Can LLMs Explain AI? A Study of Explainable AI in Cybersecurity Machine Learning Models

Elton D. Freitas[1], Beloin S. F. Rodrigues[1], Emanuel B. Rodrigues[1], Rossana M. C. Andrade[1], Dário V. Conceição[2], Larisse C. Lucas[1], George P. S. Fernandes[1], Oracio C. Melo[1], Miguel F. Castro[1]

[1]Group of Computer Networks, Software Engineering and Systems (GREat)
Federal University of Ceará (UFC)
60.440 – 554 – Fortaleza – CE – Brazil

[2]Efrei Research Lab
École d'Ingénieur des Technologies de l'Information et de la Communication
94800 – Villejuif – Île-de-France – France

Abstract. *This work evaluates whether Large Language Models (LLMs) can support Explainable Artificial Intelligence (XAI) in cybersecurity, as standalone tools or combined with SHAP and LIME. Experiments with Random Forests on three intrusion detection datasets compared GPT-5 and GPT-OSS-20B explanations to SHAP and LIME. Standalone LLMs showed semantic bias and hallucinated feature importance, while SHAP/LIME-grounded prompts improved faithfulness and coherence. In a qualitative study on explainability in cybersecurity with 38 participants, LLM+SHAP/LIME explanations were preferred over SHAP/LIME-only. GPT-5 achieved the highest preference, while GPT-OSS-20B demonstrated the potential of lightweight local LLMs for XAI.*

1. Introduction

Machine Learning (ML) has become fundamental in cybersecurity, enabling anomaly detection, malicious-pattern identification, and threat classification with high accuracy in domains such as Intrusion Detection Systems (IDS) [Pinto et al. 2023], malware classification [Ahmed et al. 2023], and encrypted traffic analysis [Elmaghraby et al. 2024]. However, many ML-based cybersecurity systems still operate as "black boxes", limiting transparency and trustworthiness.

Explainable Artificial Intelligence (XAI) techniques such as SHapley Additive exPlanations (SHAP) [Lundberg and Lee 2017] and Local Interpretable Model-agnostic Explanations (LIME) [Ribeiro et al. 2016] have been adopted to improve interpretability, yet have been adopted to improve interpretability, yet surveys [Barredo Arrieta et al. 2020] and [Giarimpampa et al. 2025] indicate that many cybersecurity solutions still lack effective explainability mechanisms. More recently, Large Language Models (LLMs) have been explored to generate natural language explanations from XAI outputs, potentially improving accessibility for non-expert users [Bilal et al. 2025]. However, ensuring the reliability and consistency of these explanations remains challenging, particularly regarding semantic bias and misleading interpretations. This raises two research questions: (i) Can LLMs be reliably used as standalone explainability tools for cybersecurity ML models, or should they be combined with SHAP and LIME to ensure fidelity and reduce

semantic bias? (ii) Do LLM+SHAP/LIME explanations lead to higher user preference and perceived interpretability than SHAP/LIME-only outputs in cybersecurity contexts?

This work investigates the feasibility of using LLMs for XAI in cybersecurity, evaluating four approaches: (i) SHAP/LIME-only; (ii) LLM-only; (iii) Commercial LLM+SHAP/LIME; and (iv) Local LLM+SHAP/LIME. The main contributions are: (i) Empirical Evaluation of LLM-based XAI: demonstrates, through experiments, the limitations of standalone LLM explainability, including semantic bias and inconsistency relative to SHAP and LIME; (ii) Human-Centered Evaluation: a user study with cybersecurity professionals, researchers, and enthusiasts of diverse expertise, comparing explainability approaches for a network traffic classification model.

2. Related Work

The integration of LLMs and XAI has been investigated across several cybersecurity applications. Multiple works combine SHAP/LIME feature importance with LLM-generated natural language justifications: for malware detection decisions [Gravereaux and Islam 2025]; for LSTM-based DDoS detection in next-generation Radio Access Networks (RANs) [Chatzimiltis et al. 2026]; for IDS frameworks aligning LLM-parsed security rules with SHAP explanations [Alnahdi and Narain 2024]; for real-time intrusion detection with actionable mitigation narratives, the NIDRF framework [Baral et al. 2024]; and for general threat detection combining ML, XAI, and LLMs [Ghazal et al. 2024].

