Predicting Residual Risk Using Large Language Models and ISO/IEC 27001 Security Controls

Marek Hrabcak
Department of Informatics, Faculty of Natural Sciences and Informatics Constantine the Philosopher University in Nitra Nitra, Slovakia;
marek.hrabcak@ukf.sk

Zoltan Balogh
Department of Informatics, Faculty of Natural Sciences and Informatics Constantine the Philosopher University in Nitra Nitra, Slovakia;
Kandó Kálmán Faculty of Electrical Engineering Óbuda University
Budapest, Hungary zbalogh@ukf.sk

Jan Francisti
Department of Informatics, Faculty of Natural Sciences and Informatics Constantine the Philosopher University in Nitra Nitra, Slovakia;
jfrancisti@ukf.sk

Abstract— Previous research demonstrated that Large Language Models (LLMs) can classify cybersecurity risks with a level of consistency comparable to human assessors. However, the prediction of residual risk after the implementation of security controls remains largely unexplored. This paper investigates the capability of LLMs to estimate residual cybersecurity risk in ISO/IEC 27001-based risk management scenarios. A dataset containing 8,671 cybersecurity risk scenarios was analyzed using an LLM-based framework that considers assets, threats, vulnerabilities, security controls, and inherent risk values. Security controls were categorized according to NIST SP 800-53 control families to enable comparative analysis of risk reduction patterns. The results showed that 88.03% of evaluated scenarios exhibited measurable risk reduction after the implementation of security controls, with an average reduction of 40.2%. Significant differences were observed among control families (Kruskal-Wallis H = 32.93, p = 0.034), while a moderate positive correlation was identified between inherent risk severity and predicted risk reduction (ρ = 0.581, p < 0.001). The findings suggest that LLMs can generate internally consistent and context-sensitive residual risk estimates and may serve as decision-support tools for cybersecurity professionals. The study contributes to the growing body of research on AI-assisted cybersecurity risk management and highlights opportunities and limitations of LLM-based residual risk prediction.

Keywords— Large Language Models, Cybersecurity Risk Management, Residual Risk Prediction, ISO/IEC 27001, Artificial Intelligence.

I. Research Question and Research Objectives

While Large Language Models have been applied to cybersecurity risk identification and classification, their use for residual risk estimation remains largely unexplored. This is particularly important because residual risk directly influences risk treatment and security decision-making.

Therefore, this study investigates the following research question:

RQ1: Can Large Language Models generate consistent and context-sensitive residual cybersecurity risk estimates based on implemented security controls?

To answer this question, a dataset containing 8,671 cybersecurity risk scenarios was analyzed. Each scenario included an asset, threat, vulnerability, proposed security control, inherent risk value, and residual risk value generated by a Large Language Model. Security controls were subsequently categorized according to the NIST SP 800-53 control families, enabling statistical analysis of risk reduction patterns across different categories of controls.

The objectives of this study are threefold:

1. To evaluate the overall reduction between inherent and residual cybersecurity risk predicted by a Large Language Model.
2. To investigate whether different categories of security controls are associated with different levels of predicted risk reduction.
3. To analyze the relationship between inherent risk severity and the magnitude of predicted residual risk reduction.

By addressing these objectives, the study extends previous research on LLM-based cybersecurity risk classification toward the prediction of residual cybersecurity risk and the evaluation of security control effectiveness.

II. Related Work

A. Large Language Models in Cybersecurity Risk Assessment

Large Language Models (LLMs) have recently emerged as a promising technology for supporting a wide range of cybersecurity activities, including threat intelligence analysis, vulnerability management, incident response, compliance assessment, and cybersecurity risk managemen. (Jones et al., 2025) Due to their ability to process large volumes of unstructured information and generate context-aware responses, LLMs have attracted considerable attention from both researchers and practitioners seeking to improve the efficiency of security-related decision-making processes. (Zhang et al., 2025)

Within the domain of cybersecurity risk management, LLMs have been explored as tools for assisting with the identification, interpretation, and evaluation of cyber risks. (Divakaran & Peddinti, 2024) Their ability to analyze relationships between assets, threats, vulnerabilities, and security requirements makes them suitable candidates for supporting established risk assessment methodologies. (Rainio et al., 2024) In addition, the integration of LLMs with cybersecurity frameworks and governance models has been proposed as a means of improving the scalability and repeatability of risk assessment processes.

