The Digital Cybersecurity Expert: How Far Have We Come?

Dawei Wang[†], Geng Zhou[†], Xianglong Li[†], Yu Bai[†], Li Chen*[†], Ting Qin[†], Jian Sun[†], and Dan Li[‡]
[†]*Zhongguancun Laboratory*, [‡]*Tsinghua University*
{*wangdw, zhougeng, lixl, baiyu, chenli, qingting, sunjian*}*@zgclab.edu.cn, tolidan@tsinghua.edu.cn*

IEEE S&P 2025

Contents

1 Background

2 Motivation

3 Related Work

4 Contributions

5 Method

6 Experiments

7 Discussion

8 Conclusion

1 Background

- With the rapid development of Large Language Models (LLMs), the concept of "digital cybersecurity experts" has gradually gained attention. Companies such as Microsoft and Google have launched related tools (e.g., Copilot for Security, Gemini in Security).

- However, a key question remains: How far are current LLMs from becoming true digital cybersecurity experts? The answer to this question is crucial for understanding the capabilities and limitations of LLMs in this field and promoting their effective deployment.

1 Background

Existing studies mainly evaluate LLMs from two aspects: performance on specific security tasks and understanding of cybersecurity knowledge. However, they have the following limitations:

- Lack of a comprehensive knowledge framework for cybersecurity experts.
- Inability to identify specific knowledge gaps of LLMs
- Mismatch between question design and knowledge mastery requirements

1 Background

- To address the aforementioned limitations, the authors have developed CSEBenchmark, a fine-grained cybersecurity evaluation framework based on cognitive science. This framework aims to comprehensively assess LLMs' cybersecurity knowledge, clarify their strengths and weaknesses, and provide guidance for the effective application of LLMs in the cybersecurity domain.

2 Motivation

- To fill the gap in the existing evaluation of LLMs' cybersecurity capabilities, by constructing a systematic knowledge framework and evaluation tools, we aim to clarify the gap between LLMs and "digital cybersecurity experts" and promote their rational deployment and optimization in this field.

3 Related Work

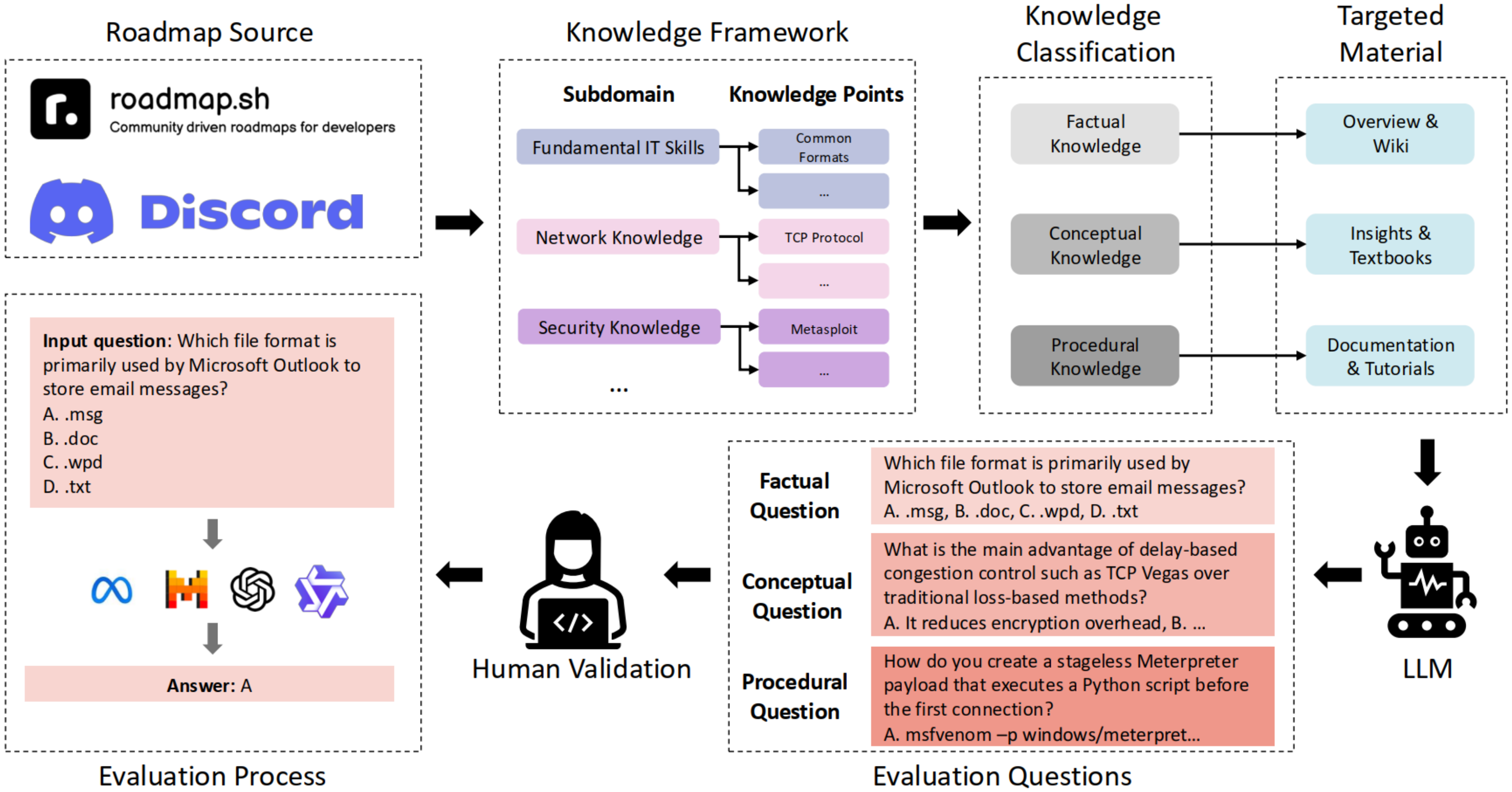
Evaluations of Large Language Models (LLMs) generally fall into two categories:

- Task-based assessments: These focus on models' performance in specific cybersecurity tasks.such as threat intelligence analysis , vulnerability management , and secure code generation. However, due to the lack of quantification of the knowledge required for tasks, such assessments struggle to clarify the reasons behind poor model performance, limiting targeted analysis.
- Knowledge-based assessments:These evaluate models' understanding of knowledge in specific cybersecurity domains through formats like multiple-choice questions.such as SecQA, CyberMetric, SecEval, and CTIBench. Nevertheless, existing studies only assess based on fragmented knowledge and lack a comprehensive modeling of the knowledge and skills required of cybersecurity experts, failing to fully answer the core question: "How far are LLMs from becoming digital cybersecurity experts?"

4 Contributions

- **New evaluation framework:**This paper proposes CSEBenchmark, the first cognitive science-based cybersecurity knowledge assessment framework, which covers 345 knowledge points across 7 subdomains and includes 11,050 multiple-choice questions. It has been made publicly available for community use.
- **New findings**： Evaluations of 12 mainstream LLMs reveal that the best-performing model achieves an accuracy of only 85.42% with existing knowledge gaps. After addressing these gaps, the correction rate for errors in relevant tasks is improved by up to 84%, and different models exhibit varying degrees of alignment with specific job roles.
- **Provide practical guidance:**Different LLMs have unique knowledge gaps, and even larger models within the same series may underperform smaller models in certain knowledge points.

