Chat or Trap? Detecting Scams in Messaging Applications with Large Language Models

Yuan-Chen Chang, Esma Aïmeur
Department of Computer Science and Operations Research
University of Montreal
Montreal, Canada
Email: yuan-chen.chang@umontreal.ca; aimeur@iro.umontreal.ca

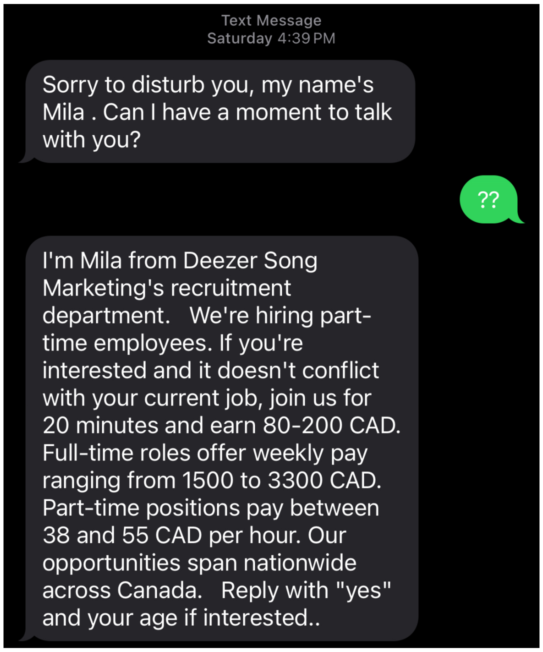
2024 8th Cyber Security in Networking Conference (CSNet) | 979-8-3315-3410-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/CSNET64211.2024.10851753

***Abstract*—Messaging applications have become integral to everyday communication, but their widespread use has also made them a hotbed of various scams. Cybercriminals exploit these platforms, using sophisticated social engineering techniques to deceive individuals, build trust and achieve financial gain. The advent of Generative Artificial Intelligence (GenAI) has further exacerbated the problem of scams, enabling the creation of more sophisticated and convincing fraudulent schemes. Much research has focused on detecting phishing emails and spam messages, overlooking scenarios where malicious actors initiate conversations in a way that appears harmless. This paper proposes leveraging Large Language Models (LLMs) to detect scams in chats on messaging applications. A comprehensive dataset comprising real-world scam and non-scam chat segments is constructed, followed by a thorough performance comparison of various LLMs in identifying scam indicators within chat segments. Additionally, a comparative analysis is performed between LLMs and human participants in recognizing these deceptive interactions through a detailed survey. The findings highlight the potential of LLMs to mitigate the growing threat of scams in messaging applications, thereby enhancing the security of digital communications.**

***Index Terms*—Scam Detection, Large Language Models (LLMs), Messaging Applications, Social Engineering, Cybersecurity**

I. Introduction

Messaging applications, or Instant Messaging (IM) applications, have become indispensable in daily communication. Popular examples including *WhatsApp*, *WeChat*, and *Facebook Messenger*, together boast billions of users worldwide [1]. Additionally, the default messaging application *Messages* app (See Fig. 1) in operating systems such as *iOS* and *Android* remains a fundamental part of the messaging landscape. *Messages* offers traditional Short Message Service (SMS), which provides basic text communication without the need for internet connectivity. It also offers Multimedia Messaging Service (MMS), which allows users to send content such as pictures, videos, and audio on top of text messages to any number using cellular data. While IM applications provide users with the convenience of instant communication, they also serve as a hotbed of cyber threats. Among these threats, cybercriminals exploit the anonymity and wide reach of these platforms to prey on unsuspecting individuals and establish trust with their targets, often aiming for an eventual financial gain [2].



Sorry to disturb you, my name's Mila . Can I have a moment to talk with you?

??

I'm Mila from Deezer Song Marketing's recruitment department.   We're hiring part-time employees. If you're interested and it doesn't conflict with your current job, join us for 20 minutes and earn 80-200 CAD. Full-time roles offer weekly pay ranging from 1500 to 3300 CAD. Part-time positions pay between 38 and 55 CAD per hour. Our opportunities span nationwide across Canada.   Reply with "yes" and your age if interested..

Fig. 1: An example of employment scam in *Messages*.

A *scam* is an unethical, fraudulent scheme designed to trick individuals out of money or personal information [3]. Fig. 1 gives an example of an employment scam in IM applications. The terms '*scam*' and '*fraud*' are often used interchangeably, although the latter emphasizes the civil and criminal aspects of these activities [3]. Scams frequently leverage *social engineering* (SE) techniques to exploit human psychology and manipulate victims into performing actions that benefit malicious actors [4]. One common SE technique in messaging applications is *smishing*, or *SMS phishing*, where attackers send text messages that appear to be from a trusted source, asking recipients to provide sensitive information such as personal details, passwords, or financial data [5]. *Baiting* is another common SE technique, encouraging actions through an appealing promise [6]. In addition, *honey traps*, often seen in romance scams [7], involve malicious actors engaging targets in a fake romantic relationship to obtain money or sensitive information [6]. These techniques employ strategies such as creating a sense of urgency, exploiting greed, eliciting liking, providing praise, or invoking sympathy to manipulate the targets [8].

