Building a Scam Narrative Corpus with Ensemble Learning

Isha Chadalavada
Northeastern University
chadalavada.i@northeastern.edu

Tianhui Huang
Northeastern University
t.tianhui.huang@gmail.com

Jessica Staddon
Northeastern University
j.staddon@northeastern.edu

***Abstract*—Users increasingly query LLM-enabled web chatbots for help with scam defense. The Consumer Financial Protection Bureau's complaints database is a rich data source for evaluating LLM performance on user scam queries, but currently the corpus does not distinguish between scam and non-scam fraud. We are developing an LLM ensemble approach to distinguishing scam and fraud CFPB complaints and describe our methodology, current performance and observations of strengths and weaknesses of LLMs in the scam defense context.**

1. Introduction

A variety of large language model-based conversational assistants (LLMs) are available via the web and are regularly relied upon by users for distilling complex information and potentially reducing the work of information retrieval via search engines. An LLM use case of growing popularity is protection from scams; this use case is endorsed by security vendors [16], [9], [11], is a natural evolution of the use of search engines for security-related tasks [8] and is particularly important for consumer protection given that financial institutions provided limited, if any, support for recovering money lost due to scams. The user-friendly nature and broad availability of LLMs makes them a particularly compelling defense tool for socially isolated users, a group that is known to be vulnerable to scams (e.g., [3], [17]).

While the scam defense opportunities of LLMs are significant, the ability of broadly accessible pretrained LLMs (e.g., ChatGPT, Google Gemini) to recognize scam markers has not been comprehensively assessed. In addition, LLMs are known to hallucinate information (e.g., [7]) and come with privacy risks due to training data leaks [1] and a potential to encourage over-sharing of personal information by users [18]. To understand how to manage these risks in the scam context, a corpus of user scam narratives with which to evaluate LLM performance, is needed.

The Consumer Financial Protection Bureau (CFPB) complaints database [2], is a promising source of scam narratives, however currently scam complaints are grouped with fraud complaints (through consumer selection of the "fraud or scam" issue or sub-issue when making a complaint). We are developing a prompting technique [10] for distinguishing CFPB complaints regarding a scam (i.e., a user has been *tricked into taking* a financially self-harming action) from those concerning fraud (i.e., a financially harmful action is *taken without the user's consent*). As part of the iterative prompt development, we have manually labeled a set of 300 CFPB complaints that consumers reported as "fraud or scam". In this work-in-progress paper, we share our best performing ensemble prompt and observations of the strengths and weaknesses of LLMs in the scam context based on prompt performance on a manually labelled set of complaints.

2. Methodology

While we experimented with 3 LLMs, GPT-4[1], Gemini[2] and Llama[3], our best-performing prompt uses only Gemini and GPT-4 so we focus on those 2 LLMs in this short paper.

Training data. To build a collection of scam complaints for training purposes, the authors independently labeled 150 CFPB complaints with the "fraud or scam" issue or subissue, relying on the definition of scam as a financial harm in which a user is tricked into taking a self-harming action [6] and following standard qualitative coding practices, e.g., [4]. The authors achieved high agreement on the set of 150, and subsequently independently labelled 50 disjoint sets each, resulting in a set of 300 CFPB complaints, 64% of which are labeled scams. We denote this set of 300 labeled narratives by L. The narratives in L range in character length from 38 to 10,975 with a mean and median of 1,416.85 and 1,157.5, respectively. 93% of the narratives in L are somewhat redacted, and on average 4.6% of the narrative characters are redacted [14].

Prompt Design and Iteration. All prompts with which we have experimented rely heavily on the definition of a scam [6] and require the LLMs to explain their labels since a large body of research shows performance can improve when LLM responses include descriptions of their "reasoning" (e.g. [15]). In addition, we have experimented with including example complaints and labels since this strategy is associated with improved performance in other contexts (e.g., [13]). Finally, we experimented with breaking prompts into multiple, simpler prompts (e.g., asking the LLM to just evaluate a single aspect of the definition and combining the LLM responses to decide on scam labels), and found this improved the performance of GPT-4.

To date, our best performance is with an ensemble model using different prompts for Gemini and GPT-4 and requiring that both LLMs predict scam for each prompt. The Gemini prompt (denoted prompt A) defines a scam, instructs Gemini to classify subsequent complaints as "scam" or "not scam" and provides to example complaints and predictions. GPT-4 receives 2 prompts, prompt B asks the LLM to determine

1. https://platform.openai.com/docs/models#gpt-4
2. https://ai.google.dev/gemini-api/docs/models/gemini#gemini-1.5-flash
3. https://ollama.com/library/llama3.1

whether money was stolen from the complainant or they were tricked into authorizing a transaction, and prompt C asks the LLM to determine the reputation of the entity who received the complainant's money or personal information and label the narrative a scam if the entity's reputation is not positive. Prompts A, B and C are in the Appendix.

