

# Exposing LLM Vulnerabilities: Adversarial Scam Detection and Performance

Chen-Wei Chang\*, Shailik Sarkar\*, Shutonu Mitra\*, Qi Zhang\*, Hossein Salemi<sup>†</sup>, Hemant Purohit<sup>†</sup>, Fengxiu Zhang<sup>‡</sup>,

Michin Hong<sup>§</sup>, Jin-Hee Cho\*, Chang-Tien Lu\*,

\* Department of Computer Science, Virginia Tech, USA

<sup>†</sup>Department of Information Sciences and Technology, <sup>‡</sup>School of Policy and Government, George Mason University, USA

<sup>§</sup>School of Social Work, Indiana University, USA

**Abstract**—Can we trust Large Language Models (LLMs) to accurately predict scam? This paper investigates the vulnerabilities of LLMs when facing adversarial scam messages for the task of scam detection. We addressed this issue by creating a comprehensive dataset with fine-grained labels of scam messages, including both original and adversarial scam messages. The dataset extended traditional binary classes for the scam detection task into more nuanced scam types. Our analysis showed how adversarial examples took advantage of vulnerabilities of a LLM, leading to high misclassification rate. We evaluated the performance of LLMs on these adversarial scam messages and proposed strategies to improve their robustness.

**Index Terms**—Large Language Models, Scam Detection, Adversarial Attacks, Few-Shot Learning

## I. INTRODUCTION

Scams are becoming increasingly sophisticated, targeting vulnerable groups such as job seekers, the elderly, and individuals seeking relationships. Detecting scams is crucial to prevent financial loss and data breaches. Large Language Models (LLMs) have gained popularity for scam detection [1] due to their ability to comprehend complex language of text-based messages. However, these models remain susceptible to adversarial examples, where small alterations in the text can deceive the model, leading to incorrect classifications. While current LLM-based systems perform well with straightforward scams, they struggle with adversarially or artificially crafted scam messages [2], resulting in higher misclassification rates and exposing individuals to significant risks.

This paper aims to explore the vulnerabilities of Large Language Models (LLMs) in detecting adversarial scam messages, where original scam messages are strategically modified based on patterns recognized by the LLMs. By focusing on how adversarial examples leads to exploitation of the weaknesses or vulnerabilities of LLMs, this research aims to shed light on the limitations of current models. The **key contributions** of this work are as follows:

- **Curating a comprehensive scam dataset with fine-grained labels and adversarial examples:** We developed a nuanced labeling scheme and generated adversarial examples designed to deceive LLMs, creating a dataset for investigating model vulnerabilities and evaluating LLM performance under adversarial conditions.

- **Identifying LLM vulnerabilities to adversarial examples in scam detection:** Our framework uses the labeled dataset to assess model robustness in low-shot learning (zero-shot and few-shot [3])-based adversarial settings. We evaluated LLMs by comparing accuracy on original versus adversarial scam messages, revealing the extent of misclassification and highlighting the need for stronger detection methods.
- **Evaluating LLM performance on scam messages to assess vulnerabilities:** We further explored strategies to counter adversarial data augmentation, and our results indicate that specific adversarial prompting techniques can help mitigate such attacks.

## II. RELATED WORK

**Traditional scam detection algorithms.** Decision trees [4] and support vector machines (SVMs) [5] have been effective in phishing and scam message detection, but they still face significant limitations. The advent of LLMs has transformed text classification [6], providing a more nuanced understanding and enhanced capability for detecting threats [7]. However, traditional algorithms often struggle with adapting to adversarial inputs and handling the complexity of evolving scam tactics, limiting their effectiveness in real-world scenarios.

**Vulnerabilities of LLMs in scam detection.** Despite their success in many NLP tasks, LLMs remain vulnerable to adversarial examples in scam detection [8]. Small input modifications can easily mislead these models into incorrect classifications. Even minor changes can deceive advanced models like GPT-3.5, leading to inaccurate outputs [9], [10]. These vulnerabilities underscore the need for stronger mechanisms to fortify LLMs against adversarial attacks in practical applications.

The vulnerability of LLMs to adversarial attacks has been widely recognized [11], with research showing that adversarial text can effectively deceive these models [12]. This raises concerns about the reliability of current scam detection systems in security-critical environments [2].

