Evaluating LLMs for Scam Detection: Safeguarding Older Adults Against Online Threats

Harshil Agarwal, Jude Lwin, Mark Seeliger, Shashank Thirumale, Shiqi Wu
University of Maryland, College Park

Sunday, December 8th, 2024

Abstract

With the large increase in online email scams during the 21st century, the demand for an accessible and accurate scam detection method has risen significantly. Many existing methods for combating scams are complex and require prior technical knowledge, making them less usable for older adults, who are particularly vulnerable to online email scams. The rapidly evolving sector of large language models (LLMs) is opening up a new opportunity to prevent online-scams using generative AI, providing older adults with a more intuitive method of identifying and avoiding email scams, without the need for detailed technical knowledge, making scam detection more accessible. This paper will compare the accuracy and effectiveness of two popular LLMs, OpenAI's GPT4.0 and Google's Gemini, in identifying and categorizing these scams, comparing their true and false positive rates in different scenarios.

1 Introduction

We would love to see a world where all people, especially older adults, do not fall prey to online scams.

Email services such as Gmail use their own methods and machine learning models to identify and tag emails as spam/scams. Despite the current methods of scam detection, older adults disproportionately fall prey to online scams. We would like to attack this problem by advancing the information in this space when related to LLM's methodology and accuracy when detecting scams. By advancing this space, we can get closer to a world where scam detection is incredibly accurate and we don't have to see people like our grandparents fall prey to irreversible damage.

There are many challenges to consider when tackling this problem. The current state of this problem starts with the number of emails that exist each day. The magnitude is much too high to be able to parse through each email manually (it would also be much too slow). Although it may be obvious to a trained human being, current models can not be 100% sure of a scam or real email. With the necessity of models to run scam detection, we run into the issue of accuracy. Due to the inaccuracy of current models, many scams are not tagged as scams, and many non-fraudulent emails are marked as scams. In the context of our paper, the false positives where scams are not tagged properly are the most important. These scams that fall through the cracks are the ones that people fall prey to.

We plan to evaluate the state of two LLMs to provide insight into their current ability. This may make weaknesses or strengths more apparent and allow for further development. Our insight is that the models are too inaccurate to be able to consider our older adults or the common people safe from scams. This field is one where accuracy is of the most importance because it is the case that one scam is all that it takes to ruin a person's life.

Following this, we will dive into related works that explain our motivation for our methodology and also differentiate our paper from previous works. We will then explain our methodologies and the results of our research.

2 Related Works

The paper *Application of AI-based Models for Online Fraud Detection and Analysis* [5] explores the potential of AI for detecting and analyzing online fraud using text data. The authors conducted a systematic literature review on AI and natural language processing techniques for online fraud detection. They synthesized findings from 223 academic records and identified the most effective AI methods currently in use. Additionally, the authors concluded that the evolving nature of scams limits the efficacy of models trained on outdated data. While this paper focuses on insights from existing literature, our work involves conducting a new experiment to evaluate specific LLMs. In another paper, *Detecting Scams Using Large Language Models* [6], Liming Jiang similarly explored the use of LLMs for scam detection. However, Jiang uses data collection, preprocessing, model selection, training, and integration into target systems in order to build an effective scam dectector using LLMs, whereas we are comparing the accuracy and effectiveness of two different LLMs. Jiang also conducted a preliminary evaluation using GPT-3.5 and GPT-4 on a duplicated email, highlighting their proficiency in identifying common signs of phishing or scam emails, and we will be testing different "scams" with our models as well.

Additionally, the paper *A Preliminary Comparison and Combination of Vision-Language Models through Qualitative Cases* [7] by Jiaqi Wang is relevant due to its comparison of Gemini and GPT models. However, our research differs in significant ways. Specifically, we evaluate GPT-4o, a more accessible model for older adults who may rely on it for scam detection. While Wang's study examines how Gemini and GPT handle multimodal inputs, our work focuses exclusively on email-based scams. By narrowing the scope, our study builds upon prior insights, applying them to a targeted real-world use case.

3 Methodology

We are going to carry out an experiment with real emails in order to measure the efficiency and precision of each model in scam detection. We want to see how capable the models will be in distinguishing fraudulent and legitimate communications, but also examine the false positives-legitimate communications that the model reports as scams-and the false negatives, scams undetected by the models. The exact methodology will be as follows.

