Personalizing Gen-AI Prompts for Good and Bad

DEREK HANSEN, Brigham Young University, United States
JERSON FRANCIA, Brigham Young University, United States
MATTEW TAYLOR, Brigham Young University, United States
SHYDRA MURRAY, Brigham Young University, United States
BEN SCHOOLEY, Brigham Young University, United States

The widespread adoption of Generative AI (Gen-AI) will no doubt transform the social media landscape and broader information ecosystem. This workshop paper discusses ongoing work related to the creation and identification of personalized spear phishing messages and disinformation, as well as the creation of prompt engineering competitions.

1 RESEARCH BACKGROUND

Dr. Derek Hansen is a Professor of Electrical & Computer Engineering at Brigham Young University (BYU) where he teaches courses on human-computer interaction, social media analysis, and cybersecurity. Dr. Hansen received his PhD in Information from the University of Michigan and taught for four years at the University of Maryland prior to joining BYU. He has received over $4 million in research grants from the NSF, IES, Idaho National Labs, Sandia National Labs, and non-profits. His book "Analyzing Social Media Networks with NodeXL" has been cited over 2,200 times [7]. His other lines of research have focused on social media's impacts on government, law, and policy [2], user-generated content and ratings in domains such as citizen science [11], news consumption [8], political bias [4], crowdsourcing of historical records [6], and community engagement [5]. He has also worked extensively on educational games and simulations that rely on user-generated content (e.g., [3]). He is working closely with Dr. Ben Schooley, doctoral student Jerson Francia, and undergraduate students Shydra Murray and Matthew Taylor on several Gen-AI projects as outlined below.

2 PERSONALIZED MANIPULATIVE AI CONTENT

In an age fraught with disinformation and foreign information manipulation and interference (FIMI) campaigns [9, 10], there is bound to be increasing use of Gen-AI to create and bolster false narratives. While fake accounts and user generated content are nothing new, Gen-AI has the potential to increase the efficiency and scale with which such content is created. Perhaps even more concerning, Gen-AI has the potential to create personalized manipulative messages that target an individual based on their own beliefs, biases, friendship networks, etc. The fact that spear phishing messages, which only account for 0.1% of all email messages, are responsible for 66% of all breaches, demonstrates the impact of personalization on deception [14]. However, we do not yet know how effective Gen-AI is in creating tailored spear phishing messages or disinformation.

Our research team at BYU has been studying the capabilities of Gen-AI in creating spear phishing messages, as compared to trained humans. We have employed a novel approach, wherein we recruited “targets” who share personal information about themselves (e.g., job title and location; personal hobby; something they have posted about online) and agree to return later for an interview. We then have humans who have been trained on how to create effective spear phishing messages create them based on the information provided by the targets. We also have Gen-AI (i.e, ChatGPT-4) generate spear phishing messages using the same prompts we gave the humans. Finally, we invite the targets back and show them 12 spear phishing messages created to target them based on the information they provided. They are asked to “sort” the messages from most compelling to least compelling and discuss why they placed them where they did. After the sorting and discussion, they are told that one or more of the messages were created by AI. They are asked to place a token (with the letters AI) on any they believe were created by AI and explain why they chose those ones and not others. We are currently analyzing the data for 28 targets. Preliminary results suggest that Gen-AI is slightly more effective at creating personalized spear phishing messages than trained humans. And this is based on an off-the-shelf Large Language Model (LLM) and a very basic prompt (“Create a spear phishing…”). Furthermore, most targets have no idea which messages were created by AI and don’t have an accurate mental model of how to even approach that question. Through this project and future projects on the creation of disinformation messages, images, and videos in the form of hypothetical social media posts, we hope to demonstrate the capabilities (and potentially limitations) of Gen-AI in personalizing disinformation. This is a necessary step in order to identify and flag such content so the risks associated with fraud and disinformation can be mitigated.

3 PROMPT ENGINEERING COMPETITIONS

Fortunately, not all uses of Gen-AI are nefarious. Many content creators with good intensions will be able to leverage the capabilities of Gen-AI to create high-quality, multimedia content more efficiently and for cheaper. However, much of this depends on their ability to write effective prompts, which constitutes an entirely new realm of knowledge. There is a significant need for techniques that help learners develop prompt engineering skills.

One strategy we are piloting at BYU is to create Gen-AI prompt engineering competitions. Competitions have been highly successful in recruiting and training in fields such as cybersecurity [1] and data science (e.g., Kaggle competitions) [13]. While there are AI art and short story competitions (e.g., Agora Worldwide Awards; AI Fables) wherein users submit the output of Gen-AI prompts, these are different than what we have in mind. Prompt engineering in practical usage seeks to identify prompts and prompt structures that can be reused to consistently generate useful content. For example, we have been comparing prompts that identify if a social media post is factual or includes misinformation. We have a single prompt structure, that takes in a “variable” which consists of a particular social media post. For example: “Tell me if this post is factual or if it includes disinformation: [message content].” We then assess the quality of different prompts by seeing how well they perform on a set of thousands of different messages included in our ground truth dataset that includes short messages labeled by humans as truthful or misinformation [12]. This more generic type of prompt is much more practical than a one-off prompt used to create a single piece of art.

