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 Information Technology, Dr. B.R. Ambedkar NIT Jalandhar, India

Presenter: Rishi Ahuja

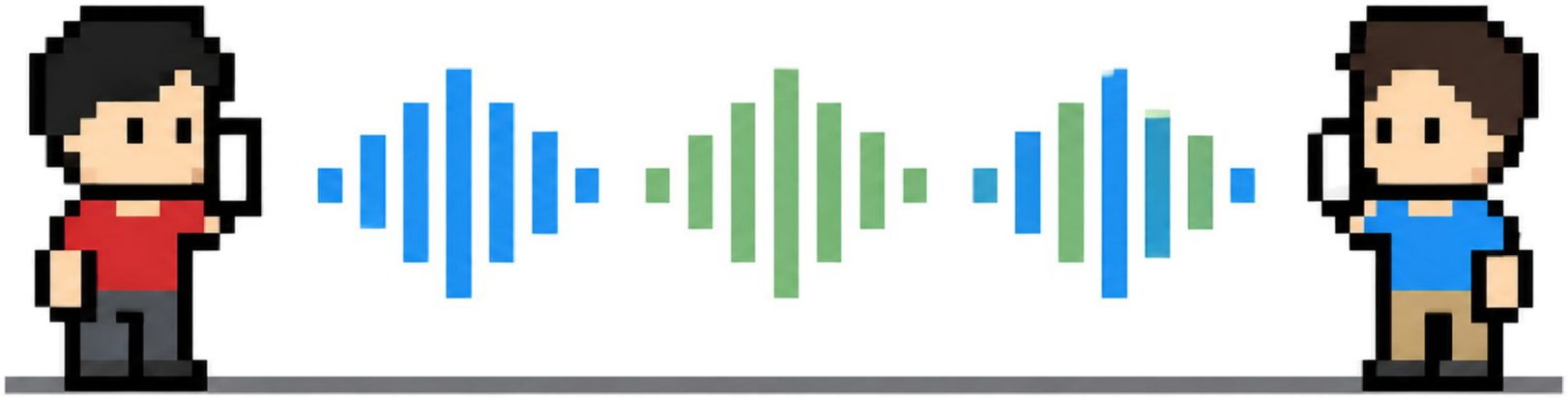
Poster: Hall 4 (Halle 4)

Conversational Fraud in India

Phone scams cause financial and emotional harm, especially for vulnerable users.



Real Indian conversations are often code-switched

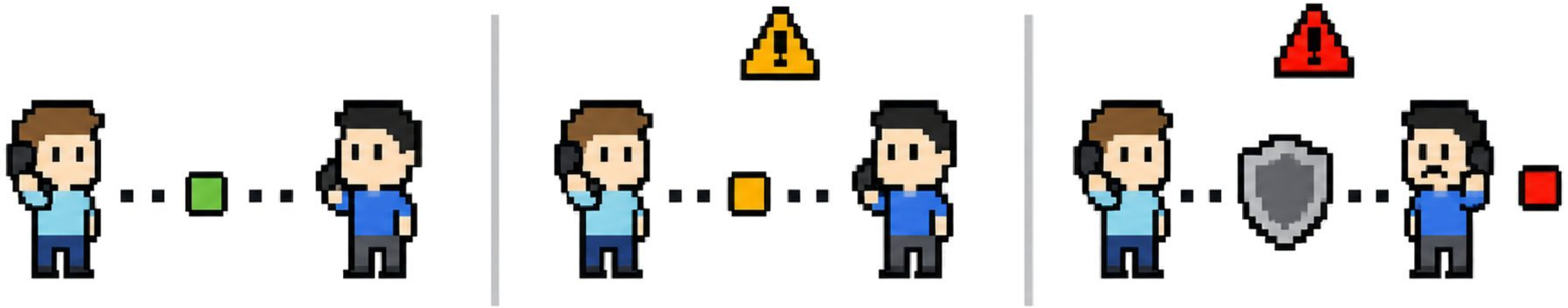


English + Hindi = Hinglish

Existing benchmarks largely miss this setting.

