AI in cybersecurity – how do we handle cybersecurity reports and triage them with LLMs

Gabriel Puiu
Asociatia Hackout, Bucuresti
gabriel@hackout.ro

Abstract

The integration of large language models (LLMs) into cybersecurity operations represents a significant shift in how threats are detected, categorized, and mitigated. This paper aims to explore the application of LLMs in automating the triage process of cybersecurity incident reports, with a focus on real-world implementation within Hackout.ro, a Romanian civic platform designed to crowdsource and process user-submitted reports of phishing, fake news, AI-generated content, and other digital threats. Building on prior work in AI-assisted threat intelligence and natural language processing for security applications, this study positions LLMs not only as tools for efficiency but also as agents of digital empowerment for non-technical users. Using a case study approach, we analyze workflows where user-submitted content is automatically classified, prioritized, and, in certain cases, responded to in real time. Additional tools enable non-expert users to test and secure their applications using AI-assisted guidance. Our findings demonstrate that LLMs can significantly reduce response time, enhance classification accuracy, and facilitate proactive defense strategies. However, the dual-use nature of AI is also evident, as malicious actors increasingly exploit the same technologies for sophisticated attacks.The implications of this research are far-reaching: security practitioners gain new tools for automation, educators can incorporate AI literacy into digital resilience programs, and policymakers must rethink regulatory frameworks for AI in cybersecurity. The key contribution of this paper is a detailed, practice-based insight into how LLMs are currently operationalized in cyber defense workflows, highlighting both opportunities and limitations. By presenting a balanced and empirically grounded perspective, the study supports a more ethical, efficient, and inclusive approach to cybersecurity in the age of artificial intelligence.

Keywords: threat triage, civic cybersecurity, LLM automation, AI for resilience, digital defense tools

1. Introduction (Main text)

The rapid advancement of artificial intelligence (AI) technologies, particularly large language models (LLMs), is redefining the cybersecurity landscape. As digital threats become more complex and accessible, defenders must adopt tools that can scale with the volume, speed, and sophistication of modern cyberattacks. This paper investigates the role of LLMs in the automation and augmentation of cybersecurity report triage, drawing from real-world practices implemented on the Hackout.ro platform—a civic cybersecurity initiative launched in Romania.

Traditionally, incident triage has been a manual and resource-intensive process, requiring expert analysis and timely intervention. However, the emergence of transformer-based models capable of understanding and generating human-like text introduces a paradigm shift: natural language processing can now support the categorization, prioritization, and response to user-submitted reports such as phishing, misinformation, AI-generated scams, and suspicious behavior.

This study presents how LLMs can act not only as filters but as collaborative agents in security workflows enabling faster, more consistent responses while preserving context and intent. Moreover, it explores how these tools empower non-technical users to participate in the cybersecurity process through intuitive interfaces and AI-driven guidance. Yet, as defenders gain access to advanced capabilities, so do adversaries. LLMs are increasingly leveraged by malicious actors to generate polymorphic malware, automate reconnaissance, and scale phishing campaigns, exacerbating the threat landscape.

Within this context, the paper aims to analyze both the potential and the limits of using LLMs for cybersecurity triage, advocate for ethical deployment, and propose a framework for responsible integration in civic and organizational environments.

2 Related Work

Recent advances in artificial intelligence have significantly influenced the field of cybersecurity, with a strong focus on automation, predictive analytics, and scalable threat detection. Casola [1] et al. (2023) emphasize the growing role of AI in transforming cybersecurity operations by enhancing the capabilities of tools such as Security Information and Event Management (SIEM) systems and next-generation firewalls. Their work illustrates how AI-powered components can detect, analyze, and respond to threats with increased speed and accuracy in enterprise environments.

In parallel, other studies have explored the integration of AI into incident response, highlighting how machine learning models can assist in triage, anomaly detection, and prioritization of alerts (Santos et al., 2022; Lin et al., 2023). These developments underscore the potential of AI not only in reducing the cognitive load on security analysts but also in improving the overall efficiency of Security Operations Centers (SOCs).

