Automating the Cybersecurity Triage Process: A Comparative Study on the Performance of Large Language Models

Pascal Bakker
p.n.bakker@student.utwente.nl
University of Twente

Jair Santanna
j.j.santanna@utwente.nl
jair.santanna@northwave-cybersecurity.com
University of Twente
Northwave Cyber Security

ABSTRACT

In Security Information and Event Management, alarms are created when certain conditions are met. Security analysts have the task to inspect alarms to filter false positives and identify their severity: triage. The problem with this process is that it is complicated and time-consuming, limiting the depth and speed of investigations. Whereas other proposed optimizations and automations appear to be very promising, rapid advancements in the development of Large Language Models (LLMs) opened up new possibilities to speed up the triage process. This research will identify ways in which LLMs can optimize triage, evaluate the performance of these techniques and offer a comparison between different LLMs. The results of this research are expected to help security teams in making informed implementation decisions when optimizing the triage process.

KEYWORDS

Cybersecurity, triage, analysis, large language model, natural language processing, automation

1 INTRODUCTION

(TODO: numbers on security incidents)

The 2023 Cost of a Data Breach Report by IBM Security [7] concludes that the average cost of a security breach in 2023 was 4.45M USD, marking an increase of 2.3% since 2022 and a 15.3% increase compared to 2020. Considering that only 1 in 3 breaches are identified by an organization's security team, and that organizations with high levels of incident response planning saved 1.49M USD on average, it shows the pressing need for security investments in training, thread detection and response technologies.

1.1 Triage

Organizations use Security Operation Centers (SOC) to respond to security incidents in real-time. SOCs consist of security analysts that investigate data from various sources such as Security Information and Event Management (SIEM) systems. The SIEM systems collect log data from a large number of sources such as network devices and applications within the organization's system. Based on rules, patterns and conditions, anomalies and suspicious activities are identified and alarms are created.

The number of alarms is large, ranging from hundreds to thousands per day. This causes security teams to experience fatigue, and it contributes to internal friction and turnover. Besides that, a large portion of the alarms are false positives or low priority [10].

Since the SOCs cannot respond to every single alarm, identifying the severity of alarms is an important step in the incident response workflow. This process, triage, involves understanding the impact of an alarm, correlating it with other alarms and identifying potential future goals of adversaries to conclude its severity. By prioritizing alarms, SOCs can focus their resources on high-severity alarms first, thus mitigating damages and reducing costs.

Although Security Orchestration, Automation, and Response (SOAR) platforms have streamlined parts of the process by automating routine tasks, many steps still require human judgment to make adequate decisions [4]. Consequentially, the triage process is prone to human error. This, in combination with the volume and complexity of alarms, presses the need for the automation of triage.

1.2 Large Language Models

Artificial Intelligence (AI) can potentially play a big role when automating triage. This field of study involves the use of machines to perform tasks that require human intelligence such as reasoning, problem-solving and learning [11]. Machine Learning (ML) is a type of AI that allows systems to solve problems by analyzing patterns in data and making predictions and decisions without explicit programming [12]. One purpose of ML is Natural Language Processing (NLP). NLP involves using computational approaches to process natural-language texts with goals such as translation, summarization, sentiment assessment or generation of texts. It plays a growing role in streamlining and automating business operations, and increasing productivity [6].

A part of NLP is Natural Language Understanding (NLU) which aims to comprehend meaning and intent in natural language. Opposite of this, Natural Language Generation (NLG) focuses on generating original human-like texts. To achieve NLU and NLG, language models make use of so-called encoders and decoders. The purpose of encoders is to turn input texts into fixed-size vectors that act as abstract representations. Language models then use decoders to turn such representations into a generated target output. This approach works well on tasks that map input sequences to output sequences (sec2sec), such as language translation [3, 15].

In 2014, Bahdanau et al. [2] introduced the concept of attention which rids the encoders of creating fixed-length vectors, allowing language models to focus on the most relevant parts of texts and enabling operations on much longer input sequences. Based on this, Vaswani et al. [16] developed the transformer architecture in 2017, setting the precedent for modern LLMs. Transformers are superior in quality, more parallelizable, thus faster, and take less time to train.

