News-Based Cyber Attack Impact Assessment: Evaluating the Performance of Large Language Models

DANIAR BAIALIEV, University of Twente, The Netherlands

Growing cyber attack threat necessitates quick and reliable impact assessment methodologies. Traditional approaches, often dependent on expert analysis and established cybersecurity frameworks, are prone to human error, manual effort, and high costs. This study investigates the novel method of integration of Large Language Models (LLMs) with traditional cybersecurity frameworks to enhance the accuracy and efficiency of cyber attack impact assessments. This research explores LLMs' capabilities in processing unstructured text data from news articles to assess impact of a cyber attack and evaluate various cost metrics. The performance of LLMs is evaluated through both quantitative analysis using the Mean Absolute Percentage Error (MAPE) and qualitative assessments via structured questionnaires comparing LLM outputs with expert evaluations. The findings indicate significant potential for this novel approach in impact assessment, though further research is necessary to prove its applicability.

Additional Key Words and Phrases: Large Language Models, cost of cybercrime, news articles, cyber attack, impact assessment, framework

1 INTRODUCTION

With the rise in digital technology and dependence on it, cyber attacks have increasingly become a major concern. However, despite the growing threat and complexity of cyber attacks, the process of assessing their impact remains a significant challenge in the cybersecurity field. Traditional methods often rely on expert analysis and established cybersecurity frameworks. These methods may not grasp all the information available, are prone to human error, require manual effort, and are costly. The study by [Yuryna Connolly et al. 2020] suggests that current impact assessment methods often include "incomplete data, skewed surveys and questionable assumptions".

According to [Zhou et al. 2023] the application of Large Language Models (LLMs) in professional fields holds significant potential especially but can be challenging. Combining this potential of LLMs, their automation capability with characteristic of news articles to provide real-time information and early warnings of cyber attacks can be valuable for the cybersecurity field.This project aims to explore new methods by examining and evaluating the abilities of Large Language Models (LLMs) in estimating the impact of cyber attacks based on the content of news articles. It seeks to integrate these models with traditional cybersecurity assessment frameworks to enhance the accuracy and efficiency of impact analyses. This study will not only contribute to the advancement of automated cyber attack impact assessment but also test the real-world applicability of integrating LLM tools with traditional assessment method.

2 PROBLEM STATEMENT

While there is existing research on cybersecurity impact assessments, there is a notable gap in studies specifically exploring the integration of LLMs into these assessments. The complexity and volume of unstructured textual data from the news articles pose significant challenges for machine learning models, which strugle without explicit categorization for effective analysis. Additionally, the traditional methods reliant on manual expert labor are costly and error-prone, underscoring the need for an automated approach.To address aforementioned issues in the field, this research aims to explore new methods by investigating whether LLMs can feasibly enhance the accuracy and efficiency of cyber attack impact assessments.

2.1 Research question

In order to address the stated problem in the research, the following research question (RQ) was formulated:

How accurately and clearly can language models estimate the impact of cyber attacks based on news articles compared to established methods?

Consequently, the following sub questions were introduced to help answer proposed RQ:

1. What cybersecurity frameworks are commonly utilized to assess the impact of cyber attacks based on news articles, and which factors most significantly influence this assessment?
2. How can existing LLMs be integrated with traditional frameworks to assess the impact of cyber attacks?
3. How do the outputs of LLMs align with expert assessments of cyber attack impacts, and what are the strengths and limitations of using LLMs for this purpose?

3 METHODOLOGY

3.1 Literature Review

For the purposes of the project, it is essential to choose relevant cyber attack impact assessment frameworks or methods. Additionally, it is important to review the literature for insights on how to integrate these frameworks with LLMs. For these purposes, a literature review was conducted using resources available through the University of Twente Library[1]. Due to the extensive collection of research available on LLMs, cyber attack impact and risk assessments, databases such as Scopus[2], IEEE Xplore[3],and ACM Digital Library[4] were used for this review. A selection of news

TScIT 41, July 5, 2024, Enschede, The Netherlands

[1] https://www.utwente.nl/en/service-portal/university-library
[2] https://www.scopus.com/home.uri
[3] https://ieeexplore.ieee.org/Xplore/home.jsp
[4] https://dl.acm.org/

articles were sourced from Google News[5] and Nexis Uni[6] and FBI's public service announcement center[7]. This step ensures that the selected framework and integration approach are grounded in established research, enhancing the credibility and relevance of the research.

(1) **Existing impact assessment frameworks:**
("cost*" OR "loss*" OR "revenue*") AND ("cyber attac*" OR "cybersecurity" OR "cybercrime") NOT ("detect*" OR "control" OR "simulation*" OR "monitor*" OR "vulnerabl*" OR "mitigat*" OR "secure*" OR "threat*")
(2) **LLM integration:**
("Large Language Models" OR "Natural Language Processing" OR "NLP" OR "LLM") AND ("prompt engineering" OR "fine tuning") AND ("text analysis" OR "text mining" OR "news articles" OR "news media")
(3) **Cyber attacks selection:**
("report*" OR "cost*") AND ("cyber attac*" OR "cybersecurity" OR "cybercrime" OR "Data breach*") NOT ("detect*" OR "control" OR "simulation*" OR "monitor*" OR "mitigat*" OR "secure*" OR "threat*")

Following the selection of the framework, an additional query was formulated to find more detailed descriptions of cyber attacks within the selected framework and to discover additional cyber attacks. *For information on the framework selection process, refer to Section 5.2.*
(4) **Framework specific cyber attack impact assessments:**
("Direct loss*" OR "Indirect loss*" OR "Defence cost*" OR "Criminal revenue" OR) AND ("cyber attac*" OR "cybersecurity" OR "cybercrime") NOT ("detect*" OR "control" OR "simulation*" OR "monitor*" OR "mitigat*")

The queries were applied to the selected set of digital libraries to define the number of relevant articles. The search resulted in 517 articles through the first phase of the search.Further, the resulting set of articles was filtered based on inclusion and exclusion criteria discussed in table 1.

