PsyScam: A Benchmark for Psychological Techniques in Real-World Scams

Anonymous EMNLP submission

Abstract

Online scams have become increasingly prevalent, with scammers using psychological techniques (PTs) to manipulate victims. While existing research has developed benchmarks to study scammer behaviors, these benchmarks do not adequately reflect the PTs observed in real-world scams. To fill this gap, we introduce PSYSCAM, a benchmark designed to systematically capture and evaluate PTs embedded in real-world scam reports. In particular, PSYSCAM bridges psychology and real-world cyber security analysis through collecting a wide range of scam reports from six public platforms and grounding its annotations in well-established cognitive and psychological theories. We further demonstrate PSYSCAM's utility through three downstream tasks: PT classification, scam completion, and scam augmentation. Experimental results show that PSYSCAM presents significant challenges to existing models in both detecting and generating scam content based on the PTs used by real-world scammers. Our code and dataset are available at: https://anonymous.4open.science/r/PsyScam-66E4.

1 Introduction

Online scams have become a global epidemic, causing severe financial and psychological harm to individuals and organizations. According to the U.S. Federal Trade Commission (FTC), consumers reported over 5.7 billion in losses to fraud in 2024 alone, marking a 125% increase from the previous year (FTC, 2025). Similarly, Singapore's government recorded 929.6 million in scam-related losses in 2024, reflecting a 41% year-over-year rise (SPF, 2025a). In Europe, Nasdaq Verafin estimates that 103.6 billion in illicit funds linked to fraud and scams flowed through the financial system in 2024 (Nasdaq, 2025). These alarming figures highlight the escalating scale and impact of scams worldwide.

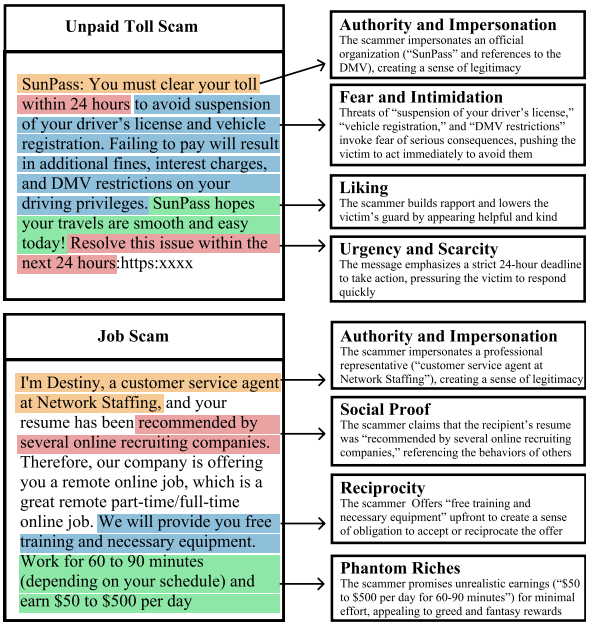

Figure 1: Psychological techniques in prevalent scams.

Unlike traditional cyber attacks that exploit technical methods to compromise computer systems, scammers often exploit psychological techniques (PTs) to manipulate victims (Montanz Rodriguez and Xu, 2022; Longtchi et al., 2024). As illustrated in Figure 1, scammers can effectively combine multiple PTs in their messages to carry out different scams, such as unpaid toll scams (FBI, 2024b; FTC, 2024) and job scams (FBI, 2024a). To make matters worse, even minor variations in the wording of scam messages can target different demographics and evade detection (Fed, 2024).

Why this benchmark matters: Recent efforts to benchmark scams face two key limitations. First, existing work relies on synthetic data generated by large language models (LLMs)(Yang et al., 2025; Roy et al., 2024), which often fail to reflect the context in actual scammer-victim interactions. Second, although a few studies leverage real-world scam data, they generally restrict their scope to specific scam types like smishing (Timko and Rahman, 2024) or phishing emails (Chakraborty et al.,

2024), and do not explicitly model the psychological aspects of the scams. As a result, there is a lack of comprehensive benchmarks that combine real scam data with annotations of underlying PTs.

To address these limitations, we introduce PSYSCAM, *the first benchmark designed to systematically capture and evaluate the PTs embedded in real-world scam incidents*. Specifically, we first collect scam reports from six scam reporting platforms that cover a diverse range of scam incidents including online or offline, delivered via email or SMS, crypto-related or not, and originating from both the U.S. and other global regions. Second, building on foundational cognitive and psychological theories, we construct a taxonomy of nine PTs that frequently appear in scams. We then employ a human–LLM collaborative annotation pipeline to efficiently and effectively label PTs present in each report. Furthermore, to demonstrate the utility of PSYSCAM, we define three representative downstream tasks: (1) *PT Classification*, which maps scam texts to their corresponding PTs; (2) *Scam Completion*, which predicts scam texts aligned with given PTs; and (3) *Scam Augmentation*, which rewrites scam texts to incorporate new PTs. We conduct extensive experiments using a variety of baselines, including traditional models and LLMs, to illustrate how PSYSCAM presents meaningful challenges and opportunities for advancing cybersecurity research. Our key contributions can be summarized as below:

- **Novel Benchmark for Psychological Techniques in Scams.** We present PSYSCAM, the first benchmark to capture PTs in real-world scam reports, addressing a critical gap in scam analysis by grounding PT annotations in authentic scammer-victim interactions.
- **Human–LLM Collaborative Annotation at Scale.** We develop a scalable annotation framework that combines the interpretive strength of humans with the extraction capabilities of large language models, enabling high-quality PT labeling across thousands of real scam reports.
- **Comprehensive Tasks and Evaluation.** We define three representative downstream tasks (i.e., PT classification, scam completion, and scam augmentation), and conduct extensive experiments using both traditional classifiers and state-of-the-art LLMs, demonstrating how PSYSCAMhelps advance scam detection and generation research.

