International Journal of Engineering, Science and Environment
Peer-Reviewed Open Access • CC BY 4.0 • Quarterly
Vol. 1 | Issue 1 | Apr–Jun 2026
Banerjee et al. | Multimodal & LLM Approaches in Cybersecurity

SYSTEMATIC REVIEW

Multimodal and Large Language Model Approaches in Cybersecurity: A Systematic Review

Trina Banerjee[1], **Piyush**[2], **Mukthikka V**[3], **Gurpreet Singh**[4*]
[1] *SAKS Global, India* [2] *University of Delhi, India* [3] *Bharath Institute of Higher Education and Research, Chennai, India*
[4] *Endicott College of International Studies, Woosong University, South Korea*
***Corresponding Author:** gurpreetsinghmcse@gmail.com

ABSTRACT

The rapid evolution of cyber threats demands increasingly sophisticated defensive mechanisms. Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) have gained traction as valuable tools across multiple cybersecurity domains, offering capabilities beyond traditional rule-based and classical machine learning approaches. This systematic review provides a detailed analysis of 55 research papers published between 2019 and 2025, examining the application of LLMs and multimodal AI across eight key cybersecurity domains: vulnerability detection, malware analysis, phishing detection, network intrusion detection, cyber threat intelligence, security operations, penetration testing, and deepfake detection. We present a unified taxonomy categorising these approaches by architectural type (encoder-only, decoder-only, and multimodal) and application domains. Our comparative analysis shows that while LLMs demonstrate strong capabilities in code comprehension, threat classification, and automated security analysis, notable challenges persist in hallucination, adversarial robustness, and dual-use implications. We further examine security vulnerabilities in LLMs themselves, including prompt injection and jailbreaking attacks, and identify open research gaps proposing future directions including agentic AI workflows, privacy-preserving security models, and domain-specific foundation models for cybersecurity.

Keywords: *cybersecurity, large language models, multimodal learning, threat detection, vulnerability analysis, deep learning, transformers*

1. INTRODUCTION

The cybersecurity landscape has changed substantially in recent years, driven by the growing sophistication of cyber threats and the steady expansion of digital attack surfaces. Traditional security mechanisms, including signature-based detection systems, static rule engines, and conventional machine learning classifiers, are becoming insufficient to address the complexity and scale of modern cyber attacks [11, 14].

The rise of advanced persistent threats (APTs), zero-day vulnerabilities, AI-generated phishing campaigns, and polymorphic malware calls for a shift toward more intelligent, adaptive, and context-aware security solutions. Models such as GPT-4, Claude, LLaMA, and Gemini have demonstrated strong capabilities in natural language understanding, code comprehension, logical reasoning, and multimodal perception [32, 53].

The intersection of LLMs and cybersecurity has given rise to a rapidly growing body of research exploring both defensive and offensive applications. On the defensive side, LLMs have been applied to vulnerability detection in source code [16, 46], automated malware analysis [36], phishing detection through multimodal webpage analysis [25, 28], network intrusion detection [15], and security operations center (SOC) automation [3].

This systematic review addresses the existing gaps by providing: (1) a taxonomy of LLM and multimodal AI applications across eight cybersecurity domains; (2) a systematic analysis of 55 peer-reviewed papers published between 2019 and 2025; (3) a comparative evaluation of architectural approaches; (4) an examination of security vulnerabilities present in LLMs themselves; and (5) identification of open research gaps and future directions.

2. BACKGROUND AND PRELIMINARIES

A. Transformer Architectures

The transformer architecture, introduced by Vaswani et al. in 2017, has become the foundation for modern language models. Its self-attention mechanism enables the model to capture long-range dependencies in sequential data. Three primary architectural types have emerged from this foundation:

Encoder-only models, exemplified by BERT and its variants, are designed for bidirectional contextual understanding. Domain-specific models such as SecureBERT [2] and CyBERT [37] demonstrate improved performance on cybersecurity NER and threat classification.