Criteria	Decision
Keywords are present in title, abstract, or keyword list	Inclusion
Publication in a scientific journal/conference	Inclusion
English language	Inclusion
Duplicate articles	Exclusion
Article which do not have open access	Exclusion

Table 1. Inclusion/exclusion criteria

After applying these criteria, the abstracts of the remaining articles were reviewed using the Zotero application. Papers that were irrelevant for this study were eliminated. This process further refined the selection, reducing the number of relevant articles to **37**.

3.2 LLM Integration

With insights from a literature review, this project looked at different Large Language Models (LLMs) to determine their suitability for text analysis and cyber risk assessment. Several LLMs, including BERT, GPT-4, Llama etc. were assessed for their capabilities in processing and analyzing unstructured text data from news articles. The selection criteria included the models' ability to handle big data, the effectiveness of their text analysis features, and their proven performance in similar domains. This step is essential to ensure that the selected model can provide accurate and relevant outputs for cyber attack impact assessments. For this purposes, BERT, known for its deep bidirectional representations, was considered for its efficiency in tasks where generative capabilities are less critical but precise classification is necessary. However, GPT-4 was selected based on its ease of integration with the chosen framework, project's time-constraints, the volume of existing studies utilizing this model for cybersecurity purposes, and its second best performance demonstrated by [Patel et al. 2024] in analyzing news articles related to cyber attacks.

To facilitate the integration of GPT-4 with the chosen cybersecurity framework and to enhance accuracy while mitigating risk of hallucinations, structured prompts were created to guide the LLM in extracting relevant information from news articles. These prompts were designed to align with the framework criteria. Prompt engineering techniques were employed to refine the outputs of the LLM. This iterative process involved adjusting the prompts based on the initial results to enhance the accuracy and relevance of the information extracted. Methods outlined by [Rodriguez et al. 2023] [Zamfirescu-Pereira et al. 2023], [Jha et al. 2023] were used, these methods include: Including examples of desired interactions to guide the model, breaking down prompts into smaller, more manageable components, using reminders or contextual information within the prompts to guide the model's responses, adding counterexamples into prompts to steer the model away from incorrect responses, using delimiters to add structure to a prompt, asking the LLM to adopt the persona of a security specialist to provide context and focus for the analysis, providing clear and direct instructions in the prompt.

Additionally, [Brown et al. 2020] state that LLMs are few-shot learners and highlight that adding a small amount of training data can substantially improve the model's accuracy. Therefore, an example cost estimation was incorporated into the prompt to assist in fine-tuning the LLM. The Anthem 2015 breach was used as training data due to availability of the most cost estimations and news articles, providing more information.

This combination of structured prompt engineering and few-shot learning techniques aimed to ensure that the LLM's outputs were both accurate and aligned with the detailed requirements of the cybersecurity assessment framework.

[5] https://news.google.com/
[6] https://ezproxy2.utwente.nl/login?url=http://www.nexisuni.com/
[7] https://www.ic3.gov/Home/ConsumerAlertsChoice

3.3 Performance analysis

To evaluate the effectiveness of using a LLM for assessing the impact of cyber attack, the LLM's output was compared with the assessments made by experts. The evaluation was divided into two main parts: quantitative analysis of numerical outputs (e.g. direct losses, defence costs) and qualitative analysis of the output.

3.3.1 Quantitative Analysis. To measure how much the LLM's estimated costs deviate from expert assessments, the Mean Absolute Percentage Error (MAPE) was used. This metric is commonly used in forecasting and helps to express the accuracy of predictions as a percentage. The formula for MAPE is:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right| \times 100$$

where:

- n is the number of observations,
- A_i is the actual value (expert assessment),
- F_i is the forecasted value (LLM output).

This formula allowed to quantify the average deviation of the LLM's cost estimates from the expert assessments, providing a clear measure of accuracy. The MAPE was chosen for its ability to express prediction accuracy as a percentage, which makes it easier to interpret and compare the results. Additionally, MAPE detects deviations between individual pairs of values, which is essential for this research. In contrast, other statistical techniques typically evaluate entire datasets when making comparisons. This research needed identifying difference or error between specific pairs of values, for instance, the LLM's estimation of direct losses from an attack and a security expert's corresponding estimation for the same attack.

3.3.2 Qualitative Analysis. To evaluate the qualitative data output from the LLM, a questionnaire was created. This questionnaire involved presenting participants with two options for each assessment: one from the LLM and one from the expert. The participants were be asked to select the better option based on clarity, relevance, and comprehensiveness. Additionally, participants were presented with the LLM's output alone to evaluate if it makes sense. This analysis aimed at assessing the structure, clarity and detect hallucinations rather then measure the accuracy of an output. The questionnaire included the following questions:

- Which assessment provides a clearer explanation of the impact of the cyber attack?
- Which assessment is more relevant to the context of the cyber attack?
- Which assessment is more comprehensive in detailing the impact?
- Does the LLM's output make sense in the context of the cyber attack?
- How would you rate the overall clarity of the LLM's output? (Scale 1-5)
- How relevant do you find the details provided in the LLM's output? (Scale 1-5)
- How comprehensive is the LLM's output in explaining the impact? (Scale 1-5)
- What specific improvements would you suggest for the LLM's output?