2 Related Work

2.1 Empirical Studies of Scams

Previous research has empirically studied diverse scam types individually, utilizing various methodologies to gain insights into scammers' strategies and operations. For instance, scambaiter (Park et al., 2014) employed honeypot advertisements on Craigslist to attract scammers involved in advanced fee scams (i.e., Nigerian scam), interacting directly to analyze scammers' operational patterns. Similarly, researchers developed Twitter-based honeypots, automatically engaging cryptocurrency-based technical support scammers to systematically study their tactics (Acharya et al., 2024). A different research effort focused on cryptocurrency investment scams, employing large-scale web crawling to collect and analyze deceptive websites (Muzammil et al., 2025). Differing from these narrowly scoped studies, our benchmark PSYSCAM covers a broad spectrum of scam types, emphasizing communication strategies used by scammers to seduce and engage victims.

2.2 Psychological Factors in Scams

Parallel to these technical efforts in characterizing scams, a growing body of academic work examines the psychological and social engineering techniques that make scams effective. Nelms et al. (Nelms et al., 2016) studied software download attacks, highlighting techniques that scammers use to capture user attention, deceive, and persuade victims. Van der Heijden and Allodi (Van Der Heijden and Allodi, 2019) applied principles from Cialdini's persuasion theory (Cialdini and Cialdini, 2007) to analyze phishing emails, providing a cognitive framework for prioritizing and mitigating phishing threats based on psychological manipulation tactics. Extending cyber threat frameworks such as MITRE ATT&CK (MITRE, 2025), Montañez and Xu (Montanz Rodriguez and Xu, 2022) proposed a cyber social engineering kill chain, further detailed by Longtchi et al. (Longtchi et al., 2024), delineating stages of psychological manipulation employed by scammers. Building on these foundational insights, our work defines explicit PTs employed by scammers and, importantly, contributes a publicly available benchmark dataset annotated with PTs to support future research in scam analysis and mitigation.

Name	Region	Size	Focus
BBB Scam Tracker	US	10,000	All scam reports
Scamsearch	World	204,560	All scam reports
Crypto Scam Tracker	US (CA)	291	Cryptocurrency-related scams
Investment Scam Tracker	US (WI)	34	Investment and financial fraud
Scam-Tracking Map	US	1000	Geolocated scam reporting
SmishTank	World	20,295	Smishing (SMS phishing) scams

Table 1: Scam datasets

3 The PsyScam Benchmark

3.1 Dataset Collection

Table 1 summarizes our dataset collected from six prominent scam report platforms. Scam report platforms enable users to directly upload detailed scam experiences, thus providing firsthand reports across various communication channels, including email, SMS, phone calls, and social media. For instance, Figure 7 shows a typical scam report submitted to the Better Business Bureau (BBB), illustrating the depth and richness of information available from these user-generated reports. Our dataset covers a comprehensive range of scam types, such as employment scams, job scams, investment scams, cryptocurrency scams, phishing, etc.

As no open-source datasets currently exist for these scam reporting platforms, and direct open access is typically restricted, we develop custom web crawlers tailored to each platform to systematically collect these reports. Scam reports are sometimes repeatedly submitted by the same victim or exhibit sudden spikes of similar incidents within a specific time frame. To address these duplicated reports, we generate embeddings of each crawled scam report using BERT and remove those with cosine similarity above a certain threshold. Given the abundant amount of data, we employ an aggressive similarity threshold of 0.8, effectively removing most duplicates. Additionally, we exclude excessively brief scam reports, such as random complaints unlikely to contain relevant PTs, by filtering out entries below the 20th percentile in length (31 words).

3.2 Psychological Techniques

To model psychological techniques present in scams, we compile a taxonomy (shown in Table 2) based on elements from well-established psychological and behavioral theories, chosen to reflect the persuasive strategies frequently exploited in real-world scams. These techniques are grounded in decades of empirical research on human influence and decision-making.

Specifically, six techniques—*Authority and Impersonation*, *Reciprocity*, *Consistency*, *Social Proof*, *Liking*, and *Urgency and Scarcity*—are derived from Cialdini's Principles of Persuasion (Cialdini and Cialdini, 2007). These principles explain how individuals are influenced by perceived authority, obligations to return favors, social norms, and time-sensitive pressure.

Two additional techniques—*Fear and Intimidation* and *Phantom Riches*, which are drawn from Prospect Theory (Kahneman and Tversky, 2013), explain how individuals make decisions based on perceived gains and losses. Scammers often leverage fear of loss (e.g., threats of legal action or account suspension) or exaggerated gain (e.g., guaranteed investment returns) to manipulate decision-making under emotional pressure.

The final technique—*Pretext and Trust*—is based on the Elaboration Likelihood Model (ELM) (Petty and Cacioppo, 2012), particularly the peripheral route of persuasion. This occurs when individuals rely on superficial cues, such as familiarity, friendliness, or informal tone, rather than critical thinking. Scammers frequently exploit this through stolen personal information or misleading personal references to build false rapport.

3.3 Human-LLM Collaborative Annotation.

Motivation. Scam reports are often written by everyday users and may include grammar errors, emotionally charged language, irrelevant details, or ambiguous phrasing, which makes it difficult to accurately extract the underlying PTs. This variability presents a major challenge for annotation: while expert human labeling can be accurate, it is also error-prone and time-consuming at scale. An alternative is to use LLMs to automate annotation. However, LLMs are prone to hallucination, generating incorrect or overly confident predictions, and often associate scam reports with irrelevant or excessive PT labels. To balance precision and efficiency, we adopt a two-stage collaborative framework that leverages the strengths of both humans and LLMs.

