JAAL: A Benchmark of Hinglish Scam Conversations

Anonymous authors
Paper under double-blind review

ABSTRACT

Scam calls unfold over many turns, yet public fraud datasets are mostly short messages, and to our knowledge none targets Hinglish, the Hindi-English mix common in Indian phone scams. We introduce JAAL, a corpus of 9,677 synthetic Hinglish scam conversations in eight categories, together with a benchmark built on it. Gemma 3 27B agents wrote the conversations from short fraud-message seeds, and automatic steps assigned the labels. RoBERTa-base reaches 92.33% Macro-F1 on the stratified synthetic test set, and TF-IDF models come within about two points. Two annotators labeled 1,045 conversations independently. They agree with the automatic labels on 86.2% and 76.6% of a synthetic sample, and their consensus matches the label in 92.5% of cases. We also test every model on 77 YouTube-sourced scam conversations. Supervised models fall to between 27% and 56% Macro-F1 there, and zero-shot models reach between 60% and 69%. The set is small and its labels are uncertain, so we read it as a stress test and not as an estimate of real-world accuracy. We also evaluate Jev, a zero-shot classifier that could suit deployment, and it matches GPT-OSS 120B on external accuracy (66.23%). We release the corpus,[1] the annotations, and the code.[2]

1 INTRODUCTION

A scam call is a conversation. The caller opens with a pretext, waits for the receiver to react, and adjusts the story until the receiver shares a code or sends money. The evidence of fraud is spread across turns, so a detector that reads one message at a time loses most of it. Most public fraud resources are still short texts such as SMS messages (Almeida et al., 2011). Surveys and call analyses describe how phone scams work (Tu et al., 2016; Wood et al., 2023), but a labeled corpus of scam conversations for classification is harder to find.

Scam calls in India are often conducted in Hinglish, which mixes Hindi and English inside one conversation. Code-switching lowers the accuracy of models built on monolingual data (Sitaram et al., 2019; Khanuja et al., 2020; Winata et al., 2021). A useful corpus therefore has to be conversational, code-switched, labeled by scam type and large enough to train a classifier. Real recordings rarely meet these conditions. Recording a private call raises consent and legal questions, and the recordings hold personal and financial details. Language models can write conversations instead (Wang et al., 2023; Long et al., 2024), but a generated corpus inherits the habits of its generator (Shumailov et al., 2024). Nobody knows in advance whether a classifier trained on it works on conversations it did not produce.

The proposed work builds such a corpus and measures that transfer. JAAL contains 9,677 synthetic conversations in eight scam categories. Two Gemma 3 27B agents, a caller and a receiver, wrote them from short fraud-message seeds, and few-shot generation extended the smaller classes. The labels come from automatic steps. We ask three questions. How well do standard models classify the corpus? How far can its labels be trusted? How much of the accuracy survives on conversations from another source? We answer with a benchmark of linear models, a small Transformer, three pretrained encoders and two zero-shot language models, with a two-annotator check of 1,045 conversations, and with a set of 77 YouTube-sourced scam conversations. Furthermore, supervised models reach 88% to 93% Macro-F1 on the synthetic test set and 27% to 56% on the external set. The best zero-shot model reaches 68.61% there. Two annotators agree with the automatic labels on 86.2% and 76.6% of a synthetic sample, and less often on the external set. The external set has 77 items and uncertain labels, so we read its numbers as a stress test with wide intervals. Even so, the linear models mostly predict one class on it, legacy telecom, and that pattern shows how the corpus can mislead a classifier.

1.1 CONTRIBUTIONS.

The specific contributions are as follows:

1. We release the JAAL corpus together with a construction record that specifies which pipeline steps are automatic and which provenance information the release does not retain, along with a characterization of the corpus that includes a duplicate and label-conflict audit.

[1] https://github.com/1007-afk/JAAL-10k-dataset
[2] https://github.com/1007-afk/JAAL-code

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2. We provide a benchmark reporting confidence intervals and human reference rows, and include Jev as a zero-shot option to reflect realistic deployment settings.
3. We provide an independent validation package consisting of a two-annotator validation of 1,045 conversations, released with per-annotator labels, and an external stress set, released as a manifest of video identifiers and labels.

2 Related Work

Scam and fraud text resources. The SMS Spam Collection covers short messages (Almeida et al., 2011). A survey of telephone spam defenses (Tu et al., 2016) and an analysis of YouTube scam-baiting calls (Wood et al., 2023) describe phone scams from a systems and content view. The second study notes that true scam victims do not in general record their conversations, which is the difficulty JAAL starts from. More recent resources are closer to ours (Table 1). TeleAntiFraud-28k pairs audio and text for telecom fraud tasks (Ma et al., 2025). PreScam builds 11,573 conversational instances from user-submitted scam reports (Sun et al., 2026). A honeypot corpus holds 10,211 real inbound calls (Traister et al., 2026). ICFD-31k contains 31,000 generated conversations with a binary fraud verdict, chunk-level labels and rationales (Ahuja et al., 2026). None of them is a Hinglish corpus with an eight-way scam taxonomy.

Table 1: Related conversational scam resources.

Resource	Size	Source	Labels or task
TeleAntiFraud-28k	28,511 audio-text pairs	Real recordings, LLM expansion, multi-agent synthesis	Scenario, fraud and fraud-type classification
PreScam	11,573 instances	User-submitted scam reports	Scam-progression prediction, 20 categories
Honeypot corpus	10,211 calls	Real inbound calls answered by a voice agent	Analysis, early-escalation prediction
ICFD-31k	31,000 conversations	Generated (Llama 3.3 70B)	Binary verdict, chunk labels, rationales
JAAL	9,677 conversations + 77 external	Generated Hinglish; YouTube-sourced stress set	Eight categories; 1,045 doubly annotated

Synthetic scam data. Two studies train scam detectors on LLM-generated conversations and test on human-written or real ones. One reports accuracy and F1 above 90% on a small set of human-written Roman Urdu transcripts (Irfan et al., 2025). The other compares single-agent and multi-agent generation and reports good results on real-world scam conversations (Gumphusiri and Triyason, 2024). Both solve a scam versus non-scam task. JAAL asks an eight-way category question and finds a large drop on its external set, so the results do not conflict, but they do show that transfer depends on the task and on the test data.

Synthetic data and its limits. Instruction and dialogue generation is now common (Wang et al., 2023; Li et al., 2023a; Ding et al., 2023), and a survey lists its usual failures: limited diversity, template reuse and narrow coverage (Long et al., 2024). Training on model output can remove the tails of a distribution (Shumailov et al., 2024). Synthetic training data helps less when the labeling task is subjective (Li et al., 2023b). A synthetic benchmark can also favor the model that generated it (Maheshwari et al., 2024). The last point matters here, because Gemma models generated JAAL and one of them is also a zero-shot evaluator.

Hinglish and code-switching. Hindi and English are mixed in everyday Indian writing, usually in Roman script (Bali et al., 2014). GLUECoS, LinCE and SemEval-2020 Task 9 exist because models trained on monolingual text do not carry over to code-mixed input (Khanuja et al., 2020; Aguilar et al., 2020; Patwa et al., 2020). Multilingual models handle code-switching unevenly (Winata et al., 2021), Romanized Hindi has no standard spelling (Roark et al., 2020), and tokenizers fragment low-resource text heavily (Rust et al., 2020). HingBERT models are pretrained on Hinglish social media text (Nayak and Joshi, 2022), and we evaluate one of them in Section 4.

