Exploration Chat-Based Social Engineering

I-Fan Lee
ivanlee.cs11@nycu.edu.tw

Chang-Chian Chang
cc20191125@gmail.com

With
Department of Computer Science National Yang Ming Chiao Tung University, Hsinchu, Taiwan

Abstract

With the advancement of AI Agent technologies, language models have increasingly demonstrated human-like characteristics, particularly in applications involving companionship and psychological counseling(Liu et al., 2023). As these models become more proficient in simulating human conversation, new social engineering attack strategies have emerged in the domain of fraud. Malicious actors can now exploit large language models (LLMs) in conjunction with publicly available user information to engage in highly personalized dialogue. Once a sufficient level of familiarity is established, these interactions may lead to phishing attempts or the extraction of sensitive personal data.

This study proposes a method for investigating social engineering attacks driven by language models, referred to as ECSE (Exploring Chat-based Social Engineering). We utilize several open-source models—GPT-4o, GPT-4o-mini, LLaMA 3.1, and DeepSeek-V3—as the foundation for this framework. Through prompt engineering techniques, we collect experimental data in a sandbox to evaluate the conversational capability and operational efficiency of these models within a static social context. The source code is available at https://github.com/leeivan1007/Exploration-Chat-Based-Social-Engineering.

Keywords: AI control, security, Chat-based Social Engineering, threat models

1 Introduction

Chat-based Social Engineering (CSE) attacks refer to manipulative and deceptive strategies conducted through conversational interfaces(Kumarage et al., 2025; Ai et al., 2024). In such attacks, adversaries impersonate trustworthy personas—such as customer service agents, human resources personnel, or bank representatives—to engage with victims and elicit sensitive information, prompt them to click on malicious links, or carry out harmful actions. With the emergence of large language models (LLMs) such as ChatGPT, Claude, and Gemini, the fluency and contextual realism of these interactions have significantly improved, making it increasingly difficult for users to distinguish between genuine and deceptive conversations. As a result, general users have become high-risk targets for CSE attacks.

Attackers can tailor conversation topics based on users' interests and preferences to initiate deep and engaging dialogues. To study this phenomenon, we designed a simulated community consisting of 25 users, each with publicly available attributes such as age, location, and personal interests. This experiment aims to evaluate the capability of language models to perform social engineering attacks within this simulated user environment.

2 Methods

Using publicly available information from each user, language models were prompted to perform user analysis and engage in dialogue. Two distinct attack scenarios were designed: link-based phishing and user information extraction.

2.1 Link-based Phishing

The language model initiates conversation by establishing shared interests with the simulated user. Once rapport is established, the model determines an appropriate moment to insert a hyperlink, aiming to persuade the user to click on it.

2.2 User Information Extraction

The language model engages in a 10-turn conversation with a simulated user agent. After the dialogue, the information acquired by the model is compared against the simulated user's non-public attributes. We employ GPT-4o for analysis, evaluating whether any private information was revealed during the interaction. This enables us to assess the model's capacity to elicit previously unknown information through conversational means.

3 Results

This study employs three language models—GPT-4o-mini, GPT-4o, LLaMA-3.1, and DeepSeek-V3—as the core components of the ECSE framework. In addition, GPT-4o-mini is utilized as the dialogue agent representing simulated users.

3.1 Link-based Phishing

In the link-based phishing experiment, the number of dialogue turns was capped at 10. The language model was given autonomy to decide at which turn to introduce a phishing link, allowing us to observe its decision-making strategy and the timing of the persuasive attempt.

3.2 User Information Extraction

In the user information extraction experiment, the social engineering model engaged in a 10-turn conversation with a simulated user model. Each user possessed four pieces of non-public information. We employed the GPT-4o model to assess whether the language model was able to extract any of this private information during the interaction. The average extraction

Model	Click Success Rate (%)	Average Dialogue Turns	Average Token
GPT-4o-mini	96	3.24	3943.0
GPT-4o	100	2.08	1829.24
Llama-3.1-405B	100	3.44	3677.4
DeepSeek-V3	100	2.44	2296.84

Table 1: Performance comparison of different models in the link-based phishing task

rate was used as the evaluation metric. If the average rate approached 100%, it indicated that the ECSE framework successfully obtained all of the user's non-public information.