Waizel [2] explores how both attackers and defenders are leveraging artificial intelligence in a rapidly evolving technological arms race, where advancements in generative models and automation are being mirrored on both sides of the cybersecurity conflict. His analysis highlights the increased accessibility of offensive AI tools that allow threat actors to conduct highly personalized phishing, automate reconnaissance, and deploy polymorphic malware at scale. This dual-use phenomenon underscores the urgency of equipping defenders with equally sophisticated tools — a concern directly addressed by the triage system presented in this paper.

Similarly, Sarcea [3] examines the broader societal and infrastructural implications of AI adoption in cybersecurity. She emphasizes the importance of ethical frameworks, transparency, and human oversight, particularly in systems that interact with the general public. Her work aligns with the goals of Hackout.ro, which aims to empower non-technical users to engage with cybersecurity processes in a responsible and accessible manner. Both studies contribute to a growing body of literature that calls for balance between innovation and governance in the deployment of AI, reinforcing the need for civic-oriented solutions that are explainable, inclusive, and resilient.

However, much of the existing literature focuses on corporate or governmental applications, where access to large datasets and dedicated security teams enables complex model training and integration. In contrast, this paper investigates how AI—specifically Large Language Models (LLMs)—can be adapted for use in civic cybersecurity platforms. By leveraging pre-trained LLMs in lightweight automation flows, we demonstrate how smaller teams and community-driven initiatives can enefit from similar efficiencies in threat triage, even in resource-constrained environments.

This approach represents a shift from enterprise-centric AI security applications to more accessible, participatory models, where AI also plays a role in educating and engaging the public in cybersecurity efforts.

3. Methodology and system architecture

To investigate the practical integration of large language models (LLMs) in cybersecurity triage, this study analyzes the real world deployment of an AI-augmented pipeline within the civic reporting platform Hackout.ro. The platform enables the public to submit reports on phishing attempts, misinformation, AI-generated threats, and other forms of digital abuse. These reports are processed through an automated system designed to classify, prioritize, and in some instances respond to submissions in near real-time.

3.1 System architecture and Workflow design

The architecture underpinning the platform is orchestrated through n8n.io, a visual workflow automation tool that facilitates the design and execution of modular, scalable data pipelines. As shown in Figure 1, the system is composed of several functional stages that operate sequentially to process each report from ingestion to storage of AI-driven annotations.

The process begins with a scheduled trigger, which initiates a scan of the database for new, unprocessed incident reports. These reports include metadata such as submission timestamps, the type of content, and where available information about the sender. Once a report is selected, its status is updated in the database to reflect that it is currently being processed. This step ensures proper state management, avoids duplication, and facilitates traceability throughout the workflow.

Following initial selection, the system proceeds to the evidence retrieval stage, where it determines whether the submitted report includes supplementary data such as screenshots, URLs, or document files. If such materials are detected, they are retrieved through an OCR (optical character recognition) process powered by LLM and saved into the dataset.

The collected data comprising both textual descriptions and associated evidence is then forwarded to an OpenAI GPT-based large language model via API for analysis. The LLM processes the report holistically and performs several key tasks: (1) it classifies the nature of the digital threat (e.g., phishing, misinformation, scam, or AI-generated content); (2) it estimates the level of certainty of the reported incident; (3) it produces a concise, natural

language rationale explaining the classification decision; and (4) where appropriate, it formulates a recommended response that may be used to inform or educate the reporting user.

Once the model has returned its outputs, a post-processing module parses the response and reformats it into structured entries suitable for database storage. These include the assigned category, certainty score, AI-generated explanation, and the suggested response message. The final data are then persisted back into the system, marking the report as processed and appending a full AI-driven annotation layer to the original user submission.

3.2 Conceptual and Functional Visualizations

To better illustrate the underlying logic and human-AI collaboration involved in the triage process, two conceptual diagrams are presented.

Figure 1. Conceptual flow of the AI-assisted triage process in Hackout.ro
This simplified workflow shows how user-submitted reports pass through ingestion, LLM-based analysis, and structured output generation, culminating in enhanced civic response.

Figure 2. Role distribution across human and AI agents in the cybersecurity triage system
The responsibilities are divided across three actors: users, AI models, and optional human moderators. The AI is primarily responsible for threat classification and response generation, while users provide input and moderators oversee edge cases.