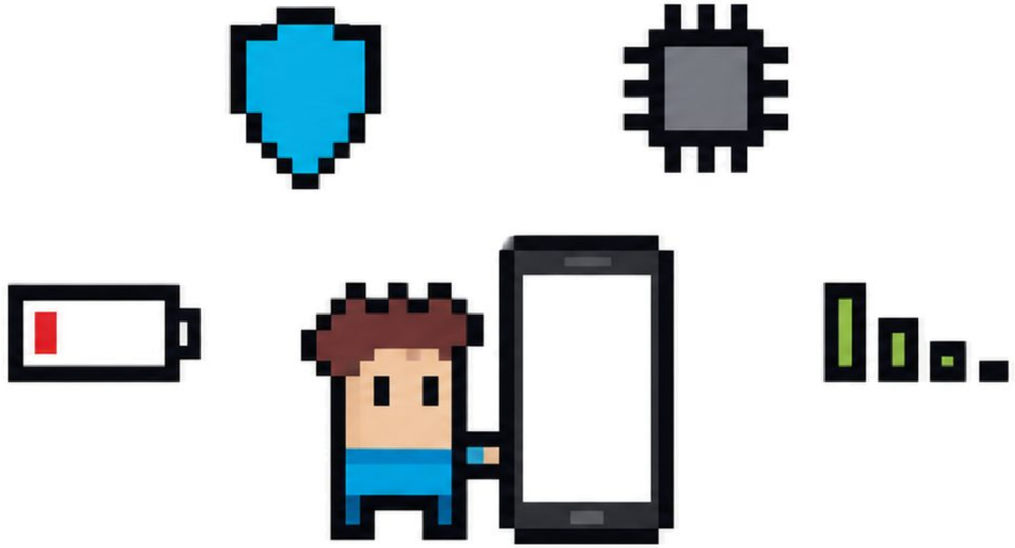
Detecting fraud after the call is too late!

Early | Suspicious | Fraud Signal



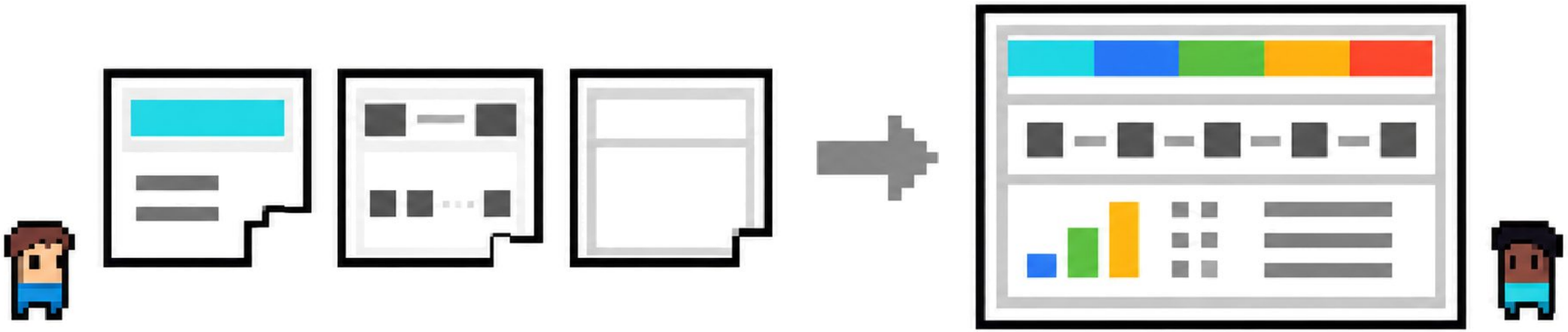
The model must decide with incomplete information.

Practical protection must be lightweight!



Designed with low-resource and mobile settings in mind.

Benchmark Design Requirements



Indian bilingual context + Streaming labels + Explainable rationales = All in one benchmark

ICFD-31k Overview

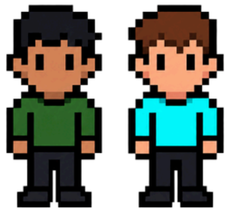
An Indian-code-mixed, real-time conversational fraud detection benchmark.



31,000+
conversations



10
fraud umbrellas



Hinglish
code switch



1.1M+
temporal chunks



Streaming Labels

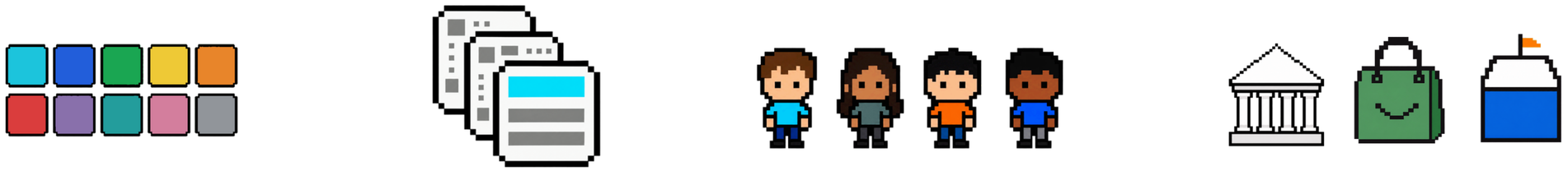


Explainable Rationales

Built for early, practical fraud detection in India.

