The Dual-Edged Sword: The Impact of Large Language Models in Network Infrastructure Security

David Debono and Anastasia Sare
Institute of Information and Communication Technology,
Malta College of Arts, Science and Technology, Triq Kordin, Paola PLA9032, Malta

Keywords: Cybersecurity, Large Language Model, GPT-3.5, GPT-4, Microsoft Copilot, Gemini, CIS Benchmark, MySQL, MongoDB.

Abstract: Large Language Models (LLMs) have become essential tools for network infrastructure and security engineers, assisting in a wide range of daily administrative tasks. However, the widespread use of these models without adequate cybersecurity expertise could potentially compromise network security. This study examines the compliance of various LLMs, including GPT-3.5, GPT-4, Microsoft Copilot, and Gemini, with CIS benchmarks. We evaluate the capabilities and limitations of these models in adhering to MySQL and MongoDB CIS benchmarks on a Linux system using both qualitative and quantitative metrics. Four distinct test cases were developed to assess the performance of GPT-3.5 and GPT-4. The first test evaluated the models' compliance and knowledge of security standards without explicitly mentioning the standards. The second test assessed the models' zero-shot knowledge when CIS benchmarks were explicitly referenced, while the third test examined the effectiveness of follow-up prompts based on the results of the second test. In the fourth test, GPT-4 was provided with the actual standard in PDF format. Additionally, the zero-shot capabilities of Gemini and Microsoft Copilot were also evaluated. Among the models tested, GPT-4 demonstrated the highest compliance with CIS benchmarks, particularly in zero-shot learning and assisted scenarios. However, challenges were noted with certain configurations, and the use of prompt engineering techniques proved crucial in maximizing compliance. With a maximum score of 76.3% compliance, the findings suggest that while LLMs can assist in providing secure configurations aligned with international standards, expert knowledge and supervision remain essential to mitigate potential vulnerabilities.

1 INTRODUCTION

Generative AI models, such as Generative Pretrained Transformer (GPT) models, have transformed communication and automation across industries. These models can generate text, music, visual art, and realistic images. In business, they streamline operations by drafting emails, generating reports, and creating marketing materials. Beyond creativity, generative AI enhances automation, with its full potential still being explored (Beheshti, 2023).

The integration of AI into information technology has led to intelligent, cost-effective solutions across various fields, particularly in cybersecurity, where AI addresses security and privacy challenges (Sarker et al., 2021).

As highlighted by Shanthi et al. (2023), early AI research in cybersecurity, began with expert systems for intrusion detection and incident response. Over time, machine learning techniques, like neural networks, were incorporated (*A New Era of Cybersecurity: The Influence of Artificial Intelligence*, 2023). By the 2000s, AI advanced to generating intrusion detection rules and identifying unknown threats (Sharma & Dash, 2023).

Today, AI enhances threat detection, vulnerability management, and security automation, with real-time threat detection and behavioral analysis revolutionising the cybersecurity domain (Sarker et al., 2021).

While AI integration in cybersecurity offers many benefits, it also presents challenges. Key issues include the need for high-quality, representative data, which is difficult to obtain. Additionally, the complexity of AI systems raises concerns about their interpretability and decision-making transparency. Other challenges include biased training data, vulnerability to adversarial attacks, and a lack of

Debono, D. and Sare, A.
The Dual-Edged Sword: The Impact of Large Language Models in Network Infrastructure Security.
DOI: 10.5220/0013090700003899

In *Proceedings of the 11th International Conference on Information Systems Security and Privacy (ICISSP 2025) - Volume 2*, pages 125-132
ISBN: 978-989-758-735-1; ISSN: 2184-4356

comprehensive regulatory frameworks (Shanthi et al., 2023).

AI's double-edged nature presents risks alongside its benefits. While advanced AI platforms like ChatGPT enhance cybersecurity, they can also be weaponized for sophisticated attacks (Sharma & Dash, 2023). This highlights the need for caution, as AI's computational power could introduce new challenges and reshape the threat landscape. Integrating AI into cybersecurity might not only amplify existing threats but also introduce new ones, requiring a balanced approach to leverage AI's potential while mitigating its risks (Al-Hawawreh et al., 2023).