The LLM responses to these prompts are combined via conjunction. That is, scam is predicted for a given complaint narrative if and only if: Gemini predicts scam in response to that narrative and prompt A *and* GPT-4 predicts #2 for that narrative and prompt B *and* GPT-4 predicts potential scam for that narrative and prompt C. We denote this ensemble model as F. On the set L, F achieves precision of .95 and recall of .84.

To begin building the corpus, we prompted the ensemble model with the 2569 CFPB complaint narratives labeled with a "fraud or scam" issue or sub-issue from January 1, 2024 through November 13, 2024[4], F identified 1,333 scam narratives (.52 of the data set). We manually evaluated a randomly selected 10% sample ($n = 133$), and measured a precision of .97.

3. LLM Performance Observations

In this section we describe some of the patterns in the LLM errors in responses to the labeled narratives, L.

Reliance on Secondary Information. When complaint narratives are unclear and common scam markers are absent, we observed that both GPT-4 and Gemini can rely on secondary characteristics such as customer service or claim denials, to make decisions, rather than declining to decide. For example, CFPB Complaint #4473515 devotes less than 14% of its 950 characters to describe the financial harm that LLMs are determining to be a scam or non-scam fraud, and 6% of the financial harm content is redacted: "Someone took money from my Citibank account of My corporation and used it to pay XXXX XXXX i have no idea who these individuals are..." The remaining 86% of the complaint describes the complainant's experience trying to resolve the harm with Citibank. While the narrative describes fraud (an unauthorized withdrawal) both GPT-4 and Gemini predict scam in response to prompt C and explain their decisions with secondary information. GPT says: "The poor customer service experience combined with the unauthorized monetary transaction strongly suggests a potential scam." and Gemini responds: "...the lack of communication from citibank, coupled with the unclear nature of the transaction and the involvement of a potentially fictitious individual or entity, points towards a potential scam."

Impact of Reputation. Scams are committed by entities of unknown or poor reputation and begin by building trust, thus persuading targets to act in the scammer's interest (e.g., [12]). Hence, a poor or unknown reputation is an indicator of a potential scammer. We did not find evidence that LLMs consider reputation when identifying scams. Rather, there is some evidence of over-reliance on official business names as indicators of positive reputation. For example, the credit repair company, Lexington Law, has been fined by the CFPB for illegal fees and deceptive advertising and there is ample web evidence of its poor reputation [5]. Lexington Law is the subject of 4 complaints in the 300 labeled narratives, but only half of those are predicted scams with our final prompt, and some LLM explanations attribute a positive reputation to Lexington Law. In particular, CFPB complaint #4189755 begins, "Over the past year I have paid Lexington law firm and infinite finance to repair my credit. The negative items on my credit are old, settled, went through a chapter XXXX in 2012 or are simply not mine. I was told by both companies that it wouldn't be hard at all to get them removed..." and both LLMs predicted non-scam; Gemini responded (via prompt A): "The user's complaint describes a common experience with credit repair companies..." and GPT-4 responded (via prompt C): "The complainant mentions dealing with Lexington Law Firm and Infinite Finance, which are known and reputable companies...".

Particularly with GPT-4 we observed sensitivity to the language used to characterize the company in the narrative. For example, in CFPB narrative #4308351 the complainant expresses some skepticism about Lexington Law ("My mother and I retained the Lexington Law Group, advertised and described as a credit repair agency.") and GPT-4 with prompt C identified it as a potential scam and responded: "Although the company is advertised as a credit repair agency, the nature of these issues might suggest that it could be a potential scam.".

Overall, we found better performance for narratives *without* company names (recall of .852, precision of .935) than for narratives that include a company name (recall of .733, precision of .846).

Performance and Narrative Length While precision initially improves with narrative length, for both GPT-4 and Gemini we found that model performance declines once narratives exceed approximately $3,000$ characters (Figure 1, Appendix). Longer narratives often included information that is not directly relevant to determining whether the harm experienced was a scam or fraud, such as descriptions of the challenges the complainant encountered when trying to resolve the harm or cut and pastes of email correspondence. As mentioned earlier, this secondary information often surfaced in model explanations (incorrectly) justifying scam/fraud predictions.

Performance on Redacted Narratives. Redaction removes information, and as is to be expected, negatively impacts LLM scam prediction performance. However, longer narratives can tolerate a higher percentage of redaction (Figure 2, Appendix). We hypothesize that this is partly driven by the fact that LLMs tend to predict scam when narratives assert a scam has occurred and the fraction of a complaint that is needed for such an assertion decreases as complaints grow. For example, CFPB complaint #4294462 has more than 10% redacted characters, but also includes an explicit scam declaration ("I soon realized that I was scammed and asked BB & T to return the money....") and is predicted a scam by F.