5 Method



Which file format is primarily used by Microsoft Outlook to store email messages?
A. .msg, B. .doc, C. .wpd, D. .txt

What is the main advantage of delay-based congestion control such as TCP Vegas over traditional loss-based methods?
A. It reduces encryption overhead, B. ...

How do you create a stageless Meterpreter payload that executes a Python script before the first connection?
A. msfvenom –p windows/meterpret...

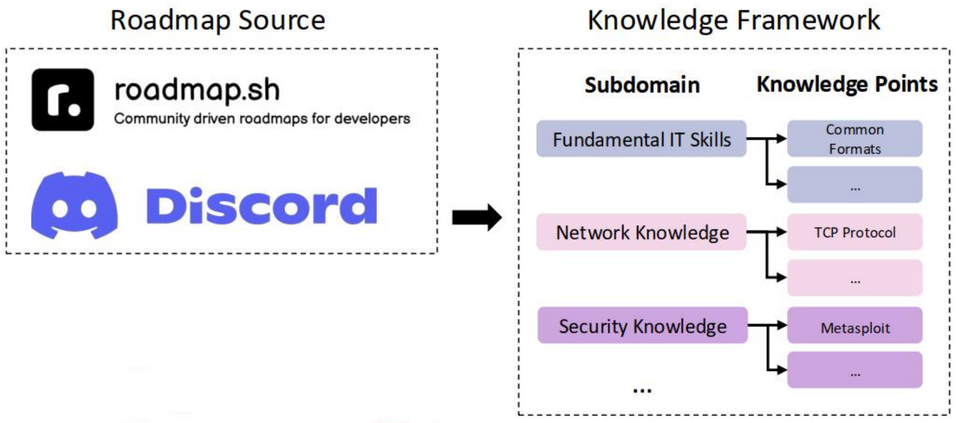
Input question: Which file format is primarily used by Microsoft Outlook to store email messages?
A. .msg
B. .doc
C. .wpd
D. .txt

Answer: A

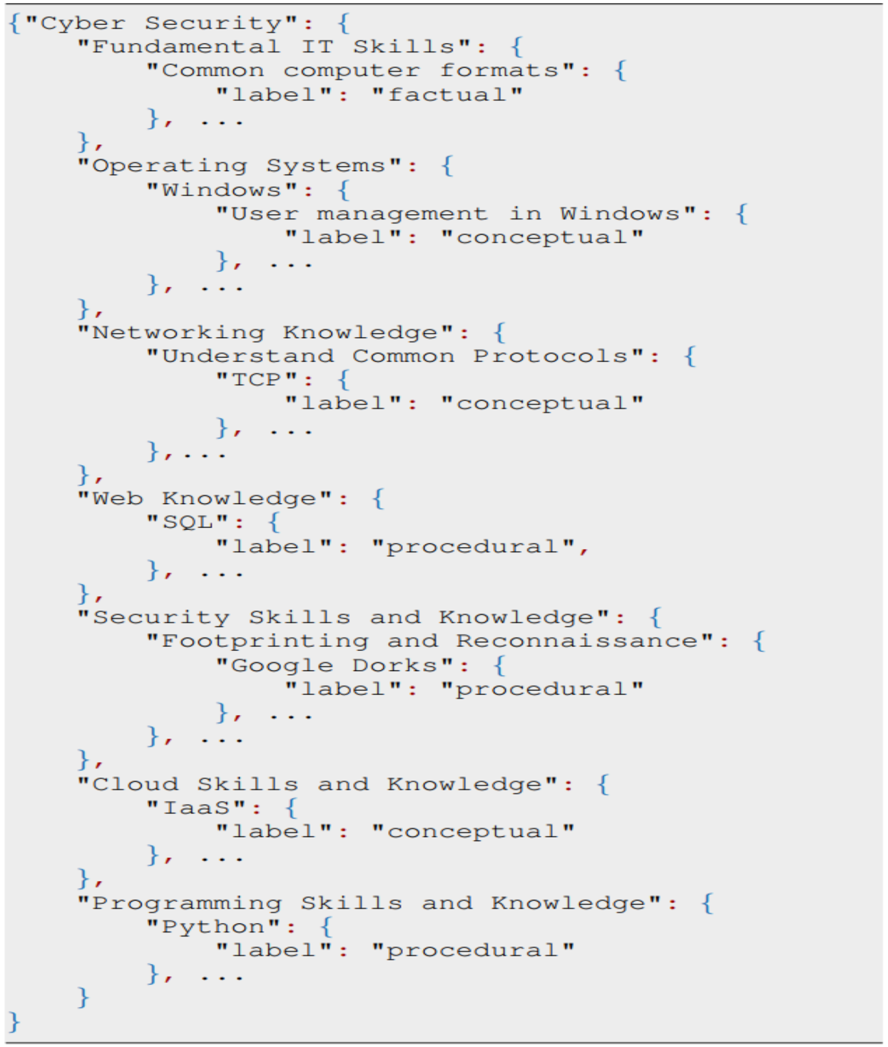
Figure 1. Overview of the construction process of CSEBenchmark.

5 Method

5.1.Knowledge Framework



- Three cybersecurity expert roadmaps as the basis: Cybersecurity Expert Roadmap and Ethical Hacker Roadmap from roadmap.sh,, and From Power Button to PWN: Computer Security Roadmap from the Hacking & Coding Discord community.
- Seven subdomains：FIS、OS、NK、WK、SSK、CSK、PSK
- This framework is organized in a hierarchical tree structure and finally forms 345 leaf nodes, representing the most specific knowledge points, enabling a fine - grained assessment of the knowledge of cybersecurity experts.

