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

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

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

Why a new benchmark?

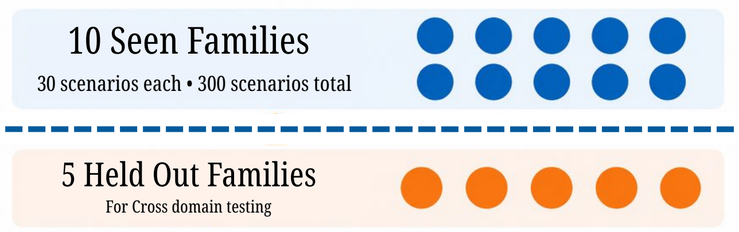
Phone scams are temporal: early turns look harmless, evidence arrives gradually, and the useful decision point is **before the call ends.**

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.

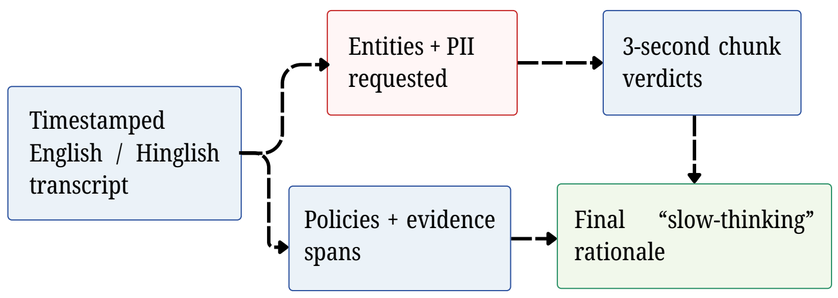
ICFD-31k turns each synthetic call into both a full-call example and a sequence of supervised, cumulative decisions.

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.

Scale, class balance, and auditability

31,000 Source conversations

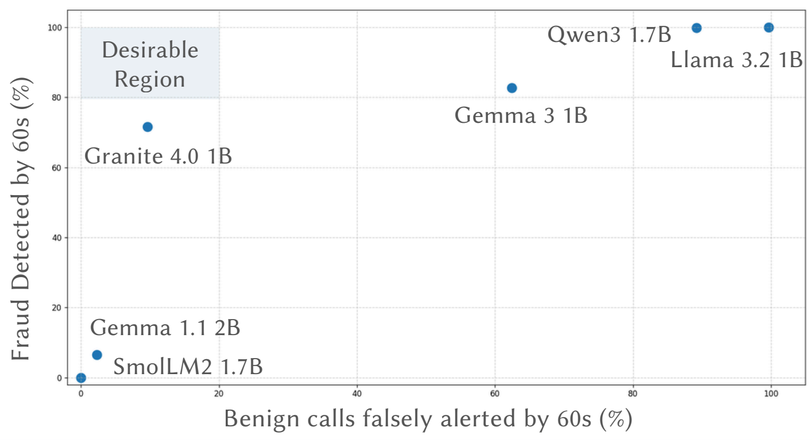
1.11M 3-second chunks

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.

Early-Detection Trade-off at 60 Seconds



Is synthetic dialogue credible?