Decoder-only models, represented by the GPT family, are autoregressive models optimised for text generation, useful for code repair [33], report generation [35], and interactive security analysis [10]. Encoder-decoder models, such as T5 and BART, combine both approaches and are particularly effective for sequence-to-sequence tasks.

B. Large Language Models in Context

The scaling of transformer models to billions of parameters has given rise to emergent capabilities relevant to cybersecurity, including in-context learning, chain-of-thought reasoning, and multi-step instruction following. The development of code-specialised LLMs — including

CodeBERT [13], StarCoder [26], and Code Llama [39] — has further advanced software security applications.

C. Multimodal Learning Fundamentals

Multimodal learning refers to the development of models that process and generate information across multiple data modalities. In cybersecurity, relevant modalities include natural language text, programming languages, visual data (screenshots, logos, CAPTCHAs), network data, and binary data. Recent advances in Vision-Language Models (VLMs) such as GPT-4V, Gemini, and LLaVA have enabled simultaneous processing of textual and visual information [25, 23].

D. Domain-Specific Security Models

SecureBERT [2] is a RoBERTa-based model pre-trained on a large corpus of cybersecurity text, demonstrating improved performance on NER, text classification, and question-answering. CyBERT [37] is fine-tuned for dense, technical cybersecurity language. CodeBERT [13] bridges the gap between natural language and programming languages.

3. METHODOLOGY

A. Search Strategy

This systematic review follows the PRISMA guidelines. We conducted a literature search across IEEE Xplore, ACM Digital Library, Springer, arXiv, USENIX, and Google Scholar, using combinations of cybersecurity and LLM-specific keywords.

B. Use of Large Language Models

In accordance with editorial policies, we disclose that Large Language Models were used as assistive tools during the literature search, synthesis, and manuscript preparation stages. All content was independently verified and critically reviewed by the authors.

C. Inclusion and Exclusion Criteria

Inclusion criteria: (1) papers published between 2019 and 2025; (2) applying LLMs, transformer-based models, or multimodal approaches to cybersecurity; (3) published in peer-reviewed venues or reputable preprint servers; (4) presenting empirical results, novel architectures, or surveys.

Exclusion criteria: papers unrelated to cybersecurity, using only classical ML without transformers, duplicate publications, papers without accessible full text, and non-English publications.

D. Paper Selection Process

The selection process proceeded in three phases: (1) initial keyword search yielding 312 candidate papers; (2) title and abstract screening reducing the pool to 98 papers; and (3) full-text review resulting in the final selection of 55 papers.

E. Categorisation Taxonomy

Selected papers were categorised along two dimensions: (1) the cybersecurity application domain, and (2) the model architecture (encoder-only, decoder-only, encoder-decoder, or multimodal). This enables cross-domain comparison and identification of architectural trends.

4. TAXONOMY OF APPLICATIONS

A. Vulnerability Detection and Code Analysis

Vulnerability detection is one of the most extensively studied applications of LLMs in cybersecurity. Thapa et al. [46] evaluated multiple transformer architectures, with RoBERTa achieving F1 of 92.1% on the Draper VDISC dataset. Fu and Tantithamthavorn [16] introduced LineVul, a transformer-based line-level approach using RoBERTa's attention mechanism. Steenhoek et al. [45] found that GPT-4 achieved 74.5% accuracy on DiverseVul.

Beyond detection, LLMs have shown promise in automated vulnerability repair. Pearce et al. [33] found that LLMs can generate correct patches for approximately 67% of synthetic vulnerability scenarios, though performance drops considerably on real-world CVEs.

B. Malware Analysis and Binary Reverse Engineering

Rahali and Akhloufi [36] introduced MalBERT, achieving detection accuracy exceeding 97% by treating malware detection as an NLP task. Xu et al. [50] introduced LLM4Decompile, which outperforms traditional decompilers such as Ghidra and IDA Pro in producing semantically accurate output. Pei et al. [34] demonstrated that GPT-4 can generate accurate YARA signatures with 78% precision.

