ICFD-31k: A Large-Scale Dataset and Benchmark for Real-Time Conversational Fraud Detection

Rishi Ahuja[1] , **Kumar Prateek**[1] and **Simranjit Singh**[1*]

[1]Department of Information Technology, Dr. B.R. Ambedkar National Institute of Technology Jalandhar, Punjab, 144008, India

{rishia.it.24, kumarprateek, singhsimranjit}@nitj.ac.in

Abstract

The proliferation of sophisticated telephone scams poses a significant societal and economic threat, impacting diverse linguistic contexts in a country like India. Furthermore, the lack of large-scale, publicly available datasets remains a critical barrier impacting research on robust, real-time countermeasures. In view of this, the proposed work introduces ICFD-31k, the first Indian Conversational Fraud Dataset, representing a new benchmark containing over 31,000 realistic conversational transcripts. ICFD-31k comprises systematically generated content, covering 10 distinct fraud umbrellas spanning from financial impersonation to job scams. ICFD-31k transcripts feature rich annotations comprising a final verdict, chunk-level streaming labels, and detailed "slow-thinking" rationales. In addition, the human-in-the-loop evaluation supports the ICFD-31k's quality, achieving a mean pairwise Cohen's Kappa of 0.534, indicating moderate inter-annotator agreement. Furthermore, the proposed work introduces two fine-tuned models based on RoBERTa: M1 for non-streaming data and M2 for streaming data. The comprehensive experiments with strong baselines (M1, M2) further demonstrate the ICFD-31k's utility. The code and reproducibility materials are available at https://github.com/SPELLAILab/ICFD-31k.

1 Introduction

AI plays a crucial role in advancing social good by safeguarding marginalized communities from exploitation. Conversational fraud uses phone calls as a primary medium, resulting in financial loss and emotional harm to millions of people. India, home to over a billion telecom users, has advanced its digital payment ecosystem. However, the payment ecosystem is threatened by evolving social engineering techniques utilized by conversational fraud. Moreover, these attacks harness urgency and trust to disproportionately affect vulnerable groups such as the elderly, populations with limited digital literacy, and individuals in linguistically diverse or low-resource environments. Also, it decreases public trust and economic progress for digital financial inclusion. Traditional AI-driven detection systems can support fraud prevention, enabling inclusive digital ecosystems. Nevertheless, such systems are hampered by the non-availability of a large-scale public dataset for conversational fraud. Existing fraud detection research has made considerable progress using graph-based and transactional approaches [Bai *et al.*, 2025; Wu *et al.*, 2024; Zou and Cheng, 2024]. However, these approaches do not address bilingual conversational fraud detection in Indian contexts. Moreover, they primarily emphasize detection performance rather than granular, conversation-level explanations. Additionally, the non-availability of annotated datasets with granular and explainable labels denies actionable insights for real-time streaming call analysis (during ongoing calls) and transparent "slow-thinking" rationales. Therefore, such scarcity of resources limits not only rigorous benchmarking and comparative evaluation of detection approaches but also the deployment of practical solutions in resource-constrained settings where protection is most urgently needed, perpetuating a cycle where those most vulnerable to exploitation have the least access to protective technologies. Hence, there is a strong need for comprehensive, multilingual conversational fraud benchmark datasets with rich annotations that enable both real-time detection and explainable decision-making in diverse, resource-constrained environments.

The main contributions of the proposed work are threefold:

- The work introduces ICFD-31k, the first large-scale Indian conversational fraud benchmark designed for bilingual (English/Hinglish) contexts, directly addressing SDG 10 (Reduced Inequalities) by enabling protective technologies for vulnerable, linguistically diverse populations disproportionately targeted by telephone scams.
- The proposed work establishes baselines with fine-tuned RoBERTa models (M1 for static, M2 for streaming), enabling chunk-level real-time fraud identification supported by rationale annotations and providing a lightweight baseline for future low-resource deployment studies.
- The study develops a complete methodology using LLMs and efficient architectures designed with low-resource and mobile deployment constraints in mind,

*Corresponding Author.

balancing accuracy, interpretability, and computational cost.

2 Related Work

This section reviews key literature on fraud detection, focusing on datasets, synthetic data generation, conversational AI applications, real-time/streaming techniques, and explainable AI approaches in security contexts.