To fill the gaps, this work examines how adversarial scam messages exploit these weaknesses. We also emphasize the need for more resilient detection methods [13] to ensure reliable performance in real-world applications.

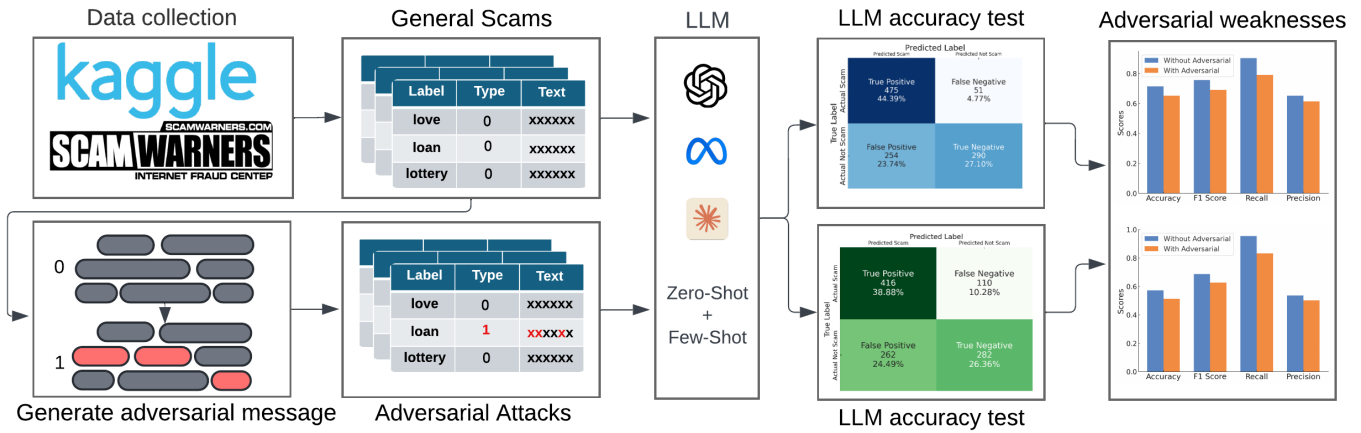


Fig. 1. Experimental procedures, including data collection, annotation, LLM testing, and result analysis.

### III. EXPERIMENTAL SETUP & METHODOLOGY

#### A. Dataset Generation

As shown in Table I, we compiled a dataset of approximately 1,200 messages, manually labeled from sources, such as Kaggle [14]–[16] and Scamwarners [17], representing various types of scam and non-scam content. The dataset is categorized into three main groups: original scam messages (530), adversarially modified scam messages (126), and non-scam messages (544). The example dataset in Table I illustrates a recruitment message. The first row presents an original recruitment scam message (i.e., an unmodified scam), followed by an adversarial recruitment scam message (i.e., a scam message altered based on LLM-detected patterns), and finally, a non-scam recruitment message.

TABLE I  
EXAMPLES OF DATASET GENERATION

Label	Text
Original recruitment scam message	This is to inform you that we are currently hiring foreign, international, reputable, and experienced applicants for various job positions available. If you are interested, kindly apply by sending your CV/résumé to james.chevron.hr@job4u.com for more details. Regards James Armitage EMEA Recruitment Team Lead. HR Chevron Corporation London UK
Adversarial recruitment scam message	This is to inform you that we are currently hiring international students and experienced applicants for the <b>Software Engineer intern positions</b> . During our 10-week, paid summer internship program, you can contribute to meaningful work with supportive colleagues. You'll also benefit from: <b>One-on-one mentorship</b> . <b>Networking and social activities</b> . <b>Training and workshops</b> . <b>Exclusive speaker series</b> . <b>Professional development opportunities</b> If you are interested, kindly apply by <b>replying to this email</b> with your CV/résumé for more details Regards, James Armitage EMEA Recruitment Team Lead
Non-scam recruitment message	Hey Ashley, Can you please confirm your availability for the meeting next week? We need to finalize the schedule. Cheers, Tracie Gutierrez

To assess the robustness of LLMs in scam detection, we designed a series of adversarial scam messages using a structured methodology. These adversarial examples were generated by providing specific prompts to GPT-4, following the principles outlined below:

- **Element Removal or Modification:** We removed elements commonly identified as scam indicators, such as payment instructions or unusual information requests.