This study investigates the effectiveness of Large Language Models (LLMs) in deploying network infrastructure resilient to cybersecurity threats. Specifically, it aims to evaluate the accuracy of LLMs, including GPT-3.5, GPT-4, Microsoft (MS) Copilot, and Gemini, in recommending secure configurations that adhere to CIS benchmarks for database deployment. We focus on the secure deployment of MySQL and MongoDB databases on Linux systems. Our contributions are as follows:

- We introduce an evaluation framework for assessing LLM compliance with CIS benchmarks across various scenarios.
- We analyze the performance of the aforementioned LLMs in recommending secure database configurations that meet industry-standard CIS benchmarks.
- We present prompt engineering techniques designed to improve the quality of LLM responses related to security benchmarks.

This paper is organized as follows: Section 2 presents related work. Section 3 explains the methodology while Section 4 discusses the data analysis and results. Conclusions and future work are presented in Section 5 and 6 respectively.

2 RELATED WORK

Generative AI, such as ChatGPT, has a profound impact on cybersecurity. Generative AI can be misused for cyberattacks, such as writing malicious code, making it harder to detect than other classic tools. Media-based AI, like image and video generation, facilitates phishing, identity theft, and deepfake fraud, contributing to rising fraud statistics. Code-based AI lowers the barrier for cyberattacks, allowing even novices to create advanced hacking tools and bypass security systems. The combination of these AI types enables more sophisticated attacks. Generative AI introduces significant cybersecurity risks, necessitating stronger security measures, governance, and education to ensure ethical use (Oh & Shon, 2023).

However, the technology is also being exploited as a counter measure to the advancements in complexity of cyberattacks. Al-Hawawreh et al. (2023) evaluated GPT-3.5's performance in cybersecurity by deploying it in a dynamic honeypot environment to simulate real-world threats. The model interacted with various attack vectors, including phishing, malware injections, SQL injection, denial-of-service, man-in-the-middle, and brute force attacks. Through this setup, they gathered both qualitative interaction logs and quantitative data on detection accuracy and response times. This mixed-methods approach allowed them to assess GPT-3.5's strengths and limitations in identifying and mitigating cyber threats effectively.

Sobania et al. (2023) assessed GPT-3.5's bug-fixing abilities using the QuixBugs benchmark, comparing it to tools like CoCoNut and Codex. They provided incorrect code to ChatGPT and manually verified its fixes. The evaluation used ChatGPT versions from December 15, 2022, and January 9, 2023, applying a generate-and-validate approach. This method emphasized the need for iterative interactions to enhance fix success rates due to performance variability. Similar variability in performance was also observed by Gupta et al. (2023).

Gupta et al. (2023) evaluated the cybersecurity capabilities of GPT-3.5, BERT, and T5 through simulated cyber-attacks like phishing, malware, data breaches, and network intrusions. Scenarios tested their ability to detect and respond to threats, varying in complexity and sophistication. LLMs were assessed on identifying phishing emails, detecting malware in legitimate-looking files, preventing data breaches, and identifying network intrusions. The study highlighted the importance of quality in the training data, with more complex scenarios showing higher variability in performance. The authors emphasized the need for robust datasets and adaptive learning techniques to improve LLMs' reliability in cybersecurity.

Ali and Kostakos (2023) introduced HuntGPT, an intrusion detection dashboard that integrates machine learning and explainable AI (XAI) to enhance cybersecurity. Using a Random Forest classifier trained on the KDD99 dataset, the system employed XAI frameworks like SHAP and LIME to provide clear, interpretable insights into detected anomalies.

The study evaluated the system's technical accuracy using questions from ISACA and CISM exams and assessed the readability of the responses using six different metrics. Their approach demonstrates how combining LLMs with XAI can improve cybersecurity tools by making threat detection outputs more understandable and actionable for analysts.

Pearce et al. (2023) studied the zero-shot vulnerability repair capabilities of GPT-3.5-Turbo and Codex, focusing on their generalization from limited data across diverse scenarios. The scenarios included real-world vulnerabilities like buffer overflows, SQL injections, cross-site scripting, and misconfigurations. Each scenario tested the LLMs' ability to understand vulnerability context, generate appropriate repair suggestions, and follow best cybersecurity practices. They used multiple iterations with varying prompt phrasings, highlighting the importance of precise prompt engineering for optimal performance. The study evaluated the LLMs using metrics like accuracy, completeness, and relevance of repair suggestions, demonstrating the impact of prompt design on their effectiveness in vulnerability repair.

