Towards SLMs Usage in Contextualized Question Generation for Security Information Audits: A Study Using the G-Eval Framework

Vinícius F. Souza[1], **Diego R. Gomes**[2], **Rafael P. B. Mota**[1], **Fernando Aires**[1]

[1] Department of Computing – Federal Rural University of Pernambuco
Recife, PE, Brazil

[2]Department of Informatics Engineering – University of Coimbra
Coimbra, Centro, Portugal

{vinicius.fsouza, rafael.perazzo, fernandoaires}@ufrpe.br, drg@dei.uc.pt

Abstract. *Conducting information security audits today is no trivial task. One key aspect is the quantity and quality of controls to be selected for the assessment, especially considering different types of organizations. This paper evaluates the ability of Small Language Models (SLMs) with Chain-of-Thought to adapt generic questions from the ISO/IEC 27001:2022 and 27002:2022 control lists to specific organizational contexts (hospital, bank, retail, development, and government). Five SLMs were tested using an automatic evaluator based on the G-Eval framework (1–5 scale) and human validation (n=25), revealing that intermediate models (3B–8B) achieve acceptable quality (mean $\geq$ 4.0). Agreement between the automatic evaluator and human evaluators varied by up to 1.2 points in some dimensions, and limitations were also observed in Portuguese. The results suggest feasibility for supporting security audits with human supervision.*

1. Introduction

In today's digital landscape, cybersecurity has become a cornerstone of any organization's success. The growing sophistication of threats, combined with the complexity of technological environments, poses ongoing challenges for security professionals. Assessing the robustness of an organization's security posture is a complex process that requires considerable expertise, time, and resources. Failures in this process can result in financial losses, reputational damage, and disruption of critical services [Spanos and Angelis 2016]. In this sense, information security maturity generally refers to a structured assessment of the level of implementation and refinement of security practices, controls, and processes within an organization [Chrissis et al. 2011].

In this context, Generative Artificial Intelligence (GAI) emerges as a promising tool for supporting cybersecurity [Ferrag et al. 2025]. GAI's ability to generate realistic synthetic data, simulate attack scenarios, and identify complex patterns can significantly optimize vulnerability detection, risk analysis, and the formulation of defense strategies [Yigit et al. 2024].

Despite this potential, there remains a research gap regarding the use of generative language models specifically to support information security maturity assessment processes, particularly in typical auditing activities such as control analysis, evidence

interpretation, and questionnaire adaptation across different organizational contexts. Furthermore, most studies on GAI for security focus on threat detection, event classification, and secure code generation, leaving open the role of these models in structured assessment tasks and decision-making based on control frameworks [Gong et al. 2026].

Large Language Models (LLMs) have become the dominant paradigm in Generative Artificial Intelligence (GAI), as they are trained on massive texts to understand and generate human-like language across a wide range of tasks [Ayyaz and Malik 2024]. In cybersecurity and auditing contexts, LLMs can assist in interpreting complex technical evidence, mapping controls to standards, and tailoring questionnaires to specific regulatory or organizational requirements [Zhang et al. 2025]. Building on these advances, Small Language Models (SLMs) emerge as a complementary approach, aiming to preserve much of this capability while enabling deployment in constrained or on-premise environments.

SLMs are a class of transformer-based language models that prioritize computational efficiency and the ability to be deployed in resource-constrained environments, such as personal devices and edge scenarios, in contrast to large models that run predominantly in data centers [Lu et al. 2025].

Given this context, this work investigates whether SLM models can support information security audit activities, particularly focused in questionnaire adaptation across different organizational contexts. The research question guiding the study is: ***How can small language models support the contextual adaptation of security audit questions across different organizational settings?***

This paper is structured as follows. Section 2 reviews the relevant literature and related works. Section 3 describes the research methodology of this paper. Section 4 presents the results of the methodology execution. Section 5 evaluates and discusses these results. Finally, Section 7 presents the conclusions and future work of this research.

