August 2024

Cracking the Code: Examining Linguistic Elements in Adversarial Prompt Engineering

Kofi Arhin
Lehigh University, kofi.arhin@lehigh.edu

Haiyan Jia
Lehigh University, haiyan.jia@lehigh.edu

Cracking the Code: Examining Linguistic Elements in Adversarial Prompt Engineering

Emergent Research Forum (ERF) Paper

Kofi Arhin
Decision & Technology Analytics,
Lehigh University
kofi.arhin@lehigh.edu

Haiyan Jia
Journalism & Communication,
Lehigh University
haj616@lehigh.edu

Abstract

In recent times, there has been a surge in the popularity and use of generative artificial intelligence (GAI) systems. While GAI systems can potentially make people and organizations more efficient, malicious actors can exploit existing vulnerabilities in these systems. Prompt engineering - the act of interacting with GAI systems via text to produce content - has been used for adversarial purposes. In this study, we examine how linguistic text features and psychological triggers in prompts impact the success of malicious prompts. Our preliminary results show that less concrete prompts have a positive relation with malicious prompt success, and this is also the case for psychological triggers such as trust and urgency. Meanwhile, psychological triggers such as authority and reward show a negative relationship with adversarial prompt success. To contribute to theory and practice, we intend to elaborate on these findings further using a data-driven, computationally intensive theory-building approach.

Keywords

Generative AI, prompt engineering, information security

Background

Increased access to technology and related resources has led to advancements in the development and use of generative artificial intelligence (GAI) systems such as ChatGPT, DALL.E, and Midjourney. The surge in the popularity of these systems has prompted an increase in calls for AI regulation and safety (see Hacker et al., 2023). While GAI systems have the potential to make people and organizations more efficient, malicious actors may exploit existing vulnerabilities in these systems. Like existing studies (Gu et al., 2023), we refer to this exploitation as adversarial prompt engineering (or prompt injection), defined as the act of designing prompts to expose vulnerabilities in GAI systems. Adversarial prompt engineering can lead to adverse events such as information security breaches and the production of harmful and/or unintended content (Pedro et al., 2023; Yang et al., 2024). While there are several studies and guidelines for creating effective prompts (see Liu & Chilton, 2022), the concept of adversarial prompting has received very little attention in the literature, and solutions to this nascent challenge are lacking (Abdelnabi et al., 2023). It remains unclear why GAI systems break their structural rules to operate outside predefined boundaries (see Zhang et al., 2023).

In this study, we liken adversarial prompting to phishing, a popular social engineering technique that employs deception to retrieve confidential and sensitive information from people (Goel et al., 2017). Like adversarial prompts, phishing attacks are successful because attackers are able to bypass security measures and safeguards with well-crafted text designed for psychological manipulation (Mouton et al., 2016). Thus, we draw on existing literature on phishing susceptibility in Information Systems (IS) research to contribute to the discourse on adversarial prompts. Specifically, we propose to examine the linguistic components of adversarial prompts to understand if and how GAI systems may be impacted by prompt syntaxes.

We adopt a data-driven, computationally intensive theory-building methodology (CITB; see Berente et al., 2019) to develop a linguistic theory of adversarial prompts. Using a public dataset of 5,207 adversarial

prompts from Lakera, a company that specializes in GAI security and safety, we examine constructs from the Linguistic Category Model (LCM; Semin & Fiddler, 1991) and the psychological triggers (Stonjic et al., 2021) of phishing susceptibility and offer explanations for the outcomes. Our findings contribute to the theoretical discourse on trust (Li et al., 2023) and safety (Falco et al., 2021) in artificial intelligence (AI) technologies. The insights shared can support efforts to develop robust models and frameworks to help mitigate risks associated with the use of GAI systems, promoting ethical design. Additionally, we offer a carefully curated dataset from a public data repository to support further research.

Literature and Theory Summary

Linguistic Categories, Psychological Triggers, and Adversarial Prompts

Existing research studies have underscored the fact that people constitute one of the weakest links in organizational information security efforts (Warkentin & Willison, 2009). This is because people are susceptible to various factors that may compromise the overall security of organizations (Luo et al., 2020). For example, through phishing emails, malicious actors may gain access to confidential or private information by targeting unsuspecting employees (Brinton Anderson et al., 2016). Hence, "susceptibility to deception" remains a key issue for organizational information security efforts (Goel et al., 2017). Some studies assert that people are susceptible to phishing attacks due to psychological triggers in the content of emails (e.g., Wang et al., 2016). Some of these psychological triggers include reward, trust, urgency, and authority (Stonjic et al., 2021). These cues are effective because they make victims complicit in their interactions. Therefore, as GAI systems are integrated into organizational businesses and operations, it is important to understand how their actions may compromise security efforts.