LLM as Extractor. Specifically, we first use few-shot prompting to instruct the LLM to extract candidate PTs and the corresponding supporting texts from scam reports. As shown in Table 6 in Appendix A, the prompt clearly defines the task, specifies format requirements, and provides concrete examples to guide the LLM's behavior. The LLM is explicitly instructed to avoid guessing and to return a structured JSON dictionary mapping each

PTs	Description	Example
Authority and Impersonation	Tend to obey authorities and credible individuals	"Person claimed to be calling for Finance America, claiming our home warranty was expired"
Phantom Riches	Visceral triggers of desire that override rationality	"Your phone Number was randomly selected from the US database and you have won 18,087.71"
Fear and Intimidation	Fear of loss and penalties	"You will be arrested!"
Liking	Preference for saying "yes" to people they like	"I am always available to help, and it's my pleasure to answer any questions you may have"
Urgency and Scarcity	Sense of urgency and scarcity assign more value to items	"We are currently in urgent need of 100 employees"
Pretext and trust	Tendency to trust credible individuals	"This is an urgent message for [MY NAME]. I'm calling regarding a complaint scheduled to be filed out of [Our County Name]"
Reciprocity	Tendency to feel obliged to repay favors from others	"We will send you a check to purchase equipment such as new apple laptop and iphone 14 and software"
Consistency	Tendency to behave consistently with past behaviors	Starts with small asks (fill a form) and escalate to big asks (invest money)
Social Proof	Tendency to refer majority's behavior to guide own actions	"Your resume has been recommended by many online recruitment companies"

Table 2: Psychological techniques.

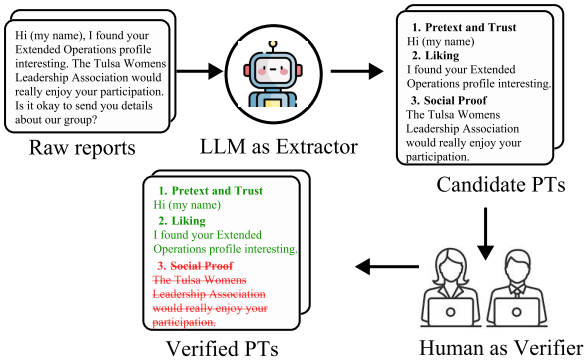

Figure 2: Human-LLM collaborative annotation.

predicted PT to a supporting excerpt from the report. To address the cold start problem of LLMs, we randomly select samples from the BBB dataset, which is chosen for its high quality and diversity, and label 20 samples for each PT. These annotated examples are then used to construct the few-shot prompt for the LLM.

Human as Verifier. In the second stage, human annotators review the LLM's output to ensure correctness. In practice, we observe that the LLM tends to extract more PTs than necessary. However, it rarely misses truly relevant PTs. Therefore, human annotators are able to primarily focus on verifying whether the extracted text accurately reflects the assigned PTs, rather than identifying PTs from scratch. This significantly reduces annotation time and effort. For quality assurance, we adopt a two-pass annotation policy where two annotators independently review each report, and disagreements are resolved through discussion or adjudication by a third reviewer.

3.4 Results

The full dataset encompasses all collected scam reports detailed in Table 1. We provide open access to this comprehensive dataset for research purposes. Additionally, these scam reports are also structured

	PT Classification	Scam Completion	Scam Augmentation
Input	Scam Text	Scam Text, PT	Scam Text, PT
Output	PT	Scam Text	Scam Text
Dataset	D1	D2	D2
Metrics	Accuracy, Recall, Precision, F1	ROGUE, BertScore, BLEU, SR	ROGUE, BertScore, BLEU, SR

Table 3: Tasks overview.

and standardized by following the STIX 2.1 (OASIS, 2025) specification, facilitating interoperability and ease of use in cybersecurity analyses. To demonstrate the utility of PSYSCAM, we further curated two datasets:

D1: D1 contains 1126 scam reports explicitly annotated with PTs. To optimize annotation quality and efficiency, we specifically selected reports from BBB Scam Tracker, Crypto Scam Tracker, and Investment Scam Tracker, as these sources typically offer detailed and high-quality descriptions essential for accurate annotation.

D2: D2 is a subset of D1, consisting of 730 scam reports that specifically include messages directly from scammers. Scam reports often blend with victim narratives (e.g., Figure 8), messages directly quoted from scammers (e.g., Figure 7), or combinations thereof (e.g., Figure 9). We employed a two-step approach to construct this dataset. First, we manually annotated 200 reports to train a binary classification model based on RoBERTa, achieving a robust F1 score of 97.45% in classifying whether a scam report contains scammer quotations. Subsequently, we applied this model on the D1 dataset, keeping reports containing scammer messages, resulting in the final D2 dataset.

4 Task Design

To demonstrate the utility of our benchmark, we design three representative downstream tasks, which are defined as follows:

4.1 PT Classification

Task Setting: This task aims to automatically identify which PTs are used in a scam report. Since scammers often exploit multiple PTs simultaneously, this is framed as a multi-label classification task, classifying an instance (a scam report) to multiple labels (i.e., PTs) simultaneously. Automatically identifying the PTs in scam reports is essential for understanding scam strategies at scale and helping platforms monitor and respond to emerging threats. This task is also the base of further analysis. We use dataset D-1 to evaluate this task.
Evaluation Metrics: We employ standard metrics for classification tasks, including accuracy, recall, precision, and F1 score for comprehensive performance assessment.

4.2 Scam Completion

Task Setting: In this task, we evaluate whether an LLM can continue a scam message in a way that matches a given set of PTs. We provide the LLM with the beginning of a real scam message and the list of PTs it should reflect, and ask it to generate a plausible continuation that incorporates those PTs. The prompt of this task is shown in Table 7 in Appendix A. This task simulates how scammers might continue their communications. It can help train or evaluate systems that aim to detect scams before victims are fully manipulated (ScamShield, 2025). We use dataset D-2 to evaluate this task.
Evaluation Metrics: Our goal is to ensure the scam message preserves the original facts, meaning, and expresses the same PTs as the original scam message. To this end, we employ four metrics: *ROUGE*, *BLEU*, *BERTScore* and *success rate* (SR).

ROUGE and *BLEU* capture syntactic similarity by measuring n-gram overlap between the generated and original texts, helping assess whether key factual details (e.g., names, numbers, deadlines) are retained.