Despite these advances, the majority of existing research focuses on risk identification, classification, and prioritization. (Ruohonen et al., 2025) Current studies primarily investigate whether LLMs can recognize cybersecurity risks and assign appropriate risk levels based on available contextual information. Consequently, the application of LLMs to more advanced stages of the risk management lifecycle remains insufficiently explored, creating opportunities for further

research into their use for supporting cybersecurity decision-making beyond initial risk assessment activities. (Divakaran & Peddinti, 2024)

B. Human and LLM-Based Risk Classification

The increasing adoption of Large Language Models in cybersecurity has raised important questions regarding their ability to perform tasks traditionally requiring expert judgment. One of the most challenging areas is cybersecurity risk assessment, where evaluators must analyze complex relationships between assets, threats, vulnerabilities, and organizational context to determine the severity of potential risks. (Bissoli et al., 2026)

Several studies have investigated the capability of LLMs to support or replicate human decision-making processes. (Loya et al., n.d.) Existing research indicates that modern LLMs can perform classification and evaluation tasks with a level of consistency approaching that of human experts in selected domains. (Liu, 2023)In cybersecurity, these capabilities have motivated research into the use of LLMs for threat analysis, vulnerability prioritization, compliance assessment, and risk evaluation. (Zhang et al., 2025)

Recent work by Hrabčák and Balogh (2026) provided empirical evidence regarding the application of LLMs in ISO/IEC 27001-based cybersecurity risk assessment. Using a dataset of 8,672 risk scenarios created within a simulated banking environment, the study compared risk classifications produced by approximately 130 human evaluators and ChatGPT using identical assets, threats, vulnerabilities, and risk assessment criteria. The results demonstrated a high degree of agreement in inherent risk classification, indicating that the LLM was capable of reproducing human reasoning patterns when evaluating cybersecurity risks. At the same time, the analysis revealed systematic differences in residual risk assessments, where the LLM tended to assign slightly more conservative risk levels than human evaluators.

These findings suggest that LLMs can provide consistent support for cybersecurity risk classification and may serve as effective decision-support tools within risk management processes. (Hashmi et al., 2024)However, existing studies primarily focus on the classification of inherent and residual risks based on predefined assessment criteria. (Ruohonen et al., 2025)They do not investigate whether LLMs can estimate the effectiveness of specific security controls or predict residual risk levels after the implementation of mitigation measures.

Consequently, while previous research has established the feasibility of using LLMs for cybersecurity risk classification, the application of LLMs to residual risk prediction remains largely unexplored. (Zhang et al., 2025) This limitation motivates the present study, which extends prior work by examining whether Large Language Models can estimate residual cybersecurity risk based on implemented security controls and whether these predictions exhibit consistent and statistically meaningful patterns.

C. Residual Risk Assessment and Security Control Effectiveness

Cybersecurity risk management extends beyond the identification and classification of risks. (Ruohonen et al., 2025) A fundamental objective of risk treatment is the reduction of inherent risk through the implementation of appropriate technical, organizational, and procedural security controls. (Brunaccioni et al., 2026)The remaining level of risk after the application of mitigation measures is commonly referred to as residual risk and represents a key factor in risk-based decision-making (Berger & Plump, 2025).

Internationally recognized frameworks such as ISO/IEC 27001, ISO/IEC 27005, and the NIST Risk Management Framework emphasize the importance of evaluating residual risk when determining whether implemented safeguards provide an acceptable level of protection. (Rananga & Venter, 2026) In practice, organizations continuously assess whether selected controls sufficiently reduce the likelihood and impact of cybersecurity threats while maintaining compliance with regulatory and business requirements. (Melo et al., 2026)

The effectiveness of security controls has been widely studied within the context of information security governance and risk management. (Papastergiou et al., 2026)Security measures such as access control mechanisms, network segmentation, cryptographic protections, incident response procedures, awareness training, and contingency planning are intended to reduce either the probability of successful attacks or the severity of their consequences. However, the actual impact of individual controls on residual risk is often difficult to quantify due to the complexity of modern information systems and the dynamic nature of cyber threats (Deaver-Vazquez et al., 2024).

