

Executive Summary

Why Phishing Remains an Open Problem

Phishing is the entry vector for 91% of cyberattacks. Existing defenses either block emails outright, preventing analyst investigation, or label them without explaining the attack mechanism.

Current tools fail analysts who need to safely inspect hostile content without exposing their IP address, OS fingerprint, or browser version to attacker infrastructure.

3.4B

PHISHING EMAILS/DAY

91%

ATTACKS START VIA EMAIL

\$4.9B

ANNUAL BUSINESS LOSS

96.2%

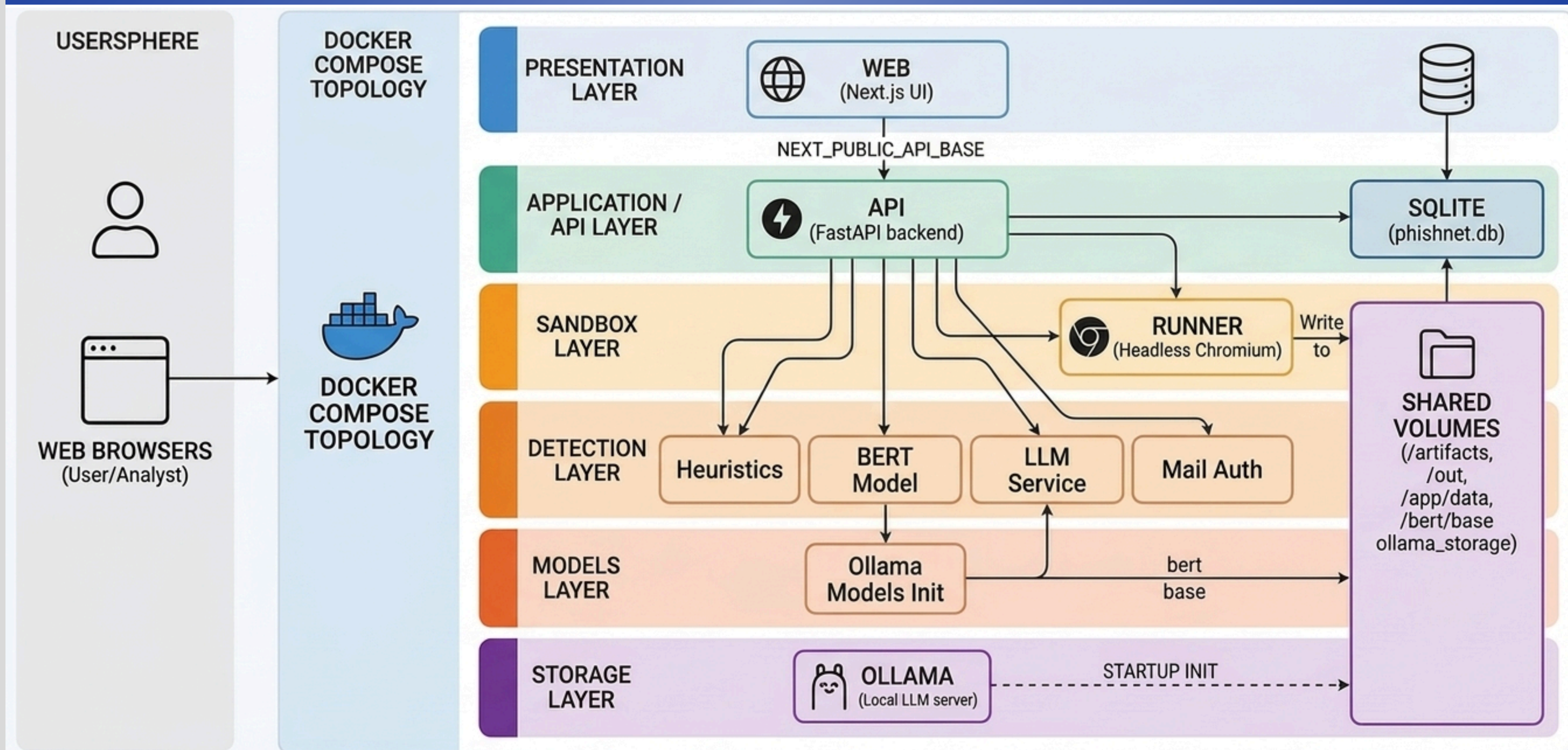
PHISHNET F1 SCORE

Research Contribution

Novel Contributions

- C1 Ensemble fusion of three heterogeneous detectors, rule-based heuristics, fine-tuned BERT, and local LLM, combined via pessimistic max-score fusion.
- C2 Adversarial sandbox for safe email rendering that blocks tracking pixels, JavaScript execution, and IP fingerprinting during analyst investigation.
- C3 Privacy-preserving architecture - all-inference runs locally via Docker; no email content leaves the analyst's machine.
- C4 Cross-correlation scoring across 5 independent signal layers; high-confidence multiplier applied at 4+ concurrent indicators.
- C5 LLM safe rewrite – hostile emails neutralized into link-stripped readable text, letting analysts understand attacker intent without risk.

Architecture



Experiment Demo

[TRAINING SAMPLE] Urgent account verification required

From: Security Training <training@example.com>

ID: f4115057-21f3-4a5f-9b85-21724e1715b4

Mail authentication (from headers)

Source: none

SPF: not stated DKIM: not stated DMARC: not stated

No Authentication-Results (or ARC) header found. The .eml may be from a client export without provider auth stamps, or headers were stripped.

Run detection Safe rewrite

Detection Results

[Progress bar]

[LLM] link mismatch

[LLM] credential request

[BERT] BERT: 63.4% phishing probability (36.6% legit)

[BERT] Model stats: acc=99.3%, f1=99.3%

ORIGINAL (TEXT ONLY)

SAFE REWRITE (LLM: NO)

Detection Methods

Heuristic Engine

Rule-based · always available

