

AI-Driven Phishing Detection: Comparing Transformer-Based Models against Traditional Heuristics

***Dr. Prashant Kohli**

Associate Professor, Department of Physics, N.M.S.N. Dass (P.G.) College, Budaun

Abstract: *Phishing remains one of the most prevalent and financially damaging categories of cyber-attack, exploiting human trust through deceptive emails, websites, and messages to harvest credentials, deliver malware, or facilitate fraud. Detection systems have historically relied on heuristic and rule-based methods — blacklists, URL pattern matching, header analysis, and handcrafted lexical features — that are fast and interpretable but brittle against novel or obfuscated attacks. The rise of transformer-based language models, pertained on massive text and code corpora, has introduced a new generation of phishing detectors capable of capturing deep semantic and contextual signals that static heuristics cannot. This paper presents a systematic comparison between transformer-based phishing detection models (including BERT-family classifiers and large generative LLMs used in zero-shot and fine-tuned configurations) and traditional heuristic-based systems (blocklists, rule engines, and classical machine learning models such as logistic regression, random forests, and support vector machines trained on handcrafted features). We evaluate both paradigms across email phishing, URL/website phishing, and SMS/smishing datasets, examining detection accuracy, false positive rates, robustness to adversarial obfuscation, latency, interpretability, and adaptability to novel campaigns. Our analysis of published benchmarks and experimental studies indicates that transformer-based models consistently achieve higher recall on novel and linguistically sophisticated phishing attempts, particularly those generated with the assistance of generative AI, while traditional heuristics retain advantages in inference speed, resource efficiency, and transparency for compliance-driven environments. We further explore hybrid architectures that combine rule-based pre-filtering with transformer-based semantic scoring to balance accuracy against operational cost. The paper concludes with a discussion of deployment considerations, the emerging arms race between generative AI-authored phishing content and AI-based defenses, and directions for future research including multimodal detection, real-time adaptation, and privacy-preserving on-device inference.*

Keywords: *Phishing detection, Transformer models, Natural language processing, Cybersecurity*

1.Introduction

Phishing attacks continue to represent one of the most common initial access vectors in cyber security incidents, exploited to steal credentials, distribute ransom ware, and perpetrate business email compromise (BEC) fraud. Industry incident reports consistently rank phishing and social engineering among the top causes of data breaches, and the average cost of a successful phishing-enabled breach has continued to climb as attackers refine their techniques. Traditional phishing detection relies heavily on heuristic mechanisms: block lists of known malicious domains, rule-based scoring of suspicious URL structures (e.g., presence of IP addresses instead

of domain names, excessive sub domains, homoglyph substitution), sender reputation databases, and handcrafted lexical or structural features fed into classical machine learning classifiers.

These heuristic approaches share a fundamental limitation: they are reactive. Block lists require a malicious domain to be observed and reported before it can be blocked, leaving a window of vulnerability during which newly registered phishing infrastructure operates undetected. Rule-based systems require security teams to manually encode new patterns as attackers evolve their tactics, a process that consistently

lags behind attacker innovation. Furthermore, the increasing use of generative AI by attackers themselves — to produce grammatically fluent, contextually tailored phishing emails free of the spelling and grammar errors that historically served as reliable heuristic signals — has begun to erode the effectiveness of lexical heuristics that depend on surface-level anomalies.

Transformer-based language models, first popularized in natural language processing through architectures such as BERT, RoBERTa, and later scaled into large generative models such as the GPT and Claude families, offer a fundamentally different detection paradigm. Rather than relying on predefined rules, these models learn distributed semantic representations of text and can be fine-tuned or prompted to recognize the subtle contextual, psychological, and structural markers of phishing content — urgency framing, impersonation of trusted brands, mismatches between display text and underlying URLs, and deceptive call-to-action phrasing — even in the absence of the surface-level errors that older heuristics depended upon. This paper undertakes a structured comparison of these two paradigms and there are three central questions: (1) How do transformer-based models perform relative to traditional heuristics across different phishing modalities (email, URL, SMS)? (2) What are the respective strengths and weaknesses of each paradigm in terms of accuracy, robustness, latency, and interpretability? (3) What hybrid or complementary architectures offer the best practical trade-off for real-world deployment? The remainder of the paper is organized as follows: Section 2 reviews related work; Section 3 describes traditional heuristic approaches in detail; Section 4 describes transformer-based approaches; Section 5 presents a comparative experimental methodology and representative findings; Section 6 analyses robustness and adversarial considerations; Section 7 discusses hybrid architectures; Section 8 considers deployment and operational factors; Section 9 discusses the

emerging generative-AI arms race; Section 10 outlines future work; and Section 11 concludes.