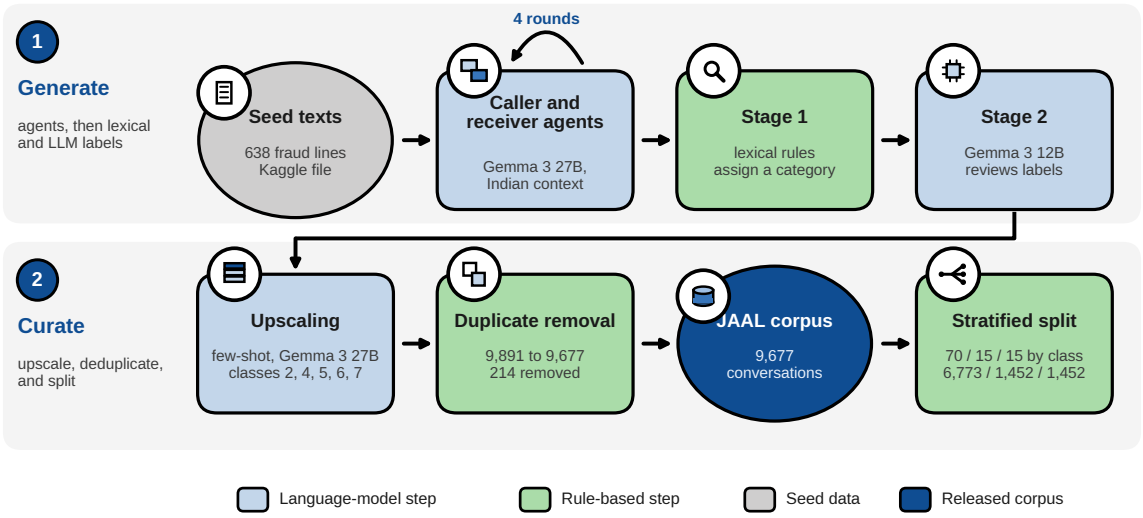


Figure 1: Generation and labeling pipeline.

Table 2: Corpus composition by class and split.

Scam category	Train	Val	Test	Total
Legacy telecom	2,078	445	446	2,969
Banking / KYC / OTP fraud	916	197	197	1,310
Digital arrest / govt. impersonation	679	146	146	971
UPI / wallet fraud	665	142	142	949
Loan / credit app scams	650	139	139	928
Dating / romance / sextortion	630	135	135	900
Delivery / customer care	598	128	128	854
Investment / task scam	557	120	119	796
Total	6,773	1,452	1,452	9,677

Documentation, annotation and distribution shift. Datasheets ask creators to state motivation, composition, collection and intended use (Gebru et al., 2021). A review of 591 text-dataset papers found subpar quality management in about 30% of them (Klie et al., 2024), and an audit of annotation reporting found that training, compensation, adjudication and agreement values are often missing (Kunilovskaya et al., 2026). Intra-annotator agreement is rarely reported (Abercrombie et al., 2023), and releasing annotator-level labels is recommended (Prabhakaran et al., 2021). Benchmark studies ask for reported uncertainty and clear scope (Reuel et al., 2024; Bowman and Dahl, 2021), and an audit of synthetic intrusion datasets found poor diversity and murky labels (Flood et al., 2024). Accuracy drops on newly collected test sets are well documented (Recht et al., 2019; Taori et al., 2020; Koh et al., 2020), and pretrained transformers lose less under shift than models trained from scratch (Hendrycks et al., 2020). We follow these practices where the data allow.

These threads leave one practical question open: can a corpus written by one model family support category classification on conversations it never produced?

3 The JAAL Corpus

Figure 1 shows the construction steps.

3.1 Seeds and generation

The seeds are 638 fraud-labeled lines (589 unique, mean 23 words) taken from the Fraud Call India file on Kaggle. They are short scam scripts and messages, not call transcripts. A simple pattern match (pound amounts, premium-rate style numbers, “txt”) flags 313 of the 638 as UK SMS-spam style text. We treat the seeds as scenario descriptions, and we make no claim about their category distribution. The caller agent

receives a seed and the dialogue so far. The receiver agent receives the dialogue only. Both run Gemma 3 27B (Gemma Team, 2025) with temperature 1.0 and top-p 0.95, and they alternate for four rounds. Both prompts ask for Hinglish, and the caller prompt asks for Indian names and details, so the conversations carry Indian context from the start. The generated conversations go straight to the first labeling stage (Appendix B).

3.2 LABELS AND UPSCALING

Stage 1 assigns a category with lexical rules. Stage 2 asks Gemma 3 12B (temperature 0) to review every assignment and to move conversations it places in another category. The smaller classes were extended after labeling by few-shot generation. Gemma 3 27B saw two or three existing conversations of the class for each new one and wrote one more, with five to six exchanges requested. All labels are automatic. Section 5 measures how far they can be trusted.

3.3 REMOVALS AND SPLITS

An earlier merged folder held 9,891 conversations. We removed 214 (2.2%) before release. Our recomputation finds that 51 of them are exact duplicates (after normalization) of a retained conversation, and 82 have cosine similarity of at least 0.90 to one. The remaining 9,677 conversations are split 70/15/15 at random, stratified by class only (Table 2). We later audited again these 9,677 conversations for train-test leakage. The audit found no exact or normalized duplicates between the training and test sets. It found only three train-test near-duplicate pairs at a cosine threshold of 0.90, indicating limited lexical overlap (Appendix D). The next step is to see what the data contain, because the benchmark scores depend on it.

3.4 CORPUS CHARACTERIZATION

The synthetic train, validation and test partitions match each other closely, so the validation and test sets are fair in-distribution checks. The external set differs from them in four ways. Figure 2 and Appendix C report the details.

Language. Token-level language identification (Gemma 4 31B) puts the synthetic partitions at about 52% English and 42% Hindi tokens. The external set has 11.84% English and 82.43% Hindi tokens. Hindi-dominant conversations make up 42.86% of the external set and 0.21% of the synthetic test set. The word-level statistics in this section use a Latin-only tokenizer, and 232 synthetic conversations (59 of them mostly Devanagari) contain Devanagari characters that it does not count.

Vocabulary. The training set has 21,184 word types. The test set has 17.00% of its word types absent from the training vocabulary, against 50.58% for the external set. An average external conversation has 6.56% unseen tokens, against 0.63% in the test set.

Length and fragmentation. Synthetic test conversations average 320 words. External conversations average 1,378 words, which is 4.3 times longer, or 3.6 times when we collapse the repeated word in one transcript (mean 1,165). The RoBERTa tokenizer splits words into 1.78 pieces on the external set and 1.52 on the test set. RoBERTa-base reads at most 512 tokens, and 46.6% of test and 92.2% of external conversations exceed that length. ModernBERT reads up to 2,048 tokens, and 0.48% of test and 53.25% of external conversations exceed that length under its own tokenizer. Figure 4 in Appendix C presents the length by class.

