Human-centric Computing and Information Sciences

January 2026 | Volume 16



KCIA
Korea Computer Industry Association

www.hcisjournai.com

HCIS | Human-centric Computing and Information Sciences

RESEARCH (Research Manuscript) Open Access

Human-centric Computing and Information Sciences (2026) 16:06
DOI: https://doi.org/10.22967/HCIS.2026.16.006
Received: December 12, 2024; Accepted: March 15, 2025; Published: January 30, 2026

Generation of Impact Factor-Driven Security Rating Questionnaire Using LLMs for AIoT Applications

Yuna Han[1], Simon Shim[2], Deva Kumar Gajulamandyam[2], Yeji Choi[1], Hyunwoo Lee[1], and Hangbae Chang[3,*]

Abstract

Industry 4.0 has transformed industries by accelerating the adoption of artificial intelligence of things (AIoT); however, it has also led to security risks, such as data leakage. Existing data protection research focuses on network layers, whereas application-layer rating systems rely on subjective evaluations, limiting consistency and applicability. This study proposes a scalable and objective AIoT security rating framework that clarifies ambiguities in the five-question rating system of the Korean Intellectual Property Office and unifies fragmented managerial and technical rating systems. By leveraging large language models (LLMs), the framework integrates a security rating model based on 14 impact factors with automated questionnaire generation. A novel percentage-based measurable constraint ensures objectivity and consistency. The fine-tuned Llama 3.1 8B Instruct model, optimized via direct preference optimization, can enhance customization and question relevance. Results across 13 metrics, including G-Eval and Security G-Eval, highlighted its superiority in aligning questions with prompts, thereby improving specificity and clarity over existing LLMs. A user survey validated its effectiveness with a score of 4.1 out of 5 for correlation and answerability, supported by a Cronbach's alpha of 0.878. This study thus introduces a robust and practical AIoT security rating framework, particularly for the manufacturing, healthcare, and transportation domains, reducing subjective biases while improving applicability.

Keywords

Security Rating Questionnaire, AIoT-based Questionnaire, Impact Factor Scaling, LLM Alignment, Security G-Eval Metric

1. Introduction

With the advent of Industry 4.0, the complementary integration of artificial intelligence (AI) and the Internet of Things (IoT) has driven the evolution of artificial intelligence of things (AIoT)-based technologies, facilitating transformative advancements in digitized intelligent industries such as smart manufacturing, healthcare, agriculture, and transportation [1]. AIoT embodies a fusion of advanced technologies, including AI, 5G/6G networks, cloud computing, and big data, characterized by hyper-

※ This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
***Corresponding Author:** Hangbae Chang (hbchang@cau.ac.kr)
[1]Department of Convergence Security, Chung-Ang University, Seoul, South Korea
[2]Department of Applied Data Science, San Jose State University, San Jose, CA, United States
[3]Department of Industrial Security, Chung-Ang University, Seoul, South Korea

connectivity and hyper-convergence [2]. However, the pervasive deployment of AIoT across diverse domains has also intensified inherent security vulnerabilities, particularly stemming from the connectivity and openness of IoT devices. Moreover, the convergence of real-time computing and AI-driven processes has led to cascading security threats that demand immediate and comprehensive mitigation strategies.

In addition to the development of data privacy protection technologies to counter security threats posed by industrial technology development [3], various approaches have been proposed to solve these problems and ensure the security and privacy of the AIoT environment; these include data encryption and the adoption of blockchain technologies for safeguarding the confidentiality and integrity of data transmitted over networks [4–6]. Despite these efforts, significant challenges persist; the main challenges include threats posed by internal employees [7] and unauthorized access [8] during AIoT management and operation. These persistent vulnerabilities highlight the crucial need for proactive strategies to protect and manage sensitive data effectively. In the Industry 4.0 era, where technological dominance is a strategic priority, information security is crucial not only for ensuring national security but also for maintaining corporate competitiveness. Hence, safeguarding sensitive data has become a paramount concern [9].

Security rating is a promising strategy to proactively manage and secure data. This approach categorizes information based on its value and utility. It was introduced to address ambiguities in traditional confidentiality, integrity, and availability (CIA) evaluations [10]. As illustrated in Table 1, the security rating model includes 14 factors aligned with the corporate data lifecycle, enabling a comprehensive and multi-dimensional assessment of data value. For instance, the Korean Intellectual Property Office (KIPO) has implemented a security rating framework based on five impact factors: cost of information, information utilization, level of information, internal utilization effects, and external leakage risks. This framework provides a structured methodology for assessing enterprise data security [11]. Furthermore, studies on security rating are categorized into managerial and technical approaches, encompassing both manual assessments by security evaluators and automated evaluations utilizing text classification algorithms and models [12].

Table 1. Corporate information security rating model reconstruction [10]

Information lifecycle	Impact factor	Subfactor
Input	Cost of information creation	Manpower, time, and capital
Use	Information utilization	Use frequency and range of use
Technical value output	Level of information	Availability, usability, level of quality, and novelty
Market value outcome	Effect of internal utilization	Value creation potential, marketability, and development maintainability
	Risk of external leakage	Business continuity and competitiveness

Despite the potential of security rating, several challenges impede its advancement at both industry and research levels. This study addresses these issues from three key perspectives. First, the disjointed application of managerial and technical approaches limits the scope of comprehensive analysis [12]. To overcome this challenge, this study proposes an integrated, automated mechanism by leveraging impact factors and sub-factors from the security rating model (Table 1) and fine-tuning large language models (LLMs) to establish a holistic security rating framework. Second, existing security rating questionnaires, such as those provided by the KIPO [11], heavily rely on the subjective judgment and expertise of evaluators. To enhance objectivity, this study introduces constraints explicitly defined in LLM prompts, enabling the generation of standardized questionnaires that can guide evaluators toward data-driven decision-making. Third, traditional metrics for evaluating the reliability and performance of security rating models, such as Cronbach's alpha for determining survey reliability and accuracy for text classification, are not optimized for the unique requirements of security rating tasks. This study proposes novel evaluation metrics tailored to security rating, emphasizing coherence, consistency, fluency, and relevance between subfactors and generated questions.

To address the aforementioned challenges, this study introduces impact factor-driven security rating questionnaires generated using LLMs and applies them within AIoT contexts. The proposed methodology integrates managerial and technical approaches, leveraging automated questionnaire generation based on 14 impact factors (subfactors). The recent questionnaire guidelines proposed by the KIPO recommend conducting security rating assessments using questionnaires based on five impact factors [11]. However, this study identifies a crucial limitation: relying solely on a simplified set of evaluation factors increases the likelihood of subjective interference by evaluators, which fundamentally undermines the accuracy of security ratings. To address this issue, this study utilizes 14 subfactors selected as evaluation criteria, providing a more granular framework. Moreover, to assess the validity of these selected evaluation criteria, the study employs user surveys, focusing on answerability and correlation. Thus, the LLM-generated questionnaires ensure objectivity through measurable constraints, while the scaling technique accommodates varying security policies by adjusting the number of sub-factors based on weighted impact factor scores specific to organizational requirements.

Given the scarcity of labeled datasets, this study employs LLM alignment to optimize questionnaire generation. Optimization is achieved by using 14 foundational questions and fine-tuning Llama 3.1 8B Instruct [13] via direct preference optimization (DPO) [14] to train preference datasets specific to security ratings. Llama 3.1 8B Instruct excels in advanced context-aware text normalization, delivering highly natural and expressive generative performance. Its integration with our security rating questionnaire framework for AIoT applications ensures high-quality output. Additional customization is possible through advanced optimization techniques, such as contrastive preference optimization (CPO) [15] and odds ratio preference optimization (ORPO) [16]. The performance of the trained models can be evaluated using the G-Eval metric [17] and its security-specific variant, the Security G-Eval metric, to assess the alignment between generated questions and prompts. To generate a security rating questionnaire for AIoT applications, this study focuses on three representative domains that handle sensitive information: manufacturing [2], healthcare [18], and smart transportation [1]. The LLM inference model is employed, integrating background information with the basic security rating prompts introduced at an earlier stage. Furthermore, 14 security rating questions are generated for each industry, and quantitative and qualitative performance evaluations are conducted to assess the proposed model and validate the generated questionnaire.

1.1 Motivation

Despite the increasing importance of security rating, its real-world application in the AIoT industry remains limited because of three key challenges: fragmented managerial and technical approaches, subjectivity in security rating questionnaires, and limitations of existing evaluation metrics. First, security rating has been traditionally performed through either manual expert assessments or automated classification models, resulting in a disjointed evaluation process. This separation hinders the development of a comprehensive and consistent security rating framework, resulting in inconsistencies in data security assessments [12]. Second, existing security rating systems rely on manually designed questionnaires, which are highly dependent on evaluator expertise, thereby reducing reproducibility and standardization. Third, traditional reliability metrics fail to capture specific security rating requirements, as they do not effectively measure coherence, consistency, fluency, and relevance in questionnaire-based assessments. To address these limitations, this study integrates managerial and technical approaches, enhances objectivity through measurable constraints, and introduces specialized evaluation metrics tailored to AIoT security ratings.

1.2 Contribution

This study contributes to the field by fostering objective evaluations based on measurable metrics instead of subjective judgment, integrating managerial and technical approaches to deliver comprehensive and

high-quality security rating questionnaires, and providing a scalable solution applicable across diverse AIoT domains.

The main contributions of this study are as follows:

- It proposes a novel security rating questionnaire generation approach based on impact factors tailored to AIoT environments by leveraging Llama 3.1 8B Instruct fine-tuned via DPO. Given the superior performance of the proposed approach across 13 metrics, including G-Eval, Security G-Eval, and reward model evaluation, it provides a scalable and reliable tool for assessing AIoT-specific security risks.
- It introduces percentage-based measurable constraints to ensure the generation of objective, consistent, and clear security rating questions. This approach addresses the limitations of the existing subjective and fragmented evaluation methods, enhancing the applicability of the generated questionnaires in general security rating tasks.
- It uses the fine-tuned LLM framework for diverse AIoT applications, including real-time manufacturing, smart healthcare, and smart transportation, with a one-shot generation prompt. Experimental results revealed significantly higher specificity, numerical precision, and subfactor clarity than state-of-the-art models, such as GPT-3.5 turbo and GPT-4o.
- It conducts a user survey to validate qualitative improvements in the proposed methodology. The results revealed high levels of user satisfaction, with an average correlation and answerability score of 4.1 out of 5. Reliability was confirmed with a Cronbach's alpha of 0.878. Moreover, analysis of robustness indicated minimal dependence on respondent experience (p=0.704).

1.3 Organization

The remaining paper is organized as follows: Section 2 reviews related work. Section 3 presents the proposed security rating questionnaire. Section 4 describes various experimental results. Finally, Section 5 covers the discussion and conclusion.

2. Related Work

This study utilizes LLMs and LLM alignment techniques for generating security rating questionnaires in AIoT environments. To establish the context and highlight the contributions of this research, related studies are examined by focusing on three perspectives: (1) existing security rating methods, (2) recent advancements in applying LLMs to question generation tasks, and (3) challenges posed by security threats in AIoT environments, and existing approaches for addressing these challenges. In Section 2.4, we present a comparative analysis of representative research on LLM-related security to elucidate the distinctive contributions of this study.

2.1 Security Rating Methods

The key considerations for security rating include sensitivity and risk assessment across various methodologies. The CIA triad is one of the most widely adopted frameworks for evaluating data sensitivity [19]. For example, it has been extensively used in the public sector to identify and manage sensitive information, facilitating the proactive mitigation of cybersecurity threats [20]. The CIA triad is valued for its simplicity and readability; however, its evaluation criteria often overlap, particularly in terms of confidentiality, resulting in ambiguities in assessing the sensitivity and importance of information.

To address these limitations, security rating methodologies have been introduced as alternative frameworks. These methodologies evaluate the inherent value of information based on various factors such as utility, novelty, and risk of leakage [10]. As shown in Table 2, two primary security rating approaches are commonly adopted based on the number of dataset labels: binary-class rating and multi-class rating. In binary-class rating, information is classified as either “confidential” or “non-confidential,”

with security requirements determined by corporate policies. The classification scores are calculated based on predefined factors. For instance, CoBAn is a binary-class rating method that clusters confidential information by keyword-based graph representation [21]. Another binary-level security rating method involves the use of the term frequency-inverse document frequency (TF-IDF) matrix for support vector machines [22]. Multi-class rating refers to the categorization of information into multiple levels, such as "confidential," "restricted," "internal," and "public," to provide more granular distinctions and overcome the limitation of binary classification. For example, CES2Vec is a confidentiality-oriented embedding method for detecting confidential information in multi-level scenarios [23]. For extracting the distinctive features of confidential information, a depth-wise separable convolutional neural network (DSCNN) is introduced to balance accuracy and time performance [24].

Both approaches typically draw on techniques from text classification tasks, including tokenization, stemming or lemmatization, feature selection, sampling, and the application of classification models [12]. However, traditional methods often rely on general-purpose text classification techniques, lacking mechanisms tailored to critical or sensitive information. This limitation reduces their effectiveness, particularly in the understanding of security-related contexts and the explainability of results.