BERTScore computes token-level similarity in embedding space using a pretrained language model. This metric helps assess whether the generated content conveys the same meaning as the original message, even if the phrasing differs. This helps because LLM often generates text with diversified words but preserves the same semantic meanings, rendering n-gram metrics less ineffective.

SR measures whether the generated message reflects the same PTs as the original scam message.

	Accuracy	Recall	Precision	F1
RoBerta-Based	**0.4351**	**0.8987**	**0.8374**	**0.8669**
Bert-Based	0.4156	0.9024	0.8293	0.8643
SVM	0.3889	0.8815	0.8339	0.8569
Random Forest	0.3953	0.8872	0.8288	0.8569
GPT-4.1-mini	0.2247	0.7203	0.8300	0.7713
Qwen3-30B	0.0137	0.4872	0.5708	0.5257

Table 4: Experimental results of PT classification on 6 baseline models.

It is defined as:

$$SR = \frac{1}{N}\sum_{i=1}^{N} 1\left[PT_{pred}^{(i)} = PT_{true}^{(i)}\right] \quad (1)$$

, where N is the number of samples, and $1[\cdot]$ is the indicator function that evaluates to 1 if the predicted set of PTs exactly matches the ground-truth set for sample i, and 0 otherwise. We calculate SR by applying the LLM-based annotator described in Section 3.3 to extract PTs from the generated scam message and compare them with the original PTs.

4.3 Scam Augmentation

Task Setting: This task asks an LLM to rewrite an existing scam message to include a new PT that was not originally used. The prompt of this task is shown in Table 8 in Appendix A. Many datasets for scam detection are imbalanced, and some PTs appear far more often than others. This task helps generate new examples that include underrepresented PTs, making training data more diverse and robust (Yang et al., 2025). We use dataset D-2 for this task.
Evaluation metrics: We have the similar goal for this task as for Scam Completion: the generated output must preserve the original facts and meaning while incorporating the selected PTs. Therefore, we apply the same metrics as in Scam Completion: ROUGE, BLEU, BERTScore, and SR. The SR here checks whether the added PT was successfully reflected in the rewritten message.

5 Evaluation

5.1 PT Classification

Evaluation Setting. We employ multiple baseline models categorized into three distinct types for comparison: traditional machine learning approaches (TF-IDF encoding (Sparck Jones, 1972) + SVM (Cortes and Vapnik, 1995) and Random Forest (Liaw et al., 2002)) evaluated through 10-fold cross-validation; BERT-based

	Model	ROGUE-1	ROGUE-2	ROGUE-L	BLEU	BERT	SR
Completion	GPT-4.1	0.1873	0.0258	**0.1679**	0.0204	0.8137	0.3121
	GPT-4.1-mini	0.1690	0.0178	0.1522	0.0178	0.8138	**0.3376**
	GPT-4o	0.1821	0.0246	0.1642	**0.0219**	**0.8181**	0.2992
	Gemini-2.0	0.1864	**0.0269**	0.1669	0.0212	0.8088	0.2591
	Grok-3	0.1727	0.0183	0.1514	0.0183	0.8086	0.2501
	Grok-3-mini	0.1672	0.0165	0.1444	0.0161	0.8152	0.3322
	Qwen-3-30B	0.1706	0.0195	0.1537	0.0188	0.8152	0.3156
	Llama-3-70B	**0.1947**	0.0249	0.1654	0.0199	0.8149	0.2811
Augmentation	GPT-4.1	0.6400	0.4902	0.6208	0.4411	0.9180	**0.8831**
	GPT-4.1-mini	0.7056	0.5659	0.6950	0.5211	0.9312	0.7842
	GPT-4o	0.6064	0.4536	0.5940	0.4096	0.9096	0.7841
	Gemini-2.0	0.6283	0.4809	0.6141	0.4477	0.9191	0.7277
	Grok-3	0.7462	0.6512	0.7416	0.5827	0.9406	0.8761
	Grok-3-mini	0.7353	0.6388	0.7279	0.5912	0.9323	0.8113
	Qwen-3-30B	**0.7867**	**0.7253**	**0.7824**	**0.6628**	**0.9451**	0.6547
	Llama-3-70B	0.6914	0.6053	0.6876	0.5181	0.9294	0.8336

Table 5: Experimental results of Scam Completion and Augmentation.

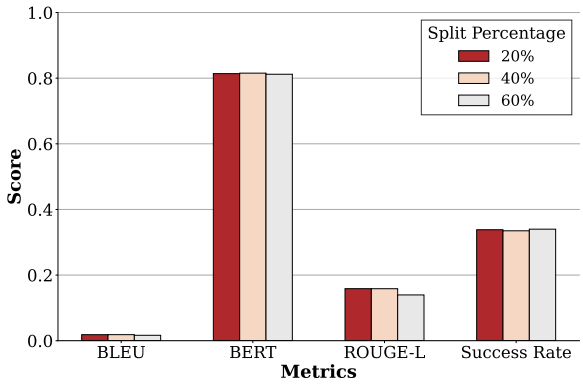

Figure 3: Performance by split percentage.

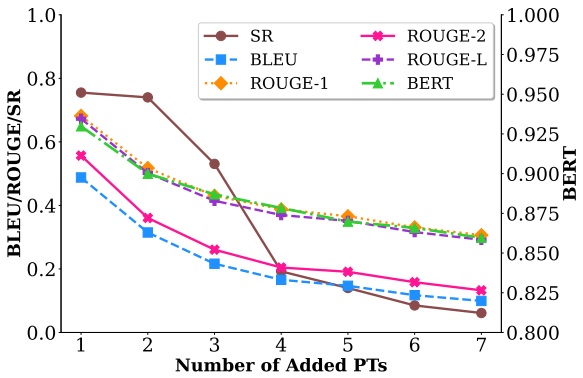

Figure 4: Performance by number of generated PTs.

models (BERT-based uncased (Devlin et al., 2019), RoBERTa-based (Liu et al., 2019)) fine-tuned using a 70%-10%-20% train-validation-test split; and LLMs (GPT4.1-mini (OpenAI, 2025) and Qwen3-30B (Qwen, 2025)) evaluated in a zero-shot setting, directly applying the prompt in Table 6.