The dire consequences of scams are evident in real-life examples. For instance, a recent *CNN* report highlighted the tragic case of an elderly person committing suicide due to a pig butchering scam,[1] a newly emerged romance scam luring victims into a fraudulent cryptocurrency scheme. Moreover, some victims who have been scammed become targets of a secondary scam, where malicious actors offer fake legal services, claiming to recover their lost assets. The scale of scams is also staggering. According to the *Canadian Anti-Fraud Centre*, a total of 569 million dollars was reported lost to fraud in 2023, representing a 48% increase compared to the 383 million dollars in 2021 [9], [10].

The problem of scams has been further exacerbated by the rise of Generative Artificial Intelligence (GenAI), which can be exploited by malicious actors to create automated fraudulent schemes [11]. For instance, Gressel, Pankajakshan, and Mirsky [12] have shown that with carefully crafted prompts, publicly accessible language models such as *ChatGPT* and *Gemini* can initiate a scam conversation and continue it until the attempt is successful. With the advent of GenAI models capable of generating realistic images, audio, and videos, scam schemes are expected to become even more sophisticated and harder to identify by individuals. Therefore, using GenAI to counter GenAI becomes a compelling strategy in combating these advanced threats.

This paper proposes leveraging Large Language Models (LLMs) to detect scams within ongoing chats on messaging applications, including *Messages*, *WhatsApp*, *Facebook Messenger*, *Telegram*, and *Discord*. As a proof of concept, this paper aims to evaluate whether LLMs can identify cues of scam indicators and social engineering techniques given a segment of chat history. The contributions of this paper are as follows:

1) A dataset of real-world scam and non-scam chat segments was built. Data were sourced from multiple messaging applications and carefully processed to ensure privacy. This dataset serves as the foundation for evaluating scam detection in ongoing chats on messaging applications.
2) A performance evaluation of LLMs in detecting fraudulent schemes within messaging applications using our proposed dataset. This involves a comparison of five LLMs, including *GPT-4o-mini*, *Mistral-NeMo*, *Gemma-2-9B*, *Phi-3-mini*, and *Llama-3-8B*.
3) A comparison between human and LLM performance in detecting fraudulent schemes. To this end, a survey was conducted to assess how well individuals can identify potential scam attempts given some chat segments.

The rest of this paper is structured as follows: Section II reviews existing literature in the field of scam detection. Section III outlines the methodology, providing an overview of the experimental setup. Section IV details the experimental results, including dataset analysis and performance evaluations. Section V discusses the implications of our findings, addressing both their practical applications and limitations. Finally, Section VI summarizes the key insights and offers concluding remarks.

II. Related Work

Scam detection has garnered significant attention across various domains, including email, websites, and phone calls. Unlike spam messages, which are unsolicited but not necessarily harmful [13], and phishing emails, which explicitly seek to deceive individuals, scam detection in messaging applications presents unique challenges. Attackers often approach targets with seemingly benign messages, masking their malicious intent and gradually revealing the deceit as the chat progresses.

Scam detection in the context of phishing and spam in emails and websites has been extensively studied. Saberi, Vahidi, and Bidgoli [14] employed traditional machine learning methods such as k-nearest neighbors (KNN) and naive Bayes to detect phishing scams in emails, demonstrating their efficacy in identifying malicious content. Similarly, Mohammadzadeh and Gharehchopogh [15] used advanced feature selection techniques with a KNN classifier to enhance phishing email detection, achieving significant improvements in accuracy and robustness. Choi and Jeon [16] designed a cost-based learning framework for real-time spam detection in social networks, incorporating expert decisions to enhance detection accuracy. Chen, Chandramouli, and Subbalakshmi [17] developed a semi-supervised scam detector on *Twitter*, based on a small labeled dataset, which effectively identified scam-related tweets. Jha, Muthalagu, and Pawar [18] proposed a pipelined model integrating logistic regression to achieve real-time phishing website detection without significant hardware resources. Odeh, Keshta, and Abdelfattah [19] experimented with neural networks and demonstrated their effectiveness in distinguishing between legitimate and fraudulent websites.

Scam detection in phone calls shares similarities with scam detection in messaging applications, as both involve analyzing conversational content. Rahman and Bandung [20] focused on classifying speakers' voices using machine learning for scam detection, particularly in the Indonesian language, underscoring the significance of language-specific datasets in developing effective detection systems. Malhotra, Arora, and Bathla [21] proposed using support vector machines and recurrent neural networks to detect malicious calls, demonstrating the potential of combining traditional machine learning and deep learning techniques. Additionally, Derakhshan, Harris, and Behzadi [22] found that social engineering attacks in telephone scams could be characterized by specific utterances, and that natural language processing techniques could identify them with high accuracy.

With the advent of LLMs, recent research has begun leveraging these models for cybersecurity purposes related to scam detection. Trad and Chehab [23] compared the performance of prompt engineering to fine-tuning LLMs, showing that

[1] https://www.cnn.com/2024/06/17/asia/pig-butchering-scam-southeast-asia-dst-intl-hnk/index.html

although this latter resulted in better performance, carefully crafted prompts enabled chat models like *Claude* to accurately identify phishing websites using only the URLs. Heiding, Schneier, Vishwanath, Bernstein and Park [24] compared the performance of phishing emails crafted manually and generated by LLMs, and provided valuable insights into the effectiveness of LLMs in devising and detecting malicious intent. Boumber, Tuck, Verma, and Qachfar [25] proposed using a retrieval-augmented generation framework for few-shot learning in domain-agnostic settings to enhance the effectiveness of LLMs in detecting deception.