As a result, residual risk assessment frequently relies on expert judgment, qualitative evaluation methods, and organizational experience. (Lv et al., 2026)Different assessors may estimate the effectiveness of the same security control differently depending on their expertise, assumptions, and interpretation of the assessed scenario. This challenge creates opportunities for intelligent decision-support mechanisms capable of providing consistent and repeatable estimates of residual risk. (Dokas, 2025)

Recent advances in artificial intelligence suggest that Large Language Models may support such activities by analyzing relationships between identified risks, implemented controls, and expected mitigation outcomes. (Papastergiou et al., 2026)Consequently, understanding how security controls influence residual risk predictions represents an important step toward the development of AI-assisted cybersecurity risk management processes.

D. Security Control Taxonomies and NIST SP 800-53

The assessment of cybersecurity risks is closely linked to the selection and implementation of security controls. To facilitate systematic risk treatment, several cybersecurity frameworks provide structured taxonomies for categorizing security measures according to their objectives and areas of application. (Khaleefah & Al-Mashhadi, 2024) Such taxonomies support consistency in risk assessment, control selection, compliance evaluation, and security governance.
Among the most widely recognized frameworks, NIST Special Publication 800-53 provides a comprehensive catalog of security and privacy controls organized into control families. These families cover a broad spectrum of cybersecurity domains, including access control, identification and authentication, configuration management, incident response, contingency planning, system integrity, physical protection, personnel security, and supply chain risk management. (Khan et al., 2023) The framework is extensively used by government agencies, critical infrastructure operators, and private-sector organizations as a

foundation for security program development and risk management activities. (Haber & Hibbert, 2018)

The family-based organization of controls offers several advantages for cybersecurity analysis. (Sánchez-García et al., 2023) It enables the aggregation of security measures into logically related categories, facilitates comparative evaluation of control effectiveness, and supports statistical analysis across heterogeneous datasets. Furthermore, grouping controls into standardized families reduces subjectivity when comparing different mitigation strategies and improves the reproducibility of research results. (Islam et al., 2026)

In the context of this study, security controls associated with individual cybersecurity scenarios were categorized according to the NIST SP 800-53 control families. This classification enabled the investigation of whether different categories of security controls are associated with different levels of predicted residual risk reduction. By utilizing an established and internationally recognized taxonomy, the analysis provides a structured foundation for evaluating the relationship between implemented controls and residual risk predictions generated by Large Language Models. (McIntosh et al., 2023)

E. Research gap

The existing body of research demonstrates that Large Language Models can support a variety of cybersecurity activities, including threat analysis, vulnerability assessment, compliance evaluation, and cybersecurity risk classification. Recent studies have shown that LLMs are capable of producing risk classifications that exhibit a level of consistency comparable to human assessors, suggesting their potential as decision-support tools in cybersecurity risk management. (Divakaran & Peddinti, 2024)

At the same time, established risk management frameworks emphasize that effective cybersecurity governance requires more than the identification and classification of risks. Organizations must also evaluate the effectiveness of implemented security controls and determine the resulting level of residual risk. (Quispe-Kobashikawa et al., 2026) While numerous studies have examined the role of security controls within frameworks such as ISO/IEC 27001, ISO/IEC 27005, and NIST SP 800-53, the assessment of residual risk remains largely dependent on expert judgment and qualitative evaluation methods (Junior & Arima, 2023).