Recent advancements have introduced word embedding techniques to enhance confidentiality detection by embedding confidentiality-specific features, instead of solely relying on conventional classification strategies [12]. While these methods show potential, they also highlight a persistent gap between managerial and technical perspectives for information classification. A unified framework that bridges these perspectives is urgently required.

Therefore, in this study, we propose a multi-faceted security rating methodology that integrates impact factors, objective indicators, and LLMs. As shown in Table 2, the existing security rating methods can be broadly categorized into managerial and technical approaches. This study aims to bridge these two strategies by enabling automated evaluations that are objective, transparent, and aligned with actionable criteria.

Table 2. Categorization of security rating methods

Category	Examples of methodology (Label)
Managerial approach (static security rating)	Security rating model for corporate information from an economic security perspective [8]
Technical approach (dynamic security rating)	
Binary-class rating	CoBAn (non-confidential and confidential)
	Neuro-fuzzy inference system for security rating (secret and restricted)
Multi-class rating	CES2Vec (unclassified, confidential, and secret)
	DSCNN (unclassified, confidential, and secret)

2.2 Question Generation using LLMs

Question generation refers to the task of generating questions designed to elicit specific responses at the sentence or paragraph level, typically guided by a reference text [25]. Recent advancements in question generation techniques have not only improved the quality of generated questions but also enabled customization for domain-specific applications [26]. As this study applies question generation to AIoT security rating, it assesses advancements in LLMs across various contexts.

For example, GPT-4 has been effectively employed to generate questions based on user comments, showcasing its ability to produce accurate and relevant outputs [27]. However, its methodology lacks support for closed-response questions—such as classification into "High," "Medium," and "Low" categories, which are crucial for security rating tasks. Similarly, in the educational domain, studies using ChatGPT to create multiple-choice questions for medical graduate examinations demonstrated high performance in terms of clarity, relevance, and task-specific suitability. These studies also revealed increased efficiency in exam preparation because of significant time savings [27, 28]. Unlike human

exam creators who draw on diverse experiences, ChatGPT often relies on constrained reference materials, which may lead to biases influenced by operator inputs. Moreover, previous studies lacked rigorous quality evaluations of the generated questions. Another study assessed the ability of generative models to create algebra-related questions and demonstrated their effectiveness in producing textbook-equivalent problems. However, the work was limited to a single chapter from a textbook, raising concerns about the generalizability of the questions across broader topics [29]. Recent innovations include the optimization of the Llama2 model through DPO alignment, which has been proven to be effective for style transfer tasks such as Socratic questioning. This approach facilitates structured and reflective question generation, underscoring the potential of alignment techniques in tailoring LLM outputs for specific domains [30].

Building on these advancements, this study adopts state-of-the-art alignment techniques to fine-tune LLMs on a domain-specific dataset tailored to AIoT security rating applications. Using this approach, this study aims to enhance the quality and relevance of the generated questions, ensuring they meet the unique requirements of the security rating domain.

2.3 Security and Privacy in AIoT Applications

The integration of AI and IoT, known as AIoT, has become a cornerstone of the Industry 4.0 transformation, fostering digitalization and enhancing operational efficiency across various industries [31]. While AIoT advancements have facilitated data utilization, they have also increased challenges related to data security and privacy. Consequently, a growing body of research has emerged to address these issues in AIoT environments [4].

In the manufacturing sector, AIoT enables proactive defect detection, thereby enhancing product quality. However, the interconnected nature of IoT devices and their reliance on open nodes can introduce significant vulnerabilities related to data confidentiality, integrity, and real-time availability. To mitigate these risks, secure data collection and management systems are crucial [2]. Traditional encryption techniques, such as RSA and ECC, are insufficient for IoT applications, necessitating the development of alternative algorithms, such as NTRUEncrypt. This algorithm enhances message encryption and decryption in IoT environments and protects against various threats, such as black hole attacks, denial-of-service attacks, and replay attacks [5].

In the smart healthcare sector, a key AIoT domain, classification models with accuracy rates exceeding 98% have been developed to detect various conditions such as epileptic seizures [32]. However, the extensive use of big data in this domain raises concerns about data contamination and the potential leakage of sensitive patient information. Although robust security measures are essential, comprehensive solutions to address these challenges remain elusive [33]. Recent research has explored network-based approaches for enhancing security. For instance, a support vector machine combined with a bee colony algorithm achieved 94.56% accuracy in detecting intrusions in smart home environments, facilitating real-time monitoring and alerting of IoT devices [34].

The transportation sector faces unique challenges because of the interconnected nature of vehicles, complicating user data management. The integration of 5G technology and AIoT has shown promise in improving information security and encryption, with threat identification accuracies exceeding 82.1%. These advancements support enhanced decision-making and outcome prediction; however, fundamental issues related to data management persist [35].

Despite these advancements, there is a lack of proactive methodologies for implementing security rating systems tailored to AIoT environments. To fill this gap, this study proposes an LLM-based security rating mechanism specifically designed for AIoT environments by integrating advanced methodologies to enhance data security and privacy management.

2.4 Comparison of LLM Security-Related Studies and This Study

With the rapid advancements in LLMs, security-related research in this domain has primarily evolved

along two distinct trajectories: (1) the development of strategies to mitigate security vulnerabilities intrinsic to LLMs (security for LLMs) and (2) the use of LLMs as tools for addressing external security threats and risks (LLMs for security) [36–41]. As summarized in Table 3, this study focuses on utilizing LLMs to tackle security challenges. Notably, this is the first study to apply a security rating task within the AIoT industry. The proposed methodology offers a significant advantage over existing approaches by enabling cost-efficient security measures that prioritize the value and utility of data, thereby optimizing resource allocation for enhanced data protection.

Table 3. Comparison of LLM-related studies in the security domain

Study	Security task	Advantage	Limitation
Salim et al. [36]	Security for LLMs	Used local differential privacy methods at the edge layer to protect against adversarial attacks in LLM-based chatbot systems, resulting in significantly better performance than models trained on contaminated data.	Faced challenges due to the additional costs incurred by separately constructing and operating edge and cloud layers for federated learning, resulting in inefficient security activities.
Zhao et al. [37]	Security for LLMs	Proposed a modified prompt approach to safeguard input text from malicious prompts, achieving enhanced data protection performance, which could address copyright violations, personal information leakage, and misinformation in LLMs.	Required substantial resources and time, with no established processes for distinguishing between usable and confidential data, resulting in inefficiency.
Ji et al. [38]	Security for LLMs	Was mitigated through a self-reflection methodology, demonstrating effectiveness across diverse medical datasets for addressing the issue of LLM hallucination.	Lacked generalizability, highlighting the need for broader frameworks applicable to diverse domains.
Ferrag et al. [39]	LLMs for security	Were comprehensively summarized, addressing the entire data lifecycle from creation to utilization with regard to security threats and LLM-based defense mechanism.	Lacked emphasis on LLM applications for proactively managing and protecting confidential data.
Zhang et al. [40]	LLMs for security	Proposed the use of ChatGPT 4.0 with test prompting, effectively identifying software bugs.	Before and after vulnerability analysis remains underexplored in current research in terms of proactive prevention and response strategies.
Kojima et al. [41]	LLMs for security	Was introduced to enhance the security of code generated by LLMs, advancing applications in code security.	Before and after vulnerability analysis remains underexplored in current research in terms of proactive prevention and response strategies.
This study	LLMs for security	Was proposed to support cost-effective security activities, demonstrating the applicability of the generated questionnaire to both AIoT and general security domains.	Required high-performance resources for the training process and application outcomes.

As shown in Table 3, existing studies have largely overlooked crucial aspects, such as the protection of confidential data and the implementation of proactive prevention mechanisms. To overcome these limitations, our proposed study introduces a cost-effective approach by developing a security rating

questionnaire, facilitating scalability and applicability across various security domains, including AIoT. Furthermore, Section 4.2 explicitly addresses further analysis on the significant resource demands associated with achieving high performance using LLMs. We explore practical trade-offs between efficiency and accuracy by evaluating various models such as Mistral 7B Instruct v0.2 and our proposed Llama 3.1 8B Instruct model, and ultimately suggest optimized solutions to balance resource efficiency and performance.

3. Proposed Methodology for Generating Security Rating Questionnaire

This study aims to automatically generate security rating questionnaires and apply the fine-tuned LLM model to the AIoT industry. Our approach provides objective evaluation items and questions, moving beyond traditional security rating methods that rely on evaluator subjectivity. Moreover, the quality and relevance of the generated questions are assessed through quantitative metrics. The high-quality LLM model for question generation is simply applied to AIoT-based industries, such as the real-time manufacturing industry, AIoT-based healthcare systems, and the smart transportation industry. The methodology is divided into four main steps Fig. 1: (1) prompt engineering for security rating question generation, (2) construction of the security rating question dataset, (3) fine-tuning of the LLM model with the alignment method depending on the number of test questions, and (4) generation of the security rating questionnaire targeting the AIoT industry via the fine-tuned model.

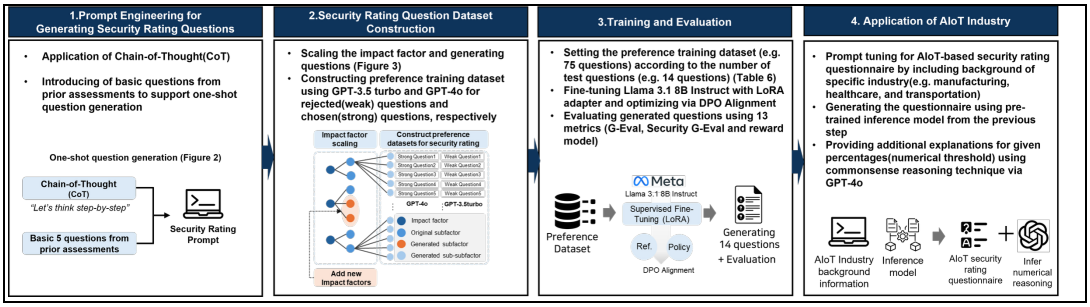


Fig. 1. Overall framework of the proposed methodology for generating security rating questionnaires targeting the AIoT industry.

In the first phase, we aim to construct domain-specific question generation prompts for the security rating task by adapting a general prompt used for typical generation tasks [42]. As shown in Fig. 2, users are instructed to avoid open-ended or vague criteria to address ambiguity and subjectivity, which is known to be a typical challenge in existing security rating tasks [43]. The zero-shot question generation prompt is further refined, by employing the chain-of-thought technique using the phrase “let’s think step by step” [44] and upgraded to a one-shot question by providing an example questionnaire for five impact factors: cost of information creation, level of information, information utilization, internal utilization effects, and risk of external leakage [10]. Although few-shot prompts are generally more effective for richer in-context learning [45], this study generates an advanced one-shot prompt because the assessment criteria provided by the KIPO are limited and pre-defined [11]. Additionally, this study addresses potential issues with traditional questionnaires, such as the lack of specific criteria (e.g., the use of vague terms, such as “high,” “medium,” and “low”), which may lead to subjective responses. Studies have shown that scaling responses with quantitative ratings can help reduce response bias [46]. To address these potential issues, this study incorporates a measurable threshold within the prompt, thereby providing quantitative evaluation criteria for more objective assessments. For instance, a percentage of the total information development time over the past year is suggested rather than measuring the time

spent on information creation and maintenance in exact hours, making the evaluation applicable across organizations with varying internal security policies.

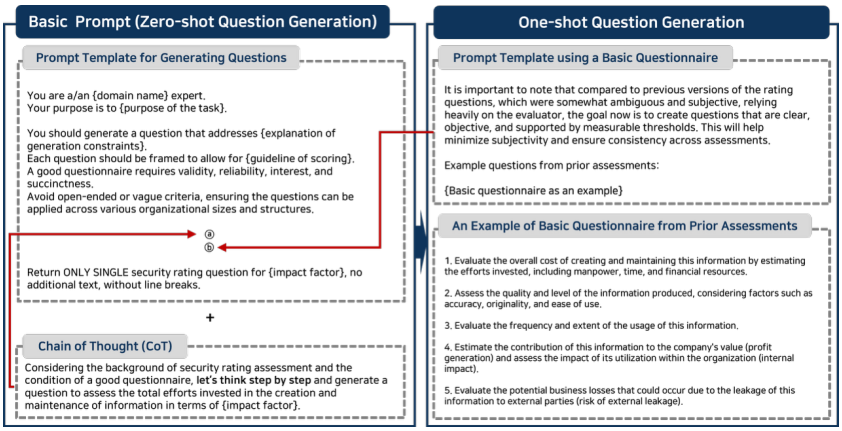


Fig. 2. Prompt template for security rating question generation.