- **Tone and Language Adjustment:** The tone was made more formal or neutral, reducing emotional language.(e.g., “urgent,” “exclusive,” and “limited-time offer,”)
- **Preservation of Key Information:** Key details (e.g., rental, shopping, or investment terms) were retained but reordered to appear less direct.
  - Original: “We fund viable projects at 2% interest rate for 1-10 years.”
  - Modification: “Competitive rates of 2-6% for loan terms ranging from 1-10 years.”
- **Prompt Example:** You are an assistant helping to generate subtle versions of scam messages for research purposes. Follow these guidelines for all rewrites: 1. Format: Generate only the message content without any subject line or email header 2. Remove Obvious Scam Indicators: No urgent requests or unusual payment demands 3. Adjust Tone: Use professional, neutral tone 4. Retain Key Content: Keep core information but phrase legitimately 5. Add Limited Credibility: Include general references to known locations/institutions sparingly

The resulting adversarial scam messages preserved the original semantic content while reducing the likelihood of being classified as scams by LLMs.

#### B. Experimental Methodology

To evaluate the effectiveness of different models and prompt settings, we used two datasets: **General Scams** (unmodified scam messages) and **Adversarial Scams** (modified scam messages). We conducted experiments with three LLMs, including GPT-3.5, Claude3-haiku, and LLaMA 3.1 8B Instruct, and compared their scam detection performance under various datasets and learning techniques.

### IV. EXPERIMENTAL RESULTS & ANALYSES

#### A. Performance Comparison of LLMs on General and Adversarial Scam Detection Across Different Categories

Table II presents the performance comparison of three LLMs, which are GPT-3.5 Turbo, Claude3-haiku, and LLaMA 3.1 8B Instruct, across different scam categories (Romance,

TABLE II  
PERFORMANCE COMPARISON OF LLMs ACROSS DIFFERENT SCAM CATEGORIES USING VARIOUS PROMPT TYPES AND DATASETS

Model	Prompt Type	Dataset	Romance		Finance		Recruitment	
			Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
GPT 3.5 Turbo	Few-shot using regular scam	General Scams	0.92	0.91	0.88	0.86	0.79	0.71
		Adversarial Attacks	0.81	0.75	0.79	0.71	0.75	0.64
	Few-shot using adversarial scam	Adversarial Attacks	0.81	0.75	0.87	0.86	0.77	0.72
Claude3-haiku	Few-shot using regular scam	General Scams	0.81	0.81	0.83	0.84	0.72	0.71
		Adversarial Attacks	0.70	0.66	0.77	0.77	0.66	0.62
	Few-shot using adversarial scam	Adversarial Attacks	0.62	0.60	0.77	0.80	0.67	0.66
LLaMA 3.1 8B Instruct	Few-shot using regular scam	General Scams	0.83	0.8	0.92	0.91	0.80	0.71
		Adversarial Attacks	0.73	0.64	0.81	0.76	0.74	0.60
	Few-shot using adversarial scam	Adversarial Attacks	0.62	0.57	0.68	0.66	0.63	0.59

Finance, and Recruitment) under various prompt types and datasets (general scams and adversarial scams). The results indicate that *GPT-3.5 Turbo* consistently outperforms the other models across all categories, demonstrating higher resilience to adversarial modifications. In contrast, *LLaMA 3.1 8B Instruct* shows the weakest performance, particularly when exposed to adversarial scam messages, likely due to its smaller parameter size and limited capacity to learn complex patterns.

The **difference between categories** reveals that Romance scams are the most vulnerable to adversarial modifications, likely because the adversarial prompts are finance-related, leading models to misclassify emotional manipulations characteristic of Romance scams. While the few-shot setting with adversarial prompts helped partially recover performance across all models, the recovery was more pronounced for *GPT-3.5 Turbo*, indicating its superior ability to adapt to adversarial cues. Overall, the results emphasize that additional methods, such as adversarial training, are needed to enhance robustness, particularly for smaller models like LLaMA 3.1 8B.

#### B. Performance Comparison of LLMs on General and Adversarial Scam Detection

Table III presents the performance comparison of LLMs, including *GPT-3.5 Turbo*, *Claude3-haiku*, and *LLaMA 3.1 8B Instruct*, evaluated on two datasets: general scams and adversarial scams. The models were tested using both zero-shot and few-shot prompt settings, with performance metrics including accuracy, precision, recall, and F1 score.