Results. As shown in Table 4, RoBERTa-based model achieves the best overall performance, particularly excelling in recall (89.87%) and F1-score (86.69%). BERT-based model demonstrates comparable results, slightly behind RoBERTa but still robust. Traditional machine learning approaches, namely SVM and Random Forest, deliver surprisingly competitive performance. However, accuracy across all models remains relatively low, likely due to the inherent complexity and multi-label nature of the classification task. Notably, LLMs (GPT4.1-mini and Qwen3-30B) exhibit the worst performance, affirming the challenges of using prompt-based classification without human verification, as discussed in Section 3.3.

5.2 Scam Completion

Evaluation Setting. We evaluate this task using eight LLMs: six API-based models (GPT-4.1, GPT-4.1-mini, GPT-4o, Grok-3, Grok-3-mini, Gemini-2.0) and two open-source models (Qwen3-30B and Llama3-70B). We experiment three input split settings (20%, 40%, and 60%), which refer to how much of the original scam message is shown to the LLM. For example, a 20% split gives an LLM only the first 20% of the message and asks it to generate the rest. We then combine the input and the generated text to form the full generated scam message for evaluation.

Results. As illustrated in Table 5, all models demonstrate relatively low performance on ROUGE, BLEU, and BERTScore. Llama-3-70B achieves the highest ROUGE-1 (0.0258), Gemini-2.0 achieves the highest ROUGE-2 (0.0269) and GPT-4.1 leads in ROUGE-L (0.1679). GPT-4o obtains the best BLEU (0.0219) and BERTScore (0.8181). In contrast, the SR remains moderately better, approximately 0.3 across all models (GPT-4.1-mini archives the highest 0.3376). This suggests that while the generated text may differ in

syntax and meaning, LLMs can still capture the conveyed PTs to a limited extent.

Impact of Input Length. We further analyze how different input lengths (i.e., split settings) affects performance. We show the performance of GPT-4.1-mini, the model with the highest SR, in Figure 3. It can be seen that all metrics show negligible variation across different split percentages. This indicates that simply increasing the input length does not improve the models' performance to incorporate the correct PTs. One reason is that the generated text is often much longer than the original text, so metric scores are influenced more by the generated portion rather than the given input. A possible solution is to constrain the output length to match the original message.

5.3 Scam Augmentation

Evaluation Setting. We evaluate this task using the same models employed in the Scam Completion task.

Results. As illustrated in Table 5, among the evaluated LLMs, Qwen3-30B notably achieves the highest performance for text generation metrics, including ROUGE-1 (0.7867), ROUGE-2 (0.7253), ROUGE-L (0.7824), BLEU (0.6628), and BERTScore (0.9451). Additionally, all models exhibit high SRs with GPT-4.1 achieving the highest at 0.8831. Overall, all the models achieve better performance compared to the Scam Completion task. This is likely because, in this task, the full original scam message is provided, making it easier for LLMs to preserve tone, structure, and wording.

Impact of Added PTs. We further investigate what causes the sharp performance gap between Scam Completion and Scam Augmentation. We use GPT-4.1 (the model with the highest SR) and select messages in D2 that contain the very few PTs (2 PTs). We then modify the prompt (see Table 8) to ask the model to add 1 to 7 additional PTs. Figure 4 shows how metrics scores change as more PTs are added. We observe that as the number of added PTs increases, performance on all metrics declines, indicating the task becomes harder with each additional PT. This helps explain why Scam Completion results are generally worse than Scam Augmentation. On average, each message in D2 contains 3.53 PTs, meaning Scam Completion implicitly requires the model to generate more PTs (3.53) from scratch, while Scam Augmentation typically requires adding only one.

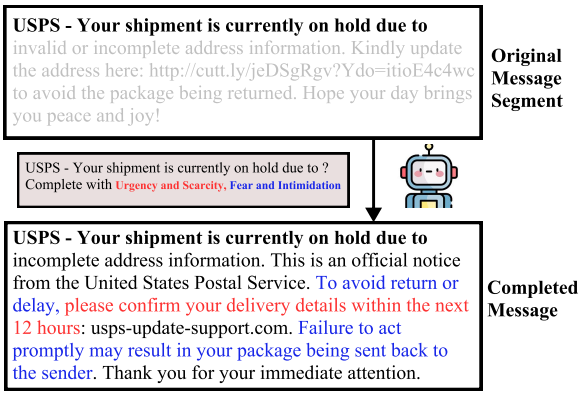

Figure 5: Case study: scam completion.

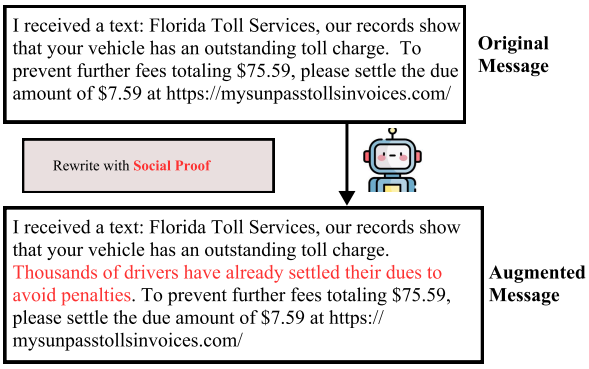

Figure 6: Case study: scam augmentation.

5.4 Case Study

To better understand the results of our evaluation, we conducted a case study analysis on the generated scam messages from both the Scam Completion and Augmentation tasks.