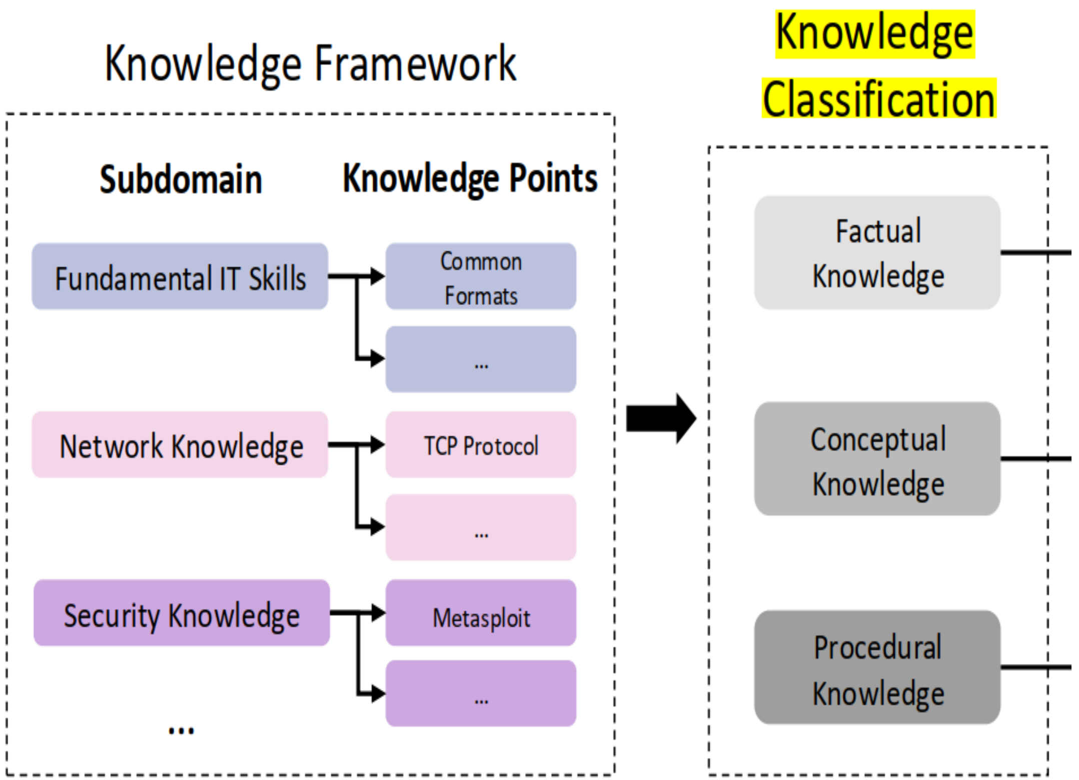


```
{"Cyber Security": {
    "Fundamental IT Skills": {
        "Common computer formats": {
            "label": "factual"
        }, ...
    },
    "Operating Systems": {
        "Windows": {
            "User management in Windows": {
                "label": "conceptual"
            }, ...
        }, ...
    },
    "Networking Knowledge": {
        "Understand Common Protocols": {
            "TCP": {
                "label": "conceptual"
            }, ...
        },...
    },
    "Web Knowledge": {
        "SQL": {
            "label": "procedural",
        }, ...
    },
    "Security Skills and Knowledge": {
        "Footprinting and Reconnaissance": {
            "Google Dorks": {
                "label": "procedural"
            }, ...
        }, ...
    },
    "Cloud Skills and Knowledge": {
        "IaaS": {
            "label": "conceptual"
        }, ...
    },
    "Programming Skills and Knowledge": {
        "Python": {
            "label": "procedural"
        }, ...
    }
}
```

Listing 1. Example of the knowledge framework.

5 Method

5.2.Knowledge Classification



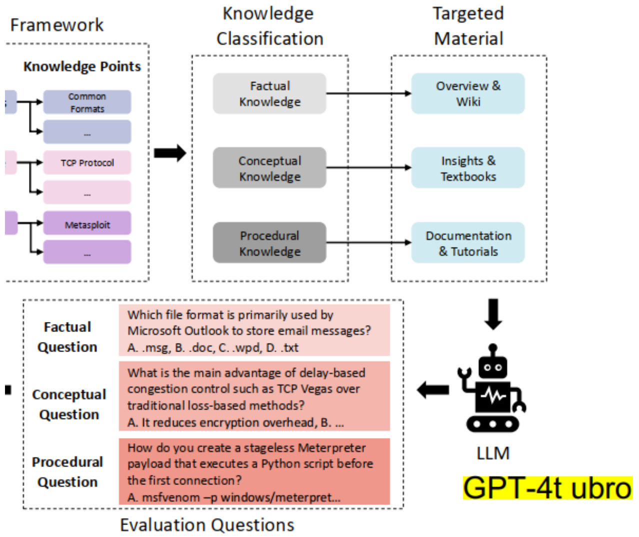
Based on the knowledge classification theory in cognitive science and combined with the characteristics of both theory and practice in the field of cybersecurity, the 345 knowledge points are divided into three categories:

- **Factual knowledge:** Specific information that needs to be memorized, with a total of 121 knowledge points.
- **Conceptual knowledge:** Theoretical knowledge that requires understanding of underlying principles, with a total of 136 knowledge points.
- **Procedural knowledge:** Skills that need hands - on practice, with a total of 88 knowledge points.

The classification is completed by two cybersecurity practitioners. In case of disagreements, a senior expert makes the final decision to ensure alignment with actual application scenarios.

5 Method

5.3.Question Generation



Which file format is primarily used by Microsoft Outlook to store email messages? A. .msg, B. .doc, C. .wpd, D. .txt

What is the main advantage of delay-based congestion control such as TCP Vegas over traditional loss-based methods? A. It reduces encryption overhead, B. ...

How do you create a stageless Meterpreter payload that executes a Python script before the first connection? A. msfvenom –p windows/meterpret...

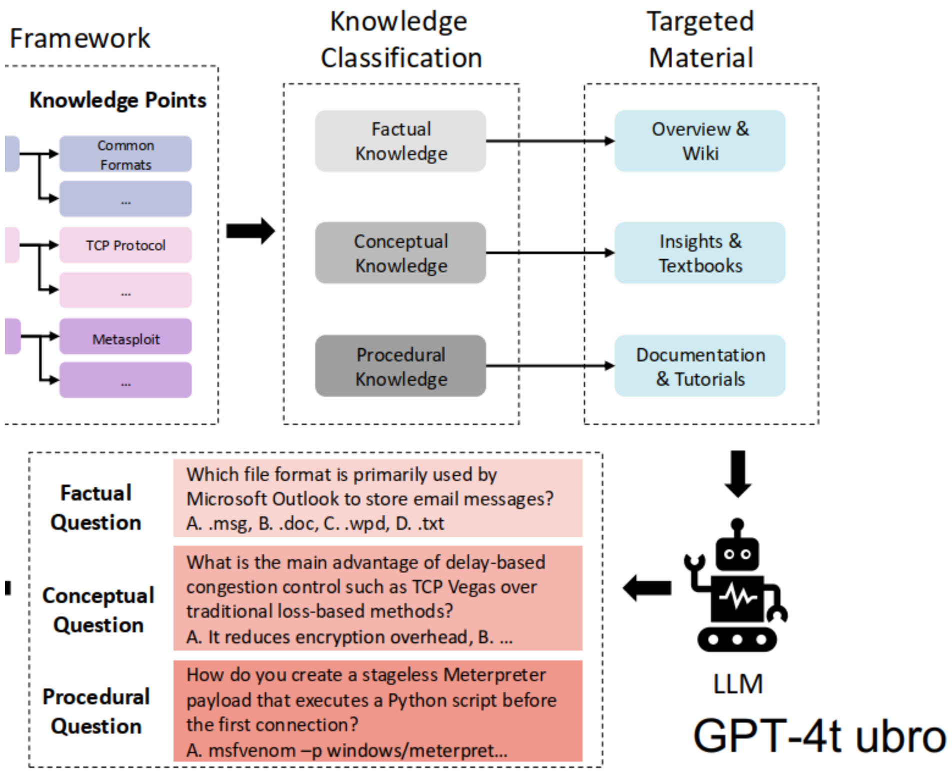
TABLE 1. DEFINITIONS FOR QUESTION GENERATION ACROSS DIFFERENT KNOWLEDGE TYPES.