2. Related Work:

2.1 Heuristic and Rule-Based Phishing Detection:

Early phishing detection research focused heavily on identifying discriminative features of malicious URLs and emails: domain age, WHOIS registration anomalies, use of URL shorteners, presence of suspicious keywords ("verify your account," "urgent action required"), mismatched anchor text and hyperlink targets, and SSL certificate irregularities. These features were commonly fed into classical machine learning classifiers such as decision trees, random forests, and support vector machines (SVMs), producing systems such as Phish Tank-integrated block lists and various academic URL-classification frameworks. Google Safe Browsing and similar industry block list services remain widely deployed today, offering low-latency, high-precision blocking of previously identified malicious infrastructure.

2.2 Classical Machine Learning

Approaches:

Beyond simple heuristic scoring, a substantial body of research applied classical supervised learning to handcrafted feature vectors derived from email headers, HTML structure, and lexical content. These systems, exemplified by tools using Naive Bayes or SVM classifiers on bag-of-words or TF-IDF representations of email text, achieved reasonably strong performance on datasets available at the time but were shown in later work to generalize poorly to phishing campaigns using vocabulary or structural patterns not represented in their training data.

2.3 Deep Learning and Transformer-Based Detection:

The introduction of word embedding (Word2Vec, GloVe) and subsequently transformer architectures (BERT, RoBERTa, Distil BERT) enabled models to capture richer

semantic and contextual representations of email and web content. Fine-tuned BERT-based classifiers applied to phishing email corpora have demonstrated strong performance gains over classical bag-of-words baselines, particularly in identifying semantically deceptive content that lacks the lexical red flags targeted by earlier heuristic systems. More recent work has explored the use of large generative LLMs (GPT-3.5/4-class and open models) in zero-shot and few-shot configurations for phishing classification, as well as for generating synthetic phishing content to augment training data and stress-test detection systems.

2.4 Multimodal and URL-Specific Transformer Models:

Given that phishing often manifests through visually deceptive websites (e.g., brand impersonation via lookalike login pages), researchers have also explored vision-transformer and multimodal architectures that jointly analyze webpage screenshots, HTML structure, and URL text. These approaches aim to detect visual brand impersonation that purely textual analysis might miss, complementing the natural-language-focused transformer detection discussed in this paper.

3. Traditional Heuristic Approaches in Detail:

3.1 Blocklist and Reputation-Based Filtering:

Blocklist systems maintain continuously updated databases of domains, IP addresses, and URLs previously confirmed as malicious, often aggregated from community reporting services, honeypots, and threat intelligence feeds. Their principal strength is extremely low false-positive rates for previously seen threats and negligible computational overhead at inference time. Their principal weakness is an inherent lag: newly registered phishing domains, which are frequently used for only hours before being abandoned, can evade

blocklist-based defenses entirely during their brief operational window.

3.2 Rule-Based Lexical and Structural Heuristics:

Rule engines apply manually engineered scoring functions to structural and lexical features: presence of urgency-inducing keywords, sender-domain mismatches, suspicious attachment types, excessive use of redirect chains, and use of homoglyphs or Unicode characters designed to visually mimic legitimate brand names. These systems are fast, fully interpretable (each flagged message can be traced to a specific triggered rule), and require no training data, making them attractive for regulated environments requiring auditability. However, rule thresholds must be manually tuned and updated, and sufficiently motivated attackers can iteratively test messages against known detection rules to identify combinations that evade them.