By comparing the LLM's outputs against expert assessments using the MAPE formula for deviations and a structured questionnaire for qualitative feedback, this study aims to extensively evaluate the performance and applicability of LLMs in the field of cybersecurity impact assessment.

4 LITERATURE REVIEW

The systematic review was performed to investigate what are currently existing frameworks and methods used for impact assessments, filter a set of well-known cyber attacks, determine a set of news articles which cover these cyber attacks, explore how LLMs are integrated in could be applied for impact assessments of the attacks, and provide insight into how to evaluate accuracy of the LLM.

4.1 Cyber attacks selection

The selection of cyber attacks for analysis was based on the credibility of the cost estimations in the referenced sources. Specifically, articles by [Internet Crime Complaint Center (IC3) 2023], [O. Kovalchuk et al. 2021], [N. Davies 2018], [Kunal et al. 2023], [CISA 2020] were used. These articles identify the most impactful cyber attacks, leading to the presumption that their cost estimations are accurate and reliable.

The table 2 includes part of the list of cyber attacks composed during the literature review. Additionally, the Table 2 preemptively includes costs associated with the selected framework which will be described next.

Company and Year	**Direct Losses ($ million)**	**Indirect Losses ($ million)**	**Defence Costs ($ million)**	**Criminal Revenue ($ million)**	**Individuals affected (million)**
Anthem (2015)	375.5	200	150	50	78.8
Yahoo (2014)	350	300	200	100	500
Merck (2017)	310	150	100	10	-
Target (2013)	292	250	150	30	70
Home Depot (2014)	252	200	100	25	56
Sony PlayStation (2011)	171	100	70	10	101.6
Equifax (2017)	164	200	120	40	145.5
Sony Pictures (2014)	43	70	30	5	0.047
Experian (2015)	20	50	15	5	15
Yahoo (2013)	16	400	50	150	1000
Ashley Madison (2015)	12.8	80	10	10	37
LinkedIn (2012)	4	30	10	1	6.5

Table 2. Costs of selected cyber attacks

4.2 Existing frameworks

- A framework proposed by [O. Kovalchuk et al. 2021] is based on an analysis of the main aspects of monetary

costs and the hidden economic impact of cybercrime. *A multifactor regression model* proposed by this articles aims to determine the contribution of the cost of the main consequences of IT incidents: business disruption, information loss, revenue loss and equipment damage caused by different types of cyberattacks worldwide in 2019 to total cost of cyberattacks. Information loss has been found to have a major impact on the total cost of cyberattacks, reducing profits and incurring additional costs for businesses.
- The systematic study by[Anderson et al. 2019] proposes a framework for analyzing cybercrime costs. They introduced several key concepts, including direct costs, indirect costs, and defense costs.
- Introduced by [Kotenko and Chechulin 2013], a comprehensive cyberattack modeling and impact assessment framework. Their work emphasized the importance of understanding attack vectors, vulnerabilities, and impacts.
- A framework by [Jones 2005] involves the identification of risk scenario, evaluation of loss event frequency and probable loss magnitude
- [Gowdanakatte et al. 2023]: Introduced a model-based risk assessment framework adapted for cyber-physical systems (CPS). Their framework emphasized the importance of modeling specific risks and attack paths unique to CPS.

A framework by [Anderson et al. 2019] was selected as target framework. It was the most referenced and used framework, most of the articles which described the selection of cyber attacks utilised this framework. Additionally, it suits well with the news articles analysis, while other frameworks were too technical or too specific, for instance focusing entirely on IOT devices. The metrics of direct losses, indirect losses, defense costs, and criminal revenue were chosen because the author highlights their importance in evaluating the financial impact of cyber attacks.

4.3 LLM integration

- [Jha et al. 2023] focus on iterative prompting architecture that uses formal methods to detect errors in the LLM response automatically.
- [Fysarakis et al. 2023] proposed PHOENI2X, a European Cyber Resilience Framework, focusing on artificial intelligence (AI)-assisted orchestration for incident response and recovery. This work highlighted the potential of integrating AI
- [Iyengar and Kundu 2023] explored the application of LLMs in cybersecurity, providing a detailed analysis of how these models could be used for impact assessments and threat identification.
- [Li and Shan 2023] introduced vulnerability detection using LLMs, providing insights into how models like GPT-4 can be used to identify potential cyber threats based on textual analysis.
- [Patel et al. 2024] developed CANAL, a news alerting language model. This model was designed specifically to categorize cyberattacks using news articles, which aligns with MALICE project.
- [Zhou et al. 2023] provides a summary of the current state and progress of large language model applications in professional settings and explores evaluation methods for such models.
- [Abdullah et al. 2018] proposed a scheme on detecting the related news about cyber-attacks and outlined important features for classification such as: threat type, threat name, threat
- [Gururangan et al. 2020] provide insights on pretraining and fine-tuning LLMs

These studies provided invaluable insights into the integration of LLMs, fine-tuning processes, and prompt engineering techniques, which will be applied in subsequent stages of this research.

