POLITECNICO DI TORINO
Repository ISTITUZIONALE

Assessing LLMs models' knowledge of automotive cyberthreats benchmarking autoISAC framework

*Original*
Assessing LLMs models' knowledge of automotive cyberthreats benchmarking autoISAC framework / Scarano, N., Mannella, L., Savino, A., Di Carlo, S.. - ELETTRONICO. - (2025), pp. 1-7. (CSCS '25: 2nd Cyber Security in CarS Workshop (CSC) Taipei (TWN) October 13-17, 2025) [10.1145/3736130.3762690].

*Availability:*
This version is available at: 11583/3004901 since: 2025-11-07T10:42:45Z

*Publisher:*
Association for Computing Machinery

*Published*
DOI:10.1145/3736130.3762690

*Terms of use:*

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

*Publisher copyright*

(Article begins on next page)

19 November 2026



# Assessing LLMs models' knowledge of automotive cyberthreats benchmarking autoISAC framework

Nicola Scarano
Politecnico di Torino
Torino, Italy
nicola.scarano@polito.it

Luca Mannella
Politecnico di Torino
Torino, Italy
luca.mannella@polito.it

Alessandro Savino
Politecnico di Torino
Torino, Italy
alessandro.savino@polito.it

Stefano Di Carlo
Politecnico di Torino
Torino, Italy
stefano.dicarlo@polito.it

## Abstract

Large Language Models (LLMs) are gaining traction in cybersecurity applications, offering both promising opportunities and potential new risks. The use of these models in sub-domains such as automotive is still in its early stages. In this work-in-progress study, we use GPT-4o from OpenAI to generate a preliminary set of domain-relevant cybersecurity questions exploiting the Automotive Information Sharing and Analysis Center (Auto-ISAC) framework, which we then refined through manual validation. We exploited the final set of 25 questions to evaluate the performance of five LLMs models. Then, these questions were administered through a survey to a group of 17 domain experts, allowing us to compare this baseline with the results from the LLMs. From our preliminary findings, we found that LLMs reached a mean of 91.2% of correct answers on the test while human experts' performance reached 64.7%. This study lays the groundwork for future investigations into the use of LLMs in the automotive-security domain and into the safe and trustworthy exploitation of LLMs.

## CCS Concepts

• **Security and privacy** → *Usability in security and privacy*.

## Keywords

Automotive, Cybersecurity, LLM assessment, auto-ISAC

**ACM Reference Format:**
Nicola Scarano, Luca Mannella, Alessandro Savino, and Stefano Di Carlo. 2025. Assessing LLMs models' knowledge of automotive cyberthreats benchmarking autoISAC framework. In *Proceedings of the 2025 Cyber Security in CarS Workshop (CSCS '25), October 13–17, 2025, Taipei, Taiwan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3736130.3762690

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
*CSCS '25, Taipei, Taiwan*
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1928-8/25/10
https://doi.org/10.1145/3736130.3762690

## 1 Introduction

The rapid transition toward interconnected smart vehicles, the growing interest in electric cars, and the widespread use of Advanced Driver Assistance Systems (ADAS) have increased the cyberthreats to which modern cars and the entire automotive sector are exposed [10, 14]. In response, cybersecurity research in automotive has generated a growing interest within the scientific community over the last years, developing in a variety of topics, from in-vehicle communication protocols like the Controller Area Network (CAN) to leveraging Open-Source Intelligence (OSINT) for threat detection and analysis [11, 15].

Large Language Models (LLMs) have gained interest from the cybersecurity community as a powerful instrument for several tasks, including automating code analysis, finding vulnerabilities, summarizing threat intelligence, and assisting with secure coding [18]. At the same time, LLMs open new research areas around their safe use and potential abuse.

Despite their common performances in general tasks, the effectiveness of LLMs in specialized domains is being studied as a fundamental step to ensure their effective deployment. In particular, the automotive sector, which is traditionally safety-oriented and highly regulated, presents unique challenges in using such technologies for cybersecurity applications. Trust and reliability are essential in these contexts; even minor factual errors or reasoning gaps can have serious implications.

Several key challenges emerge when applying LLMs to critical domains such as automotive cybersecurity. The first is the need to evaluate the extent and accuracy of their knowledge in the domain. The second, more methodological, concerns how this knowledge can be effectively assessed.

This paper aims to present the earlier results of our research where we evaluated the domain-specific knowledge of LLMs using the Automotive Information Sharing and Analysis Center (Auto-ISAC) cybersecurity framework [2]. Specifically, we developed a question-generation pipeline that leveraged the GPT-4o model by OpenAI, Auto-ISAC knowledge set, and manual verification. A set of validated questions was then collected and distributed to domain experts through a survey. Finally, we compared 5 LLMs responses to the survey with those provided by human experts, aiming to gain an initial insight into the model's understanding and knowledge of automotive cybersecurity.

Results evidence that the 5 selected state-of-the-art LLMs significantly outperform human domain experts in a developed knowledge assessment test consisting of 25 validated questions. Claude Sonnet 4 achieved perfect accuracy (100%), and the average accuracy across all evaluated LLMs was 91.2%, compared to a mean human expert performance of 64.7%.

The remainder of the paper is structured as follows. Section 2 reviews related work on cybersecurity knowledge-sharing platforms, the assessment of LLMs, and the challenges of applying these models in the safety-critical automotive domain. Section 3 outlines the methodology adopted in our study, the use of GPT-4o to generate questions based on Auto-ISAC and their validation. In Section 4, we present the results of the expert survey and the LLM performance on the same set of questions, comparing human and LLM performance. Section 5 discusses key findings, implications, and limitations. Finally, Section 6 summarizes our contributions and outlines directions for future work.

## 2 Background and Related Work

In this section, we review existing efforts in cybersecurity knowledge sharing, the application of LLMs to OSINT, and the unique characteristics of automotive cybersecurity as a safety-critical domain. This background contextualizes our work and highlights the current gaps in assessing LLMs effectiveness in structured, high-stakes environments.

### 2.1 Cybersecurity knowledge sharing platforms

Cybersecurity knowledge sharing platforms enable the collection and dissemination of information about cyber threats. Established platforms include MITRE Adversarial Tactics, Techniques, and Common Knowledge (ATT&CK) [13] and AlienVault OTX [1], both supporting collaborative threat identification and response. Sector-specific initiatives have also emerged to address domain-dependent threats and constraints, as seen in Information Sharing and Analysis Centers (ISACs) for healthcare [8], energy [4], and aviation [3]. The increasing sharing of threat information (e.g., research articles, threat incident reports, and social media data) through open-source platforms has represented in recent years a great opportunity for exploitation of Natural Language Processing (NLP) techniques for cyber threats knowledge extraction and analysis [5].

### 2.2 LLMs knowledge assessment

LLMs shown great potential in analysing a large amount of unstructured information, outperforming traditional Named Entity Recognition (NER) methods. The information available in threat-sharing communities presents a unique opportunity to exploit LLMs for proactive cybersecurity defense, like risk mitigation, attack vector information gathering, and more [6, 15]. Nevertheless, it is well known that LLMs are prone to hallucinations, generating plausible but incorrect information, which makes their outputs unreliable in high-stakes settings. In safety-critical applications, ensuring accountability, transparency, and trustworthiness is essential.

To meet the challenges of critical safety systems, the evaluation of structured models through domain-specific question answering has emerged as a valuable technique to better understand the limitations of domain-specific knowledge of LLMs. Notably, Garza et al. [7] evaluated GPT-3.5 using questions derived from the MITRE ATT&CK framework, analyzing its performance in generating and answering questions about adversary behavior and mitigations. Their findings showed that LLMs can match or even outperform non-expert human participants in certain cybersecurity domains. Similarly, Tihanyi et. al [17] introduced a broader benchmark to assess LLMs understanding across multiple cybersecurity tasks, including threat identification and categorization, providing an early glimpse into how these models generalize across structured taxonomies. Other works, such as [16], explored LLMs applications in OSINT gathering and situational awareness, highlighting both their potential and their limitations in handling ambiguous or context-sensitive threat information.

Domain-specific areas, such as automotive cybersecurity, remain largely unexplored. This creates a clear opportunity to assess how LLMs perform in safety-critical, regulation-driven environments using structured frameworks like Auto-ISAC.

### 2.3 Automotive Cybersecurity as a Safety-Critical Domain

Automotive cybersecurity demands high traceability, explainability, and compliance with standards such as ISO/SAE 21434 [9]. The Auto-ISAC framework represents an industry-driven community to share and analyze intelligence about emerging cybersecurity risks to the vehicle, and to collectively enhance vehicle cybersecurity capabilities across the global automotive industry, including light- and heavy-duty vehicle OEMs, suppliers, and the commercial vehicle sector [2]. While this framework offers structured guidance tailored to the vehicle context, there is currently no established benchmark for evaluating how well LLMs understand or support this specific data source.

## 3 Methods

This section describes the methodology adopted to generate a set of automotive cybersecurity-related questions. We first introduce the Auto-ISAC framework. Then, we detail the pipeline used to transform threat data into valid questions, exploiting GPT-4o for the generation. The pipeline includes data preprocessing,*chunk*ing strategies, prompt design, and the categorization of question types, followed by a thorough manual validation process.

### 3.1 Auto-ISAC cyberthreat matrix

Auto-ISAC presents a collection of automotive cybersecurity attacks. The automotive threat matrix, structured by Auto-ISAC, provides a broad and complete categorization of attacks in vehicle systems organized in tactics, techniques, and procedures.

Within the MITRE ATT&CK framework definition, three entities are particularly relevant: *tactics*, *techniques*, and *procedures*. *Tactics* represent the high-level objectives of an attacker; *techniques* describe the specific methods used to achieve those objectives; and *procedures* are the concrete implementations or steps taken to execute an attack. In the context of this study, we worked with *procedures* cause they represent a categorization of attack with a granularity not too wide (like topics) and not too specific (like techniques), allowing us to be more flexible in the creation of the questions. Specifically, in the Auto-ISAC database, each procedure is

structured as a JSON file containing the most relevant information related to the procedure itself and the set of techniques exploited in the specific procedure's implementation. Figure 1 shows an example based on a real-world security assessment reported by Tencent Keen Security Lab [12] and available in the Auto-ISAC threat matrix. The structure has been adapted to highlight its mapping to MITRE ATT&CK techniques and to fit in the frame. Next to the single technique, the MITRE attack ID is shown.

**Title:** Tencent Keen Security Lab: Experimental Security Assessment on Lexus Cars
**Description:** Researchers were able to remotely spawn a root shell in a component by leveraging remote code execution to automatically connect the component to a Wi-Fi hotspot.
**MITRE ID:** P0022
**Auto-ISAC ID:** ATM-P0022
**Associated Techniques:**

- **Short Range Wireless Communication (T0065)**: If available, the adversary can use the vehicle's short-range wireless capabilities (e.g., Bluetooth, Wi-Fi) post-exploitation to establish command and control or exfiltrate data.
- **Command and Scripting Interpreter (T0018)**: Adversaries may abuse scripting environments (e.g., Python, Bash) to execute arbitrary commands. See MITRE ATT&CK entry: T1059.

**Figure 1: An example of an Auto-ISAC Procedure from Tencent Keen Security Lab related to Experimental Security Assessment on Lexus Cars.**

## 3.2 Questions set generation pipeline

To generate the questions, we explored the GPT-4o model from OpenAI. As a first step, we defined a set of requirements to guide the question generation process:

- Be grounded in specific procedures and their associated techniques.
- Avoid general or introductory cybersecurity topics (e.g., definitions or abstract principles).
- Focus not on asking what a specific technique does, but on evaluating the feasibility or impact of a given attack scenario within a realistic automotive context.

*3.2.1 Chunking strategy.* In the Auto-ISAC database, there are 24 procedures. Each procedure can appear multiple times in the dataset, each time with a slightly different description and combination of techniques, usually meaning that the report or research papers are the same, but different exploits have been tested. Table 1 shows all the considered procedures and their frequency in the dataset.

To comply with the length limits of the GPT-4o prompt and to let the model during question generation exploit knowledge available in procedures with related topics, we created fixed-size sets of injectable information, *chunks*, with 16 procedures each, that can be completely fed into the GPT-4o prompt. Thus, procedures were clustered into *chunks* of semantically related categories (e.g., CAN related procedures, wireless communication, low-level software modification, and more). At the end, we grouped all the procedures in 16 *chunks*. Each of these *chunks* was then used separately as a source of knowledge from which the LLM had to generate questions (as shown in Figure 2).

**Table 1: List of Auto-ISAC procedures considered in this study.**

| ID | Procedure Title | Count |
|---|---|---|
| ATM-P0024 | Adventures in Automotive Networks and Control Units | 27 |
| ATM-P0006 | Experimental Security Assessment of BMW Cars: A Summary Report | 19 |
| ATM-P0031 | CAN Message Injection | 17 |
| ATM-P0001 | Free-fall: Hacking Tesla from wireless to CAN bus | 16 |
| ATM-P0083 | Comprehensive Experimental Analyses of Automotive Attack Surfaces | 16 |
| ATM-P0017 | Over-the-air: how we remotely compromised the gateway, ecm, and autopilot ecus of Tesla cars | 15 |
| ATM-P0076 | Experimental Security Analysis of a Modern Automobile | 14 |
| ATM-P0020 | Hacking a Tesla Model S: What we found and what we learned | 13 |
| ATM-P0073 | Remote Exploitation of an Unaltered Passenger Vehicle | 10 |
| ATM-P0022 | Tencent Keen Security Lab: Experimental Security Assessment on Lexus Cars | 7 |
| ATM-P0071 | IoT backdoors in cars | 6 |
| ATM-P0013 | Drift with Devil: Security of Multi-Sensor Fusion based Localization in High-Level Autonomous Driving under GPS Spoofing | 6 |
| ATM-P0023 | Security and Privacy Vulnerabilities of In-Car Wireless Networks: A Tire Pressure Monitoring System Case Study | 5 |
| ATM-P0070 | Exploiting Wi-Fi Stack on Tesla Model S | 4 |
| ATM-P0038 | Drive it like you hacked it | 3 |
| ATM-P0080 | Evaluating Physical-Layer BLE Location Tracking Attacks on Mobile Devices | 2 |
| ATM-P0088 | There Will Be Glitches Extracting and Analyzing Automotive Firmware Efficiently | 2 |
| ATM-P0016 | Losing the Car Keys: Wireless PHY-Layer Insecurity in EV Charging | 1 |
| ATM-P0082 | NFC Relay Attack on Tesla Model Y | 1 |
| ATM-P0175 | Driving Down the Rabbit Hole | 1 |
| ATM-P0193 | Unlocking the Drive: Exploiting Tesla Model 3 | 1 |
| ATM-P0194 | Experimental Security Analysis of a Modern Automobile | 1 |
| ATM-P0196 | Extracting SecOC secrets from an ECU | 1 |
| ATM-P0197 | New Example: Jailbreaking an Electric Car in 2022-2023 or What It Means to Hotwire Tesla's x86-Based Seat Heater | 1 |

*3.2.2 Question types.* We exploited prompting on the GPT-4o model, providing specific procedures (each time a different *chunk*) and clearly defining its question generation task. In particular, we tested the generation of two types of question formats: *multiple-choice* and *true/false*. We designed a distinct prompt for each format while maintaining the same chunking and thematic grouping strategy described above. Figures 3 and 4 represent two LLM outputs generated with our framework, one per question type. Specifically, we asked the model to create questions and options and, for validation purposes, explain each answer and the sources used to craft that specific question.

*3.2.3 Validation.* We manually reviewed each item at the end of the question generation process to identify and correct grammatical, logical, or contextual issues. The review was carried out in two stages: first, we filtered out questions that were ambiguous, too generic, or incorrect; then, we revised the remaining questions as needed to improve clarity and correctness. We focused on the

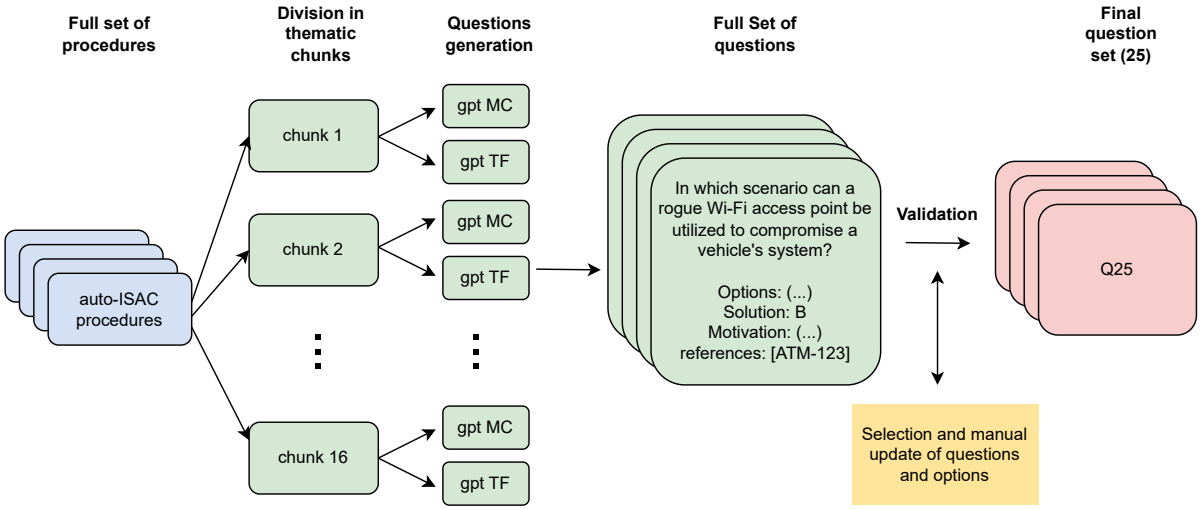



**Figure 2: The figure shows the entire pipeline we used for generating the questions. The initial set of procedures was first divided into *chunks*; each *chunk* was then used as an input prompt to GPT-4o to generate two class questions. After that, the questions were manually validated and, if necessary, modified to produce the final set of 25 correct questions.**

**Question:** A diagnostic function misuse allows the injection of CAN bus messages, which can result in unintended vehicle operations such as locking or unlocking the vehicle?
**Correct answer:** True
**Explanations**
- True: The misuse of a diagnostic function allows sending arbitrary CAN messages to affect vehicle functions like locking or unlocking

**Source procedures:** ATM-P0045

**Figure 3: Example of GPT-generated True/False question.**

answer options, ensuring that each was meaningful. Specifically, we verified that only one correct answer was identifiable and that wrong options were plausible but did not strictly concern the use case defined in the question text. Eventually, we collected the most accurate questions and crafted a survey. We asked industry experts (junior and senior) and researchers, from Ph.D. to full professor level, with competencies in automotive, cybersecurity, and/or hardware and software resilience, to answer the survey, serving as a benchmark for better understanding the relevance of the performance of LLMs on the same set of questions.

## 4 Results

This section presents the results from the question generation and assessment experiment. First, we report statistics on the number and type of questions generated and validated. Then, we present demographics of the survey respondents. Finally, we present the performance of five different LLMs on the built quiz, benchmarking the results with responses given by a group of domain experts.

**Question:** What technique involves modifying the ECU's software to run customized code permanently by bypassing update verification?
**A)** Unintended Vehicle Network Message
**B)** Modify OS Kernel, Boot Partition, or System Partition
**C)** Bypass Mandatory Access Control
**D)** Network Sniffing
**Correct answer:** B
**Explanations**
- A) Phishing involves social engineering, not ECU software modification:
- B) *Correct.* Modifying the OS kernel, boot partition, or system partition involves updating the ECU software to run customized code
- C) Bypassing mandatory access control deals with access policies, not software modification
- D) Network sniffing involves capturing data in transit, not modifying ECU software

**Source procedures:** ATM-P0148

**Figure 4: Example of GPT generated multiple choice question.**

Table 2 summarizes the outcome of the question generation process. Of the 80 multiple-choice (five per each of the 16 chunks) questions generated, 10 were kept after validation and manual update, and 70 were discarded due to ambiguity or factual errors. Similarly, for true/false questions, we kept 15 questions out of the 80 generated. This resulted in a final validated set of 25 questions: 10 multiple-choice and 15 true/false.

**Table 2: Question Generation Summary**

| Format | Kept | Modified | Discarded |
|---|---|---|---|
| Multiple Choice | 10 | all at least slightly modified | 70 |
| True/False | 15 | all at least slightly modified | 65 |

The survey attracted 17 respondents from diverse academic and industry backgrounds, as detailed in Table 3. The participant pool was well-balanced between academia and industry, with full professors representing the largest single group (23.5%), followed by junior industry experts (23.5%) and Ph.D. students (17.6%). Experience levels were notably high, with 35.3% of respondents having more than 15 years of professional experience, and an additional 23.5% having 6–15 years. Cybersecurity expertise was the strongest among participants, with 47.1% reporting "Very High" competence and only 5.9% reporting "Low" expertise. In contrast, automotive domain expertise showed a more distributed pattern, with 47.1% reporting "Low" expertise, leaving the majority of the expertise with at least a "Medium" level of knowledge on the specific subfield. Eventually, Hardware/Software Resilience expertise was more evenly distributed among the 4 expertise levels.

**Table 3: Demographics and expertise profile of survey respondents (N=17)**

| Category | Description | Count (%) |
|---|---|---|
| **Role** | | |
| | Full Professor | 4 (23.5%) |
| | Associate Professor | 2 (11.8%) |
| | Assistant Professor | 2 (11.8%) |
| | Junior Industry Expert | 4 (23.5%) |
| | Senior Industry Expert | 1 (5.9%) |
| | Post Doc | 1 (5.9%) |
| | PhD Student | 3 (17.6%) |
| **Years of Experience** | | |
| | Less than 3 years | 5 (29.4%) |
| | 3-6 years | 2 (11.8%) |
| | 6-15 years | 4 (23.5%) |
| | More than 15 years | 6 (35.3%) |
| **Automotive Domain Expertise** | | |
| | Low | 8 (47.1%) |
| | Medium | 4 (23.5%) |
| | High | 4 (23.5%) |
| | Very High | 1 (5.9%) |
| **Cybersecurity Domain Expertise** | | |
| | Low | 1 (5.9%) |
| | Medium | 4 (23.5%) |
| | High | 4 (23.5%) |
| | Very High | 8 (47.1%) |
| **HW/SW Resilience Expertise** | | |
| | Low | 4 (23.5%) |
| | Medium | 7 (41.2%) |
| | High | 3 (17.6%) |
| | Very High | 3 (17.6%) |

Figure 5 illustrates the performance distribution comparing human experts with AI models on the 25-question test. The results reveal several key findings regarding the relative capabilities of LLMs versus domain experts over the Auto-ISAC knowledge base.

All tested LLMs demonstrated significantly superior performance compared to the human expert baseline, with scores ranging from 80% to 100%. Claude Sonnet 4 achieved perfect performance (25/25 correct answers, 100%), followed closely by Gemini 2.5 Pro scoring 24/25 (96%) and DeepSeek-V3 with 23/25 (92%). GPT-4o achieved 22/25 correct answers (88%), while GPT-4o-mini, despite being the smallest model tested, still achieved 20/25 (80%). Notably, the mean of the LLMs performance, 91.2%, exceeded the human expert mean of 64.7% by a substantial margin of 26.5%.

The human experts' performances exhibited considerable variability, with scores ranging from 40% to 92% and a standard deviation of 16.8%. The distribution shows a right skew, with the majority of participants (11 out of 17, 64.7%) scoring between 10-18 correct answers (40-72%). Only three experts achieved the maximum human score of 92% (23/25 correct), which equals the performance of DeepSeek-V3.

Analyzing the competence profiles of these top three performers (Table 4), we can see that respondent 1, a Senior Industry Expert with 6-15 years of experience, compensated for a low automotive domain expertise through very high cybersecurity and hardware/software resilience knowledge. Respondent 9, an Assistant Professor with similar experience levels, achieved top performance through a balanced profile of medium automotive expertise, high cybersecurity knowledge, and medium resilience competence. Notably, Respondent 12, despite having less than 3 years of experience as a Junior Industry Expert, matched the performance of more senior colleagues through very high automotive domain expertise and resilience knowledge, combined with medium cybersecurity competence. This kind of analysis can help in future developments, allowing us to better target the domain expert sample.

**Table 4: Competence profile of top 3 performers (92% accuracy)**

| ID | Role | Exp. | Automotive | Cybersecurity | Resilience |
|---|---|---|---|---|---|
| **1** | Senior Industry Expert | >6<15 | **L** | **VH** | **VH** |
| **9** | Assistant Professor | >6<15 | **M** | **H** | **M** |
| **12** | Junior Industry Expert | <3 | **VH** | **M** | **VH** |

## 5 Discussion

This work-in-progress study offers early insights into LLMs capabilities in automotive cybersecurity. Although limited to a small sample of 25 validated questions from the Auto-ISAC knowledge base and a limited expert sample (N=17), the results indicate that models like Claude Sonnet 4 (100%) and Gemini 2.5 Pro and DeepSeek-V3 (96%) possess substantial knowledge in this domain. The human experts' mean score (64.7%) shows that the questions were challenging enough to test domain knowledge, while the LLMs clearly outperformed them, highlighting their depth of understanding.

However, the process of question generation itself revealed several challenges. Out of the 160 questions initially generated by

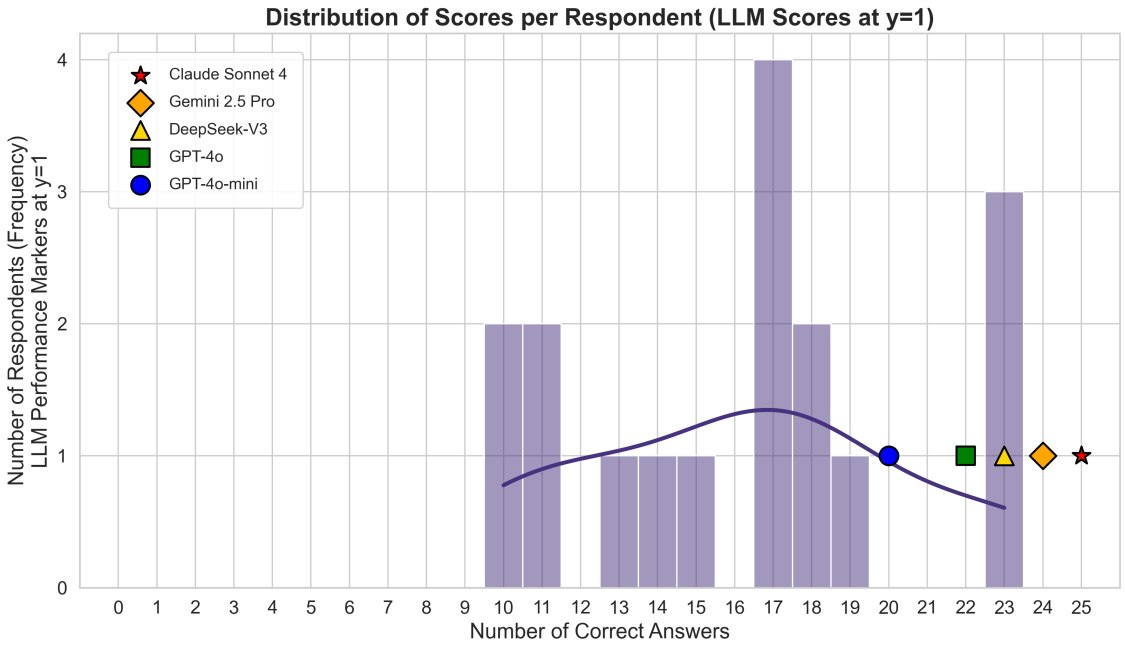



**Figure 5: Distribution of human respondent scores on the survey, overlaid with performance of the LLMs tested. Each vertical bar indicates how many respondents achieved a given score. Markers at $y = 1$ show LLMs performance: Claude Sonnet 4 (25/25), Gemini 2.5 Pro (24/25), DeepSeek-V3 (23/25), GPT-4o (22/25), and GPT-4o-mini (20/25).**

GPT-4o, only 25 survived after manual filtering and revision, implying a drastic discard rate of more than 80%. This filtering was necessary due to common issues such as ambiguity, factual inaccuracies, or reasoning errors. Such a high removal rate underlines the current fragility of relying on LLMs for question creation.

## 6 Conclusion and Future Work

This work-in-progress study provides initial evidence that state-of-the-art LLMs significantly outperform human domain experts in a developed knowledge assessment test of 25 questions in the domain of automotive cybersecurity, with Claude Sonnet 4 achieving perfect accuracy (100%) and with a mean between LLMs of 91.2% with respect to the human experts performance with mean 64.7%.

At the same time, the study highlighted the limitations of relying on LLMs for generating valid assessment items. Errors often arise from both model-side issues—such as hallucinations or shallow contextualization—and human-side factors, including prompt design or subjective filtering criteria. This interplay makes it difficult to clearly attribute responsibility, suggesting the need for hybrid strategies where automation is complemented by expert oversight.

Future work will aim to expand this preliminary study in order to strengthen the generality of the obtained results. First, we plan to enlarge the survey by increasing the number of questions and incorporating material from a broader set of automotive cybersecurity sources beyond Auto-ISAC. This expansion will be accompanied by a more stratified and diverse sample of domain experts, enabling a more accurate and representative baseline for comparison. In parallel, we will conduct a systematic error analysis to better distinguish between mistakes originating from LLMs and those introduced by human factors such as prompt design or validation bias. Finally, future research will explore more granular assessments of LLM knowledge across specific subdomains of automotive cybersecurity, providing a clearer picture of where these models excel and where human expertise remains indispensable.

We have published the LLMS-generated questions and code at https://github.com/smilies-polito/autoISAC_LLM_Knowledge

## Acknowledgments

This study was carried out within the SERICS - Security and Rights in the CyberSpace and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1556 11/10/2022, PE00000014). This manuscript reflects only the authors' views and opinions. Neither the European Union, nor the European Commission, nor the Ministry can be considered responsible for them.

## References

[1] AlienVault, now AT&T Cybersecurity. 2025. Open Threat Exchange (OTX). https://otx.alienvault.com/. [Online; accessed 18-June-2025].
[2] Auto-ISAC, Inc. 2025. Auto-ISAC: Automotive Information Sharing and Analysis Center. https://automotiveisac.com/. [Online; accessed 18-June-2025].
[3] Aviation Information Sharing and Analysis Center (A-ISAC). 2025. A-ISAC: Information Sharing and Analysis Center for Aviation. https://https:/www.a-isac.com/. [Online; accessed 18-June-2025].
[4] Energy Information Sharing and Analysis Center (E-ISAC). 2025. E-ISAC: Information Sharing and Analysis Center for the Energy Sector. https://www.eisac.com/s/. [Online; accessed 18-June-2025].

[5] Pavlos Evangelatos, Christos Iliou, Thanassis Mavropoulos, Konstantinos Apostolou, Theodora Tsikrika, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2021. Named Entity Recognition in Cyber Threat Intelligence Using Transformer-based Models. In *2021 IEEE International Conference on Cyber Security and Resilience (CSR)*. IEEE, Rhodes, Greece, 348–353. doi:10.1109/CSR51186.2021.9527981

[6] Peng Gao, Xiaoyuan Liu, Edward Choi, Sibo Ma, Xinyu Yang, Zhengjie Ji, Zilin Zhang, and Dawn Song. 2022. ThreatKG: A Threat Knowledge Graph for Automated Open-Source Cyber Threat Intelligence Gathering and Management. doi:10.48550/ARXIV.2212.10388

[7] Ethan Garza, Erik Hemberg, Stephen Moskal, and Una-May O'Reilly. 2023. Assessing Large Language Model's knowledge of threat behavior in MITRE ATT&CK. In *ACM KDD AI4Cyber: The 3rd Workshop on Artificial Intelligence-enabled Cybersecurity Analytics at KDD'23*. ACM, Long Beach, California, 1–7.

[8] Health Information Sharing and Analysis Center (H-ISAC). 2025. Health-ISAC: Information Sharing and Analysis Center for Healthcare. https://health-isac.org/. [Online; accessed 18-June-2025].

[9] ISO. 2021. ISO/SAE 21434:2021. https://www.iso.org/. https://www.iso.org/standard/70918.html).

[10] Seonghoon Jeong, Huy Kang Kim, Mee Lan Han, and Byung Il Kwak. 2024. AERO: Automotive Ethernet Real-Time Observer for Anomaly Detection in In-Vehicle Networks. *IEEE Transactions on Industrial Informatics* 20, 3 (2024), 4651–4662. doi:10.1109/TII.2023.3324949

[11] Sadek Misto Kirdi, Nicola Scarano, Franco Oberti, Luca Mannella, Stefano Di Carlo, and Alessandro Savino. 2024. CARACAS: vehiCular ArchitectuRe for detAiled Can Attacks Simulation. In *2024 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, Paris, France, 1–6. doi:10.1109/ISCC61673.2024.10733705

[12] Keen Security Lab. 2020. Experimental Security Assessment on Lexus Cars. https://keenlab.tencent.com/en/2020/03/30/Tencent-Keen-Security-Lab-Experimental-Security-Assessment-on-Lexus-Cars/. Accessed: 2025-06-28.

[13] MITRE Corporation. 2025. MITRE ATT&CK: Adversarial Tactics, Techniques, and Common Knowledge. https://attack.mitre.org/. [Online; accessed 18-June-2025].

[14] Franco Oberti, Ernesto Sanchez, Alessandro Savino, Filippo Parisi, and Stefano Di Carlo. 2021. Mitigation of Automotive Control Modules Hardware Replacement-based Attacks Through Hardware Signature. In *2021 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks - Supplemental Volume (DSN-S)*. IEEE, Taipei, Taiwan, 13–14. doi:10.1109/DSN-S52858.2021.00017

[15] Nicola Scarano, Luca Mannella, Alessandro Savino, and Stefano Di Carlo. 2024. Can social media shape the security of next-generation connected vehicles?. In *2024 IEEE 30th International Symposium on On-Line Testing and Robust System Design (IOLTS)*. IEEE, Rennes, France, 1–4. doi:10.1109/IOLTS60994.2024.10616053

[16] Samaneh Shafee, Alysson Bessani, and Pedro M. Ferreira. 2025. Evaluation of LLM-based chatbots for OSINT-based Cyber Threat Awareness. *Expert Systems with Applications* 261 (2025), 125509. doi:10.1016/j.eswa.2024.125509

[17] Norbert Tihanyi, Mohamed Amine Ferrag, Ridhi Jain, Tamas Bisztray, and Merouane Debbah. 2024. CyberMetric: A Benchmark Dataset based on Retrieval-Augmented Generation for Evaluating LLMs in Cybersecurity Knowledge. In *Proceedings of the 2024 IEEE International Conference on Cyber Security and Resilience (CSR)*. IEEE, London, United Kingdom, 296–302. doi:10.1109/CSR61664.2024.10679494

[18] Jie Zhang, Haoyu Bu, Hui Wen, Yongji Liu, Haiqiang Fei, Rongrong Xi, Lun Li, Yun Yang, Hongsong Zhu, and Dan Meng. 2025. When LLMs meet cybersecurity: a systematic literature review. *Cybersecurity* 8, 1 (2025), 55. doi:10.1186/s42400-025-00361-w

## Glossary

**ADAS** Advanced Driver Assistance Systems. 1

**ATT&CK** Adversarial Tactics, Techniques, and Common Knowledge. 2, 3

**Auto-ISAC** Automotive Information Sharing and Analysis Center. 1–3, 5

**CAN** Controller Area Network. 1, 3

**ISAC** Information Sharing and Analysis Center. 2

**LLM** Large Language Model. 1–6

**NER** Named Entity Recognition. 2

**NLP** Natural Language Processing. 2

**OSINT** Open-Source Intelligence. 1, 2