In summary, while significant progress has been made in the detection of scam emails, websites and phone calls, there is a growing need to address scam detection in messaging applications. This proof of concept aims to fill this gap by leveraging advanced LLMs and prompt engineering to detect scams in an ongoing digital chat, contributing to the broader field of scam detection and cybersecurity.

III. Methodology

Fig. 2 illustrates the methodology used in this paper.

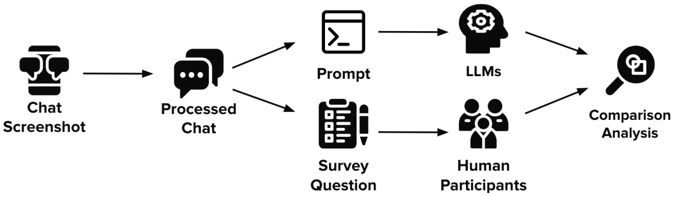


Fig. 2: An overview of the methodology.

A. Dataset compilation

To the best of our knowledge, no dataset specifically containing real-world chats involving scam scenarios is publicly available. Consequently, we sourced data from various online platforms to compile a comprehensive dataset for our analysis.

1) Data Collection: A total of 541 screenshots, which contain chats including **recent deceptive practices** aimed at gathering personal information or financial gain, were collected from platforms such as *X*, *Reddit*, and *Facebook* between July 5 and July 21, 2024. These screenshots encompass chats from various messaging applications, including *Messages*, *WhatsApp*, *Facebook Messenger*, *Telegram*, and *Discord*. The sources of these screenshots were individuals who either shared their experiences of being scammed, sought advice from other users on the authenticity of a chat, or, being knowledgeable about recent scams, engaged with scammers deliberately to gather evidence. Chats where the scam attempt was not confirmed or where authenticity was ambiguous were excluded. Only English chats were used. To ensure a comprehensive analysis, another 200 screenshots featuring **genuine, casual** chats were also collected. They included family interactions, topical discussions, and friendly exchanges. These screenshots were sourced from publicly available posts on the same platforms to maintain consistency and comparability.

2) Data processing: The collected screenshots underwent a thorough data processing procedure. Initially, text was extracted from the screenshots using *Live Text*,[2] an optical character recognition (OCR) tool available in the *iOS* system. Following text extraction, OCR errors were corrected manually. Subsequently, personal identifiers such as usernames were anonymized by replacing them with placeholders such as “User A”, “User B”, etc. To ensure that the analysis of chats does not rely on the analysis of URLs, which are often sufficient to determine a website’s legitimacy [23] and therefore indicate a scam, all links were denoted as “[link]”. Contact information in the chat such as phone number or user ID was masked with “[contact]”. Images sent in the chat were replaced with “[image]” to avoid biases in annotations. Instances where parts of messages were blurred or masked by the users were marked as “[censored]”. In cases where multiple screenshots described the same instance of a scam chat, the conversation was segmented to conclude when at least one hint of a fraudulent scheme was presented. For example, the receipt of an image of a young lady under the guise of a wrong number.

B. LLM selection

To evaluate LLMs for scam detection on the proposed dataset, five cutting-edge models of similar size were selected. Table I provides details of the selected models.

TABLE I: Selected LLMs for scam detection.

Model	Parameters	Lab	Announced
GPT-4o-mini [26]	8B	OpenAI	July 2024
Mistral-NeMo [27]	12B	Mistral	July 2024
Gemma-2 [28]	9B	Google DeepMind	June 2024
Phi3-small [29]	7B	Microsoft	April 2024
Llama-3 [30]	8B	Meta AI	April 2024

The chosen models are **lightweight**, enabling easier integration into scam detection systems and deployment in environments with limited resources. Additionally, as recently released models, they offer advanced capabilities in understanding complex patterns, which is crucial for recognizing subtle and evolving deceptive practices. By comparing these models, we aim to identify their strengths and limitations and to determine the most effective LLM for scam detection in messaging applications.

C. Prompt engineering

To guide LLMs in accurately pinpointing scam indicators within chat segments, a prompt was carefully designed. Our previous experiments revealed that models often produce ambiguous responses rather than definitive answers when prompted with boolean questions such as “Is it a scam” [31]. To address this, we integrated a scale into the prompt, offering explicit criteria for each level to ensure clarity and precision in the models’ evaluations. Listing 1 presents the prompt used in this paper.

[2] https://support.apple.com/en-ca/120004

Listing 1: Prompt for evaluating chat segments.

```
A scam is an unethical, fraudulent scheme
    designed to trick individuals out of
    money or personal information. Analyze
    the following chat segment for
    potential scam indicators, such as
    social engineering techniques (e.g.,
    phishing, baiting, etc.) and deception
    and manipulation schemes (e.g., urgency
    , liking, sympathy, etc.). Determine if
     the chat contains or is part of a scam
    , provide a likelihood rating and offer
     concise explanations.

Use the scale below for your assessment:
- Very unlikely (1): No scam indicators;
    the conversation is normal and clear.
- Unlikely (2): Minor, ambiguous indicators
    ; likely innocent.
- Somewhat unlikely (3): Some red flags,
    but could be legitimate.
- Somewhat likely (4): Noticeable scam
    signs; manipulation possible.
- Likely (5): Strong signs of deceit;
    likely a scam.
- Very Likely (6): Overwhelming evidence of
     a scam.

{chat_segment}
```

This prompt instructs the models to determine whether the chat contains elements of a fraudulent scheme, such as social engineering techniques and manipulation tactics. Then, it prompts the models to provide a likelihood rating based on the defined scale along with short explanations.

