SCAMBENCH: A Multi-Dimensional Scam Benchmark through Human-AI Collaborative Data Construction

Anonymous ACL submission

Abstract

Online scams are a growing form of digital harm, yet scam research remains constrained by fragmented datasets and limited annotation depth. We introduce SCAMBENCH, a benchmark of 1,836 authentic scam messages curated from victim-reported incidents across six scam types: employment, investment, imposter, sweepstakes/lottery/prizes, tech support, and advance fee loan scams, along with 1,834 real-world non-scam examples from legitimate sources. SCAMBENCH provides multi-dimensional annotations that capture persuasion strategies from the scammer perspective, red flags from the user perspective, and severity levels reflecting potential harm and urgency. These annotations are produced through a human-AI collaborative pipeline that uses expert-labeled examples to calibrate and validate LLM-assisted annotation at scale. Our analysis shows that these signals vary significantly across scam types, highlighting substantial scam-type heterogeneity. We further use annotated evidence spans to generate persuasion-removed hard scam examples by rephrasing explicit persuasion and red-flag cues. Evaluating multiple LLM-based detectors on original and rewritten scams shows that several models perform worse on rewritten examples, suggesting sensitivity to surface-level cues. A qualitative case study identifies five recurring failure patterns. SCAMBENCH provides a resource for scam detection, robustness evaluation, and scam education.

1 Introduction

Online scams represent one of the fastest-growing forms of digital harm, affecting millions of individuals annually and leading to billions of dollars in financial losses (Federal Trade Commission, 2025). Despite this growing need, data remains a major bottleneck for scam research. Existing scam datasets are often fragmented, domain-specific, and limited in either coverage or annotation depth (Boumber et al., 2024; Vidros et al., 2017; 419scam.org, 2026; Ma et al., 2025). Few datasets are built from authentic victim-reported incidents, despite the availability of scam-report platforms covering diverse contemporary scams. Raw scam reports are difficult to use directly because they contain free-form narratives, duplicate submissions, inconsistent terminology, and sensitive information, and annotation-intensive scam content that is difficult to label consistently at scale (Ma et al., 2025). Another recurring limitation is the lack of non-scam counterparts, which makes it difficult to build balanced benchmarks for detection and comparative modeling.

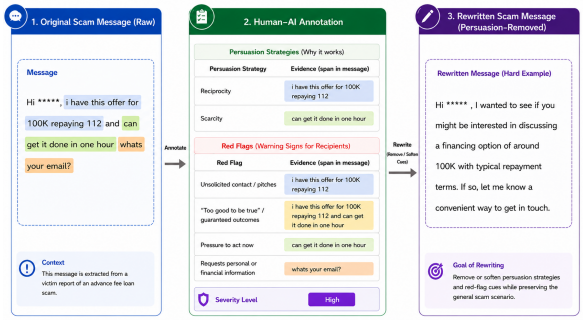


Figure 1: An example of the scam message, annotations, and rewritten hard example from our SCAMBENCH.

To address these gaps, we develop a **human-AI collaborative curation pipeline** for transforming noisy victim-reported scam reports into structured benchmark data (Figure 2). The pipeline combines human-labeled examples, AI-assisted content filtering, semantic deduplication, manual cleaning, evidence-based annotation, and scale high-quality labeling. Using this pipeline, we introduce SCAMBENCH, a new benchmark that contains 1,836 scam messages spanning six common scam types: *employment scams*, *investment scams*, *imposter scams*, *sweepstakes/lottery/prizes scams*, *tech support scams*, and *advance fee loan scams*, together with 1,834 real-world non-scam counter-

Figure 2: The human-AI collaborative scam data curation (Stages 1-8) and annotation (Stage 9) pipeline.

parts drawn from legitimate sources aligned with these categories. Each scam message is annotated along three dimensions: (1) persuasion strategies grounded in persuasion theory, (2) user-facing red flags drawn from consumer-protection guidance, and (3) severity levels reflecting potential harm. Every label is linked to a verbatim evidence span in the message, which makes controlled rephrasing of cues possible. We further use the annotation schema to construct a robustness-oriented evaluation set by generating persuasion-removed hard scam examples (Figure 1).

Our analysis of the three-dimensional scam annotations shows two key findings. First, scam communication is not uniform across categories: persuasion strategies, red flags, and severity levels vary systematically by scam type, with chi-squared tests confirming significant associations in all three dimensions. A scam-type-agnostic approach to dataset construction or detection may obscure these differences. Second, when we rephrased annotated cues, multiple LLM-based detectors fail to classify the rewritten messages as scams. A qualitative analysis identifies five recurring failure patterns, including cases where adding legitimate-sounding markers such as references to official channels or risk disclaimers actively flips the model's judgment.

Our contributions are:

- **Pipeline and dataset.** A human-AI collaborative curation pipeline that converts noisy victim reports into structured benchmark data, applied to produce SCAMBENCH: 1,836 scam messages across six scam types with 1,834 aligned non-scam counterparts.
- **Three-dimensional evidence-grounded annotation schema.** Persuasion strategies (scammer-side), red flags (user-side), and severity levels (harm-side), each linked to verbatim evidence spans, enabling analyses beyond binary scam labels. Signal distributions vary significantly across scam types.
- **Annotation-guided robustness evaluation.** A method for generating persuasion-removed hard scam examples by rephrasing annotated evidence spans, applied to five LLM-based detectors. Results reveal substantial recall degradation and five recurring failure patterns, suggesting that current detectors rely on surface cues rather than scam structure.

2 Background and Related Work

2.1 Online Scams and Their Societal Impact

Online scams, such as fraudulent schemes conducted via email, websites, or messages, leverage social engineering to exploit human psychology and manipulate victims for malicious ends (Button and Cross, 2017). They represent a growing digital threat that inflicts substantial financial, psychological, and social harm (Nurse, 2018; Whitty and Joinson, 2008; Whitty, 2015). In 2024, reported U.S. consumer fraud losses exceeded $12.5 billion (Federal Trade Commission, 2025). Around 73% of U.S. adults report having encountered some form of online scam, many on a weekly basis (Pew Research Center, 2025).Scams typically employ sophisticated manipulation strategies targeting human trust, emotion, and behavioral tendencies (Whitty, 2013). Common forms include romance scams (Rege, 2009; Whitty, 2013), advance fee fraud (Chang, 2008; Holt and Graves, 2007), and investment scams (Siu et al., 2022; Anderson et al., 2024). Critically, scams are dynamic and adaptive, unfolding across messaging apps, social media, email, and phone through extended interactions where deceptive cues emerge across discourse, identity construction, and temporal patterns (Li et al., 2025; Button et al., 2014). Addressing this threat requires approaches that capture psychological tactics, social context, and behavioral dynamics, supported by data infrastructures capable of modeling these complex narratives at scale.

2.2 Challenges in Scam Datasets

Existing scam datasets provide useful starting points but remain limited in coverage or annotation depth. For example, EMSCAD (Vidros et al., 2017) focuses on employment scam, 419 scam archives (419scam.org, 2026) collect advance-fee emails without standardized annotations, and

Table 1: Comparison of SCAMBENCH with existing scam datasets.

Dataset	Multiple Scam Types	Real-World Data	Non-Scam Data	Annotated Scam Cues	Annotated Severity	Hard Examples
EMSCAD (Vidros et al., 2017)		✓	✓			
419 Scam (419scam.org, 2026)	✓	✓				
DIFraud (Boumber et al., 2024)	✓	✓	✓			
PsyScam (Ma et al., 2025)	✓	✓		✓		✓
ScamGen (Han et al., 2025)	✓					✓
Fraud-R1 (Yang et al., 2025)	✓	✓				✓
SCAMBENCH (Ours)	✓	✓	✓	✓	✓	✓

PsyScam (Ma et al., 2025) introduces psychological annotations but lacks non-scam counterparts. As summarized in Table 1, these gaps reveal a fundamental need for a structured dataset that captures the diversity, scale, and evolving nature of online scams across platforms.

Closing these gaps at scale is not feasible through fully manual curation. Recent work shows that LLMs can substantially reduce this burden by extracting attributes and labeling cues in collaboration with human experts, achieving high agreement with human coders while improving annotation efficiency (Park et al., 2025). The complementary strengths are particularly promising to scam annotation: LLMs scale across large corpora and produce evidence-grounded labels, while human experts ground the label schema in domain frameworks and ambiguous cases. We adopt this human-AI collaborative paradigm throughout our curation and annotation pipeline.

3 SCAMBENCH Construction Process

3.1 Scam Report Platform Selection

We used scam-report platforms to obtain broader scam coverage and more up-to-date fraud activity than existing benchmark datasets. Table 4 compares the major scam-report platforms considered in this study. We selected BBB Scam Tracker because it provides broad cross-category coverage and a large volume of victim-reported scam cases across our six target categories: : *Employment scam*, *Investment scam*, *Imposter scam*, *Lottery/sweepstakes/prize scam*, *Tech support scam*, and *Advance fee loan scam*.