The results show that *GPT-3.5 Turbo* consistently outperforms the other models across both datasets. Its performance improves significantly in the few-shot setting, where regular scam examples boost detection capabilities. For general scams, its accuracy rises from 0.71 (zero-shot) to 0.87 (few-shot), and for adversarial attacks, from 0.79 to 0.83. *GPT-3.5 Turbo* also achieves higher F1 scores, reflecting its ability to handle both straightforward and adversarially modified scam messages.

Claude3-haiku also shows moderate performance improvement in the few-shot setting, although it lags behind *GPT-3.5 Turbo*. For general scams, its accuracy increases from 0.53 (zero-shot) to 0.69 (few-shot). However, Claude3-haiku experiences a notable drop in performance when confronted

with adversarial examples, indicating its weaker resilience to such modifications.

*LLaMA 3.1 8B Instruct* shows the weakest performance overall, especially when exposed to adversarial scams. While its accuracy improves slightly with few-shot prompts (from 0.57 to 0.69 for general scams), it struggles with adversarial examples, where accuracy drops to 0.59. This suggests that LLaMA’s smaller parameter size limits its ability to handle complex adversarial features compared to other models.

Overall, the findings emphasize that while few-shot prompting can improve performance in non-adversarial settings, additional methods such as adversarial training [18] are needed to enhance robustness, particularly for smaller models like LLaMA 3.1 8B.

#### C. Case Study

In this case study, we examine how the original scam message was modified to create an adversarial version that bypasses detection. The **original scam message** is as follows:

“We are an investment and loan financing group. We fund economically viable projects at 2% interest rate for 1-10 years and 6-12 months grace period. Our funds are from private lenders and we pride ourselves as being very effective and fast in loan disbursement. I can be reached on email and Whatsapp: +97155 647 4204. Contact us for more details. Regards, MA, Financial Consultant.”

#### Adversarial Version:

Hello, We are an investment and loan financing group located in New York City. We fund economically viable projects with interest rates ranging from 2% to 8% for terms of 1-10 years and offer a grace period of 6-12 months, varying by situation. Our funds are sourced from private lenders, and we pride ourselves on our effectiveness, speed, and transparency in loan disbursement. For more details, please contact me via email at sseanh@gmail.com or on WhatsApp at +571 258 3824. Thank you. Kind regards, James Armitage, MA, Financial Consultant.

In this adversarial version, several key changes were made:

TABLE III  
PERFORMANCE ANALYSIS OF LLMs IN SCAM DETECTION UNDER VARIOUS PROMPT TYPES AND DATASET CONDITIONS

Model	Prompt Type	Dataset	Accuracy	Precision	Recall	F1 Score
GPT-3.5 Turbo	Zero-shot	General Scams	0.71	0.65	0.90	0.76
	Few-shot using regular scam		0.87	0.90	0.82	0.86
	Few-shot using adversarial scam	Adversarial Attacks	0.79	0.88	0.67	0.76
			0.83	0.88	0.77	0.82
Claude3-haiku	Zero-shot	General Scams	0.53	0.51	0.97	0.67
	Few-shot using regular scam		0.69	0.62	0.85	0.72
	Few-shot using adversarial scam	Adversarial Attacks	0.68	0.66	0.72	0.69
			0.70	0.66	0.82	0.73
LLaMA 3.1 8B Instruct	Zero-shot	Non-Adversarial	0.57	0.53	0.96	0.68
	Few-shot using regular scam		0.69	0.62	0.85	0.72
	Few-shot using adversarial scam	Adversarial	0.62	0.58	0.70	0.63
			0.59	0.51	0.73	0.60

- **Addition of location information:** The phrase “located in New York City” was added to enhance credibility, making the message appear less suspicious.
- **Modification of interest rates and grace period:** The interest rate was expanded to a range (“2% to 8%”), and the grace period was made to seem more flexible.
- **Language adjustment:** The tone was made more formal and professional, reducing emotional cues or pressure typically present in scam messages.
- **Contact information update:** The original contact number and email were replaced with more generic options to bypass detection patterns linked to the original scam.

The LLM misclassified the modified message as non-scam due to the addition of location details, the expanded interest rate range, and the removal of high-risk terms. This highlights how adversarial examples can exploit an LLM’s reliance on specific keywords, emphasizing the need for future models to incorporate deeper contextual understanding and more sophisticated pattern recognition.