Type	Definition
Factual	Multiple-choice questions focusing on factual knowledge emphasize memory and recall.
Conceptual	Multiple-choice questions focusing on conceptual knowledge emphasize understanding and applying abstract concepts
Procedural	Multiple-choice questions focusing on procedural knowledge emphasize the mastery of specific operational steps and procedural skills, particularly in the context of solving targeted problems within defined scenarios.

- The purpose of this table is to guide the GPT-4-Turbo model used for problem generation, enabling it to accurately generate problems that conform to various knowledge features
- **Material segmentation method:** The study instead segments materials according to their chapter structure, ensuring that each section retains complete contextual integrity after segmentation.
- **Adaptive generation of question quantity:** Materials of the same length may vary in information density. Materials with higher information density should generate more questions, while those with lower density should generate fewer; otherwise, it may result in redundant questions or insufficient coverage. Therefore, the study defines information density as the number of topics, and uses LLMs to first extract all topics from the materials, then generate 5 questions per topic. This achieves an adaptive match between the number of questions and the information density of the materials.

5 Method

5.3.Question Generation



Which file format is primarily used by Microsoft Outlook to store email messages? A. .msg, B. .doc, C. .wpd, D. .txt

What is the main advantage of delay-based congestion control such as TCP Vegas over traditional loss-based methods? A. It reduces encryption overhead, B. ...

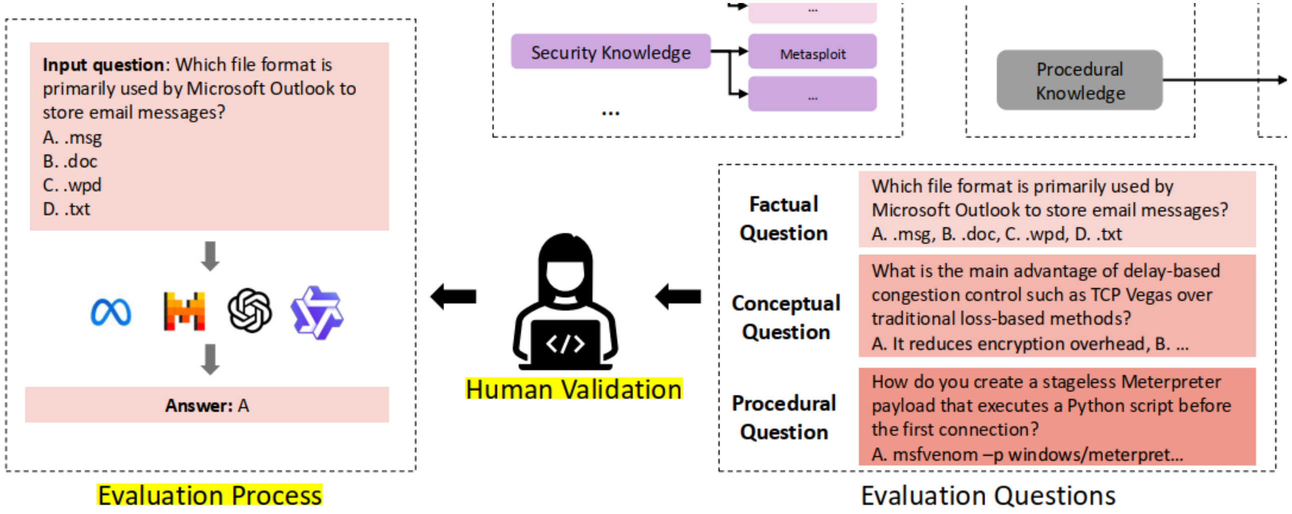
How do you create a stageless Meterpreter payload that executes a Python script before the first connection? A. msfvenom –p windows/meterpret...

Evaluation Questions

Questions were generated using GPT-4-Turbo, with each question containing 1 correct answer and 3 distractors. In total, 11,743 questions were generated for 345 knowledge points. After removing duplicates based on semantic text similarity (with a threshold of 0.85), 11,468 unique questions were obtained, with a generation cost of $234.5.

5 Method

5.4.Dataset Validation and Correction



...
Security Knowledge
Metasploit
...
...
Procedural Knowledge
Input question: Which file format is primarily used by Microsoft Outlook to store email messages?
A. .msg
B. .doc
C. .wpd
D. .txt
Answer: A
Evaluation Process
Human Validation
Factual Question
Which file format is primarily used by Microsoft Outlook to store email messages? A. .msg, B. .doc, C. .wpd, D. .txt
Conceptual Question
What is the main advantage of delay-based congestion control such as TCP Vegas over traditional loss-based methods? A. It reduces encryption overhead, B. ...
Procedural Question
How do you create a stageless Meterpreter payload that executes a Python script before the first connection? A. msfvenom –p windows/meterpret...
Evaluation Questions

Due to the hallucination issue in content generated by LLMs, a manual verification was conducted on the 11,468 questions (taking 672 person - hours). It was found that 1,726 questions had 8 types of problems, such as incorrect answers, multiple correct options, and missing context.