The network security landscape relies on standards and benchmarks to guide organizations in protecting their digital infrastructure. Key frameworks include the Centre for Internet Security (CIS) benchmarks, which provide best practices for securing IT systems, and ISO/IEC 27001, which offers a framework for information security management systems (ISMS) (*ISO/IEC 27001:2013*, n.d.). The NIST Cybersecurity Framework is also widely used to manage and reduce cybersecurity risks. However, rapidly evolving cyber threats often outpace these standards, creating challenges for organizations, especially smaller ones, in keeping up with compliance demands (Gupta et al., 2023). Generative AI, like ChatGPT, can help organizations adapt quicker to changes, but human oversight remains crucial to ensure ethical and secure implementations. Additionally, AI systems themselves must be safeguarded against potential cyber threats (Zeadally et al., 2020).

3 METHODOLOGY

For this research, GPT-3.5, GPT-4, MS Copilot, and Gemini were used to assess how well these LLMs adhered to MySQL (CIS Oracle, n.d.) and MongoDB (CIS MongoDB, n.d.) CIS benchmarks. For this study the benchmark for MySQL Community Server 5.7 v2.0 and MongoDB 5 v1.1, were used. The research pipeline depicted in Figure 1 was utilized for this study.

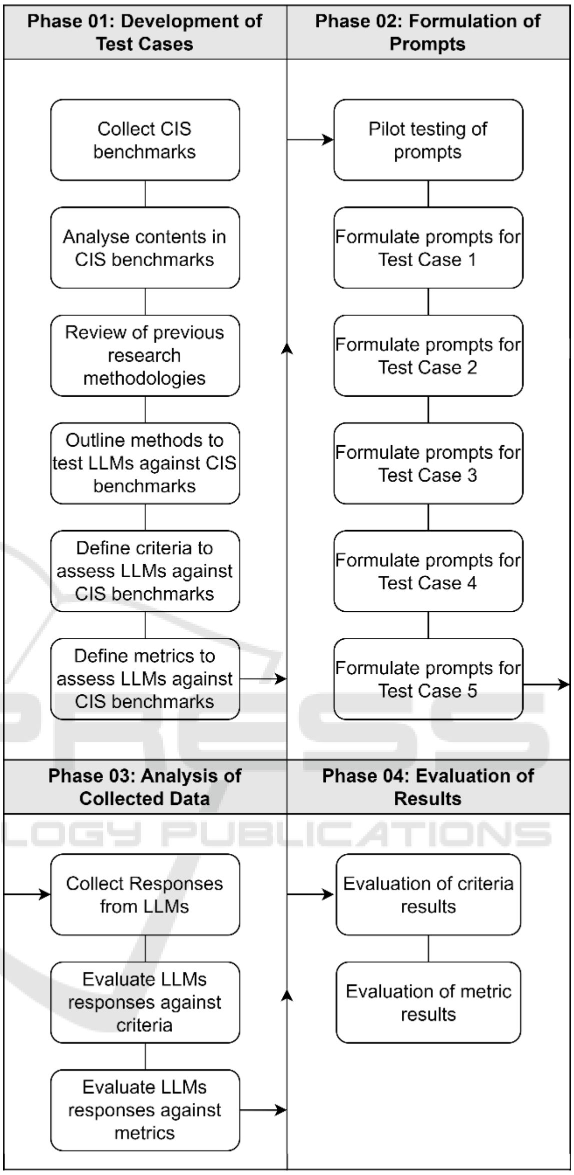
Phase 01: Development of Test Cases
Phase 02: Formulation of Prompts
Collect CIS benchmarks
Analyse contents in CIS benchmarks
Review of previous research methodologies
Outline methods to test LLMs against CIS benchmarks
Define criteria to assess LLMs against CIS benchmarks
Define metrics to assess LLMs against CIS benchmarks
Pilot testing of prompts
Formulate prompts for Test Case 1
Formulate prompts for Test Case 2
Formulate prompts for Test Case 3
Formulate prompts for Test Case 4
Formulate prompts for Test Case 5
Phase 03: Analysis of Collected Data
Phase 04: Evaluation of Results
Collect Responses from LLMs
Evaluate LLMs responses against criteria
Evaluate LLMs responses against metrics
Evaluation of criteria results
Evaluation of metric results

Figure 1: Research Pipeline.

