Hierarchical Reason-of-Contact Detection in Retail Banking Customer Interactions via LLM-Driven Taxonomy Induction

Yoages Kumar Mantri* **Sriraja Yagneswaran*** **John Hartley*** **Greig Cowan***
Raad Khraishi* **Abhinav Gupta***
NatWest AI Research
{Yoageskumar.mantri, Yagneswaran.Sriraja, John.Hartley, Greig.Cowan,
Raad.Khraishi, Abhinav.Gupta1}@natwest.com

Abstract

Retail banks handle high volumes of customer interactions across different channels that span various topics. Early and accurate detection of the intent of the customer is critical towards streamlining contact-center operations through efficient routing and handling of conversations. Mining of customer interactions leads to identification of friction points in customer journeys and offers valuable insights about customer needs. Existing approaches to define customer intents or contact reasons remain fragmented, manually maintained across organizations and relying on knowledge of specific business processes. We propose a framework that develops a dynamic hierarchical Reason-of-Contact (RoC) taxonomy to cover customer topics across hundreds of business processes. We further demonstrate the implementation of this taxonomy to a robust solution that identifies intents for all customer conversations across different channels. Our deployed system supports real time use with a 150 to 300 ms turnaround per conversation. It achieves up to 10% improvement in F1 score over baseline approaches on a reference dataset. We also detail deployment considerations, including dynamic taxonomy updates, out-of-domain detection, and auditability. Finally, we present ablations and error analyses to characterize effectiveness.

1 Introduction

In the high-stakes environment of retail banking, every customer interaction carries a significant operational and emotional weight. This remains true whether the interaction is a report of a lost card or a sensitive inquiry regarding bereavement. Accurately identifying the *Reason-of-Contact (RoC)* is the pivotal for efficient routing, regulatory compliance, and overall customer satisfaction. However, traditional intent classification systems often fail in production because they treat language as a static set of labels rather than an evolving hierarchy of needs. Rule-based systems are often too brittle to capture the linguistic diversity of thousands of daily callers. Simultaneously, standard supervised models suffer from taxonomy drift, a phenomenon where predefined categories become obsolete as new business processes emerge (Casanueva et al., 2020).

We present a deployed framework that bridges the gap between the scalability of unsupervised learning and the precision of human-in-the-loop expert systems. Our approach moves beyond flat intent detection by inducing a three-level (L_0-L_1-L_2) hierarchical taxonomy that is uniform across businesses (Zhu et al., 2025). This unification is critical; by standardizing the nomenclature, we have reduced friction in cross-departmental handovers and met with a positive response from business stakeholders.

The framework has two stages shown in Figure 1. In the offline stage, we leverage Large Language Models (LLMs) with iterative prompting to cluster and label raw conversation data into a logical hierarchy, subsequently refined by domain experts. The online stage performs dual-pass inference: identifying a *First RoC* from early utterances for immediate routing and an *Overall RoC* for post-call analytics. This architecture achieves a low-latency of 150-300 ms, enabling real-time deployment at a scale of 80,000 daily interactions.

Our contributions are:

1. A two-stage *Reason-of-contact (RoC)* taxonomy generation framework that blends unsupervised intent clustering with LLM-driven hierarchical induction. This is exemplified with the generation of a three-level retail banking taxonomy with descriptions at each level, using iterative LLM prompting and domain expert validation.

*All authors contributed equally to this work.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 1555–1566
July 2-7, 2026 ©2026 Association for Computational Linguistics

2. Development and deployment of a live *Reason-of-contact (RoC)* solution using the above taxonomy to categorize diverse customer conversations across voice and chat channels. The solution is able to easily handle high volumes of conversation ($\approx$ 80,000 per day) while exceeding existing performance baselines.

We demonstrate improved performance on real-world conversation data using our approach in comparison to standard baselines. The paper includes business scenarios, ablations, error analysis and learnings to highlight practical impact and benefits.

2 Related Work

Traditional RoC detection relies on either rule-based or supervised learning. Rule-based methods are interpretable but struggle with paraphrases and emerging intents. Supervised classifiers achieve high accuracy on fixed taxonomies but require extensive labeled data (Gupta et al., 2019) and retraining when new intent categories appear, which is poorly suited for dynamic nature of retail banking environments (Chatterjee and Sengupta, 2020).

Unsupervised intent discovery methods attempt to mitigate these labeling costs by clustering conversation logs without manual intervention (Padmasundari and Bangalore, 2018). Techniques such as ITER-DBSCAN address unbalanced intent distributions (Chatterjee and Sengupta, 2020), while frameworks like USNID focus on estimating cluster counts in open-world settings (Zhang et al., 2024). However, these approaches often produce flat clusters that lack the hierarchical interpretability required for complex banking operations (Liu et al., 2021).

LLM-assisted taxonomy induction has recently shown strong results. Chain-of-Layer (CoL) uses iterative prompting to construct structured taxonomies while mitigating hallucinations via ranking filters (Zeng et al., 2024).

Regarding deployment and inference, dense embedding matching with approximate nearest neighbor search supports efficient classification and incremental updates without model retraining (Johnson et al., 2017). While Retrieval-Augmented Generation (RAG) and hybrid methods like LARA (Liu et al., 2024) offer flexibility, they often incur significant per-token costs and lack the determinism required for regulated financial intent labeling. Our framework diverges from these by utilizing LLMs strictly for infrequent offline taxonomy induction. For online inference, we rely on interpretable nearest-neighbor matches. This design choice helps achieving 150 to 300 ms while allowing for rapid taxonomy updates through simple embedding updates rather than exhaustive knowledge base rebuilds.