2.1 Fraud Detection Datasets

The scarcity of publicly available conversational datasets has hindered the development of robust fraud detection systems. Furthermore, existing work predominantly relies on proprietary data from financial institutions or telecommunications providers, limiting reproducibility and cross-study comparison. The most closely related prior work is TeleAntiFraud-28k [Ma *et al.*, 2025], a recent dataset of 28,000 Chinese telecom fraud conversations with audio and text modalities. TeleAntiFraud-28k provides slow-thinking rationales for explainability. However, it is limited to the monolingual Chinese context, lacks chunk-level temporal annotations essential for real-time streaming detection, and fails to capture code-switching (Hinglish) or region-specific social engineering tactics prevalent in the Global South, such as in India. Moreover, real Indian telecom fraud calls are difficult to redistribute because recordings may contain sensitive personal data and raise consent and data-protection concerns, therefore necessitating an approach that can address bilingual conversational fraud detection in Indian contexts with granular, conversation-level explanations. Besides, other fraud-related datasets focus on different modalities. For instance, financial transaction datasets [Dal Pozzolo *et al.*, 2014] provide tabular format data but lack the conversational context required. Similarly, email phishing datasets [Fette *et al.*, 2007] and SMS spam corpora [Almeida *et al.*, 2011] operate on written text rather than conversational interactions, therefore necessitating a resource that can jointly address real-time conversational fraud detection and explainability requirements.

2.2 Conversational AI for Security

Recent advancements in large language models have demonstrated strong performance on various NLP tasks, including zero-shot and few-shot learning scenarios [Brown *et al.*, 2020; Chung *et al.*, 2024]. However, their application to security domains like fraud detection remains underexplored. The zero-shot baseline with Gemma 1.1 2B [Gemma Team *et al.*, 2024] is used as a compact, on-device reference point rather than a proxy for all frontier LLMs. Its 35.83% F1 on ICFD-31k shows that this lightweight prompt-only baseline is insufficient for reliable fraud screening without domain adaptation, while evaluating larger cloud LLMs remains complementary future work. Additionally, the literature includes related work on dialogue safety that focuses primarily on toxicity detection [Zhang *et al.*, 2018], hate speech identification [Davidson *et al.*, 2017], and misinformation in social media [Shu *et al.*, 2017]. Those tasks, while related, differ fundamentally from fraud detection, which requires recognizing subtle social-engineering tactics and policy violations, and reasoning about the temporal progression of deceptive behavior within a conversation.

2.3 Synthetic Data Generation

The use of LLMs for synthetic data generation has gained traction across various NLP applications [Puri *et al.*, 2020; Schick and Schütze, 2021]. Recent work [Ye *et al.*, 2022] shows that carefully controlled synthetic data can augment or even replace human-annotated datasets for certain tasks. However, concerns about distribution shift, hallucination, and lack of authentic complexity remain significant challenges. Therefore, ICFD-31k adopts a generation approach different from prior synthetic data efforts in three key aspects. First, a structured generation pipeline is employed with explicit persona and scenario combinations to ensure even coverage and diversity. Second, ICFD-31k quality is evaluated through rigorous human evaluation with a mean pairwise Cohen's Kappa of 0.534, indicating moderate inter-annotator agreement, rather than only relying on automatic metrics. Third, utility is demonstrated through comprehensive benchmarking, showing that models trained on ICFD-31k achieve strong performance on realistic fraud detection tasks.

2.4 Streaming and Real-Time Detection

Most fraud detection systems operate in batch mode, analyzing complete transactions or conversations post-hoc. However, real-world deployments require real-time detection to enable immediate intervention against fraud. In the literature, prior work on streaming classification has focused primarily on concept drift adaptation [Gama *et al.*, 2014] and online learning [Hoi *et al.*, 2021], rather than the temporal progression of fraud signals within individual conversations. Consequently, ICFD-31k employs a chunk-level annotation schema that enables research into early detection systems that must balance speed (detecting fraud quickly) against accuracy (avoiding false alarms with incomplete information).

2.5 Explainable AI for Security

The need for explainable AI in security applications has been widely recognized [Barredo Arrieta *et al.*, 2020], particularly for high-stakes decisions involving financial loss or legal consequences. Recently, several authors explored various approaches, including attention visualization [Wiegreffe and Pinter, 2019], rationale extraction [Lei *et al.*, 2016], and chain-of-thought reasoning [Wei *et al.*, 2022]. Therefore, ICFD-31k integrates explicit reasoning rationales. Each conversation in the proposed ICFD-31k contains reasoning chains detailing specific policy violations and conversational evidence using slow-thinking rationales [Ma *et al.*, 2025] presented in TeleAntiFraud-28k.

3 The ICFD-31k Dataset

The proposed work uses a novel generational pipeline to produce ICFD-31k that addresses data limitations and ensures contextual applicability for the Indian fraud scenarios.