5 EXPERIMENT

5.1 Data Cleaning and Preprocessing

For the selection of cyber attacks collected and described in the literature review section, relevant articles were selected and added to the dataset. These articles were sourced from the dataset compiled by [McCandless et al. 2024] with an addition of articles from Google News.

For instance, in the case of the cyber attack on Yahoo, which occurred in 2013 but was only disclosed in 2017, three articles in total were added to the dataset. Two articles were taken from the [McCandless et al. 2024] dataset: one from [B.B.C 2017] and another from [CNBC 2017]. Additionally, an article from [Perlroth 2017] was retrieved via Google News to provide more context and detail.

As a result, the data entry for the 2013 cyber attack on Yahoo includes the following information:

- Company Affected and Year: Yahoo (2013)
- Direct Losses (in $ million): 350
- Indirect losses (in $ million): 300
- Criminal Revenue (in $ million): 150
- Defense Costs (in $ million): 50
- Number of Individuals Affected (in million): 3000
- Source 1: "At least 500 million user accounts have been stolen from Yahoo, the company confirmed on Thursday...."
- Source 2: "Yahoo has said that all of its three billion user accounts were affected in a hacking attack dating back to 2013. The company, which was taken over by Verizon...."
- Source 3: "Verizon Communications, said on Tuesday that a previously disclosed attack that had occurred in 2013 affected all three billion ...."

The process was repeated for the rest of the dataset.

5.2 Prompt Engineering

To enhance LLM's abilities in categorisation of costs, prompts were iteratively refined.

5.2.1 Prompt 1. The first version of the prompt included only basic instructions and simply asks an LLM to provide a calculation of the costs.

Prompt 1

Here is a news article on a recent cyber attack:
[News article]
Based on the information provided in the article, calculate the following:

1. Direct Losses
2. Criminal Revenue
3. Defence Costs
4. Cost to Society
5. Number of Individuals Affected

Additionally, provide a brief evaluation or impact assessment of the cyber attack in the article.

5.2.2 Prompt 2. To improve clarity, the second prompt incorporated common prompt engineering techniques. These included establishing a clear and organized structure, adopting a specific persona to provide context and focus, and supplying precise definitions for key categories.

Prompt 2

You are a security specialist tasked with analyzing the costs and impacts of cyber attacks. Here is a news article on a recent cyber attack:
[News Article]
Based on the information provided in the article, please perform the following evaluations:

1. Estimate the Direct Losses incurred by the cyber attack. This is the monetary equivalent of losses, damage, or other suffering felt by the victim as a consequence of a cybercrime
2. Estimate the Indirect losses incurred by the cyber attack. This is the monetary equivalent of the losses and opportunity costs imposed on society by the fact that a certain cybercrime is carried out, no matter whether successful or not and independent of a specific instance of that cybercrime.
3. Determine the Defence Costs associated with responding to the cyber attack. They include the cost of development, deployment, and maintenance of prevention measures, and inconvenience and opportunity costs caused by the prevention measures.
4. Estimate the Criminal Revenue generated from the cyber attack. This is the monetary equivalent of the gross receipts from a crime.
5. Identify the Number of Individuals Affected by the cyber attack.

Finally, provide a brief evaluation or impact assessment of the cyber attack, considering its severity, scope, and long-term implications.

5.2.3 Prompt 3. To further assist the LLM, examples for each category were added to enhance it's ability to recognise different costs even better. Additionally, clear labels are introduced together with notes.

Prompt 3

You are a security specialist tasked with ...
...in the article perform the following tasks:

1. **Estimate Direct Losses**: Direct losses include the monetary equivalent of losses, damage, or other suffering felt by the victim as a consequence of a cybercrime. For example, money withdrawn from victim account; time and effort to reset account credentials (for banks and consumers); distress suffered by victims; secondary costs of overdrawn accounts or deferred purchases, inconvenience of not having access to money when needed; lost attention and bandwidth caused by spam messages, even if they are not reacted to.
2. **Estimate the Indirect losses**: Indirect losses include the monetary equivalent of the losses and opportunity costs imposed on society by the fact that a certain cybercrime is carried out, no matter whether successful or not and independent of a specific instance of that cybercrime. For example, Loss of trust in online banking, leading to reduced revenues from electronic transaction fees, and higher costs for maintaining branch staff and cheque clearing facilities; missed business opportunity for banks to communicate with their customers by email; reduced uptake by citizens of electronic services as a result of lessened trust in online transactions; efforts to clean-up PCs infected with malware for a spam sending botnet.
3. **Estimate Defence Costs**: Defence costs include the cost of development, deployment, and maintenance of prevention measures, and inconvenience and opportunity costs caused by the prevention measures. For example, security products such as spam filters, antivirus, and browser extensions to protect users; security services provided to individuals, such as training and awareness measures; security services provided to industry, such as website take-down services; fraud detection, tracking, and recuperation efforts; the inconvenience of missing messages falsely classified as spam
4. **Estimate Criminal Revenue**: Criminal revenue is the monetary equivalent of the gross receipts from a crime. For example, if phishing is advertised by email spam, the phisherman's criminal revenue is the sum of the money withdrawn from victim accounts.
5. **Identify Number of Individuals Affected**: For example, if the article states that 500,000 accounts were stolen, report that number.

