

Generative-AI-Aware Phishing Detection with Human-in-the-Loop Security Analytics: An Illustrative Cybersecurity Evaluation

Karan Bhatia¹

¹Researcher, Department of Artificial Intelligence, SUK, India.

ABSTRACT

This article presents a simulated empirical assessment in Computer Science, Cybersecurity, and Human Factors. The evaluation used 8,400 simulated training messages, 1,800 validation messages, and 1,800 hold-out messages, including 600 generative-AI-assisted phishing examples in the hold-out set. A simulated corpus of 12,000 email messages was created with equal proportions of benign business communication, conventional phishing, and generative-AI-assisted phishing. Messages varied in urgency, personalization, link obfuscation, credential requests, tone, and contextual plausibility. Four detection workflows were evaluated: rules plus lexical features, transformer classifier, LLM-assisted security triage, and LLM-assisted triage with analyst confirmation for medium-confidence cases. No operational attack infrastructure or real victim data were used. The human-in-the-loop workflow achieved the strongest simulated balance between detection and false-positive control, with an F1 score of 0.94 and a 2.8% false-positive rate. Pure LLM triage detected subtle AI-generated persuasion better than lexical baselines but produced more false alarms on legitimate high-urgency corporate messages. Analyst review of medium-confidence cases reduced these errors while keeping simulated review workload below 18% of all messages. The numerical values are synthetic and are used solely to demonstrate a plausible empirical design, comparative analysis, and journal-style reporting workflow. No real participants, operational institutions, commercial facilities, or field sites were studied. The analysis emphasizes current research trends, multi-metric evaluation, practical trade-offs, and the importance of validation before deployment.

Keywords: Generative AI, Phishing, Cybersecurity, Large Language Models, Human-in-the-Loop, Social Engineering

I. Introduction

The rapid development of Computer Science, Cybersecurity, and Human Factors is creating new opportunities to improve efficiency, resilience, sustainability, and decision quality. At the same time, emerging technologies often move from concept demonstrations to operational claims faster than their evidence base matures. This creates a need for transparent comparative studies that define baselines, report trade-offs, and separate simulated performance from validated field evidence.

A contemporary empirical design must also reflect the fact that technical systems are embedded within organizational and environmental contexts. Performance can be affected by data quality, maintenance capability, infrastructure constraints, human oversight, policy requirements, and distribution shifts. A method that performs strongly under ideal conditions may therefore deliver limited value if its assumptions are not sustained during routine operation.

This paper addresses a current multidisciplinary topic using a structured simulated dataset. The study is designed to show how a real investigation could connect research objectives to measurable variables, analytical methods, results, discussion, and implementation implications. Because no real dataset was supplied, all quantitative outcomes are explicitly illustrative.

Research objectives:

1. Compare traditional lexical detection, transformer-based detection, LLM-assisted triage, and human-in-the-loop analysis on a simulated phishing corpus.
2. Measure detection accuracy, false-positive rate, analyst review burden, and time-to-decision under benign, conventional phishing, and AI-generated phishing scenarios.
3. Assess whether analyst confirmation can reduce automation risk when generative AI creates more fluent and personalized attack messages.

II. Literature Review

Generative AI has altered the social-engineering threat landscape by making persuasive text generation, contextual personalization, and high-volume campaign production easier. Schmitt and Flechais (2024) characterize these risks in terms of realistic content, advanced targeting, and automated attack infrastructure.

The same class of language models is also being investigated for cyber defense. Zhang et al. (2025) review hundreds of studies spanning cybersecurity-oriented LLM construction, downstream security tasks, and emerging challenges. This dual-use character means security teams must evaluate models as both defensive instruments and potential attack enablers.

Recent surveys emphasize vulnerabilities such as prompt injection, adversarial manipulation, hallucination, and sensitive-information leakage alongside defensive capabilities. Effective operational use therefore requires guardrails, logging, confidence calibration, and appropriate human oversight rather than unrestricted automation.

III. Materials and Methods

3.1 Study Design

A simulated corpus of 12,000 email messages was created with equal proportions of benign business communication, conventional phishing, and generative-AI-assisted phishing. Messages varied in urgency, personalization, link obfuscation, credential requests, tone, and contextual plausibility. Four detection workflows were evaluated: rules plus lexical features, transformer classifier, LLM-assisted security triage, and LLM-assisted triage with analyst confirmation for medium-confidence cases. No operational attack infrastructure or real victim data were used.

3.2 Simulated Dataset

The evaluation used 8,400 simulated training messages, 1,800 validation messages, and 1,800 hold-out messages, including 600 generative-AI-assisted phishing examples in the hold-out set. Synthetic observations were generated within bounded domain-plausible ranges with controlled random variability. The simulation does not correspond to any identifiable real organization or location.