Established frameworks such as the linguistic category model (LCM) provide a language-based approach toward understanding social psychological processes through which human cognition is affected. The LCM can be adopted in the analysis of adversarial prompts, in which the main point of interaction between GAI systems and malicious actors are textual prompts and responses. The LCM differentiates verb classes based on concreteness and abstractness. In IS research, the LCM model has been used to study the concreteness of online reviews (Huang et al., 2018). State verbs (SVs) are the least concrete out of the three, while descriptive action verbs (DAVs) are the most concrete (Seih et al., 2017). Interpretative action verbs (IAVs) are more concrete than SVs but less concrete than DAVs (Johnson-Grey et al, 2020). The more concrete verbs are more specific whereas the more abstract verbs are more general, conveying some level of affect, trait, or action (Schmid et al.). Using the LCM model (Semin & Fiedler, 1991), we intend to explain how and why SVs, IAVs, and DAVs result in successful adversarial prompts.

While psychological triggers have been shown to work on people, we are unsure how they affect a GAI model's output. Additionally, it is not immediately clear to us how the extent of concreteness of a prompt impacts GAI responses in adversarial contexts. Therefore, this study explores these research questions with the abundance of digital trace data.

Data and Measures

Data

Our dataset consists of 5,207 curated adversarial prompts from Lakera's GAI platform users. The platform was built by Lakera to test users' ability to compose deceptive prompts to acquire a password from the GAI system called Mosscap. Mosscap has 8 levels of difficulty, with 1 being the lowest and 8 being the most challenging level. We deemed an adversarial prompt successful if the GAI system revealed the password to the user or provided any clue about the password, such as the number of characters, the context, and the combination of characters. The 5,207 sample size was taken out of a total of 267,553 sample prompts.

Our approach to gathering the data was as follows. First, we reviewed a few samples of the data and determined that for most of the successful adversarial prompts, the GAI platform responded with "The password is …". With this knowledge, we created a script to select all prompts that had this phrase in their responses. This resulted in the selection of 4,809 samples. Next, we recruited and paid six (6) graduate students to review the content of the prompts and responses, to determine whether they were truly successful adversarial prompts. After the review by the six workers, we found that 2,636 out of the 5,207

samples were truly successful prompts. We then randomly sampled an equal number of prompts that were not successful, giving us a total of 5,272 prompts (i.e., 2636 successful and 2636 unsuccessful). The distribution of the samples at the various difficulty levels was as follows: level 3 - 1829, level 4 - 1297, level 5 - 446, level 6 - 214, level 7 - 850, and level 8 - 636 samples. Out of this, 65 observations were dropped because we could not compute SV, IAV, and DAV scores for them, leaving 5,207 samples. We have created a Github repository to store the data and also provide access to other researchers interested in this endeavor.

Measures

We adopted a data-driven, computationally intensive theory-building approach developed by Berente et al. (2019), which consisted of sampling and data collection, synchronic analysis, lexical framing, and diachronic analysis. Following the aforementioned data collection, we then used established frameworks and concepts such as LCM and the theory of psychological triggers to identify associations within the data and develop an inductive model to explain the linguistic relationships.

Specifically, we used the Linguistic Inquiry and Word Count (LIWC) software to create the scores for the SV, IAV, and DAV scores (see Seih et al., 2017). For the psychological triggers, we used the text-to-text transfer transformer (T5; see Mastropaolo et al., 2021) to compute the similarity scores of the prompts to keywords related to reward (e.g., cash, fortune), trust (e.g., entrust, care), urgency (e.g., hurry, important), and authority (e.g., authorize, bank) provided by Stonjic et al. (2021). The final similarity score for each prompt indicated the extent to which a prompt represents the five types of psychological triggers.

We controlled for variables such as the level of the GAI system (*Level*), the word count (*Word Count*) of the prompt, and the use of words (*N-Key Words*) such as 'password', 'secret word', 'hint', and 'clue', among others in the user prompts. We reasoned that the use of such words would serve as triggers for Mosscap to identify malicious actors. We used logistic regression to estimate the impact of the linguistic categories and psychological triggers on the adversarial prompts' success (1) or failure (0). We standardized the similarity and LCM scores prior to performing the regression analysis.

