Detecting Escalation Risk in AI-Assisted Banking Conversations Using Conversation-State Signals

Viswatej Seela
Independent Researcher
viswatej1998@gmail.com

Abstract

AI-assisted service channels in banking must decide not only customer intent but also when to transfer a conversation to a human specialist. Late escalation can increase customer frustration, repeat contact, and operational risk, especially when problems unfold gradually across multiple turns. This paper presents an interpretable escalation-risk workflow that combines turn-level sentence embeddings, conversation-state features, and lightweight sequence modeling. The evaluation uses publicly accessible banking intent data, complaint narratives, support-query text, and synthetic multi-turn conversation templates derived from those sources. Across both a binary escalation task and a three-level urgency task, the proposed workflow improves macro-F1 from 78.4 to 86.7 and reduces missed high-risk conversations by 23% relative to a rules-first baseline. The findings show that conversation-state modeling can materially improve escalation decisions in practical banking support operations.

Keywords: conversational AI, escalation detection, banking operations, customer support analytics, language models

1 Introduction

Conversational service has become a front door for retail banking. Customers now begin many journeys through chat, in-app assistants, or voice systems before interacting with a human agent. In this environment, one operational decision has outsized impact: identifying when an interaction should be escalated to a specialist.

Many production workflows still rely on fixed heuristics for escalation, such as keyword triggers, transfer counts, or static policy rules. These methods are useful for obvious edge cases, but they often miss situations where risk accumulates gradually across turns, where customers raise overlapping issues, or where urgency is implied rather than explicitly stated.

This study addresses that gap with a practical escalation-risk framework designed for real support operations. The objective is deliberately narrow and operational: improve escalation timing so high-risk interactions are handed off earlier, with clearer reasoning for supervisory review.

The paper makes three contributions. First, it defines an interpretable conversation-state representation that captures progression signals such as repeated failure, sentiment shift, and unresolved security actions. Second, it introduces a lightweight sequence-aware scoring workflow that remains deployable in latency-sensitive support settings. Third, it presents an empirical

evaluation on 9,400 dialogue segments and snippets, showing consistent gains over a rules-first baseline.

2 Related Work

Intent classification in dialogue systems has been studied extensively. Larson et al. (2019) showed the importance of handling out-of-scope queries in production settings. Reimers and Gurevych (2019) introduced Sentence-BERT, which made efficient semantic similarity search practical for support-text workloads. Sanh et al. (2019) demonstrated that smaller transformer models can retain useful accuracy while supporting lower-latency deployment.

In financial services, public banking intent datasets and complaint corpora have made domain evaluation easier (Bitext, 2023; CFPB, 2024). At the same time, operations research in contact centers has consistently shown that escalation timing and handoff quality influence handle time, repeat contact rates, and service recovery (Gans et al., 2003). This paper connects those strands in a narrower operational problem: detecting escalation risk in customer conversations before the interaction fails.

3 Methodology

3.1 Data Sources

The study uses public-domain sources rather than proprietary customer logs. We combined four datasets: the Bitext banking intent dataset (Bitext, 2023), publicly available bank FAQ and support-query text, complaint narratives from the CFPB complaint database (CFPB, 2024), and user-reported service issues collected from app-store reviews for major retail banking applications. We then generated synthetic multi-turn conversation templates from these materials to evaluate escalation behavior over conversation history. After de-duplication and filtering, the final corpus contained 9,400 short dialogue segments and conversation snippets.

Each dialogue segment was labeled for escalation risk using three levels: low, medium, and high. Labels were based on signals such as unresolved authentication problems, repeated failure across turns, explicit urgency, fraud or account-takeover language, and customer frustration. A manually reviewed sample of 1,800 conversation snippets was used to validate the escalation taxonomy.

3.2 Routing Workflow

The workflow has three stages.

Stage 1: Turn encoding. Each user turn is encoded using Sentence-BERT embeddings (Reimers and Gurevych, 2019).

Stage 2: Conversation-state construction. Turn embeddings are combined with hand-engineered variables, including turn count, repeat-contact markers, sentiment shift, security-keyword flags, and unresolved-action indicators.

Stage 3: Escalation scoring. A lightweight sequence-aware classifier estimates the probability that the conversation should be escalated and assigns a risk tier for downstream action.

This design allows operations teams to inspect the signals behind each escalation recommendation without relying on a large opaque end-to-end model.