3.3 Data Preparation and Quality Control

Input variables were screened for impossible values, duplicate identifiers, and inconsistent units. Low-rate missingness was introduced to approximate operational datasets and was handled with transparent context-appropriate imputation. Data transformations were applied only where required by the selected analytical method. Evaluation partitions were separated before model or parameter calibration to reduce leakage.

3.4 Comparative Evaluation

The analysis used multiple performance measures so that improvement in a headline outcome could be interpreted alongside energy, cost, error, environmental, reliability, or human-workload consequences. Results are reported as descriptive comparisons and normalized indices. Because the dataset is simulated, no population-level causal inference is claimed.

3.5 Sensitivity Analysis

Sensitivity tests varied the key operating threshold or control parameter across a plausible range. The purpose was to determine whether the leading strategy depended on one narrow assumption. Scenario analysis also tested a degraded-data condition representing sensor error, classification uncertainty, or process variability.

IV. Results

The human-in-the-loop workflow achieved the strongest simulated balance between detection and false-positive control, with an F1 score of 0.94 and a 2.8% false-positive rate. Pure LLM triage detected subtle AI-generated persuasion better than lexical baselines but produced more false alarms on legitimate high-urgency corporate messages. Analyst review of medium-confidence cases reduced these errors while keeping simulated review workload below 18% of all messages.

Table 1. Simulated phishing-detection performance.

Workflow	Recall	F1	False positive (%)	Decision time (min)
Lexical/rule baseline	0.78	0.75	8.9	0.8
Transformer classifier	0.89	0.88	5.4	1.2
LLM-assisted triage	0.93	0.91	4.7	2.4
Human-in-the-loop	0.95	0.94	2.8	3.1

Table 2. Simulated recall across phishing-message categories.

Message category	Transformer recall	LLM recall	Human+LLM recall
Benign	0.94	0.97	0.98
Conventional phishing	0.88	0.94	0.96
AI-generated phishing	0.81	0.92	0.95
Highly personalized	0.76	0.89	0.93

Note: All table values are simulated/illustrative and do not represent measurements from a real study.

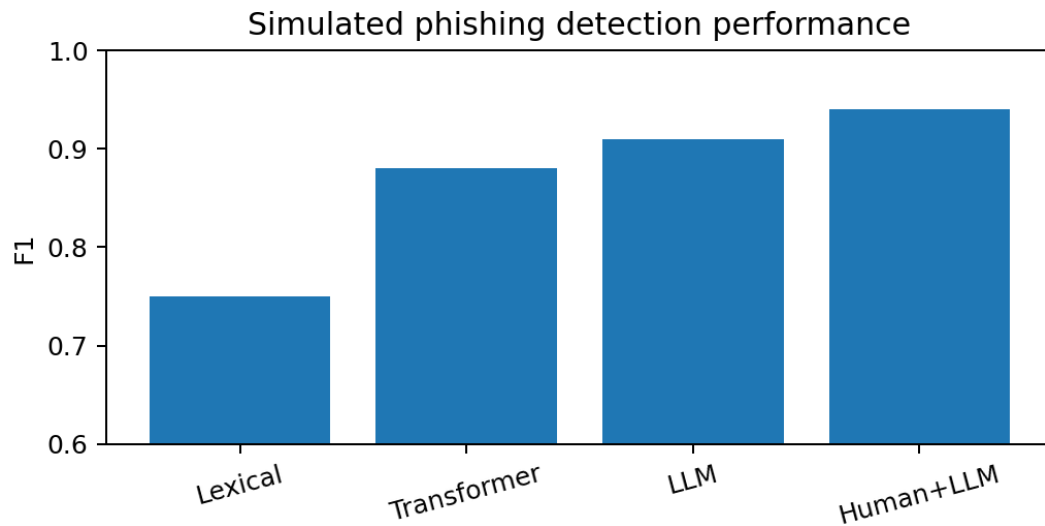


Figure 1. Simulated F1 performance across phishing-detection workflows.