After performing these calculations, provide a **brief evaluation or impact assessment** of the cyber attack. Discuss its severity, scope, and potential long-term implications.
Important: Ensure that each calculation is clearly labeled and explained. Use specific details and data from the article to support your analysis.

5.3 Fine tuning

An example cost estimation was used in fine-tuning the LLM. The Anthem 2015 breach was used as training data due to availability of the most cost estimations and news articles, providing more information.

The 2015 cyber attack on Anthem includes the following information:

- Company Affected and Year: Anthem (2015)
- Direct Losses (in $ million): 375.5
- Inirect Losses (in $ million): 200
- Criminal Revenue (in $ million): 50
- Defense Costs (in $ million): 150
- Number of Individuals Affected (in million): 78.8
- Source 1: "Massive breach at health care company Anthem Inc...."[8]
- Source 2: "Health Insurer Anthem Struck By Massive Data Breach..."[9]
- Source 3: "Millions of Anthem Customers Targeted in Cyberattack..." [10]

6 RESULTS AND DISCUSSION

6.1 Qualitative analysis

Direct losses is the most straightforward of the cost metrics, therefore it has the least percentage error of all as reported in the Table 3. Number of affected individuals is frequently highlighted in news articles, often appearing in the headlines, resulting in a negligible MAPE for this metric. Consequently, this metrics was not considered when calculating the total MAPE. However, it highlights LLMs abilities in extracting information which is clearly available. Conversely, LLM had significant challenges with identifying the criminal revenue even when few-shots were added. This is likely due to limited availability of information on criminal revenue in news articles and often the lowest value of this metric among others.

The following metrics for MAPE by [Hyndman and Koehler 2006] will be used to evaluate accuracy of LLM's output.

- Excellent: MAPE < 10%
- Good: 10% ≤ MAPE < 20%
- Reasonable: 20% ≤ MAPE < 50%
- Poor: MAPE ≥ 50%

Table 3 clearly shows that the LLMs ability is affected by the implementation techniques.

Inspecting the first prompt, the MAPE for calculable results fell into the poor category. Additionally the results were so inaccurate that the MAPE could not be calculated, as LLM sometimes gave unanalyzable output. For instance: "A substantial amount considering number of records sold" was one of the outputs for criminal revenue.

Prompt engineering techniques together with fine-tuning achieved remarkable results, bringing the total MAPE to 23.3% and resolving the issue with LLM's hallucinations.

Additional refinements to the prompts continued to enhance performance. While performance in the direct losses category remained largely unchanged, MAPE for other categories dropped down by ≈ 8.6%. indicating notable improvements.

It is important to note that mostly, LLM's error was due to incorrectly categorising a value mentioned in a news article, which had a doubled impact on the error as some value was removed from a category and simultaneously added to another. Fine-tuning and prompt engineering helped with miscategorization but didn't completely solve the problem. For instance, in one of the outputs of the third prompt LLM categorized the external consulting costs, which were identified as direct losses by security experts [CISA 2020], as defense costs. This misclassification resulted in additional error both in defence costs and direct losses category. Direct cost estimation of an LLM is available in Appendix A, Estimation of costs by [CISA 2020] is available in Appendix B.

Overall, the results show great potential in LLM implementation techniques. Additionally, this research managed to bring average error of GPT-4 LLM down to 15.6% which is considered to be good estimation.

Prompt	Direct Losses	Indirect Losses	Defence Costs	Criminal Revenue	Individuals affected	Indirect Losses
Prompt 1	32%	N/A	81.7%	67.6%	<0.1%	N/A
FS Prompt 2	10.6%	18.6%	31.1%	27.3%	<0.1%	23.3%
FS Prompt 3	10.4%	14.8%	19.2%	17%	<0.1%	15.6%

Table 3. Calculated MAEP comparison across different categories

6.2 Quantitative analysis

A total of 116 participants in total completed the survey.

According to the Figure 1 assessments made by LM (language model) were considered to be more detailed in comparison to SE (security experts) by vast majority of the participants. This could be due to the fact that the assessments made by LLM simply had at least twice as more text and were thorough in every aspect.

[8] https://eu.usatoday.com/story/tech/2015/02/04/health-care-anthem-hacked/22900925/

[9] https://www.forbes.com/sites/gregorymcneal/2015/02/04/massive-data-breach-at-health-insurer-anthem-reveals-social-security-numbers-and-more/

[10] https://www.nytimes.com/2015/02/05/business/hackers-breached-data-of-millions-insurer-says.html

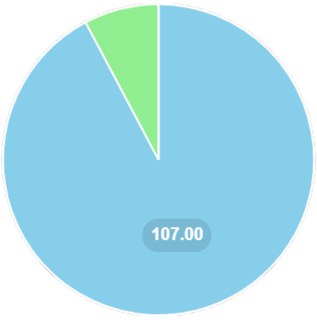


Fig. 1. Participants' responses on preference of the level of detail in LLM (Blue) and expert assessments (Green); There is a significant preference for the detailed explanations provided by LLM

On the other hand, Figure 2 shows that source of the impact assessment has almost no impact on it's relevance according to the participants.

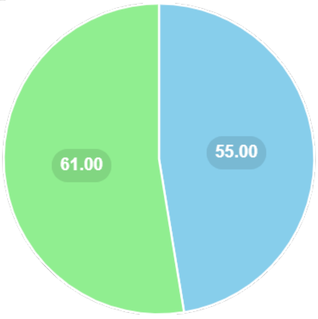


Fig. 2. Participants' responses on preference of the relevance in LLM (Blue) and expert assessments (Green). Both sources are considered equally relevant by participants

Additionally, participants generaly rated LLM's structure to be superior as can be seen on Figure 3. LLM's output often included well-organised sections and subsections clearly outlining the costs associated with each category and detailing the calculations involved.

