
Scam Detective: Using Gamification to Improve AI-Powered Scam Awareness

Angelica Marie M. Casuela
University of the Philippines Diliman

With

In collaboration with Apart Research and BlueDot Impact

Abstract

This project outlines the development of an interactive web application aimed at involving users in understanding the AI skills for both producing believable scams and identifying deceptive content. The game challenges human players to determine if text messages are genuine or fraudulent against an AI. The project tackles the increasing threat of AI-generated fraud while showcasing the capabilities and drawbacks of AI detection systems. The application functions as both a training resource to improve human ability to recognize digital deception and a showcase of present AI capabilities in identifying fraud. By engaging in gameplay, users learn to identify the signs of AI-generated scams and enhance their critical thinking abilities, which are essential for navigating through an increasingly complicated digital world. This project enhances AI safety by equipping users with essential insights regarding AI-generated risks, while underscoring the supportive functions that humans and AI can fulfill in combating fraud.

Keywords: AI safety education, interactive learning, educational games, knowledge assessment, user engagement, fraud detection, AI-generated content, digital literacy

1. Introduction

1.1 Problem Statement

The increased accessibility and capabilities of large language models has created unprecedented challenges in identifying digital scams and fraudulent

communications. While AI can generate increasingly convincing deceptive content, it also offers potential solutions for detection. This project addresses the critical question: How can we effectively educate the public about AI-generated scams while demonstrating both the capabilities and limitations of AI-powered fraud detection systems? "AI Scam Detective" creates a web-based game where users are tasked to identify scams while gaining insights into AI capabilities and limitations in the fraud detection domain.

1.2 Background and Motivation

The landscape of digital fraud has been dramatically transformed by generative AI technologies. As highlighted by PricewaterhouseCoopers (n.d.), AI now enables fraudsters to create more believable content at a faster pace with increased volume, eliminating traditional "tells" such as poor spelling and grammar that historically helped users identify scams.

Simultaneously, research on educational approaches shows that gamification presents a particularly effective method for fraud awareness training. According to Davis and Grant (2022), gamified phishing training reduces user susceptibility compared to traditional methods while maintaining significantly higher engagement levels. This finding is further supported by Arachchilage et al. (2013), whose work showed that participants who engaged with mobile game-based training were more likely to correctly identify fraudulent websites compared to those who only received text-based information.

The risk dimensions identified in OpenAI's Preparedness Framework (OpenAI, 2023), particularly focusing on persuasion risks, highlights where AI could generate highly convincing deceptive content. By creating an educational game that demonstrates both the generation and detection capabilities of AI systems, we help users develop more calibrated trust in technology while building practical resilience against emerging threats.

1.3 Threat Model and Safety Implications

This project leverages OpenAI's Preparedness Framework to address AI safety challenges by focusing on cybersecurity and persuasion risks through an interactive educational game.

In terms of cybersecurity, the game demonstrates how AI can be misused to craft convincing phishing attempts and social engineering attacks, even in scenarios with low to medium risk where explicit programming is not required. It also brings to light the limitations of current automated detection systems, showing that sophisticated, AI-generated scams can evade these defenses, and emphasizes the importance of human-AI collaboration to enhance overall security.

On the persuasion front, the game—titled "AI Scam Detective"—exposes users to a spectrum of deceptive tactics, ranging from overt scam patterns to highly tailored messages capable of deceiving even vigilant individuals.

This approach transforms abstract safety concerns into a practical, hands-on learning experience that bridges the knowledge gap between experts and everyday users, equipping them with the insights needed to navigate and mitigate real-world AI risks without requiring technical expertise.

2. Methods

2.1 Approach

This project is inspired by Lakera’s Gandalf web game that explores adversarial attacks on LLMs. By using a gamified approach on educating the public with AI-powered fraud generation, we can increase engagement in educating oneself of patterns of scam-related messages. The game presents users with a series of text messages, challenging them to identify scams. At the same time, an AI safety layer whose goal is to detect fraud, analyzes the same messages, in which the player needs to score higher against.