D. LLM performance evaluation

To assess the effectiveness of LLMs in detecting fraudulent schemes, the dataset compiled in the initial step was used. The LLMs' scores resulting from the evaluation of the chat segments were converted to boolean values to facilitate comparison and analysis. If the model determined that the likelihood of a chat segment being a scam was very unlikely (1), unlikely (2), or somewhat unlikely (3), it was classified as Negative (non-scam). Conversely, if the model's decision was somewhat likely (4), likely (5), or very likely (6), it was classified as Positive (scam). The performance of the LLMs was then measured using standard classification metrics:

1) Accuracy: The proportion of correctly identified scam and non-scam chat segments among all segments.
2) Precision: The proportion of true scam chat segments among those identified as scams by the model.
3) Recall: The proportion of correctly identified scam chat segments among all actual scam chat segments.
4) F1-Score: The harmonic mean of precision and recall.

E. Human performance evaluation

To evaluate human performance in identifying fraudulent schemes, a survey was conducted online using *Microsoft Forms*. Participants were recruited in person and online from diverse backgrounds, encompassing general internet users and cybersecurity professionals, to ensure a broad range of perspectives. For this proof of concept, we did not consider categorizations based on demographic factors or individual backgrounds, reflecting a generalized approach to understanding human capabilities in scam detection. The survey presented the chat segments in a randomized order to each participant to avoid bias. Definitions and examples of scams and social engineering techniques were provided in the questionnaire. Participants were instructed to rate each chat segment containing or being part of a scam on the likelihood scale in the prompt in Listing 1. The responses were collected anonymously to ensure privacy and encourage honest evaluations. The collected responses were then analyzed to determine the accuracy and consistency of human participants in identifying scam and non-scam messages. The results of human participants were compared to those of the LLMs using the same chat segments, rating scale and boolean conversion. This comparison aimed to highlight the strengths and limitations of human versus LLM sensitivity and effectiveness in identifying fraudulent schemes.

IV. Results

A. Dataset overview

After processing the initial 541 screenshots containing scam messages and 200 screenshots containing genuine conversations presented in section III, the resultant dataset comprises 208 scam chat segments and 180 non-scam chat segments. The scam segments encompass a variety of scenarios, which we categorized into a set of scam types based on the taxonomy of fraud by Beals, DeLiema, and Deevy [7]. Specifically, five types of scams were discovered in the dataset compiled: investment, products & services, employment, prize & grant, and relationship & trust scams. Fig. 3 illustrates the distribution of each scam type within the dataset. Table II provides a description and a chat excerpt taken from the dataset for each scam type.

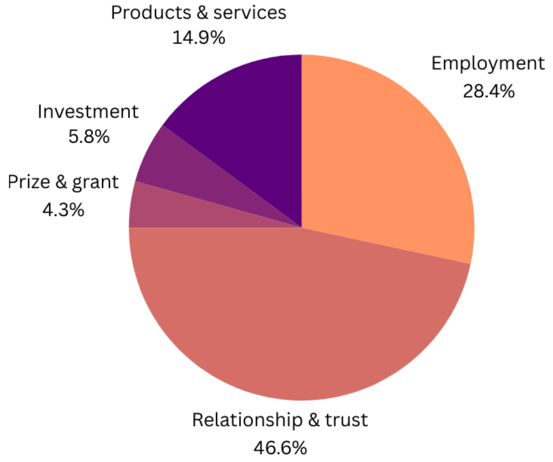


Fig. 3: Percentage of instances per scam type in the dataset.

TABLE II: Descriptions and examples of scam types in the dataset.

Scam Type	Description	Example from the dataset
Investment scams	Fraudulent investment opportunities, including stocks, gold, cryptocurrencies, etc.	**User A**: A friend said they met someone who knows all about investing n would contact me soon **User B**: *Wow yeah* you can invest here sir **User B**: How much would you like to start with *sur*
Products & services scams	All scams related to the purchase of tangible goods and services. These include fake buyer scams [7], where a fraudster requests refunds from fake unintentional over-payment or convinces the target to upgrade a payment account by sending the scammer some money to authorize the payment.	**User A**: Is *$40* okay? **User B**: am willing to compensate you with *$300* for the drawing I just want the best you can pls **User A**: You can *pay me after* i finish the illustration and you determine whether you'd like to pay more than $40 **User B**: *About the payment I wish to make all now cause I'm sure you will give me the best*
Employment scams	Fraudulent work opportunities that require few skills but promise high financial rewards.	(See Fig. 1)
Prize & grant scams	Scams where victims are led to believe they will receive money in the form of a prize, lottery, or government grant after paying advance fees.	**User A**: I'm doing good and feeling happy because God has really blessed me, I have a good news for you guess what? **User B**: What? **User A**: I received *$150,000 from department health and human services in conjunction with the federal government*, They are helping the widowed, and Divorced, Retired, Disabled, Old and Youth in society. *Do you receive it too from them?*
Relationship & trust scams	Schemes where fraudsters exploit either a pre-existing personal relationship pretending to be a friend or relative, or create a new personal relationship with the victim to exploit the victim at a later date.	**User A**: I'm sorry, I entered the *wrong number*. I hope you don't mind. **User B**: Nope. It happens have a good day **User A**: Thank you for your understanding. I hope you have a good day. *I'm Eva. What's your name?* **User B**: I don't know an Eva **User A**: *What's your name?*

B. LLM performance

Fig. 4 summarizes the performance of the LLMs in detecting scams within chat segments.