Duplicates and label conflicts. The released corpus has no exact duplicates. Character-level TF-IDF finds 11 near-duplicate pairs at cosine 0.90 or higher, in 6 clusters (largest size 3), and 2 pairs cross class boundaries. Three train-test pairs are near duplicates, and two of the three carry different labels (Appendix D). Word-level TF-IDF finds 816 pairs at 0.70 or higher, and 13.8% of them have different labels. Test conversations have a median similarity of 0.25 to their nearest training conversation, and 7.3% reach 0.70. These figures point to small template overlap. The audit found no exact or normalized duplicates between the training and test sets. It found only three train-test near-duplicate pairs at a cosine threshold of 0.90, indicating limited lexical overlap (Appendix D). The corpus is lexically narrow (Distinct-1 0.008, Distinct-2 0.134, Distinct-3 0.378), which fits domain-specific scam vocabulary.

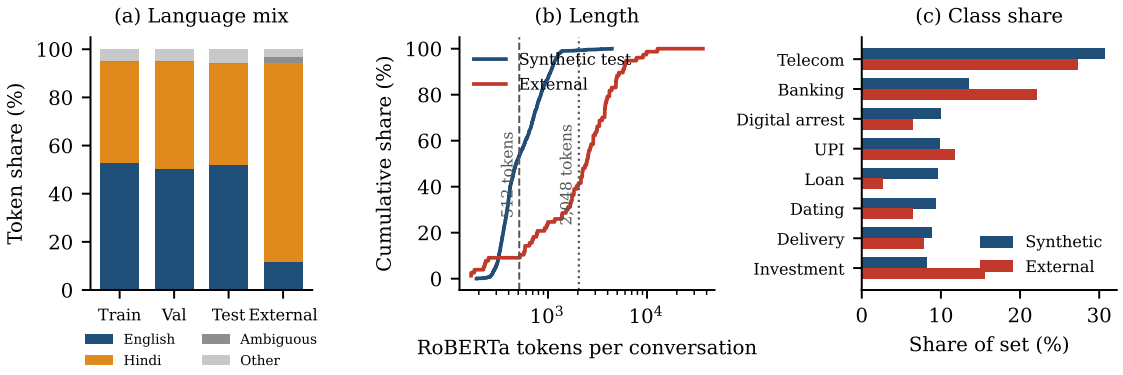


Figure 2: Language mix, length and class share of the synthetic and external sets.

Table 3: Benchmark results.

	Synthetic test		External (n = 77)	
Model	Acc.	M-F1	Acc.	M-F1
Supervised				
Logistic regression	90.56	90.27	44.16	26.56
Linear SVM	90.98	90.65	49.35	39.65
Transformer (scratch)	89.53	89.28	49.35	44.45
RoBERTa-base	**92.63**	**92.33**	59.74	50.17
ModernBERT-base	92.49	92.08	61.04	56.18
HingBERT	91.94	91.58	62.34	53.67
Zero-shot				
GPT-OSS 120B	83.06	83.32	66.23	63.75
Gemma 4 31B	86.85	86.79	**74.03**	**68.61**
Jev	–	–	66.23	60.46
Annotators (reference labels)				
Annotator 1 (practitioner)	86.16†	86.47†	67.53	61.68
Annotator 2 (PhD student)	76.55†	76.19†	53.25	48.59

4 BENCHMARK

This section describes the models used, the characteristics of the in-distribution and external datasets, the failure modes of the linear models, and the class-weighting strategy.

Models. Logistic regression and a linear SVM use TF-IDF features with a Hinglish tokenizer. A three-layer Transformer with 1.42M parameters is trained from scratch on a custom BPE vocabulary of 7,996 tokens. RoBERTa-base is fine-tuned with a 512-token input, and ModernBERT-base with a 2,048-token input, and HingBERT (Nayak and Joshi, 2022), a BERT model pretrained on Hinglish text, with a 512-token input. Gemma 4 31B, GPT-OSS 120B and Jev classify without training examples, from category definitions only. Jev is evaluated on the external set only. We report Macro-F1, and an invalid model output counts as a miss. Each model ran once with seed 42, and the validation set chose every checkpoint. Only the final evaluation used the external set. Appendix G lists the settings.

In distribution. All supervised models exceed 89% Macro-F1 on the test set (Table 3). The annotator rows marked † use the 968-conversation synthetic sample, not the test set, and they measure agreement with the automatic labels. Logistic regression and the SVM come within about two points of the best encoder. Substantial signal therefore sits in lexical cues, which fits the labeling procedure: Stage 1 assigns labels from

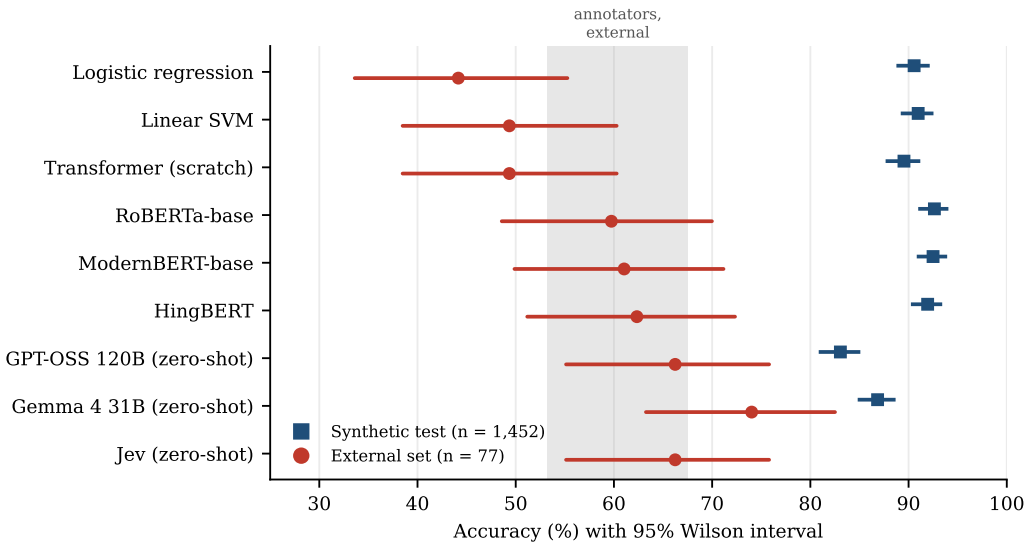


Figure 3: Accuracy on the synthetic test set and the external set.

lexical rules. The zero-shot models score lower, at 83% to 87%, because they know the category names but not the label boundaries the automatic procedure drew. The two annotators, measured on a 968-conversation synthetic sample, agree with the automatic labels less often (86.2% and 76.6%) than RoBERTa does on the test set (92.6%). The two figures come from different samples, so the comparison is only indicative. It suggests that the encoders reproduce regularities of the labeling procedure that human readers do not share.

On the external set. Supervised Macro-F1 falls to between 26.56% and 56.18%. Logistic regression loses 63.7 points, the SVM 51.0, the scratch Transformer 44.8, RoBERTa 42.2 and ModernBERT 35.9. The word-feature models lose the most, and the pretrained models lose less, which matches earlier findings on shift (Hendrycks et al., 2020). Gemma 4 31B reaches 68.61% and GPT-OSS 120B 63.75%. Figure 3 shows accuracy with 95% Wilson intervals. The external intervals are wide: Gemma 4 31B has 74.0% [63.3, 82.5] and ModernBERT 61.0% [49.9, 71.2], and HingBERT 62.3% [51.2, 72.3]. They overlap, so the comparison does not show that the zero-shot model is better than the best encoder. The comparison is also confounded. The encoders read at most 512 or 2,048 tokens, whereas the zero-shot models received the full transcript within their available context windows. HingBERT, the Hinglish-specific encoder, reaches 91.58% Macro-F1 on the test set and 53.67% Macro-F1 and 62.34% accuracy on the external set. This is the highest external accuracy among the supervised models, but with one seed per model we do not claim that it is better than the other encoders. Annotator 1 reaches 67.53% on the external set and Annotator 2 53.25%, so human agreement with these labels is modest as well. Jev is the zero-shot option we consider for deployment settings. It matches GPT-OSS 120B on accuracy (66.23%) and scores 60.46% Macro-F1 (Appendix C).

