Evaluating BEC Emails Generated by LLMs

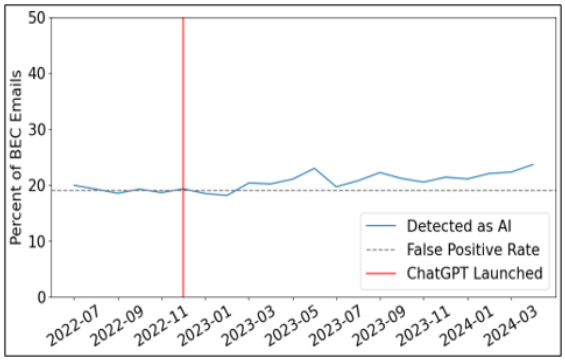
AnMei Dasbach-Prisk[2], Claire Wang[1], Wei Hao[1], Asaf Cidon[1]
[1]Columbia University NY, New York [2]Cabrillo College, CA, Aptos

Abstract: Business Email Compromise (BEC) attacks pose a significant[1] and evolving threat. This research contributes to a larger project aimed at evaluating email scams and building a comprehensive training dataset to improve scam detection. Specifically, we explore the potential of Large Language Models (LLMs) to enhance the persuasiveness of spear phishing email templates[2]. We assess the effectiveness of different prompting techniques using G-Eval[3], a framework that evaluates LLM generated output provided custom criteria. Our findings overall demonstrate LLMs' ability to improve template quality.

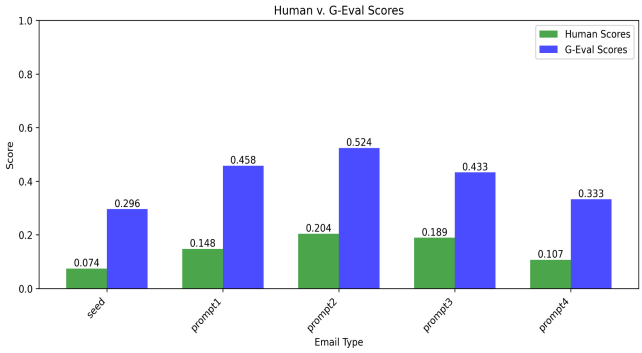
Method: Generated 480 phish from 10 starter templates, 12 profiles (made-up employees), and four different prompting methods, using gpt-4o-mini

Evaluation: Using G-Eval and the following criteria: Cialdini's principles, urgency, persuasion, relevance scored emails

Preliminary Analysis:



Results Graph:



References:

1. FBI. Business Email Compromise: The $50 Billion Scam, 2023. https://www.ic3.gov/Media/Y2023/PSA230609
2. UC Berkeley. Phishing Examples Archive, 2024. https://security.berkeley.edu/education-awareness/phishing/phishing-examples-archive
3. Liu, Yang, et al. "G-eval: Nlg evaluation using gpt-4 with better human alignment." arXiv preprint arXiv:2303.16634 (2023).

My contributions: Throughout this summer I learned a lot about prompt engineering and helped craft a prompt that was used in a real campaign to test if generative AI personalized emails are as effective as human personalized emails.