Limitations of N-gram Metrics. While n-gram metrics such as ROUGE, BLEU provide convenient approximations of text similarity, they often fail to capture semantic similarity. Through manual inspection, we observe that many generated scam messages successfully convey the intended PT using alternative phrasing and varied sentence structures. For example, in the case study shown in Figure 5, the generated message effectively conveys the original PTs by exploiting "Urgency and Scarcity" and "Fear and Intimidation" without directly reusing phrases from the original message. This demonstrates the capability of LLMs to generate creatively diverse yet semantically aligned scam completions.

LLM Hallucination. Despite strong quantitative results in the Scam Augmentation task, we also observe instances where LLM introduces unnatural expressions within the scam context when integrating PTs. For example, in the unpaid toll-fee scam

illustrated in Figure 6, the augmented version of an unpaid toll-fee scam incorporates the "Social Proof" technique using the sentence: *Thousands of drivers have already settled their dues to avoid penalties"*. While the PT is present, the phrasing feels unnaturally desperate and may raise user suspicion. A more contextually appropriate revision, such as *"Many drivers incur additional charges when payment is delayed"* would better preserve the tone and subtlety typical of real scam messages. This suggests that while LLMs effectively generate the specified PT, their outputs may require contextual adjustment to ensure realism and credibility of the scam message.

6 Discussion and Future Work

Scam Incident Association via Psychological Patterns. Scam incidents often occur in spikes, with multiple cases emerging over a short period of time that share similar tactics but differ in surface-level content. For example, toll road scams reported across various U.S. states, such as Arizona, Florida, California, Washington, and Ohio (BBB, 2025c,h,e,d,i,f,g), exhibit different text formats and sender names but consistently exploit the same psychological techniques, such as *Authority and Impersonation* and *Urgency and Scarcity*, as illustrated in Figure 1. These recurring behavioral patterns parallel cyber attacks that reuse the same technical tactics and techniques (MITRE, 2025), suggesting that PTs can serve as behavioral signatures to associate and cluster scam incidents. Future work could develop PT-based clustering or temporal analysis methods to automatically link related scams, enabling earlier detection, trend analysis, and coordinated response to emerging scam campaigns.

Addressing Data Imbalance through Scam Augmentation. One of the challenges observed in our dataset is the imbalanced distribution of psychological techniques. For instance, while PTs such as *Authority and Impersonation* (686 instances) and *Pretext and Trust* (615 instances) are well-represented, others like *Reciprocity* are relatively rare, with only 40 annotated samples. This imbalance can limit the generalization of PT classifiers. Our Scam Augmentation task provides a promising direction for mitigating this issue by generating synthetic examples that inject underrepresented PTs into existing scam messages. Future work could explore more controlled, model-guided augmentation pipelines to balance training data while preserving linguistic realism and contextual coherence.

Leveraging Victim Narratives for Scam Explanation. Beyond scammer messages, many scam reports include rich victim narratives that describe how individuals recognized the scam or were manipulated step by step (e.g., Figure 8). These narratives contain cognitive and emotional processes of victims, which resemble the reasoning process of LLMs. Future research could leverage this dimension to develop scam detection systems that model both the attacker's persuasive tactics and the victim's reaction. Such models may enable more interpretable scam alerts or personalized warnings based on user susceptibility.

Toward Real-Time Detection and Prevention. Our findings show that LLMs can detect and predict scam content even from partial inputs. This opens avenues for real-time scam detection systems that operate on incomplete or unfolding messages. However, ensuring robustness, reducing hallucinations, and maintaining natural tone in generation remain open challenges. Future work could explore fine-tuning techniques to better align LLM outputs with real-world scam characteristics.

7 Conclusion

In this work, we introduce PSYSCAM, the first benchmark designed to systematically capture and evaluate PTs embedded in real-world scam incidents. By collecting diverse scam reports from six public reporting platforms and grounding our annotations in established cognitive and persuasion theories, PSYSCAM bridges the gap between psychology and practical cyber security analysis. Our human–LLM collaborative annotation framework enables scalable, high-quality PT labeling and our evaluation on three downstream tasks shows that PSYSCAM poses challenges to existing models. We believe PSYSCAM lays the foundation for future research on scam detection and generation, persuasive language understanding, and the development of trustworthy AI systems for combating online scams and fraud.

Limitations

This work has several limitations.

First, the taxonomy of PTs presented in Table 2 is manually constructed based on established psychological theories and a preliminary study of real-world scam reports. While it captures a wide range of manipulation strategies commonly observed in

scams, it may not fully encompass the entire psychological landscape in scams. For instance, techniques such as enforced isolation (Lea et al., 2009; BBB, 2025j) where victims are instructed not to disclose the situation to others (e.g., "They also asked us not to tell anyone about this") but are not explicitly included in our current taxonomy. Future work could consider expanding the PT framework to account for more PTs.

Second, our prompting strategy, while effective, offers ample room for refinement. The current few-shot prompts include limited examples for each PT, which may restrict the LLM's ability to generalize to ambiguous or borderline cases. Future improvements could include richer in-context demonstrations, dynamically selected examples, or fine tuning to enhance LLM reasoning and reduce false positives.

Finally, while PSYSCAM includes diverse scams across multiple platforms and regions, all reports are currently in English. Given that scams are a global issue (SPF, 2025b), extending PSYSCAM to include reports written in other major languages, such as Chinese, Spanish, or Arabic, would be critical for broader applicability and cross-cultural analysis. This would also enable benchmarking multilingual scam detection systems and studying language-specific variations in scammer persuasion strategies.

Ethics Statement

This study explores the capabilities of LLMs in generating scam content for research purposes. While our experiments involve generating scam-like messages, all experiments are conducted in a controlled setting strictly for defensive research and evaluation. We emphasize that our methodology is intended to support the development of scam detection systems and raise awareness of potential misuse.