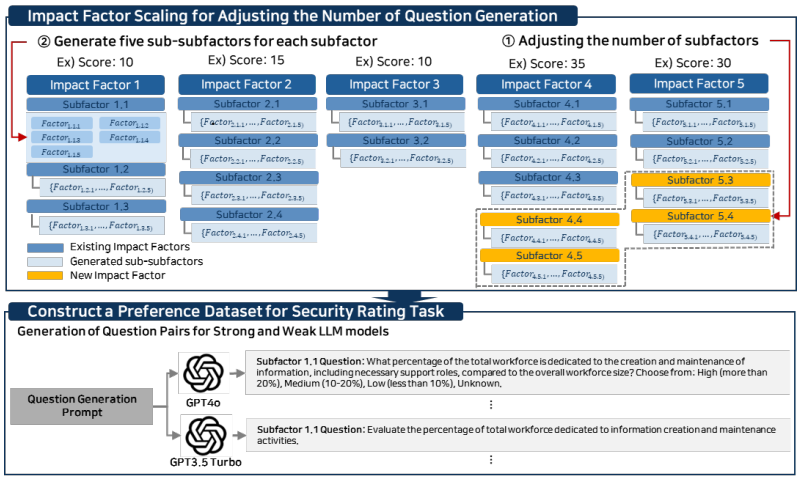


Fig. 3. Proposed impact factor scaling for adjusting the question generations number to construct a preference dataset in the security rating task.

As no prior reference or ground truth exists for security rating questions, the second phase focuses on constructing a preference dataset to determine the effective questions for use as a training dataset. As shown in Fig. 3, the preference dataset is constructed by adjusting the number of questions, subdividing the impact factors into subfactors, and referencing the structure of training datasets used in reward models [47]. Initially, the number of questions can be determined based on the complexity or difficulty of the information being evaluated. We provide 14 default impact factors from prior studies and scale the number of generated questions based on the ratio of the impact factor score to the total score. Let W_i represent the score weight of the i-th impact factor, and w_i represent the ratio of the i-th impact factor, calculated as $w_i = \frac{W_i}{Total\ Score}$. For example, the five existing impact factors account for 10, 15, 10, 35, and 30 points out of 100, with their subfactors varying in number (e.g., 3, 4, and 2). As "internal utilization effect" (the fourth impact factor) and "risk of external leakage" (the fifth impact factor) have higher weights but fewer evaluation items, weighting the value of existing information may improve reliability [48].

Therefore, we propose generating more subfactors as the number of questions increases, with the number of questions for each k-th iteration calculated as $N_k = 14 \times k$, where $k \in \mathbb{N}$ represents the iteration count. The total number of subfactors for each impact factor (I_i) is represented by M_k, which is the sum of all subfactors. The number of generated questions is then adjusted proportionally, as described in Equations (1) and (2). To construct the preference dataset for the security rating task, each subfactor generates five sub-subfactors, as shown in Fig. 3. The datasets utilized in this study are categorized into three types: a preference dataset for the training phase, a general security rating dataset for testing, and a dataset for AIoT applications. Unlike the training phase, the dataset employed in the testing phase represents generated questions derived from the inference model of the proposed fine-tuned model, Llama 3.1 8B Instruct. Furthermore, for AIoT applications, the dataset is adapted using the prompts and presented for testing on the inference model. For each question in the training process, the selected response is generated by GPT-4o, while the rejected response is generated by GPT-3.5 turbo. This distinction enables comparative evaluation between high- and low-quality responses, facilitating the refinement of the question generation process. A more detailed explanation of the source of test prompts in the general security rating tasks and AIoT applications is provided at the end of Section 3 and in Fig. 4.

$$I_i = \lceil w_i \times N_k \rceil \tag{1}$$

$$M_k = \begin{cases} 14, & k = 0 \\ \sum_{i=1}^{n} I_i, & k \geq 1 \end{cases} \tag{2}$$

In the third phase, a pre-trained LLM with a human-like linguistic structure is fine-tuned to enhance the quality of the generated questions. Additionally, alignment is performed to align human values and goals [49]. This study applies DPO, i.e., a human preference optimization without reward model training [50], to the Llama 3.1 8B Instruct model to optimize it for generating 14 security rating questions, which can be applied to general security rating tasks without being limited to a specific industry. Let us assume that π_θ and π_{ref} refer to optimal policy and reference policy, respectively. As shown in Equation (3), the DPO objective function regarding optimal policy computes the maximum likelihood for the policy with parameterization when y^+ and y^- denote the preferred and rejected items of x, respectively.

$$\mathcal{L}_{DPO} = -\mathbb{E}\left[\log \sigma(\beta(\log \frac{\pi_\theta(y^+|x)}{\pi_{ref}(y^+|x)} - \log \frac{\pi_\theta(y^-|x)}{\pi_{ref}(y^-|x)}))\right] \tag{3}$$

We also present the use of other state-of-the-art alignment techniques, such as CPO and ORPO, to support the generation of more than 14 questions tailored to address specific security policies based on the requirements of institutes or organizations. CPO alignment removes the reference model from the traditional DPO formulation, enabling preference label computation without a reference model. In contrast, ORPO adjusts reward distribution by computing penalties between rejected and chosen responses. By conducting a comparative analysis of state-of-the-art LLM alignment techniques, we aim to develop an optimal methodology for generating security rating questionnaires.

The general security rating is evaluated using the G-Eval metric, which represents the state-of-the-art assessment framework for evaluating the quality of generated questions in terms of coherence, consistency, fluency, and relevance [15]. However, G-Eval primarily emphasizes the assessment of the linguistic structure between the prompt and generated questions, making it less suitable for evaluating subfactor-based questions in our task. As a security rating questionnaire must assess whether the generated questions effectively reflect the intended subfactors, this study proposes a novel evaluation method, Security G-Eval, to address this limitation. Security G-Eval enhances the evaluation process by incorporating subcategories and generated questions within the context of security rating, ensuring a more comprehensive assessment.

Table 4 summarizes the key differences between G-Eval and Security G-Eval in terms of conceptual focus and scoring criteria. G-Eval primarily evaluates clarity, constraint adherence, grammar, and relevance between the prompt and generated questions, while Security G-Eval places greater emphasis on the relationship between the generated questions and the underlying subfactors or impact factors, evaluating each item accordingly. For example, for the coherence metric, this study emphasizes that "the question should not only be clear but also directly tie into the subcategory's impact factors, reflecting how the factors influence the evaluation process" (Fig. A1). For the consistency metric, the focus shifts to the following: "emphasize the evaluation of subcategories and generated questions rather than just the prompt and question" (Fig. A2). For the fluency metric, we expand the evaluation scale from 3 to 4 to assess both grammatical accuracy and the expression of subcategories (Fig. A3). Finally, for the relevance metric, we add "should directly relate to how the subcategory influences the overall evaluation" (Fig. A4).

Table 4. Comparison of evaluation metrics between G-Eval and Security G-Eval

Metric			Explanation
Coherence	G-Eval	Definition	Evaluates the clarity of the generated question and its alignment with the prompt
		Score criteria	1–5 (A higher score means better structure and clarity in logical flow)
	Security G-Eval	Definition	Evaluates the alignment between the generated question and impact factors, focusing on subfactor alignment
		Score criteria	1–5 (A higher score means better alignment with security subcategories and impact factors)
Consistency	G-Eval	Definition	Evaluates the degree of compliance with instructions or constraints provided in the prompt of the generated question
		Score criteria	1–5 (A higher score means better adherence to guidelines and constraints)
	Security G-Eval	Definition	Evaluates compliance with constraints related to subfactors and impact factors in the generated question
		Score criteria	1–5 (A higher score means better adherence to measurable security criteria and constraints)
Fluency	G-Eval	Definition	Evaluates any issues related to grammar, spelling, and sentence structure.
		Score criteria	1–3 (A higher score means better grammatical accuracy and readability)
	Security G-Eval	Definition	Evaluates the expression of the subcategory and impact factor in the generated question, ensuring alignment not only in grammar but also in coherence with the subcategory and impact factor
		Score criteria	1–4 (A higher score means better readability and use of security-appropriate language)
Relevance	G-Eval	Definition	Evaluates the relevance between the prompt and the generated question
		Score criteria	1–5 (A higher score means better alignment with the main topic, avoiding redundancy)
	Security G-Eval	Definition	Evaluates how well the generated question relates to the subcategory and impact factor
		Score criteria	1–5 (A higher score means better focus on security-specific aspects and impact factors)

In other words, the key strength of Security G-Eval lies in its ability to assess the relevance of generated questions related to specific subcategories, enabling a more nuanced evaluation that goes beyond the capabilities of G-Eval. For instance, consider the following question: “What percentage of your organization’s competitive advantage might be lost if sensitive company information was leaked externally? Rate as High if above 50%, Medium if between 20% and 50%, or Low if below 20%.” While this question achieves the highest possible score (3 points) under the G-Eval criteria, it lacks explicit indicators for competitiveness, resulting in an average score (1.5 points) under the Security G-Eval criteria. In contrast, consider the revised version of the aforementioned question: “What percentage of your organization's competitive advantage, specifically in terms of market share or proprietary technology, might be lost if sensitive company information was leaked externally? Rate as High if above 50%, Medium if between 20% and 50%, or Low if below 20%.” This explicitly incorporates key competitiveness-related indicators such as market share and proprietary technology, thereby aligning more closely with the relevant impact factors. Consequently, this version can achieve the highest possible score (4 points) under Security G-Eval.

All specific scoring methodologies with examples for each Security G-Eval metric are presented in Appendix A.

In the fourth phase, we finally focus on generating a security rating questionnaire specifically tailored to AIoT applications by leveraging the fine-tuned LLM from the previous step. Following the research trend of applying LLM inference directly to real-world scenarios [51], this study extends this approach to the AIoT domain without requiring additional training, as illustrated in Fig. 4. The base prompt structure is adapted from the general security rating questionnaire prompt in Fig. 2, with modifications made to include AIoT-specific content. The prompt begins by specifying the target AIoT industry and incorporating background information drawn from relevant literature, forming a one-shot example prompt that employs a CoT approach. This method enables the generation of questionnaires that capture the unique characteristics of each industry highlighted in the background statement. For instance, Figs. B1, B2, and B3 in Appendix B showcase industry-specific features, such as “machine operation” for real-time manufacturing, “healthcare records” for smart healthcare systems, and “autonomous vehicles” for smart transportation. Moreover, percentage-based measurable criteria for each generated question can be explained by the search functionality of GPT-4o to help security experts make evaluation decisions.

Your task is to generate security rating assessment questions in **{AIoT industry}** that cover 5 impact factors and 14 subcategories. These factors include the cost of information creation at the input stage (manpower, time, capital), the level of information at the output stage (availability, usability, quality, novelty), information utilization (frequency of use, range of use), the internal utilization impact (value creation potential, marketability, development maintainability), and the risk of external leakage (business continuity, competitiveness).

Consider the background of **{AIoT industry}**, as follows: **{Background of AIoT industry}**

Now, let's think step by step and generate a question to assess the effect of **{Impact factor}** of **{Subcategory}** in **{AIoT industry}.**

Each question should be framed to allow for High/Medium/Low responses with clearly defined thresholds or percentages. A good questionnaire requires valid, reliable, interesting and succinct. Avoid open-ended or vague criteria, ensuring the questions can be applied across various organizational sizes and structures.

Return ONLY SINGLE security rating question for **{Subcategory}** no additional text, without line breaks.

Fig. 4. Test prompt for AIoT industry applications.

We aim to enhance the credibility and transparency of the results by providing a reasonable justification for the numerical threshold used in the AIoT security rating questionnaire. To achieve this, we use the question “What could be a reasonable justification for categorizing the percentage of the following generated question by considering the originally generated question prompt?” along with the background in the prompt to derive reasoning based on commonsense [52]. By leveraging the deep search capabilities

of GPT-4o, we can expand the search function and provide a range of justifications based on rich information. A detailed example is provided in Appendix C (Fig. C1).

4. Experimental Results

In this study, we generated information security rating questions using custom training and test datasets based on identified impact factors [17]. The baseline dataset consisted of 75 training samples and 14 test samples. The training dataset, as described in Section 3, was the preference dataset consisting of instruction–response pairs, including accepted and rejected responses. The test dataset was initially composed of 14 security rating questions, as the goal was to use it for AIoT applications. However, to ensure broader applicability across various organizations instead of being limited to a specific company or institution, this study proposes a training dataset based on the impact factor scaling technique. This approach accommodates scenarios where additional test questions are required, providing a flexible solution for different organizational requirements.

The experiments were conducted using an NVIDIA L4 GPU and an A100 server, featuring 80GB RAM and an Intel Xeon Platinum 8375C processor, in a Colab environment. The system was configured using Ubuntu 20.04 LTS and Python 3.10. For the baseline model, we employed the Llama 3.1 8B Instruct model from the Unsloth library, fine-tuned with the LoRA adapter. For aligning the LLM, we employed DPO with the following hyperparameters: batch size = 8, epoch = 15, learning rate = 5×10^{-6}, logging step = 1, maximum sequence length = 1024, and maximum prompt length = 512. For CPO and ORPO, we modified the settings to batch size = 2, epoch = 15, and logging step = 57. The primary focus of the experiments was to evaluate the suitability of the constructed training dataset, select the most appropriate pre-trained LLM, and align the selected LLM with the custom training data.