2. Related Works

In the field of information security maturity, Schmitz et al. highlight that practitioners often require additional support to perform reliable maturity-level assessments, as unaided evaluations may be prone to inaccuracies. In their study, professionals assessed predefined maturity levels for a hypothetical scenario based on ISO/IEC 27002 controls and COBIT maturity levels, and the results showed significant variation in assessment quality. The authors also found that participants holding security-related certifications tended to achieve lower deviations from the reference maturity levels, indicating better assessment performance than non-certified participants [Schmitz et al. 2021].

Previous analyses by Rabii et al. highlight the ISO/IEC 27001 and ISO/IEC 27002 standards as recurring references in maturity assessments, while the 2025 review by Brezavšcek and Baggia shows that the sector still faces three practical challenges: complex models, poor sector-specific adaptation, and limited support for smaller organizations; Therefore, the authors argue for simpler frameworks that are easier to adapt to specific sectors and supported by tools that reduce manual effort, such as the use of artificial intelligence [Rabii et al. 2020, Brezavšček and Baggia 2025].

Recent studies have begun to examine the role of LLMs in external auditing,

highlighting their efficacy in routine tasks, but also identifying significant limitations in producing comprehensive, contextually relevant risk assessments. These findings underscore a critical challenge: while LLMs demonstrate potential for time-saving, they often struggle with the depth and contextual nuance required for complex auditing standards [Fotoh and Mugwira 2025]. This reinforces the need for a shift in focus from general-purpose LLMs to the applicability of Small Language Models (SLMs).

Rather than focusing solely on maximizing model size, research on SLMs emphasizes lighter architectures, optimized training strategies, and compression techniques that preserve strong reasoning and comprehension capabilities while keeping memory, latency, and energy costs compatible with everyday use and privacy-sensitive applications [Nguyen et al. 2025]. In this context, SLMs have been gaining prominence in both commercial devices and specialized domains, offering a practical trade-off between performance and computational cost, which makes them particularly attractive for information security and auditing scenarios, where on-premises execution and control over sensitive data are central requirements.

3. Methodology

The methodology of this work is based on the G-Eval framework and was organized into four main stages, as illustrated in Figure 1: (i) definition of data and question-generation tasks based on ISO/IEC 27001:2022 and 27002:2022 controls; (ii) use of Small Language Models (SLMs) with Chain-of-Thought as generative models; (iii) automatic evaluation of the questions by a judge based on the G-Eval framework, following explicit quality criteria; and (iv) human evaluation of a subset of the outputs, used to validate the automatic judge and interpret the results.

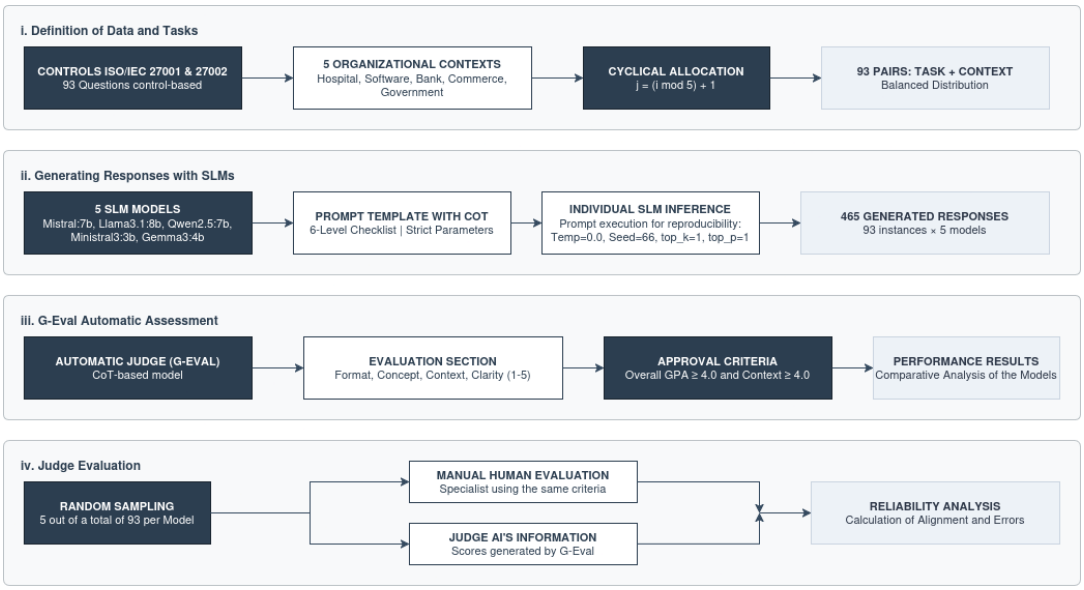

Figure 1. Flowchart of the proposed methodology.