C. Phishing and Social Engineering Detection

Modern phishing attacks are inherently multimodal. Lee et al. [25] proposed a two-phase approach using multimodal LLMs, achieving 98.6% detection accuracy. Li et al. [28] introduced KnowPhish, integrating multimodal knowledge graphs with LLMs for zero-shot detection. Koide et al. [24] found ChatGPT achieves 92.3% accuracy in identifying phishing URLs.

D. Network Intrusion Detection

Ferrag et al. [15] proposed a privacy-preserving BERT-based model for IoT/IIoT network security, achieving F1-scores exceeding 95%. Lin et al. [29] introduced MIND-IoT, a hybrid transformer-CNN architecture achieving 98.14% accuracy. Liu et al. [30] developed an LLM-based framework producing interpretable detection decisions in natural language.

E. Cyber Threat Intelligence

SecureBERT [2] and CyBERT [37] are notable contributions for CTI extraction from unstructured sources. The integration of Retrieval-Augmented Generation (RAG) with LLMs has emerged as an effective approach: Li et al. [27] introduced TechniqueRAG for adversarial technique annotation, while Abdeen et al. [1] developed RAGIntel, substantially reducing hallucination compared to standalone LLM approaches.

F. Security Operations and Incident Response

Alam et al. [3] explored multimodal LLMs for SOC operations. Sahin et al. [40] found that LLMs notably accelerate the initial stages of investigation but require careful human oversight during remediation. Chuvakin et al. [7] noted that approximately 15% of LLM-generated SOC reports contain factual errors.

G. Penetration Testing and Offensive Security

Deng et al. [10] introduced PentestGPT, guiding penetration testers through reconnaissance, vulnerability identification, and exploitation. Fang et al. [12] demonstrated that GPT-4 could autonomously exploit 87% of tested one-day vulnerabilities when provided with CVE descriptions — a finding with serious implications for automated attacks.

H. Deepfake Detection and Visual Security

Jia et al. [23] found that GPT-4V achieves zero-shot deepfake detection accuracy of 73.2%, improving to 84.7% with targeted prompting. Shi et al. [43] introduced SHIELD, revealing that even advanced MLLMs achieve only 61% robustness against state-of-the-art visual attacks.

5. COMPARATIVE ANALYSIS

A. Summary of Reviewed Papers

Table 1 presents a summary of representative works across all eight cybersecurity domains, categorised by model architecture, task type, and key performance metrics.

Table 1. *Summary of Representative LLM and Multimodal Approaches Across Cybersecurity Domains*

Reference	Domain	Architecture
Thapa et al. (2022)	Vuln. Detection	Encoder
Fu & Tantithamthavorn (2022)	Vuln. Detection	Encoder
Steenhoek et al. (2024)	Vuln. Detection	Decoder
Pearce et al. (2023)	Vuln. Repair	Decoder
Rahali & Akhloufi (2023)	Malware	Encoder
Xu et al. (2024)	Binary RE	Decoder
Lee et al. (2024)	Phishing	Multimodal
Li et al. (2024)	Phishing	Multimodal
Koide et al. (2024)	Phishing	Decoder
Ferrag et al. (2024)	Network IDS	Encoder
Alkhatib et al. (2022)	Network IDS	Encoder
Lin et al. (2024)	Network IDS	Hybrid
Aghaei et al. (2023)	CTI	Encoder
Abdeen et al. (2024)	CTI	RAG+LLM
Deng et al. (2024)	Pen Testing	Decoder
Fang et al. (2024)	Pen Testing	Decoder
Jia et al. (2024)	Deepfake	Multimodal

B. Architectural Trends

Analysis of the reviewed literature reveals several architectural trends. Encoder-only models are predominant in classification-oriented tasks such as malware detection, vulnerability classification, and network intrusion detection. Decoder-only models are preferred for generative tasks. A notable trend is the rise of hybrid architectures and RAG-based systems [1, 27].