5 Method

5.4.Dataset Validation and Correction

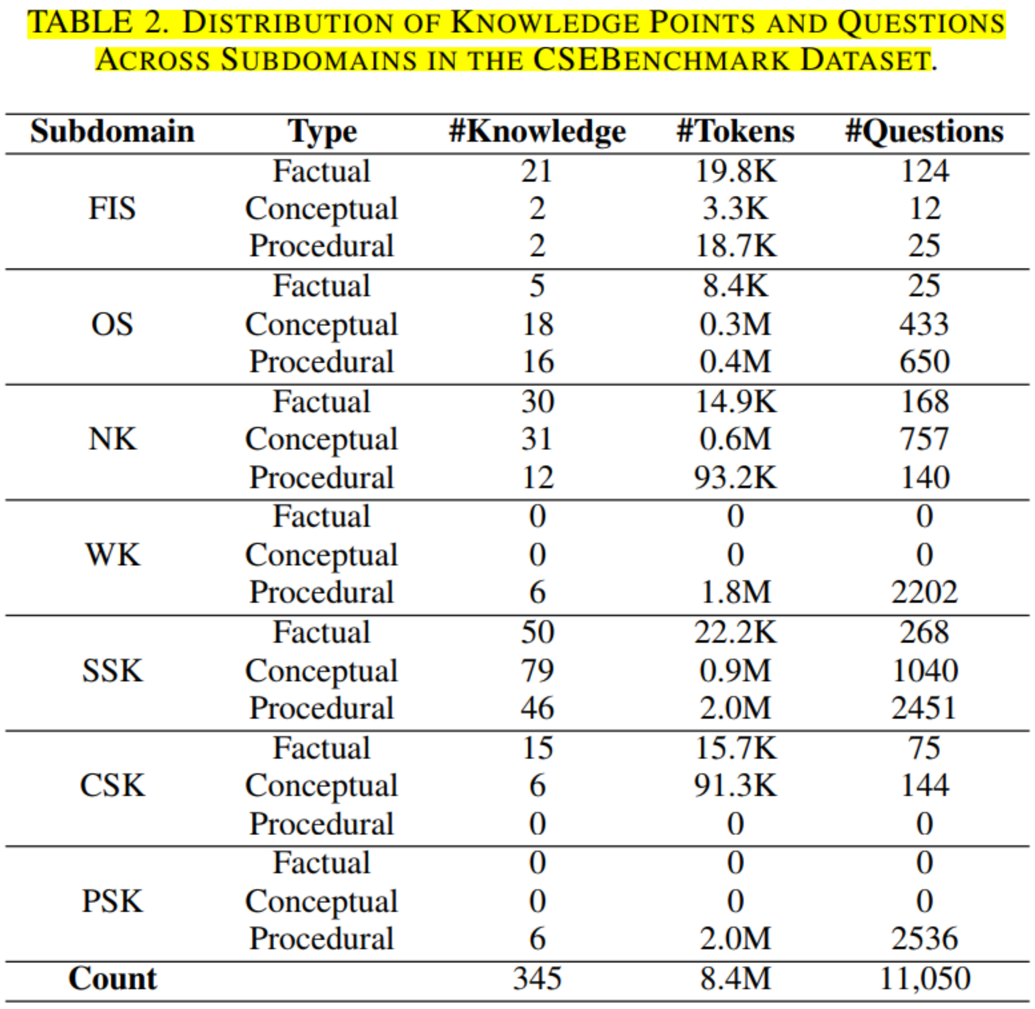


TABLE 2. DISTRIBUTION OF KNOWLEDGE POINTS AND QUESTIONS ACROSS SUBDOMAINS IN THE CSEBENCHMARK DATASET.

Subdomain	Type	#Knowledge	#Tokens	#Questions
FIS	Factual	21	19.8K	124
	Conceptual	2	3.3K	12
	Procedural	2	18.7K	25
OS	Factual	5	8.4K	25
	Conceptual	18	0.3M	433
	Procedural	16	0.4M	650
NK	Factual	30	14.9K	168
	Conceptual	31	0.6M	757
	Procedural	12	93.2K	140
WK	Factual	0	0	0
	Conceptual	0	0	0
	Procedural	6	1.8M	2202
SSK	Factual	50	22.2K	268
	Conceptual	79	0.9M	1040
	Procedural	46	2.0M	2451
CSK	Factual	15	15.7K	75
	Conceptual	6	91.3K	144
	Procedural	0	0	0
PSK	Factual	0	0	0
	Conceptual	0	0	0
	Procedural	6	2.0M	2536
Count		345	8.4M	11,050

Through manual correction (such as replacing incorrect answers and generating new distractors) and removing invalid questions, 11,050 high-quality multiple-choice questions were finally obtained.The dataset covers 7 subdomains. The distribution of questions is skewed due to differences in knowledge points designed by the roadmap and variations in corpus size. The specific distribution is as follows: 161 questions for Fundamental IT Skills (FIS), 1108 questions for Operating Systems (OS), 1065 questions for Network Knowledge (NK), 2202 questions for Web Knowledge (WK), 3759 questions for Security Skills and Knowledge (SSK), 219 questions for Cloud Skills and Knowledge (CSK), and 2536 questions for Programming Skills and Knowledge (PSK).

6 Experiments

6.1. Experiment Settings

- LLM selection and configuration

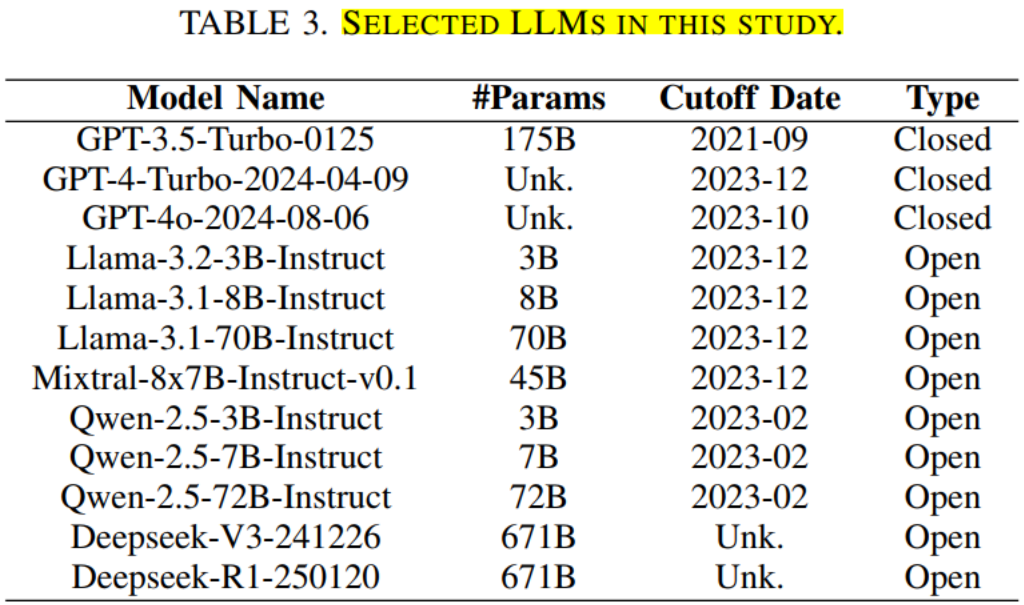


TABLE 3. Selected LLMs in this study.

Model Name	#Params	Cutoff Date	Type
GPT-3.5-Turbo-0125	175B	2021-09	Closed
GPT-4-Turbo-2024-04-09	Unk.	2023-12	Closed
GPT-4o-2024-08-06	Unk.	2023-10	Closed
Llama-3.2-3B-Instruct	3B	2023-12	Open
Llama-3.1-8B-Instruct	8B	2023-12	Open
Llama-3.1-70B-Instruct	70B	2023-12	Open
Mixtral-8x7B-Instruct-v0.1	45B	2023-12	Open
Qwen-2.5-3B-Instruct	3B	2023-02	Open
Qwen-2.5-7B-Instruct	7B	2023-02	Open
Qwen-2.5-72B-Instruct	72B	2023-02	Open
Deepseek-V3-241226	671B	Unk.	Open
Deepseek-R1-250120	671B	Unk.	Open

Twelve mainstream LLMs were selected, with parameter scales ranging from 3B to 671B. They cover closed-source models (such as the GPT series) and open-source models (such as the Llama series and Deepseek series), including mixture-of-experts models (Mixtral 8×7B) and inference models (Deepseek-R1). These models were invoked through APIs or open-source frameworks, with the temperature parameter set to 0.2 to reduce the impact of random outputs.

6 Experiments

6.1. Experiment Settings

- **Interaction methods:** Three methods were adopted, namely Zero-shot, Few-shot (5-shot), and Chain-of-Thought (CoT). The best result was taken as the upper limit of the model's knowledge.