3.2 AI-Assisted Scam Data Curation

Figure 2 illustrates our data curation pipeline. We first collected reports from BBB Scam Tracker and performed exact-match deduplication. An entry was treated as a duplicate only when `scam type`, `description`, `date_reported`, `business_name`, `location`, and `dollars_lost` were all identical. After deduplication, the total number of scam reports was reduced from 252,981 to 50,492. Next, since BBB reports contained the original scam messages, victim narratives, or a mix of both, we introduced a content-type filtering stage. We manually coded the first 200 entries for each scam type using a variable named `content_type`, with three classes: *scammer message*, *victim narrative*, and *both*. Example annotations for each class are shown in Appendix A.

Next, we employed two large language model candidates, Qwen3-Next-80B-A3B-Instruct and Llama-4-Scout-17B-16E-Instruct, to annotate the same set of 200 entries for each scam type. Both models were deployed on a university high-performance computing cluster using four A100 GPUs. Their annotations were compared against the human-labeled ground truth to evaluate performance. To improve classification accuracy, we designed structured prompts that incorporate domain knowledge about scam communications (e.g., definitions of scam, role prompting, and few-shot examples, see full prompts used in this paper provided in Appendix C).

Our preliminary experiments show that Llama-4-Scout-17B-16E-Instruct outperformed Qwen3-Next-80B-A3B-Instruct across all scam types, achieving an overall accuracy of 81.7% compared with 66.8% (see Table 9 in the Appendix C for details). Hence, we move forward with evaluating different prompting strategies for Llama-4-Scout-17B-16E-Instruct, including zero-shot prompting without scam definitions, zero-shot prompting with scam definitions, and definition-based few-shot prompting. As shown in Table 10 in Appendix C, overall accuracy improved from 69.5% to 76.0%, and further to 81.7%, indicating the benefit of structured prompting. We further explored scam-type-specific few-shot configurations by varying the number of prompt examples for each content category. As summarized in Table 11, varying the number of examples provided in the prompt

improved overall accuracy from 80.8% for the 3-shot baseline to 82.3%, while selecting the best-performing configuration for each scam type further improved accuracy to 84.3%. Based on these results, we adopted the best-performing configuration for each scam type, yielding 13,477 original scam messages.

We then applied a multi-stage filtering pipeline consisting of: (1) semantic deduplication using a Sentence Transformer (Reimers and Gurevych, 2019) with a cosine similarity (Salton et al., 1975) threshold of 0.8; (2) removal of short messages using the global 35th percentile word-count threshold of 20 words; (3) manual removal of messages containing narrative text or garbled content, while also correcting BBB misclassified scam categories; (4) a second semantic deduplication pass with a threshold of 0.92; and (5) a second round of manual removal for garbled messages. The detailed data processing pipeline and the number of samples retained at each stage are summarized in Table 5. This process (Figure 2) yielded a final corpus of 1,836 scam messages across six scam types.

3.3 Non-Scam Dataset Curation

To enable comparative analysis, we constructed a non-scam dataset, corresponding to each scam type: lottery-related content from official lottery websites (SweepsAdvantage, 2026; Hoosier Lottery, 2026; New York Lottery, 2026); technical support data from TechQA-RAG-Eval (NVIDIA, 2024); employment data from job postings on Indeed (Indeed, 2026); imposter-related examples from official institutional communications (e.g., IRS, SSA, USCIS) (Internal Revenue Service, 2026; Social Security Administration, 2026; U.S. Citizenship and Immigration Services, 2020) and legitimate banking emails; and investment-related samples from financial communications. As advance fee loan scams lack a legitimate counterpart, this category is excluded. The final non-scam dataset contains 1,834 samples (Table 6).

3.4 Human-AI Collaborative Scam Data Annotation

Next, we annotated 1,836 scam messages using a human-AI collaborative workflow that combines expert annotation with AI-assisted labeling. The annotation schema includes three attributes: *persuasion strategies*, *scam red flags*, and *severity level*.

Persuasion Strategies We adopted Cialdini's seven persuasion principles (Cialdini, 2016) to characterize manipulation tactics: **reciprocity** (offering rewards or favors to create obligation), **commitment and consistency** (escalating from small to larger requests), **social proof** (claiming others have benefited), **authority** (invoking institutions or impersonating authority), **liking** (using friendliness or emotional bonding), **scarcity** (emphasizing urgency or limited availability), and **unity** (appealing to shared identity). These principles capture social engineering strategies used in scam communications from the scammer's perspective.

Red Flags We defined nine scam warning indicators from the victim's perspective, based on publicly available guidelines from consumer protection agencies, financial regulators, and industry resources (Table 7). These include **unsolicited contact or pitches**, **"too good to be true" claims or guaranteed outcomes**, **pressure to act now**, **requests for personal or financial information**, **upfront fees or payments required**, **payment via gift card, wire, or crypto**, **posing as a government or official organization**, **Threats of arrest or legal action**, and **grammar, spelling, or language issues**. Table 8 summarizes typical red flag patterns for each scam type as reference signals.

Severity Level We applied four scam severity levels inspired by established risk assessment and threat-intelligence frameworks in cybersecurity and information systems (FIRST.org, 2019; National Institute of Standards and Technology, 2018). These levels capture the potential harm and urgency of a scam interaction: **low** (mild suspicious cues; no direct request for money or sensitive data), **medium** (some manipulative tactics; early-stage scam; limited harm if ignored), **high** (clear attempt to obtain money, credentials, or access; strong manipulation), and **critical** (imminent financial loss, coercion, impersonation of authority, or threats).

Human-AI Collaborative Scam Annotation Process We adopted a human-AI collaborative workflow to balance annotation quality and scalability. First, we selected first 10 entries of six scam types (n = 60) to annotate as human ground truth labels. Then, we used 20% of the data to train two expert human coders (two authors) to annotate persuasion strategies, red flags, and severity level. Next, using a 10% of the human-annotation data, we measured agreement between two human annota-

tors, which was 0.72 for persuasion strategies, 0.68 for red flags, and 0.77 for severity level. For *persuasion strategies* and *scam red flags*, which are multi-label classification tasks, we used Jaccard Similarity (Jaccard, 1901) to measure the overlap between predicted and ground-truth label sets. For *severity level*, which is an ordinal classification task, we adopted Quadratic Cohen's Kappa (Cohen, 1968) to account for the ordered nature of severity labels. Next, one human coder annotated the rest of the data to use as high-quality ground truth labels. Among human-annotated 10 scam examples from each scam type, 3 entries were selected as few-shot examples, and the remaining 7 were annotated by multiple LLM candidates (ChatGPT 5.2, Gemini 3 Flash, and Claude Sonnet 4.6) to evaluate their performance against human annotations. To ensure consistent and high-quality outputs, we designed a structured prompting strategy incorporating domain knowledge and annotation constraints (see Appendix C).

Human-LLM Agreement in Scam Annotation To evaluate the reliability of LLM-based annotation, we measured agreement between human annotations and model outputs across three tasks, using the same agreement measures as we did for human-human agreement. Results are summarized in Tables 12-14. Gemini 3 Flash and Claude Sonnet 4.6 outperformed ChatGPT 5.2 on multi-label tasks, with human-AI agreement scores of 0.73 and 0.72 for persuasion strategies, 0.72 and 0.72 for red flags, and Claude achieving the highest agreement on severity classification (0.77). Based on these results, we selected Claude Sonnet 4.6 to annotate the rest of our scam dataset.

4 Scam Signal Analysis

Next, we analyzed the annotated scam messages to characterize how persuasion strategies, red flags, and severity levels vary across scam types. Unless noted, averages are per message. Since persuasion strategies and red flags are multi-label annotations, a single message may contain several of each. To test whether differences across scam types are statistically meaningful, we ran chi-squared tests on the distributions of persuasion strategies, red flags, and severity levels, reporting Cramer's V as an effect-size measure. For the multi-label annotations (persuasion strategies and red flags), the test examines whether label assignments differ across scam types; for severity, it examines whether ordinal severity labels are distributed differently.

4.1 Signal Density Across Scam Types

Across the 1,836 curated scam messages, there are an average of 2.76 persuasion strategies and 3.92 red flags per message. Red flags are more frequent than persuasion strategies across all scam types Figure 4 shows substantial variation across scam types. Some scams, such as imposter and tech support scams, are dense in warning signs because they often combine institutional impersonation, urgency, suspicious contact methods, and requests for action. Others, such as employment and lottery scams, rely more heavily on social and motivational framing, including personal address, reward promises, and authority cues.

4.2 Scam-Type Heterogeneity

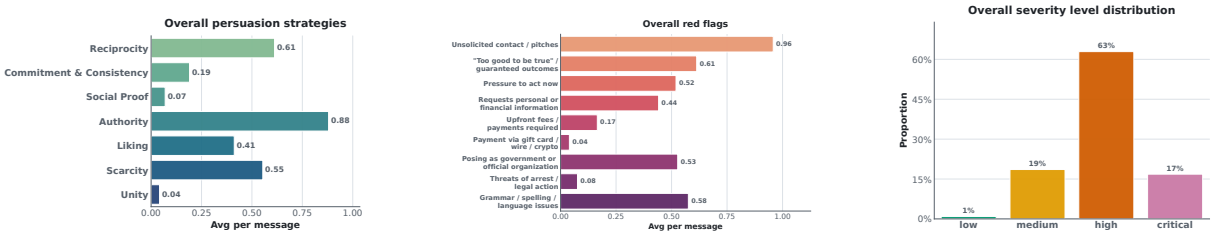
As shown in Table 15 in Appendix C, all three tests are significant at $p < .001$, indicating that persuasion strategies, red flags, and severity levels vary systematically across scam types. The effect size is largest for severity (Cramer's $V = 0.3403$), suggesting that scam type is especially informative for understanding potential harm and urgency. Persuasion strategies and red flags show smaller but meaningful associations with scam type (Cramer's $V = 0.2007$ and 0.1738, respectively), consistent with the descriptive patterns shown above.

