Explainable AI and Large Language Models for Cybersecurity: A Narrative Review of Detection, Explanation, and Evidence-Grounded Interpretation

Vishakha R. Chougale[1], Janhavi P. Sathe[2], Aaditi T. Patil[3], Shamali V. Chavan[4]
[1,2,3,4]*Department of Artificial Intelligence & Machine Learning, New Institute of Technology, Kolhapur, Maharashtra, India*

***Abstract*— Machine learning (ML) underpins intrusion detection, malware analysis and alert triage, but its opacity limits analyst trust and error diagnosis. Explainable AI (XAI) methods such as SHAP and LIME attribute predictions to features, while large language models (LLMs) can render technical evidence as readable narratives. This review synthesises eight recent surveys and primary studies on XAI for cybersecurity, LLMs for cybersecurity, and their intersection. It separates four functions the literature often conflates: detection, explanation, narrative, and operational prioritisation. Three findings recur. XAI research concentrates on IoT intrusion detection with post-hoc attribution and is rarely validated with analysts. LLMs perform near chance as flow classifiers and are about 7,000 times slower than lightweight models, yet with verified context they can explain alerts; even the best measured configuration was fully correct for only 80% of explanations. XAI pipelines also enlarge the attack surface, and defences are scarce. We propose an evidence-grounded framework that evaluates each layer separately. This is a narrative review, not a systematic review or a new experiment.**

***Index Terms*— Explainable AI, Large Language Models, Intrusion Detection, SHAP, LIME, Hallucination, Alert Triage**

I. INTRODUCTION

Security operations centres (SOCs) face tens of thousands of alerts a day, most of them false positive [2] ML detectors find patterns that fixed rules miss, but a bare label gives analysts little to verify or challenge. Yan et al. [3] argue that explanations let developers validate model behaviour and diagnose misclassifications, and that users trust security systems more when predictions come with explanations.

LLMs add a different capability: turning structured telemetry into prose. But fluency is not accuracy. Da Silva and Westphall [6] name hallucination and context loss as central unresolved problems, and Xu et al. [5] list interpretability and trustworthiness as open challenges.

This review asks four questions.

RQ1. How is XAI used and evaluated in cybersecurity detection?

RQ2. What can LLMs do for detection and explanation, and where do they fail?

RQ3. What new risks arise when XAI outputs meet LLM narratives?

RQ4. What must be measured to call an integrated workflow trustworthy?

II. SCOPE AND METHOD

This is a focused narrative review of eight sources: three XAI surveys or reviews [1]–[3], one XAI-based IDS framework [4], two LLM surveys [5],[6] and two primary studies on LLM-generated explanations of network alerts [7], [8]. The source papers cover 53 studies (2018–2025) [1], 50 papers [2], and 185 papers selected from over 40,000 collected [5]. No database search, deduplication or risk-of-bias scoring was performed, so coverage claims are limited to these sources. Results from different studies use different datasets and metrics and are reported side by side, never pooled.

III. BACKGROUND

An intrusion detection system (IDS) labels activity as benign or as an attack class, and is judged by precision, recall, F1, and false-positive rate. Prasad et al. [1]

stress that missed attacks usually cost far more than false alarms, so recall and F1 matter more than accuracy.

Charmet et al. [2] separate intrinsically explainable models (linear and logistic regression, decision trees, random forests, gradient boosting) from methods that explain opaque models. The latter include local methods (LIME, Anchors, counterfactuals, LEMNA, ICE), global methods (PDP, ALE, global surrogates), and gradient-based techniques for neural networks. SHAP averages a feature's marginal contribution across feature coalitions and serves both local and global explanation. Yan et al. [3] organize the field into ante-hoc explanations, built into the model, and post-hoc explanations generated afterwards. Post-hoc SHAP and LIME dominate the IDS literature [1].

LLMs are transformer models pre-trained on large text corpora and adapted by fine-tuning, prompting, retrieval or tool use [5], [6]. They are applied to vulnerability detection, malware analysis and network intrusion detection, with a growing trend toward autonomous agents [5].