Since G-Eval is a rubric-based evaluation framework that supports task-specific criteria, we defined four dimensions to reflect the requirements of contextualized question generation for security audits: Format, Concept, Context, and Clarity. Format captures

adherence to the required structure, Concept checks semantic and maturity alignment, Context measures adaptation to the organizational scenario, and Clarity assesses whether the response is concise and distinguishable.

To answer the research question, this paper adopted an operational performance criterion. For each dimension evaluated, scores are assigned on a scale of 1 to 5. We consider a model capable of performing the task with acceptable quality when it achieves an overall average of 4.0 or higher and, simultaneously, an average of 4.0 or higher in the Context dimension. Values below this threshold indicate a need for more intensive human review or the model's unsuitability for direct use in assessments. In addition, we later perform a sensitivity analysis by varying this threshold between 3.5 and 4.5 on the validation subset to examine how stricter or more lenient cutoffs affect agreement between the LLM-Judge and the human evaluator.

3.1. Data and tasks

The dataset used in this study served as input for a contextualized question-generation task. Each ISO control was paired with one of five organizational scenarios, prompting the models to generate tailored security audit questions. The dataset comprised 93 generic questions, corresponding to the 93 controls defined in ISO/IEC 27001 and ISO/IEC 27002 [ISO 2022a, ISO 2022b]. To contextualize these tasks, five organizational domains were considered: hospitals, software companies, banking institutions, commercial enterprises, and government agencies. The final prompt structure followed the format task + scenario, distributed according to the following formula:

Let $T = \{t_1, \ldots, t_{93}\}$ be the set of tasks (one control per item) and $C = \{c_1, \ldots, c_5\}$ be the set of contexts. Each evaluation instance is a pair (t_i, c_j), where the context is assigned cyclically according to

$$j = ((i - 1) \bmod 5) + 1,$$

ensuring that all contexts are used in a balanced manner across the 93 tasks. Example of input: "About the equipment maintenance process:, government agency context."

3.2. Generative Models

Small Language Models were chosen for their feasibility of execution in limited computational environments and for reducing the need to expose sensitive data to large-scale external providers. Accordingly, five small-scale language models are evaluated on tasks derived from the controls of ISO/IEC 27001:2022 and 27002:2022, using an experimental setup with Chain-of-Thought and an automated evaluator inspired by the G-Eval framework.

Five widely used generative language models were evaluated: Mistral 7B[1], Llama 3.1 8B[2], Qwen 2.5 7B[3], Ministral 3 3B[4] and Gemma 3 4B[5].

[1] `https://ollama.com/library/mistral`
[2] `https://ollama.com/library/llama3.1`
[3] `https://ollama.com/library/qwen2.5`
[4] `https://ollama.com/library/ministral-3`
[5] `https://ollama.com/library/gemma3`

To assign tasks to these models, a Chain-of-Thought (CoT) scheme was incorporated into the prompt, since recent studies show that explicit exposure to chains of reasoning consistently improves the performance of smaller models on complex reasoning tasks and alignment with larger models [Magister et al. 2023, Ranaldi and Freitas 2024, Liu et al. 2023]. Thus, after iterative cycles of refinement, the following CoT template was developed and applied consistently across all five models.