Similar integration patterns appear in cyber-risk management, where CodeBERT-based vulnerability prioritization is justified through SHAP and correlation heatmaps aligned with NIST 800-53 [Papastergiou et al. 2026]; in phishing detection, where the EXPLICATE framework combines LIME/SHAP with DeepSeek v3 explanations [Lim et al. 2025] and HuntGPT combines a Random Forest classifier, SHAP, and GPT-3.5 Turbo for analyst-facing narratives[Ali and Kostakos 2023]; and in software vulnerability analysis, using CodeBERT with Model-Agnostic Meta-Learning for few-shot detection [Basheer et al. 2025]; and instruction-tuned LLMs (LLMVulExp) for detection and remediation explanations [Mao et al. 2025]. Table 1 summarizes the main characteristics of the related works and highlights the positioning of the proposed approach in comparison with the current literature.

3. Methodology

To investigate the role of LLMs in XAI for cybersecurity, an experimental methodology was designed to evaluate the fidelity and interpretability of LLM-generated explanations, comparing traditional XAI techniques with LLM-based explanations. An ML model was trained to detect cyberattacks and abnormal network traffic, following paradigms adopted in recent studies [Choubisa et al. 2022, Wu et al. 2022]. The primary focus is the explainability mechanisms applied to the model rather than the model itself, so the methodology is model-agnostic and applicable to different ML algorithms and cybersecurity scenarios.

3.1. Model Training, Evaluation and Explainability using SHAP and LIME

The ML algorithm adopted was Random Forest, widely used in IDS for its robustness to overfitting on large, multi-feature datasets [Hozouri et al. 2025] and identified as one

Table 1. Comparison with the main related works

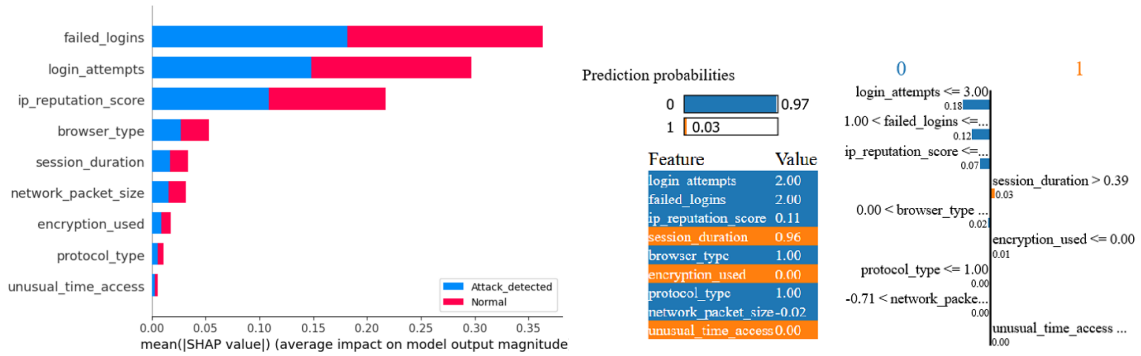
Work	LLM-only for XAI	LLM + SHAP/LIME for XAI	Human Evaluation
[Gravereaux and Islam 2025]	-	✓	-
[Chatzimiltis et al. 2026]	-	✓	-
[Alnahdi and Narain 2024]	-	✓	-
[Baral et al. 2024]	-	✓	-
[Ghazal et al. 2024]	-	✓	-
[Papastergiou et al. 2026]	-	✓	-
[Lim et al. 2025]	-	✓	-
[Ali and Kostakos 2023]	-	✓	-
[Basheer et al. 2025]	-	✓	-
[Mao et al. 2025]	✓	-	✓
This work	✓	✓	✓

of the most common algorithms in monitoring tools used in Security Operations Centers (SOCs) [Giarimpampa et al. 2025]. Three publicly available datasets hosted on the Kaggle repository were used. The Random Forest model was trained separately on each dataset using a 70/30 train/test split. Performance was evaluated using accuracy, precision, recall, F1-score, and AUC. For Dataset 1, the model achieved an accuracy, precision, recall, and F1-score of 0.9977, with an AUC of 1.00. For Dataset 2, the corresponding values were 0.8944, 0.9099, 0.8944, and 0.8919, respectively, with an AUC of 0.88. Dataset 3 yielded an accuracy, precision, recall, and F1-score of 0.9967, with an AUC of 1.00.