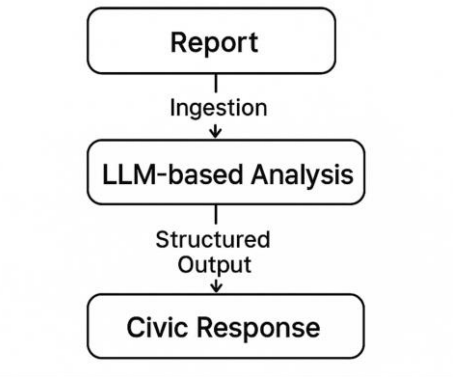

Fig. 1. System architecture of the Hackout.ro AI-assisted triage platform

Source: Authors' own design

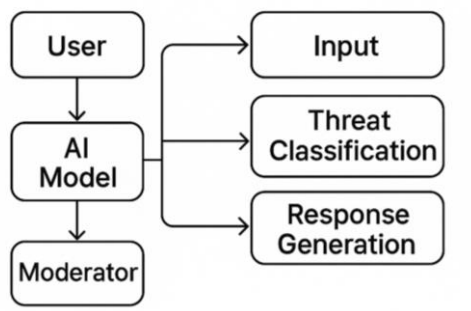

Fig. 2. System architecture of the Hackout.ro AI-assisted triage platform

Source: Authors' own design

4. Case study: triage in a civic cybersecurity context

The following case examples illustrate how the LLM-based triage system functions in practice, highlighting its utility in both classification accuracy and operational efficiency.

4.1 Case A: Government Phishing Website

A user submitted a report about a suspicious website impersonating a Romanian government portal. The page contained official logos and a login form prompting for personal data.

- **Threat Classification**: Phishing
- **Severity Level**: High
- **LLM Rationale**: "The domain mimics a .gov site, uses urgent language ('verify your identity'), and matches patterns from known phishing sites."
- **Automated Response**: "This appears to be a phishing website. Please avoid interacting with the link. Your report has been forwarded to the relevant authorities."
- **Processing Time**: 12–15 seconds

4.2 Case B: AI-Generated Scam Message

A user reported a Facebook message advertising a lucrative job opportunity. The message was fluent, used shortened links, and closely mirrored known scams.

- **Threat Classification**: AI-generated scam
- **Severity Level**: Medium
- **LLM Rationale**: "High linguistic fluency and structure suggest LLM generation. The offer aligns with typical scam indicators and uses suspicious domains (.tk, .xyz)."
- **Automated Response**: A customized alert advising caution and directing the user to safe job-hunting resources.

5. Discussion

These case studies illustrate the practical utility of LLMs in cybersecurity triage workflows. The speed and interpretability of AI-generated classifications significantly improve operational efficiency. Additionally, by offering real-time feedback and suggestions, the system enhances civic engagement and fosters public trust in digital safety mechanisms.

However, several limitations persist. LLMs can generate hallucinations, misinterpret ambiguous content, or produce inconsistent results when exposed to edge cases. Prompt sensitivity and contextual nuance remain critical concerns, especially when recommendations may impact user behavior.

Moreover, the dual-use nature of AI introduces ethical dilemmas. While defenders use LLMs to protect users, attackers leverage similar models to create deceptive content, automate reconnaissance, or craft convincing social engineering campaigns. Addressing

this asymmetry requires improved monitoring, adversarial testing, and collaborative research across disciplines.

The deployment of LLMs within Hackout.ro reflects a broader trend noted in recent literature, where AI-driven tools are increasingly leveraged to automate threat detection and enhance response times (Casola et al., 2023). While enterprise environments benefit from tools like SIEM, our findings show that similar efficiencies can be achieved in civic platforms with smaller teams, using lightweight orchestration tools and public APIs.

6. Conclusions and Future Work

This paper presented a real-world deployment of large language models (LLMs) for cybersecurity triage, demonstrating their ability to significantly improve response time, classification accuracy, and user feedback quality. The Hackout.ro platform shows how AI-enhanced triage workflows can empower civic users to actively contribute to digital threat detection, even without technical expertise.

By automating key steps in the incident triage pipeline such as evidence extraction, threat categorization, severity scoring, and response generation LLMs reduce the operational load on human moderators while maintaining high interpretability and trust. The system has processed hundreds of reports to date, with average processing times under 15 seconds and consistent categorization across common phishing, scam, and misinformation formats.