3.1 Data Collection

We originally planned to collect emails and text messages samples from our own email and spam folders. However, we decided to deviate from this original decision due to advisor feedback. Our original plan would not be as representative of our target audience, as the data in our spam folders would not be similar to that of emails targeting older adults. We switched to a dataset containing both fraudulent and non-fraudulent emails with the type of fraudulent emails being Nigerian Letters or "419" Fraud [8]. This is one of the best known types of fraudulent emails especially for older adults. The dataset contains 11,958 data points each being labeled 0 for "Non-Fraudulent" and 1 for "Fraudulent" with the fraudulent data dating from 1998 to 2007.

3.2 Model Selection

We will be using two advanced LLMs: ChatGPT-4o and Google Gemini 1.5 Pro. We choose these models in light of the performance of both models regarding state-of-the-art natural language understanding and classification tasks.

3.3 Testing Process

Each model was provided the same prompt with the .csv file containing the email dataset attached. "Read column one, which contains email contents, and evaluate if it is a scam or if it is legitimate. Create another .csv, equal columns to the one given one and label each email as 0 if you evaluate it as legitimate, or 1 if you evaluate it as a scam. Do not use prior context to evaluate, evaluate each email as an independent test, similar to a real world situation. "

This resulting .csv now contained the original labels and the models labels. We then compared the original and new labels to identify True Positives, True Negatives, False Positives, and False Negatives.

These new identifications were used in the creation of confusion matrices and 4 key metrics (explained in results section) for both models. The confusion matrix allows for better visualization of the performance of our models with respect to our predicted labels (the datasets' labels) and true labels (the models output).

3.4 False Positive and False Negative Analysis

We will compare the models' predictions to the ground truth labels to identify instances where the models either classify legitimate communications as scams or fail to detect actual scams. Specifically:

- **False Positives:** Legitimate emails flagged as scams.
- **False Negatives:** Scam emails that were not detected by the models.

These classifications will allow us to evaluate the performance of the models using their precision and recall (this will be explained later).

4 Results

The performance of ChatGPT-4o and Gemini-1.5-Pro in detecting scam emails was evaluated using four key metrics: accuracy, precision, recall, and F1 score [9].

1. Accuracy: the proportion of all classifications that were correct, whether positive or negative.
2. Precision: the proportion of all the model's positive classifications that are actually positive.
3. Recall: the proportion of all actual positives that were classified correctly as positives
4. F1 Score: the harmonic mean (a kind of average) of precision and recall.

The results of these evaluations are summarized in Table 1.

Table 1: Performance Metrics for ChatGPT-4o and Gemini-1.5-Pro

Metric	**ChatGPT-4o**	**Gemini-1.5-Pro**
Accuracy	0.84	0.87
Precision	0.85	0.87
Recall	0.84	0.87
F1 Score	0.83	0.87

4.1 ChatGPT-4o Performance

ChatGPT-4o achieved an accuracy of **0.84**, indicating that it correctly classified 84% of emails as either scam or legitimate. The system demonstrated a precision of **0.85**, meaning that 85% of emails identified as scams were indeed scams. Recall, which measures the ability to identify all scam emails, was recorded at **0.84**. The F1 score, a harmonic mean of precision and recall, was **0.83**, reflecting a balanced performance across these metrics.

4.2 Gemini-1.5-Pro Performance

Gemini-1.5-Pro outperformed ChatGPT-4o across all metrics. It achieved an accuracy of **0.87**, demonstrating higher overall classification performance. Its precision, recall, and F1 score were all **0.87**, indicating consistent and superior performance in identifying scam emails without sacrificing precision or recall.

4.3 Comparative Analysis

While both systems demonstrated strong performance, Gemini-1.5-Pro consistently scored higher than ChatGPT-4o in all four metrics. The differences, though modest, suggest that Gemini-1.5-Pro offers better reliability and consistency in detecting scam emails. Notably, the F1 score of **0.87** for Gemini-1.5-Pro indicates a more balanced trade-off between precision and recall compared to the F1 score of **0.83** for ChatGPT-4o.

In summary, while ChatGPT-4o performs well in detecting scam emails, Gemini-1.5-Pro exhibits a measurable edge in terms of accuracy, precision, recall, and F1 score, making it the more effective tool for this task based on the evaluated metrics.