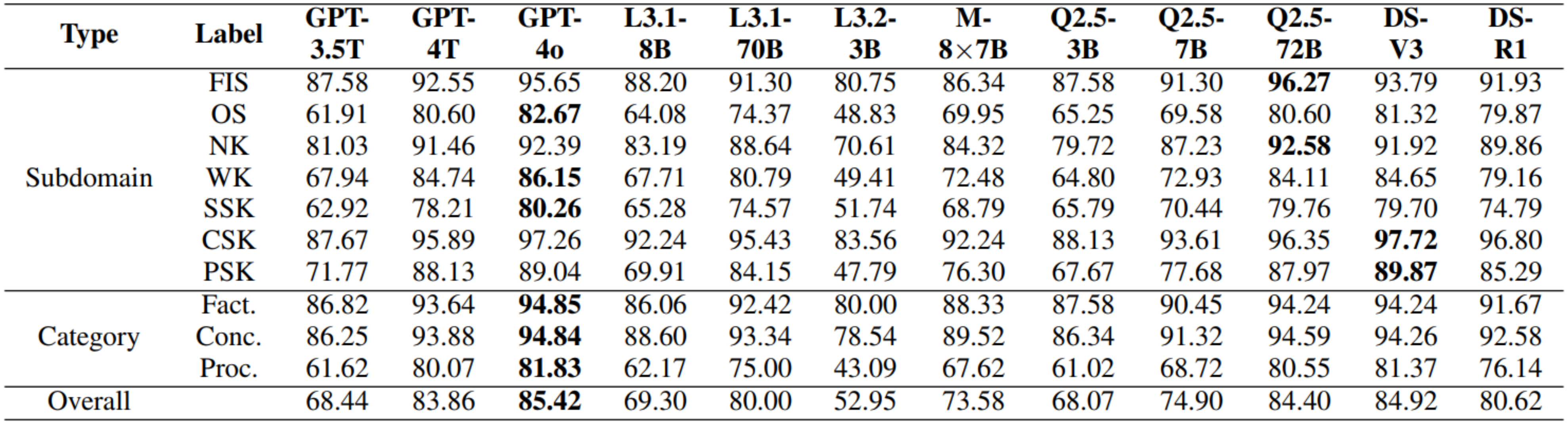
- **Measurement methods:** Each question underwent 5 independent reasoning processes, and it was considered correct only when all reasoning results were correct. The system rotated the positions of options to ensure that the model relied on understanding rather than guessing. The xFinder tool was used to extract answers, with an accuracy rate of 92.47%.

- **Evaluation metrics:** The accuracy rate of questions associated with knowledge points was used as the metric, which was divided into four levels (100%, [90%,100%), [80%,90%), <80%), representing complete mastery, near mastery, partial mastery, and weak links respectively.

6 Experiments

6.2. LLM Cybersecurity Expertise Assessment

TABLE 4. ACCURACY OF THE TESTED LLMS ACROSS SEVEN SUBDOMAINS AND THREE KNOWLEDGE CATEGORIES (ACRONYMS USED).



Type	Label	GPT-3.5T	GPT-4T	GPT-4o	L3.1-8B	L3.1-70B	L3.2-3B	M-8×7B	Q2.5-3B	Q2.5-7B	Q2.5-72B	DS-V3	DS-R1
Subdomain	FIS	87.58	92.55	95.65	88.20	91.30	80.75	86.34	87.58	91.30	**96.27**	93.79	91.93
	OS	61.91	80.60	**82.67**	64.08	74.37	48.83	69.95	65.25	69.58	80.60	81.32	79.87
	NK	81.03	91.46	92.39	83.19	88.64	70.61	84.32	79.72	87.23	**92.58**	91.92	89.86
	WK	67.94	84.74	**86.15**	67.71	80.79	49.41	72.48	64.80	72.93	84.11	84.65	79.16
	SSK	62.92	78.21	**80.26**	65.28	74.57	51.74	68.79	65.79	70.44	79.76	79.70	74.79
	CSK	87.67	95.89	97.26	92.24	95.43	83.56	92.24	88.13	93.61	96.35	**97.72**	96.80
	PSK	71.77	88.13	89.04	69.91	84.15	47.79	76.30	67.67	77.68	87.97	**89.87**	85.29
Category	Fact.	86.82	93.64	**94.85**	86.06	92.42	80.00	88.33	87.58	90.45	94.24	94.24	91.67
	Conc.	86.25	93.88	**94.84**	88.60	93.34	78.54	89.52	86.34	91.32	94.59	94.26	92.58
	Proc.	61.62	80.07	**81.83**	62.17	75.00	43.09	67.62	61.02	68.72	80.55	81.37	76.14
Overall		68.44	83.86	**85.42**	69.30	80.00	52.95	73.58	68.07	74.90	84.40	84.92	80.62

Overall performance: GPT-4o achieved the highest accuracy (85.42%), Deepseek-V3 took the lead among open-source models (84.92%), followed by Qwen-2.5-72B (84.40%); the worst-performing model, Llama-3.2-3B, had an accuracy of only 52.95%.

6 Experiments

6.2. LLM Cybersecurity Expertise Assessment

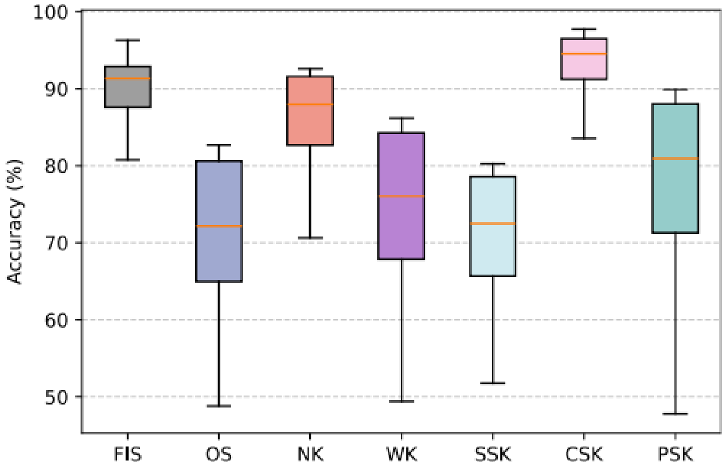


Figure 2. Accuracy distribution of LLMs across subdomains.

Subdomain differences: LLMs perform well in Fundamental IT Skills (FIS), Network Knowledge (NK), and Cloud Skills (CSK) with accuracy rates exceeding 90%, but show significant deficiencies in Operating Systems (OS), Web Knowledge (WK), Security Skills (SSK), and Programming Skills (PSK), with a median accuracy of approximately 72%.