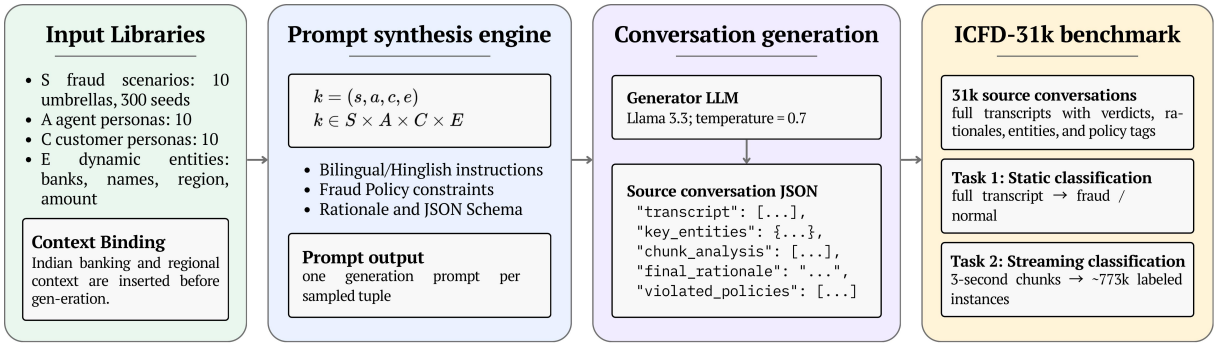

Figure 1: The systematic data generation pipeline for ICFD-31k. Foundational components (left) are combined via the Prompt Synthesis Engine ($k \in S \times A \times C \times E$) to produce diverse prompts for the Llama 3.3 70B.

3.1 Design Principles

The study covers 10 fraud categories ("umbrellas") such as bank & payment fraud, E-commerce scams, job scams, government impersonation, and many more, with each umbrella containing 30 scenarios with varying difficulty.

3.2 Data Generation Pipeline

The generation of ICFD-31k followed a systematic, two-stage pipeline as in Figure 1. First, it employs foundational components: 10 agent and customer personas, over 300 scenarios across 10 fraud umbrellas (each with description, case type, and policy rules), and extensive entity lists including Indian banks, e-commerce platforms, and region-specific names. Thereafter, the pipeline uses the cross-product of personas and scenarios, yielding approximately 31,000 structured JSON conversations using an LLM-based generation system. Finally, in the second stage, each conversation expands into temporal chunks at 3-second intervals, producing 1,111,071 training samples for real-time streaming models.

3.3 Annotation Schema

Each[1] conversation is stored as structured JSON (Listing 1) containing timestamped transcripts with speaker turns. The schema includes two specific innovations for training next-generation detection systems as discussed below:

Chunk-level streaming labels. The schema provides verdict and rationale annotations at approximately 3-second intervals throughout each conversation, enabling training and evaluation of real-time detection systems that must make progressive decisions with incomplete information.

Slow-thinking rationales. Each conversation includes a detailed reasoning chain explaining the final verdict [Ma *et al.*, 2025], citing specific policy violations and conversational evidence. Furthermore, it facilitates the development of explainable AI systems that can provide transparent justifications. Besides, the schema also captures entity extraction (`pii_requested`, organizations, products), violated policies, scam outcomes, and the generative context (scenario, personas) for comprehensive analysis.

[1]Large language models were used only to assist with grammar refinement and manuscript editing. All scientific content, experimental results, and conclusions are the sole work of the authors.

Listing 1. Complete annotation schema for a single conversation in ICFD-31k.

```
{
  "transcript": [
    {"speaker": "Agent", "text": "Namaste...", "
     timestamp_end": 5.0},
    {"speaker": "Customer", "text": "Achha...", "
     timestamp_end": 10.0}
  ],
  "multimodal_analysis": {
    "dominant_emotion": "Trust", "pace": "Fast",
    "confidence_score": 0.8
  },
  "key_entities": {
    "organization": ["Canara Bank"], "product": ["Card"],
    "pii_requested": ["Card number", "Expiry", "CVV"]
  },
  "chunk_level_analysis": [
    {"timestamp": 10.0, "verdict_at_chunk": "NO",
     "rationale_at_chunk": "Introduction"},
    {"timestamp": 25.0, "verdict_at_chunk": "YES",
     "rationale_at_chunk": "Card number requested (Rule A)
     "}
  ],
  "final_slow_thinking_rationale": "Agent requested full
    card...",
  "final_verdict": "YES",
  "violated_policies": ["Rule A: no full card requests"],
  "scam_outcome": "Successful",
  "session_id": 213,
  "agent_persona": "Empathetic helper persona",
  "customer_persona": "Busy professional persona",
  "scenario": {
    "scenario_id": 3, "description": "Debit card CVV scam"
    ,
    "case_type": "Clear Fraud", "policy": "Rule A: no card
     data"
  }
}
```