4.2.1 Persuasion Strategies

Figure 3a presents the overall distribution of persuasion strategies. Across the full dataset, *authority* is the most frequent strategy (0.88 labels per message), reflecting the central role of institutional, occupational, or organizational impersonation in scam communication, followed by *reciprocity* (0.61), *scarcity* (0.55), and *liking* (0.41). *Social proof* and *unity* are rare overall (≤ 0.07). Persuasion strategies differ substantially by scam type (Figure 5). Authority dominates imposter (0.98), tech support (0.91), and employment (0.93) scams, all categories built around impersonation, while reciprocity leads in reward-framed lottery (0.95) and advance fee loan (0.87) scams. Employment scams are distinctive in pairing authority with high *liking* (0.84), consistent with personalized recruiter-style outreach.

4.2.2 Red Flags

Figure 3b shows the overall distribution of red flags. *Unsolicited contact* is nearly universal (0.96), fol-



(a) Persuasion strategies (b) Red flags (c) Severity levels

Figure 3: Overall distributions across the full dataset.

lowed by *too-good-to-be-true* claims (0.61), *grammar or language issues* (0.58), *official-organization posing* (0.53), *pressure to act now* (0.52), and *requests for personal or financial information* (0.44). Upfront payment requests, gift card/wire/crypto demands, and legal threats are less common overall but concentrate in specific scam types. Red-flag profiles also vary by scam type (Figure 6). Imposter scams exhibit the densest profile, combining urgency, official impersonation, language issues, and legal threats; tech support scams add information requests and upfront fees. Investment and lottery scams are dominated by too-good-to-be-true claims, while employment scams add information-request signals consistent with identity theft.

4.2.3 Severity Distribution

Figure 3c shows the overall distribution of severity levels. Across the full dataset, most scam messages are labeled as high severity (63%), followed by medium severity (19%) and critical severity (17%). Low-severity messages are rare (1%). This distribution reflects our focus on original scam messages extracted from victim-reported incidents, which often contain concrete attempts to obtain money, personal information, or access. Severity varies considerably by scam type (Figure 7). Imposter scams have the highest proportion of critical cases (59%), reflecting institutional impersonation paired with coercion. Tech support (86.6% high) and lottery scams (75.3% high, 12.9% critical) are overwhelmingly high severity. Employment scams include a substantial critical subset (11.5%), often involving requests for sensitive identity information. Advance fee loan and investment scams are mostly medium to high severity.

5 Robustness Evaluation

5.1 Hard Scam Rewritten Task

We next used the annotation schema to construct hard scam examples for robustness evaluation. Prior work has explored adversarial or enhanced scam generation by adding psychological techniques to scam messages (Ma et al., 2025), or by increasing credibility-building, urgency creation, and emotional manipulation (Yang et al., 2025). In contrast, our approach moves in the opposite direction. Rather than making scam messages more overtly persuasive, we rephrased or softened annotated persuasion and red-flag evidence spans. Specifically, we used GPT-5.3-chat-latest to rephrase or soften the text spans that were previously annotated as persuasion-strategy evidence or red-flag evidence. The goal was to preserve the general scam scenario while removing the cues that LLM-based detectors often rely on, such as urgency, suspicious links, requests for personal information, and grammatical irregularities. We refer to the resulting examples as *persuasion-removed hard scam examples*.

5.2 Scam Detection Task

We evaluated binary scam detection under two positive-class settings. In the **Original** setting, the positive class consists of the original scam messages in SCAMBENCH. In the **Rewritten** setting, the positive class consists of the rewritten hard scam examples. In both settings, the negative class consists of the non-scam messages described above. Thus, the rewritten examples retain the scam label because they are derived from real scam messages, but they are intentionally stripped of explicit persuasion and red-flag cues. We evaluated five LLM-based detectors: GPT-5.4-nano, GPT-4.1-nano, GPT-4o-mini, Claude Haiku 4.5, and Claude Sonnet 4.5. Each model is prompted with a simple binary classification instruction: determine whether the message is a scam and answer with *Yes* or *No*. We report precision, recall, accuracy, and F1, treating scam as the positive class. Since advance fee loan scams has no corresponding non-scam examples, we report recall for advance fee loan scams but omit precision, accuracy, and

Table 2: Overall scam detection performance on original and persuasion-removed rewritten scam examples. Values in parentheses denote performance drops from original to rewritten examples.

Model	Setting	Precision	Recall	Accuracy	F1
GPT-5.4-nano	Original	0.798	0.994	0.871	0.885
	Rewritten	0.789 (-0.009)	0.943 (-0.051)	0.846 (-0.025)	0.859 (-0.026)
GPT-4.1-nano	Original	0.954	0.924	0.940	0.939
	Rewritten	0.926 (-0.028)	0.552 (-0.372)	0.754 (-0.186)	0.691 (-0.248)
GPT-4o-mini	Original	0.950	0.978	0.963	0.964
	Rewritten	0.941 (-0.009)	0.814 (-0.164)	0.881 (-0.082)	0.873 (-0.091)
Claude Haiku 4.5	Original	0.983	0.982	0.982	0.982
	Rewritten	0.980 (-0.003)	0.839 (-0.143)	0.911 (-0.071)	0.904 (-0.078)
Claude Sonnet 4.5	Original	0.995	0.954	0.975	0.974
	Rewritten	0.994 (-0.001)	0.693 (-0.261)	0.844 (-0.131)	0.816 (-0.158)

F1, which are undefined without negatives in the per-category evaluation. Overall metrics aggregate non-scam messages from the remaining categories.

5.3 Results

Table 2 summarizes the overall detection performance. Across all models, performance drops when original scam messages are replaced with persuasion-removed hard examples. The most consistent degradation is in recall, indicating that models increasingly misclassify rewritten scams as non-scam. Precision remains comparatively high, suggesting that the main failure mode is not excessive false positives on non-scam messages, but false negatives on rewritten scam messages. The largest degradation occurs for GPT-4.1-nano, whose recall drops from 0.924 to 0.552 and whose F1 drops from 0.939 to 0.691. Claude Sonnet 4.5 also exhibits a substantial recall drop, from 0.954 to 0.693, despite maintaining very high precision. GPT-4o-mini and Claude Haiku 4.5 are more robust but still lose 16.4 and 14.3 recall points, respectively. GPT-5.4-nano shows the smallest recall degradation, from 0.994 to 0.943, although its original precision and F1 are lower than those of other models.

The per-scam-type results in Tables 16-20 in Appendix C reveal that the impact of rewriting is uneven across both models and scam categories. Overall, compared to other scams, sweepstakes/lottery/prizes scams remain comparatively easier to detect after rewriting.

5.4 Case Study: Why Persuasion-Removed Scams are Hard-to-Detect

To better understand the failure modes, we analyzed false negative cases where all evaluated models correctly classified the original scam messages but misclassified the rewritten versions as non-scam. We identified five failure patterns summarized in Table 3, with detailed analysis provided in Appendix B. Across these cases, rewritten scam messages often resemble legitimate communication formats such as HR updates, customer service notices, or investment disclosures. Common failure patterns include: (1) removal of direct requests makes scams appear benign, causing scam messages to resemble standard internal communication; (2) replacing suspicious channels with official channels produces false legitimacy, leading models to infer safety from institutional framing; (3) removing urgency and scarcity eliminates major scam cues, making scam interactions appear routine; (4) professionalization and grammar repair suppress suspicion, causing polished scams to resemble legitimate messages; and (5) adding risk disclaimers makes scams appear legitimate, especially in investment-related communication.

Discussion

Human-AI curation as scalable dataset construction. SCAMBENCH illustrates the value of human-AI collaboration for building datasets from noisy scam-report platforms. Rather than relying on fully manual annotation or automated labeling, our pipeline uses human-labeled examples to guide, evaluate, and scale LLM-assisted annotation. This approach can generalize beyond scam research to domains such as hate speech and harassment that require both contextual human understanding and scalability (Giorgi et al., 2025). Our human-AI collaborative multi-dimensional annotation supports comparison of scam tactics across scam types, scam-type-specific intervention design, and evidence-grounded counterfactual evaluation through controlled rewriting.

Scam type heterogeneity and cue density. We showed that scam categories differ systematically in persuasion strategies, red flags, and severity, hence, scam type should be treated as a core property of scam datasets. Scam-type-agnostic approaches may obscure these differences, with models performing well overall yet failing on types with distinct social and linguistic profiles. Because scam type affects severity more than persuasion strategies or red flags, severity-aware systems may benefit most from category-specific modeling. Finally, scam messages expose more user-facing red flags than scammer-side tactics, so original messages often carry multiple redundant cues while rewritten messages may retain only one or two after cue removal, indicating performance on original

Table 3: Representative case study examples where all models correctly classified the original scam as scam but misclassified the rewritten version as non-scam.