4.1 Results of Adjusting the Training Dataset based on Impact Factor Scaling

In this study, we generated 14 questions based on the security rating model initially proposed in the original research paper [10], focusing on flexibly adjusting the number of test questions according to the proposed impact factor scaling technique. This approach aims to accommodate cases requiring more detailed evaluations tailored to each organization's security policy. To generate the security rating questions introduced in this study, it is crucial to determine the number of test samples and estimate the required training data accordingly. Based on the guidelines outlined in Section 3, we established our training dataset to be approximately five times larger than the test dataset. As shown in Table 5, this ratio ensures a balance: a small dataset could degrade question generation performance, while an excessively large dataset could weaken the correlation between subfactors (test data) and sub-subfactors (training data). Furthermore, we constructed the training and test datasets independently through redundancy reviews, ensuring no overlap between the datasets.

Table 5. Results of adjustment based on the impact factor scaling method

Iteration (k)	Before application (N_k)	After application (M_k)	Training dataset (M)	Impact factor 1 (D_1)	Impact factor 2 (D_2)	Impact factor 3 (D_3)	Impact factor 4 (D_4)	Impact factor 5 (D_5)
0	14	14	75	3	4	2	3	2
1	14	18	95	3	4	2	5	4
2	28	28	145	3	4	3	10	8
3	42	42	215	4	6	4	15	13
4	56	57	290	5	8	6	20	17
5	70	71	360	7	11	7	25	21

Before evaluating the proposed method across 13 metrics for generating a security rating questionnaire, we analyzed various subfactors to assess the potential for overfitting within the derived subfactors. For instance, overfitting can be indicated by a weakened correlation between the generated questions and associated prompts. The experimental results demonstrated comprehensively high correlations when the number of subfactors increased to 71 (Table 6). However, a significant decline in correlation was noted when the number of subfactors increased to 83. These findings suggest that the utilization of 83 subfactors can lead to overfitting, thereby reducing the generalizability of their relationship with the prompts. For example, for the first impact factor, the 71 test questions were directly associated with the "cost of information," encompassing various aspects such as outsourcing and resource allocation. In contrast, the 83 test questions included other factors, such as "carbon footprint impact," a subfactor from an environmental perspective that is less relevant to the impact factor. Similarly, for the second impact factor, the 71 test questions focused on data granularity, comprehensiveness, and adaptability, which are key determinants of information quality. In contrast, the 83 test questions included other factors, such as "cultural adaptability," which is not related to information quality. Thus, the subfactors derived from the 71 questions are more directly related to the impact factors, while those derived from the 83 questions are either indirectly related or unrelated. Consequently, 14 to 71 test questions appear to be optimal for inclusion in the questionnaire.

Table 6. Derivation results of subfactors for 71 and 83 question datasets

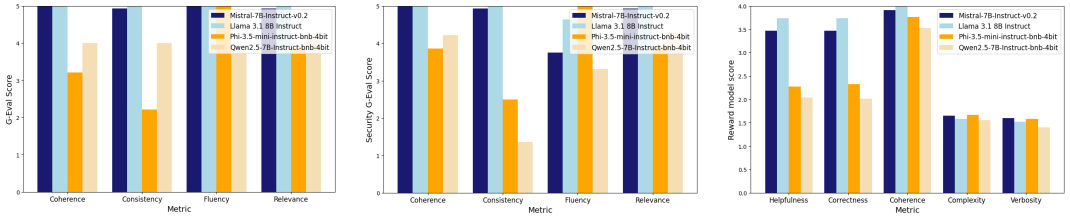
No.	Impact factors	Question	Derived Subfactors
1	Cost of information creation	71	Resource allocation efficiency, technological dependency, and Outsourcing requirements
		83	Carbon footprint impact
2	Level of information	71	Data granularity, information comprehensiveness, contextual adaptability, cross-system compatibility, and innovation relevance
		83	Information relevance consistency and cultural adaptability
3	Information utilization	71	User engagement, automation potential, and data sharing frequency
		83	User feedback integration
4	Effect of internal utilization	71	Process optimization, employee satisfaction, decision-making enhancement, strategic alignment, cost benefit realization, organizational resilience, talent retention, cross-functional collaboration, operational continuity, and competitive advantage
		83	Adaptability to regulatory changes, process flexibility, real-time data update capability, collaborative culture support
5	Risk of external leakage	71	Supply chain disruption, customer attrition, security compliance costs, regulatory scrutiny, brand equity damage, loss of innovation edge, investor confidence erosion, and vendor trust loss
		83	Reputational consistency, data ownership complexity, geopolitical risk impact, and supply chain resilience

Based on the findings presented in Table 4, we refined the number of subfactors within each impact factor to align with 14 to 71 test questions. The results are summarized in Table 1. For example, when generating 18 adjusted questions, we introduced two additional subfactors, namely "innovation potential" and "operational efficiency" under impact factor 4. Similarly, we added "reputation damage" and "legal liability" under impact factor 5. Notably, the amount of training dataset required increased proportionally with the number of test questions in the test dataset D.

4.2 Performance Evaluation based on Pre-trained LLMs

To generate security rating questions without ground truth references, we used LLMs pre-trained on

language sequences by referencing various data sources. As shown in Fig. 5, we fine-tuned four state-of-the-art LLM models, namely Mistral 7B Instruct v0.2 [20], Llama 3.1 8B Instruct [13], Phi-3.5-mini Instruct-bnb-4bit [53], and Qwen 2.5-7B Instruct-bnb-4bit [54], which are widely used in question generation tasks [55], via DPO alignment. The evaluation, conducted using 14 test questions, utilized 13 metrics: G-Eval [17] was used to assess coherence, consistency, fluency, and relevance; Security G-Eval was used to assess coherence, consistency, fluency, and relevance; and the Nemotron-4 340B-based reward model [50] was used to assess helpfulness, correctness, coherence, complexity, and verbosity (Fig. 5). We used all 13 metrics to assess the quality of the generated language. G-Eval and Security G-Eval focus on analyzing the overall structure of sentences and the relationship between prompts and generated questions. Notably, Security G-Eval builds upon the core principles of G-Eval by incorporating a finer-grained subcategory focus, enabling a more nuanced evaluation of security-related aspects. Meanwhile, the reward model extends the evaluation by assessing practical applicability, such as helpfulness and correctness, in security rating questionnaires. In other words, G-Eval and Security G-Eval primarily evaluate contextual quality, while the reward model employs a probabilistic scoring mechanism based on alignment with expected outputs. As Security G-Eval is also based on an LLM-driven evaluation, we carefully considered potential biases. Furthermore, as the reward may penalize well-explained yet lengthy responses, potentially affecting fairness in evaluation, we employed multiple complementary metrics to ensure a balanced assessment.



(a) (b) (c)

Fig. 5. Comparison results of state-of-the-art LLM models with DPO alignment assessments: (a) G-Eval Score (b) Security G-Eval (Ours), and (c) Reward models.

To ensure consistency in evaluation and analysis, we normalized all metrics to a maximum score of 5 in G-Eval and Security G-Eval. As depicted in Fig. 5(a) and 5(b), the Llama 3.1 8B Instruct model outperformed the other three state-of-the-art LLM models. Mistral 7B Instruct v0.2 was the second-best performing model; it exhibited better training efficiency than the Llama 3.1 8B Instruct model when trained on an A100 server. While both models had the same number of trainable parameters (167,772,160) under identical resource conditions, Mistral 7B Instruct v0.2 exhibited a 7.7% faster training speed, with a training time of 556.83 seconds (compared to 603.28 s for Llama 3.1 8B Instruct), a steps-per-second rate of 0.242 (compared to 0.224), and a samples-per-second rate of 2.02 (compared to 1.865). This suggests that Mistral 7B Instruct v0.2 is more time efficient. However, the trade-off between training speed and output quality implies that reducing the training time can slightly degrade performance and compromise the quality of the generated questions.

Meanwhile, the superior performance of the Llama 3.1 8B Instruct model suggests that it more accurately reflects prompts and generates well-structured questions with clear articulation of subcategories, demonstrating enhanced linguistic structure and grammatical accuracy. Furthermore, as shown in Fig. 5(c), the Llama-based model effectively generates helpful, accurate, and coherent security rating questions while minimizing sentence complexity and verbosity.

Consequently, rather than prioritizing efficiency, this study aims to ensure inference performance by generating high-quality questionnaires, thereby facilitating applications across various AIoT industry domains. Therefore, we selected the Llama 3.1 8B Instruct model for training and testing in the security rating questionnaire.

4.3 Experimental Results and Analysis of LLM Alignment based on the Number of Questions Generated

To customize the proposed dataset, we applied the state-of-the-art effective LLM alignment methods DPO, CPO, and ORPO to Llama 3.1 8B Instruct. As shown in Tables 7 and 8, DPO generally outperformed on most metrics with 14 test questions, while CPO and ORPO exhibited superior performance with 71 test questions. This finding suggests that each method has its strengths depending on the size of the dataset. These performance differences stem from the inherent characteristics of the alignment methods, DPO optimizes by increasing the probability of selected questions using the cross-entropy loss function but is prone to overfitting with larger datasets because of the lack of penalty for rejected responses. This tendency becomes more pronounced as the number of generated questions increases from 14 to 71 and then to 83. Meanwhile, CPO and ORPO perform better with larger sets of generated questions, such as 71 or 83 questions, than with a smaller set of 14 questions. Notably, at the optimal number of 71 questions, both methods exhibit relatively low performance degradation. Furthermore, ORPO maximizes the gap between selected and rejected questions via the odds ratio, providing clearer insights as the dataset quality improves with growth. These results suggest that the generation of 83 questions excessively fragments the impact factor into subfactors, potentially diminishing effectiveness. Therefore, the selection of the appropriate alignment method depends on the number of evaluation points and requires careful consideration of the dataset size for optimal results. As this study aims to generate 14 security rating questions for AIoT applications, we used DPO for LLM alignment.

Table 7. LLM alignment evaluation based on the number of questions generated, as assessed using G-Eval and Security G-Eval

Model	Q.	Coherence		Consistency		Fluency		Relevance	
		G-Eval	Security G-Eval	G-Eval	Security G-Eval	G-Eval	Security G-Eval	G-Eval	Security G-Eval
DPO	14	5	5	5	5	3	3	5	5
	71	4.9718	4.9718	4.8802	4.8802	2.8732	2.8732	4.9718	4.9718
	83	4.9518	4.7710	2.8433	4.8795	4.8072	4.8855	3.6385	4.8915
CPO	14	4.9285	4.9642	4.9285	4.9642	3	3.9285	4.9285	4.9285
	71	5	5	4.9436	4.9647	3	4	4.9859	4.9859
	83	5	4.9638	2.9759	4.9879	4.9819	4.9819	3.9397	4.9759
ORPO	14	4.9285	4.9285	4.9285	4.9285	2.8571	2.8571	5	5
	71	5	5	4.8802	4.8802	3	3	5	5
	83	4.9879	4.9036	3	4.9277	4.9698	4.9578	3.9879	4.9518

Table 8. LLM alignment evaluation based on the number of questions generated as assessed using the reward model

Model	Q.	Helpfulness	Correctness	Coherence	Complexity	Verbosity
DPO	14	3.7377	3.7399	3.9877	1.5842	1.5206
	71	3.5773	3.5689	3.9308	1.5641	1.4718
	83	2.1389	2.1554	3.7008	1.4560	1.2773
CPO	14	3.7310	3.7421	3.9888	1.6093	1.5373
	71	3.7638	3.7964	3.9837	1.6764	1.5560
	83	3.7592	3.7821	3.9585	1.6583	1.5404
ORPO	14	3.4397	3.4628	3.9386	1.6735	1.5943
	71	3.8153	3.8369	3.9903	1.6638	1.5210
	83	1.9643	2.0266	3.6890	1.5674	1.4119

Table 9. Example of the general security rating questionnaire

No.	Impact factor	Generated question
1	Manpower	What percentage of the total workforce is directly engaged in creating, maintaining, and updating the information assets, considering that more than 20% is High, 5%–20% is Medium, and below 5% is Low?
2	Time	What percentage of the total information development time is spent on information creation and maintenance within the last year, which is above the industry average of 60%? High if above 80%, Medium if between 60% and 80%, and Low if below 60%.
3	Capital	What percentage of your annual capital expenditure is directly attributed to the creation and maintenance of information assets? Low: <5%, Medium: >5%, and High: >20%.
4	Availability	What percentage of the organization's critical information is available to the relevant stakeholders at the right time? High: 90% or more, Medium: between 50% and 90%, and Low: less than 50%.
5	Usability	What percentage of your organization's employees can easily understand and use the information provided to them, considering that more than 75% of employees can use the information with minimal support is High, between 25% to 75% is Medium, and Low less than 25% is Low?
6	Level of quality	What percentage of your organization's information is of high quality, meaning it is accurate, complete, and reliable, and can be used directly in decision-making without significant rework? High if above 80%, Medium if between 50% and 80%, and Low if below 50%.
7	Novelty	What percentage of your organization's information is new or original, and not a copy or adaptation of existing information, within last year? High: above 40%, Medium: between 20% and 40%, and Low: below 20%.
8	Use frequency	What percentage of the shared information is used by the organization more than five times a month? High if more than 50%, Medium if between 26% and 50%, and Low if less than or equal to 25%.
9	Range of use	What percentage of the organization's employees use the shared information at least once a month? High if more than 50%, Medium if between 26% and 50%, and Low if less than or equal to 25%.
10	Value creation potential	What percentage of the organization's total revenue can be directly attributed to the use of the information created within the past year, and is this percentage 20% or higher? High if 40% or higher, Medium if between 20% and 39%, and Low if below 20%.
11	Marketability	What percentage of the products or services offered by your organization are based on the information you are assessing, and how much do these products or services contribute to the overall revenue of your organization, considering that above 20% is High, 10%–20% is Medium, below 10% is Low?
12	Development maintainability	What percentage of your organization's development projects could be easily maintained and updated within the last year, considering the complexity of the projects and the availability of documentation and support? High if above 80%, Medium if between 40% and 80%, and Low if below 40%.
13	Business continuity	What percentage of your business operations would need to be unavailable for more than a week due to a security breach before it would have a major impact on your company's ability to operate normally? High: >25%, Medium: 10%–25%, and Low: <10%.
14	Competitiveness	What percentage of your organization's competitive advantage do you think would be lost if a key piece of information was leaked externally? High if over 50%, Medium if between 20% and 50%, and Low if below 20%.

4.4 Results of Security Rating Questionnaire Generation based on Subfactors

This study focuses on the generation of security rating questionnaires using 14 subfactors. To optimize the number of security assessment questions, the impact factor scaling technique is proposed. This technique derives subfactors and adjusts the number of questions according to the security requirements of the organization. This section focuses on analyzing both general and AIoT-based questionnaires in terms of the practical utility of the proposed method, while qualitative results are presented in Section 4.3. To specify the analysis results, this section is divided into two subsections: Section 4.4.1 examines 14 representative security rating questions using the proposed methodology, while Section 4.4.2 discusses its application to specific AIoT industry domains. The evaluation is based on prompts presented in Fig. 4 and is compared against state-of-the-art pre-trained models, such as GPT-4o and GPT-3.5 turbo.

4.4.1 Generation of the security rating questionnaire

The proposed security rating prompts incorporate constraints to ensure that the generated questions are clear, objective, and supported by measurable thresholds, thereby mitigating evaluator subjectivity and addressing vague criteria often observed in traditional security rating methods. Table 9 provides examples of questionnaires generated for the security rating task, aligning with the specific constraints illustrated in Fig. 2.

For instance, in the case of the seventh subfactor, i.e., “novelty,” the generated question explicitly defines the term and provides a clear comparison metric. Specifically, “novelty” is defined as “information that is new or original, and not a copy or adaptation of existing information, within the last year.” Furthermore, the evaluation metric uses percentage units (e.g., high: above 40%, medium: between 20% and 40%, and low: below 20%) instead of ambiguous or binary categorizations, making the questionnaire applicable across diverse organizational contexts.

Therefore, the proposed approach represents a marked improvement over traditional methods, which often include vague prompts such as “Evaluate novelty” or simple categorical options, such as “high,” “medium,” and “low,” without supporting quantitative criteria. The proposed methodology provides evaluators with clear standards and facilitates objective assessment grounded in measurable evidence. Thus, the generated questions can significantly enhance practical utility by providing both clarity and actionable evaluation criteria.

4.4.2 Generation of security rating questionnaire for the AIoT industry

Table 10 summarizes the results of utilizing LLMs fine-tuned for security rating questions in the AIoT industry based on the proposed methodology. These results are benchmarked against high-performing pre-trained models to evaluate the utility and effectiveness of AIoT-specific security rating questions. The analysis focuses on three key AIoT industry domains: real-time manufacturing, AIoT-based healthcare, and smart transportation. GPT-3.5 turbo serves as the baseline, while outputs from GPT-4o and the proposed approach (Llama 3.1 with DPO alignment) are compared. Complete questionnaires for each domain, generated using the proposed methodology, are provided in Appendix B (Figs. B1, B2, and B3). Representative examples are analyzed in this section.

In the real-time manufacturing domain, the evaluation of the subfactor “competitiveness” highlighted key limitations of the existing models. Outputs from GPT-3.5 turbo lacked measurable constraints, failed to provide clear criteria for evaluating subfactors, and included verbose classifications (e.g., high, medium, and low), introducing ambiguity and potential delays in evaluation. In comparison, GPT-4o exhibited improved clarity and detail, providing well-articulated descriptions for subfactors such as encryption and access control. However, it occasionally omitted crucial thresholds. In contrast, the proposed model effectively resolved these issues by generating specific and actionable descriptions, such as “Not accessible to external competitors, such as cloud services with strict access controls or on-premises servers with firewalls.” It also introduced percentage-based thresholds for categorizing risks (High: >40%, Medium: 20%–40%, and Low: <20%), thereby enhancing both clarity and precision. This approach

ensures that evaluations are objective and directly applicable to the manufacturing domain.

Table 10. Comparison of question generation results based on state-of-the-art LLM models, such as GPT-3.5 turbo and GPT-4o, for AIoT applications

AIoT industry	Method	Example of generated question
Real-time manufacturing	GPT-3.5 turbo (Baseline)	How would you rate the potential impact of external leakage risks on the competitiveness of our real-time manufacturing system? - High: The risk of external leakage significantly threatens our competitive position, potentially leading to substantial losses or significant competitive disadvantages. - Medium: The risk of external leakage poses a moderate threat to our competitiveness, (omitted). - Low: The risk of external leakage has minimal impact on our competitiveness, (omitted).
	GPT-4o	What percentage of your real-time manufacturing data, critical to maintaining competitive advantage, employs encryption and access controls to prevent unauthorized access and external leakage?
	Our methods (Llama 3.1 + DPO)	What percentage of your manufacturing data is stored in a secure environment that is not accessible to external competitors, such as cloud services with strict access controls or on-premises servers with firewalls? Less than 20% or more than 80% would be considered High risk, between 20% and 80% would be considered medium risk, and less than or equal to 20% would be considered Low risk.
Smart healthcare	GPT-3.5 turbo (Baseline)	How would you evaluate the innovative capabilities of your AIoT solutions in creating new value within your healthcare system? - High: Demonstrates clear evidence of consistently introducing novel features or services that significantly enhance patient care and operational efficiency, leading to distinct competitive advantages. (omitted)
	GPT-4o	What percentage of processes and decision-making tasks in your AIoT-based healthcare system are enhanced or automated through AI technologies, contributing directly to increased operational efficiency and value creation? (High: 70% or more, Medium: 40%–69%, and Low: below 40%)
	Our methods (Llama 3.1 + DPO)	What percentage of patients diagnosed with diseases by the AIoT system in your healthcare organization have shown improved health outcomes within 6 months compared to those diagnosed using traditional diagnosis methods? High: > 25%, Medium: 12.5%–25%, and Low: < 12.5%.
Smart transportation	GPT-3.5 turbo (Baseline)	To what extent does the level of information usability in your AIoT system for smart transportation industry enable users to efficiently access and interpret data for making informed decisions? - High: Usability features are intuitive, are user-friendly, and enable quick and easy access to relevant information, resulting in a significant increase in operational efficiency by at least 30%. (Omitted).
	GPT-4o	What percentage of the data processed by your Industrial AIoT-enabled intelligent transportation system is consistently transformed into actionable insights that improve the decision-making process for transportation operations and customer services?
	Our methods (Llama 3.1 + DPO)	What percentage of your system's data is easily accessible and understandable by all relevant stakeholders, such as maintenance personnel, drivers, and passengers, within 1 minute of requesting it? High: >80%, Medium: 20%–80%, and Low: <20%.

In the AIoT-based healthcare domain, the evaluation of the subfactor "value creation potential" revealed further disparities among models. The output from GPT-3.5 turbo often lacked relevance to the subfactor and omitted measurable thresholds, increasing subjectivity. GPT-4o improved target specificity but exhibited limited comprehension of healthcare-specific contexts. For example, generic outputs, such as "AI technologies contributing directly to increased operational efficiency and value creation," could be broadly applied to any AIoT domain, reducing domain specificity. In contrast, the proposed model produced contextually relevant questions tailored to healthcare. It identified distinct evaluation targets, such as "Patients diagnosed with diseases," and outlined explicit criteria for assessing improvements in diagnosis rates achieved by AIoT systems compared to traditional methods within a defined timeframe. This demonstrates the capacity of the model to incorporate domain-specific nuances, ensuring higher objectivity and clarity in evaluations.

For the subfactor "usability" in the smart transportation domain, GPT-3.5 turbo incorporated measurable thresholds but failed to tailor evaluation targets to the domain, resulting in overly generic outputs. Similarly, GPT-4o lacked percentage-based criteria and clearly defined evaluation targets, weakening its performance in this context. However, the proposed model effectively addressed these challenges by considering industry-specific nuances. It generated evaluation questions targeting relevant stakeholders, such as "Maintenance personnel, drivers, and passengers," and introduced constraints by the statement "1-minute request response time." Additionally, percentage-based thresholds were applied, enabling precise and actionable scoring. This comprehensive approach significantly enhanced the utility of the generated questions for smart transportation.

Overall, the comparative analysis revealed that GPT-3.5 turbo and GPT-4o often produced verbose, generic, or incomplete outputs; however, the proposed model overcame these limitations. By leveraging fine-tuning and explicitly defined constraints, the model could generate questions that were domain-specific, objective, and practically applicable. Even with minimal background prompts referencing the definition of each industry in the survey literature [1], the proposed methodology could produce clear and actionable evaluation criteria, thereby enhancing the clarity, objectivity, and real-world applicability of security rating questionnaires in AIoT industry contexts.

To further validate the applicability of AIoT across specific cases, we explored a specific case within the AIoT industry, focusing on its practical implementation. In particular, we selected display and semiconductor industries as representative examples of real-world manufacturing systems. Instead of employing a general manufacturing background, we conducted the analysis by incorporating industry-specific contexts related to display and semiconductor manufacturing.

When generating security rating questions regarding the "manpower," impact factor, industry-specific details were reflected in the formulation of the questions. For instance, in the display industry, the question explicitly mentions key display technologies: "What percentage of your workforce is directly involved in the development of display technologies such as LCD, OLED, micro-LED, and others? High: above 20%, Medium: between 10% and 20%, and Low: below 10%." This demonstrates the focus on display technology-related workforce allocation. Similarly, in the semiconductor industry, the question targets labor costs in semiconductor production: "What percentage of your company's semiconductor production information creation process involves manual labor that costs more than 150% of the industry average? High/Medium/Low. Above 60%, High; Between 11% and 60%, Medium; and Below 11%, Low." This question emphasizes the specific economic considerations within semiconductor manufacturing. These findings suggest that industry-specific backgrounds significantly influence the contextual framing of security rating assessments, thereby enhancing the relevance of AIoT applications in distinct manufacturing sectors.

A quantitative evaluation was conducted using Security G-Eval, the evaluation metric proposed in this study, to compare the proposed model with GPT-4o, the latest state-of-the-art model. As shown in Table 11, the proposed model exhibited superior performance in terms of coherence, consistency, fluency, and relevance. This result suggests that the inference capability of the Llama 3.1 8B Instruct model, which was used to generate high-quality security rating questionnaires, exhibits high feasibility for application

across various AIoT domains.

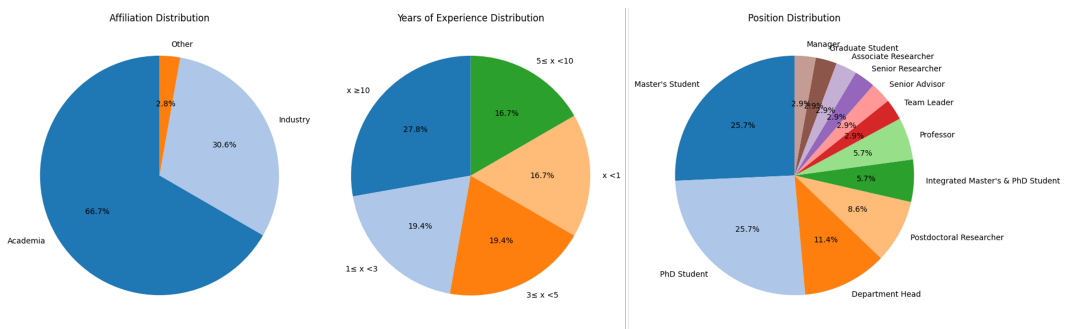
Table 11. Comparison of state-of-the-art LLM models for AIoT applications based on Security G-Eval

AIoT industry	Model	Coherence	Consistency	Fluency	Relevance
Manufacturing	Our models	5	5	4	5
	GPT-4o	4.8571	4.8928	3.5714	3.8571
Healthcare	Our models	5	5	4	5
	GPT-4o	4.3928	4.2857	3.8571	4.2142
Transportation	Our models	5	5	4	5
	GPT-4o	4.2857	4.0357	3.8571	4

4.4.3 Human evaluation to assess qualitative improvement in LLM-generated security rating questions

A human evaluation was conducted to assess the qualitative improvement in LLM-generated security rating questionnaires compared to traditional ones. This survey targeted security professionals, aiming to verify the effectiveness of AI-generated questions and evaluate their practical applicability across diverse contexts. The key aspects of this evaluation are summarized below:

- **Survey participants and experience levels:** A total of 36 respondents participated in the evaluation, representing both academia and industry professionals (e.g., those from Hyundai Motors and LG Display). As shown in Fig. 6, a detailed demographic breakdown indicated that the evaluation included respondents from various backgrounds: 66.7% from academia, 30.6% from the industry, and 2.8% from others. The participants had diverse experience levels within the security domain: 28% had more than 10 years of experience, 19% had 3 to 5 years of experience and 19% had 1 to 3 years of experience. This distribution ensures a balanced representation of expertise, incorporating both early-career professionals and seasoned experts, thereby enhancing the validity of the evaluation results.



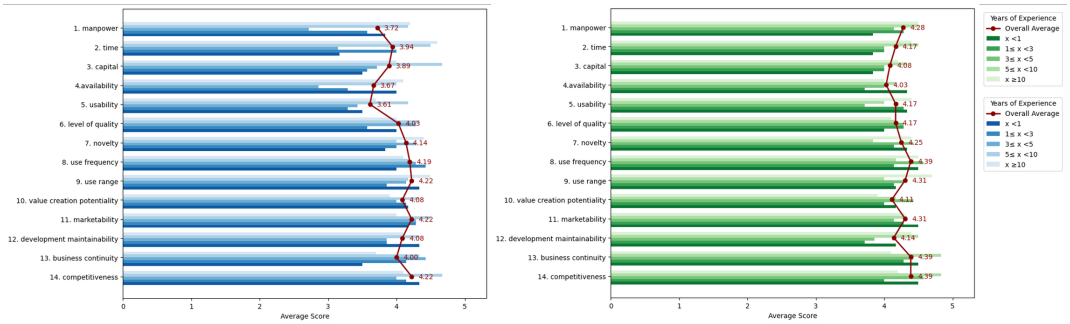
(a) (b) (c)

Fig. 6. The distribution of respondents based on experience level and job position: (a) affiliation distribution, (b) years of experience distribution, and (c) position distribution.

- **Evaluation metrics:** To evaluate the improvement in questions generated by LLMs, this survey employed two evaluation metrics: correlation and answerability. Correlation measures the extent to which LLM-generated questions align with the evaluation criteria, while answerability assesses the ease with which respondents provide respondents to provide answers to these questions. For example, to measure correlation and answerability, the respondents were asked two specific questions: “How well do LLM-generated security rating questions align with the given impact factors and subfactors compared to traditional questions?” (correlation) and “How easy are LLM-generated questions to answer compared to traditional questions?” (answerability). The survey covered 14 LLM-generated questions, as shown in Table 7. The responses were collected using a 5-point Likert scale. A higher

score indicates that LLM-generated questions exhibit significantly greater relevance (correlation) and ease of answering (answerability), while a lower score indicates that traditional questions remain more effective in these aspects. Unlike traditional evaluations that emphasize the correlation between the prompts and generated questions, as well as the grammar, fluency, and understandability of the generated questions [25, 56], this study focused on content relevance and ease of answering to better adapt the methodology for non-native English speakers.

- **Key findings:** The overall average score for correlation and answerability was 4.105. For correlation, the LLM-generated questions scored 3.99 on average, demonstrating strong relevance to the evaluation criteria. Among the evaluation items, "use range" had the highest average score (4.22), while "usability" had the lowest average score (3.572). These results suggest that questions regarding the breadth of information use are highly relevant; however, questions regarding the usability of the generated information require further refinement. For answerability, the questions received a higher overall average score of 4.22, indicating that the respondents found the LLM-generated questions easier to answer than traditional ones. The highest average score was 4.4, indicating that questions focusing on how frequently information, is used were particularly clear and straightforward. In contrast, "availability" had the lowest average score (4.03), suggesting that questions regarding the timeliness of information availability could benefit from further simplification or clarification. Reliability analysis using Cronbach's alpha yielded a high value of 0.878, as calculated by SPSS, confirming strong internal consistency across the survey. ANOVA revealed no statistically significant differences in scores based on affiliation groups (p=0.423) or experience levels (p=0.704). Traditional questionnaires have often been criticized for yielding inconsistent results influenced by the evaluators' experience levels [43]. However, our proposed methodology effectively mitigates this limitation, as the evaluation results remained consistent regardless of the participants' professional backgrounds or years of experience. As shown in Fig. 7, correlation and answerability results exhibited minimal variance across different expertise levels. Notably, respondents with more than 10 years of experience consistently rated the questions more than 4 out of 5, reinforcing the reliability and validity of the generated security rating questions. This consistent performance across various levels of expertise underscores the robustness of the proposed methodology and its capability to generate high-quality, objective, and domain-specific security rating questions.



(a) (b)

Fig. 7. Correlation and answerability results based on expertise levels: (a) correlation score and (b) answerability.

These results suggest that our proposed LLM-based security rating questionnaire demonstrates better consistency and answerability than the traditional questionnaire, making it suitable for widespread adoption without requiring extensive security expertise. Furthermore, the proposed methodology exhibits potential for multilingual adaptation, enhancing its versatility across diverse applications within the security domain.

5. Discussion and Ethical Considerations

This study focuses on leveraging LLMs to generate security rating questionnaires, aiming to overcome the limitations of traditional rating models and explore their applicability to the AIoT industry. However, as an initial study demonstrating the feasibility of AIoT security rating questionnaire generation, this research has certain limitations, particularly in terms of incorporating expert opinions from various industries and applying the methodology to diverse real-world AIoT scenarios. For instance, in the manufacturing sector, key security requirements and manufacturing processes vary across industries, such as display and semiconductor manufacturing. However, the proposed approach primarily relies on the provided background information, which may limit the generation of highly specialized rating questions requiring expert knowledge. To address this limitation, further advancements in the model are necessary to refine question generation for domain-specific applications. Future research will focus on refining the methodology by providing concrete scenarios and conducting focus group interviews with industry experts. Additionally, both quantitative and qualitative analyses will be performed to assess the practical applicability of the proposed framework in real-world settings. These efforts are expected to enhance the generation of security rating questionnaires tailored to specific subfields, incorporating domain expertise.

This study has additional limitations, including the relatively high computational resources required for question generation and the lack of detailed explanations for numerical thresholds, such as percentage constraints. Future research will address these limitations by improving model efficiency, integrating explainable AI techniques (e.g., attention-based explanation and commonsense reasoning [56]), and justifying numerical thresholds. These enhancements will further strengthen the applicability and scalability of the proposed framework.

Nevertheless, as the use of LLMs introduces potential risks, such as inherent biases and data privacy concerns, this study has implemented several measures to mitigate these ethical issues. To prevent security threats, all organization-specific details, such as affiliations and institutional identifiers, have been excluded from the AIoT prompts, ensuring the utilization of only generalized background information. Moreover, the output does not contain any personally identifiable information, such as names or institutional details, that could be linked to specific individuals.

This study also employs a human evaluation and considers only organizational roles, job positions, and experience levels when analyzing demographic distributions, while excluding personal data, such as names, emails, and phone numbers. These measures ensure compliance with ethical standards and minimize potential ethical concerns associated with biases and privacy risks.

6. Conclusion

Security rating serves as a cost-effective investment strategy for safeguarding and optimizing the management of confidential and high-value data, especially in the AIoT industry. However, prior research has often been fragmented, separating managerial and technical approaches, and relying on subjective and ambiguous evaluation methods.

To address these challenges, this study proposes an integrated framework that combines the strengths of managerial and technical approaches through 14 impact factors in the AIoT domain. The importance of tailored security metrics is highlighted to assess the consistency, coherence, fluency, and relevance between subcategories and the generated questions. Finally, this study proposes the fine-tuned Llama 3.1 8B Instruct model with DPO alignment for generating 14 evaluation questions.

Our proposed framework exhibited strong qualitative and quantitative performance. Specifically, the generated questions ensured objectivity by utilizing a percentage-based questionnaire approach and achieved superior outcomes across 13 metrics, including G-Eval, Security G-Eval, and a reward model, with a human evaluation score of 4.0 out of 5. Furthermore, our approach consistently outperformed

state-of-the-art models, such as GPT-4o and GPT-3.5 turbo, when adapted to AIoT-specific applications. Thus, our research provides a foundational guideline for automatically generating security rating questionnaires, which can be adapted not only to AIoT but also to other industry domains.

Nevertheless, this research has limitations, such as validating applicability in real-world AIoT scenarios, incorporating industry expert opinions, and enhancing the explainability of numerical threshold outcomes. These issues will be addressed in future research by conducting expert interviews, validating real-world scenarios, and integrating attention-based explanations and commonsense reasoning to support the interpretability and reliability of the generated security rating questionnaire.

Author's Contributions

Conceptualization, HC, YH, SS; Investigation, YH, YC, HL; Methodology, YH, DKG; Writing—overall original draft, YH; Writing—original draft of part 4.4.3, YC; Survey, YH, YC; Writing of the review and editing, YH; Formal analysis, YH; Data curation, YH.

Funding

This work was supported by the MSIT (Ministry of Science and ICT), Korea, under the Convergence security core talent training business support program (IITP-2025-RS-2023-00266605) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

Competing Interests

The authors declare that they have no competing interests.

Appendix A. Security G-Eval prompt

Evaluation Criteria:

Coherence (1-5): The logical flow and clarity of the question. The question should clearly build upon the subcategory, be well-structured, and enable evaluators to make objective, measurable decisions based on the criteria provided (High, Medium, Low, or Unknown). The question should not just be clear but also directly tie into the subcategory's impact factors, reflecting how the factors influence the evaluation process.

Evaluation Steps:

1. Review the prompt that defines the subcategory and key impact factors (e.g., manpower).
2. Read the generated question carefully. Check if it logically connects with the subcategory and its impact factors, ensuring the question is clear, well-structured, and facilitates consistent, measurable decisions.
3. Assign a score for coherence on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the evaluation criteria.

Fig. A1. Security G-Eval prompt for coherence evaluation metric.

Evaluation Criteria:

Consistency (1-5): The degree to which the generated question strictly adheres to the subcategory and impact factor guidelines outlined in the prompt. A consistent question should reflect the key impact factors or subcategory (e.g., manpower, time, capital) and use appropriate measurable criteria (e.g., percentages or proportional ranges). It should focus solely on the relevant subcategory, avoiding any unrelated elements.

Evaluation Steps:

1. Read the prompt and identify the key subcategory and impact factors specified.
2. Compare the generated question to the prompt, checking if it aligns with the subcategory and uses specific, measurable criteria.
3. Assign a score for consistency based on how closely the generated question reflects the prompt's requirements and whether it avoids unrelated or ambiguous factors.

Fig. A2. Security G-Eval prompt for consistency evaluation metric.

Evaluation Criteria:

Fluency (1-4): The quality of the generated question in terms of grammar, spelling, punctuation, word choice, and sentence structure, as well as the clarity and naturalness of the language when addressing the subcategory and its impact factors.

1. Poor – The question contains significant grammatical, spelling, or sentence structure errors, making it difficult to understand. It may fail to reflect the subcategory in a clear or relevant way, leading to confusion.

2. Fair – The question has some errors that affect clarity or smoothness, but it is still mostly understandable. The language may not fully align with the subcategory and impact factors, or it lacks a natural flow but is largely error-free.

3. Good – The question has few or no errors and reads clearly. However, while the language is accurate and understandable, it may not fully achieve the naturalness or clear connection with the subcategory.

4. Excellent – The question is free from errors, reads naturally, and clearly reflects the subcategory and its impact factors. The language is both smooth and appropriate, facilitating easy comprehension while being highly relevant to the evaluation.

Evaluation Steps:

1. Review the generated question for grammar, spelling, punctuation, and sentence structure issues.
2. Assess the word choice, clarity, and smoothness of the question. Ensure that the language used is both clear and natural in reflecting the subcategory and impact factors.
3. Assign a fluency score based on the 1 to 4 scale, where 1 is the lowest and 4 is the highest.

Fig. A3. Security G-Eval prompt for fluency evaluation metric.

Evaluation Criteria:

Relevance (1-5): The extent to which the generated question addresses the specific subcategory and impact factors outlined in the prompt. The question should focus solely on important and relevant aspects of the subcategory, avoiding unnecessary or redundant content. It should directly relate to how the subcategory influences the overall evaluation.

1. Poor. The question is largely irrelevant, does not align with the subcategory, or contains excessive unrelated information.
2. Weak. The question is somewhat relevant but introduces irrelevant elements or lacks focus.
3. Fair. The question is mostly relevant but may include minor irrelevant information or redundancies.
4. Good. The question is relevant to the subcategory and covers important aspects with little to no irrelevant information.
5. Excellent. The question is highly relevant, focusing solely on key aspects of the subcategory without any irrelevant or redundant information.

Evaluation Steps:

1. Read the prompt to identify the specific subcategory and key impact factors.
2. Review the generated question and assess whether it directly addresses the subcategory and 3. impact factors without introducing irrelevant or redundant information.
3. Assign a relevance score based on how well the question aligns with the prompt.

Fig. A4. Security G-Eval prompt for relevance evaluation metric.

Appendix B. Example of generated questionnaire for AIoT applications

Number	Impact Factor	Generated Question
1	Manpower	• What percentage of your manufacturing personnel are directly involved in creating or collecting data for Industrial AIoT techniques, above 80%? (High if above 80%, Medium if between 50% to 80%, Low if below 50%.)
2	Time	• What percentage of your manufacturing data is collected within the first 30 minutes of machine operation? Less than 10%, High; 10% to 50%, Medium; more than 50%, Low.
3	Capital	• What percentage of your total manufacturing capital investment is allocated to purchasing and maintaining industrial AIoT sensors and equipment, above 30%? High/Medium/Low. The answer is High if above 50%, Medium if between 30% to 50%, and Low if below 30%.
4	Availability	• What percentage of your real-time manufacturing data is available to all relevant stakeholders within a 1-hour time frame after its creation? High if above 90%, Medium if between 50% to 90%, Low if below 50%.
5	Usability	• What percentage of your manufacturing data is easily accessible and understandable by all employees involved in production processes, above 70%? High if above 70%, Medium if between 40% to 70%, Low if below 40%.
6	Level of Quality	• What percentage of your manufacturing data is of high quality, meaning it is accurate, complete, and consistent, above 95%? High if above 95%, Medium if between 90% to 95%, Low if below 90%.
7	Novelty	• What percentage of the information generated from Industrial AIoT techniques in your real-time manufacturing process is new or not previously known to your organization, above 20% but not more than 50%?
8	Use frequency	• What percentage of the manufacturing data collected from industrial sensors is used more than 30% of the time within the next 3 months? High if above 70%, Medium if between 30% to 70%, Low if below 30%.
9	Range of use	• What percentage of the manufacturing data generated by Industrial AIoT techniques is used across more than 50% of the production processes in your organization? High if above 80%, Medium if between 50% to 80%, Low if below 50%.
10	Value creation potential	• What percentage of the products manufactured in the last year have shown an increase in value due to real-time monitoring and analysis of manufacturing data, with an increase being defined as a value above 10% of the product's original price? High: Above 80%, Medium: Between 20% to 80%, Low: Below 20%.
11	Marketability	• What percentage of your organization's products or services have seen a market share increase of more than 15% within the last year due to the utilization of Industrial AIoT techniques in real-time manufacturing, and this increase is attributed to better product quality, faster time-to-market, or improved customer experience? High is above 50%, Medium is between 10% to 50%, and Low is 10% or below.
12	Development maintainability	• What percentage of the development time spent on new products in the last year was reduced due to the reuse of code and knowledge gained from real-time manufacturing data analysis, with a reduction being defined as a time below 50% of the original development time? High: Above 70%, Medium: Between 30% to 70%, Low: Below 30%.
13	Business continuity	• What percentage of your manufacturing data is stored in a cloud-based system that is not accessible to your company's IT department after working hours, and is this percentage more than 20%? High if above 20%, Medium if between 0% and 20%, Low if below 0%.
14	Competitiveness	• What percentage of your manufacturing data is stored in a secure environment that is not accessible to external competitors, such as cloud services with strict access controls or on-premises servers with firewalls? Less than 20% or more than 80% would be considered High risk, between 20% to 80% would be considered medium risk, and less than or equal to 20% would be considered Low risk.

Fig. B1. Example of a generated questionnaire for the AIoT-based real-time manufacturing industry.

Number	Impact Factor	Generated Question
1	Manpower	• What percentage of your total workforce in the AIoT department is dedicated to the creation of information for your healthcare system? Less than 10%, please respond High. What percentage is between 10% to 20%, please respond Medium. Otherwise, please respond Low.
2	Time	• What percentage of time is spent on developing AIoT solutions in the healthcare industry dedicated to testing and validation, rather than development and deployment? High: above 40%, Medium: between 20% to 40%, Low: below 20%.
3	Capital	• What percentage of the total annual investment in AIoT in your healthcare organization is spent on hardware and software infrastructure? Less than 30%, High; 30% to 50%, Medium; above 50%, Low.
4	Availability	• What percentage of the medical data in your AIoT system is available to be accessed and utilized by the healthcare professionals within 1 minute of their request? High: 90%, Medium: 50%, Low: 10%.
5	Usability	• What percentage of patients can easily access their health records on the AIoT-based healthcare system within 5 minutes? High: above 90%, Medium: between 80% to 90%, Low: below 80%.
6	Level of Quality	• What percentage of patient health records in your AIoT system are updated within the last 30 days? High: above 90%, Medium: between 80% to 90%, Low: below 80%.
7	Novelty	• What percentage of the new medical insights and discoveries generated by your AIoT system in the healthcare industry are novel, i.e., not reported or implemented by other healthcare organizations in the past year? High: above 70%, Medium: between 30% to 70%, Low: below 30%
8	Use frequency	• What percentage of your organization's total healthcare data is used at least monthly? High: 80% or above, Medium: Between 20% to 80%, Low: Below 20%.
9	Range of use	• What percentage of your organization's healthcare-related data is utilized across more than half of your departments? High: above 80%, Medium: between 20% to 80%, Low: below 20%.
10	Value creation potential	• What percentage of patients diagnosed with diseases by the AIoT system in your healthcare organization have shown improved health outcomes within 6 months compared to traditional diagnosis methods? High: > 25%, Medium: 12.5% - 25%, Low: < 12.5%.
11	Marketability	• What percentage of your organization's AIoT-based healthcare products and services are sold to external customers within 6 months of product/service launch? High: above 70%, Medium: between 20% to 70%, Low: below 20%.
12	Development maintainability	• What percentage of your AIoT based healthcare system's codebase is updated or refactored within the last 3 months? High: Above 80%, Medium: Between 20% to 80%, Low: Below 20%.
13	Business continuity	• What percentage of your organization's critical healthcare data is available within 1 hour in case of a disaster? High: above 95%, Medium: between 80% to 95%, Low: Below 80%.
14	Competitiveness	• What percentage of your organization's sensitive data in AIoT based healthcare system is protected by robust and unique authentication methods such as multi-factor authentication to prevent unauthorized access from external parties? Less than 20%, High; 20% to 60%, Medium; More than 60%, Low.

Fig. B2. Example of a generated questionnaire for the AIoT-based healthcare industry.

Number	Impact Factor	Generated Question
1	Manpower	What percentage of your transportation-related AIoT projects require more than 50% of your team to work overtime to meet the project deadline? High if above 30%, Medium if between 15% and 30%, Low if below 15%.
2	Time	What percentage of your organization's time is spent collecting and processing data for autonomous vehicles maintenance and predictive analytics in the smart transportation industry, High > 30%, Medium = 15-30%, Low < 15%?
3	Capital	What percentage of your annual capital investment in smart transportation infrastructure is spent on AIoT-related projects? High > 40%, Medium = 21-40%, Low < 21%.
4	Availability	What percentage of the real-time traffic data in your smart transportation system is available for analysis within 5 minutes of its creation? High: Above 90%, Medium: Between 70% to 90%, Low: Below 70%.
5	Usability	What percentage of your system's data is easily accessible and understandable by all relevant stakeholders, such as maintenance personnel, drivers, and passengers, within 1 minute of requesting it? High: >80%, Medium: 20-80%, Low: <20%.
6	Level of Quality	What percentage of data in smart transportation industry is accurate and reliable? High: Above 95%, Medium: Between 80% to 95%, Low: Below 80%.
7	Novelty	What percentage of your newly created information in smart transportation is novel, rather than a repetition of existing knowledge, within the last year? High: Above 70%, Medium: Between 30% to 70%, Low: Below 30%.
8	Use frequency	What percentage of the data collected from Industrial AIoT devices in the smart transportation industry is used more than once a week? High > 80%, Medium = 50-80%, Low < 50%.
9	Range of use	What percentage of your organization's total transportation data is used in at least 75% of its applications? High: Above 90%, Medium: Between 50% to 90%, Low: Below 50%.
10	Value creation potential	What percentage of your smart transportation system's data-driven insights lead to cost savings or revenue increases within the last year? High >20%, Medium=5-20%, Low <5%.
11	Marketability	What percentage of your smart transportation system's new features are adopted by at least 70% of your target market within 6 months after launch? High: above 80%, Medium: between 50% to 80%, Low: below 50%.
12	Development maintainability	What percentage of your smart transportation system's codebase is updated within a month after a new feature is developed to ensure development maintainability? High: Above 80%, Medium: Between 20% to 80%, Low: Below 20%.
13	Business continuity	What percentage of your organization's critical data and systems are available and accessible within 2 hours in case of a disaster? High >95%, Medium 38-95%, Low <38%.
14	Competitiveness	What percentage of your company's data related to smart transportation systems is stored on cloud services that are not controlled by your organization? High if above 60%, Medium if between 20% to 60%, Low if below 20%.

Fig. B3. Example of a generated questionnaire for the AIoT-based transportation industry.

Appendix C. Example of explainable AI for justifying numerical threshold via commonsense reasoning question

Commonsense reasoning question:

"What could be a reasonable justification for categorizing the percentage of the following generated question by considering the original given generated question prompt?"

* Generated question: {}
* Generated question prompt: {}

Response:

1. Alignment with Industry Standards and Compliance
Healthcare organizations are mandated to comply with regulations like the Health Insurance Portability and Accountability Act (HIPAA), which requires the implementation of safeguards to protect electronic protected health information (ePHI). Conducting thorough risk assessments and implementing security measures, including MFA, are essential components of HIPAA compliance. The Office for Civil Rights emphasizes the importance of evaluating risks and vulnerabilities to ePHI and implementing appropriate security measures to mitigate identified risks.
* Reference: Guidance on Risk Analysis (HIPPA)

2. Current Adoption Rates of Multi-Factor Authentication
Adoption rates of MFA in the healthcare sector provide a benchmark for setting risk categories. Reports indicate that approximately 61% of consumers enable MFA for online healthcare portals and applications.
This statistic suggests that organizations with MFA implementation below this threshold may be at higher risk, justifying the "Less than 20%: High risk" category.
* Reference: Multi-Factor Authentication & Smishing (Office of Information Security Securing One HHS)

3. Risk of External Data Breaches
The healthcare industry has experienced a significant number of data breaches, often due to inadequate authentication measures. Implementing MFA significantly reduces the risk of unauthorized access, thereby decreasing the likelihood of data breaches. Organizations with higher percentages of sensitive data protected by MFA are better positioned to prevent unauthorized access, supporting the "More than 60%: Low risk" classification.

4. Impact on Business Continuity and Competitiveness
Protecting sensitive data is crucial for maintaining business continuity and competitiveness in the healthcare sector. Data breaches can lead to operational disruptions, financial losses, and reputational damage. Organizations that implement robust authentication methods, such as MFA, are better equipped to prevent such incidents, thereby maintaining their competitive edge. This consideration aligns with the original prompt's emphasis on assessing the risk of external leakage concerning competitiveness.

5. Proactive Risk Management Strategies
Regular data audits and risk assessments are proactive measures that help identify and address vulnerabilities in healthcare systems. By categorizing the extent of MFA implementation, organizations can prioritize their security efforts and allocate resources effectively to mitigate risks associated with unauthorized access. In summary, the proposed categorization reflects a comprehensive approach to assessing and mitigating risks associated with unauthorized access to sensitive data in AIoT-based healthcare systems. It emphasizes the importance of robust authentication measures in maintaining data security, regulatory compliance, and organizational competitiveness.
* Reference: Data Protection Strategies for Healthcare Providers (AMWINS)

Fig. C1. Example of an explainable AI result for the "competitiveness" impact factor in the smart healthcare industry.

References

[1] K. S. Awaisi, Q. Ye, and S. Sampalli, "A survey of industrial AIoT: opportunities, challenges, and directions," *IEEE Access*, vol. 13, pp. 96946-96996, 2024. https://doi.org/10.1109/ACCESS.2024.3426279

[2] A. Matin, M. R. Islam, X. Wang, H. Huo, and G. Xu, "AIoT for sustainable manufacturing: overview, challenges, and opportunities," *Internet of Things*, vol. 24, article no. 100901, 2023. https://doi.org/10.1016/j.iot.2023.100901

[3] S. R. Jermiah, J. Ha, S. K. Singh, and J. H. Park, "PrivacyGuard: collaborative edge-cloud computing architecture for attribute-preserving face anonymization in CCTV networks," *Human-centric Computing and Information Sciences*, vol. 14, article no. 43, 2024. https://doi.org/10.22967/HCIS.2024.14.043

[4] P. Arora and R. Makani, "Blockchain Integration with AIoT data security and privacy for sustainability," *Journal of Artificial Intelligence and Systems*, vol. 6, pp. 112-123, 2024. https://doi.org/10.33969/AIS.2024060108

[5] M. Hammad, R. M. Jillani, S. Ullah, A. Namoun, A. Tufail, K. H. Kim, and H. Shan, "Security framework for network-based manufacturing systems with personalized customization: an industry 4.0 approach," *Sensors*, vol. 23, no. 17, article no. 17, 2023. https://doi.org/10.3390/s23177555

[6] S. K. Singh, L. T. Yang, and J. H. Park, "FusionFedBlock: fusion of blockchain and federated learning to preserve privacy in Industry 5.0," *Information Fusion*, vol. 90, pp. 223-240, 2023. https://doi.org/10.1016/j.inffus.2022.09.027

[7] Y. Han and H. Chang, "XA-GANomaly: an explainable adaptive semi-supervised learning method for intrusion detection using GANomaly," *Computers, Materials and Continua*, vol. 76, no. 1, pp. 221-237, 2023. https://doi.org/10.32604/cmc.2023.039463

[8] Y. Han and H. Chang, "Insider threat surveillance system via multimodal-based physical leak detection and optimal imbalanced augmentation in research and development environments," *Human-Centric Computing and Information Sciences*, vol. 14, article no. 73, 2024. http://dx.doi.org/10.22967/HCIS.2024.14.073

[9] B. Herzberg, "Data classification: a critical piece to your data security puzzle," 2022 [Online]. Available: https://www.datanami.com/2022/12/05/data-classification-a-critical-piece-to-your-data-security-puzzle/.

[10] O. Na, L. W. Park, H. Yu, Y. Kim, and H. Chang, "The rating model of corporate information for economic security activities," *Security Journal*, vol. 32, pp. 435-456, 2019. https://doi.org/10.1057/s41284-019-00171-z

[11] Korean Intellectual Property Office, "Trade secret security rating guidelines," c2024 [Online] Available: https://www.tradesecret.or.kr/secret-resources/files/GUIDE.pdf.

[12] Y, Han, J. Lee, J. Lee, and H. Chang, "A review of text-based information security rating: fundamental concepts, methods, datasets, challenges, and future works," *Human-centric Computing and Information Sciences*, vol. 15, article no. 32, 2025. https://doi.org/10.22967/HCIS.2025.15.032

[13] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, et al., "The Llama 3 herd of model," 2024 [Online]. Available: https://arxiv.org/abs/2407.21783.

[14] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *Advances in Neural Information Processing Systems*, vol. 36, pp. 53728-53741, 2023.

[15] H. Xu, A. Sharaf, Y. Chen, W. Tan, L. Shen, B. V. Durme, K. Murray, and Y. J. Kim, "Contrastive preference optimization: pushing the boundaries of LLM performance in machine translation," 2024 [Online]. Available: https://arxiv.org/abs/2401.08417.

[16] J. Hong, N. Lee, and J. Thorne, "Reference-free monolithic preference optimization with odds ratio," 2024 [Online]. Available: https://arxiv.org/abs/2403.07691.

[17] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-eval: NLG evaluation using GPT-4 with better human alignment," 2023 [Online]. Available: https://arxiv.org/abs/2303.16634.

[18] S. K. Singh, S. Chauhan, A. Alsafrani, M. Islam, H. I. Sherazi, and I. Ullah, "Optimizing healthcare data quality with optimal features driven mutual entropy gain," *Expert Systems*, vol. 42, no. 2, article no. e13737, 2025. https://doi.org/10.1111/exsy.13737

[19] D. Hiza, "Assessing the significance of CIA triad security model in establishing ICT security controls in the public sector," Institute of Accountancy Arusha (IAA), Tanzania, 2022.

[20] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplet, D. D. L. Casas, et al., "Mistral 7B," 2023 [Online]. Available: https://arxiv.org/abs/2310.06825.

[21] G. Katz, Y. Elovici, and B. Shapira, "CoBAn: A context based model for data leakage prevention," *Information Sciences*, vol. 262, pp. 137-158, 2014. https://doi.org/10.1016/j.ins.2013.10.005

[22] E. Alparslan, A. Karahoca, and H. Bahsi, "Security-level classification for confidential documents by using adaptive neuro-fuzzy inference systems," *Expert Systems*, vol. 30, no. 3, pp. 233-242, 2013. https://doi.org/10.1111/j.1468-0394.2012.00634.x

[23] J. Jiang, Y. Lu, M. Yu, G. Li, C. Liu, S. An, and W. Huang, "CES2Vec: a confidentiality-oriented word embedding for confidential information detection," in *Proceedings of the 2020 IEEE Symposium on Computers and Communications (ISCC)*, Rennes, France, 2020, pp. 1-7. https://doi.org/10.1109/ISCC50000.2020.9219702

[24] Y. Lu, J. Jiang, M. Yu, C. Liu, C. Liu, W. Huang, and Z. Lv., "Depthwise separable convolutional neural network for confidential information analysis," in *Knowledge Science, Engineering and Management*. Cham, Switzerland: Springer, 2020, pp. 450-461. https://doi.org/10.1007/978-3-030-55393-7_40

[25] A. Ushio, F. Alva-Manchego, and J. Camacho-Collados, "Generative language models for paragraph-level question generation," 2022 [Online]. Available: https://arxiv.org/abs/2210.03992v1.

[26] N. Mulla and P. Gharpure, "Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications," *Progress in Artificial Intelligence*, vol. 12, no. 1, pp. 1-32, 2023. https://doi.org/10.1007/s13748-023-00295-9

[27] T. Fukuma, K. Noda, T. U. K. Hoso, Y. Ichikawa, K. Kambe, Y. Masubuch, and F. Toriumi, "QANA: LLM-based question generation and network analysis for zero-shot key point analysis and beyond," 2024 [Online]. Available: https://arxiv.org/abs/2404.18371.

[28] B. H. H. Cheung, G. K. K. Lau, G. T. C. Wong, E. Y. P. Lee, D. Kulkarni, C. S. Seow, R. Wong, and M. T. Co, "ChatGPT versus human in generating medical graduate exam multiple choice questions: a multinational prospective study (Hong Kong SAR, Singapore, Ireland, and the United Kingdom)," *PLOS One*, vol. 18, no. 8, article no. e0290691, 2023. https://doi.org/10.1371/journal.pone.0290691

[29] S. Bhandari, Y. Liu, Y. Kwak, and Z. A. Pardos, "Evaluating the psychometric properties of ChatGPT-generated questions," *Computers and Education: Artificial Intelligence*, vol. 7, article no. 100284, 2024. https://doi.org/10.1016/j.caeai.2024.100284

[30] N. A. Kumar and A. Lan, "Improving socratic question generation using data augmentation and preference optimization," 2024 [Online]. Available: https://arxiv.org/abs/2403.00199.

[31] O. E. L. Castro, X. Deng, and J. H. Park, "Comprehensive survey on AI-based technologies for enhancing IoT privacy and security: trends, challenges, and solutions," *Human-centric Computing and Information Sciences*, vol. 13, article no. 39, 2023. https://doi.org/10.22967/HCIS.2023.13.039

[32] A. A. Malibari, M. K. Nour, F. N. Al-Wesabi, R. Alabdan, A. Mohamed, M. Al Duhayyim, A. Alkhayyat, and D. Gupta, "Metaheuristics with deep learning enabled epileptic seizure classification for smart healthcare on cyborg robots," *Human-centric Computing and Information Sciences*, vol. 13, article no. 39, 2023. https://doi.org/10.22967/HCIS.2023.13.039

[33] M. Badawy, "Integrating artificial intelligence and big data into smart healthcare systems: a comprehensive review of current practices and future directions," *Artificial Intelligence Evolution*, vol. 4, no. 2, pp. 133-153, 2023.

[34] Y. D. Al-Otaibi, "CYBROG-based smart home security and power-saving control system using hybrid technique," *Human-centric Computing and Information Sciences*, vol. 13, article no. 52, 2023. https://doi.org/10.22967/HCIS.2023.13.052

[35] H. Qiao, X. Xu, and Y. Zang, "Security and privacy protection technology of AIoT intelligent transportation teaching system with 5G technology," in *Proceedings of the 3rd International Conference on Big Data, Information and Computer Network (BDICN)*, Sanya, China, 2024, pp. 115-158. https://doi.org/10.1109/BDICN62775.2024.00038

[36] M. M. Salim, X. Deng, and J. H. Park, "A privacy-preserving local differential privacy-based federated learning model to secure LLM from adversarial attacks," *Human-centric Computing and Information Sciences*, vol. 14, article no. 57, 2024. https://doi.org/10.22967/HCIS.2024.14.057

[37] J. Zhao, K. Chen, X. Yuan, Y. Qi, W. Zhang, and N. Yu, "Silent guardian: protecting text from malicious exploitation by large language models," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 8600-8615, 2024. https://doi.org/10.1109/TIFS.2024.3455775

[38] Z. Ji, T. Yu, Y. Xu, N. Lee, E. Ishii, and P. Fung, "Towards mitigating hallucination in large language models via self-reflection," 2023 [Online]. Available: https://arxiv.org/abs/2310.06271.

[39] M. A. Ferrag, F. Alwahedi, A. Battah, B. Cherif, A. Mechri, and N. Tihanyi, "Generative AI and large language models for cyber security: all insights you need," 2024 [Online]. Available: https://dx.doi.org/10.2139/ssrn.4853709.

[40] Y. Zhang, W. Song, Z. Ji, and N. Meng, "How well does LLM generate security tests?," 2023 [Online]. Available: https://arxiv.org/abs/2310.00710.

[41] A. Kavian, M. M. Pourhashem Kallehbasti, S. Kazemi, E. Firouzi, and M. Ghafari. "LLM security guard for code," in *Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering*, Selerno, Italy, 2024, pp. 600-603. https://doi.org/10.1145/3661167.3661263

[42] X. Pan, J. van Ossenbruggen, V. de Boer, and Z. Huang, "A RAG approach for generating competency questions in ontology engineering," 2024 [Online]. Available: https://arxiv.org/abs/2409.08820.

[43] G. Hong, Y. Han, and W. Yun, "A study on the classification of display technical documents using graph embedding," *Korean Journal of Industry Security*, vol. 13, no. 1, pp. 1-25, 2023.

[44] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in Neural Information Processing Systems*, vol. 35, pp. 22199-22213, 2023.

[45] U. Lee, H. Jung, Y. Jeon, Y. Sohn, W. Hwang, J. Moon, and H. Kim, "Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in English education," *Education and Information Technologies*, vol. 29, no. 9, pp. 11483-11515, 2024. https://doi.org/10.1007/s10639-023-12249-8

[46] P. A. Glasow, "Fundamentals of survey research methodology," 2005 [Online]. Available: https://www.mitre.org/news-insights/publication/fundamentals-survey-research-methodology-0.

[47] Z. Wang, Y. Dong, O. Delalleau, J. Zeng, G. Shen, D. Egert, J. J. Zhang, M. N. Sreedhar, and O. Kuchaiev, "HelpSteer2: open-source dataset for training top-performing reward models," 2024 [Online]. Available: https://arxiv.org/abs/2406.08673.

[48] Z. Meng, Z. Zhang, G. Li, and D. Zhang, "An active weight learning method for efficient reliability assessment with small failure probability," *Structural and Multidisciplinary Optimization*, vol. 61, pp. 1157-1170, 2020. https://doi.org/10.1007/s00158-019-02419-z

[49] Z. Wang, B. Bi, S. K. Pentyala, K. Ramnath, S. Chaudhuri, S. Mehrotra, et al., "A comprehensive survey of LLM alignment techniques: RLHF, RLAIF, PPO, DPO and more," 2024 [Online]. Available: https://arxiv.org/abs/2407.16216.

[50] B. Adler, N. Agarwal, A. Aithal, D. H. Anh, P. Bhattacharya, A. Brundyn, et al., "Nemotron-4 340B technical report," 2024 [Online]. Available: https://arxiv.org/abs/2406.11704.

[51] B. Li, Y. Jiang, V. Gadepally, and D. Tiwari, "LLM inference serving: survey of recent advances and opportunities," 2024 [Online]. Available: https://arxiv.org/abs/2407.12391.

[52] N. F. Rajani, B. McCann, C. Xiong, and R. Socher, "Explain yourself! Leveraging language models for commonsense reasoning," 2019 [Online]. Available: https://arxiv.org/abs/1906.02361.

[53] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach, et al., "hi-3 technical report: a highly capable language model locally on your phone," 2024 [Online]. Available: https://arxiv.org/abs/2404.14219.

[54] B. Hui, J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang, et al., "Qwen2.5-coder technical report," 2024 [Online]. Available: https://arxiv.org/abs/2409.12186.

[55] G. Biancini, A. Ferrato, and C. Limongelli, "Multiple-choice question generation using large language models: methodology and educator insights," in *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, Cagliari, Italy, 20254, pp. 584-590. https://doi.org/10.1145/3631700.3665233

[56] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Win, and M. Du, "Explainability for large language models: a survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 2, article no. 20, 2024. https://doi.org/10.1145/3639372