3.3 Classical Machine Learning on Handcrafted Features:

Feature-engineered classical models (logistic regression, random forest, gradient boosting, SVM) trained on structured feature vectors combining URL lexical statistics, WHOIS metadata, HTML DOM structure counts, and email header anomalies offer a middle ground between pure rule engines and deep learning. These models can learn non-linear combinations of heuristic features and often achieve strong benchmark accuracy, but remain fundamentally limited by the expressiveness of the handcrafted feature set: any semantic nuance not captured by the chosen features (e.g., a subtly deceptive turn of phrase) is invisible to the model regardless of the underlying learning algorithm's sophistication.

4. Transformer-Based Detection Approaches:

4.1 Fine-Tuned Encoder Models:

The dominant transformer-based approach for phishing classification fine-tunes an encoder-

only model (BERT, RoBERTa, Distil BERT, or domain-specific variants) on labelled phishing/legitimate corpora, using the pooled representation of the input sequence (email body, subject line, or URL string) as input to a classification head. This approach benefits from transfer learning: the base model's pertaining on massive general-domain text provides a strong semantic foundation that fine-tuning specializes toward phishing-specific cues, typically requiring substantially less labelled data than training a comparable model from scratch.

4.2 Prompt-Based Classification with Generative LLMs:

Large generative LLMs can be applied to phishing detection through zero-shot or few-shot prompting, in which the model is presented with the suspect content and asked to classify it as phishing or legitimate, optionally providing a natural-language justification (e.g., identifying specific manipulative phrasing, brand impersonation, or suspicious links). This approach requires no task-specific training, generalizes flexibly across languages and phishing sub-types (email, SMS, social media messages), and produces interpretable, human-readable rationale an advantage over both classical heuristics (which surface a rule ID rather than a rationale) and fine-tuned encoder models (which typically output only a classification score).

4.3 URL and Domain-Specific Transformer Variants:

Specialized transformer architectures have been developed to operate directly on URL character sequences, treating the URL as a token sequence and learning to detect deceptive lexical patterns (typo squatting, excessive subdomain nesting, brand-name substring insertion) without relying on external WHOIS or reputation data. These character-level transformer models are particularly valuable for detecting newly registered domains that have no block list history.

4.4 Retrieval-Augmented and Agentic Approaches:

More sophisticated pipelines combine an LLM with retrieval mechanisms that pull in real-time domain registration data, certificate transparency logs, or brand-impersonation reference databases, allowing the model to ground its classification in current external evidence rather than relying solely on static pertaining knowledge. Agentic systems may additionally render a suspect URL in a sandboxed browser, capture a screenshot, and pass both the visual rendering and underlying HTML/text to a multimodal model for joint visual-textual analysis, closely mirroring how a human analyst might investigate a suspicious link.

5. Comparative Experimental Methodology and Findings:

5.1 Benchmark Datasets:

Comparative evaluation typically draws on several standard datasets:

Nazario Phishing Corpus and Enron-derived legitimate email sets, commonly combined to create labelled email phishing/legitimate benchmarks.

- **Phish Tank and Open Phish URL feeds:** providing continuously updated real-world phishing URL samples paired with legitimate URL samples from sources such as Alexa/Tranco top-site lists.
- **UCI Phishing Websites Dataset:** a widely used structured-feature benchmark for classical ML comparison.
- **SMS Spam Collection and smashing-specific corpora:** used to evaluate detection performance on SMS-based phishing.
- **Synthetically generated LLM-authored phishing content:** increasingly used in recent studies to evaluate detection performance against AI-generated attacks specifically

designed to lack traditional lexical red flags.

5.2 Evaluation Metrics:

Standard classification metrics apply: precision, recall, F1 score, and area under the ROC curve (AUC-ROC). Given the asymmetric cost of phishing misclassification (a missed phishing email can lead to credential theft, while a false positive on a legitimate email disrupts business communication), false positive rate at a fixed high-recall operating point is often reported as a primary metric alongside overall F1. Latency (inference time per message) and computational cost are also reported when comparing operational feasibility, particularly for high-throughput email gateway deployments.