3.1 Case 1: Knowledge Evaluation

This test aimed to assess GPT-3.5 and GPT-4's adherence to CIS benchmarks and their knowledge of database configuration best practices. The models were asked general questions related to database configuration, without explicitly mentioning the benchmarks, as suggested by Al-Hawawreh et al. (2023), to explore ChatGPT's potential in suggesting secure configurations. The recommendations were compared to CIS benchmarks to identify gaps based on three evaluation criteria as detailed in Table 1.

Table 1: Evaluation criteria and weights.

Criteria	Weight
Accuracy - Does the response correctly identify and explain the best practices in accordance with the CIS benchmarks?	40%
Relevance - Is the information provided directly relevant to the query?	30%
Actionability - Does the response provide practical, actionable advice?	30%

The CIS benchmarks for MySQL (*CIS Oracle MySQL Benchmarks*, n.d.) and MongoDB (*CIS MongoDB Benchmarks*, n.d.) for the Linux Environment are divided into nine and seven categories respectively as depicted in Table 2.

Table 2: CIS Benchmark Categories.

MySQL	MongoDB
OS Configuration	Installation and Patching
Installation and Planning	Authentication
File Permissions	Authorisation
General	Data Encryption
Permissions	Audit Logging
Audit and Logging	OS Hardening
Authentication	File Permission
Network	
Replication	

For each category a prompt was crafted. For example, to check the MySQL category for File Permissions, a prompt of the form "Describe the best practices and recommendations related to file permissions for MySQL Community Server", was used. All prompts followed a similar pattern. The answers were qualitatively evaluated using the criteria in Table 1 and the rating scale in Table 3. This scale was inspired by Sobania et al. (2023) and Ali and Kostakos (2023). Weights in Table 1 were applied to achieve a final score using equation (1).

$$s = (acc * 0.4) + (rel * 0.3) + (act * 0.3) \qquad (1)$$

3.2 Case 2: Zero-Shot Knowledge

Test Case 2 aimed to evaluate GPT-3.5, GPT-4, Gemini and MS Copilot's ability to respond to database configuration queries based on CIS benchmarks without direct exposure to the benchmarks. The goal was to assess their zero-shot learning capabilities, like Pearce et al. (2023) evaluation of GPT-3.5-Turbo and Codex for code vulnerability repair. The study emphasized the significance of prompt phrasing and context in generating accurate responses. The aim was to explore whether the models could deduce and apply MySQL and MongoDB security standards from CIS benchmarks, despite not being specifically trained on them. This case tested the LLMs general understanding of database security, derived from a broad dataset, and assessed their practical utility in real-world scenarios as suggested by Ye et al. (2023). Multiple LLMs were used based on Gupta et al. (2023) suggestions regarding the need to understand the capabilities and limitations of different models in handling complex cybersecurity tasks.

Table 3: Rating Scale.

Value	Rating
1	**Very Poor** - Does not meet the criterion defined in Table 1 at all.
2	**Poor** - Slightly meets the criterion defined in Table 1 but is largely inadequate.
3	**Average** - Adequately meets the criterion defined in Table 1 with some improvements needed.
4	**Good** - Meets the criterion defined in Table 1 well with minor improvements needed.
5	**Excellent** - Fully meets the criterion defined in Table 1.

A total of 76 CIS MySQL benchmarks and 23 MongoDB benchmarks were evaluated using a Hit or Miss approach. The LLM's were asked generic questions that covered the categories in Table 2, and the responses were evaluated based on whether the benchmarks present in the categories were mentioned or not. An example of a prompt used to test the benchmarks related to MySQL authentication took the form "How should I set up account policies on MySQL securely on Ubuntu in accordance with CIS benchmarks?".

3.3 Case 3: Zero-Shot Continuation

Test Case 3 evaluated GPT-3.5 and GPT-4's ability to correct deviations from MySQL and MongoDB CIS benchmarks through follow-up prompts, after the answers from Case 2 were analysed. Similar to Sobania et al. (2023) and Pearce et al. (2023), who showed that additional context and iterative prompts improve LLM performance, this study aimed to assess if similar methods could enhance compliance accuracy with CIS standards.