The prompt template instructed the models to generate a maturity assessment checklist with six response categories, where Level 0 indicates no meaningful response and Levels 1-5 represent increasing maturity, each containing 1-3 sentences and a specific contextual example, following a fixed checkbox output format. The complete template and instructions are available in the project repository (`https://github.com/ViniSnoop/pibic-slm-auditoria`)

Each of the 93 task–context combinations described in Section 3.1 was inserted into this standardized template and sent individually to each of the five generative models, resulting in a total of 465 responses (93 instances per model). The configuration used was as follows: temperature: 0.0, seed: 66, top_k: 1, and top_p: 1, to ensure the reproducibility of the experiment.[6]

3.3. Automatic Judge and Evaluation

To evaluate the results, the G-Eval methodology [Liu et al. 2023] was used, configuring a judge model based on Chain-of-Thought. This judge was executed and after several cycles of prompt refinement, the following CoT was arrived at and used in the judge model.

The judge was configured with explicit rubrics for four dimensions scored 1-5: (1) Format (exact 6-level structure, no extra text); (2) Concept (correct infosec maturity progression, no hallucinations); (3) Context (tailored examples for the specific organization); (4) Clarity (distinct, concrete, concise levels). The complete prompt is available in the project repository. These scores provided the quantitative basis for the comparative analysis of the generative models, enabling calculation of performance variation.

3.4. Judge Evaluation

To evaluate the behavior of the judge constructed using the G-Eval methodology, five task–context combinations were randomly selected from among the 93 evaluated instances. Each of these instances, along with the respective model responses, was manually evaluated using the same rubric adopted by the judge: the four dimensions (Format, Concept, Context, and Clarity) with integer scores from 1 to 5, as defined in the CoT. To assess the judge's reliability, we compared automated scores with human judgments. A confusion matrix was constructed based on a binary threshold, where an overall average score of ≥ 4.0 combined with a context dimension score of ≥ 4.0 denotes 'Pass', while any value below this threshold denotes 'Fail'. This approach allows quantification of alignment and identification of qualitative discrepancies between the AI judge and the human evaluator.

[6]The experiment was designed to maximize reproducibility; however, the use of temperature = 0.0 may influence model behavior by reducing output diversity and making responses more deterministic.

4. Results

4.1. Performance of the tested models on the full dataset (automatic evaluator)

On the full set of 465 outputs, the G-Eval judge appears to rate Ministral-3:3b and Gemma3:4b as meeting the operational criterion, while Llama3.1-8B is borderline due to a slightly lower Context score and Mistral-7B and Qwen2.5-7B fall below the acceptance threshold. Ministral-3:3b attains the highest overall average (4.53) and the strongest Context score (4.7), which suggests, under this specific judge and rubric, that a relatively small model can consistently adapt ISO controls to different organizational scenarios when combined with Chain-of-Thought prompting. Gemma3:4b shows a similar pattern, with balanced scores across dimensions.

Table 1. Average scores of G-Eval judge by dimension and model on the full dataset (n=465)

Model	Format (SD)	Concept (SD)	Context (SD)	Clarity (SD)	Average
Gemma3:4b	4.9 (0.5)	4.2 (0.4)	4.5 (0.7)	4.2 (0.4)	**4.44**
Llama3.1:8b	4.7 (0.6)	4.0 (0.2)	3.9 (0.6)	3.9 (0.3)	**4.14**
Ministral-3:3b	4.8 (0.8)	4.3 (0.5)	4.7 (0.7)	4.3 (0.5)	**4.53**
Mistral:7b	4.3 (0.9)	3.9 (0.4)	3.7 (0.7)	3.9 (0.3)	**3.95**
Qwen2.5:7b	4.3 (0.7)	3.8 (0.4)	3.6 (0.6)	3.6 (0.5)	**3.84**

In contrast, the lower Context means for Llama3.1-8B, Mistral-7B and Qwen2.5-7B may indicate potential weaknesses in scenario-specific adaptation, even when Format and Concepts are reasonably strong. These global results should be interpreted as an initial indication of how intermediate SLMs behave under the chosen G-Eval configuration, rather than as definitive rankings, and later sections examine judge–human discrepancies and sensitivity to the acceptance threshold in more detail.