From bilingual conversations to real-time, explainable decisions.

Scenario Diversity and Controlled Generation



10 Fraud Categories | 300+ Scenarios | Personas | Indian Entities



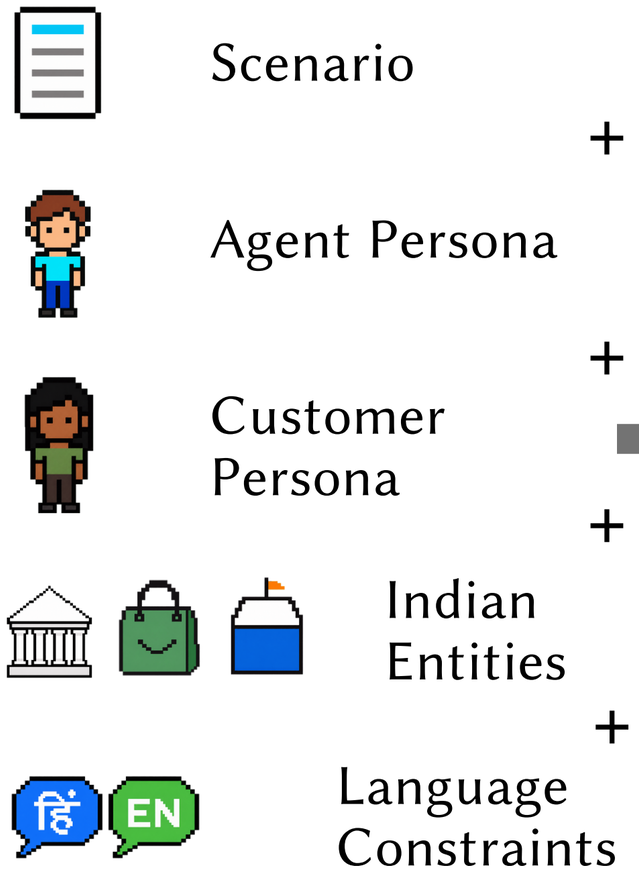
Controlled Generation

~31K configurations



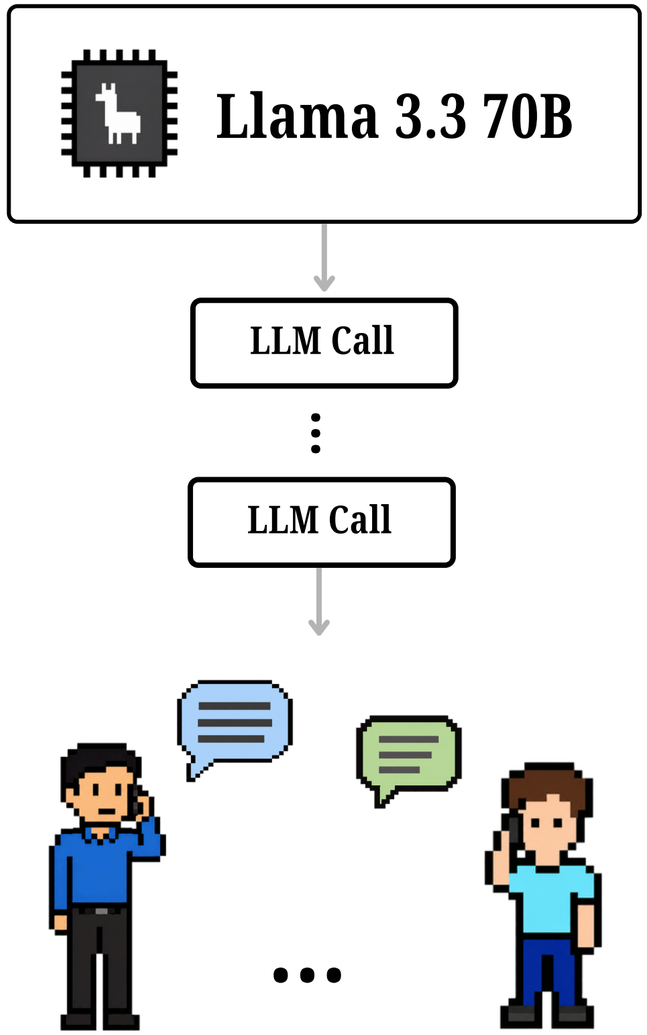
1.96% average 3-gram overlap

Conversation Generation Overview

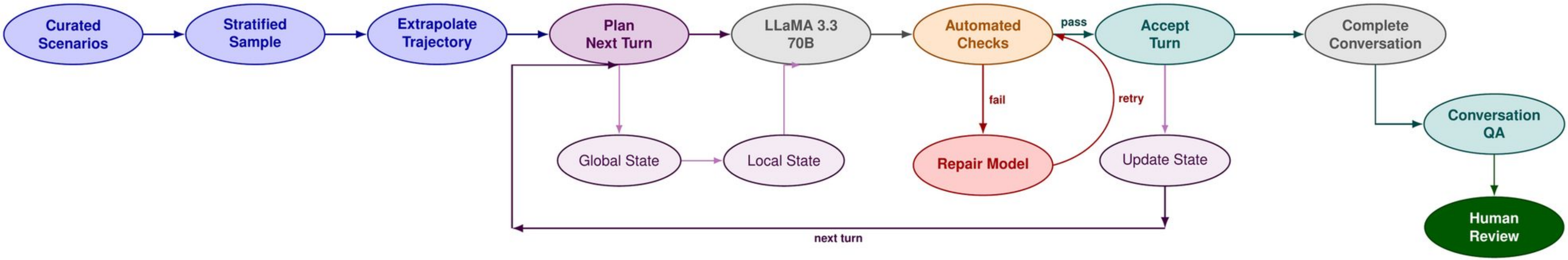


Turn Planner (example)

01	Introduce Context
02	Establish interaction
03	Advance the scenario
04	Customer response
05	Resolve / Conclude

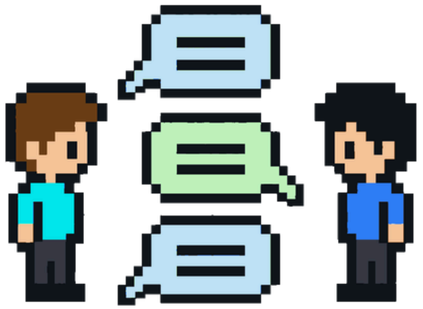


Generation Pipeline



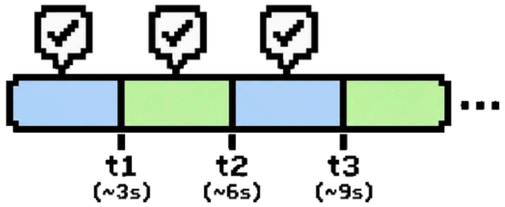
Benchmark Tasks: Static, Streaming, and Unseen fraud

Static



Full conversation
One fraud / benign prediction
One final rationale

Streaming



Cumulative conversation every
~3 seconds
Verdict + rationale at each point

Cross Domain



Train on 10 main umbrellas
Test on 5 unseen fraud types
1,000 unseen conversations