Despite the growing interest in applying artificial intelligence to cybersecurity risk management, current research focuses primarily on inherent risk identification, classification, prioritization, and compliance-related tasks. (Hashmi et al., 2024)To the best of our knowledge, no prior study has systematically investigated whether Large Language Models can predict residual cybersecurity risk based on implemented security controls and whether such predictions exhibit statistically consistent patterns across different categories of controls. (Papastergiou et al., 2026)

This limitation represents an important research gap, particularly given the increasing complexity of cybersecurity environments and the demand for scalable risk assessment approaches. (Francisco & Linnér, 2023) Therefore, the objective of this study is to evaluate the capability of a Large Language Model to estimate residual cybersecurity risk after the implementation of security controls, analyze the resulting risk reductions, and investigate the relationship between residual risk predictions and standardized NIST SP 800-53 control families. (Berger & Plump, 2025)

III. METHODOLOGY

A. Residual Risk Prediction Framework

The objective of this study was to evaluate the capability of a Large Language Model (LLM) to predict residual cybersecurity risk after the implementation of security controls. The proposed methodology was designed to investigate whether LLM-generated residual risk estimates exhibit consistent and statistically meaningful patterns across a large set of cybersecurity scenarios.

The overall framework consisted of four consecutive phases: (1) cybersecurity scenario preparation, (2) security control categorization, (3) residual risk prediction, and (4) statistical evaluation. In the first phase, cybersecurity risk scenarios were collected from ISO/IEC 27001-based risk assessment exercises performed within a simulated banking environment. Each scenario contained an identified asset, threat, vulnerability, proposed security control, and an inherent risk value represented on a five-level ordinal scale. In the second phase, security controls were categorized according to the NIST SP 800-53 control families. This categorization provided a standardized taxonomy that enabled comparative analysis of different categories of cybersecurity controls. In the third phase, the Large Language Model was used to estimate the residual risk for each scenario after considering the proposed mitigation measure. The model received a structured input consisting of the asset description, threat description, vulnerability description, security control, and the corresponding inherent risk value. Residual risk predictions were generated using the same five-level qualitative scale employed during the original risk assessment process, ranging from 1 (Very Low Risk) to 5 (Very High Risk). Identical prompt structures and evaluation criteria were applied across all scenarios to ensure consistency of the generated assessments.

The effectiveness of individual controls was evaluated using the risk reduction metric (ΔRisk), defined as:

$$\Delta\text{Risk} = \text{Inherent Risk} - \text{Residual Risk}$$

A positive ΔRisk value indicates a reduction in risk following the implementation of a security control, while a value of zero indicates no predicted change. Negative values were excluded during dataset validation because security controls are expected to maintain or reduce risk rather than increase it.

In the final phase, the resulting dataset of 8,671 cybersecurity scenarios was statistically analyzed. Descriptive statistics were used to evaluate the overall distribution of risk reductions. Comparative analysis of NIST SP 800-53 control families was performed to identify differences among categories of security controls. The Kruskal-Wallis test was applied to determine whether these differences were statistically significant, while Spearman correlation analysis was used to evaluate the relationship between inherent risk severity and the magnitude of predicted risk reduction.

B. Dataset Construction

The dataset used in this study originated from ISO/IEC 27001-based cybersecurity risk assessment exercises conducted within a simulated banking environment referred to as

HackMeBank. The environment was documented through a comprehensive set of organizational processes, assets, and security controls aligned with the requirements of ISO/IEC 27001 and relevant cybersecurity regulations.

A total of approximately 120 MSc students specializing in Applied Informatics participated in the identification and assessment of cybersecurity risks. For each identified scenario, participants specified the affected asset, threat, vulnerability, proposed security control, and corresponding inherent and residual risk values using a qualitative five-level risk scale. The resulting dataset contained 8,671 unique cybersecurity risk scenarios covering a broad spectrum of organizational, technical, physical, and procedural security issues.

C. Experiment 1: Prompt Design and LLM Configuration

The residual risk prediction process was implemented using an OpenAI Large Language Model accessed through the API interface. To ensure consistency and reproducibility, all cybersecurity scenarios were evaluated using a standardized prompt structure and identical assessment criteria. For each scenario, the model received the following input attributes: Asset description, Threat description, Vulnerability description, Proposed security control, Inherent risk value. The model was instructed to estimate the residual cybersecurity risk after the implementation of the proposed mitigation measure. Residual risk was evaluated using the same five-level qualitative scale employed throughout the study, where 1 represented Very Low Risk and 5 represented Very High Risk. The prompt was designed to emulate expert cybersecurity risk assessment reasoning by requiring the model to consider the relationship between the identified threat, the exploited vulnerability, the affected asset, and the expected effectiveness of the proposed security control.