5.3 Representative Comparative Findings:

Synthesizing results across the literature, several consistent patterns emerge:

1. **On established, well-represented phishing patterns, heuristic and transformer-based approaches achieve comparable accuracy:** For phishing emails and URLs containing classic red flags — misspelled brand names, mismatched sender domains, known malicious infrastructure — both rule-based heuristics and transformer models achieve high F1 scores (often above 0.95), since these are precisely the patterns heuristic systems were designed to catch.
2. **Transformer-based models show a clear advantage on novel and linguistically sophisticated phishing content:** When evaluated against phishing emails generated or refined using generative AI — which tend to be grammatically correct, contextually plausible, and free of the classic lexical anomalies heuristics rely on — fine-tuned transformer models consistently outperform rule-based and classical ML
3. **Heuristic systems retain a significant edge in inference latency and resource efficiency:** Rule engines and lightweight classical ML models can classify messages in sub-millisecond time on commodity hardware, whereas transformer-based inference — particularly with larger generative LLMs — introduces latency and computational cost that can be prohibitive for high-throughput, real-time email gateway filtering unless carefully optimized (e.g., via model distillation or quantization).
4. **False positive behavior differs qualitatively between paradigms:** Heuristic systems tend to produce false positives on legitimate messages that happen to trigger surface-level rules (e.g., a legitimate marketing email using urgency language), while transformer-based models tend to produce false positives on legitimate messages with atypical but benign phrasing patterns not well-represented in training data. Neither paradigm eliminates false positives entirely, reinforcing the value of human review or hybrid escalation workflows for borderline cases.
5. **Classical ML on handcrafted features occupies a middle ground:** Random forest and gradient boosting models trained on structured URL and header features often match or exceed encoder-only transformer models on well-curated benchmark datasets such as the UCI Phishing Websites Dataset, since these benchmarks were themselves constructed around the specific handcrafted features such models exploit; this suggests some reported

"wins" for classical ML partly reflect benchmark-feature alignment rather than a general robustness advantage.

6. **Zero-shot generative LLM classification is competitive but inconsistent:** Without fine-tuning, general-purpose LLMs prompted to classify phishing content achieve respectable but variable accuracy, with performance sensitive to prompt phrasing and occasionally producing overcautious classifications (flagging benign marketing or transactional emails as phishing due to superficial similarity in tone), indicating that prompt engineering and few-shot calibration meaningfully affect deployment reliability.

6. Robustness and Adversarial Considerations:

6.1 Adversarial Evasion of Heuristic Systems:

Rule-based and block list systems are particularly vulnerable to systematic evasion once an attacker understands the detection logic: minor URL perturbations (adding benign-looking subdomains, using URL shorteners, employing homoglyph substitution), rotating infrastructure faster than block list update cycles, and avoiding known trigger keywords can each individually or jointly defeat heuristic detection with modest attacker effort.

6.2 Adversarial Evasion of Transformer-Based Systems:

Transformer-based classifiers are not immune to adversarial manipulation. Research on adversarial text perturbation has demonstrated that carefully crafted paraphrasing, synonym substitution, or insertion of semantically neutral filler text can shift a transformer classifier's output across the decision boundary while preserving the phishing email's functional intent to a human reader. Additionally, because transformer models learn statistical associations

from training data, sufficiently novel social engineering framings attack narratives not well-represented during training or fine-tuning can evade detection even without deliberate adversarial crafting.

6.3 The Generative AI Arms Race:

A defining feature of the current phishing landscape is the use of generative AI by attackers themselves to author phishing content, effectively removing many of the lexical anomalies (poor grammar, awkward phrasing, inconsistent tone) that both heuristic rules and earlier machine learning classifiers relied upon as discriminative signals. This dynamic has accelerated the shift toward transformer-based detection, since defending against AI-authored attacks increasingly requires detection systems capable of the same depth of contextual and semantic reasoning used to generate the attacks in the first place. This arms-race dynamic suggests that the performance gap between heuristic and transformer-based approaches will likely widen over time as generative AI-assisted phishing becomes more prevalent, rather than narrow.

7. Hybrid Architectures:

Given the complementary trade offs identified above, a layered hybrid architecture is increasingly favoured in both research prototypes and commercial deployments:

Stage 1 - High-Speed Heuristic Pre-

Filtering: Incoming messages or URLs are first screened against block lists and lightweight rule engines, immediately blocking known-malicious infrastructure and flagging obvious high-confidence indicators at minimal computational cost. This stage handles the majority of traffic with negligible latency.

