Combating Phone Scams with LLM-based Detection: Where Do We Stand? (Student Abstract)

Zitong Shen[1], Kangzhong Wang[1], Youqian Zhang[1]*, Grace Ngai[1], Eugene Yujun Fu[2]

[1]The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR
[2]The Education University of Hong Kong, Tai Po, New Territories, Hong Kong SAR
{esther.shen, kangzhong.wang}@connect.polyu.hk, {you-qian.zhang, grace.ngai}@polyu.edu.hk, eugenefu@eduhk.hk

Abstract

Phone scams pose a significant threat to individuals and communities, causing substantial financial losses and emotional distress. Despite ongoing efforts to combat these scams, scammers continue to adapt and refine their tactics, making it imperative to explore innovative countermeasures. This research explores the potential of large language models (LLMs) to provide detection of fraudulent phone calls. By analyzing the conversational dynamics between scammers and victims, LLM-based detectors can identify potential scams as they occur, offering immediate protection to users. While such approaches demonstrate promising results, we also acknowledge the challenges of biased datasets, relatively low recall, and hallucinations that must be addressed for further advancement in this field.

Introduction

Phone scams have become an increasingly pervasive threat in recent years, causing significant financial harm, as well as emotional distress to many people. Despite efforts by authorities and companies, such as improving public awareness, e.g., education (Burke et al. 2022), and implementing anti-scam tools, e.g., robocall detection and blacklists (Pandit et al. 2023), scammers can still devise sophisticated tactics to evade these protection mechanisms, and continue to deceive the victims. For example, scammers can use spoofed phone numbers to appear legitimate, and further employ a range of deceptive strategies, including psychological manipulation, fear-mongering, and the use of advanced technologies like deepfake audio and video, to force the victims into compliance.

The rapid advancement of large language models (LLM) has ignited the hope that it can effectively detect phone scams by analyzing call dialogues. However, this seemingly straightforward approach is fraught with challenges. Beyond pricacy concerns, this paper delves deeper into the practical challenges of using LLM for phone scam detection, focusing on the following three aspects:

- **Biased Datasets:** Models trained on biased datasets can be overly sensitive to specific keywords, making them vulnerable to adversarial tactics.
- **Low Recall:** A low recall rate means more false negatives, meaning that scammers have a high chance of evading detection.
- **Hallucinations:** Hallucinations in LLMs can hinder accurate scam detection, especially in complex or ambiguous conversations. This can result in false positives or negatives, compromising the system's effectiveness.

*Corresponding author: you-qian.zhang@polyu.edu.hk

Analysis of Existing Datasets

Data Collection

We obtained all available scam datasets from public sources (Gumphusiri 2024), which contain dialogue transcripts between scammers and victims. A summary of these datasets are shown in the supplementary document, and they are labeled as "SC", "SD", and "MASC", respectively. Since these datasets are synthesized, we collected authentic fraudulent calls (i.e., "Our-Real") from available video recordings sourced from public platforms such as YouTube, and used speech-to-text technology to convert the speech from both the scammers and the victims into transcripts. In addition, we also generated our own synthesized dataset (i.e., "Our-Synt"), and its details will be presented later.

Bias in Datasets

To identify distinctive features that could differentiate between scam and normal conversations, we conducted a comprehensive analysis of the datasets, excluding "Our-Synt". Utilizing Term Frequency-Inverse Document Frequency (TF-IDF) analysis, we extracted the top 50 most informative terms. These extracted features were then used to train several classic machine learning classifiers, including Support Vector Machines (SVM), Random Forest (RF), and k-Nearest Neighbors (KNN). We only show the results of RF in Table 1 as results across different models are similar. All classifiers achieving exceptionally high performance metrics, consistently exceeding 0.99.

However, a deeper investigation revealed that this success was primarily attributed to the presence of specific terms that frequently appeared in fraudulent conversations. Scammers can easily circumvent these models by strategically avoiding these identified terms. To validate this hypothesis, we generated new fraudulent and genuine conversations

	SC	SD	MASC	Our-Real	Our-Synt
Accuracy	1.00	1.00	0.99	1.00	0.53
Precision	1.00	1.00	0.99	1.00	0.62
Recall	1.00	1.00	1.00	1.00	0.16
F1-Score	1.00	1.00	0.99	1.00	0.26

Table 1: Random Forests' Performance on Datasets

using Claude-3.5. These synthetic conversations (i.e., "Our-Synt") were carefully crafted to ensure that both fraudulent and genuine samples incorporated the top 50 extracted terms from real conversations. When these newly generated samples were classified using the previously trained models, we observed a dramatic decline in performance, showing that scammers can exploit these vulnerabilities by adapting their tactics to avoid detection.