The model-family comparison provides another view of this result. Gemma 4 31B is from the same model family as the Gemma 3 models used to generate the synthetic corpus, whereas GPT-OSS 120B is from a different family. On the in-distribution test set, neither zero-shot model exceeds the supervised encoders or word-feature models, despite Gemma 4 sharing a model family with the generators. On the external set, however, Gemma 4 31B achieves the highest accuracy and Macro-F1 among the evaluated models. The external conversations come from YouTube transcripts rather than from the synthetic generation process, so this result is not explained by direct access to the generated conversations. Gemma 3 was used only in the external-set filtering pipeline, which may still introduce some family-specific influence.

Where the linear models fail. Logistic regression predicts legacy telecom for 62 of 77 external conversations, and the SVM does so for 56. Legacy telecom accounts for 41 of the 43 errors of logistic regression and 35 of the 39 errors of the SVM. Four classes get an F1 of 0 from logistic regression: UPI, investment, digital

Table 4: Annotator agreement with the reference labels.

Subset	N	A1 acc.	A2 acc.	A1 labeled	A1 / A2 M-F1	κ [95% CI]
Overall	1,045	84.78	74.83	86.69	84.84 / 74.51	0.701 [0.670, 0.733]
Synthetic	968	86.16	76.55	87.79	86.47 / 76.19	0.721 [0.689, 0.752]
External	77	67.53	53.25	72.22	61.68 / 48.59	0.448 [0.312, 0.581]

arrest and loan. The same pattern appears in distribution, where the largest confusions send investment (36 conversations), delivery (20), dating (15) and banking (14) conversations to legacy telecom (Appendix C). Legacy telecom is the largest class (30.7%), and the annotators rated its synthetic conversations least realistic (Section 5).

Class weighting. We also trained class-weighted variants. The encoder and Transformer variants also change label smoothing and dropout, and the Transformer variant changes input length and batch size as well (Appendix C).

The benchmark measures agreement with automatic labels. The next question is how good those labels are.

4.1 External Stress Set

The external set holds 77 YouTube-sourced scam conversations. Bilingual YouTube Data API queries (Devanagari and English, India region) returned videos between 30 seconds and 30 minutes long, 50 transcripts for each of the eight categories. Gemma 3 12B decided whether each transcript held a scammer-to-victim exchange and transliterated the Hindi parts to Roman script. The transcript text is therefore partly model-produced. A pre-filter then removed short transcripts and tutorial-style ones, and Gemma 4 31B judged the rest as conversation, re-enactment, awareness, tutorial, narration or unclear. A transcript entered the set if the judge said conversation with confidence of at least 0.75 and set the participant and back-and-forth flags. This left 77 of 400 transcripts (19.25%), between 2 and 21 per category (Appendix C). The category of each transcript is inherited from the query that found it. No separate classifier or person assigned it, so the labels are uncertain, and Section 5 measures how uncertain. The set also contains flaws we found late. One transcript has a single word repeated 16,379 times, which is 89% of its words. A partial audit found at least two items that are not scam calls, a helpline demonstration and a personal call. The set is a stress test for the models, not a sample of real scam calls.

Despite these constraints, constructing an external set is difficult. Starting from 400 retrieved transcripts, we applied multiple LLM-based screening stages to identify scammer-to-victim exchanges and then classify the remaining transcripts by conversation type before retaining only 77 items. This filtering reduced the candidate pool to 19.25%, showing the practical difficulty of obtaining external conversations that satisfy the required interaction, language, and scam-context criteria simultaneously.We do not redistribute the transcript text. The recordings were uploaded by other people, and the people in them did not agree to research use (Ethics Statement). The release holds video identifiers, query categories, judge outputs and both annotators' labels. Reviewers can obtain the transcripts through the review supplement.

5 Human Validation

Design. We drew a stratified 10% sample of the synthetic corpus (968 conversations, stratified by class) and added all 77 external conversations. The 1,045 conversations were shuffled together under blind identifiers, and the source was not shown. Annotator 1 is a fraud-detection practitioner, and Annotator 2 is a PhD student. Each worked independently from a separate printed copy of the conversations and a separate answer sheet, following a written guide with the eight definitions and tie-breaking rules (Appendix H). They chose one category, a confidence level and a realism rating, and they submitted labels before any comparison. We did not adjudicate disagreements.

Annotator 1 left 23 conversations unlabeled (18 synthetic, 5 external). The headline figures count an unlabeled conversation as an error, and Table 4 also reports accuracy on labeled items only. Kappa treats "unlabeled" as a ninth class (Cohen, 1960).

Agreement. The practitioner matches the automatic labels more often than the student does (Table 4). Agreement between the annotators is substantial on the synthetic sample ($\kappa = 0.721$) and moderate on the external set ($\kappa = 0.448$). Kappa is 0.737 and 0.487 when only labeled pairs count. The annotators agree on 787 of 1,022 conversations that both labeled, and that consensus equals the automatic label in 92.5% of cases (synthetic 92.9%, external 86.0%). The figure is 94.1% for the 528 conversations where both are also highly confident. Consensus is 100% for digital arrest and loan, and lowest for investment (81.0%), delivery (84.6%) and UPI (84.8%). Both annotators agree against the automatic label on 59 conversations, which makes these the first candidates for correction. The annotators disagree on 235 conversations. The automatic label equals the practitioner's in 67.2% of cases and the student's in 17.0%. The student also used low confidence far more often (20.5% of items against 1.3%). We report these as observed differences. They do not show which annotator is right.

Where labels are weak. Both annotators under-assign investment and dating relative to the automatic labels, and they disagree most on banking versus telecom (25 conversations), UPI versus telecom (18) and dating versus telecom (15) (Appendix E, Figure 6). Two of the three train-test near-duplicate pairs carry different labels, so the boundary between investment, delivery and telecom is fuzzy in the data itself. Annotator 1 gave the same category on 46 of 48 conversations labeled twice in a second pass over 50 items, which is 95.8% [86.0, 98.8].

Realism. We asked whether each conversation reads as a real call or as generated text. The two annotators barely agree (weighted kappa 0.29). Annotator 2 rated 94.8% of external conversations realistic against 37.7% of synthetic ones. Annotator 1 shows no such separation and rated 48.7% of external conversations as not realistic, against 35.8% of synthetic ones. Both annotators rated legacy telecom the least realistic synthetic class (81% and 36% not realistic). We report realism as an exploratory measure to provide additional context on the generated conversations. It is not intended to establish a distinction between real and synthetic conversations.

The ratings show where the labels are firm and where they are not. The next section states what that permits us to claim.