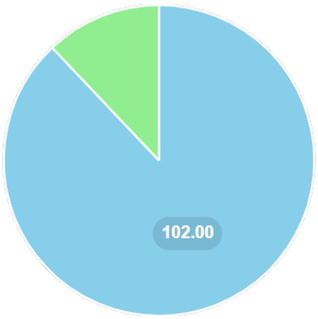
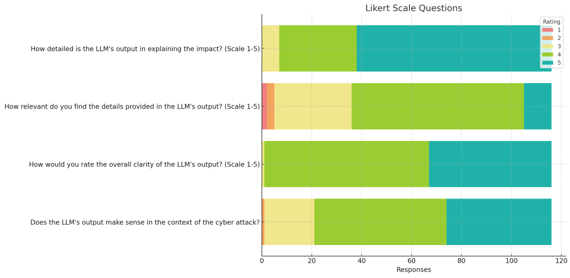


Fig. 3. Participants' perceptions of clarity in LLM (Blue) and expert assessments (Green). The chart shows a significant preference for the structured responses of the LLM

The results of a Likert scale questions on Figure clearly show that participants were generally satisfied with the quality of LLM's output. Negative responses were minimal and neutral responses were negligible for majority of the questions except for those that asked relevance and coherence of the assessment. However, due to further examination of participant's comments in open questions, it was evident that these participants were dissatisfied the volume of text generated by the LLM.



4

Fig. 4. Likert scale questions Stacked Bar Chart showing satisfaction with the LLM's responses

6.3 Reflection on results

Despite the promising results, the study also highlighted several challenges. One significant issue was the LLM's tendency to misclassify certain costs, which affected the overall accuracy. While prompt engineering and fine-tuning mitigated some of these errors, further refinement is necessary to fully address this limitation. However, it's important to not add too much information to the prompts as was mentioned in the articles reviewed in literature review section due to LLM's tendency to include redundant data when prompts require diverse output.

Additionally, the reliance on the availability and quality of news articles presents a constraint, as not all cyber attacks are reported comprehensively in the media. The reliance on news articles as the primary data source introduced variability in the data quality, which occasionally affected the LLM's performance. The use of news articles was necessitated by the unavailability of detailed information required for more accurate impact assessments, as such information is often sensitive and not revealed to the public.

Furthermore, the prompt engineering and fine-tuning processes, although effective, were time-consuming and required significant manual intervention. Future research should explore more automated and scalable approaches to LLM fine-tuning and prompt engineering.

Furthermore, the quantitative analysis using MAPE provided a clear measure of accuracy, but it not grasp full context of the data. The qualitative analysis partially addressed this gap, but a more integrated approach could yield better insights.

7 CONCLUSION

7.1 Key findings

The potential of LLMs in text analysis, specifically in impact estimation based on news articles is evident and clear. The LLM managed to extract costs categorised by commonly

used impact assessment framework. These estimations were within acceptable range of error from the baseline cost estimations made by experts for research. Certain metrics such as number of individuals affected were straightforward enough for the LLM to almost make no errors, as the information was often clearly, but reliability is still impressive. The LLM's output was rated as superior in terms of structure and level of detail. Clarity of the ouput is also acceptable. Additionally, it is important to remember that LLMs' the capability in interpreting extensive datasets quickly makes them a valuable tool in cyber attack impact assessments.
However, these results were achieved by adopting common LLM implementation techniques such as fine-tuning and prompt engineering. Initial results revealed that the LLMs abilities in impact estimations are limited, and it's not viable to use this novel method for these purposes. Nevertheless, after the refinement process improved significantly, achieving a MAPE of 15.6%.This highlights the critical importance of these techniques.
Overall, the paper clearly demonstrates LLM's promising abilities in cyber attack impact estimation. However, further research needs to be performed to validate the effectiveness of LLMs in impact assessments.

7.2 Future research

In order to further inspect the viability of application of LLMs in cyber attack impact estimation, several directions for future research are proposed:

- Exploration of More Models: This study focused on a specific LLM: GPT-4, but future research should evaluate a variety of models, such as BERT and T5. Comparative studies will help identify the most suitable models for different types of cyber threats and data sources.
- Incorporation of More Frameworks: Integrating LLMs with additional cybersecurity frameworks can provide a more elaborate evaluation. Research should explore the compatibility of LLMs with a broader range of frameworks to identify the most effective combinations.
- Utilization of New Datasets: This study only focused on news articles due to their richness in text data and availability. Future studies should try diverse datasets which improve the generalizability of the proposed methods across different contexts.
- Power Consumption: One of the disadvantages of LLMs is that they require substantial computational resources. Due to the importance of optimizing power consumption for sustainability, this aspect needs to be investigated.

REFERENCES

Mohamad Syahir Abdullah, Anazida Zainal, Mohd Aizaini Maarof, and Mohamad Nizam Kassim. 2018. Cyber-Attack Features for Detecting Cyber Threat Incidents from Online News. In *2018 Cyber Resilience Conference (CRC)*. 1–4. https://doi.org/10.1109/CR.2018.8626866

R Anderson, C Barton, R Bohme, R Clayton, C Gañán, T Grasso, Michael Levi, T Moore, and Marie Vasek. 2019. Measuring the changing cost of cybercrime.