Expert Annotation Study

5 expert annotators



100 conversations sampled
fraud / normal
clear / subtle / ambiguous

15 shared conversations
rated independently by all 5 experts

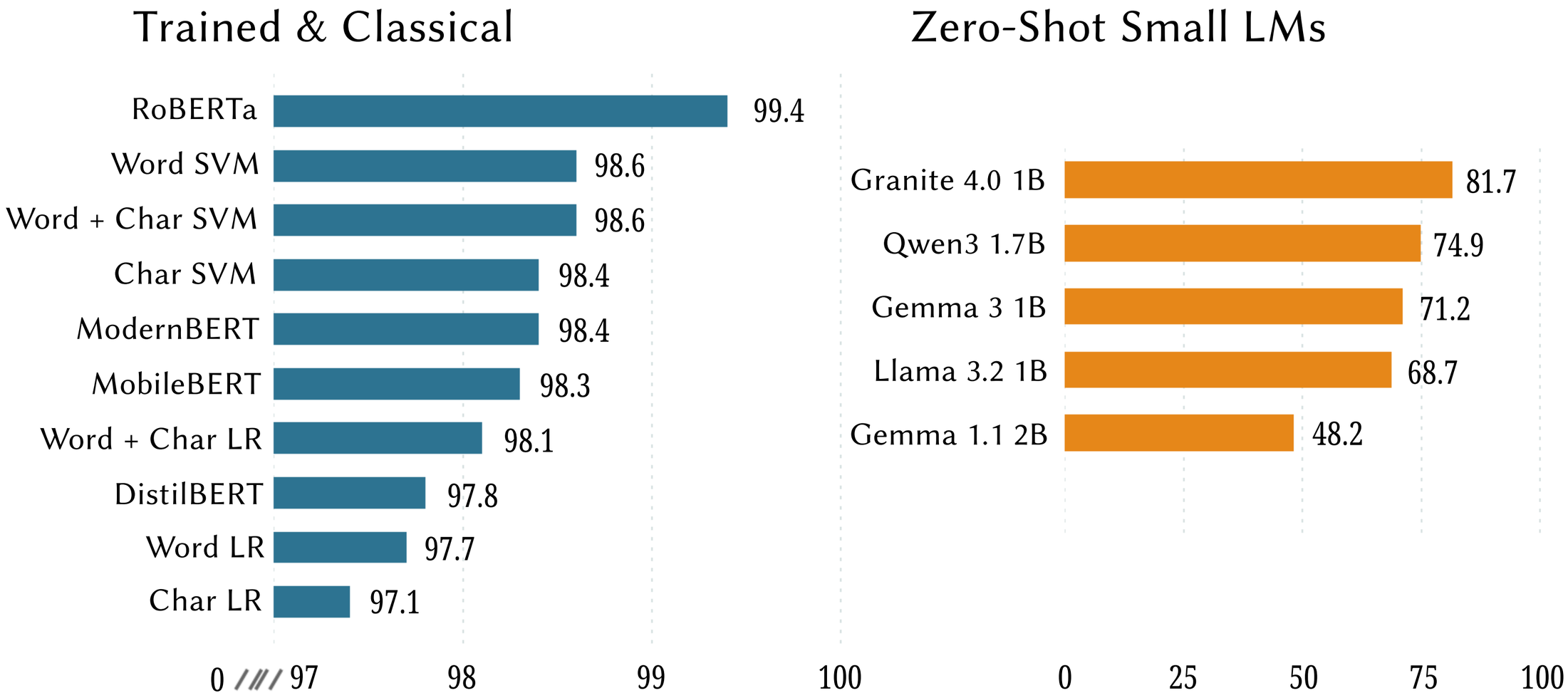
Each expert provided
fraud / normal verdict