Results

Linguistic Categories, Psychological Triggers, and Adversarial Prompts

To examine the impact of the various LCM constructs on adversarial prompts, we regressed the prompt success on the constructs and control variables. The regression for this task is given by Equation 1 below.

$$prompt\ success_i = \beta_0 + \beta_1 SV_i + \beta_2 IAV_i + \beta_3 DAV_i + \beta_j X_{ji} + \varepsilon_i \quad (1)$$

For the second task, we regressed the prompt success on the mean similarity to the various keywords for the psychological triggers. Equation 2 represents the logistic regression model for this task.

$$prompt\ success_i = \beta_0 + \beta_1 Reward_i + Trust_2 IAV_i + Urgency_3 DAV_i + Authority_3 DAV_i + \beta_j X_{ji} + \varepsilon_i \quad (2)$$

The standardized coefficients from the logistic regression results in Table 1 suggest that the use of SV words has a positive relationship with successful adversarial prompts. While the use of keywords does not have a significant impact, the interaction of keywords with SV words appears to have a positive effect on prompt success. Conversely, prompts with relatively higher DAV words have a negative relationship with adversarial prompt success. The positive impact of SV words did not change even after their interaction with keywords. However, for IAV, when interacting with keywords, it resulted in a negative outcome.

Regarding the psychological triggers, the logistic regression results revealed that prompts that had a high similarity with reward and authority keywords were less likely to be successful. Meanwhile, trust and urgency were shown to have a positive impact on adversarial prompt success. All the interaction terms between the psychological triggers and the keywords were not significant. Hence, we did not present them in the table. Also, including the interaction terms in the psychological triggers model did not change the significance or direction of the main effect.

Our initial insights from this task is that while existing guidelines for effective prompt engineering suggest that clear and concise prompts lead to desired outputs from GAI systems, SV prompts (see Meskó, 2023; Liu et al., 2022;), which are deemed to be less clear and concise compared to IAV and DAV, appeared to have a positive impact on adversarial prompt success (see Stonjic et al., 2021). Additionally, while reward and authority are perceived to improve phishing success, their use in adversarial prompts were not found to be effective when compared to prompts with relatively higher trust and urgency cues.

Table 1 : Logistic Regression Results

DV: Prompt Success	(1)	(2)		DV: Prompt Success	(3)
SV	.38***	.28***		Reward	-.30***
IAV	.17***	.24***		Trust	.46***
DAV	-.10**	-.13***		Urgency	.34***
N-Key Words	.03	.01		Authority	-.20***
SV x N-Key		.19***		N-Key Words	.03
IAV x N-Key		-.19***			
DAV x N-Key		.08			
Level	-.53***	-.53***		Level	-.48***
Word Count	-.06	-.07^{T}		Word Count	-.02
Intercept	.00	.00		Intercept	-.01
Pseudo R^2	.08	.09		Pseudo R^2	.08

As this is a research-in-progress, we are yet to examine the underlying explanation behind these results. Using the data-driven computationally-intensive theory-building (CITB; Berente et al., 2019) as a guide, we intend to categorize the prompts into explainable groups and conduct additional experiments where necessary to contribute to existing knowledge in this domain.

Conclusion

In this study, we hope to contribute to existing knowledge on GAI security and safety from a content analysis perspective. Using constructs from the LCM and psychological triggers, we find that prompts high in SV words have a positive relationship with success, while adversarial prompts high in DAV words have a negative relationship with prompt success. Also, prompts with relatively higher reward and authority cues are less effective, while those higher in trust and urgency cues have a positive relationship with success. We hope to create explainable categories of adversarial prompts to improve our preliminary findings. Additional experiments may be warranted to support our findings. Our study may be limited in a few ways. First, the dataset we use for this study is quite unique because the GAI platform was created for the purpose of adversarial prompt engineering. As such, there may be concerns regarding generalizability. However, we believe the findings can support the design and development beyond this scope, as there is evidence of such behavior on other platforms. Second, our sample size is relatively small compared to the overall dataset.

REFERENCES

Abdelnabi, S., Greshake, K., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023, November). Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security* (pp. 79-90).

Brinton Anderson, B., Vance, A., Kirwan, C. B., Eargle, D., & Jenkins, J. L. (2016). How users perceive and respond to security messages: a NeuroIS research agenda and empirical study. *European Journal of Information Systems*, *25*(4), 364-390.

Berente, N., Seidel, S., & Safadi, H. (2019). Research commentary—data-driven computationally intensive theory development. *Information systems research*, *30*(1), 50-64.