4.2. Performance of the tested models on human validation subset (automatic evaluator)

Table 2 shows the averages assigned by the automated judge on the 25-item validation subset. All models except Qwen2.5:7b satisfy the operational criterion on this subset, with Ministral-3:3b and Llama3.1-8B obtaining the highest overall and Context scores and relatively low variability. Gemma3:4b and Mistral-7B also meet the threshold but exhibit higher dispersion in some dimensions, especially Format and Context for Gemma3:4b, suggesting occasional deviations from the rubric even when average quality remains acceptable. Qwen2.5-7B falls below the threshold in all dimensions, confirming its weaker behaviour under the judge on this subset.

Table 2. Average scores of G-Eval judge by dimesion and model on the human validation subset (n=25)

Model	Format (SD)	Concept (SD)	Context (SD)	Clarity (SD)	Average
Gemma3:4b	4.2 (1.79)	4.2 (0.45)	4.0 (1.73)	4.4 (0.55)	**4.20**
Llama3.1:8b	4.8 (0.45)	4.0 (0.00)	4.2 (0.84)	4.0 (0.00)	**4.25**
Ministral-3:3b	5.0 (0.00)	4.2 (0.45)	4.8 (0.45)	4.2 (0.45)	**4.55**
Mistral:7b	4.6 (0.55)	4.0 (0.00)	4.0 (0.71)	3.8 (0.45)	**4.10**
Qwen2.5:7b	3.8 (1.10)	3.8 (0.45)	3.4 (0.55)	3.6 (0.55)	**3.65**

It is noteworthy that Ministral-3:3b, despite having the fewest parameters among the successful models, achieved the highest absolute score in context (4.8), suggesting that architectural efficiency and Chain-of-Thought (CoT) are more critical to accuracy in auditing tasks than simply scaling parameters. In contrast, models such as Qwen2.5:7b performed below expectations across all metrics, highlighting difficulties in structuring responses that strictly follow the required checklist format. These quantitative variations underscore the need to carefully select the base model based on the specificity of the target organizational domain.

4.3. Performance of the tested models on human validation subset (human evaluator)

Table 3 presents the scores given by the human researcher. Here, most models meet the criteria, with Gemma3:4b obtaining the highest overall average and Llama3.1-8B and Mistral-7B following closely. The main discrepancy appears in Ministral-3:3b, which received the highest overall score in the judge's evaluation and one of the lowest averages under human assessment, on par with Qwen2.5-7B; however, Qwen2.5-7B still failed the human evaluation due to a low score in Context.

Table 3. Average scores of Human Validation by dimesion and model on the human validation subset (n=25)

Model	Format (SD)	Concept (SD)	Context (SD)	Clarity (SD)	Average
Gemma3:4b	4.6 (0.55)	5.0 (0.00)	5.0 (0.00)	5.0 (0.00)	**4.90**
Llama3.1:8b	5.0 (0.00)	5.0 (0.00)	4.0 (1.41)	5.0 (0.00)	**4.75**
Ministral-3:3b	4.0 (0.00)	4.2 (1.79)	5.0 (0.00)	4.4 (1.34)	**4.40**
Mistral:7b	4.4 (0.55)	5.0 (0.00)	4.6 (0.89)	5.0 (0.00)	**4.75**
Qwen2.5:7b	4.0 (0.00)	5.0 (0.00)	3.8 (0.84)	4.8 (0.45)	**4.40**

4.4. Alignment between the judge's and human validations, and the sensitivity threshold

The alignment between the judge and the human evaluator is partial: both identify Qwen2.5-7B as the weakest model and agree that intermediate SLMs can reach acceptable quality, but they differ on which model should be considered the strongest, with the judge favouring Ministral-3:3b and the human assessment favouring Gemma3:4b. Ministral-3:3b, with strong scores in Format and Context under the judge, illustrates that small