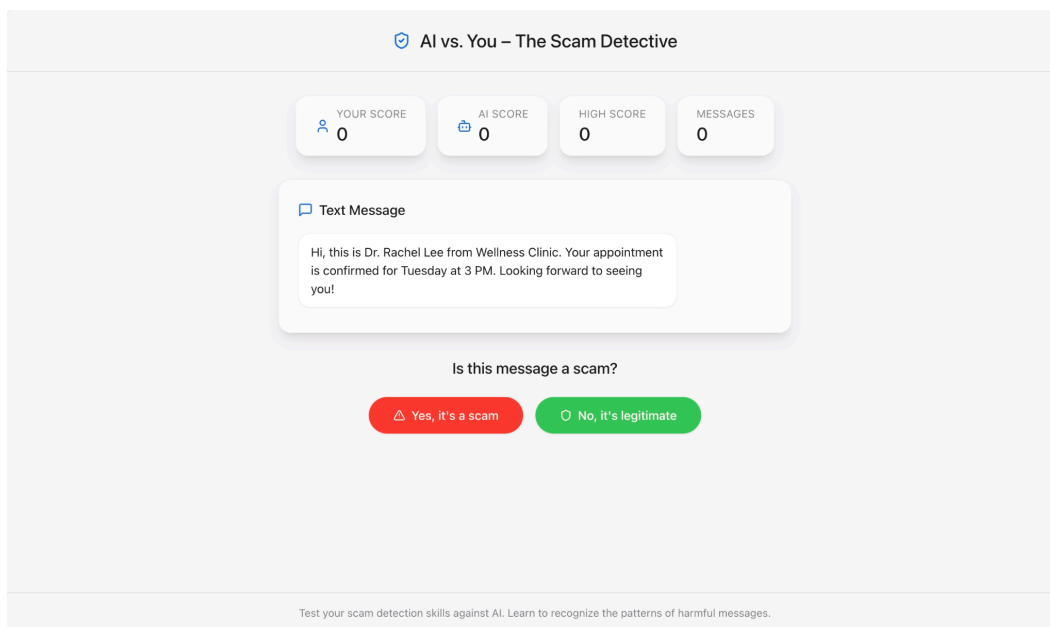


Figure 1 - The landing page of the Scam Detective game
(<https://ai-scam-detect-game.pages.dev>)

Users receive immediate feedback on their choices alongside an explanation by the AI safety layer, fostering an interactive learning experience. This game becomes an interactive AI risk demonstration presented by

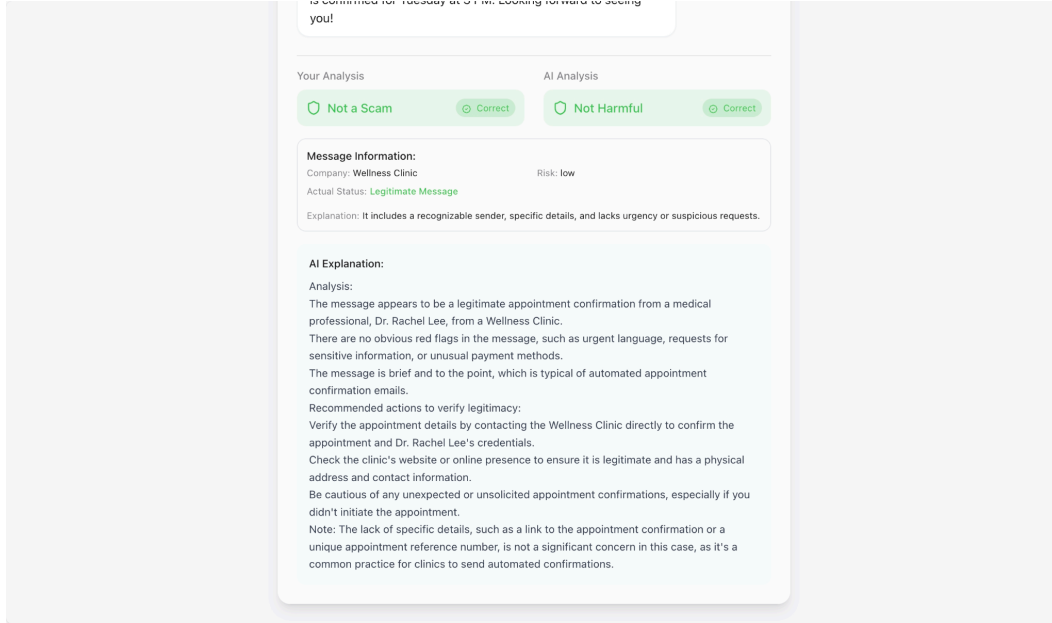
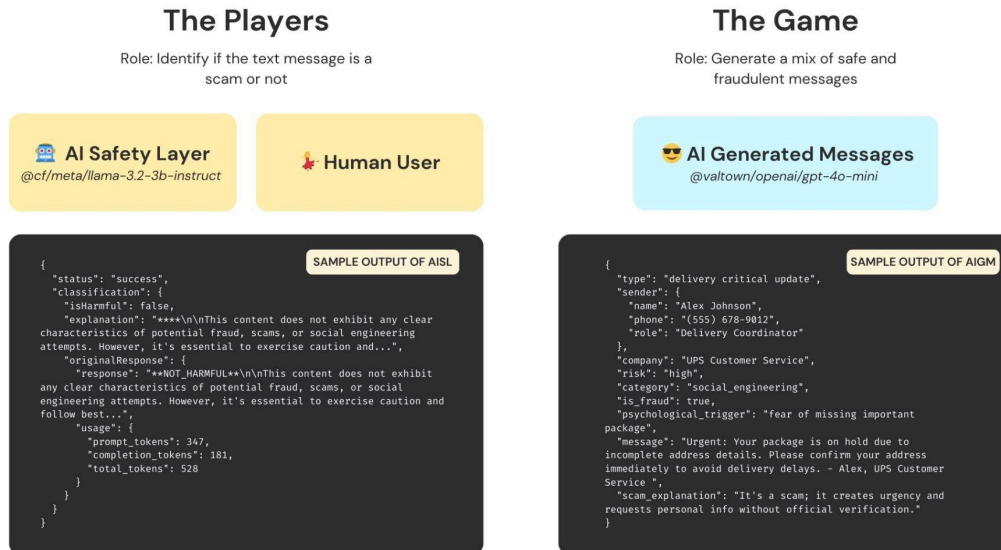