3.4 Dataset Statistics

The dataset ICFD-31k expands 31,000 source conversations into 1,111,071 temporal chunks. Specifically, the 3-second interval was selected to approximate short conversational turns and support research into early intervention. Further, in order to control a fraud-heavy training expansion, all non-Clear-Normal cases are fully chunked while Clear Normal calls are sparsely sampled (1–2 chunks). It may be noted that the asymmetric construction is explicit, should not be read as deployment prevalence, and consequently motivates future interval/sampling ablations. Moreover, the ICFD-31k dataset

Metric	Value
Total Source Conversations	31,000
- Main Set	30,000
- Cross-Domain Test Set	1,000
Total Training Chunks	∼1,111,071
Fraud Umbrellas	10 + 5
Avg. Conversation Length	115.8 seconds
Class Distribution (Fraud)	∼70%
Class Distribution (Normal)	∼30%

Table 1: Overall statistics of the ICFD-31k dataset.

Fraud Umbrella	Conversation Count
Bank & Payment Fraud	∼3,000
E-commerce & Marketplace	∼3,000
Utility Bill & Service	∼3,000
Job & Employment	∼3,000
Loan Shark & Harassment	∼3,000
Tech Support & Account	∼3,000
Personal Emergency & Extortion	∼3,000
Insurance & Healthcare	∼3,000
Travel & Lottery	∼3,000
Government Impersonation	∼3,000

Table 2: Distribution of source conversations across the 10 primary fraud umbrellas.

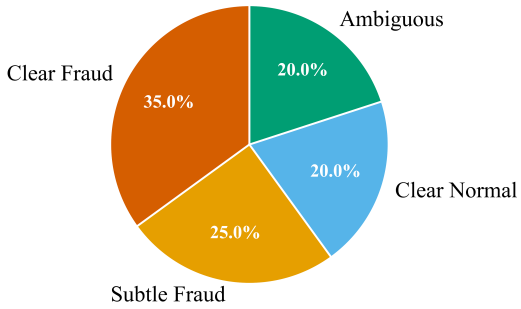

Figure 2: Distribution of case types (Clear Fraud, Clear Normal, Ambiguous, and Subtle Fraud) across the dataset.

Case Type	Mean PPL	Median PPL	Std. Dev.
Normal	20.44	19.47	6.10
Fraud	22.37	20.80	8.41

Table 3: Perplexity scores across conversation types, demonstrating high linguistic fluency.

is structured across 10 primary fraud umbrellas to ensure a diverse and balanced benchmark. Specifically, each umbrella contributes approximately 3,000 source conversations, ensuring no single fraud typology dominates the ICFD-31k dataset as shown in Table 2. Furthermore, the main training set consists of these 10 umbrellas, while an additional 1,000 conversations across 5 completely unseen domains (e.g., Cryptocurrency, Romance scams) were generated to create a dedicated cross-domain test set for evaluating model generalization. In addition, within each umbrella, the distribution of scenario complexity is maintained according to the adopted design principles as illustrated in Figure 2, providing a rich mix of clear, ambiguous, and subtle cases. Besides, Table 1 summarises the overall statistics for ICFD-31k.

4 Quantitative Dataset Analysis

The study evaluates four corpus properties to test whether ICFD-31k is fluent, diverse, and not merely templated: (1) linguistic fluency, (2) structural novelty, (3) semantic complexity, and (4) vocabulary distribution. Subsequently, these diagnostics directly target synthetic-artifact risks and thereby complement, rather than replace, real-call validation.

4.1 Linguistic Fluency (Perplexity)

A core requirement for a synthetic dataset is that it is fluent and natural-sounding. Hence, ICFD-31k measures fluency using Perplexity (PPL) [Jelinek *et al.*, 1977], the standard metric for evaluating language model naturalness, where a lower score indicates better performance. Specifically, it calculates the PPL of conversations using a pre-trained GPT2 model. The ICFD-31k dataset achieves a low mean PPL of 20.44 for normal conversations and 22.37 for fraud conversations as described in Table 3. The obtained scores provide model-based evidence that the Llama 3.3 70B [Meta, 2024] generator produced linguistically fluent text, while human evaluation provides complementary evidence regarding perceived realism and naturalness.

4.2 Structural Novelty (N-gram Overlap)

A major risk in synthetic generation is "templating," where the model produces conversations that are structurally identical but with different entities. Therefore, the proposed work measures the N-gram Overlap between 10,000 random pairs of conversations for validation. The result shows a vanishingly small overlap, providing evidence that ICFD-31k is not simply "copy-pasting" templates as summarized in Table 4. Additionally, the obtained average 3-gram overlap is only 1.96%, suggesting that the used generation pipeline, with its 300 unique scenarios and 20 persona combinations, produced structurally varied conversations.