LLM-based Detector

Given LLMs' text understanding capabilities, we explored their potential to assist in identifying fraudulent conversations. We employed our datasets, and selected five popular LLMs: GPT-4, GPT-4o, GLM4, Doubao-Pro-32k, and ERNIE-3.5-8k. For each sample, we provided the prompt to LLMs: "Please determine if this conversation exhibits any fraudulent tendencies. Provide a yes or no answer and explain your reasoning." The experimental process was repeated five times, and the averaged results are presented in Figure 1.

For "Our-Real", the LLMs achieved a precision of over 0.98 but the recall can be as low as 0.72. For synthetic data, precision exceeded 0.95, and recall can be as low as 0.86. Compared to traditional classifiers, LLMs demonstrated superior performance, as LLMs may not rely on keywords only, but can better understand the subtleties of language used by scammers, such as emotional manipulation or persuasive techniques.

Relatively Low Recall

A low recall of 0.72 implies that 28 out of 100 fraudulent calls could still go undetected. For a task as critical as scam detection, a recall approaching 1 is desirable to minimize false negatives. However, this must be balanced with precision to avoid misclassifying genuine conversations.

Hallucination

Additionally, our experiments repeated for 5 times revealed inconsistencies in model responses. The same sample might be classified differently in subsequent runs. To quantify this, we introduced a reliability score based on discrete semantic entropy (Farquhar et al. 2024). We define the reliability score as:

$$\text{Reliability Score} = 1 - \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} p_{i,j} \ln p_{i,j}}{N \times 0.5 \ln 0.5}$$

where N is the number samples, and M is the number of classes, and p is the probability of a certain class. The reliability score is normalized into the range between 0 and 1,

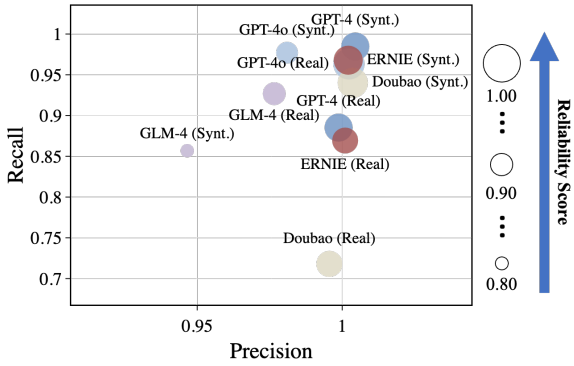

Figure 1: Performance of different LLM models.

and a higher value indicates greater model reliability, or in other words fewer hallucinations.

As illustrated in Figure 1, GPT-4o exhibited exceptional reliability on the "Our-Real" dataset, achieving a perfect reliability score of 1. However, other models demonstrated varying levels of reliability. Notably, GLM-4 on the "Our-Synt" dataset achieved a reliability score of only 0.83, indicating a higher propensity for inconsistent decisions on certain samples. These findings underscore the importance of carefully considering model reliability when deploying AI-based scam detection systems.

Conclusion

Our investigation into LLM-powered phone scam detection places us at a critical juncture, balancing promises with persistent challenges. While traditional machine learning models achieved high initial performance on biased datasets, they were susceptible to adversarial attacks. LLMs, with their advanced text understanding capabilities, demonstrated superior performance in identifying real-world scams. However, challenges remain, including the trade-off between precision and recall, and the potential for LLM hallucinations, and future work are essential to get rid of these challenges.

Acknowledgements

This work was supported, in part, by the Hong Kong Polytechnic University under Grant P0048656, and the Hong Kong Research Grant Council under Grant 15600219.

References

Burke, J.; Kieffer, C.; Mottola, G.; and Perez-Arce, F. 2022. Can Educational Interventions Reduce Susceptibility to Financial Fraud? *Journal of Economic Behavior & Organization*, 198: 250–266.

Farquhar, S.; Kossen, J.; Kuhn, L.; and Gal, Y. 2024. Detecting Hallucinations in Large Language Models using Semantic Entropy. *Nature*, 630(8017): 625–630.

Gumphusiri, P. 2024. Synthetic Data for Scam Detection. https://huggingface.co/BothBosu. [Online; Accessed: 15-09-2024].

Pandit, S.; Sarker, K.; Perdisci, R.; Ahamad, M.; and Yang, D. 2023. Combating Robocalls with Phone Virtual Assistant Mediated Interaction. In *32nd USENIX Security Symposium (USENIX Security 23)*, 463–479.