A typical prompt structure was defined as follows:

You are a cybersecurity risk assessment expert.
Evaluate the residual cybersecurity risk after implementing
the proposed security control.
Consider: Asset, Threat, Vulnerability, Security Control, Inherent Risk.
Estimate the resulting residual risk on a scale from 1 to 5, where:
1 = Very Low Risk, 2 = Low Risk, 3 = Medium Risk, 4 = High Risk, 5 = Very High Risk
The residual risk should reflect the expected level of risk remaining after the security control has been fully implemented.
Return only the residual risk value.

An example evaluation request is shown below:

Asset: Web Application
Threat: Unauthorized Access
Vulnerability: Weak Authentication Mechanism
Security Control: Implement Multi-Factor Authentication (MFA)
Inherent Risk: 4

Expected output:

Residual Risk: 2

The same prompt structure was applied to all 8,671 cybersecurity scenarios.

D. Experiment 2: NIST Control Family Classification and Validation

To support control-family-based analysis, all security controls were categorized according to the NIST SP 800-53 control families. The classification was performed using a Large Language Model configured with descriptions of all supported control families and explicit classification instructions. For each security control, the model assigned a single primary NIST family representing the dominant security objective of the mitigation measure.

To ensure consistency, the model was executed using a deterministic configuration (temperature = 0) and an identical prompt structure. Controls that could not be reliably mapped to a predefined family were assigned to the category OTHER.

The NIST SP 800-53 taxonomy provided a standardized framework for grouping heterogeneous security controls into comparable categories, enabling subsequent statistical analysis across control families. Although the classification process was automated, the resulting categories were used primarily for aggregate analysis rather than individual control evaluation. Therefore, occasional misclassifications are unlikely to materially affect the overall trends observed across large control-family groups. Future work should include expert validation of control-family assignments to further improve classification reliability.

Examples of NIST control families include AC (Access Control), AU (Audit & Accountability), CM (Configuration Management), CP (Contingency Planning), IA (Identification & Authentication), and IR (Incident Response).

E. Dataset overview

The dataset contained 8,671 cybersecurity risk scenarios, each including an inherent risk value, an LLM-predicted residual risk value, and the corresponding risk reduction metric (ΔRisk). As shown in Table 1, the average inherent risk decreased from 3.41 to 1.99 after applying security controls, representing a mean reduction of 1.42 risk levels (40.2%).

TABLE I. DATASET OVERVIEW (*SOURCE: AUTHORS*)

	Metric	***Value***
1	Total Scenarios	8,671
2	Mean Inherent Risk	3.41
3	Median Inherent Risk	3
4	StdDev Inherent Risk	0.92
5	Mean Residual Risk	1.99
6	Median Residual Risk	2
7	StdDev Residual Risk	0.81
8	Mean ΔRisk	1.42
9	Median ΔRisk	1
10	StdDev ΔRisk	0.86
11	Mean Reduction (%)	40.2
12	Median Reduction (%)	40.0
13	StdDev Reduction (%)	21.2
14	Minimum ΔRisk	0
15	Maximum ΔRisk	4
16	Negative ΔRisk Count	0
17	Scenarios with ΔRisk > 0	88.03%
18	Mean Reduction (%)	40.2%

The predicted residual risks were generally lower than the corresponding inherent risks, with no observed risk increases.

These results indicate internally consistent residual risk estimates suitable for subsequent statistical analyses.

IV. Results

A. Risk Reduction by Control Family

Security controls were grouped according to the NIST SP 800-53 control families and analyzed using average risk reduction metrics. Table 2 shows that CP, MA, IA, and SI controls achieved the highest average reductions, whereas CM, SR, and AT controls exhibited the lowest.