Stage 2 - Classical ML Scoring on

Structured Features: Messages not resolved with high confidence in Stage 1 are passed through a classical ML model operating on structured features (URL statistics, header metadata, sender reputation history), providing

a fast secondary filter that further narrows the set of messages requiring deeper analysis.

Stage 3 - Transformer-Based Semantic Analysis: The remaining ambiguous or high-stakes messages (e.g., those addressed to high-privilege accounts, referencing financial transactions, or exhibiting borderline heuristic scores) are escalated to a fine-tuned transformer classifier or LLM-based analysis pipeline, which evaluates deeper semantic and contextual cues and produces both a classification and a natural-language rationale.

Stage 4 - Multimodal Verification for Web-Based Phishing: For URL-based threats specifically, a sandboxed rendering step captures a screenshot and DOM structure, which is jointly analysed by a multimodal model to detect visual brand impersonation not evident from URL or HTML text analysis alone.

Stage 5 - Human Analyst Review and Feedback Loop: Findings that remain ambiguous after automated analysis, or that involve high-impact targets, are escalated to human security analysts, whose determinations are fed back into the training pipeline for both the classical ML and transformer components, enabling continuous adaptation to emerging campaign patterns.

This staged architecture allows the majority of traffic to be filtered at minimal computational cost through heuristic and classical ML stages, reserving the higher latency and cost of transformer-based semantic analysis for the subset of messages where its additional accuracy provides the greatest marginal value directly addressing the latency limitation identified in Section 5.

8. Deployment and Operational Considerations:

8.1 Latency and Throughput:

Enterprise email gateways must process extremely high message volumes with minimal added latency; even a small per-message delay compounds significantly at scale. Techniques such as model distillation (training a smaller model to mimic a larger transformer's outputs), quantization, and caching of previously seen sender/domain scores are commonly employed to make transformer-based components operationally viable within the staged architecture described above.

8.2 Interpretability and Compliance:

Regulated industries (finance, healthcare, government) often require auditable justification for automated security decisions, particularly when a blocked message affects business operations or when compliance frameworks mandate explainability. Rule-based heuristics offer straightforward interpretability by design, while transformer-based systems require additional engineering — such as generating natural-language rationale alongside classification, or applying post-hoc explainability techniques (e.g., attention visualization, SHAP values on auxiliary structured features) — to meet similar transparency requirements.

8.3 Data Privacy:

Fine-tuning or deploying transformer models on sensitive corporate email content raises data privacy and confidentiality concerns, particularly when using third-party API-based LLMs that process content outside an organization's security perimeter. On-device or on-premises deployment of smaller, locally hosted transformer models has emerged as a common mitigation, trading some accuracy or capability for stronger data governance guarantees.

8.4 Continuous Retraining and Drift:

Phishing campaign characteristics evolve rapidly, and both classical ML and transformer-based models are subject to performance degradation (concept drift) as attacker tactics shift. Effective deployment requires a

continuous feedback and retraining pipeline, informed by analyst-reviewed false negatives and false positives, to maintain detection performance over time — an operational requirement common to both paradigms but particularly important for maintaining transformer model performance against evolving generative-AI-authored content.

9. The Emerging Generative AI Arms Race:

The dynamic between AI-generated phishing content and AI-based defenses represents one of the most consequential trends shaping this field. As generative LLMs lower the barrier to producing fluent, contextually tailored, and even personalized phishing content at scale, defenders face an increasingly symmetric technological contest: sophisticated attackers using generative AI to craft attacks are increasingly met by defenders using comparably sophisticated AI to detect them. This dynamic has several implications: first, it accelerates the obsolescence of purely lexical heuristics, since AI-generated content is specifically free of the surface-level tells those heuristics depend upon; second, it raises the stakes for adversarial robustness research, since attackers may eventually use generative models not only to author phishing content but to iteratively probe and evade specific detection systems; and third, it underscores the importance of multimodal and behavioral signals (sender history, login pattern anomalies, cross-referencing with authenticated communication channels) that remain valid regardless of how linguistically sophisticated the phishing text itself becomes.