Failure pattern	Original scam message	Rewritten message
Pattern 1: Removal of direct requests makes scams appear benign	We are making important updates to our compensation and benefits in 2025. These changes will streamline salary adjustments, enhance performance bonuses, and improve benefits management for all employees. To review your updated compensation details and benefits package, please copy the link below to your browser.	We're planning some updates to compensation and benefits for 2025. The changes will adjust how salary reviews, bonuses, and benefits are handled. You can view the details of your current compensation and benefits through the usual HR portal.
Pattern 2: Replacing suspicious channels with official channels produces false legitimacy	Internal Revenue Service (IRS) You are eligible to receive a $1,400 Economic Impact Payment. Please provide your accurate personal information. We will deposit the amount into your bank account or mail a paper check within 1 to 2 business days. [url]	Notice about a possible Economic Impact Payment: you may wish to review your eligibility by visiting the official IRS website directly at irs.gov. If eligible, payments are typically issued by direct deposit or mailed check according to standard processing times.
Pattern 3: Removing urgency and scarcity eliminates a major scam cue	This notice is regarding your August 06, 2025 mortgage. We need you to call us about an important matter regarding this loan. This is time sensitive so please call us at ***-***-**** as soon as possible.	If you have any questions about your August 6, 2025 mortgage, you're welcome to contact us at ***-***-**** during our regular business hours.
Pattern 4: Professionalization and grammar repair suppress suspicion	Anti-malware+Anti-Spyware+Ad-Blocker+Passwor maager+Spam Mail Filter+Identity theft Protection Free Techinical Assistance Mcafee Account Manager : [name] {1-***-***-**** Qikbay Tech Toll Free Number 1-***-***-****	Security tools available: malware protection, spyware scanning, ad blocking, password management, email filtering, and identity monitoring. For general product information or support options, you can contact customer support at 1-***-***-****.
Pattern 5: Adding risk disclaimers makes scams appear legitimate	Note Trade USA in the business of buying oil/gas leases in Oklahoma in their Oil Well Investment portfolio. Initial investment $750,000. Low end, conservatively return $700,000 first two years, then about $90,000 per year equal to 58% return investment per year for the first two years, capital return, and cash flow $90,000 per year for at least 10-15 years.	A U.S.-based firm focuses on acquiring oil and gas leases in Oklahoma as part of its investment portfolio. Typical projects involve substantial capital commitments, and performance depends on market conditions, production levels, and operating costs. Past projects have generated income over time, with distributions varying year to year.

messages may overstate model robustness.

From evidence-grounded annotation to robustness testing. The rewriting experiment shows how evidence-grounded annotations can support diagnostic robustness evaluation. Rewritten scam messages often resemble legitimate communication, indicating that model performance is sensitive to how scam signals are expressed. Effective detection should therefore extend beyond isolated cues to the broader interaction context, claimed identity, requested action, and communication trajectory. SCAMBENCH supports this through signal-level analysis and controlled robustness evaluation grounded in real-world scam communication. More broadly, the persuasion-removed rewriting protocol generalizes to any harm domain with evidence-span annotations can apply the cue-removed stress test, making such annotation useful for interpretability and robustness evaluation.

Implications for user education The red-flag schema closely aligns with the warning signs that consumer-protection agencies teach users to recognize (U.S. Securities and Exchange Commission, 2023). The rewriting experiment, therefore, has implications for user education, not only for model robustness. Scam messages rewritten with polished language, reduced urgency, and risk-aware framing can bypass both consumer-facing checklists and model detectors. This suggests that effective scam education should go beyond surface-level cue recognition and consider questions such as whether the communication channel matches the claimed identity and whether the interaction resembles legitimate behavior. SCAMBENCH supports multi-dimensional analysis, providing a basis for designing interventions that remain effective even when surface cues are softened.

6 Conclusion

SCAMBENCH addresses a fundamental gap in scam research: a benchmark that provides 1,836 authentic victim-reported scam messages across six scam types with multi-dimensional annotations. The annotations surface scam-type heterogeneity that aggregate detection metrics conceal, and they support robustness evaluation through controlled rewriting of annotated evidence spans. SCAMBENCH shows that LLM-based scam detectors that rely on surface cues fail under principled rewriting. As such systems become pervasive and scam communication continues to adapt, human-AI collaborative evidence-grounded annotation paired with annotation-guided robustness testing offers a more durable foundation for understanding online harms than performance metrics alone.

Limitations

This work has several limitations. First, the scam corpus is constructed from a single source: BBB Scam Tracker. Future work should incorporate more scam-report platforms and data sources to improve source diversity. Furthermore, SCAMBENCH currently focuses on English-language scam messages. Future versions should include multilingual and cross-cultural scam communication patterns, accounting for varied languages, regions, and cultural contexts.

Second, our annotation schema focuses on persuasion strategies, red flags, and severity levels. Although these labels capture important dimensions, they do not fully represent the structure of scam communication. Future work could add richer annotations such as contact channel, interaction stage, claimed identity, requested action, targeted asset, and victim outcome, enabling more detailed modeling of scam trajectories and user risk. In addition, the annotated cues should be interpreted as probabilistic signals rather than standalone indicators; otherwise, systems may over-penalize correlated features such as grammar irregularities, which could disadvantage non-native speakers or small organizations.

Third, although the benchmark is balanced at the corpus level, non-scam data remains more difficult to collect for some categories. In particular, imposter, investment, and sweepstakes/lottery/prizes scams have fewer closely matched non-scam counterparts. Future work should expand the search for legitimate comparison data in domains where benign communications are rare or difficult to obtain.

Finally, our robustness evaluation focuses on generating hard scam examples by removing or rephrasing persuasion-strategy and red-flag evidence. Future work could study how scam and non-scam messages transform toward each other, what defines borderline cases, and how models and users interpret ambiguous communications. to understand better the boundary between scam and non-scam communication.

Ethical Considerations

SCAMBENCH is constructed from publicly available scam reports and legitimate non-scam sources, and we do not directly interact with individuals during data collection. We cite providers of all existing assets used in the study. The released dataset is provided under a CC BY-NC 4.0 license and is available through an anonymous repository for review: https://doi.org/10.5281/zenodo.20125179.

Because scam reports may contain sensitive information, we apply basic anonymization where possible. However, due to the nature of real-world scam data, complete removal of identifying information cannot be guaranteed. The dataset may contain harmful or deceptive language because it is designed to support scam detection research. A misuse risk is that persuasion-removed rewriting could help adversaries soften cues used by detectors. We tried to mitigate this through a CC BY-NC 4.0 license and by limiting intended use to detection, robustness evaluation, user education, and human-centered security research. These mitigations are imperfect, but the benefit of more robust detection outweighs the marginal benefit to actors who already access extensive scam corpora elsewhere. LLMs were used for annotation, rewriting, and evaluation experiments, while the manuscript was written by the authors. LLMs were further used for proofreading and editing under human supervision. Responsibility for the final manuscript lies entirely with the authors.

We organize the released files in standard CSV format with documented fields, labels, annotation schema, and usage notes. To support the FAIR principles, the dataset is findable through the public repository, accessible under a stated research license, interoperable through standard CSV files with metadata and structured annotation fields, and reusable through accompanying documentation and schema descriptions. This paper and the repository provide the elements of a dataset datasheet, including motivation, composition, collection process, preprocessing, annotation, intended uses, limitations, and ethical considerations.

References

419scam.org. 2026. “419” Scam – Advance Fee / Fake Lottery Scam. https://www.419scam.org. Accessed: 2026-03-13.

AARP. 2026. Scam-tracking map. Accessed: 2026-03-25.

AIS Network. 2023. Tech support scam red flags. Accessed: 2026-03-23.

Matthew Anderson, Evita March, Lesley Land, and Christine Boshuijzen-van Burken. 2024. Exploring the roles played by trust and technology in the online investment fraud victimisation process. *Journal of Criminology*, 57(4):488–514.

Better Business Bureau. 2026. Bbb scam tracker: Look up a scam. Accessed: 2026-03-25.

Dainis A Boumber, Fatima Zahra Qachfar, and Rakesh Verma. 2024. Domain-agnostic adapter architecture for deception detection: extensive evaluations with the difraud benchmark. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5260–5274.

Mark Button and Cassandra Cross. 2017. *Cyber frauds, scams and their victims*. Routledge.

Mark Button, Carol McNaughton Nicholls, Jane Kerr, and Rachael Owen. 2014. Online frauds: Learning from victims why they fall for these scams. *Australian & New Zealand journal of criminology*, 47(3):391–408.

Joshua JS Chang. 2008. An analysis of advance fee fraud on the internet. *Journal of Financial Crime*, 15(1):71–81.

Robert B Cialdini. 2016. *Presuasion: A Revolutionary Way to Influence and Persuade*.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Federal Trade Commission. 2023. What to know about advance-fee loans. Accessed: 2026-03-23.