6 Limitations, Intended Use and Validity

JAAL measures how well a classifier assigns one of eight scam categories to a Hinglish conversation drawn from this generator. It supports three uses: comparing classifiers on a fixed distribution, testing how they cope with code-switching and length, and stress-testing them on the external set. The labels are automatic, and human agreement with them is 77% to 86%, so score differences of a few points are not meaningful. The external labels are inherited from search queries, and annotators match them on 53% to 68% of items. A drop on the external set therefore combines model failure, label noise and the differences in Section 3.4. The 43 external conversations on which both annotators agree form an easier subset. The SVM reaches 69.8% against their consensus label, compared with 91.0% on the synthetic test set, so the gap shrinks but remains large. The work has the following limitations :

One generator family. Gemma models generated and labeled the conversations, filtered and transliterated the external set, identified token languages.

Weak seeds and unrecorded provenance. Roughly half of the seed lines are UK SMS-spam style text. The release does not flag upscaled files. Some conversations are degenerate, such as wrong-number loops or promotions with no scam, and 59 are mostly in Devanagari script.

7 Conclusion

A corpus written by one model family can train classifiers that score above 92% on its own test set. The same classifiers fall to between 27% and 56% on a small set of conversations from another source, and the linear models mostly answer "legacy telecom". Two annotators find the labels mostly sound (86% and 77% agreement, 92.5% when they agree), with weak spots at the boundaries between investment, delivery, UPI and telecom. JAAL reports the corpus, the annotations and the stress set to work on that gap. The most useful next

steps are a larger, cleaner outside set, a split grouped by seed and generation route, and evaluation of more Hinglish-specific encoders.

ETHICS STATEMENT

No phone call was recorded for this work. Recording private calls raises consent and legal questions in India, and we make no legal determination. The synthetic conversations were generated by models, and all names, numbers and accounts in them are invented. The seeds come from a public Kaggle file. The external set was drawn from public YouTube uploads that other people recorded, so we do not redistribute transcript text. We release video identifiers, labels and a reconstruction script instead, and we share the transcripts with reviewers only through the supplement. No personal data about the annotators is released. The transcripts are spoken by real people and may contain names or numbers, which is another reason we do not redistribute them. We share only video identifiers, and users should follow the platform's terms of service and copyright rules when they retrieve the videos. The corpus is dual-use, because it describes how scams unfold. We release it for detection, defense and evaluation research, we ask users not to build scam scripts or social-engineering tools with it, and we release the generation prompts for reproducibility. We considered the ICLR Code of Ethics when preparing this release.

REPRODUCIBILITY STATEMENT

Anonymized repositories with the corpus, the annotations, the external manifest and the code accompany the submission. The code includes a script that recomputes every human-validation number from the released annotation file. We reproduced the linear-model results, the tokenizer statistics, the length and vocabulary statistics, the duplicate audit and all human-validation tables from the released files. The encoder and zero-shot results come from the authors' runs and were not re-run for this submission. Appendix G reports settings.

AI USE STATEMENT

Large language models were used to assist with grammar refinement, manuscript editing, and to generate synthetic dataset JAAL. All scientific content, experimental results, and conclusions are the sole work of the authors.

REFERENCES

Gavin Abercrombie, Tanvi Dinkar, Amanda Cercas Curry, Verena Rieser, and Dirk Hovy. Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement. https://doi.org/10.48550/arxiv.2301.10684, 2023.

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. LinCE: A centralized benchmark for linguistic code-switching evaluation. https://doi.org/10.48550/arxiv.2005.04322, 2020.

Rishi Ahuja, Kumar Prateek, and Simranjit Singh. ICFD-31k: A large-scale dataset and benchmark for real-time conversational fraud detection. https://doi.org/10.24963/ijcai.2026/786, 2026.

Tiago A. Almeida, José María Gómez Hidalgo, and Akebo Yamakami. Contributions to the study of SMS spam filtering. https://doi.org/10.1145/2034691.2034742, 2011.

Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. "I am borrowing ya mixing?" an analysis of English-Hindi code mixing in facebook. https://doi.org/10.3115/v1/w14-3914, 2014.

Samuel R. Bowman and George E. Dahl. What will it take to fix benchmarking in natural language understanding? https://doi.org/10.18653/v1/2021.naacl-main.385, 2021.

Jacob Cohen. A coefficient of agreement for nominal scales. https://doi.org/10.1177/001316446002000104, 1960.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. https://doi.org/10.18653/v1/2023.emnlp-main.183, 2023.

R. Flood, Gints Engelen, David Aspinall, and Lieven Desmet. Bad design smells in benchmark NIDS datasets. https://doi.org/10.1109/eurosp60621.2024.00042, 2024.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, and Kate Crawford. Datasheets for datasets. https://doi.org/10.1145/3458723, 2021.

Gemma Team. Gemma 3 technical report. https://doi.org/10.48550/arxiv.2503.19786, 2025.

Pitipat Gumphusiri and Tuul Triyason. Synthetic data for scam detection: Leveraging LLMs to train deep learning models. https://doi.org/10.1109/wi-iat62293.2024.00141, 2024.

Dan Hendrycks, Xiaoyuan Liu, Eric W. Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Xiaodong Song. Pretrained transformers improve out-of-distribution robustness. https://doi.org/10.18653/v1/2020.acl-main.244, 2020.

Sameed Irfan, Aswad Sheeraz, and Muhammad Hasnain. Using LLM-generated data to create a Roman Urdu scam call detector. https://doi.org/10.26710/sbsee.v7i3.3496, 2025.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. GLUECoS: An evaluation benchmark for code-switched NLP. https://doi.org/10.18653/v1/2020.acl-main.329, 2020.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. Analyzing dataset annotation quality management in the wild. https://doi.org/10.1162/coli_a_00516, 2024.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard L. Phillips, Irena Gao, Tong Lee, Étienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. https://doi.org/10.48550/arxiv.2012.07421, 2020.

Maria Kunilovskaya, Gagan Bhatia, Lisa Sophie Albertelli, Yanran Chen, Christian Greisinger, Lotta Kiefer, Christoph Leiter, Subhadeep Roy, Tewodros Achamaleh, Muhammad Arslan Manzoor, Sebastian Pohl, Yufang Hou, and Steffen Eger. Who annotates in NLP? a large-scale assessment of human annotation reporting between 2018 and 2025. https://doi.org/10.48550/arxiv.2606.02255, 2026.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. https://doi.org/10.52202/075280-2264, 2023a.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. Synthetic data generation with large language models for text classification: Potential and limitations. https://doi.org/10.18653/v1/2023.emnlp-main.647, 2023b.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Hua Ding, Gang Chen, and Haobo Wang. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. https://doi.org/10.18653/v1/2024.findings-acl.658, 2024.

Zhiming Ma, Peidong Wang, Minhua Huang, Jinpeng Wang, Kai Wu, Xiangzhao Lv, Ya-Ju Pang, Yin Yang, Wenjie Tang, and Yuchen Kang. TeleAntiFraud-28k: An audio-text slow-thinking dataset for telecom fraud detection. https://doi.org/10.1145/3746027.3755835, 2025.

Gaurav Maheshwari, Dmitry Ivanov, and Kevin El Haddad. Efficacy of synthetic data as a benchmark. https://doi.org/10.48550/arxiv.2409.11968, 2024.

Ravindra Nayak and Raviraj Bhuminand Joshi. L3Cube-HingCorpus and HingBERT: A code mixed Hindi-English dataset and BERT language models. https://doi.org/10.48550/arxiv.2204.08398, 2022.

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas Pykl, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. SemEval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. https://doi.org/10.18653/v1/2020.semeval-1.100, 2020.