Automated quality

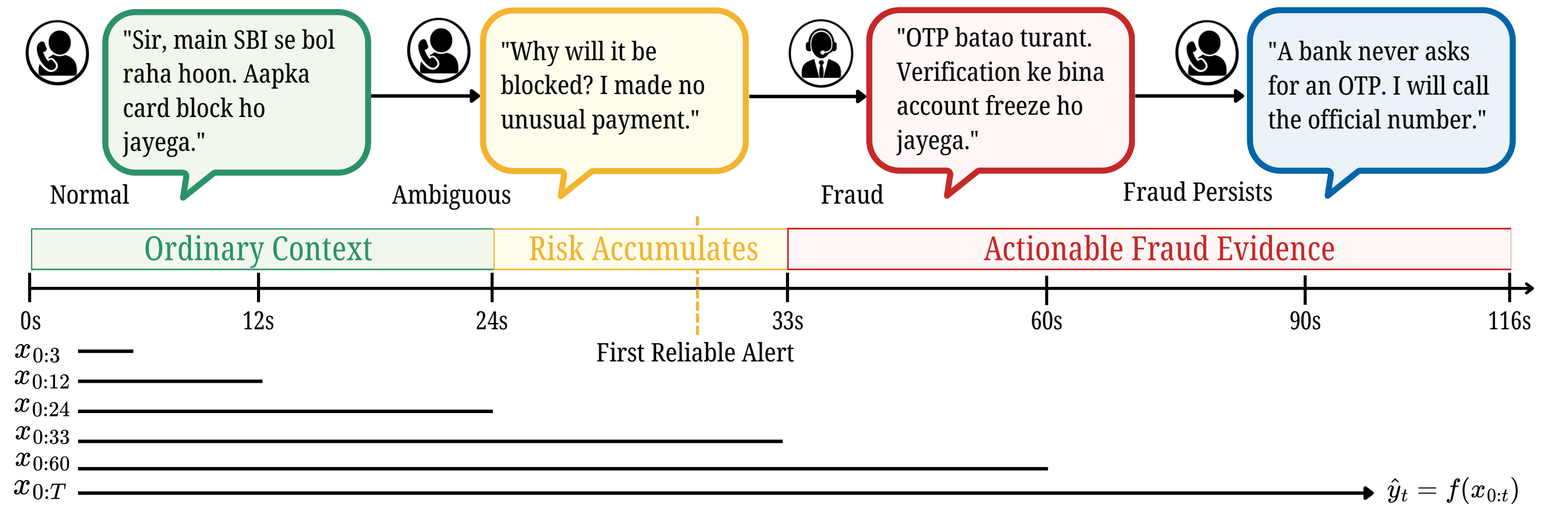
Perplexity	20.44/22.37
Distinct-3	0.611
3-gram overlap	1.96%
Semantic separation	0.017

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.