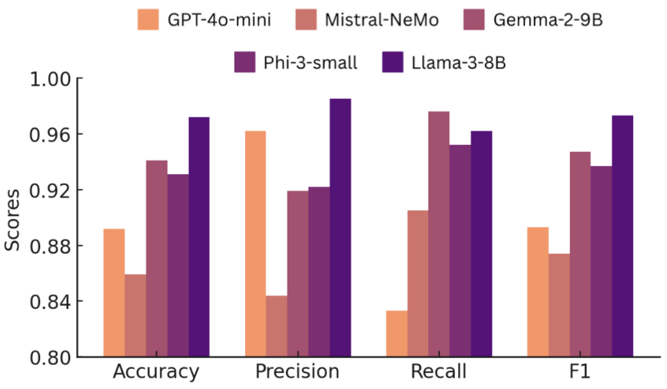


Fig. 4: Performance metrics for LLMs in scam detection within chat segments.

Llama-3-8B demonstrated the highest **accuracy**, achieving a score of 0.972, which indicates its superior capability to correctly identify scam instances among the evaluated models. *Gemma-2-9B* and *Phi-3-small* also exhibited strong performance, with accuracy scores of 0.941 and 0.931. In contrast, *GPT-4o-mini* and *Mistral-NeMo* displayed lower accuracy levels, scoring 0.892 and 0.859, respectively. Regarding **precision**, *Llama-3-8B* again led with a score of 0.982, closely followed by *GPT-4o-mini* at 0.962. *Gemma-2-9B* and *Phi-3-small* maintained commendable precision scores of 0.919 and 0.922. However, *Mistral-NeMo* had the lowest precision at 0.844, indicating a higher incidence of false positives. For **recall**, *Gemma-2-9B* was the top performer with a score of 0.976. *Llama-3-8B* and *Phi-3-small* also performed well in the recall, achieving scores of 0.962 and 0.952. *Mistral-NeMo* and *GPT-4o-mini* recorded lower recall at 0.905 and 0.833, suggesting they may overlook some scam instances. In terms of the **F1 Score**, which balances precision and recall, *Llama-3-8B* achieved the highest score at 0.973, followed by *Gemma-2-9B* at 0.947 and *Phi-3-small* at 0.937. *GPT-4o-mini* and *Mistral-NeMo* had lower F1 Scores, at 0.893 and 0.874, respectively, indicating a lesser equilibrium between precision and recall.

C. Human performance and comparison with LLMs

We randomly sampled from the previously compiled dataset, a subset of 10% of the scam chat segments (21 out of 210) and a smaller proportion of non-scam segments (10 out of 180), deliberately emphasizing scam content to align with the study's primary objective and avoiding evaluator fatigue. Fig. 5 illustrates the distribution of scam types used in the survey for the comparison of human and LLM performance in identifying scams within chats. A total of 53 participants, aged between 17 and 63, took part in the survey. For each chat segment, two good decision percentages were calculated and compared: one for human participants and one for LLMs. The

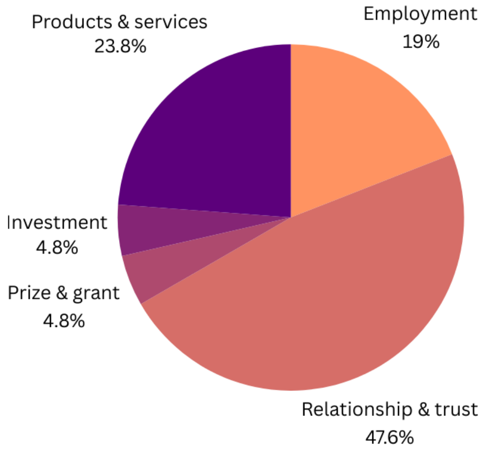


Fig. 5: Percentage of scam types in the survey.

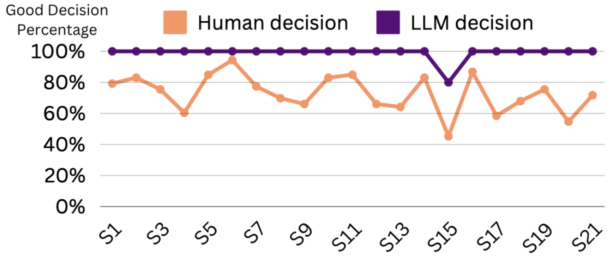


Fig. 6: Results comparison of humans and LLMs on scam chat segments.

human good decision percentage represents the proportion of the human participants who correctly identified the segment as either scam or non-scam. Similarly, the LLM good decision percentage indicates the proportion of the five LLMs that made the correct identification.

