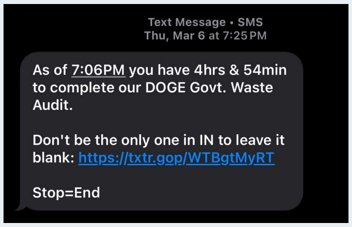
Lie to Me, Not to My AI: Unmasking Scams with Large Language Models

Yuan-Chen Pearl Chang, Esma Aïmeur

Université de Montréal

INTRODUCTION

- **GenAI** makes scams easier to launch and harder to detect [1].
- Existing solutions are fragmented and struggle with scams that unfold over time [2].



Text Message • SMS
Thu, Mar 6 at 7:25 PM

As of 7:06PM you have 4hrs & 54min to complete our DOGE Govt. Waste Audit.

Don't be the only one in IN to leave it blank: https://txtr.gop/WTBgtMyRT

Stop=End

Figure 1. An example of a scam.

CONTRIBUTIONS

LinkLynx: A Modular Scam Website Detection System

SCOUT: A **S**cam **C**hat **O**bservation & **U**nderstanding **T**ool

DECEPT: A **D**ataset for **E**valuating **C**hat and **E**mbedded **P**hishing **T**hreats

METHODOLOGY

- **LinkLynx** analyzes the **URL** structure, **WHOIS** data, and webpage **text + screenshot**.
- Each module uses an **LLM** for risk scoring and explanation.

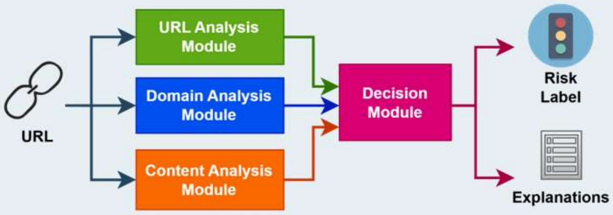


Figure 2. **LinkLynx** Architecture.

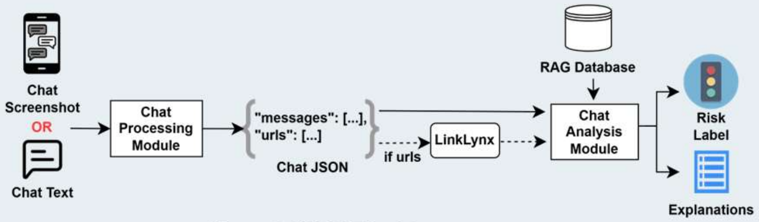


Figure 3. **SCOUT** Architecture.

- **SCOUT** uses a multimodal LLM to extract and format chat messages and URLs.
- Sends URLs to **LinkLynx**.
- Scores chat risk using an **LLM** + **R**etrieval-**A**ugmented **G**eneration.

EXPERIMENT & RESULTS

Dataset Overview

DECEPT-URL
- Raw URLs
- WHOIS data
- Webpage text + screenshots

DECEPT-Chat
- Messages
- Image captions
- URL risk

Table 1. DECEPT Statistics.

Dataset	Samples	Scam %	Legit %
DECEPT-URL	15000	49%	51%
DECEPT-Chat	1504	53%	47%

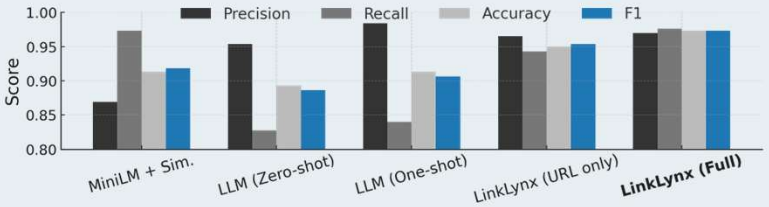


Figure 4. Performance of Baseline Models vs. LinkLynx.

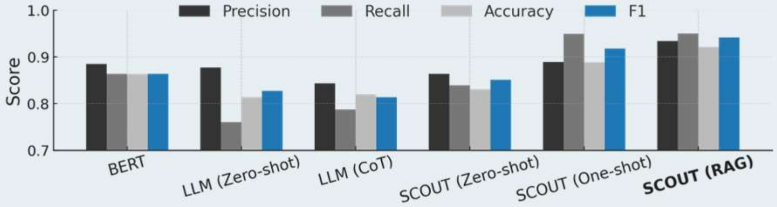


Figure 5. Performance of Baseline Models vs. SCOUT.

CONCLUSION

- **LLMs + context = more effective scam detection.**
- Next steps: Human evaluations, scam-adaptive feedback loop, robustness against adversarial attacks.

REFERENCES

[1] Chang, Y.-C., & Aïmeur, E. (2024). "Is this site legit?": LLMs for scam website detection. In Web Information Systems Engineering – WISE 2024.

[2] Chang, Y.-C., & Aïmeur, E. (2024). Chat or trap? Detecting scams in messaging applications with large language models. In 2024 8th Cyber Security in Networking Conference (CSNet).