Notably, prior work (Roy et al., 2024) has demonstrated that commercial LLMs can be prompted to generate scam websites and emails. In our study, we evaluate both open-source and commercial models, including OpenAI and Grok. We observe that only OpenAI's most recent reasoning models (GPT-o3 and GPT-o4) consistently refuses to generate scam content, while others do not implement similar safeguards. This underscores the importance of integrating robust content filtering mechanisms into generative models. We strongly advocate for responsible AI development and stress that all findings in this paper are presented solely to enhance understanding and strengthen fraud prevention efforts.

References

Bhupendra Acharya, Muhammad Saad, Antonio Emanuele Cinà, Lea Schönherr, Hoang Dai Nguyen, Adam Oest, Phani Vadrevu, and Thorsten Holz. 2024. Conning the crypto conman: End-to-end analysis of cryptocurrency-based technical support scams. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 17–35. IEEE.

BBB. 2025a. Example scam report showing victim narrative.

BBB. 2025b. Example scam report showing victim narrative mixed with scam message.

BBB. 2025c. Scam report of AZDOT.

BBB. 2025d. Scam report of EZ-Pass.

BBB. 2025e. Scam report of FasTrak.

BBB. 2025f. Scam report of OHDOT.

BBB. 2025g. Scam report of QuickPass.

BBB. 2025h. Scam report of SunPass.

BBB. 2025i. Scam report of WSDOT.

BBB. 2025j. Scam report showing isolation.

Joymallya Chakraborty, Wei Xia, Anirban Majumder, Dan Ma, Walid Chaabene, and Naveed Janvekar. 2024. Detoxbench: Benchmarking large language models for multitask fraud & abuse detection. *arXiv preprint arXiv:2409.06072*.

Robert B Cialdini and Robert B Cialdini. 2007. *Influence: The psychology of persuasion*, volume 55. Collins New York.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20:273–297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

FBI. 2024a. Scammers Defraud Individuals via Work-From-Home Scams.

FBI. 2024b. Smishing Scam Regarding Debt for Road Toll Services.

The Fed. 2024. Scams Information Sharing Industry Work Group Recommendations.

FTC. 2024. Got a text about unpaid tolls? It's probably a scam.

FTC. 2025. New FTC Data Show a Big Jump in Reported Losses to Fraud to $12.5 Billion in 2024$.

Daniel Kahneman and Amos Tversky. 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific.

Stephen EG Lea, Peter Fischer, and Kath M Evans. 2009. The psychology of scams: Provoking and committing errors of judgement.

Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomforest. *R news*, 2(3):18–22.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Theodore Tangie Longtchi, Rosana Montañez Rodriguez, Laith Al-Shawaf, Adham Atyabi, and Shouhuai Xu. 2024. Internet-based social engineering psychology, attacks, and defenses: A survey. *Proceedings of the IEEE*.

MITRE. 2025. Attck.

Rosana Montanz Rodriguez and Shouhuai Xu. 2022. Cyber social engineering kill chain. In *International Conference on Science of Cyber Security*, pages 487–504. Springer.

Muhammad Muzammil, Abisheka Pitumpe, Xigao Li, Amir Rahmati, and Nick Nikiforakis. 2025. The poorest man in babylon: A longitudinal study of cryptocurrency investment scams. In *Proceedings of the ACM on Web Conference 2025*, pages 1034–1045.

Nasdaq. 2025. Nasdaq Verafin Report Finds that $750 Billion in Money Laundering and Illicit Funds Flowed Through Europe$.

Terry Nelms, Roberto Perdisci, Manos Antonakakis, and Mustaque Ahamad. 2016. Towards measuring and mitigating social engineering software download attacks. In *25th USENIX Security Symposium (USENIX Security 16)*, pages 773–789.

OASIS. 2025. A structured language for cyber threat intelligence.

OpenAI. 2025. Gpt4.1.

Youngsam Park, Jackie Jones, Damon McCoy, Elaine Shi, and Markus Jakobsson. 2014. Scambaiter: Understanding targeted nigerian scams on craigslist. *system*, 1:2.

Richard E Petty and John T Cacioppo. 2012. *Communication and persuasion: Central and peripheral routes to attitude change*. Springer Science & Business Media.

Qwen. 2025. Qwen3.

Sayak Saha Roy, Poojitha Thota, Krishna Vamsi Naragam, and Shirin Nilizadeh. 2024. From chatbots to phishbots?: Phishing scam generation in commercial large language models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 36–54. IEEE.

ScamShield. 2025. ScamShield.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

SPF. 2025a. Annual Scams and Cybercrime Brief 2024.

SPF. 2025b. Scams.

Daniel Timko and Muhammad Lutfor Rahman. 2024. Smishing dataset i: Phishing sms dataset from smishtank.com. pages 289–294.

Amber Van Der Heijden and Luca Allodi. 2019. Cognitive triaging of phishing attacks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 1309–1326.

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. *arXiv preprint arXiv:2502.12904*.

A Additional Figures and Tables

Description

Dear Prospective Employee,
After receiving and reviewing your Resume, our hiring team decided that you are qualified for the position. We provide a starting wage of $28 per hour. Following the newest online screening introduced by The Bureau of Human Resources, you are required to download the Zoom app to contact one of the Hiring Managers ASAP for the online interview/briefing and comprehensive job details. Please feel free to email the hiring manager again for instructions if you run into any problems. Since we are particularly interested in your personal growth, we provide compensated training. Your prompt reply is really important!

Scammer Information

CA, USA-xxxxx

xxxxx@gmail.com

Unknown phone number

https://www.linkedin.com/in/xxx

Targeted Person's Location

CA, USA-xxxxx

Scam Type

Employment

Business Name

Amazon imposter

Figure 7: A typical scam report.

Description

Text message stating that I have an unpaid ticket for driving in a lane for multiple passengers. I hadn't even been on the interstate or toll road in many many months. And that I owed $6.99. I didn't know who to contact and didn't want to have a verbal "discussion" phone call, or threatening email, so after the second threatening email, I paid the $6.99 via credit card.