TABLE II. Risk reduction by NIST SP 800-53 control family. (*Source: Authors*)

Family	*N*	*Avg ΔRisk*	*Avg Reduction %*
CP	482	1.53	41.97%
MA	161	1.51	43.63%
IA	720	1.49	40.66%
SI	502	1.49	40.94%
PE	511	1.46	42.48%
RA	128	1.45	40.40%
PT	173	1.44	40.18%
IR	321	1.43	38.98%
SA	244	1.43	39.13%
PM	300	1.42	41.33%
MP	377	1.42	40.71%
CA	263	1.41	40.51%
AC	812	1.41	39.19%
PL	135	1.39	40.00%
OTHER	362	1.38	39.81%
PS	274	1.38	40.82%
AU	792	1.37	39.51%
AT	660	1.37	39.50%
SR	114	1.34	38.25%
CM	512	1.33	39.43%

The results indicate that predicted residual risk varies across control families, suggesting that the model differentiates between categories of cybersecurity controls when estimating mitigation effects.

B. Distribution of Risk Reduction

Table 3 presents the distribution of risk reduction (ΔRisk) across 8,671 cybersecurity scenarios. Most scenarios exhibited a reduction of one (45.6%) or two (31.6%) risk levels, whereas reductions of three (10.1%) and four (0.8%) levels were less frequent.

TABLE III. Distribution of predicted risk reduction (ΔRisk) across all evaluated cybersecurity scenarios (*Source: Authors*)

Delta	*Count*	*Percent*
0	1038	11,97
1	3951	45,57
2	2737	31,56
3	877	10,11
4	68	0,79

Table 3 shows a positive relationship between inherent risk level and average risk reduction, with higher-risk scenarios exhibiting larger predicted reductions after the implementation of security controls.

Average risk reduction increased consistently with inherent risk severity, ranging from 0.05 at risk level 1 to 2.19 at risk level 5. Overall, 88.03% of the 8,671 evaluated scenarios exhibited measurable risk reduction, with an average reduction of 40.2%, indicating that the model adapts its predictions to the severity of the assessed risk.

C. Statistical Comparison of Control Families (Kruskal-Wallis)

The Kruskal-Wallis test confirmed statistically significant differences in predicted risk reduction among NIST SP 800-53 control families ($H = 32.93$, $p = 0.034$). In addition, Spearman correlation showed a moderate positive relationship between inherent risk and risk reduction ($\rho = 0.581$, $p < 0.001$), suggesting that higher-risk scenarios were associated with larger predicted reductions.

D. Correlation Analysis

Table IV shows a positive relationship between inherent risk and average risk reduction. The average reduction increases consistently from 0.05 at inherent risk level 1 to 2.19 at inherent risk level 5. This observation is supported by a statistically significant Spearman correlation ($\rho = 0.581$, $p < 0.001$), suggesting that the model adapts residual risk estimates according to the severity of the original risk scenario.

TABLE IV. Positive relationship between inherent risk and average risk reduction (*Source: Authors*)

RISK_INH_VALUE	*AVG_RISK_DELTA*
1	0.05381166
2	0.73862623
3	1.12219287
4	1.80541012
5	2.18699187

V. Discussion

The most significant finding is that 88.03% of the 8,671 evaluated cybersecurity scenarios exhibited measurable risk reduction after the implementation of security controls, with an average reduction of 40.2%. This result suggests that Large Language Models can support not only cybersecurity risk classification but also residual risk estimation by reflecting the expected mitigation effects of implemented controls.

Most scenarios experienced a reduction of one or two risk levels, while larger reductions were relatively rare, indicating that the model does not systematically overestimate control effectiveness. Significant differences were also observed among NIST SP 800-53 control families ($H = 32.93$, $p = 0.034$), suggesting that the model differentiates between categories of security controls when estimating residual risk. Furthermore, a moderate positive correlation was identified between inherent risk and predicted risk reduction ($\rho = 0.581$, $p < 0.001$). Higher-risk scenarios were associated with larger reductions, indicating that the generated estimates are context-sensitive and responsive to the severity of the assessed risk.

Overall, the findings suggest that LLMs can generate internally consistent residual risk estimates that may support cybersecurity risk assessment and decision-making processes. Although the objective accuracy of individual predictions was not evaluated, the results demonstrate meaningful relationships between security controls, inherent risk severity, and predicted residual risk.