4. At the time of writing, there are a little more than 2,700 narratives in the same time window, perhaps due to processing delays.

References

[1] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and Alina Oprea. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[2] The Consumer Financial Protection Bureau (CFPB). Consumer complaints database. *https://www.consumerfinance.gov/data-research/consumer-complaints/*.

[3] Gilbert Gimm and Scott Beach. Financial exploitation vulnerability and social isolation in older adults: Results from a longitudinal survey. *Innovation in Aging*, 4(Suppl 1):29, 2020.

[4] Daniel J Hruschka, Deborah Schwartz, Daphne Cobb St. John, Erin Picone-Decaro, Richard A Jenkins, and James W Carey. Reliability in coding open-ended data: Lessons learned from hiv behavioral research. *Field methods*, 16(3):307–331, 2004.

[5] Truman Lewis. $1.8 billion going to victims of credit repair scam. *ConsumerAffairs.com*, 2024.

[6] David Modic and Stephen EG Lea. Scam compliance and the psychology of persuasion. *Available at SSRN 2364464*, 2013.

[7] OpenAI. GPT-4 Model Card. https://cdn.openai.com/papers/gpt-4-system-card.pdf, 2024. Accessed: 2024-05-12.

[8] Elissa M. Redmiles, Amelia R. Malone, and Michelle L. Mazurek. I think they're trying to tell me something: Advice sources and selection for digital security. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 272–288, 2016.

[9] Seth Ruden. How to level up your fraud defense with chatgpt's own ai suggestions. *BioCatch Blog Channel*, January 27, 2023.

[10] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024.

[11] Sarah. Using chatgpt for real-time fraud detection in finance. *Dataleon. Ai Blog, December 11*, 2023.

[12] Azianura Hani Shaari, Mohammad Rahim Kamaluddin, Wan Fariza Paizi Fauzi, and Masnizah Mohd. Online-dating romance scam in malaysia: An analysis of online conversations between scammers and victims. *GEMA Online Journal of Language Studies*, 19(1), 2019.

[13] Zhen Tan, Alimohammad Beigi, Song Wang, Ruocheng Guo, Amrita Bhattacharjee, Bohan Jiang, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*, 2024.

[14] The Consumer Financial Protection Bureau (CFPB). OFFICE OF CONSUMER RESPONSE. Narrative scrubbing standard version 6.6. *https://files.consumerfinance.gov/f/documents/cfpb_narrative-scrubbing-standard_2023-05.pdf*.

[15] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[16] Lance Whitney. How to use norton's free ai-powered scam detector. *ZDNet, September 14*, 2023.

[17] Tong Xing, Fei Sun, Kaipeng Wang, Jiawei Zhao, Mengxuan Wu, and Jie Wu. Vulnerability to fraud among chinese older adults: Do personality traits and loneliness matter? *Journal of Elder Abuse & Neglect*, 32(1):46–59, 2020.

[18] Zhiping Zhang, Michelle Jia, Hao-Ping Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. "it's a fair game", or is it? examining how users navigate disclosure risks and benefits when using llm-based conversational agents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–26, 2024.

Appendix A. Ensemble Prompt and Performance

The prompts are:

1) Prompt A [used with Gemini][5]: *"A scam is an attempt to defraud a person or group after first gaining their trust. Confidence tricks exploit victims using a combination of the victim's credulity, naïveté, compassion, vanity, confidence, irresponsibility, and greed. Researchers have defined confidence tricks as "a distinctive species of fraudulent conduct intending to further voluntary exchanges that are not mutually beneficial", as they "benefit con operators ('con men') at the expense of their victims (the 'marks')". Classify the following customer complaint as either "scam" or "not scam" (NO OTHER OPTIONS), then explain your decision.*
    - *[Example 1] if the user inputs: "I was asked to transfer money to an account after receiving a suspicious email. It claimed to be from my bank, but when I called my bank, they said they never sent such an email." The following answers should be generated: "Answer: scam Explanation: The customer describes a classic phishing attempt where a fraudulent email tries to convince them to send money, which was confirmed to be false by their bank."*
    - *[Example 2] if the user inputs: "Eight unauthorized transactions totaling $5800 were made from my account without my knowledge. I filed a claim with my bank, but they denied it twice, even after I provided proof of suspicious texts." The following answers should be generated: "Answer: not scam Explanation: The user describes unauthorized withdrawals from their bank account and attempts to recover the funds through their bank, which denied the claim. The user is likely a victim of financial fraud, not a scam perpetrator."*