SLMs can generate high-quality questions for adapting ISO controls, although the discrepancies observed in the validation subset highlight the need for human supervision

threshold	tp	tn	fp	fn	accuracy	sensitivity	specificity	kappa	n_pass_judge	n_pass_human
3.5	15.0	2.0	4.0	4.0	0.68	0.789	0.333	0.123	19.0	19.0
3.6	15.0	2.0	4.0	4.0	0.68	0.789	0.333	0.123	19.0	19.0
3.7	15.0	2.0	4.0	4.0	0.68	0.789	0.333	0.123	19.0	19.0
3.8	15.0	2.0	4.0	4.0	0.68	0.789	0.333	0.123	19.0	19.0
3.9	15.0	2.0	4.0	4.0	0.68	0.789	0.333	0.123	19.0	19.0
4.0	15.0	2.0	4.0	4.0	0.68	0.789	0.333	0.123	19.0	19.0
4.1	8.0	6.0	2.0	9.0	0.56	0.471	0.75	0.179	10.0	17.0
4.2	8.0	6.0	2.0	9.0	0.56	0.471	0.75	0.179	10.0	17.0
4.3	8.0	6.0	2.0	9.0	0.56	0.471	0.75	0.179	10.0	17.0
4.4	8.0	6.0	2.0	9.0	0.56	0.471	0.75	0.179	10.0	17.0
4.5	2.0	8.0	1.0	14.0	0.4	0.125	0.889	0.011	3.0	16.0

Figure 2. Sensitivity analysis for thresholds from 3.5 to 4.5 on the validation subset.

Figure 2 summarizes the sensitivity analysis across thresholds from 3.5 to 4.5. The judge's decisions remain unchanged between 3.5 and 4.0, but become more conservative from 4.1 onward.

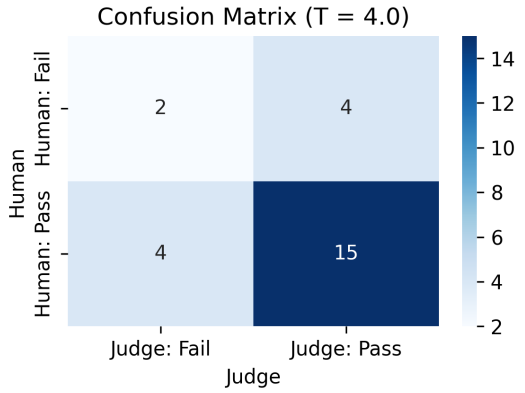

Figure 3. Confusion matrix for the validation subset at threshold $T = 4.0$.

At $T = 4.0$, the judge and the human evaluator approve the same number of instances, but both false positives and false negatives are still present.

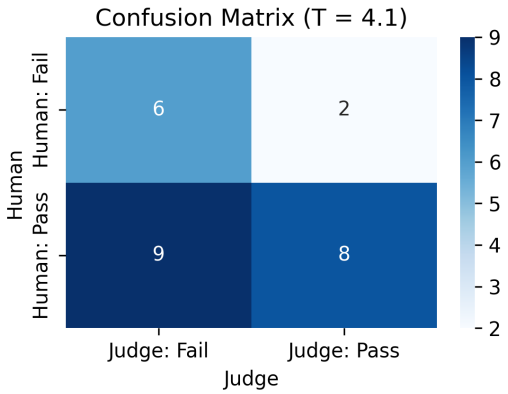

Figure 4. Confusion matrix for the validation subset at threshold $T = 4.1$.

At $T = 4.1$ Figure 4, the judge becomes more conservative, reducing the number of approved instances relative to the human evaluator. The confusion matrix shows that this change is primarily associated with an increase in false negatives, while false positives become less frequent. This behaviour is consistent with the trend observed in Figure 2 and illustrates the trade-off between stricter acceptance criteria and agreement with the human evaluation.

At the strictest threshold $T = 4.5$, the judge rejects substantially more responses than the human evaluator, confirming the increasingly conservative behaviour observed in the sensitivity analysis.

Taken together, the three confusion matrices illustrate how progressively stricter thresholds modify the judge's behaviour. While higher thresholds reduce false positives, they also increase false negatives and decrease agreement with the human evaluator. These findings reinforce the choice of T=4.0 as a balanced operational threshold for the exploratory analysis.