After the analysis of the responses from Case 2, the benchmarks that were marked as Miss were further investigated by asking the LLMs specific questions using keywords in the given benchmark. A maximum of two further questions were used. The questions used had the form "What further steps should I take to configure [database] on Ubuntu for [keyword] in accordance with the CIS benchmarks?".

3.4 Case 4: Zero-Shot Assisted

This test was designed to evaluate GPT-4's ability to respond to database configuration queries when provided with the MySQL and MongoDB CIS benchmarks in PDF format. The aim was to assess whether access to CIS benchmark documents could improve GPT-4's accuracy and compliance. The responses were evaluated using the same Hit or Miss approach from Case 2. The same questions used in Case 2 were used.

4 DATA ANALYSIS AND RESULTS

4.1 Case 1: Knowledge Evaluation

Table 4: Case 1 Results for MySQL.

Category	GPT-3.5	GPT-4
1. OS Configuration	2.3	3.0
2. Installation and Planning	2.7	2.9
3. File Permissions	3.3	3.2
4. General	3.3	2.6
5. Permissions	3.3	2.6
6. Audit and Logging	3.6	2.9
7. Authentication	2.9	2.9
8. Network	2.9	2.9
9. Replication	2.9	2.9
Total	**27.2**	**25.9**

Table 4 presents the results obtained using the weighted evaluation framework proposed for GPT-3.5 and GPT-4 for MySQL (responses and analysis available at https://github.com/ddebon2004/LLM). From the results, it was observed that GPT-3.5 obtained a slightly better overall score than GPT-4. Upon analysing further the scores obtained, it was observed that in categories such as 4, 5, and 6, GPT-3.5's answers were more relevant and aligned with the CIS benchmarks, whereas the answers provided by GPT-4 mentioned a lot of points which were not directly related to the particular category. In categories 1 and 2, GPT-4 had a higher score, due to a more practical and actionable response and a wider range of points covered for these categories. The responses also included guidance on how to implement the recommendations as shown in Table 5.

Table 5: GPT-3.5 and GTP-4 Case 1 Answers Extract.

Prompt:	Describe the best practices and recommendations related to securing the operating system on which the MySQL Community Server is running.
GPT-3.5	Implement file system permissions: Set appropriate file system permissions to restrict access to MySQL data directories, configuration files, and other sensitive files. Ensure that only authorized users and processes have read and write permissions to these files.
GPT-4	4. File System Security Secure the MySQL data directory (/var/lib/mysql by default) by setting appropriate file system permissions. Only the MySQL user and group should have access. Consider using file system encryption for data at rest, especially for sensitive data.

Table 6: Case 1 Results for MongoDB.

Category	GPT-3.5	GPT-4
1. Installation and Patching	3.6	4.6
2. Authentication	4.0	2.6
3. Authorisation	4.3	4.3
4. Data Encryption	4.3	5.0
5. Audit Logging	5.0	3.0
6. OS Hardening	3.0	5.0
7. File Permission	5.0	5.0
Total	**29.2**	**29.5**

Table 6 presents the results obtained for MongoDB. The difference between the two LLMs was statistically negligible considering that the metrics used are subjective in nature. Similar behaviours as observed when analysing the MySQL responses were observed between GPT-3.5 and GPT-4.

4.2 Case 2: Zero-Shot Knowledge

In this section we present the results for GPT-3.5, GPT-4, MS Copilot, and Gemini when the LLMs were explicitly asked for database security instructions related to CIS benchmarks. Table 7 shows the *total hits / total benchmarks* for MySQL.

GPT-4 had an edge over the other LLMs when asked about security policies related to CIS benchmarks with 31 benchmarks out of the total 76, being mentioned by the model with certain categories including Installation and Planning, Audit and Logging and Networking almost covering all the points mentioned in the standard. MS Copilot had a very similar performance to GPT-3.5, whilst Gemini although giving valid answers, missed a lot of details included in the benchmarks.

Table 7: Case 2 Results for MySQL.