4.3 Semantic Complexity (t-SNE & Similarity)

The study designs a robust benchmark for fraud detection that deliberately emphasizes difficulty. Consequently, fraudulent messages display subtle linguistic patterns that overlap substantially with legitimate conversations. Such semantic complexity was evaluated using two distinct methods. First, the proposed work computes the Semantic Similarity (cosine similarity of `all-MiniLM-L6-v2` embeddings)

N-gram	Mean Overlap (%)	Std. Dev.
1-gram	35.92	6.31
2-gram	7.29	3.52
3-gram	1.96	1.86

Table 4: N-gram overlap analysis demonstrating structural novelty across conversations.

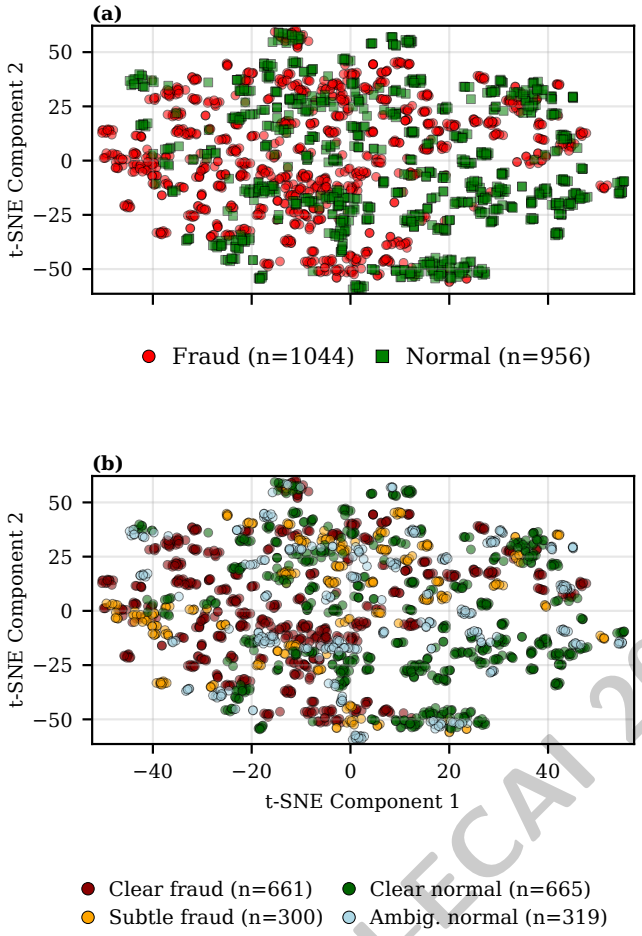

Figure 3: t-SNE visualization of semantic embeddings. Top: Binary fraud vs. normal classes show significant overlap. Bottom: Four case types reveal the semantic complexity.

Metric	Mean Similarity	Std. Dev.
Intra-Class (Fraud)	0.421	0.108
Intra-Class (Normal)	0.381	0.113
Inter-Class (Fraud-Normal)	0.383	0.110
Separation Score	**0.017**	—

Table 5: Semantic similarity scores showing high class overlap, indicating dataset complexity.

within classes (e.g., fraud-to-fraud) and *between* classes (e.g., fraud-to-normal). Also, the obtained result, as shown in Table 5, indicates that the dataset contains a low separation score of 0.017. It suggests that fraudulent and normal conversations are linguistically close, making the task difficult for models that rely only on shallow lexical cues. Additionally, the study visualizes such complexity in Figure 3 using a t-SNE [van der Maaten and Hinton, 2008] plot. The "Fraud vs Normal" panel (top) shows substantial class overlap. The "By Case Type" panel (bottom) is even more revealing: the "Subtle Fraud" (orange) and "Ambig. normal" (light blue) cases are semantically mixed in the "battleground" area between the clear cases. The results provide quantitative evidence that ICFD-31k constitutes a difficult benchmark, demanding models to perform deeper contextual reasoning beyond keyword matching.

Metric	Score
Distinct-1	0.0653
Distinct-2	0.3371
Distinct-3	0.6113

Table 6: Linguistic diversity metrics on a 10k sample, showing controlled vocabulary for domain fidelity.

4.4 Vocabulary Distribution (Distinct-n & Zipf's Law)

The study analyzes the ICFD-31k's vocabulary. Specifically, the Linguistic Diversity (Distinct-n) scores described in Table 6 (e.g., Distinct-1: 0.0653, Distinct-2: 0.3371, Distinct-3: 0.6113) indicate a trade-off between lexical diversity and repeated domain-specific terminology. In this regard, the controlled temperature (T=0.7) may contribute to the frequent occurrence of terms such as 'OTP', 'bank', and 'CVV'. Furthermore, the vocabulary distribution is further examined using Zipf's Law [Zipf, 1949], as presented in Figure 4. Accordingly, the MSE of 1.81 and the curve's deviation from the ideal line suggest a vocabulary distribution shaped by recurring fraud-domain terminology. Collectively, these findings should be interpreted alongside the structural-diversity and human-evaluation results.