5. Discussion

Although the results favor the use of AI for contextualizing generic information security questions, Spearman metrics show negative correlations in the Concept (ρ = -0.58) and Clarity (ρ = -0.31) dimensions. Given the small validation sample (n=25), these coefficients should be interpreted as exploratory indicators rather than definitive statistical evidence. A prime example is Ministral-3:3b, where the G-Eval judge assigned high scores to semantically superficial or technically incoherent responses.

During the experiments, it was observed that SLMs exhibit lower stability in Portuguese than in English, which justifies prioritizing the latter language. When tested in Portuguese, the models frequently failed to respond or repeated the instructions instead of generating appropriate responses. Studies such as Sectum [Santos 2024], which developed an information security chatbot based on Llama-7B, previously indicated these limitations in Portuguese, an observation corroborated during the experiments in this study.

6. Limitations and Future Work

This study should be interpreted as exploratory. First, the human validation was limited to a small subset of responses and a single evaluator, which constrains the strength of the

Table 4. Representative cases: Judge vs. Human

ID	Context	Judge Avg	Human Avg	Label
G4B-L27	Hospital	5.0	5.0	Perfect example
M3B-L44	Bank	5.0	3.0	Judge conceptual over-optimism
G4B-L68	Software	2.5	4.8	Incorrect judge penalty
L8B-L92	Bank	4.5	4.2	Judge genericness bias
Q7B-L92	Bank	2.8	4.5	Human leniency bias
M7B-L83	Commerce	3.6	4.5	Contextual deficiency alignment

Full evaluation instances available in the project repository.

comparison with the LLM-based judge and increases the possibility of subjective bias. A more robust design would involve multiple evaluators with relevant expertise and explicit agreement measures, such as Cohen's Kappa or the Intraclass Correlation Coefficient.

Second, the acceptance rule based on an average score greater than or equal to 4 should be viewed as an operational choice rather than a universal cutoff. The sensitivity analysis showed that the judge's decisions remain stable between 3.5 and 4.0 and become more conservative from 4.1 onward, supporting 4.0 as a pragmatic threshold in this setting. However, this choice may not generalize to other datasets or evaluation rubrics, and future work should examine principled threshold calibration strategies.

Third, the study does not yet include an external baseline, such as a larger LLM or expert-written questions, which limits the contextualization of the observed results. The comparison among compact models is informative, but it does not establish how much performance is retained relative to stronger systems or manual adaptation. In addition, the divergences observed between G-Eval and human judgment indicate that automated evaluation should be treated as a scalable screening mechanism, not as a substitute for expert review.

Future work should expand the validation set, include multiple human evaluators, and report agreement statistics more systematically. It would also be valuable to compare compact models with larger LLMs and specialist baselines, as well as to test alternative rubrics and thresholding strategies. Extending the framework to other cybersecurity tasks and organizational contexts would further clarify the generalizability of the approach.

7. Conclusions

This work provides evidence that SLMs with Chain-of-Thought can support the adaptation of generic questions from ISO/IEC 27001:2022 and 27002:2022 controls to specific organizational contexts, with models such as Llama3.1:8b and Gemma3:4b meeting acceptable quality criteria according to both automatic and human evaluation. Despite the limitations discussed, such as misalignment between the automatic evaluator and human evaluators in certain dimensions and difficulties with Portuguese, the results indicate promising evidence of practical applications in security audits, provided they are conducted in English and especially in combination with human supervision.

Acknowledgements

The author used DeepL and an AI tool to support translation and language refinement; the final content was reviewed and edited by the author, who takes full responsibility for the manuscript.

References

[Ayyaz and Malik 2024] Ayyaz, S. and Malik, S. M. (2024). A comprehensive study of generative adversarial networks (gan) and generative pre-trained transformers (gpt) in cybersecurity. In *2024 Sixth International Conference on Intelligent Computing in Data Sciences (ICDS)*, pages 1–8. IEEE.

[Brezavšček and Baggia 2025] Brezavšček, A. and Baggia, A. (2025). Recent trends in information and cyber security maturity assessment: A systematic literature review. *Systems*, 13(1).