Category	GPT-3.5	GPT-4	Copilot	Gemini
1	6 / 16	9 / 16	4 /16	6 / 16
2	5 / 8	7 / 8	5 / 8	0 / 8
3	0 / 9	1 / 9	2 / 9	1 / 9
4	2 / 9	2 / 9	1 / 9	1 / 9
5	1 / 10	1 / 10	0 / 10	0 / 10
6	1 / 5	3 / 5	1 / 5	1 / 5
7	3 / 11	4 / 11	1 / 11	1 / 11
8	1 / 3	3 / 3	1 / 3	0 / 3
9	1 / 5	1 / 5	1 / 5	1 / 5
Total	**20 / 76**	**31 / 76**	**16 / 76**	**11 / 76**

Table 8: Case 2 Results for MongoDB.

Category	GPT-3.5	GPT-4	Copilot	Gemini
1	1 / 1	1 / 1	0 /1	0 / 1
2	1 / 3	1 / 3	1 / 3	1 / 3
3	3 / 5	4 / 5	0 / 5	3 / 5
4	2 / 5	2 / 5	1 / 5	0 / 5
5	4 / 4	4 / 4	1 / 4	2 / 4
6	0 / 3	0 / 3	0 / 3	0 / 3
7	2 / 2	2 / 2	1 / 2	2 / 2
Total	**13 / 23**	**14 / 23**	**4 / 23**	**7 / 23**

For MongoDB, both versions of GPT had similar results as depicted in Table 8. It was observed that these LLMs had a better performance in adhering to CIS policies for this database. This is mainly due to the MongoDB benchmarks being similar to common generic database security best practices. Table 9 highlights the benchmarks for the second category. The first benchmark represents a very common authentication measure and was mentioned by the two LLMs, however both had problems with the other benchmarks due to the very specific nature of the recommendations to MongoDB.

Table 9: CIS MongoDB Authentication Benchmarks.

2.1 Ensure Authentication is configured
2.2 Ensure that MongoDB does not bypass authentication via the localhost exception
2.3 Ensure authentication is enabled in the sharded cluster

4.3 Case 3: Zero-Shot Continuation

Tables 10 and 11 show the results for GPT-3.5 and GPT-4 after follow-up questions were posed in response to the answers from Case 2, utilizing prompt engineering techniques.

Table 10: Case 3 GPT-3.5 MySQL Iterations.

Category	Prompt 1	Prompt 2	Prompt 3
1	7 / 16	13 / 16	16 / 16
2	4 / 8	8 / 8	8 / 8
3	0 / 9	0 / 9	0 / 9
4	1 / 9	2 / 9	4 / 9
5	1 / 10	2 / 10	10 / 10
6	2 / 5	2 / 5	2 / 5
7	5 / 11	7 / 11	11 / 11
8	0 / 3	1 / 3	2 / 3
9	1 / 5	1 / 5	4 / 5
Total	**21 / 76**	**36 / 76**	**57 / 76**

Table 11: Case 3 GPT-4 MySQL Iterations.

Category	Prompt 1	Prompt 2	Prompt 3
1	3 / 16	11 / 16	12 / 16
2	5 / 8	8 / 8	8 / 8
3	0 / 9	0 / 9	0 / 9
4	2 / 9	4 / 9	5 / 9
5	1 / 10	10 / 10	10 / 10
6	2 / 5	2 / 5	4 / 5
7	7 / 11	10 / 11	11 / 11
8	3 / 3	3 / 3	3 / 3
9	1 / 5	1 / 5	5 / 5
Total	**24 / 76**	**49 / 76**	**58 / 76**

It was noted that when using the same prompt as in test Case 2, the LLMs often produced entirely different responses in 55% of the cases, indicating that the models exhibit a degree of randomness in their behaviour. By using additional prompts and specifying security terms more clearly within the benchmarks, the models provided more responses aligned with the CIS standard. With a maximum of three prompts, both models covered 76% of the benchmarks. It was also observed that, in general, GPT-4 produced better responses by the second prompt compared to GPT-3.5. The results emphasize the importance of asking precise questions and having a solid understanding of the security standard. Table 12 provides an example of the prompts used for Category 9.

Table 12: Iterative prompts for MySQL Category 9.

