TAED: Trust-Aware Explainable Defense Against Phishing Emails

Tam Nguyen, Gabriella Campos, Riley Phan, Rahul Shrestha, Jayaraman Srinivas; Advisor: Dr. Robin Chataut
Department of Computer Science, Texas Christian University

INTRODUCTION & MOTIVATION

The Problem

~16% of data breaches involve phishing — avg. cost **$4.8M** per incident

1M+ phishing attacks in Q1 2025 alone — **33% surge** in BEC incidents

Current AI detectors can be fooled by small, invisible text changes

High model confidence does not always mean a correct decision

Research Questions & Contributions

RQ1 Are complex AI models more easily fooled than simpler ones when emails are slightly altered?

RQ2 Can analyzing the reasoning behind a model's decision help catch attacks that standard methods miss?

RQ3 Does TAED's approach outperform existing detection systems when facing these manipulated emails?

ADVERSARIAL DATASET CONSTRUCTION

4,673 Adversarial Samples | 4 Attack Vectors

To test how well TAED holds up under pressure, we created a dataset of adversarial emails — messages that look normal to a person but are designed to fool AI detectors. Each attack strategy subtly alters the email in a different way:

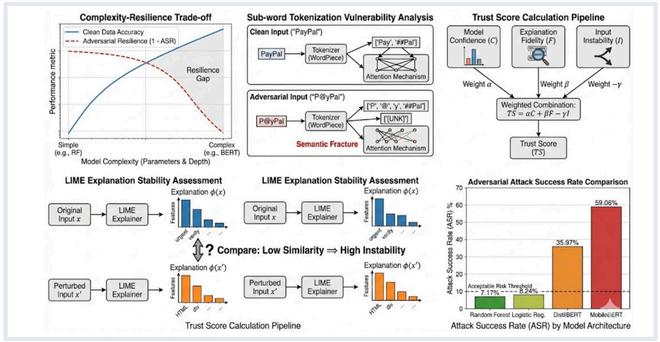
Attack Vector	Clean Example (Before)	Adversarial Example (After)
Homoglyph	Confirm your **PayPal** account.	Confirm your **PayPal** account. (Cyrillic 'а' U+1EA1 breaks tokenizer)
Paraphrasing	Please click below to login.	Immediate authentication is required below. (Generated via T5 model, alters n-grams)
URL Obfuscation	paypal.com/signin	bit.ly/secure-login-882 (Shortened link hides true destination)
Noise Injection	Dear Customer, regarding...	D#e@r C!u$tomer, re-ga_rding... (High-frequency character perturbation)

All modifications were tested at 10%, 30%, and 50% intensity while preserving the email's readability to human recipients

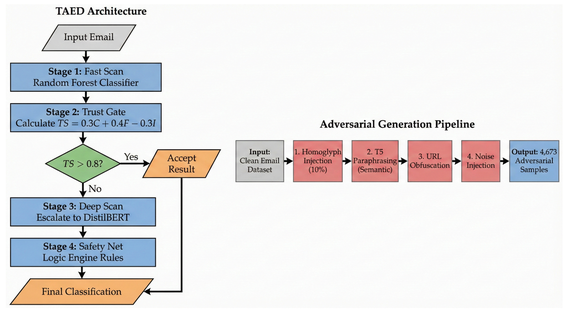
Data Sources

Phishing Email Dataset (Alam/Kaggle) • **Enron Email Dataset** • **Phishing Curated** (Zenodo 8339691)

METHODOLOGY: TAED FRAMEWORK



4-Stage Detection Pipeline



RESULTS

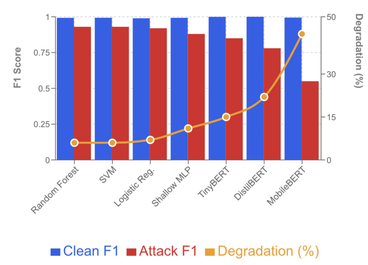
7.89% TAED Attack Success Rate

59.1% MobileBERT Attack Success

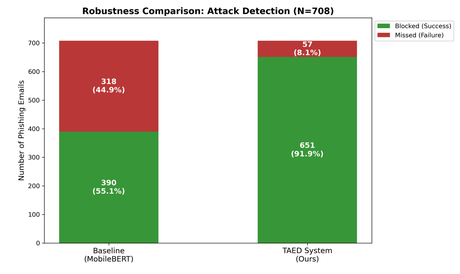
99.3% Clean Data Accuracy

28ms Stage 1 Latency

Attack Success Rate by Architecture



Robustness Recovery Analysis



DISCUSSION

Does Every Component Actually Matter?

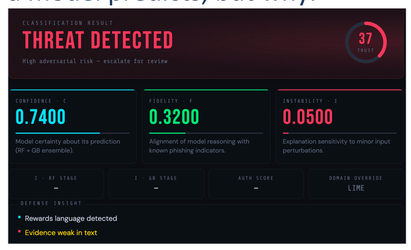
We tested what happens when we remove each part of the Trust Score one at a time. The results show that all three components work together — removing any one of them makes the system significantly weaker:

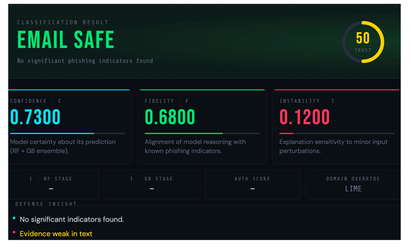
Configuration	Components	Failure Mode	ASR	FER
Confidence Only	$TS = C$	Accepts high-confidence adversarial errors	29.4%	–
C + Fidelity	$TS = 0.5C + 0.5F$	Misses brittle reasoning; accepts $F > 0.3$ attacks	18%	–
C + Instability	$TS = 0.5C - 0.5I$	Cannot validate semantic grounding; over-escalates	–	12%
F + Instability	$TS = 0.5F - 0.5I$	No efficiency filter; ignores model certainty	–	>25%
Full (TAED)	$TS = 0.3C + 0.4F - 0.3I$	**None**	**7.89%**	**4%**

Key insight: When content alone cannot distinguish a legitimate urgent email from a phishing attempt, domain context provides the missing signal that makes the difference.

CONCLUSION

TAED shows that building trustworthy AI means checking not just what a model predicts, but why.





Key Findings

- Simpler models (Random Forest) proved more resilient than large transformers under adversarial attack
- Confidence scores alone are not enough — TAED's Trust Score reduced attack success from 59% to 7.89%
- Explanation quality (fidelity + stability) is a reliable signal for detecting adversarial manipulation
- Domain context resolves edge cases that content analysis alone cannot handle