C. Cross-Domain Observations

Domain adaptation is critical: SecureBERT [2] outperforms base BERT by 8–12 percentage points on CTI tasks. Scale does not guarantee superiority: fine-tuned RoBERTa [16] achieves higher vulnerability detection accuracy than zero-shot GPT-4 [45]. Multimodal approaches show the highest improvement margins, particularly in phishing detection [25, 28].

6. SECURITY OF LARGE LANGUAGE MODELS

A. Prompt Injection Attacks

Prompt injection is the most widely studied vulnerability class in LLM-integrated systems. Greshake et al. [18] showed that adversarial instructions embedded in external data sources can hijack the LLM's behaviour. In cybersecurity contexts, this is especially concerning for SOC tools that use LLMs to process security logs.

B. Jailbreaking Attacks

Wei et al. [47] provided a systematic analysis of how LLM safety training fails, identifying two fundamental failure modes: competing objectives and mismatched generalisation. Zou et al. [55] introduced the Greedy Coordinate Gradient (GCG) attack, generating universal and transferable adversarial suffixes that can jailbreak multiple aligned LLMs.

C. Implications for Cybersecurity Deployment

SOC automation systems processing untrusted input must be hardened against indirect prompt injection. RAG-based systems must validate retrieved content to prevent knowledge base poisoning. LLM-generated security patches must undergo independent verification. Multi-layer defence combining input sanitisation, output filtering, and human oversight is essential.

7. DISCUSSION

A. Key Findings

This systematic review reveals that LLMs have been successfully applied to every major cybersecurity domain. The dual-use nature of LLMs is a defining characteristic: the same capabilities enabling automated vulnerability detection also enable automated exploit development. Multimodal approaches represent an active frontier with the most substantial performance improvements.

B. Limitations of Current Approaches

Hallucination remains a pervasive concern in generative applications. Context window limitations constrain analysis of large codebases. Adversarial robustness of LLM-based security tools remains largely untested. Evaluation standardisation is lacking, making direct comparison of approaches difficult across the field.

C. Ethical Considerations

The deployment of LLMs in cybersecurity raises important ethical questions. The automation of offensive security capabilities lowers the barrier to entry for malicious actors. Responsible research practices, including careful

consideration of dual-use implications and appropriate access controls, are essential.

8. FUTURE DIRECTIONS

Agentic AI for Cybersecurity: The shift toward autonomous AI agents capable of performing multi-step security tasks represents an important research direction [10, 12]. Domain-Specific Foundation Models: The success of SecureBERT [2] suggests that larger, comprehensive security foundation models could be highly effective. Privacy-Preserving Security AI: Federated learning, differential privacy, and on-premise deployment strategies are required for sensitive security data. Standardised Evaluation Frameworks would enable consistent cross-study comparison including adversarial settings and temporal shifts.

9. CONCLUSIONS

This systematic review provided a detailed analysis of 55 research papers examining the application of LLMs and multimodal AI across eight cybersecurity domains. Our analysis shows that LLMs and multimodal models are substantially changing cybersecurity practice, offering strong capabilities in code comprehension, threat detection, automated analysis, and decision support.

The key contributions include: (1) a unified taxonomy spanning eight cybersecurity domains; (2) a comparative analysis showing that domain-specific and multimodal approaches consistently outperform general-purpose models; (3) an examination of security vulnerabilities present in LLMs themselves; and (4) identification of future research directions. The path forward must balance innovation with responsibility, ensuring these technologies are deployed safely, ethically, and with appropriate human oversight.

DECLARATIONS

Competing Interests

The authors declare no competing interests.

Data Availability

Not applicable. This systematic review is based on previously published research papers that are publicly available through the cited sources.

Author Contributions

Trina Banerjee, Piyush, and Mukthikka V contributed equally to conceptualisation, literature review, data analysis, and writing. Gurpreet Singh supervised the study, critically reviewed the manuscript for intellectual content, and approved the final version for submission.

REFERENCES

[1] Abdeen, B., Lakhani, A., Al-Shaer, E., & Aghaei, E. RAGIntel: RAG-based LLM system for cyber attack investigation. PeerJ Computer Science, vol. 10, e2517, 2024.