News B.B.C. 2017. Yahoo 2013 data breach hit 'all three billion accounts'. *BBC News* (Oct. 2017). https://www.bbc.com/news/business-41493494

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. https://arxiv.org/abs/2005.14165 _eprint: 2005.14165.

CISA. 2020. *Cost of a Cyber Incident: Systematic Review and Cross-Validation.* Technical Report. Cybersecurity and Infrastructure Security Agency (CISA).

CNBC. 2017. Yahoo data breach is among the biggest in history. *CNBC* (2017). https://www.cnbc.com/2016/09/22/yahoo-data-breach-is-among-the-biggest-in-history.html

Konstantinos Fysarakis, Alexios Lekidis, Vasileios Mavroeidis, Konstantinos Lampropoulos, George Lyberopoulos, Ignasi Garcia-Milà Vidal, José Carles Terés i Casals, Eva Rodriguez Luna, Alejandro Antonio Moreno Sancho, Antonios Mavrelos, Marinos Tsantekidis, Sebastian Pape, Argyro Chatzopoulou, Christina Nanou, George Drivas, Vangelis Photiou, George Spanoudakis, and Odysseas Koufopavlou. 2023. PHOENI2X – A European Cyber Resilience Framework With Artificial-Intelligence-Assisted Orchestration, Automation & Response Capabilities for Business Continuity and Recovery, Incident Response, and Information Exchange. In *2023 IEEE International Conference on Cyber Security and Resilience (CSR)*. 538–545. https://doi.org/10.1109/CSR57506.2023.10224995

Shwetha Gowdanakatte, Indrakshi Ray, and Mahmoud Abdelgawad. 2023. Model Based Risk Assessment and Risk Mitigation Framework for Cyber-Physical Systems. In *2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*. 203–212. https://doi.org/10.1109/TPS-ISA58951.2023.00034

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *CoRR* abs/2004.10964 (2020). https://arxiv.org/abs/2004.10964 arXiv: 2004.10964.

Rob J. Hyndman and Anne B. Koehler. 2006. Another look at measures of forecast accuracy. *International Journal of Forecasting* 22, 4 (2006), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Internet Crime Complaint Center (IC3). 2023. 2023 Internet Crime Report. https://www.ic3.gov/Media/PDF/AnnualReport/2023_IC3Report.pdf

Arun Iyengar and Ashish Kundu. 2023. Large Language Models and Computer Security. In *2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*. 307–313. https://doi.org/10.1109/TPS-ISA58951.2023.00045

Susmit Jha, Sumit Kumar Jha, Patrick Lincoln, Nathaniel D. Bastian, Alvaro Velasquez, and Sandeep Neema. 2023. Dehallucinating Large Language Models Using Formal Methods Guided Iterative Prompting. In *2023 IEEE International Conference on Assured Autonomy (ICAA)*. 149–152. https://doi.org/10.1109/ICAA58325.2023.00029

Jack A. Jones. 2005. An Introduction to Factor Analysis of Information Risk (FAIR). (2005). Publisher: Norwich University.

Igor Kotenko and Andrey Chechulin. 2013. A Cyber Attack Modeling and Impact Assessment framework. In *2013 5th International Conference on Cyber Conflict (CYCON 2013)*. 1–24.

Kunal, Muskaan Rana, Diya Sharma, and Anurag. 2023. Understanding Cyber-Attacks and their Impact on Global Financial Landscape. In *2023 International Conference on Circuit Power and Computing Technologies (ICCPCT)*. 1452–1456. https://doi.org/10.1109/ICCPCT58313.2023.10245828

Hongping Li and Li Shan. 2023. LLM-based Vulnerability Detection. In *2023 International Conference on Human-Centered Cognitive Systems (HCCS)*. 1–4. https://doi.org/10.1109/HCCS59561.2023.10452613

David McCandless, Tom Evans, and Paul Barton. 2024. World's Biggest Data Breaches & Hacks. https://docs.google.com/spreadsheets/d/1i0oIJJMRG-7t1GT-mr4smaTTU7988yXVz8nPlwaJ8Xk/edit?gid=2#gid=2

J. Milne N. Davies, A. S. Dawson. 2018. *Understanding the Costs of Cyber Crime: A Report of the Home Office.* Technical Report Horr 96. Home Office. https://assets.publishing.service.gov.uk/media/5a82d166e5274a2e8ab59814/understanding-costs-of-cyber-crime-horr96.pdf

O. Kovalchuk, M. Shynkaryk, and M. Masonkova. 2021. Econometric Models for Estimating the Financial Effect of Cybercrimes. In *2021 11th International Conference on Advanced Computer Information Technologies (ACIT)*. 381–384. https://doi.org/10.1109/ACIT52158.2021.9548490 Journal Abbreviation: 2021 11th International Conference on Advanced Computer Information Technologies (ACIT).

Urjitkumar Patel, Fang-Chun Yeh, and Chinmay Gondhalekar. 2024. CANAL - Cyber Activity News Alerting Language Model : Empirical Approach vs. Expensive LLMs. In *2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC)*. 1–12. https://doi.org/10.1109/ICAIC60265.2024.10433839

Nicole Perlroth. 2017. Yahoo says 3 billion accounts were affected by 2013 attack. *The New York Times* (Oct. 2017). https://www.nytimes.com/2017/10/03/technology/yahoo-hack-3-billion-users.html

Alberto D. Rodriguez, Katherine R. Dearstyne, and Jane Cleland-Huang. 2023. Prompts Matter: Insights and Strategies for Prompt Engineering in Automated Software Traceability. In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*. 455–464. https://doi.org/10.1109/REW57809.2023.00087 ISSN: 2770-6834.