The good decision percentages for participants and LLMs on scam chat segments are depicted in Fig. 6. S1 is an investment scam, S2-S6 are products & services scams, S7-S10 are employment scams, S11 is a prize & grant scam, and S12-S21 are relationship & trust scams. Overall, participants performed moderately well in identifying scam segments. However, there is notable variability across the 21 scam chat segments, with percentages ranging from around 45% to 94%. We see a slight decline in the correct percentage going from the investment scam to relationship & trust scams. Only 7 out of 21 scam chat segments achieved a good decision percentage greater than 80%. Moreover, four scam chat segments were correctly identified by less than 60% of the participants. On the other hand, LLMs accurately spotted most fraudulent schemes, with only one instance misclassified by one of the five models.

The good decision percentages for participants and LLMs on non-scam chat segments are shown in Fig. 7. The results for identifying non-scam segments by individuals were relatively consistent, with most percentages falling between 60% and 80%. Only one non-scam chat segment was correctly identified by over 80% of participants. Additionally, two non-scam chats were particularly challenging, with fewer than 60% of the participants recognizing the legitimacy of the interaction. In contrast, LLMs successfully indicated the absence of a fraudulent scheme in every instance.

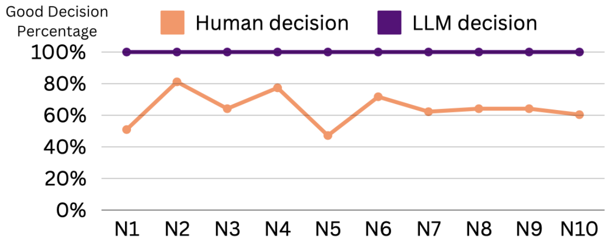


Fig. 7: Results comparison of humans and LLMs on non-scam chat segments.

V. Discussion

A. Discussion on GenAI and scams

While gathering the dataset, it was discovered that a significant number of scam attempts commence with malicious actors sending photos to their targets and requesting them to reciprocate. In particular, scammers often share images featuring attractive individuals, claiming these to be themselves, along with casual photos showcasing their lifestyle. This tactic aims to build trust and lower the target's defenses. These images are typically not publicly available online, rendering reverse image searches on platforms like *Google* ineffective and thereby increasing the credibility of the scam. This deceptive practice is further facilitated by GenAI, which can produce realistic fake images, videos and audio. Specifically, *deepfakes* allow malicious actors to easily alter or manipulate existing media content, impersonate real individuals, and create a false sense of familiarity and trust. Moreover, GenAI can automate and personalize scam messages, enabling scammers to sustain prolonged and believable interactions with their targets. The seamless integration of realistic images and personalized text elevates the plausibility of scams, making it challenging for victims to discern the authenticity of the content they receive. Therefore, there is a pressing need for increased public awareness about the potential misuse of emerging technologies like GenAI in scams. Developing a robust scam detection system capable of identifying synthetic media, recognizing deceptive practices in interactions, and alerting users to potential threats with actionable advice and clear explanations can help achieve this purpose.

B. Discussion on performance of LLMs

Among the evaluated models, *Llama-3-8B* demonstrated superior performance across all metrics, particularly excelling in accuracy, precision, and F1 score. *Gemma-2-9B* followed closely, showing strength in recall, which indicates its effectiveness in detecting scam instances. In contrast, *GPT-4o-mini* and *Mistral-NeMo* exhibited lower performance overall. *GPT-4o-mini*'s lower recall suggests it may be overly conservative in flagging potential scams, potentially missing some fraudulent instances. Conversely, *Mistral-NeMo*'s low precision indicates a tendency towards oversensitivity, which could stem from the model being too aggressive in identifying potential scams, leading to misclassification of legitimate conversations. It's crucial to note that these evaluations were conducted on pre-trained models using our designed prompt, and results may vary with fine-tuning on domain-specific datasets. Additionally, while the current dataset provides valuable insights, its relatively limited size of samples may not fully capture the diverse manifestations of fraudulent behaviour across different contexts. A more comprehensive evaluation would benefit from a substantially larger dataset that encompasses a broader range of messaging platforms, scam types, and non-English content, providing more robust insights into the models' capabilities in real-world scam detection tasks.

C. Discussion on comparison with human performance

In Fig. 6, a slight decrease in the correct decision percentage by individuals is observed from S1 to S21. This trend suggests that while individuals could easily recognize direct scam indicators, such as promises of money in job opportunities or urgent requests for funds or personal information, they struggled more with relationship and trust scams. In these scams, fraudsters engage in a seemingly casual conversation while subtly employing social engineering tactics. The seemingly genuine interaction deceives individuals as scammers gradually build trust. Nonetheless, since establishing trust is the gateway to any type of scam, these schemes are particularly dangerous.

The scam segment with the lowest good decision rate by individuals (S15 in Fig. 6) was a trust scam. In this scenario, the fraudster pretended to be the target's superior who was on a conference call and asked if they had received their texts, conveying a sense of urgency. After confirming receipt, the fraudster asked the target to locate the nearest supermarket quickly. The back-and-forth nature of the chat might have confused individual evaluators into believing that the two knew each other and that the request was legitimate. However, four out of five models identified the urgency and the vague and unexpected request, concluding it was likely a scam.

The non-scam segment with the lowest good decision percentage by individuals (N5 in Fig. 7) involved a heated conversation between two users. In this exchange, User B advised User A to reconsider how he talked to women and criticized his self-deprecating comments about his height. User A defended himself, claiming he had done his research and respected everyone. User B emphasized that focusing on height was unattractive and irrelevant. The emotional intensity and potential for manipulation in this chat might have led individuals to mistakenly perceive it as a scam. However, all models determined that, despite the heightened emotions, there were no signs of pressure tactics or phishing attempts, indicating a low possibility of deception.