confidence & comments

Cohen's κ = 0.566

(meanwise kappa) moderate agreement on deliberately difficult cases

Full-Context Classification

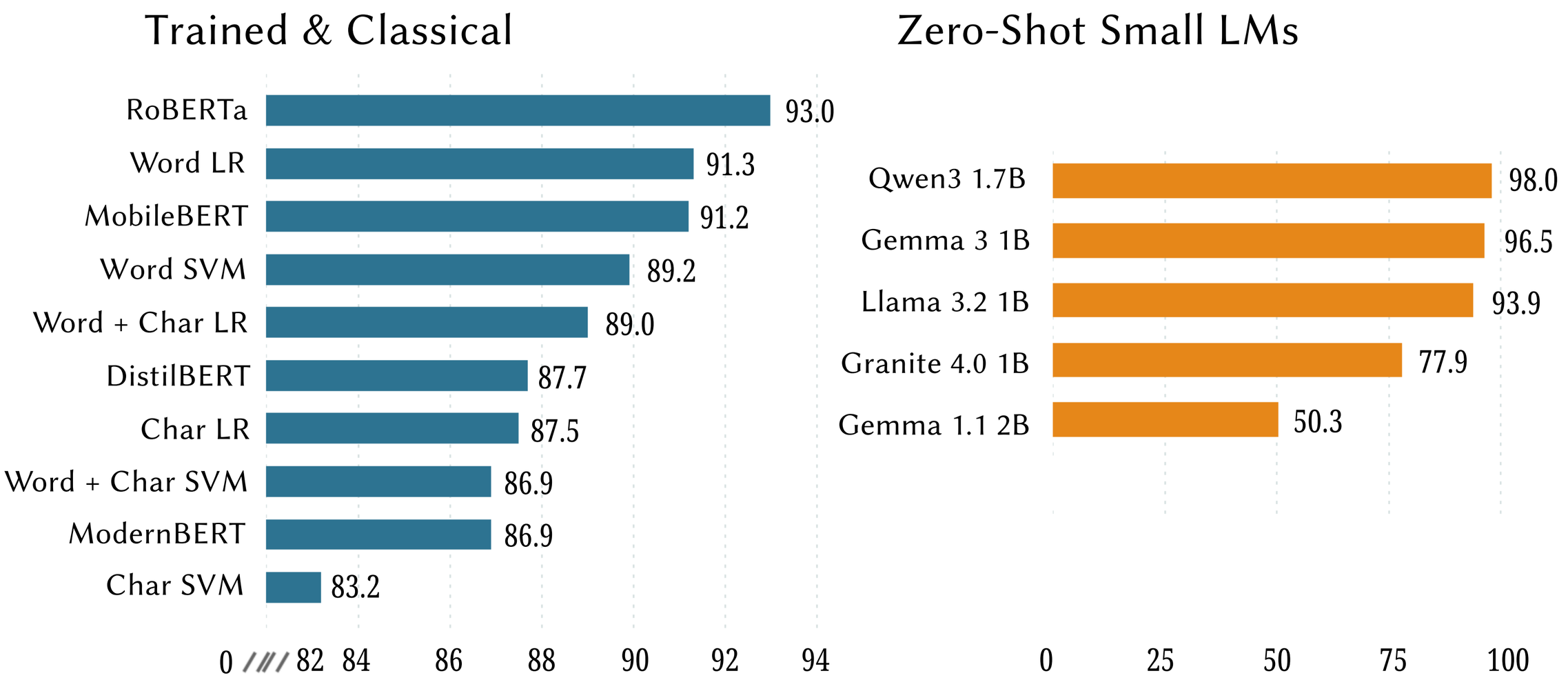
Static classification · full conversation available · F1 (%)