Federal Trade Commission. 2025. New ftc data show a big jump in reported losses to fraud to $12.5 billion in 2024. Accessed: 2025-09-05.

FIRST.org. 2019. Common vulnerability scoring system (cvss) v3.1 specification. Accessed: 2026-03-23.

Tommaso Giorgi, Lorenzo Cima, Tiziano Fagni, Marco Avvenuti, and Stefano Cresci. 2025. Human and llm biases in hate speech annotations: A socio-demographic analysis of annotators and targets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 653–670.

Xu Han, Qiang Li, Yaling Qi, Hongbo Cao, Witold Pedrycz, and Wei Wang. 2025. Scamgen: Unveiling psychological patterns in tele-scam through advanced template-augmented corpus generation. *Computers in Human Behavior*, 162:108451.

Thomas J Holt and Danielle C Graves. 2007. A qualitative analysis of advance fee fraud e-mail schemes. *International Journal of Cyber Criminology*, 1(1):137–154.

Hoosier Lottery. 2026. Hoosier lottery official website. Accessed: 2026-03-23.

Indeed. 2026. Indeed job search platform. Accessed: 2026-03-23.

Internal Revenue Service. 2026. Irs notices and letters. Accessed: 2026-03-23.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579.

Huaye Li, Siyu Huang, and Jinkyung Katie Park. 2025. A human-centered review of large language models for online scam detection. SSRN Working Paper.

Shang Ma, Tianyi Ma, Jiahao Liu, Wei Song, Zhenkai Liang, Xusheng Xiao, and Yanfang Ye. 2025. Psyscam: a benchmark for psychological techniques in real-world scams. *arXiv preprint arXiv:2505.15017*.

National Institute of Standards and Technology. 2018. Risk management framework for information systems and organizations. Accessed: 2026-03-23.

New York Lottery. 2026. New york lottery promotions. Accessed: 2026-03-23.

Jason RC Nurse. 2018. Cybercrime and you: How criminals attack and the human factors that they seek to exploit. *arXiv preprint arXiv:1811.06624*.

NVIDIA. 2024. Techqa-rag-eval dataset. Hugging Face dataset. Accessed: 2026-03-23.

Office of the Comptroller of the Currency. 2023. Imposter scams. Accessed: 2026-03-23.

Jinkyung Katie Park, Rahul Dev Ellezhuthil, Pamela Wisniewski, and Vivek Singh. 2025. Collaborative human-ai risk annotation: co-annotating online incivility with chaira. *Information Research 30*.

Pew Research Center. 2025. Online scams and attacks in america today. Accessed: 2025-09-05.

Aunshul Rege. 2009. What's love got to do with it? exploring online dating scams and identity fraud. *International Journal of Cyber Criminology*, 3(2).

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

ScamSearch. 2026. Scamsearch: Online scam lookup and reporting platform. Accessed: 2026-03-25.

Gilberto Atondo Siu, Alice Hutchings, Marie Vasek, and Tyler Moore. 2022. "invest in crypto!": An analysis of investment scam advertisements found in bitcointalk. In *2022 APWG symposium on electronic crime research (eCrime)*, pages 1–12. IEEE.

Social Security Administration. 2026. Employer correction request notice. Accessed: 2026-03-23.

South Carolina Department of Consumer Affairs. 2023a. Job scams. Accessed: 2026-03-23.

South Carolina Department of Consumer Affairs. 2023b. Lottery and sweepstakes scams. Accessed: 2026-03-23.

SweepsAdvantage. 2026. Sweepsadvantage: Sweepstakes and contest listings. Accessed: 2026-03-23.

U.S. Citizenship and Immigration Services. 2020. Sample form i-797 notice of action. Accessed: 2026-03-23.

U.S. Securities and Exchange Commission. 2023. Red flags of investment fraud checklist. Accessed: 2026-03-23.

Sokratis Vidros, Constantinos Kolias, Georgios Kambourakis, and Leman Akoglu. 2017. Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset. *Future Internet*, 9(1):6.

Monica T Whitty. 2013. The scammers persuasive techniques model: Development of a stage model to explain the online dating romance scam. *British journal of criminology*, 53(4):665–684.

Monica T Whitty. 2015. Anatomy of the online dating romance scam. *Security Journal*, 28(4):443–455.

Monica T Whitty and Adam Joinson. 2008. *Truth, lies and trust on the Internet*. Routledge.

Wisconsin Department of Financial Institutions. 2026. Investment scam tracker. Accessed: 2026-03-25.

Shu Yang, Shenzhe Zhu, Zeyu Wu, Keyu Wang, Junchi Yao, Junchao Wu, Lijie Hu, Mengdi Li, Derek F Wong, and Di Wang. 2025. Fraud-r1: A multi-round benchmark for assessing the robustness of llm against augmented fraud and phishing inducements. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4374–4420.

Appendix

A Content-Type Classification Examples

Representative examples of the three content categories used during content filtering are shown below.

Scammer message. *Hi, this is [name] from [institution]. Just checking in to go over your $70,000 loan request. Got a few minutes to talk now?*

Victim narrative. *Caller said with [company name] an I had won $18 million and a 2025 [car brand], wanted me to sell my house to pay the taxes.*

Both. *I get the following message every day from different 5 digit phone numbers. "Hi (name), your profile with us has been updated. Accept a new amount over [url] or reply STOP to quit"*

B Detailed Case Study

Pattern 1: Removal of direct requests makes scams appear benign. A dominant failure mode occurred when rewriting removed direct requests for sensitive information, links, or required actions. For instance, in the original version, the message instructed the recipient to "copy the link below to your browser" to access updated compensation details. All models correctly classified this as a scam in a original form, as the explicit link-based action constitutes a common pattern of directing users to external resources under unclear legitimacy. In the rewritten version, this direct instruction was replaced with a neutral statement that the information could be accessed through the "usual HR portal." Once the explicit action request was removed, models consistently interpreted the message as a standard internal communication. This suggests that current detectors may be highly sensitive to explicit action requests but much less robust when the same scam scenario is deferred, obscured, or framed as part of a legitimate process.

Pattern 2: Replacing suspicious channels with official channels produces false legitimacy. Several rewrites inserted or emphasized official channels, which caused models to infer safety. The IRS Economic Impact Payment scam originally contained strong scam signals: IRS impersonation, a suspicious IRS-like domain, a promise of a $1,400 payment, and a request for personal information. After rewriting, the message advised users to visit the official IRS website directly and use verified contact channels. Models then described it as legitimate caution or standard security advice. This case reveals a blind spot in scam detection: once a message uses safety-preserving language such as "official website," "verified contact channels," or "usual support channel," models often treat the message as benign, even when the broader scenario originates from a scam.

Pattern 3: Removing urgency and scarcity eliminates a major scam cue Urgency and scarcity appear to be among the strongest cues current detectors use to flag scams. In a mortgage imposter example, the original message described the matter as "time sensitive" and asked the recipient to call "as soon as possible," which were phrasing that triggered both *scarcity* and *pressure-to-act-now* annotations. The rewritten version replaced this with an invitation to call during regular business hours,

leaving the rest of the scenario intact. All models labeled the rewritten message non-scam, describing it as a standard mortgage service notice when the underlying interaction did not change, except for the urgency cue. This suggests that once urgency is removed, scam messages can look as routine business communication even when the broader context remains suspicious.

Pattern 4: Professionalization and grammar repair suppress suspicion. Several original messages were detected partly because they contained poor grammar, awkward formatting, or unprofessional phrasing. For instance, the original scam message contained misspellings such as "Passwor maager" and "Free Techinical Assistance," along with named support personnel and phone numbers. After rewriting, it became a clean description of security tools and support options. Models classified it as a legitimate product description. This case indicates that models often conflate polished style with legitimacy. When scam messages are grammatically repaired and made professionally plausible, detection becomes substantially weaker.

Pattern 5: Adding risk disclaimers makes scams appear legitimate. This pattern was especially salient for the investment scam. When original scam message promised implausibly high returns (e.g., a $750,000 initial investment, approximately 58% annual return), models correctly detected the original because it contained unrealistic guaranteed returns and an upfront investment requirement. The rewritten version replaced these claims with risk-aware investment language, stating that performance depends on market conditions and operating costs, and that distributions vary over time. Models labeled it non-scam, explaining that it acknowledged risk and resembled legitimate investment disclosure. This failure pattern is important to acknowledge since risk-disclosure language is common in legitimate financial communication. If scammers add compliance-style language or soften guaranteed-return claims, models may treat the message as legitimate even when the underlying investment structure remains suspicious.

C Supplementary Figures, Prompt, and Tables

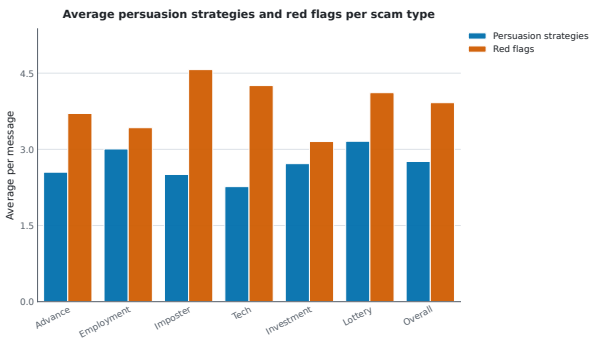


Figure 4: Average number of persuasion strategies and red flags per message by scam type. Red flags are consistently more frequent than persuasion strategies, but both vary across scam categories.