VI. Limitations

This study has several limitations. First, the residual risk values and NIST SP 800-53 control-family assignments were generated by a Large Language Model and were not independently validated by cybersecurity experts. Therefore, the results should be interpreted as an evaluation of the model's consistency rather than as evidence of the objective

accuracy of individual predictions. Future research should include expert validation and inter-rater agreement analysis. Second, the dataset was based on predefined cybersecurity scenarios and may not fully reflect the complexity of real-world environments. Third, the study evaluated a single LLM configuration and prompting strategy; different models or prompts may produce different results. Finally, security controls were assessed individually, whereas real-world risk reduction often results from combinations of interacting technical, organizational, and procedural controls.

VII. Conclusion and Future Work

This paper evaluated the use of Large Language Models for residual cybersecurity risk prediction based on implemented security controls. Analysis of 8,671 cybersecurity risk scenarios showed that 88.03% of cases exhibited measurable risk reduction, with an average reduction of 40.2%. Significant differences were observed among NIST SP 800-53 control families, and a moderate positive correlation was identified between inherent risk severity and predicted risk reduction ($\rho = 0.581$, $p < 0.001$). The findings suggest that LLMs can generate internally consistent and context-sensitive residual risk estimates that may support cybersecurity risk management activities. Future research should focus on expert validation, cross-model evaluation, and hybrid risk assessment approaches integrating expert knowledge and historical cybersecurity data.

Acknowledgment

This work was supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic (ME SR) and of Slovak Academy of Sciences (SAS) under the contract No. VEGA 1/0492/26

References

[1] Berger, B. J., & Plump, C. (2025). Automatic security-flaw detection - towards a fair evaluation and comparison. Software and Systems Modeling. https://doi.org/10.1007/s10270-025-01300-6

[2] Bissoli, A., Mollaeefar, M., Landuyt, D. V., & Ranise, S. (2026). Benchmarking the effectiveness of multi-agent LLMs in collaborative privacy threat modeling with LINDDUN GO. Journal of Information Security and Applications. https://doi.org/10.1016/j.jisa.2026.104489

[3] Brunaccioni, L., Fusco, F. de, & Vallerio, M. (2026). Dynamic SHAP: Time resolved explainable AI for batch processes. Computers and Chemical Engineering. https://doi.org/10.1016/j.compchemeng.2026.109717

[4] Deaver-Vazquez, C., Taylor, E., Rowley, D., & Langis, B. (2024). A quantitative approach to assessing and managing cybersecurity risks. Edpacs. https://doi.org/10.1080/07366981.2024.2340849

[5] Divakaran, D. M., & Peddinti, S. T. (2024). Large Language Models for Cybersecurity: New Opportunities. IEEE Security & Privacy, 2–9. https://doi.org/10.1109/msec.2024.3504512

[6] Dokas, I. M. (2025). From hallucinations to hazards: benchmarking LLMs for hazard analysis in safety-critical systems. Safety Science. https://doi.org/10.1016/j.ssci.2025.107056

[7] Francisco, M., & Linnér, B.-O. (2023). AI and the governance of sustainable development. An idea analysis of the European Union, the United Nations, and the World Economic Forum. Environmental Science and Policy. https://doi.org/10.1016/j.envsci.2023.103590

[8] Haber, M. J., & Hibbert, B. (2018). Risk Management Frameworks. 241–247. https://doi.org/10.1007/978-1-4842-3627-7_20

[9] Hashmi, E., Yamin, M. M., & Yayilgan, S. Y. (2024). Securing tomorrow: a comprehensive survey on the synergy of Artificial Intelligence and information security. AI and Ethics. https://doi.org/10.1007/s43681-024-00529-z

[10] Islam, S., Sardar, B., Kalogeraki, E. M., Lampropoulos, K., & Papastergiou, S. (2026). Hybrid AI-Based dynamic risk assessment framework with explainable AI practices for composite product cybersecurity certification. International Journal of Information Security. https://doi.org/10.1007/s10207-026-01218-0

[11] Jones, L. B., Thornton, J., & Silva, D. D. (2025). Limitations of risk-based artificial intelligence regulation: a structuration theory approach. Discover Artificial Intelligence. https://doi.org/10.1007/s44163-025-00233-9