1	How should I set up replication on MySQL securely on Ubuntu in accordance with CIS benchmarks?
2	What further steps should I take to configure MySQL on Ubuntu for ensuring secure ssl server certificate, master repository, super privilege, and wildcard hostnames in accordance with the CIS benchmarks?
3	What further steps should I take to configure MySQL on Ubuntu for ensuring secure master ssl

	verify server certificate, master info repository, super privilege for replication users, and wildcard hostnames for replication users in accordance with the CIS benchmarks?

Another interesting point were the responses for category 3, File Permissions, where both LLMs provided file permissions that were completely misaligned with the benchmarks, even after additional prompts were used.

Table 13: Case 3 GPT-3.5 MongoDB Iterations.

Category	Prompt 1	Prompt 2	Prompt 3
1	1 / 1	1 / 1	1 / 1
2	1 / 3	3 / 3	3 / 3
3	3 / 5	3 / 5	5 / 5
4	2 / 5	5 / 5	5 / 5
5	4 / 4	4 / 4	4 / 4
6	0 / 3	3 / 3	3 / 3
7	2 / 2	2 / 2	2 / 2
Total	**13 / 23**	**21 / 23**	**23 / 23**

Table 14: Case 3 GPT-4 MongoDB Iterations.

Category	Prompt 1	Prompt 2	Prompt 3
1	1 / 1	1 / 1	1 / 1
2	1 / 3	3 / 3	3 / 3
3	3 / 5	4 / 5	5 / 5
4	2 / 5	5 / 5	5 / 5
5	4 / 4	4 / 4	4 / 4
6	0 / 3	3 / 3	3 / 3
7	2 / 2	2 / 2	2 / 2
Total	**13 / 23**	**22 / 23**	**23 / 23**

In the MongoDB case, with the use of additional prompts, both LLMs successfully covered all benchmarks, requiring a maximum of three prompts, as shown in Tables 13 and 14. In this case, the randomness of the answers for the first prompt was negligible and the information given was more consistent with the responses given in Case 2. Consistent with the previous case, both LLMs exhibited similar knowledge with this type of database.

4.4 Case 4: Zero-Shot Assisted

Table 15 presents the results obtained when GPT-4 was asked about security benchmarks and provided with the actual PDF file of the benchmarks. The table also provides a summary of the results obtained for Case 2, for comparison reasons.

Table 15: GPT-4 Case 2 and Case 4 Results.

	MySQL		**MongoDB**	
Category	**Case 2**	**Case 4**	**Case 2**	**Case 4**
1	9 / 16	13 / 16	1 / 1	1 / 1
2	7 / 8	7 / 8	1 / 3	3 / 3
3	1 / 9	5 / 9	4 / 5	5 / 5
4	2 / 9	5 / 9	2 / 5	5 / 5
5	1 / 10	10 / 10	4 / 4	4 / 4
6	3 / 5	5 / 5	0 / 3	0 / 3
7	4 / 11	4 / 11	2 / 2	2 / 2
8	3 / 3	3 / 3		
9	1 / 5	5 / 5		
Total	**31 / 76**	**57 / 76**	**14 / 23**	**20 / 23**

The results demonstrate that supplying the model with the actual document had a positive impact on the responses, yielding a level of detail comparable to using additional prompts, as seen in Case 3. This highlights the model's ability to analyse and learn from the information provided. However, it was also observed that the knowledge is retained only within the specific session. Another interesting point is that is that the model demonstrated improved accuracy for MySQL File Permissions in Category 3 compared to Case 3, but it did not provide sufficient details for Category 7, authentication.

5 CONCLUSIONS

The methodology presented in this study has proven effective in evaluating LLMs' compliance with CIS benchmarks across different scenarios. While subjective in nature, the evaluation framework can be applied to assess LLMs' responses in any domain.

Similar to the work of Ali and Kostakos (2023), who evaluated ChatGPT's general knowledge using ISACA and CISM questions, the first test in this study focused on assessing LLMs' knowledge of security configuration concepts for MySQL and MongoDB databases on Linux servers. When posed with general security questions about these systems, GPT-3.5 provided more accurate and relevant information, while GPT-4 excelled in offering actionable advice.