IV. XAI FOR CYBERSECURITY (RQ1)

Applications. Charmet et al. [2] group applied work by purpose: transparency and trust, improving classifier performance, and explaining errors. Examples include SHAP explanations of IDS decisions, autoencoder anomaly explanation, decision-tree surrogates for neural IDS, and attribution analysis to understand why malware detectors change after updates. LEMNA was designed for security, handling feature dependency in malware detectors. Prasad et al. [1] report that most of their 53 studies target IoT and related settings (IIoT, IoMT) and anomaly detection, with cloud-native infrastructure, threat intelligence and adaptive defence underexplored. They attribute this to heterogeneous devices, public IoT benchmarks (N-BaIoT, BoT-IoT, ToN-IoT) and real-time pressure.

Efficiency. Post-hoc explanation can be too slow for operations. Nugraha et al. [4] combine an ANOVA selector with global SHAP scores to keep ten features (about 70% reduction), and retrain a lightweight XGBoost detector. They explain it with SHAP and LIME, cross-validating the two. On CIC-DDoS2019, CICIoT2023 and 5G PFCP traces they report F1 of at least 99% while LIME time fell from 36 s to 4.9 s (87%). Their ablation shows ANOVA-only selection slightly lowered F1 on Portmap traffic (0.9981 versus 0.9994), suggesting statistical filtering alone can miss interaction effects. These attack classes are highly separable (SYN Flood F1 near 0.9999), so transfer to subtler attacks should not be assumed.

Evaluation gaps. Three gaps recur across [1]–[4].

1. No standard for explanation quality. Accuracy and F1 measure prediction, not explanation; faithfulness, consistency, complexity and sensitivity are proposed but not standardized [1]. Explanation evaluation is a thin strand, only three works in the taxonomy of [2].
2. Little human evaluation. Only two of the 53 studies in [1] assessed how analysts perceived or used explanations.
3. Old, static benchmarks. Charmet et al. [2] note a study that used a 2009 dataset, which raises questions about its relevance to current threats.

Attribution is also not causation. Correlated features undermine PDP and motivate ALE, and LIME's instability across repeated runs motivated Bayesian variants [2]. An explanation can be unstable and still look authoritative.

V. LLMs FOR DETECTION AND EXPLANATION (RQ2)

Broad surveys. The surveys [5], [6] show LLMs being used for vulnerability detection and repair, penetration-testing assistance, phishing detection, policy writing and log analysis. Recurring problems are context loss in long interactions, hallucinated actions, safety filters interfering with legitimate use, and dependence on prompt quality [6]. Mitigations include fine-tuning (risking catastrophic forgetting), in-context learning (consuming context window) , retrieval-augmented generation (retrieved text may be related but not relevant). Chain-of-verification (reduces but does not remove hallucination, at higher cost) [6]. The mixture-of-experts design proposed in [6] is not empirically validated.

LLMs as classifiers. Houssel et al. [7] fed NetFlow records, encoded as text, to Llama3-8B-Instruct and GPT-4 on NF-UNSW-NB15-v2 and NF-CSE-CIC-IDS2018-v2. Zero-shot macro precision and recall were roughly 40% to 53%, around or below chance. Fine-tuning Llama3 with KTO or ORPO on up to 50,000 samples helped only slightly (best about 55%

precision and 59% recall), far below conventional classifiers. Inference took about 14,000 µs per flow versus about 2 µs for a random forest, a roughly 7,000-fold gap. The authors conclude that LLMs are impractical as stand-alone NIDS and are better used as complementary tools.

LLMs as explainers. In the same study, manual inspection of Llama3 explanations found them consistent with the flow's features and able to name traffic types such as DNS. There were factual errors: wrong IP geolocation, misidentified protocol numbers, and port 0 treated as inherently malicious. The model treated suspicious features as independent clues rather than reasoning over combinations [7]. This analysis was qualitative.

The follow-up eX-NIDS study [8] quantifies the problem and a mitigation. A baseline explainer gives the LLM only the flagged flow and an instruction. eX-NIDS adds a Prompt Augmenter that supplies feature definitions, protocol-ID-to-name mappings, and IP context (threat intelligence, geolocation, earlier related flows) from a conventional database rather than RAG. Two domain experts scored explanations of 50 malicious flows from NF-CSE-CIC-IDS2018-v2.