[12] Junior, A. S. C., & Arima, C. H. (2023). Cyber risk management and iso 27005 applied in organizations: a systematic literature review. Foco, 16(02), e1188–e1188. https://doi.org/10.54751/revistafoco.v16n2-215

[13] Khaleefah, A. D., & Al-Mashhadi, H. M. (2024). Methodologies, Requirements, and Challenges of Cybersecurity Frameworks: A Review. Iraqi Journal of Science. https://doi.org/10.24996/ijs.2024.65.1.38

[14] Khan, A. N., Bryans, J., Sabaliauskaite, G., & Jadidbonab, H. (2023). Integrated Attack Tree in Residual Risk Management Framework. Information. https://doi.org/10.3390/info14120639

[15] Liu, Y. (2023). The Importance of Human-Labeled Data in the Era of LLMs. https://doi.org/10.24963/ijcai.2023/802

[16] Loya, M., Sinha, D., & Futrell, R. (n.d.). Exploring the Sensitivity of LLMs' Decision-Making Capabilities: Insights from Prompt Variations and Hyperparameters. 3711–3716. https://doi.org/10.18653/v1/2023.findings-emnlp.241

[17] Lv, X., Yin, Z., Chen, S., Liu, Y., Yao, H., Ma, Y., Wang, C., Ma, Q., Li, X., & Zhou, S. (2026). Risk assessment model for hydrate blockage in deep-sea oil and gas multiphase pipelines based on XGBoost and SHAP. Chemical Engineering Science. https://doi.org/10.1016/j.ces.2026.124205

[18] McIntosh, T. R., Liu, T., Sušnjak, T., Alavizadeh, H., Ng, A. H.-M., Nowrozy, R., & Watters, P. A. (2023). Harnessing GPT-4 for generation of cybersecurity GRC policies: A focus on ransomware attack mitigation. Computers & Security, 134, 103424–103424. https://doi.org/10.1016/j.cose.2023.103424

[19] Melo, C., Maza, J. de la, & Recabarren, M. (2026). Validating AI-generated classroom observations: Reliability, accuracy, and limits of LLM-based pedagogical judgment. Computers and Education: Artificial Intelligence. https://doi.org/10.1016/j.caeai.2026.100612

[20] Papastergiou, S., Basheer, N., Lampropoulos, K., Verrios, P., & Islam, S. (2026). Explainable AI based dynamic cybersecurity risk management for cyber insurability. International Journal of Information Security. https://doi.org/10.1007/s10207-025-01189-8

[21] Quispe-Kobashikawa, G., Zuloaga-Estrada, C., Castaneda, P., Mansilla-Lopez, J., & Garcia-Nunez, A. D. (2026). Enhancing Information Security in Technology Small and Medium-Sized Enterprises: A Metrics-Driven Model Based on ISO/IEC 27001:2022. Engineering, Technology & Applied Science Research. https://doi.org/https://doi.org/10.48084/etasr.17248

[22] Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. Scientific Reports. https://doi.org/10.1038/s41598-024-56706-x

[23] Rananga, N., & Venter, H. S. (2026). An Anti-Sheriff Cybersecurity Audit Model: From Compliance Checklists to Intelligence-Supported Cyber Risk Auditing. Applied Sciences. https://doi.org/10.3390/app16052315

[24] Ruohonen, J., Rindell, K., & Busetti, S. (2025). From cyber security incident management to cyber security crisis management in the European Union. Computers & Security. https://doi.org/10.1016/j.cose.2025.104689

[25] Sánchez-García, I. D., Gilabert, T. S. F., & Calvo-Manzano, J. A. (2023). Countermeasures and their taxonomies for risk treatment in cybersecurity: A systematic mapping review. Computers & Security. https://doi.org/10.1016/j.cose.2023.103170

[26] Zhang, R., Li, H.-W., Qian, X.-Y., Jiang, W.-B., & Chen, H.-X. (2025). On large language models safety, security, and privacy: A survey. Journal of Electronic Science and Technology. https://doi.org/10.1016/j.jnlest.2025.100301

[27] "Treatment episode data set: discharges (TEDS-D): concatenated, 2006 to 2009." U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Office of Applied Studies, August, 2013, DOI:10.3886/ICPSR3012