SHAP and LIME were applied to generate feature-based explanations: SHAP provided global feature importance via Shapley values, while LIME generated local, instance-level explanations. As shown in Table 2, both techniques produced highly consistent results across datasets, with only minor ranking differences due to LIME's local nature. This agreement suggests that the extracted explanations provide a reliable reference for evaluating LLM-generated outputs. As an example, Figure 1 illustrates both global and local explainability perspectives provided by SHAP and LIME for Dataset 2.

Table 2. Comparison of the top-3 most important features identified by SHAP and LIME across the evaluated datasets

Dataset	Number of features	Top-3 features	
		SHAP	LIME
Dataset 1	6	1. *payload_size* 2. *port* 3. *status*	1. *port* 2. *payload_size* 3. *status*
Dataset 2	9	1. *failed_logins* 2. *login_attempts* 3. *ip_reputation_score*	1. *login_attempts* 2. *failed_logins* 3. *ip_reputation_score*
Dataset 3	41	1. *src_bytes* 2. *dst_bytes* 3. *flag*	1. *src_bytes* 2. *dst_bytes* 3. *num_failed_logins*



(a) Global SHAP feature importance **(b) LIME explanation of a random instance**

Figura 1. Global SHAP feature importance and LIME local explanations for randomly selected instances for Dataset 2

3.2. Model Explainability using LLMs

Two configurations were evaluated: (i) LLM-only, in which models generated explanations without SHAP/LIME information, and (ii) LLM+SHAP/LIME, in which SHAP importance values and LIME decision rules were incorporated into the prompts. For each dataset, LLMs received a structured JSON prompt containing the model description, feature descriptions and encodings, a 50-row data sample, and a 50-row prediction sample (predicted vs. real labels), and were asked to explain which features most influenced predictions. Following prompt-engineering practices [Boonstra 2025], Chain-of-Thought (CoT) was adopted as the primary prompting strategy for its effectiveness in multi-step reasoning [Wei et al. 2023]. All experiments used multi-turn prompts with a 12,288-token maximum, with temperature and top-p left at provider defaults. Two LLMs were evaluated: GPT-5, a large-scale commercial model, and GPT-OSS-20B, an open and locally executable model, used as a lightweight self-hosted alternative. In the LLM+SHAP/LIME configuration, SHAP importance values and LIME decision rules were added to the prompt context to evaluate whether structured explainability information improves fidelity and consistency.

3.3. Human-Centered Evaluation

To address the second research question, a human-centered evaluation was conducted with cybersecurity professionals, researchers, enthusiasts, and non-technical participants. Dataset 1 was selected for its lower complexity and more accessible feature space. Since LLM-only explanations showed significant fidelity issues (Section 4.1), the evaluation focused on SHAP/LIME-grounded approaches. After an overview of the dataset, model, and evaluation results, participants were exposed to three explainability approaches: (1) SHAP/LIME-only, feature importance plots (as shown in Figure 1); (2) LLM+SHAP/LIME using GPT-5, following the methodology in subsection 3.2 but additionally incorporating SHAP/LIME outputs; and (3) LLM+SHAP/LIME using GPT-OSS-20B, an approach similar to the previous one, but using an open LLM running locally. For each approach, participants rated ease of understanding, perceived technical complexity, and the quantity/relevance of information, then selected their preferred approach with a written justification.

4. Results and Discussion

4.1. Can LLMs be reliably used as standalone tools for XAI in cybersecurity ML models?

When prompted without SHAP or LIME, GPT-5 and GPT-OSS-20B produced explanations that diverged from the SHAP/LIME ground truth (Table 2). In Dataset 1, both models overvalued the

user_agent feature despite its near-zero SHAP importance; GPT-OSS-20B ranked it top-3 for all classes, GPT-5 for two of three. Similar divergence occurred in Dataset 3. Dataset 2 was the only case where GPT-5 fully agreed with the SHAP/LIME reference; GPT-OSS-20B was partially coherent but over-ranked *session_duration*.