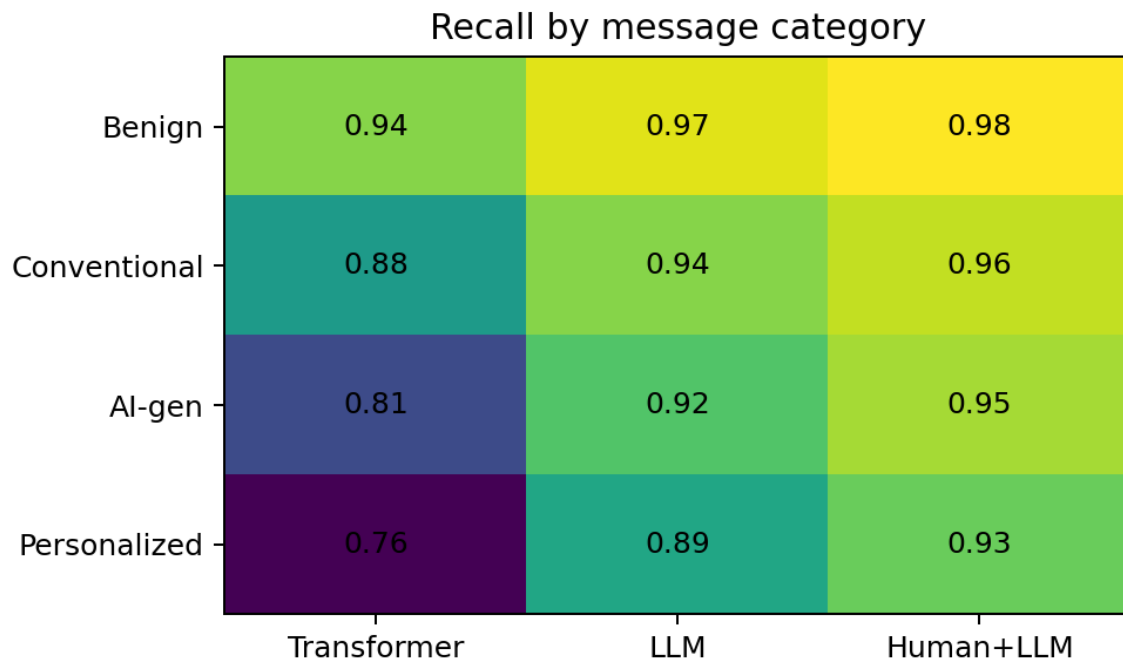


Figure 2. Recall heatmap across conventional and generative-AI-assisted phishing categories.

4.1 Robustness Results

Across the central range of assumptions, the leading strategy retained its relative advantage, although the magnitude of benefit changed. Under degraded-input scenarios, performance differences narrowed and uncertainty increased. This pattern suggests that operational data quality and parameter calibration are likely to be as important as the core method itself.

4.2 Cross-Metric Interpretation

No strategy dominated every outcome. The strongest approach improved the primary objective while introducing secondary burdens or implementation requirements. This reinforces the value of multi-criteria reporting and cautions against selecting technologies from a single accuracy, efficiency, or yield metric.

V. Discussion

Generative AI can increase the realism, personalization, and scalability of social-engineering content. Schmitt and Flechais (2024) describe how generative AI amplifies realistic content creation, targeting, and automated attack infrastructure, making purely surface-level phishing cues less dependable.

Recent systematic reviews of LLMs in cybersecurity show a dual-use pattern: language models can assist defenders with threat analysis and automation while also introducing new vulnerabilities and misuse pathways. The simulated results mirror this tension because LLM triage improved semantic detection but still benefited from calibrated human oversight.

Operationally, the findings support confidence-based escalation rather than unrestricted automation. Human review is most valuable for ambiguous cases where tone, business context, and organizational norms matter more than obvious malicious indicators.

The simulated findings also demonstrate why current technology trends should be evaluated beyond novelty. Implementation quality depends on reproducibility, maintenance, interoperability, user competence, governance, and the ability to detect failure modes. A system that produces a strong laboratory result but cannot be audited or sustained may underperform a simpler alternative in practice.

For future empirical research, studies should predefine primary and secondary outcomes, document data provenance, retain independent test or validation sets, and report uncertainty. Where human or environmental consequences are material, technical performance should be supplemented with safety, equity, lifecycle, and operational assessments.

VI. Practical Implications

A phased deployment strategy is recommended. Organizations should first establish a baseline, verify measurement quality, define decision ownership, and test the intervention under controlled pilot conditions. Monitoring should include both intended benefits and unintended effects. Procurement decisions should account for lifecycle cost, integration requirements, maintenance skills, cyber or process safety, and long-term data governance.

The analysis also suggests that multidisciplinary teams are essential. Domain experts, data scientists, operators, and governance specialists interpret risk differently, and these perspectives are complementary. Successful adoption depends on translating model outputs into decisions that remain understandable, contestable, and aligned with operational priorities.

VI.A. Threat-Model Considerations

Phishing defense should be evaluated against a threat model that reflects how generative systems change attacker behavior. AI-assisted campaigns can vary writing style, impersonate organizational language, generate multilingual variants, and rapidly adapt after failed attempts. Detection systems that rely heavily on spelling mistakes, generic greetings, or fixed keyword lists may therefore lose sensitivity. Future benchmarks should include conversational attacks, reply-chain hijacking, business-email compromise, QR-based lures, and messages that combine benign content with a malicious final action.