*Includes additional post-acceptance benchmarking on ICFD-31k.

Cross-Domain Evaluation on Unseen Fraud Types

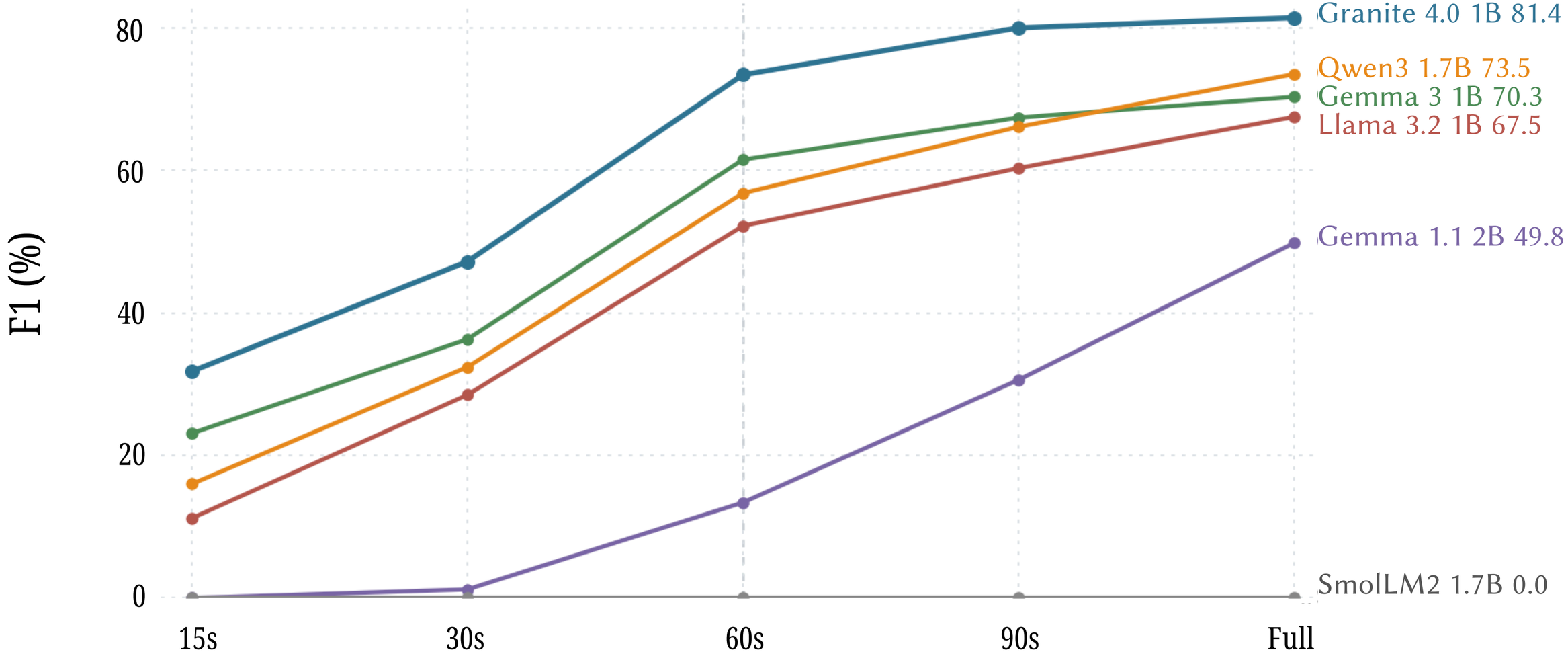
Cross-domain evaluation · Train on 10 fraud umbrellas · Test on 5 unseen fraud types · F1 (%)