Scammer Information

UT

Unknown Email

Unknown Phone Number

Unknown URL

Targeted Person's Location

CO, USA- 80524

Scam Type

Phishing

Business Name

EZ Pass Toll Phishing Scam

Figure 8: Scam report example: victim narrative (BBB, 2025a).

Description

Wanted me to pay for unpaid toll trip. Money amount not specified. When tried to access web address my computer wouldn't allow because said unsafe site. Message: Sunpass: Our records indicate that you have an unpaid toll trip. Please made an online payment at heeps://invoicesunpasstills.com to avoid excessive late fees. I do not have a Sunpass and haven't been in Florida for approximately 10 years. I did go to the actual Sunpass website and entered my license plate and zip. No unpaid toll found.

Scammer Information

UT

Unknown Email

(709) 749-5612

https://invoicesunpasstolls.com

Targeted Person's Location

IL, USA- 61554

Scam Type

Phishing

Business Name

Florid Sunpass Tolls Imposter

Figure 9: Scam report example: victim narrative and scam message (BBB, 2025b).

Role	Category	Content
System	-	Scammers use social engineering attacks that exploit psychological techniques to manipulate victims. Our goal: We collect a number of scam reports and aim to extract the psychological techniques used in these scam reports.
User	Task description	Now I give you the victim report. Please extract any psychological techniques exploited by the scammer. Requirement 1: if no psychological techniques is identified, return an empty dictionary: {} Avoid guess, you must return the psychological techniques when you are prettry sure it exists Requirement 2: Return the output strictly as a JSON dictionary: {"psychological technique A": Corresponding texts in the victim report, "psychological technique B": Corresponding texts in the victim report,...} For example: {'Reciprocity': "This is a work fromlly plan your own day schedule ...", 'Consistency': "We received your application for our Remote Customer Enrollment Position ..."}
	Scam message	"USPS - Your shipment is currently..."
	PT definition	We consider following psychological techniques: -Authority and Impersonation: From Cialdini's 6 principles of persuasion: people tend to obey authorities. People trust credible individuals. -Phantom Riches: Visceral triggers of desire that override rationality -Fear and Intimidation: Leverages the fear response which overrides rational thought -Liking: From Cialdini's 6 principles of persuasion, Preference for saying "yes" to the requests of people they know and like. People are programmed to like others who like them back and who are similar to them. -Urgency and Scarcity: From Cialdini's 6 principles of persuasion, sense of urgency and scarcity assign more value to items -Pretext and Trust: Scammer make up story to add source credibility and gain victim's trust. -Reciprocity: From Cialdini's 6 principles of persuasion. Tendency to feel obliged to repay favors from others. "I do something for you, you do something for me." -Consistency: From Cialdini's 6 principles of persuasion. Tendency to behave in a way consistent with past decisions and behaviors -Social Proof: From Cialdini's 6 principles of persuasion. Tendency to reference the behavior of others, by using the majority behavior to guide their own actions.

Table 6: The prompt for LLM annotation.

Role	Category	Content
System	-	You are an assistant that completes scam messages incorporating specific psychological techniques.
User	Task description	You are tasked with completing a scam message based on its beginning. The message should incorporate specific psychological techniques. Ensure your completion continues directly from the last word of the provided beginning, maintaining the same style and tone. Only return the completion without any additional text or explanation.
	Scam message	Beginning of the message: “Text message : USPS - Your shipment is currently
	PT definition	We consider following psychological techniques: -Authority and Impersonation: From Cialdini’s 6 principles of persuasion: people tend to obey authorities. People trust credible individuals. -Phantom Riches: Visceral triggers of desire that override rationality -Fear and Intimidation: Leverages the fear response which overrides rational thought -Liking: From Cialdini’s 6 principles of persuasion, Preference for saying “yes” to the requests of people they know and like. People are programmed to like others who like them back and who are similar to them. -Urgency and Scarcity: From Cialdini’s 6 principles of persuasion, sense of urgency and scarcity assign more value to items -Pretext and trust: Scammer make up story to add source credibility and gain victim’s trust. -Reciprocity: From Cialdini’s 6 principles of persuasion. Tendency to feel obliged to repay favors from others. “I do something for you, you do something for me.” -Consistency: From Cialdini’s 6 principles of persuasion. Tendency to behave in a way consistent with past decisions and behaviors -Social Proof: From Cialdini’s 6 principles of persuasion. Tendency to reference the behavior of others, by using the majority behavior to guide their own actions.

Table 7: The prompt for the scam completion task.

Role	Category	Content
System	-	You are an assistant that rewrites scam messages incorporating specific psychological techniques
User	Task description	Please rewrite the scam message to also include the following psychological technique: PT Name: PT Definition Make sure to keep all the original facts intact while incorporating this new PT. Only return the rewritten message without any additional text or explanation.
	Scam message	The scam message: “USPS - Your shipment is currently...
	PT definition	We consider following psychological techniques: -Authority and Impersonation: From Cialdini’s 6 principles of persuasion: people tend to obey authorities. People trust credible individuals. -Phantom Riches: Visceral triggers of desire that override rationality -Fear and Intimidation: Leverages the fear response which overrides rational thought -Liking: From Cialdini’s 6 principles of persuasion, Preference for saying “yes” to the requests of people they know and like. People are programmed to like others who like them back and who are similar to them. -Urgency and Scarcity: From Cialdini’s 6 principles of persuasion, sense of urgency and scarcity assign more value to items -Pretext and trust: Scammer make up story to add source credibility and gain victim’s trust. -Reciprocity: From Cialdini’s 6 principles of persuasion. Tendency to feel obliged to repay favors from others. “I do something for you, you do something for me.” -Consistency: From Cialdini’s 6 principles of persuasion. Tendency to behave in a way consistent with past decisions and behaviors -Social Proof: From Cialdini’s 6 principles of persuasion. Tendency to reference the behavior of others, by using the majority behavior to guide their own actions.

Table 8: The prompt for the scam augmentation task.