Lena Yuryna Connolly, David S Wall, Michael Lang, and Bruce Oddson. 2020. An empirical study of ransomware attacks on organizations: an assessment of severity and salient factors affecting vulnerability. *Journal of Cybersecurity* 6, 1 (Dec. 2020), tyaa023. https://doi.org/10.1093/cybsec/tyaa023 _eprint: https://academic.oup.com/cybersecurity/article-pdf/6/1/tyaa023/35198032/tyaa023.pdf.

J. D. Zamfirescu-Pereira, Richmond Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 21. https://doi.org/10.1145/3544548.3581388 event-place: Hamburg, Germany.

Mingji Zhou, Wei Chen, Senliang Zhu, Tianyang Cai, Ji Yu, and Guoyu Dai. 2023. Application of large language models in professional fields. In *2023 11th International Conference on Information Systems and Computing Technology (ISCTech)*. 142–146. https://doi.org/10.1109/ISCTech60480.2023.00033

A EXAMPLE LLM'S IMPACT ASSESSMENT

Estimate Direct Losses

Direct losses for the Equifax cyber attack include:

- **Monetary losses:** Equifax spent $55.5 million for credit monitoring.
- **Distress and inconvenience:** The inconvenience to users had to be compensated which costed Equifax $14.9 million.

Quantifying these:
$55500000 + $1490000 = $70,400,000 = $70.4 million

The total direct losses are estimated to be $70.4 million.

B EXAMPLE SECURITY EXPERT IMPACT ASSESSMENT

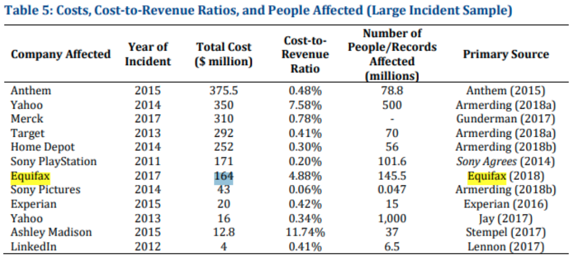
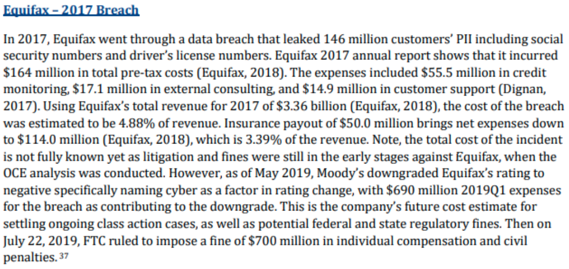


Table 5: Costs, Cost-to-Revenue Ratios, and People Affected (Large Incident Sample)

Company Affected	Year of Incident	Total Cost ($ million)	Cost-to-Revenue Ratio	Number of People/Records Affected (millions)	Primary Source
Anthem	2015	375.5	0.48%	78.8	Anthem (2015)
Yahoo	2014	350	7.58%	500	Armerding (2018a)
Merck	2017	310	0.78%	-	Gunderman (2017)
Target	2013	292	0.41%	70	Armerding (2018a)
Home Depot	2014	252	0.30%	56	Armerding (2018b)
Sony PlayStation	2011	171	0.20%	101.6	*Sony Agrees* (2014)
Equifax	2017	164	4.88%	145.5	Equifax (2018)
Sony Pictures	2014	43	0.06%	0.047	Armerding (2018b)
Experian	2015	20	0.42%	15	Experian (2016)
Yahoo	2013	16	0.34%	1,000	Jay (2017)
Ashley Madison	2015	12.8	11.74%	37	Stempel (2017)
LinkedIn	2012	4	0.41%	6.5	Lennon (2017)

Table 4. Example direct cost estimation by [CISA 2020]



Equifax – 2017 Breach

In 2017, Equifax went through a data breach that leaked 146 million customers' PII including social security numbers and driver's license numbers. Equifax 2017 annual report shows that it incurred $164 million in total pre-tax costs (Equifax, 2018). The expenses included $55.5 million in credit monitoring, $17.1 million in external consulting, and $14.9 million in customer support (Dignan, 2017). Using Equifax's total revenue for 2017 of $3.36 billion (Equifax, 2018), the cost of the breach was estimated to be 4.88% of revenue. Insurance payout of $50.0 million brings net expenses down to $114.0 million (Equifax, 2018), which is 3.39% of the revenue. Note, the total cost of the incident is not fully known yet as litigation and fines were still in the early stages against Equifax, when the OCE analysis was conducted. However, as of May 2019, Moody's downgraded Equifax's rating to negative specifically naming cyber as a factor in rating change, with $690 million 2019Q1 expenses for the breach as contributing to the downgrade. This is the company's future cost estimate for settling ongoing class action cases, as well as potential federal and state regulatory fines. Then on July 22, 2019, FTC ruled to impose a fine of $700 million in individual compensation and civil penalties.[37]

Fig. 5. Example impact assessment of Equifax(2017) cyber attack by [CISA 2020]