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

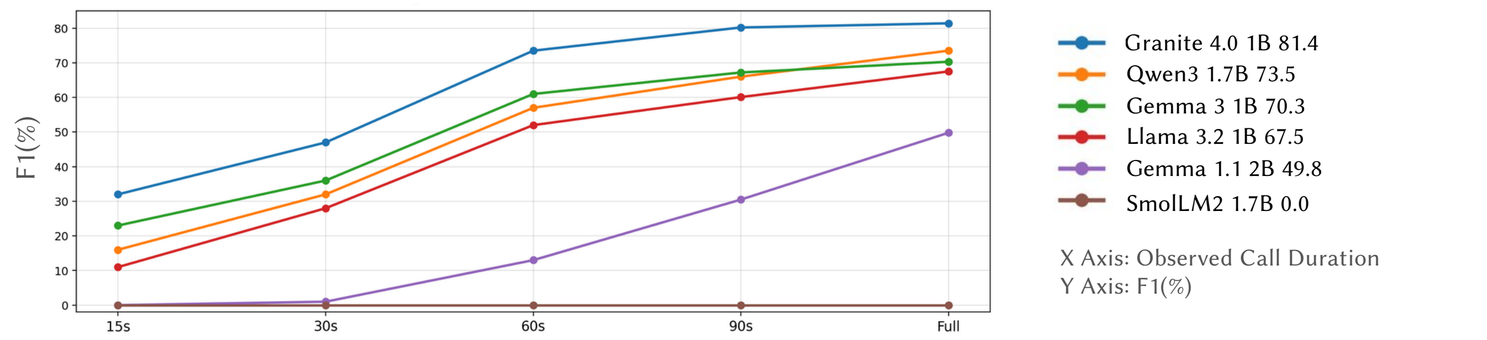
STRUCTURED GENERATION

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

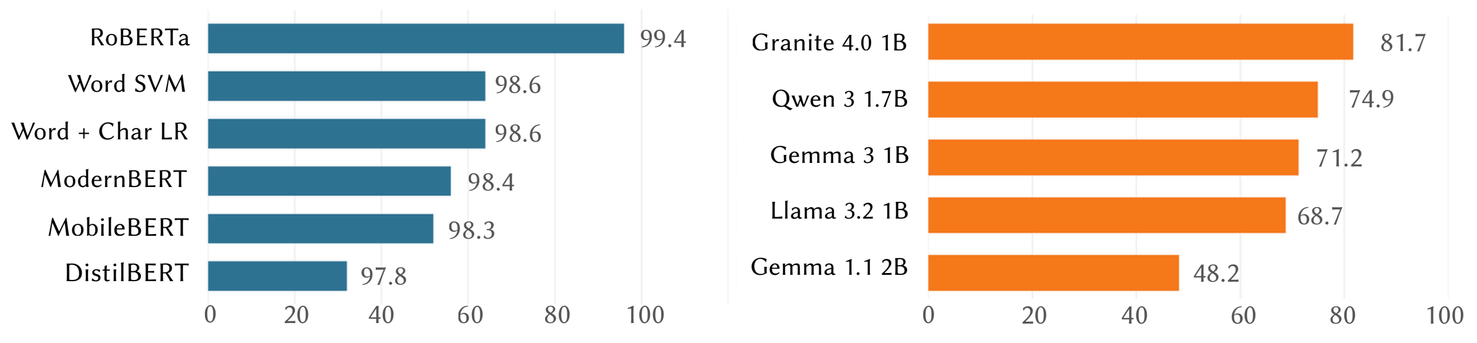
ICFD-31K Benchmark

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

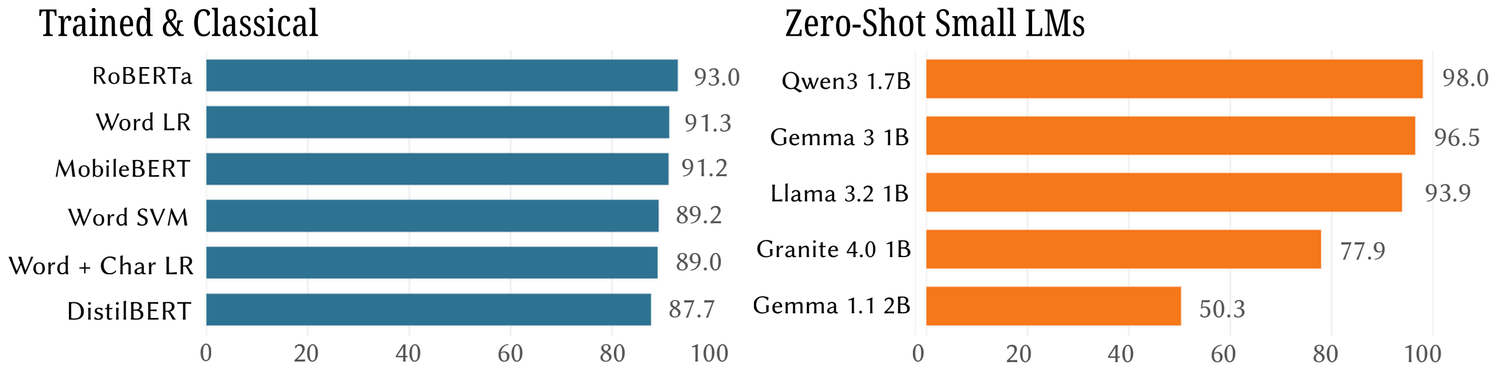
Streaming Zero-Shot Performance Over Time



Full Context Classification · F1 (%)



Cross-Domain Evaluation on Unseen Fraud Types · F1 (%)



Contact and Resources

GitHub & Code

Dataset

Contact Authors

Acknowledgements

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.