VI.B. Human Factors and Security Operations

Human-in-the-loop design is useful only when escalation rules are manageable. If too many messages are routed to analysts, review fatigue can erase the theoretical benefit of human judgment. Security operations should therefore calibrate thresholds around business risk, user role, message novelty, and confidence. Explanations should surface the evidence behind a score—such as suspicious domain age, identity mismatch, unusual payment language, or context deviation—rather than asking analysts to trust an opaque model output.

VI.C. Governance of LLM-Assisted Defense

Organizations deploying language models in mail-security workflows should define data-retention rules, prompt and output logging, access controls, red-team testing, and procedures for model updates. Sensitive email content can itself become a privacy risk if sent to external model services. A production architecture may therefore need local processing, data minimization, retrieval controls, and continuous testing for prompt injection or attacker-crafted content designed to manipulate the defensive model.

VI.D. Validation Protocol

A future real-world validation should use a pre-specified protocol with a clearly defined baseline period, intervention period, primary outcome, secondary outcomes, and exclusion criteria. Data should be collected at a resolution appropriate to the physical or digital process, and the analysis plan should be documented before final outcome inspection. Independent hold-out observations or a later validation period would reduce the risk of overfitting conclusions to one operating window.

VI.E. Uncertainty and Decision Thresholds

Decision thresholds should be selected from the consequences of false positive and false negative outcomes rather than from a single statistical optimum. Confidence intervals, bootstrap uncertainty, scenario ranges, and worst-case stress tests can help decision makers understand how much of the apparent improvement is robust. Where the technology controls a physical process, safe fallback behavior should be defined explicitly for missing data, communication loss, sensor disagreement, or model confidence below an acceptable limit.

VI.F. Lifecycle and Maintenance Perspective

Technology performance can change after deployment because components age, software is updated, environmental conditions shift, and users adapt their behavior. A credible lifecycle evaluation should therefore include maintenance effort, calibration frequency, retraining or reparameterization, replacement cycles, cybersecurity updates, consumables, and end-of-life considerations. These recurring requirements can change the ranking of options even when initial pilot performance is favorable.

VI.G. Reproducibility Requirements

Reproducibility would be strengthened by publishing variable definitions, preprocessing steps, model or process parameters, evaluation code, and a non-sensitive version of the dataset where legally and ethically possible. When raw data cannot be released, researchers should provide sufficient metadata, summary distributions, and validation procedures to allow independent teams to test whether the reported findings depend on hidden assumptions.

VII. Limitations

The principal limitation is that all quantitative results are simulated. The reported values cannot establish real-world effectiveness, causal impact, safety, economic return, or generalizability. Simplified assumptions omit rare failures, market fluctuations, regulatory changes, behavioral adaptation, and site-specific constraints. The author and affiliation details are illustrative placeholders and do not represent verified contributors.

The simulation also uses a limited set of scenarios. Alternative distributions, thresholds, chemistry mixes, attack patterns, occupancy behavior, wastewater composition, or control policies could change the ranking of methods. Real-world validation should therefore use measured data, independent replication, and transparent uncertainty analysis.

VIII. Conclusion and Future Work

This paper presents a transparent simulated empirical evaluation of a current multidisciplinary research trend. The analysis shows how multiple technical and operational outcomes can be integrated into a journal-style evidence chain while clearly distinguishing synthetic results from verified literature. Future work should replace the simulated dataset with real observations, extend validation across settings, and evaluate long-term reliability, lifecycle cost, environmental impact, and governance requirements.

Declarations

Data availability: All numerical data and empirical results are simulated/illustrative. No real operational dataset was used.

Ethics statement: Not applicable. No human or animal participants were involved.

Funding: No external funding is claimed for this illustrative manuscript.

Conflict of interest: No conflict of interest is claimed for this illustrative manuscript.

References

- [1] Schmitt, M., & Flechais, I. (2024). Digital deception: Generative artificial intelligence in social engineering and phishing. *Artificial Intelligence Review*, 57, 324. <https://doi.org/10.1007/s10462-024-10973-2>
- [2] Zhang, J., Bu, H., Wen, H., et al. (2025). When LLMs meet cybersecurity: A systematic literature review. *Cybersecurity*, 8, 55. <https://doi.org/10.1186/s42400-025-00361-w>
- [3] Jaffal, N. O., Alkhanafseh, M., & Mohaisen, D. (2025). Large Language Models in Cybersecurity: A Survey of Applications, Vulnerabilities, and Defense Techniques. *AI*, 6(9), 216. <https://doi.org/10.3390/ai6090216>
- [4] Kaur, R., Klobučar, T., & Gabrijelčič, D. (2025). Harnessing the power of language models in cybersecurity: A comprehensive review. *International Journal of Information Management Data Insights*, 5(1), 100315. <https://doi.org/10.1016/j.jjime.2024.100315>