To quantify agreement, a Top-n Feature Agreement Score (TFAS) was defined (Equation 1) as the proportion of overlap between the top-n LLM features and the union of the SHAP/LIME top-n reference features (n = 3). As summarized in Table 3, GPT-5 achieved perfect agreement only in Dataset 2, with an average TFAS of 0.67 across classes and datasets; GPT-OSS-20B averaged 0.33. These results show that LLM-only explanations remain only partially aligned with actual classifier behavior, while incorporating SHAP/LIME information into the prompts substantially improved fidelity and consistency. For this reason, the human-centered evaluation focused exclusively on SHAP/LIME-grounded explanations.

$$TFAS = \frac{|F(n)_{LLM} \cap F(n)_{XAI}|}{n} \tag{1}$$

Table 3. TFAS results across datasets and classes for top-3 feature importance

Dataset	Class	GPT-5	GPT-OSS-20B
Dataset 1	BotAttack	0.67	0.33
	Normal	0.67	0.33
	Port Scan	0.67	0.33
Dataset 2	Normal	1.00	0.67
	Attack	1.00	0.67
Dataset 3	Normal	0.33	0.00
	Abnormal	0.33	0.00
Average	–	**0.67**	**0.33**

4.2. Do LLM+SHAP/LIME explanations lead to higher user preference and perceived interpretability compared to SHAP/LIME-only outputs in cybersecurity contexts?

A total of 38 participants completed the evaluation. Regarding cybersecurity knowledge, 44.7% (n=17) reported intermediate and 39.5% (n=15) basic knowledge, while advanced/specialist participants totaled 15.8% (n=6). For XAI familiarity, 31.6% (n=12) reported superficial knowledge and 31.6% (n=12) no prior knowledge, while only 10.5% (n=4) reported advanced/specialist familiarity, indicating that evaluators included both experts and non-specialists.

Table 4 summarizes the main results of the human-centered evaluation. LLM+SHAP/LIME explanations outperformed SHAP/LIME-only in comprehension: GPT-5 was perceived as easiest to understand, followed by GPT-OSS-20B, while SHAP/LIME-only visualizations were often considered difficult. Regarding coherence (asked only for LLM-based explanations, since SHAP/LIME-only outputs contained no generated text), GPT-5 achieved the highest perceived consistency, with GPT-OSS-20B receiving few inconsistency reports. For most evaluation criteria, participants assessed all three explainability approaches independently, meaning that the total number of responses could exceed the number of participants. For example, a participant could classify all approaches as easy to understand. The only exception was the "XAI approach preference" criterion, in which participants were required to select a single preferred approach. When asked to select the single most useful approach, GPT-5 achieved the highest preference (57.9%, n=22), followed by GPT-OSS-20B (34.2%, n=13), with SHAP/LIME-only

selected by only 7.9% (n=3). Overall, GPT-5 achieved the highest preference for clarity and usefulness, while GPT-OSS-20B produced highly competitive results and, as a locally deployable model, avoids privacy and connectivity limitations associated with cloud-based proprietary APIs, a relevant advantage for sensitive, restricted cybersecurity environments.

Table 4. Summary of main explainability evaluation results

Evaluation category	**SHAP/LIME**	**GPT-5**	**GPT-OSS-20B**
Ease of understanding	36.84% (n=14)	65.79% (n=25)	52.63% (n=20)
High information relevance	39.47% (n=15)	47.37% (n=18)	52.63% (n=20)
Non-technical adequacy	21.05% (n=08)	39.47% (n=15)	42.10% (n=16)
Technical adequacy	55.26% (n=21)	50.00% (n=19)	60.53% (n=23)
User corroboration	N/A	97.37% (n=37)	89.47% (n=34)
XAI approach preference	**7.89% (n=3)**	**57.89% (n=22)**	**34.21% (n=13)**

5. Limitations of the Study and Ethical Issues

Although three distinct datasets were used for robustness, they may not fully capture real-world complexities such as dynamic traffic, encrypted communications, and evolving attacks. The human-centered evaluation's sample size and expertise distribution may also limit generalizability. Computational overhead, energy consumption, and cost of LLM+SHAP/LIME systems were not evaluated. The evaluation was voluntary, anonymous, and collected no personally identifiable, financial, or health-related data; participants were informed beforehand and could withdraw at any time.

6. Conclusion

This work investigated LLMs, SHAP, and LIME for XAI in cybersecurity ML models. Across three intrusion detection datasets, standalone LLM explanations showed semantic bias and inconsistency with classifier behavior, while SHAP/LIME-grounded explanations aligned substantially better. The human-centered evaluation found LLM+SHAP/LIME explanations easier to understand and more useful than SHAP/LIME-only, especially for non-technical users. GPT-5 achieved the highest preference, while GPT-OSS-20B remained competitive while avoiding cloud-based privacy and operational limitations, suggesting local LLMs are viable for explainable cybersecurity systems. Future work includes combining visual XAI with LLM narratives via data storytelling, adaptive systems that adjust explanation complexity to user expertise, and further study of LLM+SHAP/LIME robustness against prompt injection, adversarial manipulation, and hallucinations.