Table 1. Explanation quality, baseline vs eX-NIDS [8] (% of 50 explanations)

Model	Method	Correctness	Feature consists.	Factual consist.	Avg.
Llama3-70B-Instruct	Basic prompt	26	84	42	50.7
Llama3-70B-Instruct	eX-NIDS	36	100	90	75.3
GPT-4	Basic prompt	40	96	78	71.3
GPT-4	eX-NIDS	80	100	92	90.7

Context raised every metric for both models, but correctness stayed weakest: at best, one explanation in five contained an incorrect argument. The design has clear limits: 50 flows, one dataset, two raters, malicious flows only, and no analyst-outcome measure. Local Llama3 explanation averaged about 4 s (baseline) and 6 s (augmented), and estimated API cost rose from $5.75 to $10.37 per 1,000 queries because augmented prompts were roughly five times longer [8].

Together, [7] and [8] support a division of labour: a conventional detector decides, and an LLM with verified context explains. They do not show that LLM explanations improve analyst decisions.

VI. SECURITY AND TRUST RISKS (RQ3)

Charmet et al. [2] devote half their survey to the security of XAI: 27 papers across fairness, integrity, privacy, confidentiality and robustness. The findings most relevant to explanation pipelines are these.

- Manipulation. Attacks on saliency maps, LIME, SHAP and counterfactuals can change an explanation while leaving the prediction unchanged. Where the model owner controls explanation generation, an auditor cannot rely on it.
- Leakage. Explanations have been used to reconstruct private images, infer training-set membership and extract models with fewer queries; richer explanations leak more, creating a privacy-explainability trade-off.
- Fragility. Small perturbations can alter interpretations without changing classifications, and some saliency methods are insensitive to the model or data, so visual plausibility does not validate an explanation.
- Few defences. Only 6 of the 27 papers describe countermeasures, mostly for image or text models rather than security telemetry.

LLMs add hallucination and context degradation [5]–[8], data-governance risk when logs go to hosted models, and cost tied to prompt length [8]. The reviewed sources treat prompt injection and leakage only generally. Evidence for attacks on LLM alert explanations is thin; we flag this as a gap, not a finding. Charmet et al. [2] conclude that XAI adds to the attack surface and that method choice should weigh explanatory quality against confidentiality. They also note that external data such as honeypot results and threat intelligence can reinforce explanations, which is what eX-NIDS operationalises [8].

VII. SYNTHESIS

The literature often treats a detector's output, an attribution and a narrative as one thing. We separate four layers, each with its own failure mode.

1. Detection: missed attacks, false alarms, benchmark accuracy that does not transfer [1], [2], [7].
2. Explanation: unstable, unfaithful or manipulable attributions; computational cost [1], [2], [4].
3. Narrative: hallucinated facts, misread feature semantics, fluent unsupported inference [6]–[8].
4. Operational: explanations that are correct but unusable, or that invite over-trust [1].

Each transition can lose evidence or uncertainty. Supplying definitions and context reduces narrative failures [8]; feature selection cuts explanation cost without hurting detection [4]; the operational layer is the least validated [1].

Table 2. What the evidence supports

Claim	Support	Confidence
LLMs are unsuitable as stand-alone flow classifiers	[7]: Near-chance zero-shot, minor fine-tuning gains, ~7,000× slower	Moderate: two models, two datasets, small test sets
Context-enriched prompts improve LLM explanation accuracy	[8]: All metrics improved for both models	Moderate: 50 flows, manual scoring
Feature selection makes post-hoc explanation near-real-time	[4]: F1 ≥ 99%, LIME 36 s → 4.9 s	Moderate: separable attacks
Security XAI lacks human-centred validation	[1]: Two of 53 studies	Good for that corpus
XAI pipelines are attackable and under-defended	[2]: 27 papers, 6 with countermeasures	Good, but mostly non-security data
LLM explanations improve analyst accuracy or speed	None found	No evidence

VIII. PROPOSED EVALUATION FRAMEWORK (RQ4)

We propose, as a conceptual framework requiring empirical testing, a pipeline in which every LLM statement must trace to a verifiable source.

1. Detect: The model outputs label, score, timestamp and version.
2. Explain: SHAP, LIME or a surrogate produces an attribution; stability and faithfulness checks run before it is passed on.
3. Package evidence: Alert, log excerpts, feature values and definitions, explanation and threat context are bundled into a structured schema, following [8].
4. Narrate: The LLM summarises only the package, separating facts from inference and stating what is missing.
5. Validate: Deterministic checks verify cited values and identifiers; unsupported claims are flagged or removed.
6. Decide: An analyst reviews; response actions remain under existing approval policy.
7. Monitor: Drift, explanation stability, latency, hallucination rate and analyst outcomes are tracked.