Early well-known examples of such transformer-based models are BERT (Bidirectional Encoder Representations from Transformers) [5], and GPT (Generative Pre-trained Transformer) [13].

- BERT was specifically designed as a pre-trained model to be easily fine-tuned for a wide range of tasks such as answering questions and natural language inference, without the need of task-specific architecture. In the cybersecurity domain, BERT models have been fine-tuned to detect malicious software [14] and phishing emails [9].
- GPT is pre-trained on a large amount of unlabeled data and designed to generate coherent context-specific text. Like BERT, it requires fine-tuning to adapt the model to specific tasks.

(TODO: disadvantages fine-tuned models vs general models)

1.3 Research Structure

This research aims to explore the potential of LLMs in optimizing the triage process, as well as evaluate the performance of different LLMs (e.g., GPT-4 [1], Llama 3 [?], Mistral [8]) and establish a comparison of these models. To pursue our goal, we define the following research questions (RQ) as the basis of our research:

- **RQ1:** How can LLMs be integrated into the existing incident response workflow to streamline the triage process?
- **RQ2:** What suitable evaluation metrics should be used to assess the performance of LLMs in cybersecurity triage?
- **RQ3:** How do different LLMs compare in performance when optimizing the cybersecurity triage process?

(TODO: Structure)

2 OPTIMIZING TRIAGE USING LLMS

Explain structure of this section.

2.1 Steps of Triage

What is triage in detail?

2.2 Existing Optimizations

What are the existing methods of optimization?

2.3 Use of LLMs

What relevant things are LLMs used for?

2.4 LLMs in Context of Triage

Answering RQ1.

3 EVALUATION OF LLMS

Research question 2

4 COMPARISON OF LLMS

Research question 3

5 CONCLUSION

Conclusions

6 DISCLAIMER ON THE USE OF AI

Besides the use of LLMs as needed for this research, during the preparation of this work, the authors will use generative AI tools such as ChatGPT, LLama3, Grammarly, JetBrains Grazie and JetBrains full line code completion for the following purposes:

- Find definitions of terms and concepts when conventional tools and search engines are unsatisfactory.
- Check and correct the spelling of words and grammar of sentences.
- Improve the readability of sentences and paragraphs through rewording and restructuring.
- Use code completion functionality to speed up programming tasks.

After using these tools/services, the authors will review and edit the content as needed and take full responsibility for the content of the work. The services will not be used to produce scientific insights, create figures, draw conclusions or provide recommendations.

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[3] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[4] Anton Chuvakin. 2019. *The Uphill Battle of Triaging Alerts*. Dark Reading. https://www.darkreading.com/threat-intelligence/the-uphill-battle-of-triaging-alerts

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] IBM Corporation. n.d.. *What is security information and event management (SIEM)?* IBM Corporation. https://www.ibm.com/topics/siem

[7] IBM Security. 2023. Cost of a Data Breach Report 2023. https://www.ibm.com/reports/data-breach

[8] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).

[9] Younghoo Lee, Joshua Saxe, and Richard Harang. 2020. CATBERT: Context-aware tiny BERT for detecting social engineering emails. *arXiv preprint arXiv:2010.03484* (2020).

[10] Orca Security. 2022. 2022 Cloud Security Alert Fatigue Report. https://orca.security/wp-content/uploads/2022/03/Orca-2022-Cloud-Security-Alert-Fatigue-Report.pdf
[11] Oxford English Dictionary 2024. artificial intelligence, n. In *Oxford English Dictionary*. Oxford University Press. https://doi.org/10.1093/OED/3194963277
[12] Oxford English Dictionary 2024. machine learning, n. In *Oxford English Dictionary*. Oxford University Press. https://doi.org/10.1093/OED/2166790335
[13] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. (2018).
[14] Abir Rahali and Moulay A Akhloufi. 2021. MalBERT: Using transformers for cybersecurity and malicious software detection. *arXiv preprint arXiv:2103.03806* (2021).
[15] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems* 27 (2014).
[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).