*Includes additional post-acceptance benchmarking on ICFD-31k.

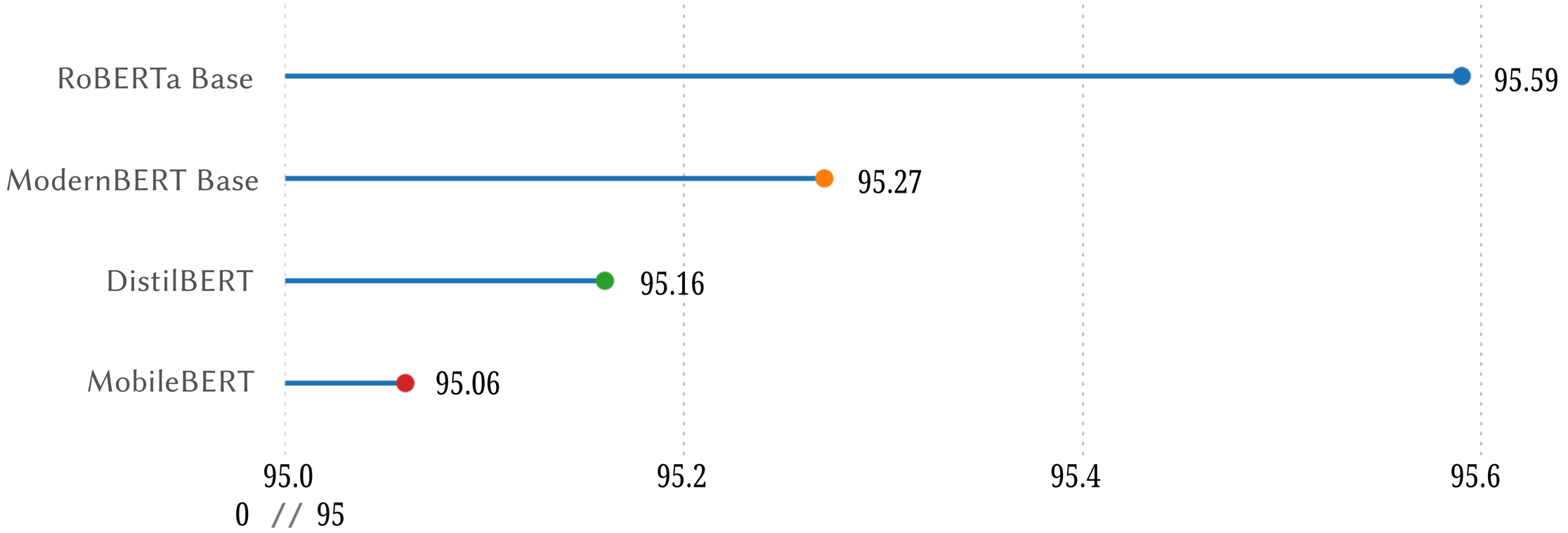
Streaming Zero-Shot Performance Over Time