6 Experiments

6.2. LLM Cybersecurity Expertise Assessment

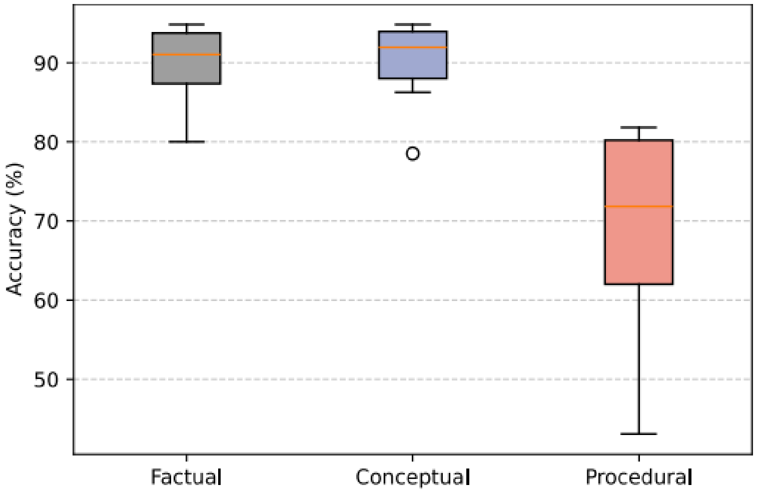


Figure 3. Accuracy distribution of LLMs across knowledge categories.

Differences in knowledge types: Factual knowledge (median accuracy 92%) and conceptual knowledge (92%) are well mastered, but there is a significant weakness in procedural knowledge (71.86%), especially in the use of professional tools and practical skills

6 Experiments

6.2. LLM Cybersecurity Expertise Assessment

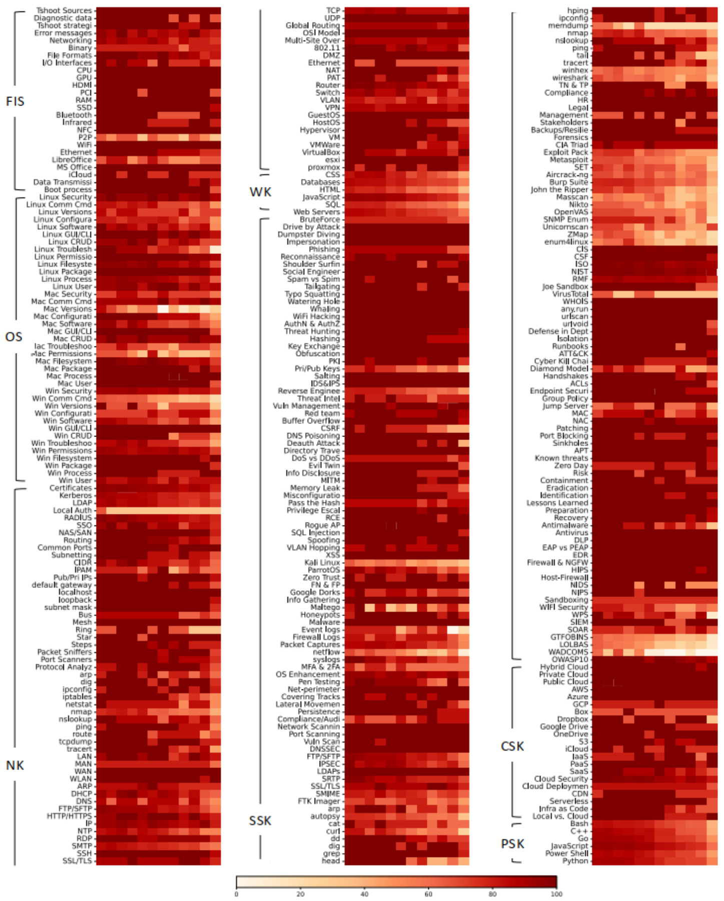


Figure 4. Heatmap of accuracy across 345 knowledge points for 12 models. The y-axis labels denote individual knowledge points, with subdomain names in parentheses for grouped items. Each section contains 12 columns representing models from left to right: GPT-4o, Deepseek-V3, Qwen-2.5-72B, GPT-4-Turbo, Deepseek-R1, Llama-3.1-70B, Qwen-2.5-7B, Mixtral-8x7B, GPT-3.5-Turbo, Llama-3.1-8B, Qwen-2.5-3B, Llama-3.2-3B.

Knowledge point coverage: 241 knowledge points have reached the expert level (100% accuracy), and 35 are close to the expert level. These mainly include factual knowledge, conceptual knowledge, and high-frequency procedural knowledge, accounting for 80% of the total.

6.3. LLM Knowledge Gap Assessment

- **Overview of knowledge gaps:** Although LLMs have reached or approached the level of cybersecurity experts in 276 knowledge points, there are still significant gaps in 69 knowledge points.

- **Partially mastered knowledge points (accuracy rate of 80% -90%):**There are a total of 40 knowledge points that LLMs have partially mastered but still need to improve, including:

(1) 11 factual knowledge (such as basic coding, differences in operating system versions, threat intelligence, etc.);

(2) 11 conceptual knowledge (such as MacOS permission management, DNS, VPN, DDoS attacks, and other core concepts);

(3) 18 procedural knowledge (such as Linux/Windows system installation and configuration, use of nmap scanning tools, log analysis, introduction to reverse engineering, etc.).

- **Weak knowledge points (accuracy below 80%):**There are a total of 29 knowledge points, and LLMs have significant room for improvement, including:

(1) 4 factual knowledge (such as P2P) and one conceptual knowledge (the difference between brute force cracking and password spraying), even if seemingly simple, the model understanding is still not solid (such as frequently answering incorrectly questions about security enhancement measures for local authentication);

(2) 24 procedural knowledge focused on the use of professional network security and forensic tools (such as Windows commands SQL).Due to their strong professionalism and insufficient representativeness in the pre training corpus, these tools are difficult for the model to effectively grasp..

6 Experiments

6.3. LLM Knowledge Gap Assessment

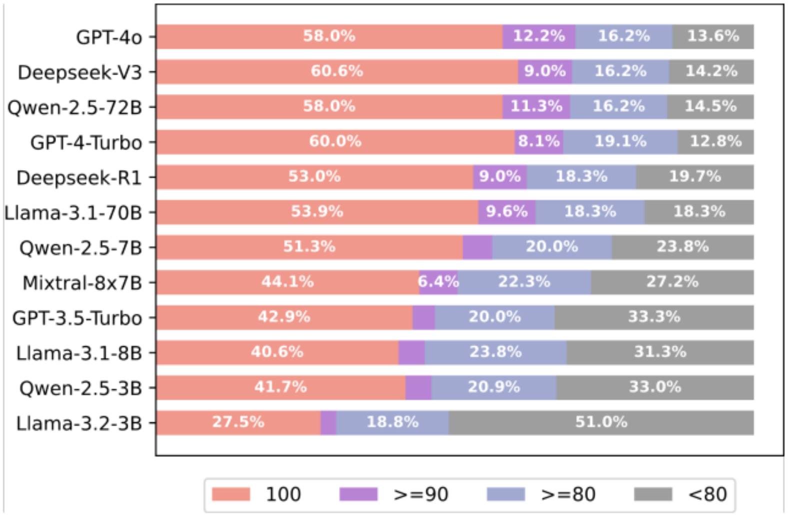


Figure 5. Proportion of knowledge points across four accuracy ranges for each LLM.

Model-specific gaps: The gaps of different LLMs are distinctive. Even large models in the same series may be inferior to smaller models in terms of certain knowledge points (for example, Llama-3.1-70B performs worse than Llama-3.1-8B in the use of tcpdump).