There are several possible domains for these more general Gen-AI prompt engineering competitions. Food advertisers could ask for a prompt that creates consistent, high-quality images of a {food} (where {food} is a variable that gets replaced with different types of food). Patient advocates could ask for a prompt that would accurately summarize and distill key findings from a {medical article} for non-trained members of online medical support groups. And eventually filmmakers may upload scenes from a film and ask for a prompt that creates a trailer video from the scenes. The possibilities are extensive and touch nearly every area of user generated content. Coordinating cross-institution competitions in this space would be a great outcome of the workshop.

To support such prompt engineering competitions, we envision a platform composed of 3 core components:

1. A Gen-AI prompt engineering use-case description and field for participants to test their prompts, view the output and scores (if available), and submit prompts when they are ready.
2. A prompt scoring engine that evaluates each submitted prompt. This would rely on a sponsor-provided ground truth dataset, if one exists. Or it could include a scoring engine based on expert voting and/or an AI-based scoring algorithm using criteria defined by the sponsor.
3. A leaderboard that shows the top submitters and evaluation scores. A more collaborative “remix competition” could show the prompts and allow users to remix them into new prompts, while keeping track of the provenance of each portion of the prompt. AI could even be used to remix the top prompts.

4 CONCLUDING THOUGHTS

We have spotlighted two projects related to Gen-AI and user generated content. Dr. Hansen is hoping to attend in person to discuss these projects, as well as others such as: using social network analysis of social media datasets to identify Gen-AI created content; legal and policy issues related to Gen-AI content; methodologies for assessing the quality and impact of Gen-AI content; and the use of Gen-AI content in educational simulations.

REFERENCES

[1] Tyler Balon and Ibrahim (Abe) Baggili. 2023. Cybercompetitions: A survey of competitions, tools, and systems to support cybersecurity education. *Educ Inf Technol* 28, 9 (September 2023), 11759–11791. https://doi.org/10.1007/s10639-022-11451-4

[2] John Carlo Bertot, Paul T. Jaeger, and Derek Hansen. 2012. The impact of polices on government social media usage: Issues, challenges, and recommendations. *Government information quarterly* 29, 1 (2012), 30–40. https://doi.org/10.1016/j.giq.2011.04.004

[3] Elizabeth Bonsignore, Derek Hansen, Kari Kraus, June Ahn, Amanda Visconti, Ann Fraistat, and Allison Druin. 2012. Alternate Reality Games: platforms for collaborative learning. (2012).

[4] Jennifer Golbeck and Derek Hansen. 2011. Computing political preference among twitter followers. In *Proceedings of the 2011 annual conference on Human factors in computing systems* (*CHI ’11*), 2011. ACM, New York, NY, USA, 1105–1108. https://doi.org/10.1145/1978942.1979106

[5] Derek L. Hansen, Jes A. Koepfler, Paul T. Jaeger, John C. Bertot, and Tracy Viselli. 2014. Civic action brokering platforms: facilitating local engagement with ACTion Alexandria. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (*CSCW ’14*), February 15, 2014. Association for Computing Machinery, New York, NY, USA, 1308–1322. https://doi.org/10.1145/2531602.2531714

[6] Derek L Hansen, Patrick J Schone, Douglas Corey, Matthew Reid, and Jake Gehring. 2013. Quality control mechanisms for crowdsourcing: peer review, arbitration, & expertise at familysearch indexing. 2013. 649–660.

[7] Derek Hansen, Ben Shneiderman, and Marc A. Smith. 2010. *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*. Morgan Kaufmann.

[8] Silvia Knobloch-Westerwick, Nikhil Sharma, Derek L Hansen, and Scott Alter. 2005. Impact of popularity indications on readers’ selective exposure to online news. *Journal of broadcasting & electronic media* 49, 3 (2005), 296–313.

[9] Wojciech Mazurczyk, Dongwon Lee, and Andreas Vlachos. 2023. Disinformation 2.0 in the Age of AI: A Cybersecurity Perspective. https://doi.org/10.48550/arXiv.2306.05569

[10] Nicolas Hénin. 2023. *FIMI: towards a European redefinition of foreign interference*. EU DisInfoLab. Retrieved September 13, 2023 from https://www.disinfo.eu/publications/fimi-towards-a-european-redefinition-of-foreign-interference/

[11] Dana Rotman, Jenny Preece, Jen Hammock, Kezee Procita, Derek Hansen, Cynthia Parr, Darcy Lewis, and David Jacobs. 2012. Dynamic changes in motivation in collaborative citizen-science projects. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (*CSCW '12*), 2012. ACM, New York, NY, USA, 217–226. https://doi.org/10.1145/2145204.2145238

[12] Pawan Kumar Verma, Prateek Agrawal, Ivone Amorim, and Radu Prodan. 2021. WELFake: Word Embedding Over Linguistic Features for Fake News Detection. *IEEE Transactions on Computational Social Systems* 8, 4 (August 2021), 881–893. https://doi.org/10.1109/TCSS.2021.3068519

[13] Jenny Lena Zimmermann. 2021. Data Competitions: Crowdsourcing with Data Science Platforms. In *The Machine Age of Customer Insight*, Martin Einhorn, Michael Löffler, Emanuel de Bellis, Andreas Herrmann and Pia Burghartz (eds.). Emerald Publishing Limited, 183–197. https://doi.org/10.1108/978-1-83909-694-520211017

[14] 2023 Spear-Phising Trends. *Barracuda Networks*. Retrieved September 12, 2023 from https://www.barracuda.com/reports/spear-phishing-trends-2023