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

[1]Rishi Ahuja • [1]Kumar Prateek • [1]Simranjit Singh

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

Why a new benchmark?

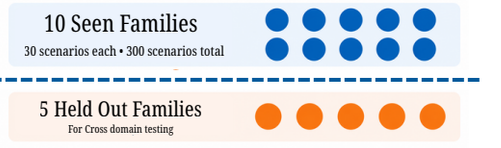
Phone scams are temporal: early turns look harmless, evidence arrives gradually, and the useful decision point is **before the call ends.**

Real-data barrier. Victim recordings carry PII, consent risks, and active-scam details; large public corpora are difficult to release safely.

Benchmark gap. Existing datasets are often English-only, post-call, or lack realistic Indian entities and Hinglish dialogue.

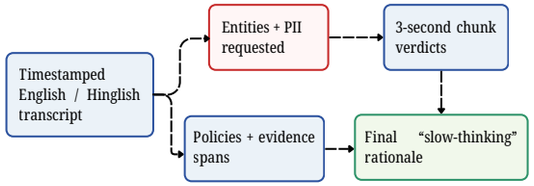
ICFD-31k turns each synthetic call into both a full-call example and a sequence of supervised, cumulative decisions.

Coverage is designed, not sampled



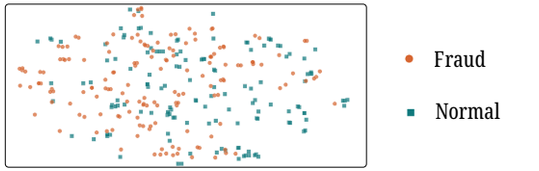
Four deliberate difficulty modes: Clear Fraud, Subtle Fraud, Clear Normal, and Ambiguous Normal. Suspicious words can occur in normal calls; decisive evidence may arrive late.

One call, a rich supervision graph



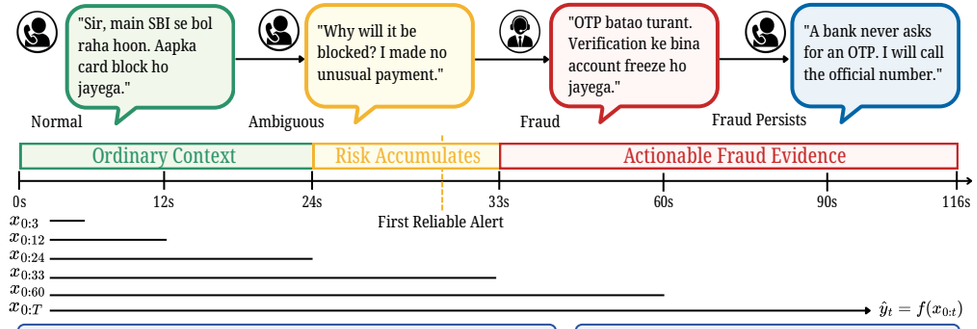
The dataset supports detection, early-warning latency, policy-grounded explanations, and future work on faithful rationales, not just a binary verdict.

Hard negatives overlap in language space



Fraud and normal samples are intentionally interleaved rather than cleanly clustered (semantic separation = 0.017). A detector must use conversational context and policy evidence; surface keywords are not a reliable shortcut.

A live call becomes cumulative evidence every 3 seconds



Streaming task. At time t, predict y_t from the cumulative transcript x0:t. Labels arrive every 3 seconds; evaluation measures both accuracy and when a reliable fraud warning first appears.

99.82% fraudulent calls flagged

32.2 s mean first-detection latency

From controlled synthesis to two benchmark views

CONTROLLED INPUTS
- 10 fraud umbrellas
- 300 scenarios
- 10 agent × 10 customer personas
- Indian banks, names, regions

PROMPT FACTORIAL

$k \in S \times A \times C \times E$

English / Hinglish instruction policy constraints balanced outcomes

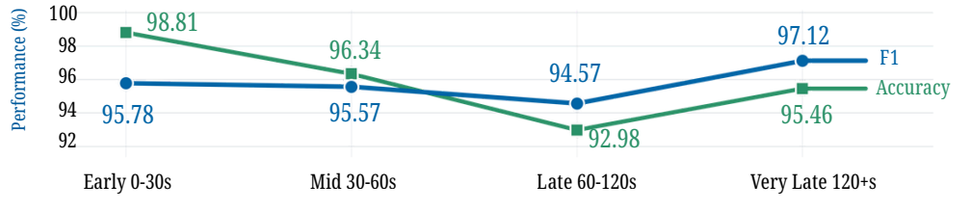
STRUCTURED GENERATION

Llama 3.3 70B timestamped JSON verdict + evidence rationale + policy tags

ICFD-31K Benchmark

31,000 full calls
1,111,071 cumulative chunks
- M1 static full-call
- M2 real-time streaming

Accuracy remains stable as calls unfold



Scale, class balance, and auditability

31,000 Source conversations

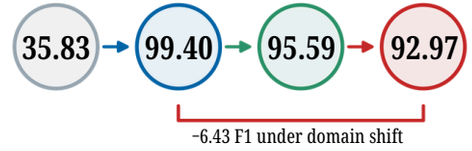
1.11M 3-second chunks

115.8s Mean call length

70:30 Fraud : Normal

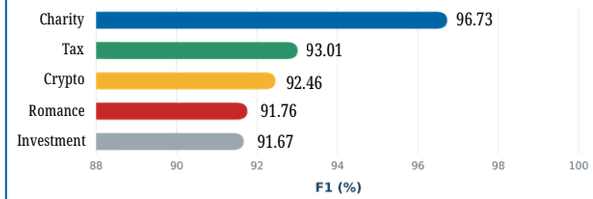
Auditable by design: Every chunk retains its source call, timestamp, cumulative transcript, verdict, rationale, and policy evidence.

The benchmark is a stress ladder



The scientific result is the gap, not the near-perfect score. Static in-domain F1 is 99.40, but unseen scams reduce F1 to 92.97 and recall by 11.56 points.

Transfer varies across unseen scams



Charity transfers best; investment and romance remain hardest. A benchmark should expose where new data collection and adaptation are needed.

Is synthetic dialogue credible?

Automated quality

Perplexity	20.44/22.37
Distinct-3	0.611
3-gram overlap	1.96%
Semantic separation	0.017

Five-expert audit

Fluency	4.19 / 5
Realism	3.96 / 5
Confidence	4.29 / 5
Consensus	80%

Fraud and normal calls overlap heavily in embedding space, consistent with intentionally hard negatives. Mean pairwise Cohen's κ = 0.534; disagreement concentrates on Subtle Fraud and Ambiguous Normal.

Contact and Resources

Github & Code

Dataset

Contact Authors