[2] Aghaei, E., Niu, X., Shadid, W., & Al-Shaer, E. SecureBERT: A domain-specific language model for cybersecurity. In Proc. SecureComm, pp. 257-275, 2023.

[3] Alam, M., Dey, N., Huang, Y., & Bozorgi, A. Looking beyond text: Reducing visual parroting with multimodal LLMs for SOC. arXiv Preprint, 2024.

[4] Alkhatib, B., Rass, S., & Zhauniarovich, Y. Can BERT understand network traffic? IEEE Trans. Network Service Mgmt., vol. 19, no. 4, pp. 4946-4957, 2022.

[5] Bethany, M., et al. Large language models for automated social engineering attacks and defense. arXiv Preprint, 2024.

[6] Chen, Y., et al. DiverseVul: A new vulnerable source code dataset. In Proc. RAID, pp. 654-668, 2023.

[7] Chuvakin, A., Simorjay, F., & Wei, Y. LLMs for security operations centers: A comprehensive survey. arXiv Preprint, 2025.

[8] Coccomini, D. A., et al. Combining EfficientNet and vision transformers for video deepfake detection. In Proc. ICIAP, pp. 219-229, 2022.

[9] Demirkiran, F., et al. An ensemble of pre-trained transformer models for malware classification. Computers & Security, vol. 121, 102846, 2022.

[10] Deng, G., et al. PentestGPT: An LLM-empowered automatic penetration testing tool. In Proc. USENIX Security Symposium, 2024.

[11] Divakaran, D. M., & Gupta, S. Large language models in cybersecurity: A survey. IEEE Access, vol. 12, pp. 179576-179609, 2024.

[12] Fang, R., et al. LLM agents can autonomously exploit one-day vulnerabilities. arXiv:2404.08144, 2024.

[13] Feng, Z., et al. CodeBERT: A pre-trained model for programming and natural languages. In Findings EMNLP, pp. 1536-1547, 2020.

[14] Ferrag, M. A., et al. Generative AI and large language models for cyber security. arXiv:2405.12750, 2024.

[15] Ferrag, M. A., et al. Revolutionizing cyber threat detection with LLMs: Privacy-preserving BERT-based model for IoT/IIoT. arXiv:2306.14263, 2024.

[16] Fu, M., & Tantithamthavorn, C. LineVul: A transformer-based line-level vulnerability prediction. In Proc. MSR, pp. 608-620, 2022.

[17] Goodman, K., et al. A transformer-based framework for payload maliciousness detection. In Proc. NDSS, 2024.

[18] Greshake, K., et al. Compromising real-world LLM-integrated applications with indirect prompt injection. In Proc. AISec, pp. 79-90, 2023.

[19] Hao, S., et al. Halligan: VLM agent for solving unseen visual CAPTCHAs. In Proc. USENIX Security Symposium, 2024.

[20] Happe, A., & Cito, J. Getting pwn'd by AI: Penetration testing with LLMs. In Proc. ESEC/FSE, pp. 2082-2086, 2023.

[21] Heijden, A. van der, & Allod, L. Cognitive triaging of phishing attacks. In Proc. USENIX Security Symposium, 2019.

[22] Hou, X., et al. Large language models for software vulnerability detection: A survey. arXiv:2403.08345, 2024.

[23] Jia, S., et al. Can LLMs and VLMs detect deepfakes? arXiv Preprint, 2024.

[24] Koide, T., et al. Detecting phishing sites using ChatGPT. arXiv:2306.05816, 2024.

[25] Lee, J., et al. Multimodal large language models for phishing webpage detection. In Proc. eCrime, 2024.

[26] Li, R., et al. StarCoder: May the source be with you! arXiv:2305.06161, 2023.

[27] Li, S., Wen, Z., & Kang, D. TechniqueRAG: Adversarial technique annotation with RAG. In Findings ACL, 2025.

[28] Li, Y., et al. KnowPhish: LLMs meet multimodal knowledge graphs. In Proc. USENIX Security Symposium, 2024.