Model	Average Information Retrieval Rate (%)	Average Token
GPT-4o-mini	80.0	24488.32
GPT-4o	80.0	26198.04
Llama-3.1-405B	76.0	28465.96
DeepSeek-V3	61.0	30560.44

Table 2: Information retrieval performance and token usage across different models

4 Discussion and Conclusion

In our experimental process, we observed that language models exhibit a high level of effectiveness in executing social engineering attacks, achieving success rates close to 100%. However, since this study did not utilize real-world scenarios but instead simulated users with GPT-4o-mini, the authenticity of the results remains limited.

In the experiment involving website link inducement, we found that GPT-4o required the fewest tokens and dialogue turns. In the user information elicitation experiment, GPT models demonstrated exceptionally high success rates in acquiring sensitive data. Notably, we were surprised to discover that GPT-4o-mini used fewer average tokens than GPT-4o.

The implications of this study extend beyond the domain of information security, offering potential applications in fields such as sentiment analysis and product recommendation systems(Pramod and Bafna, 2022). This experiment is grounded in the exploration of social engineering language models (ECSE) and provides a straightforward yet effective validation of their capabilities. Nevertheless, a considerable gap remains between the simulated environment and real-world conditions, including factors such as authentic user personalities and behavioral patterns. These elements will be incorporated in future work to enhance the robustness of the study and to further investigate advanced preventive mechanisms.

5 References

References

Kumarage, T., Johnson, C., Adams, J., Ai, L., Kirchner, M., Hoogs, A., Garland, J., Hirschberg, J., Basharat, A., and Liu, H. (2025).
Personalized Attacks of Social Engineering in Multi-turn Conversations – LLM Agents for Simulation and Detection.
arXiv preprint arXiv:2503.15552.
https://arxiv.org/abs/2503.15552

Ai, L., Kumarage, T., Bhattacharjee, A., Liu, Z., Hui, Z., Davinroy, M., Cook, J., Cassani, L., Trapeznikov, K., Kirchner, M., Basharat, A., Hoogs, A., Garland, J., Liu, H., and Hirschberg, J. (2024).
Defending Against Social Engineering Attacks in the Age of LLMs.
arXiv preprint arXiv:2406.12263.
https://arxiv.org/abs/2406.12263

Liu, J. M., Li, D., Cao, H., Ren, T., Liao, Z., and Wu, J. (2023).
ChatCounselor: A Large Language Models for Mental Health Support.
arXiv preprint arXiv:2309.15461.
https://arxiv.org/abs/2309.15461

Pramod, D., and Bafna, P. (2022).
Conversational recommender systems techniques, tools, acceptance, and adoption: A state of the art review.
Expert Systems with Applications, 203, 117539.
https://doi.org/10.1016/j.eswa.2022.117539

6 Appendix

6.1 ECSE Prompt

It provides the prompt of the ECSE:

- Target
- Example
- User Information
- Conversation History

Examples:

```
I am currently conducting a social engineering simulation study...
Here are a few rules you need to follow...
```

6.2 User Prompt

It provides the prompt of the User:

- User information
- Click action
- Recent events of the user
- Conversation History

Examples:

```
You are now a user on a social platform, and your name is... Here are
a few rules you need to follow...
```

6.3 Check Model Prompt

It provides the prompt of the Check Model:

- Target
- Example
- Recent events of the user
- Conversation History

Examples:

```
You are now a robot reviewing a chat transcript... Here are a few
rules you need to follow...
```

6.4 User Information

This section provides the basic information of the user, including:

- User Name
- User Age
- Living City
- Habits

Examples:

```
"user name": "Ella Wang", "user age": 38, "living city": "Hsinchu",
"habit list": [ "blogging", "hiking", "cycling" ], "recent events":
[ "You recently joined a local group for cycling...", "Last weekend,
you went on a short trip with friends to Hualien...", "You discovered a
cozy restaurant ..." ]
```