Sender analysis Brand spoofing URL inspection

Punycode / IDN Raw IP links Free hosting

Urgency language Cross-correlation

Negative signals

BERT Classifier

DistilBERT · 96.2% F1

Fine-tuned 5K samples 3 epochs Phishing prob.

0-100 score In-process CPU ~268 MB weights

Git LFS No GPU needed

LLM Analysis

Llama 3.2 1B · Ollama

Intent analysis Context-aware Narrative logic

Social engineering Fully local No external calls

~1.3 GB model Safe rewrite Explainable

Experimental Results

Total Samples

54,804

balanced

Train Split

46,583

samples

Holdout

8,221 samples (12%)

Test Split

4,111 samples

Hyperparameters:

6 Epochs 32 Batch 1500 chars limit 516 Batch 512 max len

Test-Set Metrics

Accuracy ~99.29% Macro F1 ~99.30% Precision ~99.17% Recall ~99.42%

Based on reported values in training_meta.json, evaluated on test split

Inference & Post-Hoc Rules

Structured text input (subject, From, body max ~1,500 chars, up to 10 URLs)

[SEP]-style joining, max length 512

Technical URL signals (raw IP, punycode, deep subdomain) boost scores < 60 to at least 65

Thresholds ≥ 65 Phishing ≥ 35 Suspicious else Benign

Conclusions

SYSTEM	ACC.	LOCAL	SANDBOX	EXPLAIN.
Chowdhury et al. [1]	~92%	No	No	Partial
Thapa et al. [2]	~94%	No	No	Yes
El-Rashidy et al. [4]	99.7%	No	No	No
Alhuzali [3]	~98%	No	No	Partial
PhishNet (ours)	96.2% F1	Yes	Yes	Yes

References

Phishing as a widespread threat

- [1] Chowdhury et al. (2023). Phishing Detection Using NLP and ML. SMU Data Science Review, 6(2). <https://scholar.smu.edu/datasciencereview/vol6/iss2/14>

Multi-layer / ensemble detection (core to PhishNet's approach)

- [2] Thapa et al. (2025). Evolution of Phishing Detection with AI. arXiv:2507.07406. <https://arxiv.org/abs/2507.07406>
- [3] Alhuzali (2025). In-Depth Analysis of Phishing Email Detection. Applied Sciences, MDPI. <https://doi.org/10.3390/app15063396>

BERT specifically for phishing

- [4] El-Rashidy et al. (2024). Advancing Phishing Email Detection. Sensors, MDPI. <https://doi.org/10.3390/s24072077>

[5] ealvaradob/phishing-dataset. HuggingFace Datasets Hub. 5,000 balanced samples.