5 Experimental Evaluation

The performance of ICFD-31k is evaluated through benchmarking with a strong baseline. In particular, the experimental design considers two questions: 1) How effectively does the model classify in-domain conversations? 2) How well does the model generalize to novel scam types? The results reveal that the fine-tuned classifier obtains near-perfect in-domain performance, while performance decreases on unseen scam types.

5.1 Experimental Setup

The developed models are evaluated on three tasks to test both real-world deployment and offline analysis capabilities.

Task 1: Static Classification. Given a complete conversation transcript, predict whether the interaction is fraudulent. This represents the best-case scenario where all information is available for batch processing.

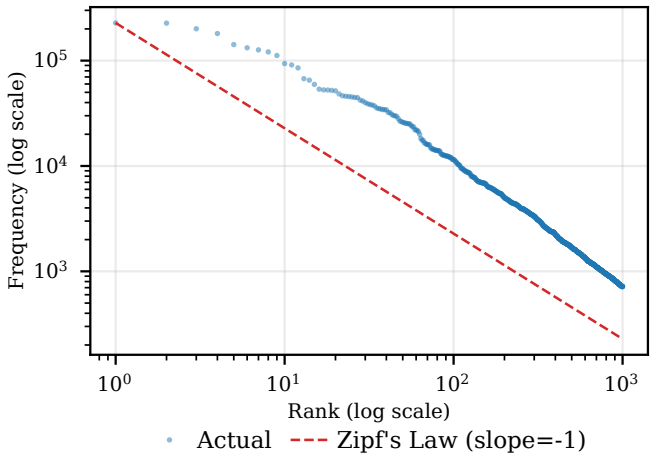

Figure 4: Zipf's Law analysis showing deviation from the ideal power law distribution (MSE: 1.81).

Model & Task	F1	Prec. / Rec.
Gemma 2B (Zero-shot)	35.83	93.58 / 22.16
RoBERTa (In-domain)	**99.40**	99.22 / 99.59
RoBERTa (Cross-domain)	92.97	98.49 / 88.03
RoBERTa (Streaming)	95.59	95.50 / 95.67

Table 7: Performance comparison across baseline models and evaluation tasks.

Stage	Timestamp	F1	Accuracy	Samples
Early	0-30s	95.78	98.81	23,833
Mid	30-60s	95.57	96.34	33,442
Late	60-120s	94.57	92.98	46,832
Very Late	120s+	97.12	95.46	25,009

Table 8: Streaming classification performance across conversation timeline.

Task 2: Streaming Classification. Given conversation chunks at 3-second intervals, predict fraud status at each timestamp, simulating real-time detection where decisions must be made with incomplete information.

Task 3: Cross-Domain Generalization. Evaluation on completely unseen fraud types to test the model's ability to adapt to novel scam typologies that emerge after training.

ICFD-31k is partitioned as follows: Training set contains 21,000 conversations (70%) from the main 10 fraud umbrellas; Validation set contains 4,500 conversations (15%); In-domain test set contains 4,500 conversations (15%) from the same 10 umbrellas; Cross-domain test set contains 1,000 conversations from 5 completely unseen fraud types (cryptocurrency scams, romance scams, investment fraud, charity scams, and tax impersonation). The proposed work ensures that for streaming experiments, all chunks inherit the split of their source conversation. Consequently, it yields 1,111,071 temporal chunks while preventing chunks from the same session from appearing in both training and test sets. Furthermore, the obtained F1-score, along with Precision and Recall are reported in Table 7. Additionally, Table 8 reports the measurement report for early detection rate (percentage of fraud cases detected within the first 30 seconds) and mean detection latency (average time to first fraud prediction) for streaming classification.

5.2 Baseline Models

The proposed work establishes three benchmarks: a compact zero-shot LLM reference (Gemma 1.1 2B), Fine-Tuned RoBERTa (static), and streaming RoBERTa. Specifically, Gemma is prompted with the full transcript and policy rules to measure prompt-only performance under an on-device model scale. Subsequently, RoBERTa-base [Liu *et al.*, 2019] is fine-tuned on full transcripts for static classification and on cumulative 3-second chunks for streaming classification. In this regard, the RoBERTa models use a 512-token limit, a binary classification head, 5 epochs with early stopping, learning rate 2e-5, and effective batch size 16, thereby prioritizing a lightweight baseline for future low-resource deployment studies over architectural novelty.