[Chrissis et al. 2011] Chrissis, M., Konrad, M., and Shrum, S. (2011). *CMMI for Development: Guidelines for Process Integration and Product Improvement*. SEI Series in Software Engineering. Pearson Education.

[Ferrag et al. 2025] Ferrag, M. A., Alwahedi, F., Battah, A., Cherif, B., Mechri, A., Tihanyi, N., Bisztray, T., and Debbah, M. (2025). Generative ai in cybersecurity: A comprehensive review of llm applications and vulnerabilities. *Internet of Things and Cyber-Physical Systems*, 5:1–46.

[Fotoh and Mugwira 2025] Fotoh, L. E. and Mugwira, T. (2025). Exploring large language models in external audits: Implications and ethical considerations. *International Journal of Accounting Information Systems*, 56:100748.

[Gong et al. 2026] Gong, C., Li, Z., and Li, X. (2026). Information security based on llm approaches: A review.

[ISO 2022a] ISO (2022a). Iso/iec 27001: 2022, information security, cybersecurity and privacy protection–information security management systems, requirements.

[ISO 2022b] ISO (2022b). Iso/iec 27002: 2022, information security, cybersecurity and privacy protection information security controls.

[Liu et al. 2023] Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023). G-eval: NLG evaluation using gpt-4 with better human alignment. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

[Lu et al. 2025] Lu, Z., Li, X., Cai, D., Yi, R., Liu, F., Zhang, X., Lane, N. D., and Xu, M. (2025). Small language models: Survey, measurements, and insights.

[Magister et al. 2023] Magister, L. C., Mallinson, J., Adamek, J., Malmi, E., and Severyn, A. (2023). Teaching small language models to reason. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781, Toronto, Canada. Association for Computational Linguistics.

[Nguyen et al. 2025] Nguyen, C. V., Shen, X., Aponte, R., Xia, Y., Basu, S., Hu, Z., Chen, J., Parmar, M., Kunapuli, S., Barrow3, J., Wu, J., Singh, A., Wang, Y., Gu, J., K. Ahmed, N., Lipka, N., Zhang, R., Chen, X., Yu, T., Kim, S., Deilamsalehy, H., Park, N., Rimer, M., Zhang, Z., Yang, H., Mathur, P., Wu, G., Dernoncourt, F., Rossi, R., and Nguyen, T. H. (2025). A survey on small language models. In Angelova,

G., Kunilovskaya, M., Escribe, M., and Mitkov, R., editors, *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 807–821, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

[Rabii et al. 2020] Rabii, A., Saliha, A., Khadija, O. T., and Roudies, O. (2020). Information and cyber security maturity models: a systematic literature review. *Information & Computer Security*, ahead-of-print.

[Ranaldi and Freitas 2024] Ranaldi, L. and Freitas, A. (2024). Aligning large and small language models via chain-of-thought reasoning. In Graham, Y. and Purver, M., editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1812–1827, St. Julian's, Malta. Association for Computational Linguistics.

[Santos 2024] Santos, M. (2024). Sectum: O chatbot de segurança da informação. In *Anais Estendidos do XXIV Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais*, pages 161–168, Porto Alegre, RS, Brasil. SBC.

[Schmitz et al. 2021] Schmitz, C., Schmid, M., Harborth, D., and Pape, S. (2021). Maturity level assessments of information security controls: An empirical analysis of practitioners assessment capabilities. *Computers & Security*, 108:102306.

[Spanos and Angelis 2016] Spanos, G. and Angelis, L. (2016). The impact of information security events to the stock market: A systematic literature review. *Computers & Security*, 58:216–229.

[Yigit et al. 2024] Yigit, Y., Buchanan, W. J., Tehrani, M. G., and Maglaras, L. (2024). Review of generative ai methods in cybersecurity.

[Zhang et al. 2025] Zhang, J., Bu, H., Wen, H., Liu, Y., Fei, H., Xi, R., Li, L., Yang, Y., Zhu, H., and Meng, D. (2025). When llms meet cybersecurity: A systematic literature review. *Cybersecurity*, 8(1):55.