Prompt Engineering for Imposter Scam Detection by Resourced-Constrained Organizations

Online scams are a large and growing safety risk. The Federal Trade Commission reports that consumers lost more than $12.5 billion to scams, the most common of which were imposter scams (e.g., a scammer posing as a representative of the target's bank) [FTC25]. A primary data source for understanding scammer techniques and measuring their financial impact is text narratives, typically in the form of user complaints to federal or local agencies. LLMs have shown promise in automating the detection of scams in text narratives when fine-tuned to create specialized models (e.g., [NP24]). However, detection strategies for off-the-shelf, pre-trained LLMs, such as GPT, Gemini, Llama, have only been developed for the binary classification of scams versus other financial harm [CHS25, SYZLNF25], yet the identification of individual scam *types* is essential for recognizing emerging threats and prioritizing scam defense measures. Agencies may not be staffed to develop and maintain customized LLM solutions and so prompt engineering guidance for pre-trained LLMs is essential given the possibility of errors ("hallucinations") and prompt sensitivity (e.g., "jailbreaking").

This paper presents ongoing work developing prompt engineering strategies for detecting imposter scams in consumer complaints. We utilized Consumer Financial Protection Bureau (CFPB) consumer complaint data from 2024-25 and identified 167 positive examples (i.e., imposter scams) that we randomly split into approximately equally sized training, validation and test sets. We conducted 21 experiments across four distinct prompt engineering workflows using Gemini-2.0-flash, with 10 retries per experiment and executed each experiment at least twice; all reported performance numbers are averages across executions of an experiment.

Through systematic experimentation with Gemini-2.0-flash, we evaluated three distinct prompt engineering workflows: (1) example-only approaches using training data examples (best precision 0.84/recall 1.0 with 5 positive/5 negative examples with author-written reasons), (2) definition-focused approaches incorporating authoritative definitions from CFPB, FTC, Federal Deposit Insurance Corporation (FDIC) and other federal agencies (best precision 0.93/recall 0.96), and (3) hybrid definition + example combinations (best precision 0.88/recall 0.96 with 2 positive/2 negative examples with author-written reasons). Our experiments demonstrate that precision and recall exceeding .9 can be achieved with prompts *based only on definitions from authoritative sources* (e.g., the FTC and FDIC); out-performing the precision of prompts relying on examples. This is an encouraging finding for resource-constrained agencies that may have limited or no training data with which to develop prompts.

[CHS25] Chadalavada, Isha, Tianhui Huang, and Jessica Staddon. "Building a Scam Narrative Corpus with Ensemble Learning." Extended abstract at the 9th Workshop on Technology and Consumer Protection (ConPro `25).

[FTC25] Federal Trade Commission. "New FTC Data Show a Big Jump in Reported Losses to Fraud to $12.5 Billion in 2024". March 10, 2025.

[NP24] Nicholas, Poh Yi Jie, and Pai Chet Ng. "ScamDetector: Leveraging Fine-Tuned Language Models for Improved Fraudulent Call Detection." In *TENCON 2024-2024 IEEE Region 10 Conference (TENCON)*, pp. 422-425. IEEE, 2024.

[SYZLNF25] Shen, Zitong, Sineng Yan, Youqian Zhang, Xiapu Luo, Grace Ngai, and Eugene Yujun Fu. "" It Warned Me Just at the Right Moment": Exploring LLM-based Real-time Detection of Phone Scams." In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1-7. 2025.