Fabián Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron J. Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. https://doi.org/10.48550/arxiv.1201.0490, 2012.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Díaz. On releasing annotator-level labels and information in datasets. https://doi.org/10.18653/v1/2021.law-1.14, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? https://doi.org/10.48550/arxiv.1902.10811, 2019.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. https://doi.org/10.52202/079017-0685, 2024.

Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J. Mielke, Cibu Johny, Işin Demirşahin, and Keith Hall. Processing South Asian languages written in the Latin script: the Dakshina dataset. https://doi.org/10.48550/arxiv.2007.01176, 2020.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. https://doi.org/10.48550/arxiv.2012.15613, 2020.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. https://doi.org/10.18653/v1/p16-1162, 2016.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. https://doi.org/10.1038/s41586-024-07566-y, 2024.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W. Black. A survey of code-switched speech and language processing. https://doi.org/10.48550/arxiv.1904.00784, 2019.

Weixiang Sun, Shang Ma, Yiyang Li, Tianyi Ma, Zehong Wang, Colby Nelson, Xusheng Xiao, and Yanfang Ye. PreScam: A benchmark for predicting scam progression from early conversations. https://doi.org/10.48550/arxiv.2605.12243, 2026.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. https://doi.org/10.48550/arxiv.2007.00644, 2020.

Ethan Traister, Ankit Raj, Jiaqi Gan, Xingyu Shen, Tyler Wu, Yuchen Zhou, Tommy Duong, Kidus Zewde, Siying Chen, and Simiao Ren. Anatomy of a scam call: What 10,000 real scam and spam calls reveal about how phone scammers operate. https://doi.org/10.48550/arxiv.2608.24127, 2026.

Huahong Tu, Adam Doupé, Ziming Zhao, and Gail-Joon Ahn. SoK: Everyone hates robocalls: A survey of techniques against telephone spam. https://doi.org/10.1109/sp.2016.27, 2016.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. https://doi.org/10.18653/v1/2023.acl-long.754, 2023.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. Are multilingual models effective in code-switching? https://doi.org/10.18653/v1/2021.calcs-1.20, 2021.

I. G. Wood, Michal Kepkowski, Leron Zinatullin, Travis Darnley, and Mohamed Ali Kâafar. An analysis of scam baiting calls: Identifying and extracting scam stages and scripts. https://doi.org/10.48550/arxiv.2307.01965, 2023.

A DATASHEET

Motivation. JAAL supports research on eight-way scam-category classification of Hinglish conversations and on how models trained on generated conversations behave on other sources. Its authors are a research group at an Indian institution (anonymized for review). No funding statement is reported here for the same reason.

Composition. The release holds 9,677 synthetic conversations (Table 2), one text file each, with one turn per line prefixed `caller:` or `receiver:`. Conversations average 320 words. It also holds a manifest of 77 external items (video identifier, query category, judge output, length) and a table of both annotators' labels for 1,045 conversations. Names, numbers and accounts in synthetic conversations are invented. Some conversations are degenerate or in Devanagari script (Section 6).

Collection and preprocessing. Sections 3 and 4.1 describe generation, labeling, upscaling, removal of 214 conversations, splitting and external-set construction. Speaker labels were normalized to `caller` and `receiver`. No other cleaning was applied to the synthetic text.

Intended and discouraged uses. Intended: classifier comparison, code-switching robustness, label-quality and synthetic-data studies. Discouraged: estimating real-world scam detection accuracy, training generators of scam scripts, and any use that adds personal data.

Distribution and license. The synthetic data, manifest and annotations are released under CC BY-NC 4.0, and the code under the MIT license. The seed file is third-party data, and users should check its terms. Transcript text of the external set is not distributed.

Maintenance. Version 1.0 is the state described here. Corrections to the labels (for instance the 59 conversations on which both annotators disagree with the automatic label) and provenance flags for upscaled files would be released as new versions with a change log. The maintainers can be reached through the repository issue tracker.

B PROMPTS

The prompts below are shortened. The code release holds the exact text.

Caller agent.

```
Act as a caller with this specific context: {call_context}. Current
History: {history}. Turns remaining: {responses}. Rules: (1)
Language: Hinglish; tone should match the context. (2) Output only a
valid JSON: ``message'': ``your response''. (3) No placeholders;
invent a realistic Indian name or detail. (4) If the history is empty,
start with a greeting. (5) As {responses} gets lower, try to wrap up
or reach the goal of the context.
```

Receiver agent.

```
Act as a person who just picked up a call from an unknown number. Your
language must be Hinglish (natural mix of Hindi and English). History:
{history}. Rules: (1) Output only a valid JSON: ``message'': ``your
response''. (2) Be organic; use fillers such as ``uhm'', ``haan'',
``boliye''. (3) No placeholders; ask ``Aap kaun bol rahe ho?'' or
assume a common name. (4) If the history is empty, say something like
``Hello? Kaun?''.
```

Upscaling. The prompt names the scam type with a description, shows two or three example conversations of the class, and asks for one new conversation. The caller speaks first, each turn is labeled, and five to six exchanges are requested. The new conversation must differ from the examples in scenario details, names, amounts, platforms and tactics.

Zero-shot classification. The system message lists the eight categories with definitions and asks the model to answer only with a JSON object {"label": "<category>"}. The user message wraps the conversation between begin and end markers. Jev receives the same definitions as criteria through its structured choice endpoint. Two oversized transcripts keep about 12,000 characters from each end.

C Additional Statistics

Table 5 presents the composition of the external set, and Table 6 its length by category. Table 7 presents language composition and Table 8 tokenizer statistics. Table 9 compares the test set with the external set. Table 10 lists per-class F1 for the linear models, and Table 11 lists the class-weighted variants. Table 12 reports Jev results by class. Figure 4 shows length by class, Figure 5 the errors of the linear models, and Figure 6 per-class annotator agreement and realism ratings.

Table 5: External set composition.

Category	Retrieved	Accepted	Accept. (%)	Share (%)
Legacy telecom	50	21	42.00	27.27
Banking / KYC / OTP fraud	50	17	34.00	22.08
Investment / task scam	50	12	24.00	15.58
UPI / wallet fraud	50	9	18.00	11.69
Delivery / customer care	50	6	12.00	7.79
Digital arrest / govt. impersonation	50	5	10.00	6.49
Dating / romance / sextortion	50	5	10.00	6.49
Loan / credit app scams	50	2	4.00	2.60
Total	400	77	19.25	100.00

Table 6: Word length by category.

	External			Synth. train	Ext. >2,048 tok. (%)	
Category	N	Mean	Median	Mean	RoBERTa	ModernBERT
Legacy telecom	21	1,452	1,208	417	71.43	66.67
Banking / KYC / OTP	17	963	418	327	47.06	41.18
Investment / task	12	2,605	1,005	334	75.00	75.00
UPI / wallet	9	1,200	1,044	203	66.67	55.56
Delivery / care	6	640	538	267	16.67	16.67
Digital arrest / govt.	5	1,408	432	262	60.00	40.00
Dating / romance	5	985	933	318	60.00	40.00
Loan / credit app	2	696	696	215	50.00	50.00

The investment mean of 2,605 words is driven by the transcript with the repeated word. The mean falls to 1,240 words once that loop is collapsed.