Figure 2 - The results of the player and AI's classification

2.2 Implementation

This web-based game utilizes the following LLMs with different respective utilities. The AI message generator makes use of OpenAI's GPT-4o-mini provided by Val.Town. The AI safety layer, the other player that the human is to play against, uses Meta's Llama-3.2-3b-instruct model provided by Cloudflare's Workers AI.



A general gameplay is a user is provided with a text message that invokes various social engineering tactics against the receiver of the message to either scam or

deceive. The AI message generator is instructed to provide a mix of safe and fraudulent messages to provide variety. The human player and the AI player receive points when answering correctly, and the game ends when the human player gets the classification wrong.

3. Results

3.1 Analysis and Findings

Initial feedback from testers show that the game presents low difficulty in classifying messages that contain fraud or scams. On the other hand, the AI safety layer can vary in performance on fraud detection that can be attributed to lower capabilities. Given that the message generator utilizes a more advanced model, it can produce convincing or sneaky attempts to scam the receiver by using a variety of social-engineering related tactics which the AI safety layer fails to detect.

3.2 Impact Assessment

The game effectively raises awareness about AI-generated scams and enhances users' detection capabilities. By engaging users in a competitive and educational environment, it contributes to broader efforts in cybersecurity education and fraud prevention.

4. Discussion and Conclusion

"Scam Detective" demonstrates a proof-of-concept of gamified learning in building public resilience against AI-generated threats. By directly experiencing both the capabilities and limitations of AI in scam generation and detection, users develop practical skills applicable to their daily digital interactions.

Users not only learn about scams but also develop a nuanced understanding of AI capabilities - where it excels and where human judgment remains superior. This "calibrated trust" in AI systems represents an important safety awareness outcome, as over-reliance on imperfect detection systems presents its own risks.

This game demonstration shows potential to be modified for other AI safety domains, helping users understand concepts like algorithmic bias, recommendation system manipulation, or content moderation challenges. By making these complex topics tangible through interactive experiences, we can build broader public literacy around AI systems and their potential risks.

As AI capabilities continue to advance, the gap between technical understanding and public awareness creates significant safety vulnerabilities. This informed perspective is crucial for developing appropriate trust in technology while maintaining necessary vigilance against potential misuse, ultimately building a more resilient digital society.

5. References

Arachchilage, N.A., Love, S., & Maple, C. (2013). Can a Mobile Game Teach Computer Users to Thwart Phishing Attacks? ArXiv, abs/1511.01622.

- Davis, N., & Grant, E.S. (2022). Simulated Phishing Training Exercises versus Gamified Phishing Education Games. 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), 1-8.
- OpenAI. (2023, December 18). *Preparedness Framework (Beta)*. OpenAI.
<https://cdn.openai.com/openai-preparedness-framework-beta.pdf>
- PricewaterhouseCoopers. (n.d.). The emerging threat of AI-powered fraud. PwC.
<https://www.pwc.com/ng/en/publications/the-emerging-threat-of-ai-fraud.html>