3 Methodology

Our framework comprises offline and online stages. In the offline stage, we generate and refine a hierarchical RoC taxonomy from historical conversation data. In the online stage, we classify live interactions against this taxonomy to determine both a First RoC and an Overall RoC (see Figure 1).

3.1 Offline Taxonomy Generation

The offline stage transforms historical customer interactions into a structured, three-level RoC hierarchy. This process moves from raw signal to semantic clusters, and finally to a validated taxonomy through a combination of unsupervised learning, LLM-driven induction, and Subject Matter Expert (SME) validation.

3.1.1 Data Preparation

We source historical data from both chat and voice channels. While chat interactions are characterized by brevity and typographical errors, voice transcripts often contain longer sentences and frequent speaker interruptions (see Appendix A.2 and A.3 for sample chats). To balance computational tractability with intent coverage, we sample $K = 50{,}000$ conversations. As primary intent is typically established early, we retain only the first ten customer utterances per interaction.

Pre-processing involves standardizing text through acronym expansion, URL removal, and NER-based masking of personally identifiable information (PII). We utilize a curated lexicon to filter non-intent-bearing turns, such as *hello* or *thank you*. Crucially, to handle multi-intent turns, we segment utterances at punctuation and conjunction boundaries to create a set of atomic utterances U. Exact duplicates are removed to prevent templated system responses from biasing subsequent cluster formation. Finally, SMEs verify intent preservation by comparing U with original transcripts on a stratified subsample of conversations.

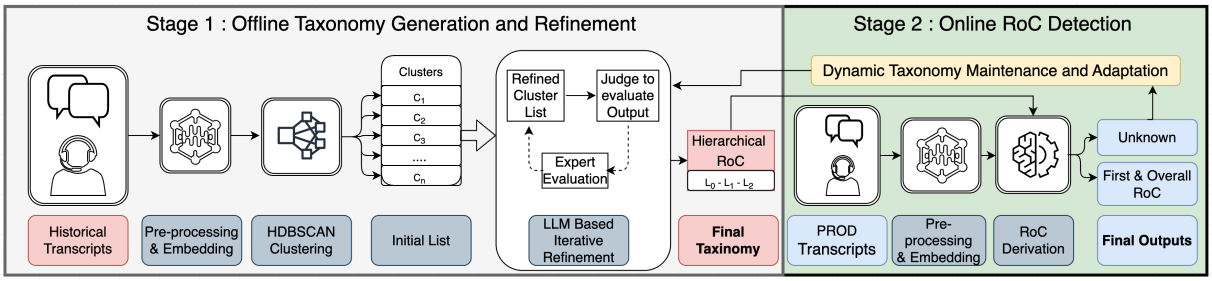

Figure 1: Framework Overview: two-stage RoC detection pipeline.

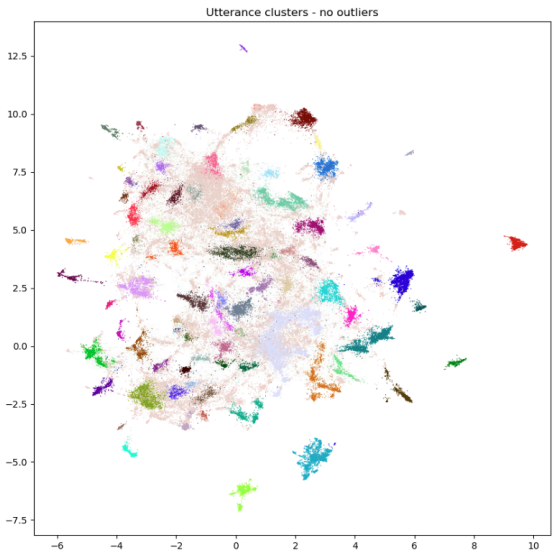

Figure 2: A 2-dimensional t-SNE plot of utterance clusters, with each color indicating a distinct cluster.

3.1.2 Thematic Clustering via Embeddings

Each atomic utterance $u \in U$ is embedded using the *all-MiniLM-L6-v2* model from Sentence-BERT (Reimers and Gurevych, 2019), yielding a dense semantic vector set $V = \{v_i \in R^{384} \mid u_i \in U\}_{i=1}^{|U|}$. We chose this model based on an empirical study of ten candidates, prioritizing the balance between semantic fidelity and the sub-millisecond inference speeds required for production. We observe the induced intents are semantically grounded, as they reside within the dense regions of the sampled utterances from which they were extracted. See Figure 3 in the Appendix A.5.

To identify intent boundaries, we apply UMAP (McInnes et al., 2020) for dimensionality reduction followed by HDBSCAN clustering (McInnes et al., 2017) (see Experiment 1 in Appendix A.1). Figure 2 illustrates the high-dimensional structure of the resulting clusters, projected into two dimensions using t-SNE.

3.1.3 LLM-Driven Hierarchical Taxonomy

Manual inspection revealed that these clusters represent broad themes rather than discrete actions. A single direct debit cluster may contain distinct intents such as *cancel direct debit*, *reinstate direct debit*, and *query direct debit amount*, indicating that clusters capture thematic similarity and cannot be converted into intents via simple cluster summarization (Lee and Lee, 2025).

Intent frequency within clusters follows a long-tail distribution. Although many intents are rare, the absolute interaction volume in retail banking makes coverage of these tail intents operationally important. We therefore apply LLM-based intent extraction per cluster (Wan et al., 2024). To maximize coverage, we use hybrid sampling over cluster embeddings , combining furthest point sampling (semantic diversity) with uniform random sampling (high-frequency coverage). The sampled utterances are then passed to an LLM to enumerate and label distinct intents.

We use an LLM to organize extracted intents into a three-level (L_0-L_1-L_2) hierarchical taxonomy under the MECE principle. The final taxonomy $\mathcal{T}$ has 14 L_0 nodes to more than 200 granular L_2 nodes with appropriate descriptions at each node (see sample rows in Table 4 in Appendix A.5). Each L_2 has a parent L_1 which has a parent L_0. Parent and child levels are induced jointly to produce coherent parent-child structure and non-overlapping siblings. SME feedbacks are incorporated into subsequent prompts until the taxonomy is semantically coherent and business-aligned.

3.1.4 Refinement and SME Alignment

Our induction process is iterative. LLMs are prompted to generate the hierarchy, which is then refined by SMEs to resolve three specific observations:

1. **Cross-branch Collisions:** Ensuring that near-duplicate labels (e.g., Update Address) are

uniquely assigned to the most relevant business parent.

2. **Granularity Balancing:** Collapsing vague or unsupported categories while preserving sensitive, high-stakes long-tail intents.

3. **Operational Alignment:** Grouping rare, low-frequency topics into a curated *General Inquiries* bucket to maintain taxonomy parsimony.

3.2 Online RoC Detection

The online stage categorizes live customer interactions against the three-level taxonomy $\mathcal{T}$. This stage provides two distinct outputs: a *First RoC* for early-stream intent estimation and an *Overall RoC* to capture the dominant intent of the complete interaction. These RoCs are calculated at each level of $\mathcal{T}$ and consolidated later.

3.2.1 First RoC Detection

The *First RoC* is estimated using the initial N customer utterances of the conversation. These utterances are embedded using the same SBERT model defined above to obtain v_{first}. The predicted node t^* is identified by maximizing the cosine similarity between the utterance embedding and the candidate node embeddings e_t.

3.2.2 Overall RoC Detection

To identify the *Overall RoC*, we employ a sliding window approach that aggregates evidence across the entire conversation. We slide a window of size N over customer utterances. A single embedding is generated for the collated N customer utterances in a window. For every position, we compute the similarity against all nodes in $\mathcal{T}$ and record the top M_k candidate nodes.

As the window moves through the transcript, we maintain a list of candidates. For each candidate node t_j that appears in the top M_k lists, we calculate its match frequency O_j and its mean similarity S_j. Nodes are then ranked by O_j to assign a rank $r_j \in [1, 5]$. The final importance of a node is determined by a rank-aware scoring function (Järvelin and Kekäläinen, 2002):

$$\text{Score}(t_j) = 2^{-r_j+1} \cdot O_j \cdot S_j$$

The *Overall RoC* is the node with the highest score. In the event of a tie, the system prioritizes the node with the higher frequency O_j, followed by the higher mean similarity S_j.

3.2.3 Hierarchical Alignment of taxonomy

To ensure the final classification is logically sound and safe for a regulated banking environment, we apply two post-processing steps:

1. **Hierarchical Consistency:** We enforce a top-down lineage to ensure RoCs computed at each level are consistent. If L_2 prediction is not a descendant of a L_1 prediction, the L_2 node takes precedence due to its higher specificity. L_0, L_1 labels are then realigned to match the L_2 parentage. For example, if L_2 identifies *Credit Card Fraud* but L_1 incorrectly suggests *Debit Card Services*, the framework corrects the L_1 label to *Credit Card Services*.

2. **Out-of-Domain (OOD) Detection:** To prevent forced misclassifications, we apply a similarity score threshold τ_{OOD} (=0.3). If the maximum similarity score at any level fails to meet this threshold, the interaction is labeled as UNKNOWN. These cases are routed to fallback queues or flagged for manual review, ensuring that the system does not generate high-confidence labels for anomalous or unsupported topics.

3.3 Dynamic Taxonomy Maintenance and Adaptation

We monitor the volume of utterances labeled UNKNOWN by tracking the frequency of cosine similarity scores falling below τ_{OOD}. If this volume exceeds the threshold τ_{unk} (=10%), we trigger an incremental realignment rather than re-running the offline stage.

This adaptation follows a four-step process. First, we gather and re-cluster all UNKNOWN utterances. Second, we compute the cosine similarity between each new cluster centroid and existing L_2 node embeddings. If the maximum similarity is below a merge threshold τ_{merge} (=0.5), the cluster is marked as a candidate for a new L_2 node; otherwise, it is absorbed into the closest existing node. Third, we invoke an LLM to generate labels and descriptions for these new clusters, followed by a brief human review. Finally, the embeddings for newly validated nodes are inserted into the FAISS index. (See all the thresholds in Table 5 in the Appendix A.5)

3.3.1 Operational Observations

The performance of the framework is influenced by specific linguistic characteristics inherent to customer service interactions. These observations directly inform our maintenance strategy and threshold tuning, particularly at the L_2 level of the taxonomy.

One prominent characteristic observed is the presence of compound-intent utterances, where customers combine multiple requests into a single utterance. E.g., an interaction may simultaneously involve a fund transfer and a password reset. To maintain high-precision routing, our framework identifies only the most dominant intent. In this configuration, secondary intents are intentionally subsumed for operational simplicity, representing an inherent property of the interaction data.

Another characteristic observed is semantic proximity between sibling nodes in the taxonomy. Certain intents are naturally proximal and share overlapping lexical cues. For example, *query direct debit amount* and *query direct debit status* often appear in nearly identical conversational contexts. This proximity is a natural feature of granular taxonomies. These overlaps are most visible at the L_2 level, where label boundaries are at their finest and descriptions are closely situated in the embedding space.

We navigate this density through the hierarchical consistency logic and OOD thresholds described in Section 3.2.3. These domain-specific traits reinforce the value of our two-stage approach, as periodic maintenance allows the system to refine these boundaries as the underlying data distribution matures.

4 Experiments and Results

We validate the coverage of our hierarchical taxonomy using 375,280 utterances extracted from diverse banking business processes. Coverage is quantified by calculating the cosine similarity between each utterance and its nearest intent in the taxonomy $\mathcal{T}$. At an increased similarity score threshold τ_{OOD} of 0.7, we observed that less than 5% of the utterances were flagged as UNKNOWN. This high coverage rate validates the effectiveness of the offline stage described in Section 3.1.

For the core RoC detection task, we evaluate the framework on 5,000 ground-truth-annotated conversations per channel (chat and voice) across ten distinct business lines. We report both First RoC and Overall RoC performance across the L_0, L_1, and L_2 levels. Our baselines include three representative approaches: (i) a BERT-based classifier fine-tuned at the L_2 level with conversational context, (ii) a CRF turn-level sequence model where the final intent is determined by the highest frequency, and (iii) a system based on predefined Business Rules. The BERT model serves as the primary neural baseline for both First and Overall RoC evaluation.

Table 1 summarizes results. Our system outperforms all baselines at every level, with the largest gains at L_2, where fine-grained intent separation is required. Classification with the Business rules perform worst due to limited contextual robustness. The CRF baseline is competitive at L_0 but degrades with taxonomy depth. These gains stem from a flexible hierarchical taxonomy and embedding-based matching. Supervised models struggle with many classes and limited labels, whereas our approach leverages unlabeled structure plus LLM-guided labeling. Macro-F1 is consistently lower than micro-F1 due to long-tail intents, and our method improves tail recall through cluster-driven intent extraction.

Ablations: We ablate key design choices (Table 2): taxonomy depth (two levels only), rank weighting (uniform weights instead of exponential decay), and LLM prompting (with and without iterative refinement). Adding L_2 increases specificity without material loss at higher levels. Uniform rank weights reduce performance, showing that rank-aware weighting improves aggregation. Removing iterative taxonomy refinement causes the largest drop, confirming that feedback-driven prompting is essential for taxonomy coherence and label quality.

5 Deployment and Operational Impact

The Reason-of-Contact (RoC) detection framework is currently deployed in a cloud production environment, processing approximately 35,000 to 40,000 chat and voice interactions on average per day. By transitioning from a legacy flat taxonomy to our hierarchical approach, we achieved a runtime of 150 to 300 ms per conversation. This represents over 30% improvement in processing efficiency over the previous solution. We have implemented several optimizations in the RoC detection pipeline like that of use of FAISS, to maintain low-latency performance at scale. Sample runtime logs for both channels are provided in Figures 4 and 5 in the

Model	L_0		L_1		L_2	
	Macro-F1	**Micro-F1**	**Macro-F1**	**Micro-F1**	**Macro-F1**	**Micro-F1**
Business Rules (overall)	0.55	0.60	0.42	0.48	0.35	0.40
CRF (overall)	0.68	0.70	0.62	0.64	0.54	0.54
BERT (first)	0.70	0.72	0.66	0.70	0.52	0.58
BERT (overall)	0.75	0.78	0.68	0.72	0.61	0.66
Our system (first)	**0.72**	**0.74**	**0.69**	**0.72**	**0.65**	**0.66**
Our system (overall)	**0.80**	**0.81**	**0.75**	**0.76**	**0.71**	**0.74**

Table 1: Comparison of RoC detection performance by model and taxonomy level.

Variants	L_0 **F1**	L_1 **F1**	L_2 **F1**
Full model	0.80	0.75	0.71
Depth = 2 (no L_2)	0.80	0.74	–
Uniform Rank Weights	0.72	0.68	0.60
No LLM iterations	0.60	0.52	0.42

Table 2: Ablation study results (Overall RoC Macro F1).

Appendix A.5 show-casing daily batch processing time taken.

5.1 Post-Deployment Solution Management

To ensure the solution is compliant with the various governance and regulatory requirements, several guardrails and utilities have been developed. These include modules for detection of PII, redaction verification layers and monitoring of model drift. The operational and performance metrics are monitored via dashboards and provide up-to-date information to relevant stakeholders. Documented details are provided to perform updates of various thresholds, OOD drift review cadence, rollback execution, and post-incident analyses. Anomalous traffic exceeding τ_{unk} (=0.5) triggers near-real-time alerts to the operations team. This mechanism has reduced instances of conversations being mis-routed, with certain teams reporting improvements of at least 20%. Operational teams have been appreciative of how the uniform taxonomy has enabled better hand-offs.

The robustness of the solution was evidenced through an ad-hoc update of the taxonomy, necessitated by the introduction of newer services that meant relevant customer queries being miss-classified. This process of identifying new nodes, inserting them at the right place in the hierarchy, validation and deployment was completed within 2 weeks without impacting any of the existing process.

5.2 Practical Considerations

The production roll-out revealed practical gaps between business expectations and the linguistic boundaries of customer data.

One primary observation involved requests for hyper-granularity. Certain business units requested the classification of over 100 sub-processes within a single product line. However, empirical analysis showed that these processes were not semantically separable within raw utterances. Attempting to force the taxonomy to this depth caused unstable classification and a significant drop in performance.

A second observation concerned cross-divisional intent duplication. Some departments requested separate intent nodes for identical processes, such as mobile number changes, to simplify their specific reporting. Segmenting these identical intents by business unit increased label collisions and degraded overall accuracy. To resolve this, we standardized these as global intents and used downstream metadata for routing. These findings highlight that successful RoC detection needs a balance between business specificity and separability.

Other examples can be seen in Appendix A.4.

6 Conclusion and Future Work

We presented a hierarchical RoC detection pipeline for retail banking dialogues that combines SBERT-HDBSCAN clustering with LLM-driven taxonomy induction and FAISS-based inference. The dual-mode classification approach enables both early routing and end-of-conversation labeling. The system offers a practical, explainable, and cost-efficient solution for enterprise-grade RoC detection, while maintaining safety, compliance, and governance standards.

For future work, we want to extend the framework to Agentic-AI as the taxonomy is uniform across. This is expected to help with faster taxonomy maintenance and evolution with day-to-day trends, address the coverage of niche conversation topics and allow for flexibility of granularity. We want to extend our current framework to support multi-intent handling instead of just single-label classification.

Acknowledgments

The authors wish to express their gratitude to Karen Dewar, Chief Data and Analytics Officer, and Graham Smith, Head of Data Science and Innovation, for their invaluable support and for fostering a constructive research environment that was essential throughout this project. Additionally, we thank Grant Falconer for his support and enablement that helped to manage the challenges and ensured smooth delivery of the solution. Lastly, we would like to extend our appreciation to all colleagues who contributed to the development of the project.

Ethical Considerations

This work presents a deployed Reason-of-Contact (RoC) detection system operating within a regulated retail banking environment. Because the system processes large volumes of customer interaction data, considerations around privacy, fairness, transparency, and responsible deployment are central to its design.

Data Privacy and Regulatory Compliance Customer conversations are processed under established governance controls and in compliance with applicable regulations (e.g., GDPR). Personally identifiable information (PII) is masked during pre-processing using NER-based detection before clustering, embedding, or taxonomy induction. Large Language Models (LLMs) are used only in the offline taxonomy generation stage on curated, privacy-filtered samples. Online inference relies on embedding similarity search and does not transmit raw customer conversations to external LLM services. Access to data and logs is restricted through role-based controls with audit trails for traceability.

Bias and Fairness The system performs intent classification and does not explicitly model demographic or protected attributes. Embedding-based similarity search is designed to capture semantic content rather than user identity. Nevertheless, linguistic variation correlated with demographic factors may influence performance. To mitigate this risk, we avoid encoding protected attributes, periodically review performance across available segments where appropriate, monitor UNKNOWN rates and misclassifications patterns, and incorporate Subject Matter Expert (SME) validation during taxonomy updates. The system is not a dedicated fairness auditing tool and operates under human oversight.

Transparency and Human Oversight The framework prioritizes interpretability and auditability. It employs a human-validated hierarchical taxonomy, explicit similarity scoring, and out-of-domain (OOD) thresholds to prevent forced misclassifications. Predictions are logged with candidate matches and confidence scores to support review and root-cause analysis. RoC outputs assist routing and analytics but do not directly trigger irreversible customer-facing decisions. Human override and controlled update procedures remain in place.

Limitations and Responsible Use The framework currently supports single-label classification and prioritizes dominant intents in compound utterances. Similarity-based decisions are sensitive to threshold calibration and semantically proximal labels. Emerging intents may be temporarily misclassified prior to taxonomy updates. Responsible deployment therefore requires continuous monitoring of drift, structured taxonomy maintenance, SME involvement, and clearly defined operational playbooks. RoC outputs should be interpreted as routing signals rather than definitive representations of customer intent.

References

Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. 1999. Optics: ordering points to identify the clustering structure. *SIGMOD Rec.*, 28(2):49–60.

Aymeric Beauchamp, Thi-Bich-Hanh Dao, Samir Loudni, and Christel Vrain. 2023. Incremental Constrained Clustering by Minimal Weighted Modification. In *29th International Conference on Principles and Practice of Constraint Programming (CP 2023)*, volume 280 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 10:1–10:22, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. *Preprint*, arXiv:2003.04807.

Ajay Chatterjee and Shubhashis Sengupta. 2020. Intent mining from past conversations for conversational agent. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4140–4152, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Arshit Gupta, John Hewitt, and Katrin Kirchhoff. 2019. Simple, fast, accurate intent classification and slot

labeling for goal-oriented dialogue systems. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 46–55, Stockholm, Sweden. Association for Computational Linguistics.