The largest test confusions send investment, delivery, dating and banking conversations to legacy telecom (logistic regression 36, 20, 15 and 14; SVM 32, 18, 15 and 10). The two models predict legacy telecom for 62 and 56 of the 77 external conversations.

* Label smoothing 0.05, dropout 0.15 (from 0.1), input length 2,048 (from 4,096), batch size 32 (from 64). † Label smoothing 0.05 and classifier dropout 0.15. Only the two linear rows change class weights alone, so the other rows do not isolate the effect of weighting.

Jev matches GPT-OSS 120B on accuracy (66.23%). Both models were queried through an API and we did not benchmark latency, so the paper makes no speed comparison. Its Macro-F1 is 3.29 points below GPT-OSS 120B and 8.15 points below Gemma 4 31B.

Figure 4 shows that external conversations are longer than synthetic ones in every class. The digital-arrest box is wide because that class has only five external items. Figure 5 shows that the test errors of both linear models collapse into legacy telecom (class 8), mostly from investment (class 3), and that both models send most external conversations to class 8. Figure 6 shows the annotators agree most with the automatic labels on

Table 7: Language composition (%).

	Train	Val	Test	External
Conversations	6,773	1,452	1,452	77
Token occurrences				
English	52.64	50.46	51.79	11.84
Hindi	42.30	44.50	42.23	82.43
Ambiguous	0.30	0.23	0.56	2.73
Other	4.76	4.81	5.42	3.00
Conversation-level label				
Hinglish mixed	69.98	72.73	70.45	57.14
English dominant	29.44	26.52	29.06	0.00
Hindi dominant	0.44	0.62	0.21	42.86
Other dominant	0.13	0.14	0.28	0.00

Table 8: Subword statistics.

Tokenizer	Metric	Synth. train	Synth. test	External
RoBERTa	Mean encoder tokens	607.05	610.32	3,309.87
	Mean subwords per word	1.52	1.52	1.78
	Words split $\geq$2 (%)	40.14	40.29	61.66
	Conv. >1,024 tok. (%)	10.94	11.64	76.62
	Conv. >2,048 tok. (%)	0.65	0.69	59.74
	Conv. >4,096 tok. (%)	0.19	0.14	22.08
ModernBERT	Mean encoder tokens	595.41	597.81	2,818.62
	Mean subwords per word	1.51	1.51	1.74
	Words split $\geq$2 (%)	40.09	40.19	58.60
	Conv. >1,024 tok. (%)	10.69	11.16	75.32
	Conv. >2,048 tok. (%)	0.32	0.48	53.25
	Conv. >4,096 tok. (%)	0.00	0.07	18.18

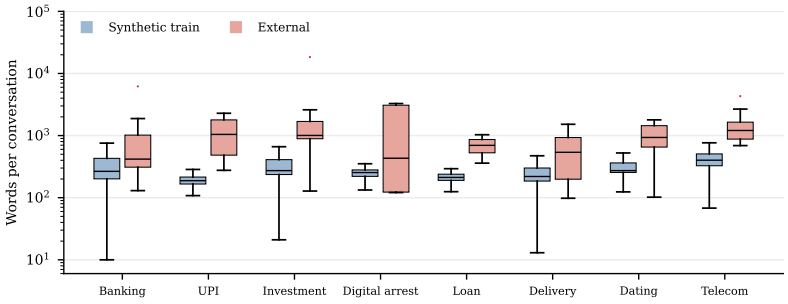


Figure 4: Words per conversation by class.

loan and digital arrest, and least on investment. Annotator 2 rates external conversations as realistic far more often than synthetic ones, while Annotator 1 does not.

D Dataset Audit

Table 13 lists the audit checks. The three train-test pairs are 3_000455 (train) with 6_000288 (test), 8_001042 (train) with 3_000150 (test), and 3_000455 (train) with 3_000457 (test). The first two carry different labels.

Table 9: Synthetic test set versus external set.

Property	Synth. test	External
English / Hindi tokens (%)	51.79 / 42.23	11.84 / 82.43
Hindi-dominant conversations (%)	0.21	42.86
Unique word types	10,799	7,133
Unique-type OOV vs. train (%)	17.00	50.58
Mean per-conversation OOV (%)	0.63	6.56
Repeated-token rate (%)	46.42	60.07
Mean words per conversation	320	1,378 (1,165 collapsed)
Median words per conversation	273	915
Maximum words per conversation	2,282	18,381 (6,201 collapsed)
RoBERTa subwords per word	1.52	1.78
RoBERTa conv. >2,048 tokens (%)	0.69	59.74
ModernBERT conv. >2,048 tokens (%)	0.48	53.25

Table 10: Per-class F1 of the linear models.

		Bank.	UPI	Inv.	Dig.	Loan	Deliv.	Dating	Telec.
Test	LR	90.1	95.7	73.1	97.9	97.4	87.7	90.7	89.6
	SVM	91.3	96.5	73.1	98.3	97.8	87.9	90.3	90.1
External	LR	66.7	0.0	0.0	0.0	0.0	66.7	28.6	50.6
	SVM	69.0	20.0	15.4	0.0	66.7	66.7	25.0	54.5

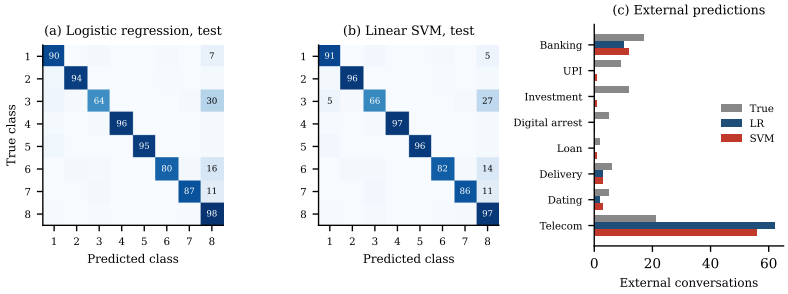


Figure 5: Linear-model errors on the test and external sets.

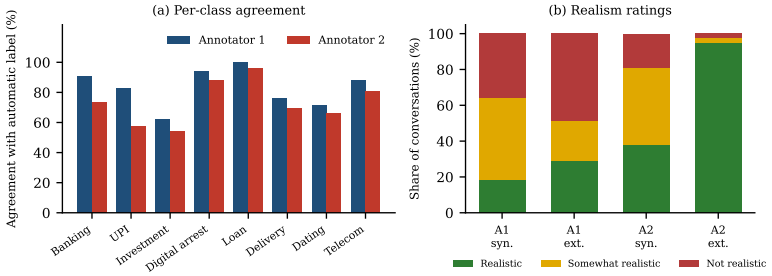


Figure 6: Annotator agreement and realism ratings.

E HUMAN VALIDATION DETAIL

Table 14 reports per-class agreement of each annotator with the automatic labels, pooled over 1,045 conversations. Table 15 shows where the two annotators disagree.

Label counts (automatic, A1, A2) for classes 1 to 8 are 148/179/172, 104/104/84, 92/62/67, 102/101/96,

Table 11: Class-weighted variants.

	Synthetic test		External	
Variant	Acc.	M-F1	Acc.	M-F1
Logistic regression, weights only	91.53	91.30	48.05	30.21
Linear SVM, weights only	91.53	91.25	51.95	41.42
Transformer, weights + changes*	88.22	88.04	55.84	54.86
RoBERTa, weights + changes†	92.22	91.99	59.74	50.87
ModernBERT, weights + changes†	92.36	91.99	57.14	53.23

Table 12: Jev results by class on the external set.