5.3 Main Results

Table 7 presents the core performance comparison across all baseline models and evaluation scenarios. Specifically, Gemma 2B achieves 35.83% F1 with high precision but low recall (22.16%), showing that this compact zero-shot baseline misses most fraudulent cases under the prompt-only setting and is unsuitable where false negatives carry high societal cost. In contrast, the fine-tuned RoBERTa achieves 99.40% F1 with near-perfect precision and recall in-domain; five-fold cross-validation yields 99.51% $\pm$ 0.07% F1. Accordingly, we interpret the in-domain result as an upper-bound synthetic benchmark score, not evidence that real-world fraud detection is solved. Furthermore, the streaming RoBERTa achieves 95.59% F1, a 3.81% decrease, quantifying the cost of incomplete information. Notably, the model detects 99.82% of fraud cases with an average latency of 32.2 seconds. In addition, Table 8 shows performance varies across timeline: early chunks (0–30s) achieve 95.78% F1, while very late chunks (120s+) reach 97.12% F1 as context accumulates.

5.4 Analysis of the Generalization Gap

The 99.40% F1-score on in-domain data appears near-perfect, but it should be read together with the cross-domain stress test. Specifically, when the same RoBERTa model is evaluated on the cross-domain test set containing 1,000 conversations from 5 completely unseen fraud types (cryptocurrency scams, romance scams, investment fraud, charity scams, and tax impersonation), the performance degrades by 6.43% from 99.40% to 92.97% F1 as discussed in Table 9. Accordingly, the drop suggests a genuine shift in fraud language and tactics rather than a solved artifact-matching task, thereby motivating continual benchmark extension as scams evolve. In this regard, static classifiers trained on historical data may degrade as fraud tactics evolve, consequently making historical accuracy an incomplete predictor of future utility.

Fraud Type	F1	Accuracy	Recall
Charity Scams	96.73	93.75	94.87
Cryptocurrency	92.46	86.36	86.79
Investment Fraud	91.67	85.00	84.62
Romance Scams	91.76	85.00	87.43
Tax Impersonation	93.01	87.50	86.93
Overall	**92.97**	**87.20**	**88.03**
In-Domain (Reference)	*99.40*	*99.38*	*99.59*
Generalization Gap	**-6.43**	**-12.18**	**-11.56**

Table 9: Cross-domain generalization results on 5 unseen fraud types.

5.5 Expert Validation and Stakeholder Analysis

The study employs a multidisciplinary expert panel comprising cybersecurity analysts, computational linguists, and fraud-detection practitioners. Their expertise respectively covered fraud logic and policy violations, Indian English/Hinglish code-switching, and social-engineering realism, collectively shaping scenario coverage, boundary cases, and the human validation protocol. Specifically, each expert evaluated 25 conversations (15 overlapping) across quality dimensions (Fluency, Coherence, Realism, Emotion) and fraud classification. Consequently, expert validation yielded a mean pairwise Cohen's Kappa [Cohen, 1960] of 0.534, indicating moderate inter-annotator agreement, alongside 80% consensus, realism 3.96/5, and confidence 4.29/5. Notably, disagreements were concentrated in Subtle Fraud and Ambiguous Normal cases, indicating boundary-case difficulty rather than annotation collapse. Accordingly, ICFD-31k should therefore be interpreted as a controlled benchmark for conversational fraud research rather than a replacement for real-call deployment validation. In this regard, important remaining gaps include ASR errors, noisy audio, accent and dialect variation, realistic deployment class priors, and the asymmetric chunk-sampling protocol. In order to address these gaps, future evaluation will couple the benchmark with speech-to-text front ends, simulate ASR perturbations, and run interval/sampling ablations to quantify the trade-off between earlier warnings and false alarms. Furthermore, for societal use, the intended deployment role is a human-in-the-loop warning layer, not autonomous call blocking or punitive decision-making. In particular, the stakeholder path is to work with cybercrime support teams, telecom safety groups, and digital-literacy organizations to collect privacy-preserving feedback on emerging scam patterns, warning language, and intervention timing. Moreover, since no comparable Indian bilingual fraud benchmark currently exists, the released cross-domain split provides an initial reproducible stress test until independent real-world evaluation sets become available. Collectively, the feedback loop can keep the benchmark current while preserving the restricted defensive-use license and auditability requirements.