Kurt Hornik, Ingo Feinerer, Martin Kober, and Christian Buchta. 2012. Spherical k-means clustering. *Journal of Statistical Software*, 50(10):1–22.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *Preprint*, arXiv:1702.08734.

Jihyun Lee and Gary Lee. 2025. The limits of post-hoc preference adaptation: A case study on DSTC12 clustering. In *Proceedings of the Twelfth Dialog System Technology Challenge*, pages 36–43, Avignon, France. Association for Computational Linguistics.

Junhua Liu, Yong Keat Tan, Bin Fu, and Kwan Hui Lim. 2024. Lara: Linguistic-adaptive retrieval-augmentation for multi-turn intent classification. *Preprint*, arXiv:2403.16504.

Pengfei Liu, Youzhang Ning, King Keung Wu, Kun Li, and Helen Meng. 2021. Open intent discovery through unsupervised semantic clustering and dependency parsing. *Preprint*, arXiv:2104.12114.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction. *Preprint*, arXiv:1802.03426.

A Padmasundari and Srinivas Bangalore. 2018. Intent discovery through unsupervised semantic text clustering. In *Interspeech 2018*, pages 606–610.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

SBERT Team. 2023. Sentencetransformers utility functions documentation. `https://sbert.net/docs/package_reference/util.html`. Accessed July 8, 2025.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing Management*, 45(4):427–437.

V. A. Traag, L. Waltman, and N. J. van Eck. 2019. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233.

Mengting Wan, Tara Safavi, Sujay Kumar Jauhar, Yujin Kim, Scott Counts, Jennifer Neville, Siddharth Suri, Chirag Shah, Ryen W White, Longqi Yang, Reid Andersen, Georg Buscher, Dhruv Joshi, and Nagu Rangan. 2024. Tnt-llm: Text mining at scale with large language models. *Preprint*, arXiv:2403.12173.

Qingkai Zeng, Yuyang Bai, Zhaoxuan Tan, Shangbin Feng, Zhenwen Liang, Zhihan Zhang, and Meng Jiang. 2024. Chain-of-layer: Iteratively prompting large language models for taxonomy induction from limited examples. *Preprint*, arXiv:2402.07386.

Hanlei Zhang, Hua Xu, Xin Wang, Fei Long, and Kai Gao. 2024. A clustering framework for unsupervised and semi-supervised new intent discovery. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):5468–5481.

Kun Zhu, Lizi Liao, Yuxuan Gu, Lei Huang, Xiaocheng Feng, and Bing Qin. 2025. Context-aware hierarchical taxonomy generation for scientific papers via LLM-guided multi-aspect clustering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15616–15634, Suzhou, China. Association for Computational Linguistics.

Appendix

A.1 Additional Experiments:

Experiment 1. Clustering Approaches. We tested other clustering approaches with the aim of achieving Weighted Average Intra-Cluster Cosine Similarity (WAICCS) higher than HDBSCAN (refer to Table 3).

Experiment 2. Taxonomy Drift Simulation. To evaluate resilience against concept drift, we conducted three simulations: (1) Introduction of new intents using LLM-generated utterances on emerging topics (e.g., Cryptocurrency), (2) taxonomy evolution through label splits and merges, and (3) paraphrase drift using LLM-based rephrasing. Performance degradation was measured using UNKNOWN assignment rates and OOD flagging accuracy. The system consistently detected new or modified intents at the L_2, with the majority of mis-assignments occurring in low-separability clusters, highlighting the need for well-separated intent boundaries. This exercise also helped with impact analysis of drift evolution, revealing structural weak points in the framework.

A.2 Sample Chat Transcript:

Chatbot: Hi there, welcome to Bank chat support. How can I help you today?

Approach	WAICCS(50)
HDBSCAN (McInnes et al., 2017)	0.905
Agglomerative Clustering (Beauchamp et al., 2023)	0.685
Spherical K-Means (Hornik et al., 2012)	0.420
ST Comm. Det. (SBERT Team, 2023)	0.367
OPTICS (Ankerst et al., 1999)	0.235
Leiden Clustering (Traag et al., 2019)	0.144

Table 3: WAICCS for top 50 clusters across various clustering methods.

Customer: Hi, I need to update my address for my savings account.

Chatbot: Sure, I can help with that. For security, please confirm your date of birth.

Customer: 14th March 1985.

Chatbot: Thank you. Now, please choose one of the following options to proceed:
1. Update residential address
2. Update mailing address
3. Update both residential and mailing address

Customer: Select option 1

Chatbot: Got it. Please type in your new residential address.

Customer: 22 Greenfield Lane, London, W1A 4ZZ.

Chatbot: Thank you. Your residential address will be updated within 24 hours. Would you like us to send a confirmation letter to your new address? (Yes/No)

Customer: Yes

Chatbot: Perfect. A confirmation will reach you within 5 to 7 business days. Is there anything else I can help you with today?

Customer: No, that's all. Thanks!

Chatbot: You're welcome! Have a great day!

A.3 Sample Voice Transcript:

Agent: Good afternoon, thank you for calling Bank customer support. This is Sarah speaking, how may I assist you today?

Customer: Uh hi, yeah, umm, I recently moved and I need to update my address, uh, for my account.

Agent: Of course, I can help with that. May I have your full name and date of birth for verification, please?