10. Future Research Directions:

Multimodal fusion at scale: Continued development of architectures that jointly reason over text, visual webpage rendering, sender metadata, and behavioral signals (e.g., unusual login timing or geographic anomalies) is likely to yield further gains beyond text-only transformer classification.

Real-time adaptive learning: Research into online or continual learning methods that allow deployed models to adapt to emerging campaign patterns without requiring full retraining cycles would help address the concept-drift challenge discussed above.

Privacy-preserving and on-device detection:

Further work on efficient, privacy-preserving transformer architectures suitable for on-device or on-premises deployment would help address the data confidentiality concerns associated with cloud-based LLM inference on sensitive communications.

Standardized adversarial robustness

benchmarks: The field would benefit from standardized benchmarks specifically evaluating detection performance against AI-generated and adversarial perturbed phishing content, enabling more consistent comparison of robustness claims across studies.

Explain ability standards for regulated

deployment: Continued research into generating faithful, human-verifiable rationale for transformer-based phishing classifications — beyond post-hoc attention visualization — would support broader adoption in compliance-sensitive industries.

Cross-organizational threat intelligence

sharing: Given that both heuristic blocklists and transformer models benefit from timely exposure to emerging campaign data, research into privacy-preserving federated learning approaches that allow organizations to collaboratively improve shared detection models without exposing sensitive internal communications represents a promising direction.

11. Conclusion:

This paper has examined the comparative strengths and weaknesses of transformer-based language models and traditional heuristic approaches for phishing detection across email, URL, and SMS modalities. Traditional

heuristics block lists, rule engines, and classical machine learning on handcrafted features remain valuable for their speed, low computational cost, and interpretability, and continue to perform well against previously observed and lexically obvious phishing patterns. Transformer-based models, by contrast, demonstrate a clear and likely widening advantage in detecting novel, linguistically sophisticated phishing content, particularly attacks authored or refined using generative AI that lack the lexical anomalies heuristic systems depend upon. Neither paradigm alone offers a complete solution: heuristics lag behind novel attacker tactics, while transformer-based approaches introduce latency, cost, and interpretability challenges that complicate deployment at scale. The evidence reviewed here supports a layered, hybrid architecture that uses fast heuristic and classical ML filtering to handle the bulk of traffic, reserving transformer-based semantic analysis for ambiguous or high-stakes cases, supported by continuous retraining and human analyst review. As the generative AI arms race between attackers and defenders continues to intensify, the ongoing evolution of transformer-based detection supported by multimodal reasoning, adversarial robustness research, and privacy-preserving deployment models — is likely to become an increasingly central pillar of enterprise phishing defense.

References

1. Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2007). A Comparison of Machine Learning Techniques for Phishing Detection. *Proceedings of the Anti-Phishing Working Groups eCrime Researchers Summit*.
2. APWG (Anti-Phishing Working Group). Phishing Activity Trends Reports.
3. Certificate Transparency Project. Log Monitoring and Domain Verification Documentation.
4. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT*.
5. Google Safe Browsing. Transparency Report and Technical Documentation.
6. Le, H., Pham, Q., Sahoo, D., & Hoi, S. C. H. (2018). URLNet: Learning a URL Representation with Deep Learning for Malicious URL Detection. *arXiv preprint*.
7. Liu, Y., Ott, M., Goyal, N., et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint*.
8. MITRE ATT&CK Framework. Phishing (T1566) Technique Documentation.
9. PhishTank and OpenPhish. Community-Reported Phishing URL Databases.
10. Roy, S. S., Naragam, K. V., & Nilizadeh, S. (2023). Generative AI for Phishing: Evaluating LLM-Assisted Social Engineering Content. *arXiv preprint*.
11. Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine Learning Based Phishing Detection From URLs. *Expert Systems with Applications*.
12. UCI Machine Learning Repository. Phishing Websites Dataset.
13. Verizon. Data Breach Investigations Report (DBIR), Annual Editions

*Corresponding Author: Dr. Prashant Kohli

E-mail: prashant.kohli01@gmail.com

Received: 14 October,2025; Accepted: 26 November,2025. Available online: 30 November, 2025

Published by SAFE. (Society for Academic Facilitation and Extension)

This work is licensed under a Creative Commons Attribution-Noncommercial 4.0 International License