7. Reproducibility and Artifact Availability

Source code, trained models, prompts, and datasets are available via GitHub at: Can LLMs Explain AI? A Study of Explainable AI in Cybersecurity Machine Learning Models

8. Acknowledgments

The authors acknowledge the use of Generative Artificial Intelligence tools for minor editorial revisions, mainly to shorten paragraphs for page limits and to translate excerpts from Portuguese into English.

References

Ahmed, M., Afreen, N., Ahmed, M., Sameer, M., and Ahamed, J. (2023). An inception v3 approach for malware classification using machine learning and transfer learning. *International Journal of Intelligent Networks*, 4:11–18.

Ali, T. and Kostakos, P. (2023). Huntgpt: Integrating machine learning-based anomaly detection and explainable ai with large language models (llms).

Alnahdi, A. and Narain, S. (2024). Towards transparent intrusion detection: A coherence-based framework in explainable ai integrating large language models. pages 87–96.

Baral, S., Saha, S., and Haque, A. (2024). An adaptive end-to-end iot security framework using explainable ai and llms.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Basheer, N., Islam, S., Alwaheidi, M., Mouratidis, H., and Papastergiou, S. (2025). Large language model based hybrid framework for automatic vulnerability detection with explainable ai for cybersecurity enhancement. *Integrated Computer-Aided Engineering*, 33.

Bilal, A., Ebert, D., and Lin, B. (2025). Llms for explainable ai: A comprehensive survey.

Boonstra, L. (2025). Prompt engineering. *Google, https://www. kaggle. com/whitepaper-prompt-engineering*.

Chatzimiltis, S., Shojafar, M., Mashhadi, M. B., and Tafazolli, R. (2026). Ai-on-ran for cyber defense: An xai-llm framework for interpretable anomaly detection. *IEEE Transactions on Network Science and Engineering*, 13:3301–3319.

Choubisa, M., Doshi, R., Khatri, N., and Kant Hiran, K. (2022). A simple and robust approach of random forest for intrusion detection system in cyber security. In *2022 International Conference on IoT and Blockchain Technology (ICIBT)*, pages 1–5.

Elmaghraby, R. T., Abdel Aziem, N. M., Sobh, M. A., and Bahaa-Eldin, A. M. (2024). Encrypted network traffic classification based on machine learning. *Ain Shams Engineering Journal*, 15(2):102361.

Ghazal, T. M., Janjua, J. I., Abushiba, W., Ahmad, M., Ihsan, A., and Al-Dmour, N. A. (2024). Cybersecurity revolution via large language models and explainable ai. In *2024 17th International Conference on Security of Information and Networks (SIN)*, pages 1–6.

Giarimpampa, D., Meier, R., Bissyande, T. F., Lenders, V., and Klein, J. (2025). Exploring the role of artificial intelligence in enhancing security operations: A systematic review. *ACM Comput. Surv.*, 58(3).

Gravereaux, S. C. and Islam, S. R. (2025). Accuracy and efficiency trade-offs in llm-based malware detection and explanation: A comparative study of parameter tuning vs. full fine-tuning.

Hozouri, A., Mirzaei, A., and Zhu, Effatparvar, M. (2025). A comprehensive survey on intrusion detection systems with advances in machine learning, deep learning and emerging cybersecurity challenges. *Discover Artificial Intelligence*.

Lim, B., Huerta, R., Sotelo, A., Quintela, A., and Kumar, P. (2025). Explicate: Enhancing phishing detection through explainable ai and llm-powered interpretability.

Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting model predictions.

Mao, Q., Li, Z., Hu, X., Liu, K., Xia, X., and Sun, J. (2025). Towards explainable vulnerability detection with large language models.

Papastergiou, S., Basheer, N., Lampropoulos, K., Verrios, P., and Islam, S. (2026). Explainable ai based dynamic cybersecurity risk management for cyber insurability. *Int. J. Inf. Secur.*, 25(1).

Pinto, A., Herrera, L.-C., Donoso, Y., and Gutierrez, J. A. (2023). Survey on intrusion detection systems based on machine learning techniques for the protection of critical infrastructure. *Sensors*, 23(5).

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models.

Wu, T., Fan, H., Zhu, H., You, C., Zhou, H., and Huang, X. (2022). Intrusion detection system combined enhanced random forest with smote algorithm. *EURASIP Journal on Advances in Signal Processing*.