6 Conclusion

The paper tackles a major bottleneck in conversational fraud detection research - the lack of large-scale, public, contextually relevant benchmarks - by introducing ICFD-31k, a novel dataset comprising over 31,000 richly annotated bilingual (English/Hinglish) conversations across 10 prevalent Indian fraud typologies. ICFD-31k features a novel annotation schema with chunk-level labels for real-time streaming analysis and detailed slow-thinking rationales to support next-generation explainable and streaming-capable AI systems. The paper also presents a comprehensive benchmark with hierarchical baselines to validate the ICFD-31k's quality. The obtained results show that fine-tuned classifiers achieve near-perfect in-domain performance (99.40% F1) in static settings. However, performance degrades substantially on novel, unseen threats, exposing a critical generalization gap underscoring the limitations of relying solely on historical accuracy and highlighting the need for more adaptable, robust models in evolving fraud landscapes.

Ethical Statement

ICFD-31k is released for defensive research only under a restricted license to reduce dual-use risk. All names, organizations, account details, phone numbers, and other PII are synthetic and do not correspond to real victims or real call recordings. Expert annotators participated with informed consent and were briefed on the defensive purpose of the study. The dataset is intended to support fraud prevention, safety research, and auditable countermeasures; offensive use, scam-script generation, or malicious social-engineering training is prohibited.

Acknowledgments

The authors acknowledge the support of the Ministry of Education, Government of India, and the Department of Information Technology, Dr. B. R. Ambedkar National Institute of Technology Jalandhar, Punjab, India for academic and institutional facilities.

References

[Almeida *et al.*, 2011] Tiago A Almeida, José María Gómez Hidalgo, and Akebo Yamakami. Contributions to the study of sms spam filtering: new collection and results. In *Proceedings of the 11th ACM Symposium on Document Engineering*, pages 259–262. ACM, 2011.

[Bai *et al.*, 2025] Haitao Bai, Pinghui Wang, Ruofei Zhang, Ziyang Zhou, Juxiang Zeng, Yulou Su, Li Xing, Zhou Su, Chen Zhang, Lizhen Cui, Jun Hao, and Wei Wang. Mutationguard: A graph and temporal-spatial neural method for detecting mutation telecommunication fraud. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)*, pages 9547–9554, 2025.

[Barredo Arrieta *et al.*, 2020] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, et al. Explainable artificial

intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[Chung *et al.*, 2024] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[Cohen, 1960] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[Dal Pozzolo *et al.*, 2014] Andrea Dal Pozzolo, Olivier Caelen, Yann-Aël Le Borgne, Serge Waterschoot, and Gianluca Bontempi. Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, 41(10):4915–4928, 2014.

[Davidson *et al.*, 2017] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515, 2017.

[Fette *et al.*, 2007] Ian Fette, Norman Sadeh, and Anthony Tomasic. Learning to detect phishing emails. In *Proceedings of the 16th International Conference on World Wide Web*, pages 649–656. ACM, 2007.

[Gama *et al.*, 2014] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):1–37, 2014.

[Gemma Team *et al.*, 2024] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology, 2024.

[Hoi *et al.*, 2021] Steven C.H. Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. Online learning: A comprehensive survey. *Neurocomputing*, 459:249–289, 2021.

[Jelinek *et al.*, 1977] Frederick Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63, 1977.

[Lei *et al.*, 2016] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117. Association for Computational Linguistics, 2016.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[Ma *et al.*, 2025] Zhiming Ma, Peidong Wang, Minhua Huang, Jingpeng Wang, Kai Wu, Xiangzhao Lv, Yachun Pang, Yin Yang, Wenjie Tang, and Yuchen Kang. Teleantifraud-28k: An audio-text slow-thinking dataset for telecom fraud detection, 2025.

[Meta, 2024] Meta. Llama 3.3 model card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md, 2024. Accessed: 2026-06-07.

[Puri *et al.*, 2020] Raul Puri, Ryan Spring, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. Training question answering models from synthetic data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5811–5826. Association for Computational Linguistics, 2020.

[Schick and Schütze, 2021] Timo Schick and Hinrich Schütze. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951. Association for Computational Linguistics, 2021.

[Shu *et al.*, 2017] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

[van der Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022.

[Wiegreffe and Pinter, 2019] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20. Association for Computational Linguistics, 2019.

[Wu *et al.*, 2024] Jiasheng Wu, Xin Liu, Dawei Cheng, Yi Ouyang, Xian Wu, and Yefeng Zheng. Safeguarding fraud detection from attacks: A robust graph learning approach. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)*, pages 7500–7508, 2024.

[Ye *et al.*, 2022] Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. Zerogen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669. Association for Computational Linguistics, 2022.

[Zhang *et al.*, 2018] Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361. Association for Computational Linguistics, 2018.

[Zipf, 1949] George Kingsley Zipf. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press, Cambridge, MA, 1949.

[Zou and Cheng, 2024] Yao Zou and Dawei Cheng. Effective high-order graph representation learning for credit card fraud detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)*, pages 7581–7589, 2024.