Customer: Sure, it's John Smith, and my date of birth is 14th of March, 1985.

Agent: Thank you, Mr. Smith. And which account would you like to update-savings, current, or both?

Customer: Just my savings account, please.

Agent: Alright. Could you give me your new address, including the postcode?

Customer: Yeah, it's, umm, 22 Greenfield Lane, London... W1A 4ZZ.

Agent: Got it. That's 22 Greenfield Lane, London, W1A 4ZZ-correct?

Customer: Yes, that's right.

Agent: Perfect. I've submitted the request. The change will take effect within 24 hours, and we'll also send a confirmation letter to your new address.

Customer: Great, thank you so much.

Agent: You're very welcome! Is there anything else I can assist you with today?

Customer: No, that's it. Appreciate your help.

Agent: My pleasure. Thanks for calling Bank, and have a wonderful day ahead!

Customer: You too, bye.

A.4 Production Case Examples:

We document four anonymized production incidents where the deployed RoC detection system successfully flagged or rerouted risky or ambiguous cases, thereby preventing potential escalation errors. Each case was surfaced through the model's out-of-distribution (OOD) alerting mechanism or through improved L2 granularity matching, leading to proactive resolution by contact center supervisors. Here, the prior model refers to the prior Entity-Action pair-based supervised system.

Case 1: Digital Currency-related Inquiry Misassigned as General Transaction A customer asked: *"How can I store bitcoins safely in my*

wallet?" - a previously unseen topic. The legacy system misrouted this to "Recent Transactions," while our system flagged it as OOD (similarity below τ_{OOD}), prompting manual triage. This led to a discussion over taxonomy extension to include emerging digital currency topics.

Case 2: Paraphrased Security Alert Routed Incorrectly The utterance *"My account got blocked right after I clicked a strange link-need help!"* was originally being routed to *Card PIN Reset*. Our model matched it to *Suspicious Activity* at L_2 with high confidence, enabling fraud triage. A similar error in the prior model had previously caused a delay in incident escalation.

Case 3: Address Update Mis-routed Due to Label Ambiguity A customer stated: *"I need to change where you send my monthly credit card bills."* The prior model mapped this to *Update Contact Details*, but our updated taxonomy with split L_2 nodes (*Update Address-Account* vs. *Update Address - Card*) correctly resolved it to the card-specific path, reducing inter-queue transfers.

Case 4: Incomplete Query Prevented from Escalation An utterance without context (e.g.,*Need to fix this*) was routed by the prior model to "Service Complaint" which triggered an unnecessary supervisor review. Our approach tagged it correctly by considering nearby utterance in the Overall RoC detection and mapped it correctly to the *Update Address* intent.

These examples illustrate how our system's hierarchical taxonomy, OOD detection, and fast retrieval have improved routing precision and reduced operator burden in ambiguous or emerging scenarios.

A.5 Other Formulae, Tables and Figures:

(A) Macro- and Micro-F1. We report performance using both macro-F1 and micro-F1 metrics (Sokolova and Lapalme, 2009). Let C be the set of classes, and for each class $c \in C$, let Precision_c and Recall_c denote the per-class precision and recall. The macro-F1 score is the unweighted mean of the per-class F1-scores:

$$\text{Macro-F1} = \frac{1}{|C|} \sum_{c \in C} \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}.$$

The micro-F1 score is computed globally by aggregating counts of true positives (TP), false positives (FP), and false negatives (FN) over all classes:

$$\text{Micro-F1} = \frac{2 \cdot \text{TP}_\mu \cdot \text{Precision}_\mu \cdot \text{Recall}_\mu}{\text{Precision}_\mu + \text{Recall}_\mu}$$

$$= \frac{2 \cdot \text{TP}_\mu}{2 \cdot \text{TP}_\mu + \text{FP}_\mu + \text{FN}_\mu}.$$

Here, $\text{Precision}_\mu = \frac{\text{TP}_\mu}{\text{TP}_\mu + \text{FP}_\mu}$ and $\text{Recall}_\mu = \frac{\text{TP}_\mu}{\text{TP}_\mu + \text{FN}_\mu}$.

Macro-F1 treats all classes equally, thereby highlighting performance on minority classes, whereas micro-F1 treats each observation equally, thus favoring performance on majority classes.

(B) ICSS and WAICCS Intra-cluster cosine similarity (ICCS) is calculated as :

$$\bar{s}_k = \frac{1}{n_k(n_k - 1)} \sum_{\substack{i,j \in C_k \\ i \neq j}} \text{sim}(x_i, x_j) \qquad (1)$$

where, n_k is the number of points in cluster C_k, and $\text{sim}(x_i, x_j)$ is the cosine similarity between embeddings x_i and x_j The weighted average over the top K clusters WAICCS(K) is computed as:

$$\bar{S}_{\text{weighted}} = \frac{1}{N} \sum_{k=1}^{K} n_k \cdot \bar{s}_k \qquad (2)$$

where, $N = \sum_{k=1}^{K} n_k$ is the total number of points across the top K clusters, and $\bar{s}_k$ is derived from Equation (1).