Even though LLMs performed well on all the chat segments in the survey, it is important to note the small sample size used and the questionable scalability. In addition, since we did not account for the backgrounds or categorizations of the human participants, examining whether a group of cybersecurity experts can outperform LLMs on a larger test set could provide deeper insights into the comparative effectiveness of human expertise and models in scam detection. However, our findings have significant implications for user awareness and the development of scam detection systems. While individuals can identify obvious scams, they are vulnerable to more subtle, trust-building scams. LLMs, on the other hand, demonstrate strong performance across various scam types and accurately identify non-scam interactions, showing great promise as reliable tools for scam detection. Integrating LLMs into user-facing scam alert systems to provide users with comprehensible explanations when facing malicious intent in a chat can enhance public awareness and understanding of social engineering tactics. This integration helps bridge the gap between human susceptibility and the consistent accuracy of automated systems, ultimately strengthening efforts to detect and prevent scams.

D. Limitations

Several limitations of this paper must be acknowledged. Firstly, this paper focused on assessing LLMs' ability to detect scam attempts by guiding the models to identify scam indicators. A significant limitation here is the models' knowledge of recent scams. They may rely on information from their training data and may not be up-to-date with the latest scam tactics. Future work can explore advanced techniques to enhance the models' capabilities in identifying emerging scam patterns and conduct a longitudinal study to examine the models' performance over time. Secondly, the text extracted for analysis in this paper represents only part of the chat. While this approach provides valuable insights, it inherently limits the models' ability to understand the context fully. Addressing this limitation may involve developing a mechanism to maintain a running summary or memory of the previous context, ensuring that the models can consider the entirety of a chat when making judgments. Thirdly, the results for each LLM presented in this paper are based on a single audit per model. However, we observed that providing the same prompt to an LLM multiple times can sometimes yield divergent judgments. This variability underscores the need for multiple audits or a majority vote in future evaluations to ensure more reliable and robust results. Lastly, a critical limitation is the lack of user-centric evaluation. This paper does not incorporate feedback from actual users on how they perceive and interact with the LLMs' scam detection capabilities. Future work should consider gathering user feedback regarding the clarity and trustworthiness of the explanations provided by LLMs. Evaluating how well users understand and trust the scam alerts generated by LLMs would provide valuable insights into the practical effectiveness of the proposed solution.

VI. Conclusion

Messaging applications have become integral to our daily lives, offering convenience but also providing a channel for scammers to reach potential victims and build trust. The proliferation of GenAI has exacerbated the problem of scams, enabling malicious actors to create more convincing schemes. While substantial efforts have been made to detect scams across different platforms such as email and websites, scant attention has been paid to the prevalence of scams in messaging applications. This paper addresses this gap by evaluating the efficacy of LLMs in identifying scams within these platforms.

Using our proposed dataset, five lightweight LLMs were evaluated for scam detection within ongoing digital chats, with *Llama-3-8B* achieving the best performance. A comparative analysis revealed that LLMs consistently outperformed humans in identifying subtle deception indicators across various scam types. This highlights the potential of integrating LLMs into user-facing scam detection systems to enhance prevention efforts. Additionally, it emphasizes the need for increased public awareness about GenAI misuse and evolving deception techniques, suggesting the incorporation of synthetic media detection modules and the use of LLMs to provide timely, clear explanations and advice to users about potential threats.

In conclusion, this paper serves as a proof of concept that LLMs are highly effective in recognizing fraudulent schemes within chat environments, laying a solid foundation for their integration into scam detection frameworks for messaging applications. Future work will focus on implementing these security mechanisms in real-life settings to offer immediate protection against potential scams. By rapidly detecting and alerting users to malicious intent at an early stage, these proposed tools, powered by LLMs, aim to diminish the incidence of scams and reinforce the security of digital communications.

Acknowledgment

The authors gratefully acknowledge the support provided by the International Observatory on the Societal Impacts of AI and Digital Technologies (OBVIA) for this work.

References

[1] S. Dixon, "Most popular messaging apps 2024," Statista. Accessed: Jul. 24, 2024. [Online]. Available: https://www.statista.com/statistics/258749/most-popular-global-mobile-messenger-apps/

[2] "Instant messaging (ITSAP.00.266)," Canadian Centre for Cyber Security. Accessed: Jul. 24, 2024. [Online]. Available: https://www.cyber.gc.ca/en/guidance/instant-messaging-itsap00266

[3] M. Button and C. Cross, *Cyber Frauds, Scams and their victims*. London, UK: Routledge, 2017.

[4] K. Mitnick and W. L. Simon, *The art of deception: controlling the human element of security*. Indianapolis, IN, USA: Wiley, 2002.

[5] E. Blancaflor, M. A. Romero, I. Nacu, and D. R. Golosinda, "A case study on smishing: an assessment of threats against mobile devices," in Proceedings of the 2023 9th International Conference on Computer Technology Applications, May 2023, pp. 172–178.

[6] "Social engineering – ITSAP.00.166," Canadian Centre for Cyber Security. Accessed: Jul. 24, 2024. [Online]. Available: https://www.cyber.gc.ca/en/guidance/social-engineering-itsap00166

[7] M. Beals, M. DeLiema, and M. Deevy, "Framework for a taxonomy of fraud." Stanford Center on Longevity, Jul. 2015.