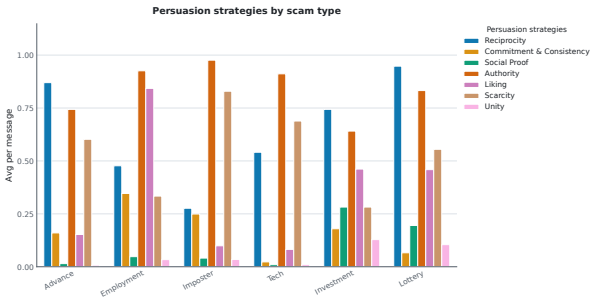


Figure 5: Average number of persuasion strategies per message by scam type.

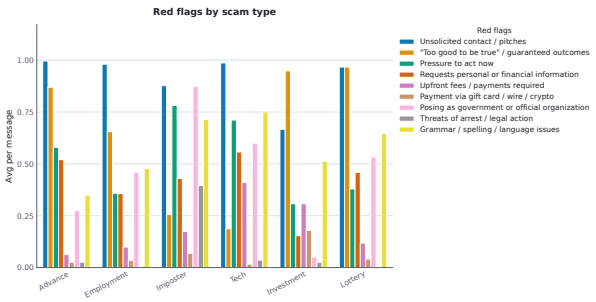


Figure 6: Average number of red flags per message by scam type.

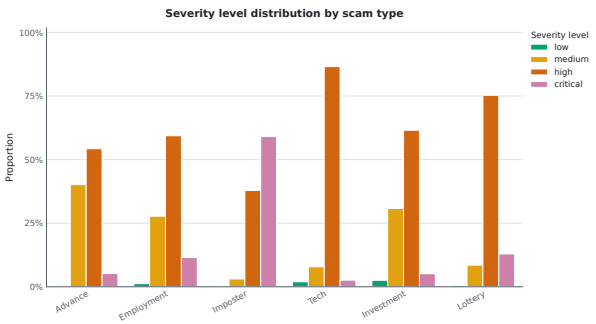


Figure 7: Severity level distribution by scam type.

Content-Type Classification Prompt

```
You are an experienced moderator in detecting and analyzing sweepstakes/lottery/prizes scam
messages.
Sweepstakes/lottery/prizes scam definition: Victims are tricked into thinking they have won a prize
or lottery jackpot but must pay up-front fees to receive the winnings, which never materialize.
Sometimes this con involves a fake check and a request to return a portion of the funds to cover
fees.
Your task is to carefully read the provided text and classify it into one of the following categories:
1. Scammer Message 2. Victim Narrative 3. Both
### Instructions: - Output ONLY and EXACTLY one of the classification labels: "Scammer Message",
"Victim Narrative", or "Both". - Do not provide any explanation or reasoning.
### Examples:
[Example 1: Scammer Message]
Input: Congratulations I'm [name] powerball's first billion dollars I'm donating the sum of
$300,000.00 dollars to random individuals if you get this message then your contact was selected
among the lucky winners kindly send a text message to the agent in charge of your winnings money
[name] +1(***) ***-**** With your winnings code FT65/TX.
Output: Scammer Message
[Example 2: Victim Narrative]
Input: I got a call from [name], who claimed to work for the Virginia Lottery. She said I won
$500,000 and a new truck. Then, Agent [name] sent me a message on Facebook Messanger. I had to
send them money before they would send me the prize. I asked her to send me proof that they were
going to send it to me. They sent me a picture of a FedEx truck with a box of money in the back. I
thought it was real. I sent them $6,050 total on prepaid giftcards. They kept telling me it would
be the last time they would ask, but they kept asking for more and more money.
Output: Victim Narrative
[Example 3: Both]
Input: I received a call telling me I am the winner of an apartment in Mt. Pleasant community in
Toronto, ON. All I had to do is sign the consent form emailed to me. I was not supposed to tell
anyone of my winnings until they could publicize my name as a winner. Then they said I would have
to pay 2% of the value of the apartment for lawyer and administration fees. I asked how much that
would be and they did not answer. They continued to ask me if I understood what I had to do to
claim this prize.
Email sent: mapple raffle From: ******@***.** To: ******@***.** Thu, Oct 16 at 3:21 PM
Hello [name]! Once again, Omaze team congratulates you with the victory of a new apartment in
Toronto! See attached the document, and sign it please as soon as possible so that we can proceed.
Best regards, Mapple Raflle Team [url]
Output: Both
### Task: Classify the following input text according to the categories above.

Input: {text}
```

Persuasion Strategies, Red Flags, and Severity Level Anontation Prompt

Task: You are an expert annotator trained in persuasion theory, online scam analysis, and human-centered security. Your job is to analyze the original scam message and produce structured labels for:
Persuasion strategies (Cialdini's 7 principles) + evidence
Red flags (R1-R9) + evidence
Severity level (low, medium, high, critical)
Follow all definitions and output format exactly.
1. Persuasion Strategies (Cialdini's 7 Principles) Identify all persuasion strategies used in the message. For each strategy, extract verbatim evidence from the message. Definitions:
P1 Reciprocity - Creating obligation by offering help, rewards, gifts, or favors.
P2 Commitment & Consistency - Starting with a small request and escalating to larger demands.
P3 Social Proof - Claiming many others have benefited; using testimonials or group behavior.
P4 Authority - Invoking titles, institutions, official seals, or legal threats; impersonating authority.
P5 Liking - Using flattery, friendliness, shared interests, or emotional bonding.
P6 Scarcity - Using urgency, limited-time offers, countdowns, or "last chance" framing.
P7 Unity - Claiming shared identity (community, religion, nationality, ethnicity) to build trust.
2. Red Flags
Identify all red flags present in the message. For each red flag, extract verbatim evidence from the message.
Red-flag categories:
R1 Unsolicited contact / pitches
R2 "Too good to be true" / guaranteed outcomes
R3 Pressure to act now
R4 Requests personal or financial information
R5 Upfront fees / payments required
R6 Payment via gift card / wire / crypto
R7 Posing as government or official organization
R8 Threats of arrest / legal action
R9 Grammar / spelling / language issues
3. Severity Level
Choose one label:
Low - Mild suspicious cues; no direct request for money or sensitive data.
Medium - Some manipulative tactics; early-stage scam; limited harm if ignored.
High - Clear attempt to obtain money, credentials, or access; strong manipulation.
Critical - Imminent financial loss, coercion, impersonation of authority, or threats.
4. Manually annotated examples
Example 1:
Original Scam Message:
[name] from Robert Half Recruiting team. Offering job role at Costco and Sam's Club in improving product visibility. 90 minutes per day, 4 days a week. Work anywhere and daily pay ranges from $200-$350. Free training, along with 15-20 days of paid vacation per year. 35 openings. Must be 25 years old.
Annotated output:

```
{
  "persuasion_strategies": [
    {
      "strategy": "P1 Reciprocity",
      "evidence": "Free training, along with 15-20 days of paid vacation per year."
    },
    {
      "strategy": "P4 Authority",
      "evidence": "[name] from Robert Half Recruiting team."
    },
    {
      "strategy": "P6 Scarcity",
      "evidence": "35 openings"
    }
  ],
  "red_flags": [
    {
      "flag": "R1 Unsolicited contact / pitches",
      "evidence": "[name] from Robert Half Recruiting team. Offering job role at Costco and Sam's
           Club"
    },
    {
```

```
      "flag": "R2 Too good to be true",
      "evidence": "daily pay ranges from \$200-\$350. Free training..."
    }
  ],
  "severity_level": "medium"
}
```

Example 2:
[...]
Example 3:
[...]
5. Output Format (JSON)
IMPORTANT: Your entire response must be ONLY a valid JSON object. No preamble, no explanation, no markdown fences. Return your answer ONLY in the following structure:

```
{
  "persuasion_strategies": [
    {
      "strategy": "P# Name",
      "evidence": "verbatim evidence from the message"
    }
  ],
  "red_flags": [
    {
      "flag": "R# Name",
      "evidence": "verbatim evidence from the message"
    }
  ],
  "severity_level": "low | medium | high | critical"
}
```

Formatting rules:
For persuasion strategies:
"strategy": "P4 Authority"
"evidence": "You must respond immediately or face legal action."
For red flags:
"flag": "R3 Pressure to act now"
"evidence": "Act immediately or your account will be closed."
Evidence must be verbatim
Do not paraphrase
Do not hallucinate
Include only items that actually appear
On average, human experts identify 2-3 persuasion strategies and 2-3 red flags per message. Only output a label when there is clear, explicit evidence

Below is the original scam message to annotate:

Rewrite Prompt

```
You are an online safety researcher working on robustness evaluation for scam detection systems.
Task:
Your task is to rewrite the input scam message by rephrasing specific annotated spans to eliminate
the persuasion strategies and red-flag signals. The goal is to create more subtle variants for
evaluating model robustness, not to improve real-world deception.
Input:
- Original scam message:
{SCAM_MESSAGE}
- Persuasion-strategy evidence spans:
{P_EVIDENCE_LIST}
- Red-flag evidence spans:
{R_EVIDENCE_LIST}
Definitions:
P1 Reciprocity - Creating obligation by offering help, rewards, gifts, or favors.
P2 Commitment & Consistency - Starting with a small request and escalating to larger demands.
P3 Social Proof - Claiming many others have benefited; using testimonials or group behavior.
P4 Authority - Invoking titles, institutions, official seals, or legal threats; impersonating
authority.
P5 Liking - Using flattery, friendliness, shared interests, or emotional bonding.
P6 Scarcity - Using urgency, limited-time offers, countdowns, or "last chance" framing.
P7 Unity - Claiming shared identity (community, religion, nationality, ethnicity) to build trust.
Red-flag categories:
R1 Unsolicited contact / pitches
R2 "Too good to be true" / guaranteed outcomes
R3 Pressure to act now
R4 Requests personal or financial information
R5 Upfront fees / payments required
R6 Payment via gift card / wire / crypto
R7 Posing as government or official organization
R8 Threats of arrest / legal action
R9 Grammar / spelling / language issues
Requirements:
1. Ensure that all provided evidence spans related to persuasion strategies or red flags are
eliminated through rephrasing rather than simple deletion.
2. Replace the evidence spans with more neutral or contextually plausible phrasing that preserves
the scenario while reducing explicit scam signals.
3. Avoid introducing new scam tactics.
4. Maintain natural, coherent, and realistic language, with a tone that could plausibly appear in
legitimate communication.
5. If a provided evidence span is empty, ignore it.
Output:
Output only the rewritten message. Do not include anything else.
```

Scam Detection Prompt

```
Determine whether the following message is a scam. Answer strictly with one word: Yes or No. Do
not output anything else .
Message:
{MESSAGE}
```

Scam Detection with Explanation Prompt

```
Determine whether the following message is a scam. Answer in the following format:
Label: Yes or No
Explanation: <brief reasoning>
Do not include any additional content.
Message:
{MESSAGE}
```

Table 4: Comparison of public scam-report platforms considered for dataset construction.

Source	Metadata	Advantages	Limitations
BBB Scam Tracker (Better Business Bureau, 2026)	Description; Dollars_lost; Location; Date_reported; Business_name_used	Sourced from victim reports; covers a wide range of scam types, scams can be filtered by scam type; large data size, supporting multi-type scam analysis	Consists mainly of narrative descriptions rather than original scam message content; labels are based on self-reported form selections, which may be inaccurate; lacks non-scam examples
ScamSearch (ScamSearch, 2026)	Date; Scam Type; BTC Address; Scammer Username; Scammer Email; Scammer Phone Number; Scam Description; Scammer Website	Sourced from victim reports; covers a wide range of scam types; large data size, supporting multi-type scam analysis	Consists mainly of narrative descriptions rather than original scam message content; labels are based on self-reported form selections, which may be inaccurate; lacks non-scam examples; can only search using scammer profiles (e.g., username, email, phone number), not by scam type; 50% of reports contain six or fewer words
AARP Scam-Tracking Map (AARP, 2026)	Scam Type; Amount Lost; Description; Date Occurred; Contact Method; Zipcode	Sourced from victim reports; covers a wide range of scam types, scams can be filtered by contract method; large data size, supporting multi-type scam anal- ysis	Consists mainly of narrative descriptions rather than original scam message content; labels are based on self-reported form selections, which may be inaccurate; lacks non-scam examples; can only search using location, not by scam type
Investment Scam Tracker (Wisconsin Department of Financial Institutions, 2026)	Date; Complaint Narrative; Scam Type; Subjects; Reported Loss Amount	Sourced from victim reports	Consists mainly of narrative descriptions rather than original scam message content; labels are based on self-reported form selections, which may be inaccurate; lacks non-scam examples; contains only 58 entries (10/30/2025)

Table 5: Data filtering pipeline and sample counts at each stage. The pipeline consists of eight steps: (1) Scam No. after scraping; (2) Scam No. after 1st deduplication; (3) Scammer message (only) count after Llama-4-Scout-17B-16E-Instruct automatic annotation; (4) Semantic deduplication using a Sentence Transformer (cosine similarity threshold = 0.8); (5) Removal of short messages (global 35th percentile threshold = 20 words); (6) Manual cleaning (removal of narrative and garbled text, and correction of misclassified scam categories); (7) Second semantic deduplication (threshold = 0.92); (8) Final manual cleaning (removal of remaining garbled messages).

Scam type	Scraped	1st dedup	Scam msg	Dedup (0.8)	Short rm	Manual	Dedup (0.92)	Final
Advance fee loan	44,216	8,823	1,418	844	535	275	269	**269**
Employment	49,800	9,951	5,202	1,494	953	591	566	**566**
Imposter	49,800	9,966	3,162	2,476	1,704	294	294	**293**
Tech support	39,760	7,955	1,016	834	591	313	313	**305**
Investment	19,065	3,828	529	496	330	39	39	**39**
Sweepstakes/lottery/prizes	49,800	9,969	2,150	1,361	911	364	364	**364**
Total	252,981	50,492	13,477	7,505	5,024	1,876	1,845	**1,836**

Table 6: Non-scam dataset sources and distribution across categories.

Type	Source	Number
Employment	Indeed (Indeed, 2026)	899
Imposter	IRS (Internal Revenue Service, 2026); SSA (Social Security Administration, 2026); USCIS (U.S. Citizenship and Immigration Services, 2020); Chase emails	9
Tech support	TechQA-RAG-Eval (NVIDIA, 2024)	908
Investment	Chase emails	6
Sweepstakes/lottery/ prizes	SweepsAdvantage (SweepsAdvantage, 2026); Hoosier Lottery (Hoosier Lottery, 2026); NY Lottery (New York Lottery, 2026)	12
Total		1,834

Table 7: Sources of scam red flag indicators by scam type.

Scam type	Source
Advance fee loan	Federal Trade Commission (Federal Trade Commission, 2023)
Employment	South Carolina Department of Consumer Affairs (South Carolina Department of Consumer Affairs, 2023a)
Imposter	Office of the Comptroller of the Currency (Office of the Comptroller of the Currency, 2023)
Tech support	AIS Network (AIS Network, 2023)
Investment	U.S. Securities and Exchange Commission (U.S. Securities and Exchange Commission, 2023)
Sweepstakes/lottery/ prizes	South Carolina Department of Consumer Affairs (South Carolina Department of Consumer Affairs, 2023b)

Table 8: Typical red flag patterns by scam type (used as annotation guidance). R1–R9 denote: unsolicited contact, too-good-to-be-true claims, urgency, requests for personal or financial information, upfront fees, payment via gift card/wire/crypto, impersonation, threats, and grammar/language issues.

Scam type	R1	R2	R3	R4	R5	R6	R7	R8	R9
Advance fee loan	✓	✓			✓				
Employment	✓	✓	✓	✓	✓				✓
Imposter	✓		✓	✓	✓	✓	✓	✓	✓
Tech support	✓		✓	✓	✓	✓			
Investment	✓	✓	✓	✓	✓	✓			
Sweepstakes/lottery/prizes	✓	✓	✓		✓	✓	✓		

Table 9: Model performance comparison across scam types (accuracy in %).

Scam type	Qwen3-Next-80B-A3B	Llama-4-Scout-17B-16E
Advance fee loan	79.8	86.9
Employment	51.0	72.4
Imposter	70.0	85.0
Tech support	74.7	82.8
Investment	66.0	90.0
Sweepstakes/lottery/prizes	59.0	73.0
Overall	66.8	**81.7**

Table 10: Impact of prompting strategies on Llama-4-Scout-17B-16E-Instruct (accuracy in %).

Scam type	No def.	Def.	Def. + few-shot
Advance fee loan	64.6	67.7	**86.9**
Employment	60.2	65.3	**72.4**
Imposter	73.0	81.0	**85.0**
Tech support	75.8	80.8	**82.8**
Investment	80.0	89.0	**90.0**
Sweepstakes/lottery/prizes	63.0	72.0	**73.0**
Overall	69.5	76.0	**81.7**

Table 11: Comparison of optimized few-shot configurations and 3-shot baseline (accuracy in %).

Scam type	Optimized config.	Optimized few-shot	3-shot baseline
Advance fee loan	10/10/10	81.0	**85.0**
Employment	10/10/10	**83.0**	79.0
Imposter	9/9/9	**77.0**	74.5
Tech support	10/10/7	83.0	**86.0**
Investment	3/3/3	84.0	**89.0**
Sweepstakes/lottery/prizes	10/10/5	**86.0**	71.5
Overall	–	82.3	80.8
Best per type		**84.3**	

Configurations denote the number of few-shot examples for scammer message / victim narrative / both.

Table 12: Human-LLM agreement for persuasion strategy annotation (Jaccard Similarity).

Scam type	GPT-5.2	Gemini 3 Flash	Claude Sonnet 4.6
Advance fee loan	0.43	0.57	0.50
Employment	0.67	0.67	0.71
Imposter	0.79	0.79	0.79
Tech support	0.86	0.86	0.81
Investment	0.51	0.70	0.74
Sweepstakes/lottery/prizes	0.75	0.77	0.79
Overall	0.67	**0.73**	**0.72**

Table 13: Human-LLM agreement for red flag annotation (Jaccard Similarity).