[29] Lin, Z., Li, J., & Ren, Y. MIND-IoT: Multimodal IoT network traffic classification. PeerJ Computer Science, vol. 10, e2326, 2024.

[30] Liu, X., Yang, Y., & He, Z. Large language model for network intrusion detection. arXiv Preprint, 2024.

[31] Lu, G., et al. Grace: Empowering LLM-based software vulnerability detection. arXiv:2411.03592, 2024.

[32] Motlagh, F. N., et al. Large language models in cybersecurity: State-of-the-art. arXiv:2402.00891, 2024.

[33] Pearce, H., et al. Examining zero-shot vulnerability repair with LLMs. In Proc. IEEE S&P, pp. 2339-2356, 2023.

[34] Pei, K., et al. Exploiting large language models for malware analysis. arXiv Preprint, 2024.

[35] Perrina, F., Siracusano, G., & Zanero, S. AGIR: Automating cyber threat intelligence reporting. In Proc. ARES, 2023.

[36] Rahali, A., & Akhloufi, M. A. MalBERT: Malware detection using transformers. IEEE Access, vol. 11, pp. 88495-88511, 2023.

[37] Ranade, P., et al. CyBERT: Contextualized embeddings for the cybersecurity domain. In Proc. IEEE BigData, pp. 3334-3342, 2021.

[38] Roy, S. S., Nath, S., & Sisodia, D. ChatBots to PhishBots? Preventing phishing attacks using LLMs. DMNLP@AAAI, 2024.

[39] Roziere, B., et al. Code Llama: Open foundation models for code. arXiv:2308.12950, 2024.

[40] Sahin, M., et al. Integrating LLMs into security incident response. In Proc. USENIX Security Symposium, 2025.

[41] Satvat, K., Gjomemo, R., & Venkatakrishnan, V. N. EXTRACTOR: Extracting attack behavior from threat reports. In Proc. IEEE EuroS&P, pp. 598-615, 2021.

[42] Schulhoff, S., et al. Ignore this title and HackAPrompt: Exposing systemic weaknesses of LLMs. In Proc. EMNLP, 2023.

[43] Shi, Y., et al. SHIELD: An evaluation benchmark for face spoofing with MLLMs. In Proc. ECCV, 2024.

[44] Siracusano, G., Ferroni, A., & Zanero, S. Enhancing SOC efficiency through multi-model integration. IEEE Trans. Inf. Forensics Security, 2024.

[45] Steenhoek, B., et al. A comprehensive study of LLMs for vulnerability detection. arXiv:2403.17218, 2024.

[46] Thapa, C., et al. Transformer-based language models for software vulnerability detection. In Proc. ACSAC, pp. 481-496, 2022.

[47] Wei, A., Haghtalab, N., & Steinhardt, J. Jailbroken: How does LLM safety training fail? In NeurIPS, vol. 36, 2024.

[48] Wu, Y., et al. How effective are neural networks for fixing security vulnerabilities? In Proc. ISSTA, pp. 1282-1294, 2023.

[49] Xu, H., et al. Large language models for cyber security: A systematic literature review. arXiv:2405.04760, 2024.

[50] Xu, H., et al. LLM4Decompile: Decompiling binary code with LLMs. In Proc. EMNLP, 2024.

[51] Yang, J., et al. AutoAttacker: A LLM guided system for automatic cyber-attacks. arXiv Preprint, 2024.

[52] Yao, Y., et al. A survey on LLM security and privacy. High-Confidence Computing, vol. 4, no. 2, 100211, 2024.

[53] Zhang, G., et al. A survey of large language models in cybersecurity. arXiv:2402.01854, 2024.

[54] Zhou, X., Zhang, T., & Lo, D. LLM for vulnerability detection: Emerging results. In Proc. ICSE-NIER, pp. 47-51, 2024.

[55] Zou, A., et al. Universal and transferable adversarial attacks on aligned language models. arXiv:2307.15043, 2023.