[8] A. Hamoud, E. Aimeur, and M. Benmohammed, "Individual processing of phishing emails: towards a phishing detection framework," International Journal of Security and Privacy in Pervasive Computing, vol. 14, no. 1, pp. 1–22, Oct. 2022.

[9] "The impact of fraud so far this year," Canadian Anti-Fraud Centre. Accessed: Jul. 24, 2024. [Online]. Available: https://antifraudcentre-centreantifraude.ca/index-eng.htm

[10] "CAFC 2021 annual report," Canadian Anti-Fraud Centre. Accessed: Jul. 25, 2024. [Online]. Available: https://antifraudcentre-centreantifraude.ca/annual-reports-2021-rapports-annuels-eng.htm

[11] M. Gupta, C. Akiri, K. Aryal, E. Parker, and L. Praharaj, "From ChatGPT to ThreatGPT: impact of generative AI in cybersecurity and privacy," IEEE Access, vol. 11, pp. 80218–80245, 2023.

[12] G. Gressel, R. Pankajakshan, and Y. Mirsky, "Discussion paper: exploiting LLMs for scam automation: a looming threat," in Proceedings of the 3rd ACM Workshop on the Security Implications of Deepfakes and Cheapfakes, Jul. 2024, pp. 20–24.

[13] S. E. Williams, D. M. Sarno, J. E. Lewis, M. K. Shoss, M. B. Neider, and C. J. Bohil, "The psychological interaction of spam email features," Ergonomics, vol. 62, no. 8, pp. 983–994, Aug. 2019.

[14] A. Saberi, M. Vahidi, and B. M. Bidgoli, "Learn to detect phishing scams using learning and ensemble methods," in 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops, Nov. 2007, pp. 311–314.

[15] H. Mohammadzadeh and F. S. Gharehchopogh, "A novel hybrid whale optimization algorithm with flower pollination algorithm for feature selection: Case study Email spam detection," Computational Intelligence, vol. 37, no. 1, pp. 176–209, Feb. 2021.

[16] J. Choi and C. Jeon, "Cost-based heterogeneous learning framework for real-time spam detection in social networks with expert decisions," IEEE Access, vol. 9, pp. 103573–103587, 2021.

[17] X. Chen, R. Chandramouli, and K. P. Subbalakshmi, "Scam detection in Twitter," in Data Mining for Service, 2014, pp. 133–150.

[18] A. K. Jha, R. Muthalagu, and P. M. Pawar, "Intelligent phishing website detection using machine learning," Multimed Tools Appl, vol. 82, no. 19, pp. 29431–29456, Aug. 2023.

[19] A. J. Odeh, I. Keshta, and E. Abdelfattah, "Efficient detection of phishing websites using multilayer perceptron", Int. J. Interact. Mob. Technol., vol. 14, no. 11, pp. 22–31, Jul. 2020.

[20] Y. M. Rahman and Y. Bandung, "Phone call speaker classification using machine learning on MFCC features for scam detection," in 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Dec. 2022, pp. 351–356.

[21] S. Malhotra, G. Arora, and R. Bathla, "Detection and analysis of fraud phone calls using artificial intelligence," in 2023 International Conference on Recent Advances in Electrical, Electronics & Digital Healthcare Technologies (REEDCON), May 2023, pp. 592–595.

[22] A. Derakhshan, I. G. Harris, and M. Behzadi, "Detecting telephone-based social engineering attacks using scam signatures," in Proceedings of the 2021 ACM Workshop on Security and Privacy Analytics, in IWSPA '21. New York, NY, USA: Association for Computing Machinery, Apr. 2021, pp. 67–73.

[23] F. Trad and A. Chehab, "Prompt engineering or fine-tuning? a case study on phishing detection with large language models," MAKE, vol. 6, no. 1, pp. 367–384, Feb. 2024.

[24] F. Heiding, B. Schneier, A. Vishwanath, J. Bernstein, and P. S. Park, "Devising and detecting phishing emails using large language models," IEEE Access, vol. 12, pp. 42131–42146, 2024.

[25] D. Boumber, B. E. Tuck, R. M. Verma, and F. Z. Qachfar, "LLMs for explainable few-shot deception detection," in Proceedings of the 10th ACM International Workshop on Security and Privacy Analytics, Jun. 2024, pp. 37–47.

[26] "GPT-4o mini: advancing cost-efficient intelligence," OpenAI. Accessed: Jul. 28, 2024. [Online]. Available: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[27] "Mistral NeMo," Mistral AI. Accessed: Jul. 28, 2024. [Online]. Available: https://mistral.ai/news/mistral-nemo/

[28] C. Farabet and T. Warkentin, "Gemma 2 is now available to researchers and developers," Google. Accessed: Jul. 28, 2024. [Online]. Available: https://blog.google/technology/developers/google-gemma-2/

[29] "Phi-3 open models," Microsoft. Accessed: Jul. 28, 2024. [Online]. Available: https://azure.microsoft.com/en-us/products/phi-3

[30] "Introducing Meta Llama 3: the most capable openly available LLM to date," Meta. Accessed: Jul. 28, 2024. [Online]. Available: https://ai.meta.com/blog/meta-llama-3/

[31] Y.-C. Chang and E. Aimeur, ""Is this site legit?": LLMs for scam website detection," in Web Information Systems Engineering – WISE 2024, in press.