The second test case evaluated the zero-shot knowledge of four different LLMs. As noted by Pearce et al. (2023), LLMs can provide relevant security configurations without directly referencing CIS benchmarks. GPT-4 demonstrated a stronger zero-shot learning capability than the other models, especially in terms of contextual appropriateness, covering 40% of the benchmarks. The models also struggled when challenged with highly specific and technical configurations.

When posed with the same repeated questions, LLMs displayed a degree of randomness in their responses. The study underscores the significance of prompt engineering and engaging with the models to achieve improved results. GPT-4 converges more quickly to the desired answer and generally requires fewer prompts. Additionally, the model demonstrated strong capabilities in learning from documents, though this learning is limited to a single session, as noted by Al-Hawawreh et al. (2023).

While the models proved effective in suggesting security configurations in compliance with international standards, the potential for missing or incorrect responses highlights their limitations. This suggests that, although helpful, domain expertise is still necessary.

6 FUTURE WORK

Future research should broaden the evaluation scope to include a wider range of databases and operating systems, improving the generalisability of the findings and uncovering strengths and weaknesses across diverse platforms. Additionally, developing more objective evaluation criteria using advanced metrics and automated tools can reduce subjective bias and enhance accuracy in assessing LLM compliance. Finally, exploring advanced prompt engineering techniques could further refine LLM performance, particularly in complex scenarios.

REFERENCES

Beheshti, A. (2023). *Empowering Generative AI with Knowledge Base 4.0: Towards Linking Analytical, Cognitive, and Generative Intelligence*. doi:10.1109/icws60048.2023.00103

Sarker, I. H., Furhad, M. H., & Nowrozy, R. (2021). AI-Driven Cybersecurity: An Overview, security intelligence modeling and research directions. *SN Computer Science*, *2*(3). doi:10.1007/s42979-021-00557-0.

Shanthi, R. R., Sasi, N. K., & Gouthaman, P. (2023). *A New Era of Cybersecurity: The Influence of Artificial Intelligence*. doi:10.1109/icnwc57852.2023.10127453.

Sharma, P., & Dash, B. (2023b). *Impact of Big Data Analytics and ChatGPT on Cybersecurity*. doi:10.1109/i3cs58314.2023.10127411.

Al-Hawawreh, M., Aljuhani, A., & Jararweh, Y. (2023). Chatgpt for cybersecurity: practical applications, challenges, and future directions. *Cluster Computing*, *26*(6), 3421–3436. doi:0.1007/s10586-023-04124-5.

Oh, S., & Shon, T. (2023b). *Cybersecurity Issues in Generative AI*. doi:10.1109/platcon60102.2023.10255179.

Sobania, D., Briesch, M., Hanna, C., & Petke, J. (2023d). *An Analysis of the Automatic Bug Fixing Performance of ChatGPT*. doi:10.1109/apr59189.2023.00012.

Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. *IEEE Access*, *11*, 80218–80245. doi:10.1109/access.2023.3300381.

Ali, T., & Kostakos, P. (2023, September 27). *HuntGPT: Integrating Machine Learning-Based Anomaly Detection and Explainable AI with Large Language Models (LLMs)*. arXiv.org. doi:10.48550/arXiv.2309.16021

Pearce, H., Tan, B., Ahmad, B., Karri, R., & Dolan-Gavitt, B. (2023). *Examining Zero-Shot Vulnerability Repair with Large Language Models*. doi:10.1109/sp46215.2023.10179324.

ISO/IEC 27001:2013. (n.d.). https://www.iso.org/obp/ui/#iso:std:iso-iec:27001:ed-2:v1:en

Zeadally, S., Adi, E., Baig, Z., & Khan, I. A. (2020). Harnessing artificial intelligence capabilities to improve cybersecurity. IEEE Access, 8, 23817–23837. doi:10.1109/access.2020.2968045.99

CIS Oracle MySQL Benchmarks. (n.d.). CIS. https://www.cisecurity.org/benchmark/oracle_mysql

CIS MongoDB Benchmarks. (n.d.). CIS. https://www.cisecurity.org/benchmark/mongodb

Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., Cui, Y., Zhou, Z., Gong, C., Shen, Y., Zhou, J., Chen, S., Gui, T., Zhang, Q., & Huang, X. (2023). A comprehensive capability analysis of GPT-3 and GPT-3.5 series models. *arXiv (Cornell University)*. doi:10.48550/arxiv.2303.10420.