L_0	L_0 **Description**	L_1	L_1 **Description**	L_2	L_2 **Description**
Transactions	Managing payments, transfers, and transaction-related issues.	Refunds and Cancellations	Handling refunds and cancellations of payments.	Request Refund	Process for initiating a refund for a transaction, including required documentation and timeframes.
Statements	Managing account statements and documents.	Statement Management	Handling account statements.	Generate Statement	Assisting with generating or accessing account statements for specific periods, formats, or delivery methods.
Cards	Managing debit and credit cards, including issues and security.	Credit Card limits	Providing information about credit card limits	Increase credit card limit	Procedures to request an increase in existing credit card limits, including eligibility criteria and approval times.
Security	Managing security and preventing fraud.	Fraud Reporting	Reporting fraudulent activities and potential scams.	Report Fraud	Assisting customers in reporting unauthorized transactions or suspicious account activities.
Loans	Handling loans and mortgage-related services.	Loan Services	Applying for and managing personal loans.	Early Loan Repayment Details	Information on early repayment options, potential penalties, and how to calculate outstanding balances.

Table 4: Sample rows from the hierarchical RoC taxonomy.

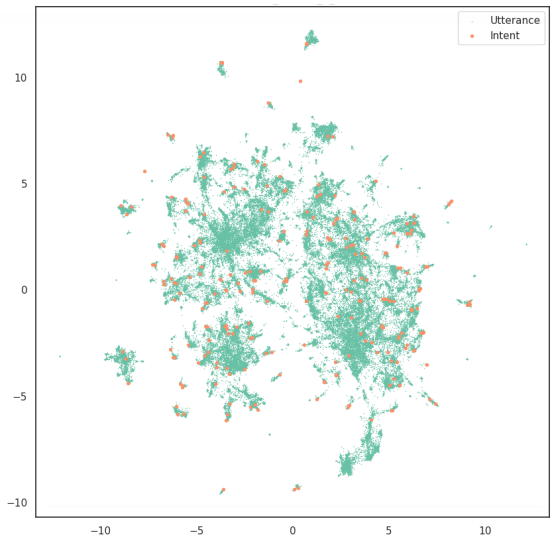

Figure 3: Plot showing a 2-dimensional t-SNE projection of utterance and intents embeddings, with utterances shown in green and intents in orange.

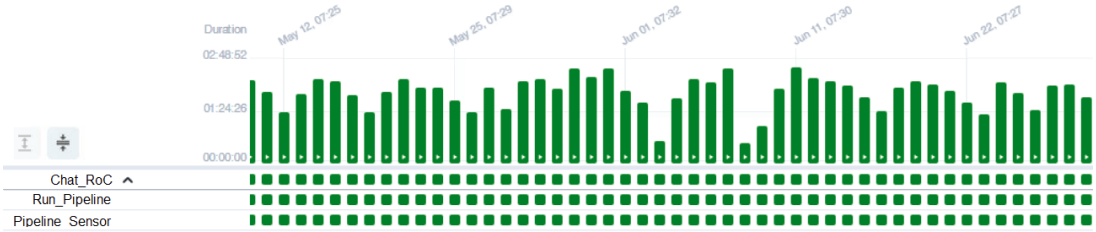

Figure 4: Cumulative run-time plot for chat channel with 40,000 average conversations per day.

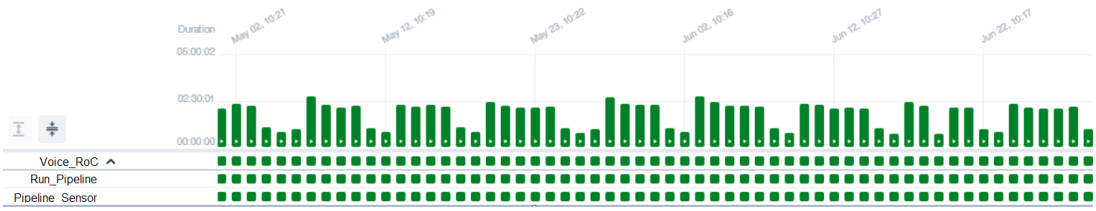

Figure 5: Cumulative run-time plot for voice channel with 35,000 average conversations per day.

Parameter	Value	Use Case	Empirical Procedure	Observed Trade-offs	Sensitivity Analysis
N_{chat}, N_{voice}	3, 10	1) First RoC – optimal number of initial turns, and 2) Overall RoC – sliding window size	Annotated 100 transcripts for each channel; varied window size and measured earliest match to ground-truth. Selected value with $\geq$90% match within selected turns.	Larger windows resulted in more confusion and lower match scores, while smaller gave more wrong results; selected value gave the best balance. Domain experts agreed with this.	Yes
τ_{OOD}	0.3	Threshold for out-of-domain (OOD) detection.	Collected manually labelled utterances; computed cosine similarity to known taxonomy nodes. Threshold chosen at acceptable error rate.	Lower values increased false positives; 0.3 offered best segregation.	Planned for future
τ_{unk}	10%	Threshold for UNKNOWN utterance filter for Overall RoC.	Requirement predominantly set by users to trigger alerts for any topic surges.	No significant trade-offs observed. Periodic reviews planned should be able to handle any gaps.	Planned for future
τ_{merge}	0.5	Threshold for dynamic taxonomy maintenance.	Monitored semantic drift in new utterances across 4 weeks; tested merge thresholds on cosine similarity. Validated merged nodes via manual annotation for semantic coherence. New clusters below this might be new intents.	Lower threshold caused poor merges and reduction in cluster WAICCS.	Yes

Table 5: Empirically set parameters with selection methodology and trade-offs.