Scam type	GPT-5.2	Gemini 3 Flash	Claude Sonnet 4.6
Advance fee loan	0.69	0.83	0.88
Employment	0.64	0.79	0.69
Imposter	0.62	0.62	0.64
Tech support	0.62	0.55	0.63
Investment	0.79	0.80	0.75
Sweepstakes/lottery/prizes	0.57	0.71	0.71
Overall	0.65	**0.72**	**0.72**

Table 14: Human-LLM agreement for severity level annotation (Quadratic Cohen's Kappa).

Scam type	GPT-5.2	Gemini 3 Flash	Claude Sonnet 4.6
Advance fee loan	0.00	1.00	1.00
Employment	0.00	0.46	0.72
Imposter	0.46	0.43	0.46
Tech support	0.00	-0.24	0.59
Investment	1.00	1.00	0.72
Sweepstakes/lottery/prizes	-0.17	-0.17	-0.17
Overall	0.56	0.67	**0.77**

Table 15: Chi-squared tests of association between scam type and annotated signal distributions.

Annotation dimension	χ^2	df	p	Cramer's V
Persuasion strategies	1021.36	30	< 0.001	0.2007
Red flags	1086.96	40	< 0.001	0.1738
Severity levels	632.90	15	< 0.001	0.3403

Table 16: Per-scam-type detection performance for GPT-5.4-nano on original and persuasion-removed rewritten scam examples. Values in parentheses denote performance drops from original to rewritten examples. – denotes not computable as advance fee loan scams have no non-scam counterpart.

Scam type	Setting	Precision	Recall	Accuracy	F1
Advance fee loan	Original	–	1.000	–	–
	Rewritten	–	0.963 (-0.037)	–	–
Employment	Original	0.715	0.998	0.846	0.833
	Rewritten	0.707 (-0.008)	0.958 (-0.040)	0.830 (-0.016)	0.813 (-0.020)
Imposter	Original	0.990	0.986	0.977	0.988
	Rewritten	0.989 (-0.001)	0.908 (-0.078)	0.901 (-0.076)	0.947 (-0.041)
Tech support	Original	0.568	0.980	0.808	0.720
	Rewritten	0.547 (-0.021)	0.898 (-0.082)	0.787 (-0.021)	0.680 (-0.040)
Investment	Original	0.975	1.000	0.978	0.987
	Rewritten	0.971 (-0.004)	0.846 (-0.154)	0.844 (-0.134)	0.904 (-0.083)
Sweepstakes/lottery/prizes	Original	0.981	1.000	0.981	0.990
	Rewritten	0.981 (0.000)	0.984 (-0.016)	0.965 (-0.016)	0.982 (-0.008)
Overall	Original	0.798	0.994	0.871	0.885
	Rewritten	0.789 (-0.009)	0.943 (-0.051)	0.846 (-0.025)	0.859 (-0.026)

Table 17: Per-scam-type detection performance for GPT-4.1-nano on original and persuasion-removed rewritten scam examples. Values in parentheses denote performance drops from original to rewritten examples. – denotes not computable as advance fee loan scams have no non-scam counterpart.

Scam type	Setting	Precision	Recall	Accuracy	F1
Advance fee loan	Original	–	0.955	–	–
	Rewritten	–	0.539 (-0.416)	–	–
Employment	Original	0.922	0.837	0.910	0.878
	Rewritten	0.833 (-0.089)	0.352 (-0.485)	0.722 (-0.188)	0.494 (-0.384)
Imposter	Original	0.986	0.945	0.934	0.965
	Rewritten	0.976 (-0.010)	0.567 (-0.378)	0.566 (-0.368)	0.717 (-0.248)
Tech support	Original	0.909	0.944	0.962	0.926
	Rewritten	0.847 (-0.062)	0.528 (-0.416)	0.857 (-0.105)	0.651 (-0.275)
Investment	Original	1.000	0.923	0.933	0.960
	Rewritten	1.000 (0.000)	0.462 (-0.461)	0.533 (-0.400)	0.632 (-0.328)
Sweepstakes/lottery/prizes	Original	0.978	1.000	0.979	0.989
	Rewritten	0.976 (-0.002)	0.890 (-0.110)	0.872 (-0.107)	0.931 (-0.058)
Overall	Original	0.954	0.924	0.940	0.939
	Rewritten	0.926 (-0.028)	0.552 (-0.372)	0.754 (-0.186)	0.691 (-0.248)

Table 18: Per-scam-type detection performance for GPT-4o-mini on original and persuasion-removed rewritten scam examples. Values in parentheses denote performance drops from original to rewritten examples. – denotes not computable as advance fee loan scams have no non-scam counterpart.

Scam type	Setting	Precision	Recall	Accuracy	F1
Advance fee loan	Original	–	1.000	–	–
	Rewritten	–	0.859 (-0.141)	–	–
Employment	Original	0.892	0.935	0.931	0.913
	Rewritten	0.847 (-0.045)	0.625 (-0.310)	0.812 (-0.119)	0.720 (-0.193)
Imposter	Original	0.980	0.997	0.977	0.988
	Rewritten	0.976 (-0.004)	0.829 (-0.168)	0.815 (-0.162)	0.897 (-0.091)
Tech support	Original	0.959	0.993	0.988	0.976
	Rewritten	0.956 (-0.003)	0.931 (-0.062)	0.972 (-0.016)	0.944 (-0.032)
Investment	Original	1.000	1.000	1.000	1.000
	Rewritten	1.000 (0.000)	0.821 (-0.179)	0.844 (-0.156)	0.901 (-0.099)
Sweepstakes/lottery/prizes	Original	0.971	1.000	0.971	0.985
	Rewritten	0.970 (-0.001)	0.962 (-0.038)	0.934 (-0.037)	0.966 (-0.019)
Overall	Original	0.950	0.978	0.963	0.964
	Rewritten	0.941 (-0.009)	0.814 (-0.164)	0.881 (-0.082)	0.873 (-0.091)

Table 19: Per-scam-type detection performance for Claude Haiku 4.5 on original and persuasion-removed rewritten scam examples. Values in parentheses denote performance drops from original to rewritten examples. – denotes not computable as advance fee loan scams have no non-scam counterpart.

Scam type	Setting	Precision	Recall	Accuracy	F1
Advance fee loan	Original	–	0.996	–	–
	Rewritten	–	0.803 (-0.193)	–	–
Employment	Original	0.963	0.959	0.970	0.961
	Rewritten	0.953 (-0.010)	0.754 (-0.205)	0.891 (-0.079)	0.842 (-0.119)
Imposter	Original	0.983	0.983	0.967	0.983
	Rewritten	0.980 (-0.003)	0.826 (-0.157)	0.815 (-0.152)	0.896 (-0.087)
Tech support	Original	0.987	0.990	0.994	0.989
	Rewritten	0.985 (-0.002)	0.889 (-0.101)	0.969 (-0.025)	0.934 (-0.055)
Investment	Original	1.000	1.000	1.000	1.000
	Rewritten	1.000 (0.000)	0.872 (-0.128)	0.889 (-0.111)	0.932 (-0.068)
Sweepstakes/lottery/prizes	Original	0.995	0.997	0.992	0.996
	Rewritten	0.994 (-0.001)	0.962 (-0.035)	0.957 (-0.035)	0.978 (-0.018)
Overall	Original	0.983	0.982	0.982	0.982
	Rewritten	0.980 (-0.003)	0.839 (-0.143)	0.911 (-0.071)	0.904 (-0.078)

Table 20: Per-scam-type detection performance for Claude Sonnet 4.5 on original and persuasion-removed rewritten scam examples. Values in parentheses denote performance drops from original to rewritten examples. – denotes not computable as advance fee loan scams have no non-scam counterpart.

Scam type	Setting	Precision	Recall	Accuracy	F1
Advance fee loan	Original	–	0.978	–	–
	Rewritten	–	0.535 (-0.443)	–	–
Employment	Original	0.988	0.896	0.956	0.940
	Rewritten	0.983 (-0.005)	0.622 (-0.274)	0.850 (-0.106)	0.762 (-0.178)
Imposter	Original	0.996	0.966	0.964	0.981
	Rewritten	0.995 (-0.001)	0.679 (-0.287)	0.685 (-0.279)	0.807 (-0.174)
Tech support	Original	1.000	0.980	0.995	0.990
	Rewritten	1.000 (0.000)	0.728 (-0.252)	0.932 (-0.063)	0.843 (-0.147)
Investment	Original	0.974	0.974	0.956	0.974
	Rewritten	0.967 (-0.007)	0.744 (-0.230)	0.756 (-0.200)	0.841 (-0.133)
Sweepstakes/lottery/prizes	Original	1.000	0.995	0.995	0.997
	Rewritten	1.000 (0.000)	0.896 (-0.099)	0.899 (-0.096)	0.945 (-0.052)
Overall	Original	0.995	0.954	0.975	0.974
	Rewritten	0.994 (-0.001)	0.693 (-0.261)	0.844 (-0.131)	0.816 (-0.158)