EVALUATING ADVERSARIAL LLM REWRITING ATTACKS AGAINST NLP-BASED PHISHING DETECTORS

Charles Liu, Tashi Stirewalt
Washington State University, Pullman WA

Introduction

Phishing detection systems increasingly rely on machine learning and transformer-based NLP models to identify malicious emails with high accuracy. However, strong benchmark performance does not necessarily imply robustness against adaptive attacks. Large language models (LLMs) can generate fluent, professional rewrites of phishing emails that preserve malicious intent while removing obvious spam indicators.

This project evaluates the robustness of phishing email classifiers against adversarial attacks, including token injection and LLM-generated semantic rewrites. We demonstrate that even highly accurate models can be vulnerable to realistic black-box evasion attacks generated through automated language rewriting.

Can LLM-generated semantic rewrites evade high-performing phishing email classifiers in realistic black-box settings?

Background

Phishing Emails
Phishing attacks use deceptive emails to trick users into revealing passwords, financial information, or other sensitive data.
Natural Language Processing (NLP)
Natural Language Processing (NLP) is a field of artificial intelligence focused on enabling computers to understand and analyze human language. NLP models are widely used in spam filtering, phishing detection, and text classification.
Large Language Models (LLMs)
Large language models (LLMs) such as modern chatbots can generate highly fluent and realistic text. While useful, they may also help attackers create more convincing phishing emails.

Dataset

Email Dataset
- ~82,000 labeled emails
- Binary classification:
  - Phishing
  - Benign

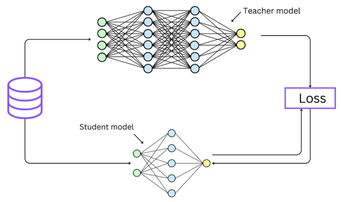
Preprocessing
- Text cleaning
- Train/validation/test split
- Tokenization and vectorization

Model

Victim Models

Baseline Model
- TF-IDF + Logistic Regression
- Accuracy: ~98.6%

Transformer Victim Model
DistilBERT
- Accuracy: ~99.7%



Threat Models

Attacker Assumptions
- Black-box access only
- No gradients or model weights
- Can modify email text

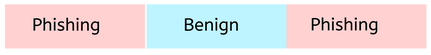
Goal
- Reduce phishing detection probability
- Preserve phishing functionality and intent

Attack Methods

Token Injection Attacks
- Insert benign-looking words or phrases
- Example:
  - “meeting”
  - “report”
  - “thanks”

Position-Based Attacks
- Compare:
  - prepend
  - append
  - middle insertion

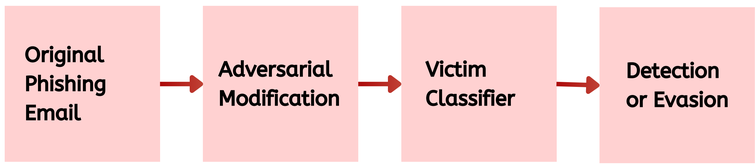
Middle Insertion Attack



LLM Rewrite Attacks
- Use local LLM (Ollama)
- Rewrite phishing emails into polished professional emails
- Preserve underlying malicious intent

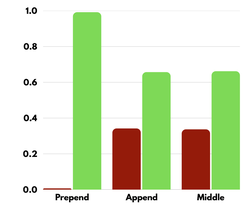
Ollama
Get up and running with large language models.

Experimental Setup



Results

Attack Type	Detection Rate	Evasion / Flip Rate
Baseline (No Attack)	99.70%	0.00%
Prepend Injection	95.90%	3.80%
Append Injection	99.00%	0.70%
LLM Rewrite Attack	88.00%	11.70%



Positional (Strongest Attack)

Analysis

The results show that high phishing detection accuracy does not necessarily imply robustness against adversarial attacks. Token injection and positional attacks significantly reduced detection rates, with prepend-based attacks proving especially effective. This suggests that transformer models may be highly sensitive to **early-token representations**.

LLM-generated semantic rewrites were also successful at evading detection while preserving phishing intent. These findings suggest that phishing classifiers rely heavily on lexical phishing indicators rather than fully understanding malicious semantic intent.

Future Work

Defensive Approaches
- Adversarial training
- Semantic consistency detection
- Robust transformer fine-tuning

Future Attacks
- Multi-turn LLM optimization
- Multilingual phishing attacks
- Persona and roleplay-based rewriting

Key Sources & Acknowledgements

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS).
[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT 2019.
[3] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
[4] Jin, D., Jin, Z., Zhou, J. T., & Szolovits, P. (2020). Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. Proceedings of the AAAI Conference on Artificial Intelligence.
[5] Wallace, E., Feng, S., Kandpal, N., Gardner, M., & Singh, S. (2019). Universal Adversarial Triggers for Attacking and Analyzing NLP. Proceedings of EMNLP-IJCNLP 2019.