Streaming zero-shot evaluation • F1 over time

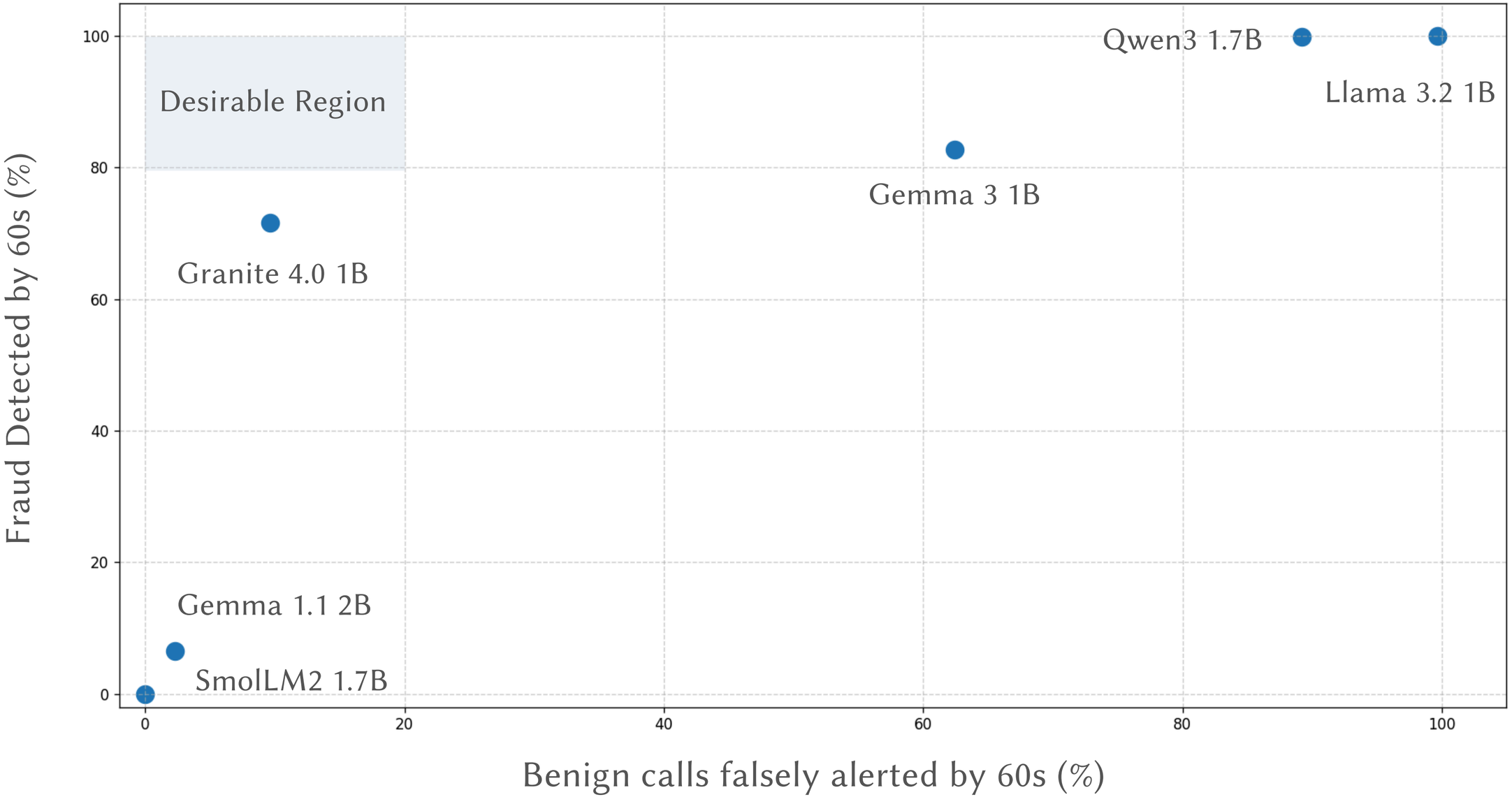


Encoder Performance in the Streaming Setting

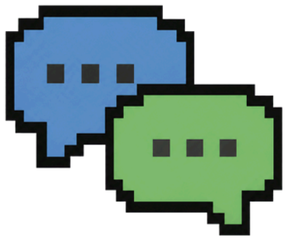
Supervised streaming evaluation • F1 (%)



Early-Detection Trade-off at 60 Seconds

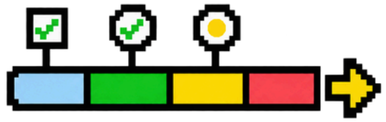


Thank You!



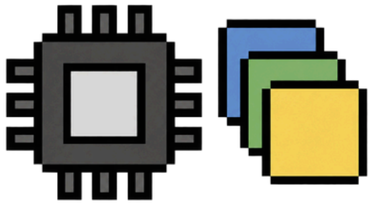
Bilingual

English + Hinglish



Realtime

Detection as call unfolds



Broad Bench

Classical + encoders + small LMs



Practical

Compact models remain competitive

Github & Code

Dataset

Contact Authors