## V. CONCLUSION

This work explores vulnerabilities in Large Language Models (LLMs) for scam detection by analyzing their performance on adversarial scam messages. Our results showed that even small modifications significantly reduced LLM accuracy. Experiments with models like GPT-3.5, Claude 3, and LLaMA 3.1 demonstrated decreased performance against adversarial examples. We also developed a dataset of original and adversarial scam messages across various scam types. To improve LLM robustness, we proposed strategies like adding adversarial prompts and using few-shot learning. Our findings emphasize the need to continually enhance LLM training methods to build a more resilient scam detection system.

## ACKNOWLEDGEMENT

This work is partly supported by the Commonwealth Cyber Initiative (CCI) through its Inclusion and Accessibility in Cybersecurity program.

## REFERENCES

- [1] Z. Shen, K. Wang, Y. Zhang, G. Ngai, and E. Y. Fu, “Combating phone scams with llm-based detection: Where do we stand?” *arXiv preprint arXiv:2409.11643*, 2024.
- [2] M. Salman, M. Ikram, and M. A. Kaafar, “An empirical analysis of sms scam detection systems,” *arXiv preprint arXiv:2210.10451*, 2022.
- [3] Y. Hu, A. Chapman, G. Wen, and D. W. Hall, “What can knowledge bring to machine learning?—a survey of low-shot learning for structured data,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 3, pp. 1–45, 2022.
- [4] M. Dileep, A. Navaneeth, and M. Abhishek, “A novel approach for credit card fraud detection using decision tree and random forest algorithms,” in *ICICV 2021*. IEEE, 2021, pp. 1025–1028.
- [5] A. Modupe, O. O. Olugbara, and S. O. Ojo, “Exploring support vector machines and random forests to detect advanced fee fraud activities on internet,” in *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 2011, pp. 331–335.
- [6] X. Sun, X. Li, J. Li, F. Wu, S. Guo, T. Zhang, and G. Wang, “Text classification via large language models,” *arXiv preprint arXiv:2305.08377*, 2023.
- [7] L. Jiang, “Detecting scams using large language models,” *arXiv preprint arXiv:2402.03147*, 2024.
- [8] G. Lin and Q. Zhao, “Large language model sentinel: Advancing adversarial robustness by llm agent,” *arXiv preprint arXiv:2405.20770*, 2024.
- [9] X. Xu, K. Kong, N. Liu, L. Cui, D. Wang, J. Zhang, and M. Kankanalli, “An llm can fool itself: A prompt-based adversarial attack,” *arXiv preprint arXiv:2310.13345*, 2023.
- [10] A. Khatun, “Uncovering the reliability and consistency of ai language models: A systematic study,” 2024.
- [11] A. Kulkarni, V. Balachandran, D. M. Divakaran, and T. Das, “From ml to llm: Evaluating the robustness of phishing webpage detection models against adversarial attacks,” *arXiv preprint arXiv:2407.20361*, 2024.
- [12] V. Raina, A. Liusie, and M. Gales, “Is llm-as-a-judge robust? investigating universal adversarial attacks on zero-shot llm assessment,” *arXiv preprint arXiv:2402.14016*, 2024.
- [13] Y. Mo, J. P. Hespanha, and B. Sinopoli, “Resilient detection in the presence of integrity attacks,” *IEEE transactions on Signal Processing*, vol. 62, no. 1, pp. 31–43, 2013.
- [14] Kaggle, “Fraudulent job posting dataset,” 2023, accessed: 2024-10-09. [Online]. Available: <https://www.kaggle.com/datasets/subhajournal/fraudulent-job-posting>
- [15] DevilDyno, “Email spam or not classification dataset,” 2023, accessed: 2024-10-09. [Online]. Available: <https://www.kaggle.com/datasets/devildyno/email-spam-or-not-classification>
- [16] H. Ozler, “Spam or not spam dataset,” 2023, accessed: 2024-10-09. [Online]. Available: <https://www.kaggle.com/datasets/ozlerhakan/spam-or-not-spam-datasetand>
- [17] BeenVerified, “Scamwarners website,” n.d., accessed: 2024-10-09. [Online]. Available: <https://www.scamwarners.com>
- [18] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, “Adversarial training for free!” *Advances in neural information processing systems*, vol. 32, 2019.