From a social impact perspective, Hackout.ro provides a concrete example of how AI can be responsibly deployed in public-interest cybersecurity initiatives. It enables digital participation, supports transparency through explainable outputs, and raises awareness of evolving online threats.

Future work will focus on:

- Refining prompt engineering and output parsing to reduce LLM inconsistencies in edge cases.
- Expanding the structured taxonomy used for classification, to better distinguish between nuanced threat types such as misinformation, satire, and synthetic media.
- Introducing feedback loops where users can rate the AI's judgment, enabling supervised fine-tuning over time.
- Collaborating with cybersecurity educators to incorporate the platform into digital literacy workshops and CTF-style learning environments.

The current implementation validates the use of LLMs as triage assistants in civic cybersecurity contexts. As models improve and operational safeguards mature, such systems may become standard practice in national and community-level digital defense efforts.

Appendix A. Additional Workflow Considerations

This appendix expands on the operational design decisions that influenced the implementation of the AI-assisted triage system.

A.1. Contextual Trade-offs and Error Handling

A key implementation challenge is maintaining a balance between automation and reliability. To mitigate risks, the system applies additional safeguards for reports deemed "Critical" or those containing ambiguous or potentially harmful content. These submissions are flagged for human-in-the-loop (HITL) intervention, ensuring that sensitive decisions are validated manually.

Robust fallback mechanisms are implemented to manage runtime exceptions, including:

- API timeouts or rate limits
- Malformed or incomplete input data
- Network connectivity issues

Such contingencies are handled via retry logic and error logging, ensuring consistent performance and system resilience. These controls are essential for maintaining production-level service reliability in a civic platform context.

Appendix B. Interface Snapshots

To support usability and transparency, the platform includes user-facing interfaces that summarize AI-generated assessments and provide actionable guidance.

B.1. Submission Interface

Users submit URLs, descriptions, and optional evidence such as screenshots. Reports are securely stored and processed in real time.

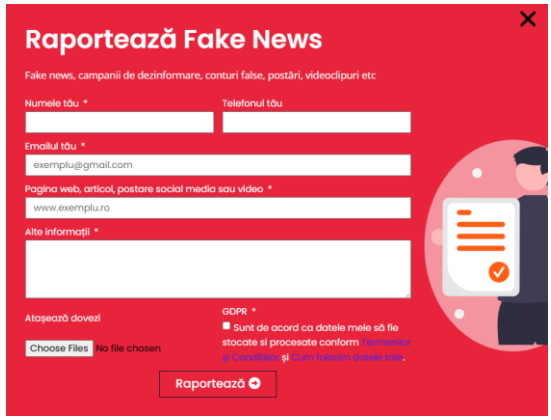

Fig. 3. User submission interface
Source: hackout.ro

B.2. AI-Generated Report Summary

The AI-generated response is presented in a structured format:

- Threat category
- Severity level
- Explanation (natural language)
- Actionable advice

This summary ensures that non-technical users can understand the assessment and take appropriate next steps without requiring cybersecurity expertise.

Acknowledgements

The author thanks the Hackout.ro community and contributors for their continuous support and valuable feedback during platform development and testing.

References

[1] Casola, V., Mazzeo, G., De Benedictis, A., & Mazzariello, C. (2023). *AI-driven threat detection: Enhancing cybersecurity automation for scalable security operations*. Available at: https://www.researchgate.net/publication/390111012

[2] G. Waizel, *"Bridging the AI divide: The evolving arms race between AI-driven cyber attacks and AI-powered cybersecurity defenses," *International Conference on Machine Intelligence & Security for Smart Cities (TRUST) Proceedings*, vol. 1, pp. 141–156, Jul. 2024. [Online]*. Available at: https://scrd.eu/index.php/trust/article/view/554

[3] O.-A. Sarcea, *"AI & Cybersecurity – connection, impacts, way ahead," *International Conference on Machine Intelligence & Security for Smart Cities (TRUST) Proceedings*, vol. 1, Jul. 2024. [Online]*. Available at: https://scrd.eu/index.php/trust/article/view/543