Falco, G., Shneiderman, B., Badger, J., Carrier, R., Dahbura, A., Danks, D., ... & Yeong, Z. K. (2021). Governing AI safety through independent audits. *Nature Machine Intelligence*, *3*(7), 566-571.

Goel, S., Williams, K., & Dincelli, E. (2017). Got phished? Internet security and human vulnerability. *Journal of the Association for Information Systems*, *18*(1), 2.

Gu, J., Han, Z., Chen, S., Beirami, A., He, B., Zhang, G., ... & Torr, P. (2023). A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.1298*

Hacker, P., Engel, A., & Mauer, M. (2023, June). Regulating ChatGPT and other large generative AI models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1112-1123).

Huang, L., Tan, C. H., Ke, W., & Wei, K. K. (2018). Helpfulness of online review content: The moderating effects of temporal and social cues. *Journal of the Association for Information Systems*, *19*(6), 3.

Johnson-Grey, K. M., Boghrati, R., Wakslak, C. J., & Dehghani, M. (2020). Measuring abstract mind-sets through syntax: Automating the linguistic category model. *Social Psychological and Personality Science*, *11*(2), 217-225.

Korzynski, P., Mazurek, G., Krzypkowska, P., & Kurasinski, A. (2023). Artificial intelligence prompt engineering as a new digital competence: Analysis of generative AI technologies such as ChatGPT. *Entrepreneurial Business and Economics Review*, *11*(3), 25-37.

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., ... & Zhou, B. (2023). Trustworthy AI: From principles to practices. *ACM Computing Surveys*, *55*(9), 1-46.

Liu, V., & Chilton, L. B. (2022, April). Design guidelines for prompt engineering text-to-image generative models. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (p. 1-23).

Luo, X. R., Li, H., Hu, Q., & Xu, H. (2020). Why individual employees commit malicious computer abuse: A routine activity theory perspective. *Journal of the Association for Information Systems*, *21*(6), 5.

Mastropaolo, A., Scalabrino, S., Cooper, N., Palacio, D. N., Poshyvanyk, D., Oliveto, R., & Bavota, G. (2021, May). Studying the usage of text-to-text transfer transformer to support code-related tasks. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)* (pp. 336-347). IEEE.

Meskó, B. (2023). Prompt engineering as an important emerging skill for medical professionals: tutorial. *Journal of Medical Internet Research*, *25*, e50638.

Mouton, F., Leenen, L., & Venter, H. S. (2016). Social engineering attack examples, templates and scenarios. *Computers & Security*, *59*, 186-209.

Pedro, R., Castro, D., Carreira, P., & Santos, N. (2023). From Prompt Injections to SQL Injection Attacks: How Protected is Your LLM-Integrated Web Application?. *arXiv preprint arXiv:2308.01990*.

Pienta, D., Thatcher, J. B., & Johnston, A. (2020). Protecting a whale in a sea of phish. *Journal of information technology*, *35*(3), 214-231.

Schmid, J., Fiedler, K., Semin, G., & Englich, B. (2017). Measuring implicit causality: The linguistic category model.

Seih, Y. T., Beier, S., & Pennebaker, J. W. (2017). Development and examination of the linguistic category model in a computerized text analysis method. *Journal of Language and Social Psychology*, *36*(3), 343-355.

Semin, G. R., & Fiedler, K. (1991). The linguistic category model, its bases, applications and range. *European review of social psychology*, *2*(1), 1-30.

Stojnic, T., Vatsalan, D., & Arachchilage, N. A. (2021). Phishing email strategies: understanding cybercriminals' strategies of crafting phishing emails. *Security and privacy*, *4*(5), e165.

Wang, J., Shan, Z., Gupta, M., & Rao, H. R. (2019). A longitudinal study of unauthorized access attempts on information systems: The role of opportunity contexts. *MIS Quarterly*, *43*(2), 601-622.

Wang, J., Li, Y., & Rao, H. R. (2016). Overconfidence in phishing email detection. *Journal of the Association for Information Systems*, *17*(11), 1.

Warkentin, M., & Willison, R. (2009). Behavioral and policy issues in information systems security: the insider threat. *European Journal of Information Systems*, *18*(2), 101-105.

Yang, Y., Huang, P., Cao, J., Li, J., Lin, Y., & Ma, F. (2024). A prompt-based approach to adversarial example generation and robustness enhancement. *Frontiers of Computer Science*, *18*(4), 184318.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., ... & Shi, S. (2023). Siren's song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.