agent when:

- The situation involves urgency, emotional distress, or potential financial harm.
- The customer's request is complex, ambiguous, or highly individualized.
- Regulatory or policy guidance is unclear or conflicting.
- There are attempts to bypass security controls or manipulate the system.

These rules ensure that the system favors conservative escalation over unsupported confidence.

Prompt Coverage and Boundary Conditions

The chatbot is designed to handle well-defined, policy-aligned inquiries, such as lost cards, unauthorized transactions, account lockouts, and questions about investigation timelines or required documentation under Regulation E. When customer messages lack sufficient specificity, the system is expected to request clarification rather than infer intent. In adversarial or unsafe scenarios, such as impersonation attempts, demands that contradict regulatory timelines, or requests to conceal fraud, the chatbot must abstain and escalate without generating substantive guidance.

Evaluation & Failure Analysis

Evaluation results show that retrieval-augmented configurations produce more stable and policy-aligned triage decisions than LLM-only approaches, particularly in high-risk and ambiguous cases. RAG configurations with k=5 and T=0.0 setup reduce the variability and limit unsupported guidance, supporting a conservative, retrieval-first system design for banking contexts.

Failures are concentrated in ambiguous or borderline cases. Hallucinations were the most common error (114 cases, 54.28%), with LLM-only configurations accounting for 57 cases (27.14%), indicating a higher risk of ungrounded responses. Retrieval-related issues remained significant, including retrieval failures (54 cases, 25.71%) and RAG ignored evidence (54 cases, 25.71%). In addition, adversarial failures (43 cases, 20.47%) and harmful responses (19 cases, 9.05%) highlight the risk of misclassification in sensitive banking scenarios. Overall, while RAG reduced hallucinations compared to LLM-only setups, the results support the need for conservative thresholds, explicit abstain rules, and human review.

To illustrate the severity of these risks, consider a scenario where a distressed customer reports a $50,000 fraudulent withdrawal. If the system prioritizes **Speed** over **Accuracy**, it might provide a "hallucinated" assurance that funds are safe instead of escalating. This false security could cause the victim to miss the legal **Regulation E** "Notice of Error" window, resulting in permanent financial loss and bank liability for negligent misrepresentation. To mitigate this, we enforce **"Abstain & Escalate"** rules that detect urgency and distress, ensuring high-stakes cases are immediately routed to human experts.