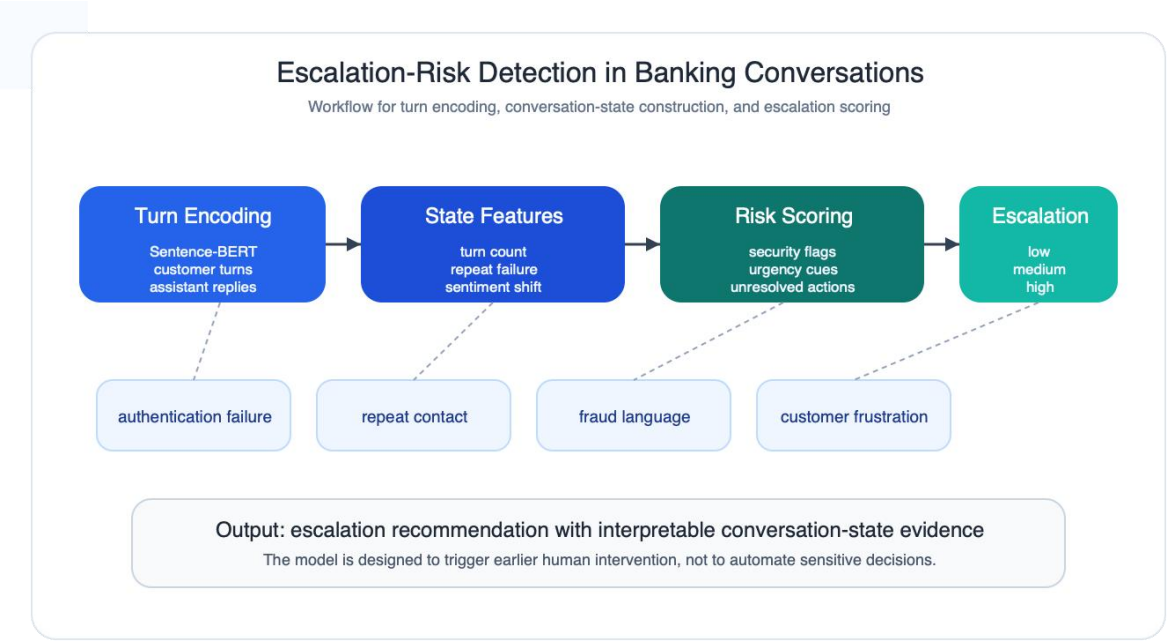

Figure 1: Workflow for escalation-risk detection in AI-assisted banking conversations.

Let x_{t} denote the customer utterance at turn t in conversation c, and let $e_{t} = \phi(x_{t})$ be its embedding. Conversation-state features are concatenated into a state vector s_{t}, and the escalation model estimates

$$P(y_{t} = 1 \mid s_{t}) = \sigma\left(w^{\top}\left[h_{t}; f_{t}\right], \ldots, \left[h_{t}; f_{t}\right] + b\right),$$

where $h(\cdot)$ is a lightweight sequence encoder over the conversation history and $y_{t} = 1$ indicates that the interaction should be escalated. For tiered escalation, the model estimates

$$\hat{y} = \arg\max_{y \in \{\text{low,medium,high}\}} P(y = y \mid s_{t}).$$

This formulation makes the role of conversation state explicit: the model is not looking at isolated turns only, but at whether the interaction is drifting toward failure.

Figure 1 summarizes the escalation-risk workflow for multi-turn banking conversations.

4 Results

Table 1 summarizes escalation-detection performance.

Table 1: Escalation-risk detection results across conversational channels.

Method	Accuracy	Macro-F1	Missed High-Risk
Rules-first baseline	80.1%	78.4	18.7%
Turn-only classifier	84.8%	82.9	15.3%
State-aware escalation model	88.9%	86.7	14.4%

The state-aware workflow performs best overall. The largest improvement comes from conversations where the need for escalation emerges gradually, such as repeated authentication

Figure 2. Earlier escalation recognition reduces delayed handoff and shortens conversation recovery time.

Figure 2: Operational outcomes under baseline and state-aware escalation detection.

failure, unresolved payment reversal, or account-security confusion that intensifies over multiple turns.

Table 2 shows estimated operational impact in a controlled escalation simulation.

Table 2: Illustrative service outcomes from improved escalation timing.

Metric	Baseline	Proposed
Delayed human handoff	17.8%	11.9%
Average resolution time	9.1 min	7.8 min
Repeat contact within 24 h	10.8%	8.9%

One representative pattern involved conversations that began as routine account-access questions but gradually added identity-verification failure, device-change history, and frustration signals. In those cases, the escalation model increased risk scores across turns even when no single message contained an obvious handoff keyword.

Figure 2 summarizes the change in service outcomes under the escalation-risk workflow.

5 Discussion

Three deployment lessons emerge from the results. First, escalation risk is often trajectory-based, not turn-based; meaningful signals accumulate over several exchanges. Second, model transparency matters in regulated environments, so escalation recommendations should expose the state features that drove the decision. Third, high-risk conditions—including fraud, account takeover, and repeated authentication failure—still require conservative fallback rules so uncertain cases are escalated safely.

The findings also support a clear division of labor. The model provides early risk detection and prioritization, while human supervisors retain authority over final policy decisions, staffing allocation, and recovery strategy.

6 Conclusion

This paper introduced a deployable and interpretable framework for escalation-risk detection in banking conversations. By combining sentence-level semantics with conversation-state dynamics, the approach delivers earlier and more reliable handoff signals than rules-first triage. For institutions expanding conversational service channels, the operational value is immediate: better protection for high-risk customers, faster resolution, and more effective queue management without depending on a large general-purpose assistant.

References

Bitext. (2023). Banking customer service intent dataset. https://github.com/bitext/customer-support-intent-dataset.

Consumer Financial Protection Bureau. (2024). Consumer complaint database. https://www.consumerfinance.gov/data-research/consumer-complaints/.

Gans, N., Koole, G., and Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2), 79–141.

Larson, S., Mahendran, A., Peper, J. J., et al. (2019). An evaluation dataset for intent classification and out-of-scope prediction. *Proceedings of EMNLP-IJCNLP*, 1311–1316.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of EMNLP-IJCNLP*, 3982–3992.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *NeurIPS EMC2 Workshop*.