Table 3. Dimensions to Measure separately

Dimension	Example measures	Basis
Detection quality	Precision, recall, F1, false-positive rate	Standard [1], [7]
Explanation faithfulness and stability	Deletion/insertion tests, top-feature overlap under perturbation	[1], [2]; SHAP–LIME cross-check in [4]
Narrative correctness and grounding	Share of explanations with all arguments valid; feature and factual consistency	Metrics of [8]
Efficiency and cost	Per-alert latency, compute, API cost	Measured in [4], [7], [8]
Privacy and robustness	Leakage, adversarial-input and injection tests	Risks in [2]; LLM-specific tests are a gap
Operational value	Time to triage, triage accuracy, calibration, missed-attack cost	Called for in [1]; unmeasured for LLM narratives

IX. RESEARCH GAPS

1. Analyst-outcome evidence. Both [1] and [8] call for user studies. A three-arm experiment would compare (A) alert only, (B) alert plus SHAP/LIME or graph explanation, (C) alert plus explanation plus evidence-grounded LLM summary, measuring triage time, accuracy, calibration and unsupported-claim rate, with datasets, models, prompts and analysis pre-registered.

2. Larger explanation benchmarks: more flows,benign and false-positive cases, subtler attacks, inter-rater agreement and cross-dataset tests.
3. Standard explanation metrics, requested in [1], [2], [8].
4. Realistic domains beyond IoT and legacy benchmarks, including cloud-native and enterprise settings [1].
5. Security of the explanation pipeline, including manipulated attributions and prompt injection, where the evidence is thin [2].
6. Efficiency at real alert volumes, including local versus hosted LLM trade-offs [4], [7], [8].

X. LIMITATIONS

The corpus is small and purposively assembled, so relevant work, especially recent primary studies on LLM alert explanation, may be missed. Quantitative statements are reproduced from the sources and not independently verified; several come from single-dataset or preliminary experiments. Two surveys [2], [3] predate widespread LLM use. The proposed framework is untested.

XI. CONCLUSION

Post-hoc XAI is established in security research but weakly evaluated and rarely tested with analysts. LLMs are poor detectors but can become credible explainers when constrained by verified context, though the best measured configuration still erred in one explanation out of five. Explanation pipelines also enlarge the attack surface. The most defensible next contribution is rigorous evaluation of the interfaces among detection, attribution, narrative, and analyst decision. The framework proposed here is a starting point for it.

Declarations. Funding: [to be completed]. Conflicts of interest: [to be completed]. Data availability: no new data were generated. Author contributions: [to be completed].

REFERENCES

[1] P. W. C. Prasad et al., “Explainable AI: Enhancing decision-making in the detection of cyber threats,” Frontiers in Computer Science, vol. 8, Art. no. 1762332, 2026.

[2] F. Charmet et al., “Explainable artificial intelligence for cybersecurity: A literature survey,” Annals of Telecommunications, vol. 77, pp. 789–812, 2022.

[3] F. Yan, S. Wen, S. Nepal, C. Paris, and Y. Xiang, “Explainable machine learning in cybersecurity: A survey,” International Journal of Intelligent Systems, 2022.

[4] B. Nugraha, A. V. Jnanashree, and T. Bauschert, “A versatile XAI-based framework for efficient and explainable intrusion detection systems,” Annals of Telecommunications, vol. 80, pp. 1095–1120, 2025.

[5] H. Xu et al., “Large language models for cyber security: A systematic literature review,” arXiv preprint arXiv:2405.04760, 2025.

[6] G. J. C. da Silva and C. B. Westphall, “A survey of large language models in cybersecurity,” arXiv preprint arXiv:2402.16968, 2024.

[7] P. R. B. Houssel et al., “Towards explainable network intrusion detection using large language models,” arXiv preprint arXiv:2408.04342, 2024.

[8] P. R. B. Houssel et al., “eX-NIDS: A framework for explainable network intrusion detection leveraging large language models,” arXiv preprint arXiv:2507.16241, 2025.