5 Future Work

The results of this study highlight important insights into the performance of ChatGPT-4o and Gemini-1.5-Pro for scam email detection. However, there are several areas where future research can expand upon this work to further improve the effectiveness and applicability of AI-driven email classification systems. Below, we outline potential directions for future work:

5.1 Integration with Real-World Email Systems

Future research could focus on integrating ChatGPT-4o and Gemini-1.5-Pro into real-world email platforms to provide real-time protection against scam emails, particularly for vulnerable populations such as the elderly. Older individuals are often disproportionately targeted by phishing scams, including fake medical bills, fraudulent insurance offers, and tech support fraud, due to a lack of familiarity with digital threats or evolving scam tactics. By embedding AI-driven detection systems directly into email platforms, scam emails could be automatically flagged, filtered, or accompanied by easy-to-understand warnings, helping elderly users identify and avoid potential fraud.

Additionally, these systems could be designed to offer specific features tailored to the elderly, such as simplified interfaces, clearer notifications, and explanations in non-technical language. Future research could also explore how these tools could detect scams that exploit the unique concerns of older adults, such as healthcare or retirement-related fraud. By focusing on accessibility and ease of use, AI-integrated email systems could empower elderly users to navigate online communication safely, reducing their vulnerability to scams and fostering greater confidence in digital platforms.

5.2 Fine-Tuning and Customization

Both ChatGPT-4o and Gemini-1.5-Pro were evaluated using their default configurations. Future studies could investigate the potential for fine-tuning these models on domain-specific datasets or incorporating user feedback to improve their accuracy, precision, and recall. This could include adapting the models to better detect scams in niche domains, such as financial services or healthcare, where the language and tactics used by scammers may differ significantly.

5.3 Analysis of Model Errors

A detailed error analysis could provide valuable insights into the limitations of both models. Future work could examine the types of scam emails that were misclassified by each model, identifying common patterns or linguistic features that the models struggled to detect. Understanding these failure cases could inform the development of more robust detection algorithms.

References

[1] Burnes, D., Henderson, C. R., Sheppard, C., Zhao, R., Pillemer, K., & Lachs, M. S. (2017). Prevalence of financial fraud and scams among older adults in the United States: A systematic review and meta-analysis. *American Journal of Public Health*, *107*(8), e13–e21. https://doi.org/10.2105/ajph.2017.303821

[2] James, B. D., Boyle, P. A., & Bennett, D. A. (2013). Correlates of susceptibility to scams in older adults without dementia. *Journal of Elder Abuse & Neglect*, *26*(2), 107–122. https://doi.org/10.1080/08946566.2013.821809

[3] IEEE Journals & Magazine — IEEE Xplore. (2024). Multi-modal comparative analysis on execution of phishing detection using artificial intelligence. https://ieeexplore.ieee.org/abstract/document/10742373

[4] Alenzi, H. Z., Aljehane, N. O., & Department of Computer Science, Tabuk University. (2020). Fraud detection in credit cards using logistic regression. *International Journal of Advanced Computer Science and Applications*, *11*(12), 540–541. https://thesai.org/Downloads/Volume11No12/Paper_65-Fraud_Detection_in_Credit_Cards.pdf

[5] Papasavva, A., Johnson, S., Lowther, E., Lundrigan, S., Mariconti, E., Markovska, A., & Tuptuk, N. (2024, September 25). Application of AI-based models for online fraud detection and analysis. *arXiv.org*. https://arxiv.org/abs/2409.19022

[6] Jiang, L. (2024). Detecting scams using large language models. *arXiv.org*. https://arxiv.org/html/2402.03147v1

[7] Qi, Z., Fang, Y., Zhang, M., Sun, Z., Wu, T., Liu, Z., Lin, D., Wang, J., & Zhao, H. (2023, December 22). Gemini vs GPT-4V: A preliminary comparison and combination of vision-language models through qualitative cases. *arXiv.org*. https://arxiv.org/abs/2312.15011

[8] Radev, D. (2008), CLAIR collection of fraud email, ACL Data and Code Repository, ADCR2008T001, http://aclweb.org/aclwiki

[9] Google Developers. "Accuracy, Precision, and Recall." Machine Learning Crash Course, https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall.