

Evaluating Large Language Models in Cybersecurity Knowledge with Cisco Certificates

Gustav Keppler¹[0000–0002–2323–0533], Jeremy Kunz¹[0009–0005–3857–2578], Veit Hagenmeyer¹[0000–0002–3572–9083], and Ghada Elbez¹[0000–0003–1137–1782]

Institute for Automation and Applied Informatics, KASTEL Security Research Labs,
Karlsruhe Institute of Technology (KIT), Germany
`{gustav.keppler,jeremy.kunz,veit.hagenmeyer,ghada.elbez}@kit.edu`

Abstract. As generative artificial intelligence evolves, understanding the capabilities in the cybersecurity domain becomes crucial. This paper examines the capability of Large Language Models (LLMs) models in solving cybersecurity certification Multiple Choice Question Answering (MCQA) exams, comparing proprietary and open-weights models. Challenges related to test-set leakage, notably on the widely used MMLU benchmark, emphasize the need for continuous validation of benchmarking results. Open-weights models, namely Mistral Large 2, Qwen 2, and Phi 3, seem to overfit the MMLU Computer Security and indicate less usability for cybersecurity knowledge tasks. The study also introduces the first visual cybersecurity MCQA benchmark, assessing the capability of Large Multimodal Models (LMMs) in interpreting and responding to visual questions. Among the tested models, the proprietary Anthropic Claude 3.5 Sonnet and OpenAI GPT-4o outperformed others in the language and vision-language setting. However, Llama 3.1 model series demonstrated significant advancement in the open-weights domain, signaling potential parity in cybersecurity knowledge with proprietary models in the near future. Code and datasets are available at: <https://github.com/GKeppler/GenAICyberSecMCQA>.

Keywords: Large Language Models (LLMs) · MCQA · Benchmarking · Cybersecurity · Large Multimodal Models (LMMs) · Visual Question Answering

1 Introduction

Generative Artificial Intelligence (AI), particularly LLMs, represents an advancement in how machines process and generate human language. These models are trained on extensive datasets comprising diverse human language inputs with trillions of words[5]. At the center of modern LLMs is the transformer architecture[24], which has evolved natural language processing (NLP) due to its effectiveness in handling large-scale language data.

One of the leading applications of this technology is OpenAI’s GPT-4[20], a model based on the Generative Pre-trained Transformer. It demonstrates profound capabilities in generating textual content, programming code, and even

poetry in various stylistic imitations. GPT-4 and competitors have also showcased their utility in the cybersecurity domain[19,28,30], outperforming traditional commercial tools in some areas, for example, automated program repair methods[27].

LMMs, which interpret and generate information across various forms of data like text, audio, and images, further expands the applicability of generative AI[13,10]. These advancements underscore the importance of understanding the characteristics and capabilities of LMMs in critical fields such as cybersecurity, where these models have not yet been explored.

MCQA is crucial for assessing LLMs capabilities, used widely for its deterministic measurement and compatibility with human testing protocols. In MCQA, models are tested for comprehension, application of pre-training knowledge, and reasoning abilities by providing answers from a set of options. Such assessments are essential for ranking LLMs in recognized benchmarks, including the Holistic Evaluation of Language Models benchmark[17].

One of the most widely used MCQA benchmarks is the Massive Multitask Language Understanding (MMLU)[12] which also features a computer security subtask of 100 questions. As LLMs advance, their performance on the MMLU has plateaued, making it difficult to discern differences in model capabilities. The MMLU-Pro[25] aims to address the plateauing performance by incorporating more complex, reasoning-intensive questions to better differentiate model capabilities in language comprehension and reasoning across various domains. However, it no longer features the computer security subtask of the MMLU.

Also, given the importance of MCQA in evaluating LLMs, it is vital to ensure that the accuracy obtained through MCQA reflects the abilities being measured. However, static benchmarks can face issues with test set leakage, which leads to accidental contamination of the training data of the LLMs or overfitting on the test set[21]. Developers who access these test sets might incorporate them into the training datasets, either unknowingly or intentionally inflating performance metrics. Widely used benchmarks, like the MMLU, pose this risk especially as the data is widespread on the internet.

This limits their reliability and suggests more benchmarks to mitigate contamination risks[9], which is addressed in this study. The key findings are summarized as follows:

- Performance in Cybersecurity Certifications: Proprietary models such as Anthropic’s Claude 3.5 Sonnet and OpenAI’s GPT-4o consistently achieved at least an 80% passing grade on Cisco certifications. In contrast, open-source models like Llama3.1 showed variability, excelling in CCNA exams more consistently than in CCNP.
- Model Overfitting: The overfitting in several models, including Qwen2 and Phi-3 as demonstrated by their inconsistent performance on CCNA, CCNP, and MMLU benchmarks is observed. In contrast, models like Llama3.1, Claude 3.5 Sonnet, and GPT-4o exhibited more consistent results across these benchmarks, suggesting more adapted training methodologies.

- Vision-based Question Answering: Both GPT-4o and Claude 3.5 Sonnet achieve high accuracy on vision-based cybersecurity exams, maintaining accuracy comparable to text-based questions and indicating their prowess across modalities.

The rest of the paper is organized as follows. Section II presents related work, while Section III presents a detailed overview of the methodology adopted for evaluating the performance on cybersecurity certification questions, including the creation and utilization of the first visual cybersecurity MCQA dataset. Section IV analyzes the data, discussing the accuracy of both open-weights and proprietary models in various testing scenarios and exploring potential biases caused by training on the test set. Sections V and VI evaluate and discuss the implications of the results for the field, emphasizing the challenges and opportunities identified through the study. Finally, Section VII concludes the paper.

2 Related Work

There is a variety of specialized benchmarks used for evaluating the cybersecurity knowledge of LLMs. From assessing the generation of insecure code to measuring the understanding and application of security principles, these benchmarks provide critical insights into the capabilities of LLMs within the cybersecurity domain.

As mentioned above, MCQA benchmarks are widely utilized. Besides the MMLU Computer Security, WMDP[16] features 3668 multiple-choice questions that evaluate LLMs’ expertise in biosecurity, cybersecurity, and chemical security. All questions were generated and controlled by experts in the field. The dataset is intended to be a proxy-benchmark that can be used to untrain highly sensitive knowledge in LLMs. SecEval[14] offers a set of multiple-choice questions covering various security domains. It utilizes OpenAI’s GPT-4 to generate questions by sourcing content from authoritative materials including open-licensed textbooks, official documentation, and industry guidelines and standards. SecQA[18] also leverages GPT-4 to generate questions based on a security textbook. However, GPT-3.5-Turbo and GPT4 reach almost 100% accuracy, indicating a lack of difficulty. The quality assessment of the generated questions is done by researchers review, but not specified further. The CyberMetric[23] benchmark series, with datasets ranging from 80 to 10,000 questions, evaluates LLMs cybersecurity knowledge using questions generated from textbooks via GPT-3.5 and Retrieval-Augmented Generation. Developed questions underwent expert validation. However, this setting reaches 96.25% GPT-4, indicating a benchmark saturation.

Overall, GPT plus Retrieval Augmented Generation (RAG) created benchmarks offer customizable, scalable, and cost-effective assessment tools that can provide extensive coverage of topics. However, they lack quality assurance of the generated questions and answers. Therefore, using industry-recognized professional certification exams provides an authentic and standardized benchmark

that is highly respected in the field. Tann et al.[22] test LLMs question-answering capabilities across Cisco certification exams. The aim is to determine if LLMs can pass these industry-recognized professional certification exams. To the best of the author’s knowledge, this only works by benchmarking the knowledge of LLMs with Cisco certification exams. However, their datasets are not publicly available and they use outdated Cisco exams.

3 Experimental Setup

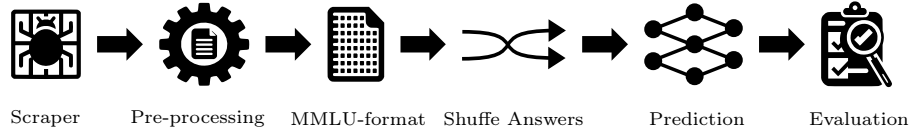


Fig. 1. Overview of the data collection, processing, and evaluation of the multiple choice questions answering benchmarks using LLMs.

In this section, the methodologies employed to evaluate the capability of LLMs in answering multiple-choice cybersecurity certification questions are outlined. This includes the shuffling of the answer possibilities[7], evaluation of an LLM overfitting on the test set. Additionally, as not only one answer per question can be correct, the score per question depends on the proportion of correct and incorrect options chosen by the LLMs. The process is shown in Figure 1

3.1 MCQA Task Definition

The MCQA task[7] for LLMs is defined as : 1) a question q ; and 2) a set of choices of varying length $\mathcal{C} = \{c_a, c_b, c_c, c_d, c_e\}$, where one or multiple are correct. Utilizing these as inputs, the LLM must give one or multiple letters of the correct option $a \in \{ (A), (B), (C), (D), (E) \}$.

In this study, a 5-shot approach is employed, where models are provided with five examples before answering each question. The standard MCQA prompt is a full prompt, including the 5 few-shot examples, the question q , and choices \mathcal{C} :

Question: q
 Choices:
 (A) c_a
 (B) c_b
 (C) c_c
 (D) c_d
 (E) c_d
 Your response should end with "The best answer is [the_answer_letter(s)]" where the [the_answer_letter(s)] is/are of A, B, C, D, E, ...
 The best answer is a

As visual question answering is also examined in this study, a separate benchmark has been included to assess LMMs capabilities. The same 5-shot templates as for the language-only multiple-choice questions benchmarks are utilized, adding the picture to the question. The few-shot examples were not adapted to the visual setting.

To assess the accuracy of the benchmark, the following considerations are made: If all correct options and no incorrect ones are selected, the answer is considered completely correct. Partial accuracy and exact accuracy are computed by averaging these scores and the number of wholly correct answers over the entire question set, respectively. This is further described in Algorithm 1.

Input: QSet: multiple-choice questions, LLMSet: large language models

Output: PartialAccuracy, ExactAccuracy: accuracy of LLMs in answering questions correctly

Function EvaluateMCQ(QSet, LLMSet):

```

foreach LLM in LLMSet do
  CorrectAnswers  $\leftarrow$  0
  PartialCorrectAnswers  $\leftarrow$  0
  foreach Q in QSet do
    Answer  $\leftarrow$  LLM(Q)
    T  $\leftarrow$  len(CorrectOptions)
    C  $\leftarrow$  number of correct options
    W  $\leftarrow$  number of incorrect options
    Score  $\leftarrow$  max(0,  $\frac{C}{T} - \frac{W}{T}$ )
    PartialCorrectAnswers  $\leftarrow$  PartialCorrectAnswers + Score
    if C == T and W == 0 then
      CorrectAnswers  $\leftarrow$  CorrectAnswers + 1
    end
  end
  PartialAccuracy[LLM]  $\leftarrow$   $\frac{\text{PartialCorrectAnswers}}{\text{len}(QSet)}$ 
  ExactAccuracy[LLM]  $\leftarrow$   $\frac{\text{CorrectAnswers}}{\text{len}(QSet)}$ 
end
return PartialAccuracy, ExactAccuracy

```

Algorithm 1: Multiple-Choice Question Answering Accuracy

3.2 Model Selection: open-weights and closed-weights models

The selection of models for the project was guided by comprehensive performance evaluation and specialization across various platforms. The HELM MMLU Leaderboard[17] offers a third-party evaluation of the accuracy for each of the 57 topics in the MMLU - including Computer Security - for a variety of open-weights and proprietary LLMs. The Chatbot Arena[11] is additionally used for model selection as it offers human-evaluated instruction following capability scoring, as all models selected for this study are instruction-following tuned. This results in the following models:

The Llama3.1[5] 405B Instruct stood out by achieving the highest scores overall on the HELM MMLU and the Chatbot Arena Leaderboard from an open-weights model. Additionally, the Llama3.1 70B Instruct and the Qwen2 72B Instruct were chosen as the medium-size models due to their performance on the MMLU HELM. The Qwen2 series are the best open-weights models available from a Chinese company. Mistral Large 2-2407 was selected as it was the top-performing open-weights LLM in the Chatbot Arena-Hard category[15] and is the most prominent model from a European provider.

In the realm of proprietary models, Anthropic Claude 3.5 Sonnet[4] with its latest update from June 20th, 2024, achieved the highest score in the HELM MMLU overall and also specifically in MMLU Computer Security. OpenAI GPT-4o[3] version from May 13, 2024, is also included in our selection. Although it ranked fourth on the HELM MMLU overall - lower than the open-weights Llama 3.1 405b Instruct - it achieved the highest scores in the Chatbot Arena[11] overall. Both models offer multi-modal capabilities, making them suitable for the vision datasets.

Furthermore the Phi-3 series models[6], Phi-3-mini-128k-instruct and Phi-3-medium-4k-instruct were selected, due to their MMLU score of 70.9% and 78% (self-reported, as not mentioned on the HELM leaderboard). Despite their relatively smaller size of only 3.8B/14B parameters, they outperform larger models such as the Llama 2 70B, which scored 69.5% in HELM results. The vision-language version of the Phi-3-mini model, the Phi3-vision-128k-instruct LMMs is used to compare to the proprietary vision-language models.

The recent release of open-weights LLMs include pre-trained, and a further instruction-tuned version of the model. In this work, only the instruction-tuned versions are compared, as the State-of-the-Art closed-weight Models are only available as instruction-tuned or chat variations. These models use a chat template with special tokens indicating the roles in a multi-turn conversation. The raw model of the proprietary models cannot be accessed. There are three distinct methods to prompt open-weights models. Firstly, the raw mode where the conversation tokens

```
<|eot_id|><|start_header_id|>assistant<|end_header_id|>
```

need to be inserted manually. Contrarily, the chat completion mode relies on a chat template, which is compatible with the OpenAI API and is therefore used

for this work as closed-weights and open-weights models can be compared. The templates used in this work can be examined in Appendix 7.

3.3 Shuffling and Parameters

Shuffling the answer options can be employed as a strategy to determine the consistency and variance in the accuracy of LLMs when operating at a temperature of zero. By rearranging the order of these choices, it can assess whether the model’s performance is stable across different presentations of the same question, thereby providing a clearer understanding of its reliability.

For open-weight models, testing is conducted in an unquantized fp16 state, as this configuration delivers superior performance compared to quantized models, although with increased computational requirements. The temperature is set to zero with a sampling rate of 1, aligning with the testing procedure used for MMLU. If the output cannot be serialized into the answer format with the used regex-pattern, there will be no resampling with for example higher temperature performed. In the scenario of 5-shot tasks, the number of shuffles is set to 5, and the maximum output tokens one can generate is set to 10. For 0-shot Chain of Thought (CoT)[26] prompting, there is no shuffling applied, and the maximum output tokens are significantly higher, set at 1000, allowing for the more expansive generation of answer explanation. This allows the model to explain its reasoning step-by-step.

When processing answers from model-generated content, the primary method involves a regex pattern ‘answer is \[?([A-J]+)\]?’ . The regex ‘.*[aA]nswer:\s*([A-J]+)’ is used in addition. For CoT prompting, first the regex ‘answer is ([^.]*)\.’ is used to match an answer sequence, and then all uppercase letters ‘[A-Z]’ are extracted from the segment.

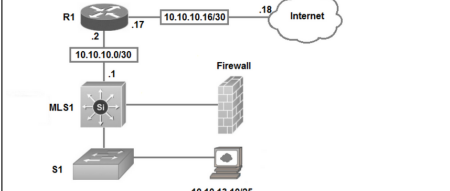
4 Dataset

The tested MCQA benchmarks in this study include the MMLU Computer Security[12], as a subtask of the widely used MMLU benchmark, as well as the CISCO Career certification exams[1] 200-301 CCNA and 350-701 SCOR are also used. The exam questions were scraped from the internet[2], as the scraping policy allows it. They were chosen as promising candidates that meet the following requirements for LLM knowledge benchmarks for the cybersecurity domain:

- Possess sufficient difficulty to prevent benchmark plateau and effectively distinguish between models with high confidence.
- Reduced test-set leakage that exists for widely used benchmarks.
- Address quality assurance issues prevalent in the GPT self-generation of datasets.

Regarding the first point, the scraped exams represent two levels of certification difficulty: The Cisco Certified Network Associate (CCNA) certification,

Table 1. Two examples from the 201-301-CCNA (orange) and 350-701-CCNP (blue) vision datasets. The correct answer is bolt and the exhibit is shown at the bottom.

201-301-CCNA Vision	350-701-CCNP Vision
Question: Refer to the exhibit. Which type of route does R1 use to reach host 10.10.13.10/32?	Question: Refer to the exhibit. What does the API do when connected to a Cisco security appliance?
Options: (A) default route (B) network route (C) host route (D) floating static route	Options: (A) create an SNMP pull mechanism for managing AMP (B) gather network telemetry information from AMP for endpoints (C) get the process and PID information from the computers in the network (D) gather the network interface information about the computers AMP sees
 <pre> R1#sh ip ro Gateway of last resort is 10.10.10.18 to network 0.0.0.0 C 10.0.0.0/8 is variably subnetted, 4 subnets, 3 masks C 10.10.10.0/30 is directly connected, FastEthernet0/1 O 10.10.13.0/25 [110/6576] via 10.10.10.1, 06:58:21, FastEthernet0/1 C 10.10.10.16/30 is directly connected, FastEthernet0/24 O 10.10.13.144/28 [110/110] via 10.10.10.1, 06:58:21, FastEthernet0/1 O 0.0.0.0/0 [20/0] via 10.10.10.18, 01:17:58 </pre>	<pre> import requests client_id = 'a1b2c3d4e5' api_key = 'a1b2c3d4-e5f6-g7h8' url = 'https://api.amp.cisco.com/v1/computers' response = requests.get(url, auth=(client_id, api_key)) response_json = response.json() for computer in response_json['data']: network_addresses = computer['network_addresses'] for network_interface in network_addresses: mac = network_interface.get('mac') ip = network_interface.get('ip') ipv6 = network_interface.get('ipv6') print(mac, ip, ipv6) </pre>

though not strictly focused on cybersecurity, provides fundamentals - which include routing, switching, and security concepts - that are crucial for a cybersecurity career. There are no prerequisites for this certification, and the difficulty level ranges from entry-level to intermediate. The required exam for this certification is the 200-301 CCNA.

The advanced Cisco Certified Network Professional Security (CCNP Security) certification provides an in-depth knowledge of securing Cisco networks. This knowledge encompasses firewall technologies, VPNs, intrusion prevention, and endpoint security. To pursue this certification, a valid CCNA certification or equivalent experience is required. The core exam for this certification is the 350-701 SCOR, and a choice of an additional topic needs to be completed for certification.

In terms of test-set leakage, the lack of disclosure regarding the training data of the compared models makes it difficult to determine whether CCNA and CCNP questions were included. However, this new benchmark likely gains an advantage from its relatively limited representation in internet text. Unlike widely established benchmarks such as MMLU, the lesser-known status of this

benchmark potentially reduces the risk of test-set leakage, as its questions may be less exposed during the models' training phases.

Compared to the MMLU Computer Security exams, Cisco multiple-choice question exams present a few key differences. Cisco exams can have varying lengths of answer possibilities, allow for multiple answer choices within a single question, as indicated by prompts such as "choose two.", and include vision-based questions.

The MMMU[29] benchmark includes vision-based multi-choice questions, however, it does not include computer security questions, only computer science questions. Therefore, there is no related vision-language benchmark available for the cybersecurity domain, and the "201-301-CCNA Vision" and "350-701-CCNP Vision" are the first of their kind - example questions are shown in Table 1.

The pre-processing of the scraped questions and answers involves the following steps:

- **De-duplication:** Perform de-duplication regarding the questions.
- **Low-quality images:** Drop questions that feature low-quality images with less than 100x100 pixels.
- **Drag-and-drop questions:** Remove questions where not at least one answer possibility is given, as they are not multiple choice questions, but query-answer matching tasks.
- **Sampling:** Sample from the remaining questions for 100 questions, if available. The image-based questions are less.
- **MMLU-Format conversion:** Convert questions, answers, and choices to the MMLU format, for easy integration of the benchmark into other packages.

Table 2. An overview multiple-choice question answering benchmarks for the cybersecurity domain. The amount of questions, the source of the questions, and the reported GPT-4 accuracy are shown.

Name	Count	Development	GPT4 Acc.
MMLU Computer Security[12]	100	Domain experts	86
SecEval[14]	2126	GPT-4	79.07
SecQA[18]	210	GPT-4 + RAG	98.0
CyberMetric[23]	10000	GPT-3.5 + RAG, human verified	88.50
WMDP-Cyber[16]	1,987	Domain experts	-
201-301-CCNA	100	Domain experts	-
350-701-CCNP	100	Domain experts	-
201-301-CCNA Vision	70	Domain experts	-
350-701-CCNP Vision	15	Domain experts	-

Regarding the data quality of new benchmarks, Multiple choice questions generated by models like GPT from textbook data are susceptible to issues

of hallucination and general quality assurance problems, because these models sometimes generate content that blends learned knowledge with confabulated details. While the model attempts to understand and replicate the information style and content from textbooks, its inability to perfectly discern facts from plausibility can lead to hallucinations[8]. Therefore, without rigorous vetting and quality control, the use of such AI-generated quiz content could compromise the benchmark quality. An overview of the scraped datasets, as well as other publicly available MCQA datasets in the cybersecurity domain, is shown in Table 2.

5 Evaluation of Results

In this chapter, the accuracy of LLMs on the cybersecurity certification exams, comparing open-weights and proprietary models across different prompting strategies and modalities, while also examining potential training data leakage and dataset quality is analyzed.

5.1 Solving cybersecurity certification questions

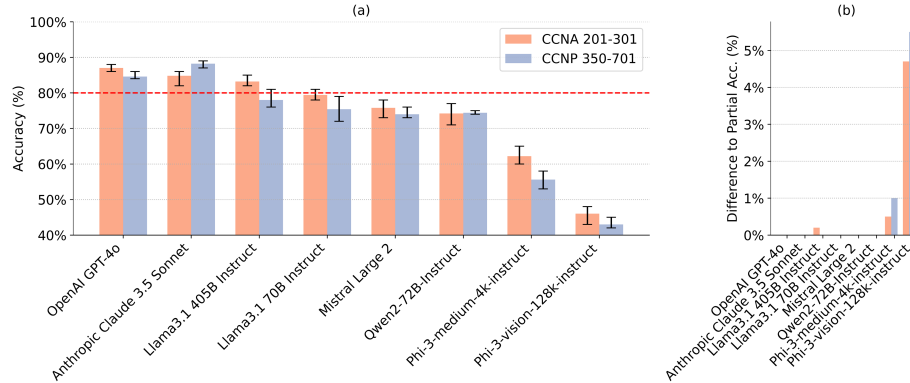


Fig. 2. (a) Overview of the accuracy of LLMs on the CCNA 201-301 and CCNP 350-701 5-shot. The error bar indicates the min. and max. accuracy of 5 shuffled runs. The red horizontal line is the assumed passing accuracy for the exams. (b) Accuracy gains if partial points are given, e.g. 0.5 point if [A] is answered, but [A,B] is correct.

In the following section, the performance of various language models on the CCNA and CCNP certification exams is evaluated by comparing 5-shot with 0-shot CoT prompting. The CCNA and CCNP have a passing grade assumed at 80% as shown in Figure 2. Among the models, Anthropic Claude 3.5 Sonnet and OpenAI GPT-4o stood out, consistently exceeding the 80% threshold across, with Anthropic Claude 3.5 Sonnet scoring as high as 88% in CCNP. However, it

is unclear why the CCNP score is higher than the CCNA for this model, as the difficulty of the CCNP exam is higher.

Llama3.1 405B is the only open-weights model passing CCNA every time, and the CCNP one out of five times. Qwen2.72B Instruct, Llama3.1 70B Instruct, and Mistral Large 2 did not meet the passing grade in the certification metrics. The Phi 3 models, Phi-3-medium and Phi-3-vision consistently fell short across all metrics, although they are more conservative when multiple answer options per question are correct, achieving partial points as shown in Figure 2 (b). The benchmark mostly states if multiple options should be selected, indicated with "choose two." Phi-3-vision often ignores this and answers with only one option, achieving partial points. This is also the case for one question in the CCNA exam and the Llama 3.1 405b model.

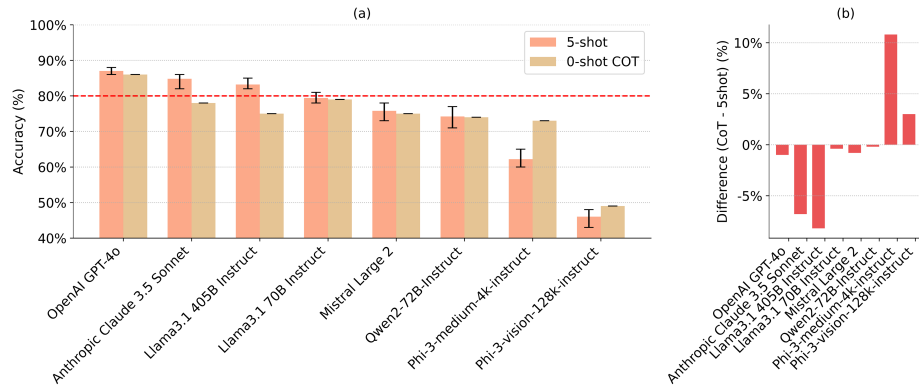


Fig. 3. (a) Overview of the accuracy comparing 5-shot vs 0-shot Chain of Thought prompting on the 200-301 CCNA benchmark. (b) Difference in the accuracy of the 5-shot vs 0-shot Chain of Thought prompting techniques.

The results presented in Figure 3 showcase the comparison of model accuracy between 5-shot and 0-shot CoT prompting. The range of accuracy over 5 shuffled runs is shown, while for the CoT prompting one unshuffled run was done. Across the models, the 5-shot setting outperforms the 0-shot CoT approach, besides the Phi-3 models. While the accuracy for the GPT-4o, Llama3.1 70b, Mistral and Qwen2 models are similar for both promoting techniques, the Claude 3.5 Sonnet and Llama3.1 405b show a notable decrease in accuracy with over 5%. The Phi-3 models benefit from the CoT prompting with the Phi-3-medium model showing over 10% accuracy increase, while the Phi-3-vision-128k-instruct model slightly surpasses the 5-shot method. Overall, the findings are highly variable.

Overall, the comparative analysis shows that only the proprietary models, namely Anthropic Claude 3.5 Sonnet and OpenAI GPT-4o, consistently achieved 80% passing grade across both datasets. Llama3.1 405B is the only open-weights model passing CCNA every time, and the CCNP one out of five times.

5.2 Trained on the test-set

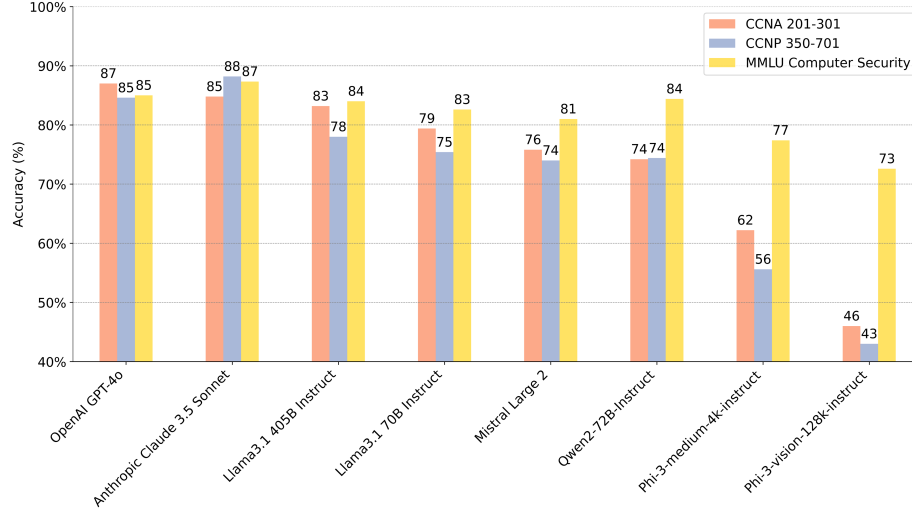


Fig. 4. Overview of the LLMs mean accuracy in the CCNA 200-301, CCNP 350-701, and the MMLU Computer Security benchmarks of 5 shuffled runs. The difference may indicate test-data leakage to the training data of the model.

The leakage of benchmark test sets on the internet poses a significant risk of contaminating training datasets in machine learning practices. As a result, models may appear to perform exceptionally well on tests not because they genuinely learned the underlying patterns and generalized well, but because they were indirectly trained on the test data. This compromises the integrity of performance evaluations and could lead to misleading conclusions about a model’s effectiveness when deployed in real-world scenarios. Figure 4 shows that the better-performing models exhibit similar accuracy on the MMLU Computer Security benchmark compared to the CCNA 201-301 and CCNP 350-701 benchmarks. Starting with the Mistral Large 2, notable declines in the certification question accuracy compared to their performance on the MMLU benchmark. These models may overfit on commonly used benchmarks or even specifically trained on the MMLU test-set. This discrepancy is larger for Qwen2, Mistral Large 2, and especially Phi-3, while it is less pronounced for Llama3.1, Claude, and GPT-4o, potentially reflecting better data handling and training practices.

5.3 Solving Visual Questions

Figure 5 highlights the vision capabilities of LMMs when tasked with solving image-based questions from the CCNA 201-301 vision and CCNP 350-701 vision datasets. The results indicate a notable variance in performance across the

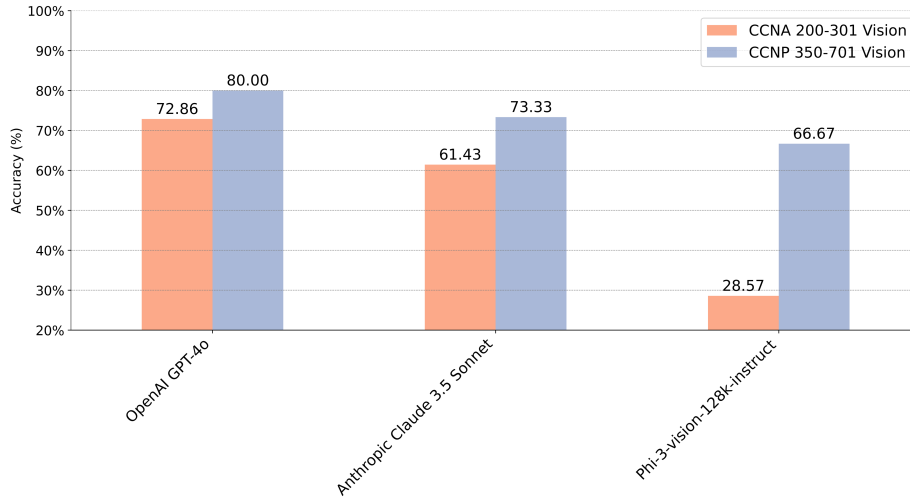


Fig. 5. Overview of the CCNA 201-301 Vision and CCNP 350-701 Vision accuracy.

models. OpenAI GPT-4o and Anthropic Claude 3.5 Sonnet demonstrate relatively high accuracy. In contrast, Phi-3-vision-128k-instruct shows significantly lower performance, particularly on the CCNA dataset. These results suggest that while advanced models like GPT-4o and Claude 3.5 Sonnet exhibit strong vision capabilities, also smaller multimodal models such as Phi-3-vision-128k-instruct achieve notable accuracy. However, the accuracy for CCNP 350-701 vision is higher for all tested models, indicating that the questions are easier to solve. This may be caused by a low number of samples in the CCNP 350-701 vision dataset of only 15 questions, where relatively easy questions may be present. In general are GPT-4o and Claude 3.5 Sonnet capable of solving vision-based questions with high accuracy, comparable to language questions. However, the overall separability for the CCNA vision benchmark is higher. Phi-3-vision shows lower accuracy.

6 Discussion

The dataset utilized was scraped from an online learning portal, as obtaining it through official channels was not feasible for publication. Despite the lack of formal validation, it benefits from being a community dataset, where the quality and relevance have been informally verified through widespread use and feedback within the community. This "community validation" offers assurance about the dataset's utility and applicability, although it may still contain inconsistencies.

Regarding the variance in prompting techniques, the results are consistent with the MMLU evaluations in the literature, where no prompting technique is advancing for all compared models[5]. Llama 3 70B and Llama 3 405B, benefit in

the MMLU benchmark from the 0-shot CoT prompting, showing improved performance over the 5-shot method, while GPT-4o and Claude 3.5 Sonnet achieve better results in the 5-shot setting. However, the high drop in accuracy for the Anthropic and LLama3.1 405b models, as well as the large increase for the Phi-3-medium model is unclear but might be due to sensitivity to the provided prompts. This is left for future work.

The leakage of benchmark test sets on the internet poses a significant risk of contaminating training datasets in machine learning practices. As a result, models may appear to perform exceptionally well on tests not because they genuinely learned the underlying patterns and generalized well, but because they were indirectly trained on the test data. This compromises the integrity of performance evaluations and could lead to misleading conclusions about a model’s effectiveness when deployed in real-world scenarios. This study demonstrated a varying gap in accuracy for the different proprietary and open-weights LLMs on the evaluated benchmarks. However, the aim is not to call one company out with the accusation of test-set training. Instead, the focus is on increasing transparency in both benchmarking the capability in knowledge tasks for cybersecurity and developing more robust MCQA benchmarks. The training data that is used for the LLMs is disclosed, however, to prevent overfitting on commonly used benchmarks for Llama 3, they state that the pre-training data was sourced and processed by an independent team incentivized to avoid benchmarks[5]. As other teams do not highlight, this may be unique for the Llama 3 series models compared to the open-weights competitors.

The visual question-solving capabilities of different LMMs indicate that although GPT-4o and Claude 3.5 Sonnet showcased similar vision capabilities, the smaller Phi-3-vision-128k-instruct managed to achieve higher accuracy compared to the language-only CCNP exam. Also, a generally higher performance across all models on the CCNP 350-701 vision dataset is shown, likely influenced by its smaller size and possibly less complex content, contrary to CCNP language-only questions. The higher separability of the CCNA 200-301 vision benchmark, enables a better model capability judgement.

In general, the introduced CCNA and CCNP benchmarks show higher separability compared to the MMLU computer security dataset, which ensures sufficient complexity to prevent benchmark plateauing and effectively distinguish between models. These characteristics show that the CCNA and CCNP exams provide a challenging testing environment that measures the capabilities of LLMs and LMMs. The findings from this study underscore the significant role that benchmarks play in assessing the domain-specific capabilities of LLMs. By examining the performances across different models and prompting techniques, this research offers an additional data point. Practically, these insights can guide the selection and application of LLMs in cybersecurity settings.

Future plans include expanding the visual datasets to cover a broader array of topics within the field of cybersecurity and evaluating them with more open-weights models. The influence of prompting techniques and their impact on accuracy, as well as improving the interpretability of the models’ decision-

making processes is also left for future work. Moreover, it can be investigated to differentiate between knowledge-based questions and those that require analytical reasoning. While certain questions draw predominantly on the retrieval of factual data, others necessitate a more intricate, stepwise reasoning approach to derive solutions. As the domain advances, there is a shift towards prioritizing reasoning-based tasks. This shift highlights the growing imperative of constructing models capable of not only the recall of information but also the emulation of human cognitive processes, including logical analysis and sophisticated problem-solving strategies.

7 Conclusion

This study evaluates the capabilities of LLMs and LMMs in solving cybersecurity certification questions, highlighting the stronger performance of proprietary models Anthropic Claude 3.5 Sonnet and OpenAI GPT-4o in exceeding accuracy thresholds necessary for certification. However, open-weights models - especially the Llama 3.1 models - are improving performance, demonstrating their potential to bridge the gap between open-sourced and proprietary models in the near future. The introduced CCNA and CCNP benchmarks show high separability compared to the MMLU computer security dataset, making them suitable for measuring cybersecurity knowledge in language-only and language-vision tasks.

The research also reveals potential overfitting to familiar datasets such as the MMLU, underscoring the need for ongoing attention to the validity of benchmark results. The training data for the models is disclosed, but to avoid benchmark overfitting the pre-training data of models should intentionally avoid benchmark data. Additionally, this study introduces the first visual cybersecurity multiple-choice question-answering dataset. Exploring vision integration in problem-solving shows varying success, with proprietary models demonstrating significant capabilities in multi-modal contexts.

Acknowledgments. This work was supported by funding from the topic Engineering Secure Systems of the Helmholtz Association (HGF) and by KASTEL Security Research Labs (structure 46.23.02).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Appendix

Cisco 5-SHOT

""<|start_header_id|>user<|end_header_id|>

The following are multiple choice questions (with answers) about network fundamentals, network access, security fundamentals, automation and programmability.

Question: Which two options are the best reasons to use an IPV4 private IP space? (Choose two.)

- A. to enable intra-enterprise communication
- B. to implement NAT
- C. to connect applications
- D. to conserve global address space
- E. to manage routing overhead

Your response should end with \"The best answer is [the_answer_letter(s)]\" where the [the_answer_letter(s)] is/are of A, B, C, D, E, F,...

<|eot_id|><|start_header_id|>assistant<|end_header_id|>

The best answer is AD.<|eot_id|><|start_header_id|>user<|end_header_id|>

The following are multiple choice questions (with answers) about network fundamentals, network access, security fundamentals, automation and programmability.

Question: Security Group Access requires which three syslog messages to be sent to Cisco ISE? (Choose three .)

- A. IOS-7-PROXY_DROP
- B. AP-1-AUTH_PROXY_DOS_ATTACK
- C. MKA-2-MACDROP
- D. AUTHMGR-5-MACMOVE
- E. ASA-6-CONNECT_BUILT
- F. AP-1-AUTH_PROXY_FALLBACK_REQ

Your response should end with \"The best answer is [the_answer_letter(s)]\" where the [the_answer_letter(s)] is/are of A, B, C, D, E, F,...

<|eot_id|><|start_header_id|>assistant<|end_header_id|>

The best answer is BDF.<|eot_id|><|start_header_id|>user<|end_header_id|>

The following are multiple choice questions (with answers) about network fundamentals, network access, security fundamentals, automation and programmability.

Question: Which two authentication stores are supported to design a wireless network using PEAP EAP-MSCHAPv2 as the authentication method? (Choose two.)

- A. Microsoft Active Directory
- B. ACS
- C. LDAP
- D. RSA Secure-ID
- E. Certificate Server

Your response should end with \"The best answer is [the_answer_letter(s)]\" where the [the_answer_letter(s)] is/are of A, B, C, D, E, F,...

<|eot_id|><|start_header_id|>assistant<|end_header_id|>

The best answer is AB.<|eot_id|><|start_header_id|>user<|end_header_id|>

The following are multiple choice questions (with answers) about network fundamentals, network access, security fundamentals, automation and programmability.

Question: The corporate security policy requires multiple elements to be matched in an authorization policy. Which elements can be combined to meet the requirement?

- A. Device registration status and device activation status
- B. Network access device and time condition
- C. User credentials and server certificate
- D. Built-in profile and custom profile

Your response should end with \"The best answer is [the_answer_letter(s)]\" where the [the_answer_letter(s)] is/are of A, B, C, D, E, F,...

<|eot_id|><|start_header_id|>assistant<|end_header_id|>

The best answer is B.<|eot_id|><|start_header_id|>user<|end_header_id|>

The following are multiple choice questions (with answers) about network fundamentals, network access, security fundamentals, automation and programmability.

Question: Which three posture states can be used for authorization rules? (Choose three.)

- A. unknown

B. known
 C. noncompliant
 D. quarantined
 E. compliant
 F. no access
 G. limited

Your response should end with `\`"The best answer is [
 the_answer_letter(s)]\`" where the [the_answer_letter(s
)] is/are of A, B, C, D, E, F,...

`<|eot_id|><|start_header_id|>assistant<|end_header_id|>`

The best answer is ACE.`<|eot_id|><|start_header_id|>user
 <|end_header_id|>`

The following are multiple choice questions (with answers
) about network fundamentals, network access, security
 fundamentals, automation and programmability.

Question: {Exam_Question}
 {Exam_Choices}

Your response should end with `\`"The best answer is [
 the_answer_letter(s)]\`" where the [the_answer_letter(s
)] is/are of A, B, C, D, E, F,.... `<|eot_id|>""`

Cisco 0-SHOT COT Template

""The following are multiple choice questions (with
 answers), choose the best answer.

Question: {Exam_Question}
 {Exam_Choices}

– For simple problems:

Directly provide the answer with minimal explanation.

– For complex problems:

Use this step-by-step format:

Step 1: [Concise description]
 [Brief explanation]

Step 2: [Concise description]
 [Brief explanation]

Regardless of the approach, always conclude with:

The best answer is [the_answer_letter(s)].

where the [the_answer_letter(s)] is one or multiple of A,
 B, C, D, E, F...

Let's think step by step.""

References

1. Current Exam List, <https://www.cisco.com/c/en/us/training-events/training-certifications/exams/current-list.html>
2. Free Exam Prep By IT Professionals | ExamTopics, <https://www.examttopics.com/>
3. Hello GPT-4o, <https://openai.com/index/hello-gpt-4o/>
4. Introducing Claude 3.5 Sonnet, <https://www.anthropic.com/news/claude-3-5-sonnet>
5. The Llama 3 Herd of Models | Research - AI at Meta, <https://ai.meta.com/research/publications/the-llama-3-herd-of-models/>
6. Abdin, M., other: Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. <https://doi.org/10.48550/arXiv.2404.14219>
7. Balepur, N., Ravichander, A., Rudinger, R.: Artifacts or Abduction: How Do LLMs Answer Multiple-Choice Questions Without the Question? <https://doi.org/https://doi.org/10.48550/arXiv.2402.12483>
8. Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q.V., Xu, Y., Fung, P.: A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. <https://doi.org/10.48550/arXiv.2302.04023>
9. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., Raffel, C.: Extracting Training Data from Large Language Models. <https://doi.org/10.48550/arXiv.2012.07805>
10. Chen, Xi et al.: PaLI-X: On Scaling up a Multilingual Vision and Language Model. <https://doi.org/10.48550/arXiv.2305.18565>
11. Chiang, W.L., Zheng, L., Sheng, Y., Angelopoulos, A.N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J.E., Stoica, I.: Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. <https://doi.org/https://doi.org/10.48550/arXiv.2403.04132>
12. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring Massive Multitask Language Understanding. <https://doi.org/10.48550/arXiv.2009.03300>
13. Li, F., Liang, K., Lin, Z., Katsikas, S.K. (eds.): Security and Privacy in Communication Networks: 18th EAI International Conference, SecureComm 2022, Virtual Event, October 2022, Proceedings, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 462. Springer Nature Switzerland; Imprint Springer, 1st ed. 2023 edn. <https://doi.org/10.1007/978-3-031-25538-0>
14. Li, G., Li, Y., Guannan, W., Yang, H., Yu, Y.: SecEval: A comprehensive benchmark for evaluating cybersecurity knowledge of foundation models, <https://github.com/XuanwuAI/SecEval>
15. Li, T., Chiang, W.L., Frick, E., Dunlap, L., Wu, T., Zhu, B., Gonzalez, J.E., Stoica, I.: From Crowdsourced Data to High-Quality Benchmarks: Arena-Hard and BenchBuilder Pipeline. <https://doi.org/10.48550/arXiv.2406.11939>
16. Li, Nathaniel et al.: The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning. <https://doi.org/10.48550/arXiv.2403.03218>
17. Liang, Percy et al.: Holistic Evaluation of Language Models <https://openreview.net/forum?id=i04LZibEqW>

18. Liu, Z.: SecQA: A Concise Question-Answering Dataset for Evaluating Large Language Models in Computer Security. <https://doi.org/10.48550/arXiv.2312.15838>
19. Motlagh, F.N., Hajizadeh, M., Majd, M., Najafi, P., Cheng, F., Meinel, C.: Large Language Models in Cybersecurity: State-of-the-Art, <http://arxiv.org/abs/2402.00891>
20. OpenAI: GPT-4 Technical Report, <https://arxiv.org/pdf/2303.08774.pdf>
21. Sainz, Oscar et al.: NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark. <https://doi.org/10.48550/arXiv.2310.18018>
22. Tann, W., Liu, Y., Sim, J.H., Seah, C.M., Chang, E.C.: Using Large Language Models for Cybersecurity Capture-The-Flag Challenges and Certification Questions, <https://arxiv.org/pdf/2308.10443.pdf>
23. Tihanyi, N., Ferrag, M.A., Jain, R., Bisztray, T., Debbah, M.: CyberMetric: A Benchmark Dataset based on Retrieval-Augmented Generation for Evaluating LLMs in Cybersecurity Knowledge. <https://doi.org/10.48550/arXiv.2402.07688>
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need, <https://arxiv.org/pdf/1706.03762.pdf>
25. Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., Chen, W.: MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. <https://doi.org/10.48550/arXiv.2406.01574>
26. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. <https://doi.org/10.48550/arXiv.2201.11903>
27. Xia, C.S., Wei, Y., Zhang, L.: Practical Program Repair in the Era of Large Pre-trained Language Models. <https://doi.org/10.48550/arXiv.2210.14179>
28. Xu, H., Wang, S., Li, N., Wang, K., Zhao, Y., Chen, K., Yu, T., Liu, Y., Wang, H.: Large Language Models for Cyber Security: A Systematic Literature Review. <https://doi.org/10.48550/arXiv.2405.04760>
29. Yue, Xiang et al.: MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. <https://doi.org/10.48550/arXiv.2311.16502>
30. Zhang, J., Bu, H., Wen, H., Chen, Y., Li, L., Zhu, H.: When LLMs Meet Cybersecurity: A Systematic Literature Review. <https://doi.org/10.48550/arXiv.2405.03644>