Category	Prec.	Rec.	F1	Supp.
Banking / KYC / OTP fraud	0.800	0.706	0.750	17
UPI / wallet fraud	0.700	0.778	0.737	9
Investment / task scam	0.333	0.083	0.133	12
Digital arrest / govt.	0.500	0.600	0.545	5
Loan / credit app scams	0.667	1.000	0.800	2
Delivery / customer care	0.714	0.833	0.769	6
Dating / romance / sextortion	0.500	0.200	0.286	5
Legacy telecom	0.714	0.952	0.816	21
Macro average	0.616	0.644	0.605	77

95/97/97, 91/90/102, 95/71/76 and 318/318/351. Confidence tracks accuracy for both annotators: Annotator 1 is 89.3% accurate at high confidence and 63.2% at medium, and Annotator 2 is 81.4% and 60.6%. Two annotators with one second pass by one of them is a small design, and we report intervals for it (Section 5).

F Benchmark Checklist

Table 16 lists benchmark practices (Reuel et al., 2024; Bowman and Dahl, 2021) and how far JAAL meets them. It is our own list, not a mapping to any published checklist.

G Reproducibility Details

Linear models use scikit-learn (Pedregosa et al., 2012): TF-IDF with a lowercasing tokenizer that keeps Devanagari runs, Latin words and single symbols, unigrams and bigrams, 50,000 features, sublinear term frequency, fitted on the training set only. Logistic regression uses C = 5.0 with the saga solver and 1,000 iterations. The linear SVM uses C = 1.0 and 2,000 iterations.

The scratch Transformer has three pre-norm encoder layers with hidden size 128, four heads and feed-forward size 256, sinusoidal positions, a custom BPE vocabulary of 7,996 tokens (Sennrich et al., 2016), a classification head on the first token, and 1,422,216 trainable parameters. It trains with AdamW (learning rate 3×10^{-4}, weight decay 0.01), batch size 64, input length 4,096, at most 20 epochs and patience 6.

RoBERTa-base and ModernBERT-base are fully fine-tuned with AdamW (learning rate 2×10^{-5}, weight decay 0.01), batch size 16, at most 10 epochs, patience 4, linear warmup over 10% of steps followed by cosine decay, and gradient clipping at 1.0. Input lengths are 512 and 2,048 tokens. HingBERT uses the checkpoint l3cube-pune/hing-bert with a 512-token input and the same recipe (seed 42). All three use the default classification head of their Hugging Face implementations.

Zero-shot models run through Ollama with temperature 0, seed 42 and JSON output. GPT-OSS 120B and Gemma 4 31B use cloud-hosted model tags queried from a laptop, and the scripts do not set a context window. Jev uses its structured decision endpoint. Training of the neural models ran on an NVIDIA H200 GPU with MIG partitioning. The authors ran the neural and zero-shot experiments, and the released environment file lists package versions.

Table 13: Dataset audit checks.

Check	Result
Exact and normalized duplicate groups (released set)	0 and 0
Near-duplicate pairs, character 3–5-gram TF-IDF, cosine ≥ 0.90	11 pairs, 6 clusters, largest 3; 2 pairs cross class
Train-test near-duplicate pairs at 0.90	3 (cosine 0.9232, 0.9160, 0.9118)
Word TF-IDF pairs at cosine ≥ 0.70 (all 9,677)	816; 390 cross-split; 113 cross-label (13.8%)
Word TF-IDF pairs at cosine ≥ 0.80	176; 86 cross-split; 17 cross-label (9.7%)
Test to nearest train, median cosine	0.25; 7.3% at 0.70 or higher
Distinct-1 / 2 / 3	0.0080 / 0.1336 / 0.3775
Random-pair Jaccard overlap of 1 / 2 / 3-gram sets	13.31% / 1.75% / 0.31%
Removed before release	214 of 9,891; 51 exact duplicates of a retained file; 82 at cosine ≥ 0.90

Table 14: Per-class agreement with the automatic labels.

		Annotator 1			Annotator 2		
Category	Supp.	P	R	F1	P	R	F1
Banking / KYC / OTP	148	0.749	0.905	0.820	0.634	0.737	0.681
UPI / wallet	104	0.827	0.827	0.827	0.714	0.577	0.638
Investment / task	92	0.919	0.620	0.740	0.746	0.544	0.629
Digital arrest	102	0.951	0.941	0.946	0.938	0.882	0.909
Loan / credit app	95	0.979	1.000	0.990	0.938	0.958	0.948
Delivery / customer care	91	0.767	0.758	0.762	0.618	0.692	0.653
Dating / romance	95	0.958	0.716	0.819	0.829	0.663	0.737
Legacy telecom	318	0.884	0.884	0.884	0.729	0.805	0.765

H ANNOTATION GUIDE (SUMMARY)

Annotators read a conversation once and chose the single category that best matched the main scam mechanism and the main action the caller wanted, not isolated keywords or brand names. They did not search for the call, use model predictions, or change labels to balance classes. Ambiguous cases followed tie-breaking rules, for example: digital arrest wins over banking when legal or police threats dominate, and dating wins over financial categories when the relationship or blackmail is central. Table 17 lists the category definitions.

Table 15: Annotator 1 versus Annotator 2 label counts.

A1 \ A2	1	2	3	4	5	6	7	8
1 Banking	124	11	2	2	1	14	0	25
2 UPI	15	59	1	0	0	9	2	18
3 Investment	1	4	49	0	0	1	2	5
4 Digital	2	1	2	91	0	1	1	3
5 Loan	1	0	1	0	92	1	2	0
6 Delivery	12	1	2	1	1	61	0	12
7 Dating	1	2	3	1	0	1	48	15
8 Telecom	14	6	7	0	2	12	14	263

Table 16: Benchmark practice checklist.

Practice	Status	Where or why
Task and scope stated	Yes	Section 6
Data and code released	Yes	Reproducibility Statement
Annotation protocol, agreement, intervals	Yes	Section 5
Intra-annotator agreement	Partly	50 items, one annotator
Statistical uncertainty on results	Partly	Wilson intervals; bootstrap only for linear models
Every reported number reproducible	Partly	Encoder and LLM runs were not re-run
Contamination and label-conflict audit	Partly	Near duplicates yes; seed-level no
Limitations and ethics stated	Yes	Section 6, Ethics Statement
Maintenance plan	Partly	Datasheet, version 1.0
Multiple seeds	No	Single seed

Table 17: Category definitions provided to annotators.

Category	Definition
1 Banking / KYC / OTP	Bank impersonation involving KYC, account verification, OTPs, card issues, account blocking or banking credentials
2 UPI / wallet	Scams built on UPI, wallets, QR codes, UPI PINs, collect requests or payment-app transactions
3 Investment / task	Investment, trading, crypto, job or task schemes that promise earnings for deposits or paid tasks
4 Digital arrest / govt.	Police, court, tax, customs or other official impersonation with legal or criminal threats
5 Loan / credit app	Fake loans, instant credit, processing fees, fraudulent loan apps, harassment
6 Delivery / customer care	Courier, e-commerce, order, refund or helpdesk scams, including remote access
7 Dating / romance / sextortion	Romantic manipulation, intimate-content threats, blackmail
8 Legacy telecom	SIM, mobile number, recharge, lottery, tower or utility claims that fit no other category