   Here is the input:[insert complaint]"
2) Prompt B [used with GPT-4]: *"Which of the following describes this complaint: #1 Money was stolen from the complainant, #2 The complainant was tricked into giving someone money, #3 neither #1 or #2. Put your answer in format: Answer: [#1/#2/#3] Explanation: [your explanation]*
   Here is the input: [insert complaint]"
3) Prompt C [used with GPT-4]: *"Which of the following describes this complaint: Potential Scam: The complainant gave money or personal information to a representative of a company of unknown reputation OR to someone impersonating a representative*

5. Includes an excerpt from https://en.wikipedia.org/wiki/Scam

of a reputable company or organization; Not Scam: The complainant gave money or personal information to a representative of a reputable company or organization. Put your answer in format: Answer: [Potential Scam/Not Scam] Explanation: [your explanation]
Here is the input: [insert complaint]"

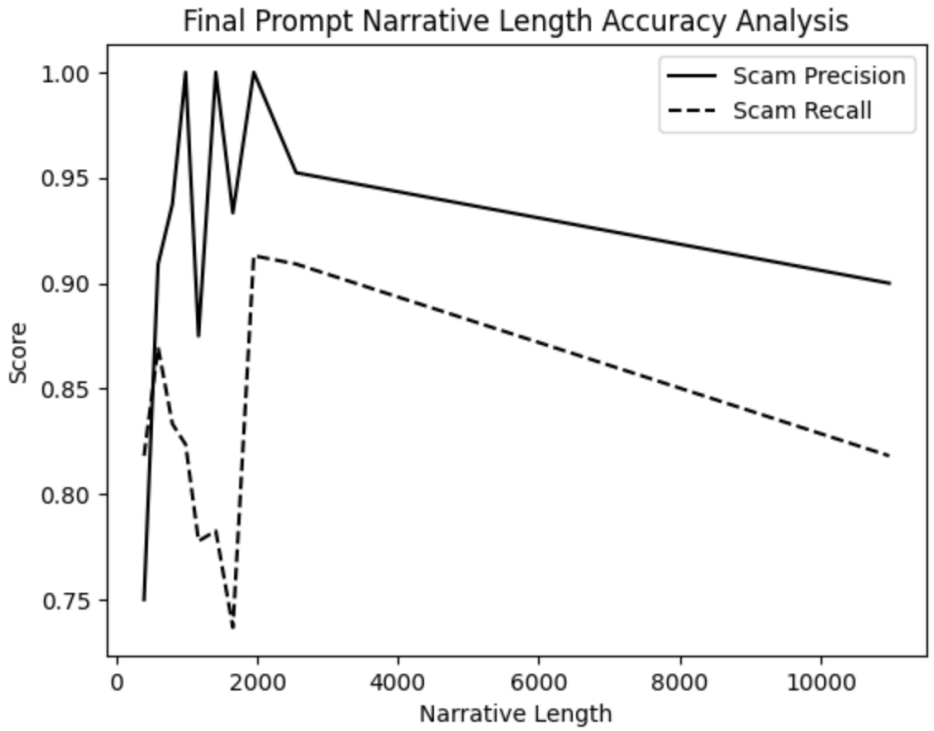

Figure 1: Precision and recall of ensemble model, F, as a function of complaint narrative length in characters.

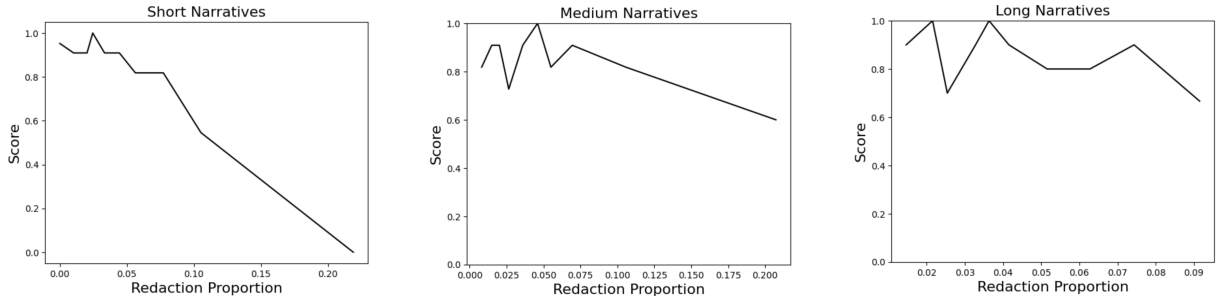

Figure 2: Each figure shows the accuracy of model F as a function of the fraction of redaction for narratives grouped by length. The narratives represented in the left most figure are the 99 messages in L of at most 875 characters ("short"); the middle figure represents the 100 narratives in L of length 875 - 1,602 characters ("medium"); the right most figure represents the 101 narratives with at lest 1,602 characters ("long"). Note that performance is more robust to redaction with longer narratives.