6 Experiments

6.4. Enhancing LLMs Through CSEBenchmark

- **Evaluate the dataset:**VuldetectBench、SecLLMHolmes、CTI-RCM

- **Experimental model:** Select the models with poor performance in the Llama series (Llama 3.1-8B, Llama 3.1-70B, Llama 3.2-3B) as the main enhancement objects; Add high-performance model GPT-4o to verify whether high-performance models can also benefit from enhancement

- **Enhancement method:** The Retrieval-Augmented Generation (RAG) technique is adopted to directionally supplement knowledge based on the knowledge gaps identified by CSEBenchmark:

(1) **Initial evaluation:** First, test the performance of the original models on the 3 datasets and record all instances of incorrect predictions.

(2)**Extraction of knowledge gaps:** Screen out knowledge points with an accuracy rate below 90% for each model from CSEBenchmark (i.e., knowledge gaps).

(3)**Construction of knowledge retrieval library:** Build a vector database for each model using Milvus, storing question-answer pairs related to knowledge gaps from CSEBenchmark; use the BGE-M3 model for embedding processing.

(4)**Directional enhanced reasoning:** When re-evaluating the previous incorrect instances, first query the vector database using task instructions, retrieve the top 5 most relevant pieces of knowledge, and incorporate them into the original prompt (with the instruction "Please use the following retrieved context to answer the question") to help the models fill in the gaps.

6 Experiments

6.4. Enhancing LLMs Through CSEBenchmark

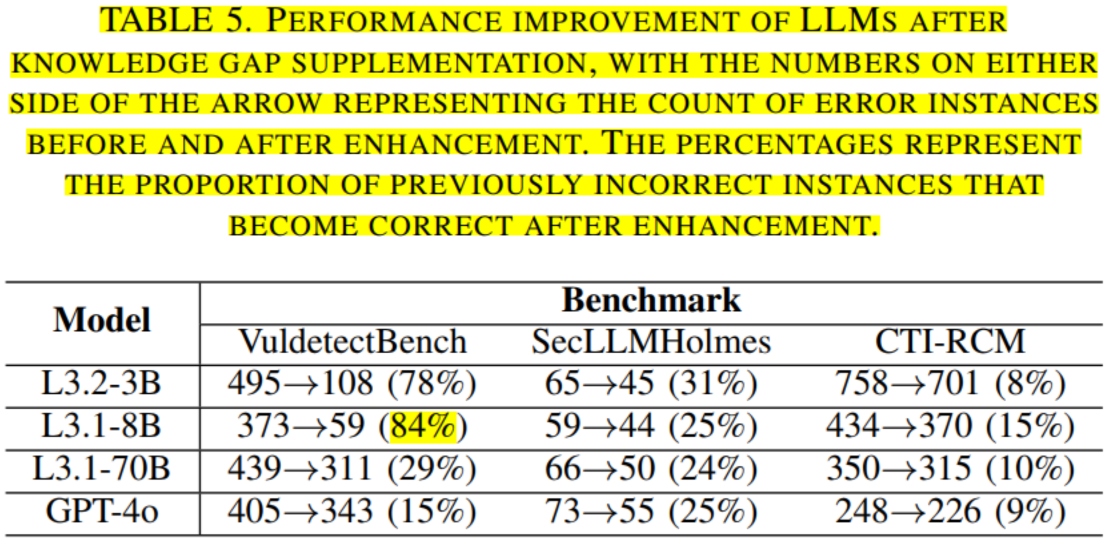


TABLE 5. Performance improvement of LLMs after knowledge gap supplementation, with the numbers on either side of the arrow representing the count of error instances before and after enhancement. The percentages represent the proportion of previously incorrect instances that become correct after enhancement.

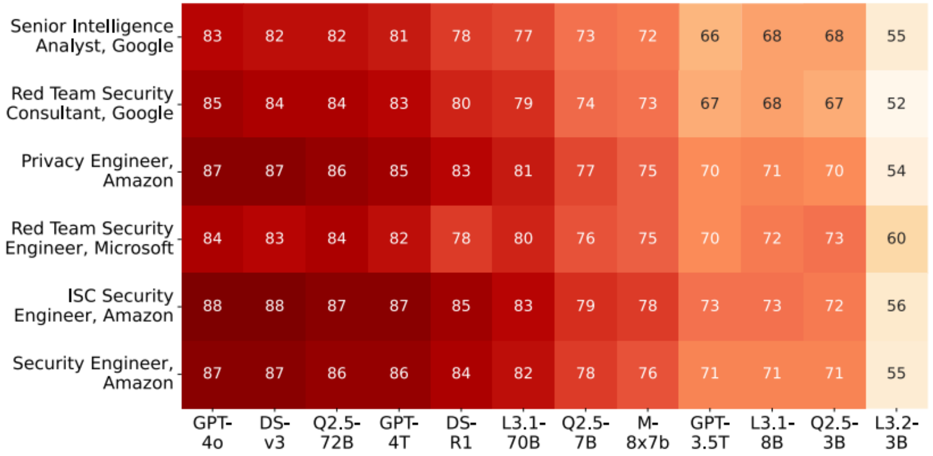
Model	Benchmark		
	VuldetectBench	SecLLMHolmes	CTI-RCM
L3.2-3B	495→108 (78%)	65→45 (31%)	758→701 (8%)
L3.1-8B	373→59 (84%)	59→44 (25%)	434→370 (15%)
L3.1-70B	439→311 (29%)	66→50 (24%)	350→315 (10%)
GPT-4o	405→343 (15%)	73→55 (25%)	248→226 (9%)

- **Experimental result:** In three network security benchmark tests (Vuldetectability Bench, SecLLMHolmes, Threat Intelligence Analysis CTI-RCM), after filling the gap, the error prediction correction rate of LLMs was as high as 84% (Llama-3.1-8B in Vuldetectability Bench), verifying the effectiveness of gap identification

- **Conclusion:**The knowledge gaps identified by CSEBenchmark can be effectively used to improve the performance of LLMs in security tasks, and even if the retrieved knowledge does not fully match the task requirements, it can still be helpful; Optimizing RAG design in the future can further enhance its effectiveness.

6 Experiments

6.5. LLM Job Role Assessment



	GPT-4o	DS-v3	Q2.5-72B	GPT-4T	DS-R1	L3.1-70B	Q2.5-7B	M-8x7b	GPT-3.5T	L3.1-8B	Q2.5-3B	L3.2-3B
Senior Intelligence Analyst, Google	83	82	82	81	78	77	73	72	66	68	68	55
Red Team Security Consultant, Google	85	84	84	83	80	79	74	73	67	68	67	52
Privacy Engineer, Amazon	87	87	86	85	83	81	77	75	70	71	70	54
Red Team Security Engineer, Microsoft	84	83	84	82	78	80	76	75	70	72	73	60
ISC Security Engineer, Amazon	88	88	87	87	85	83	79	78	73	73	72	56
Security Engineer, Amazon	87	87	86	86	84	82	78	76	71	71	71	55

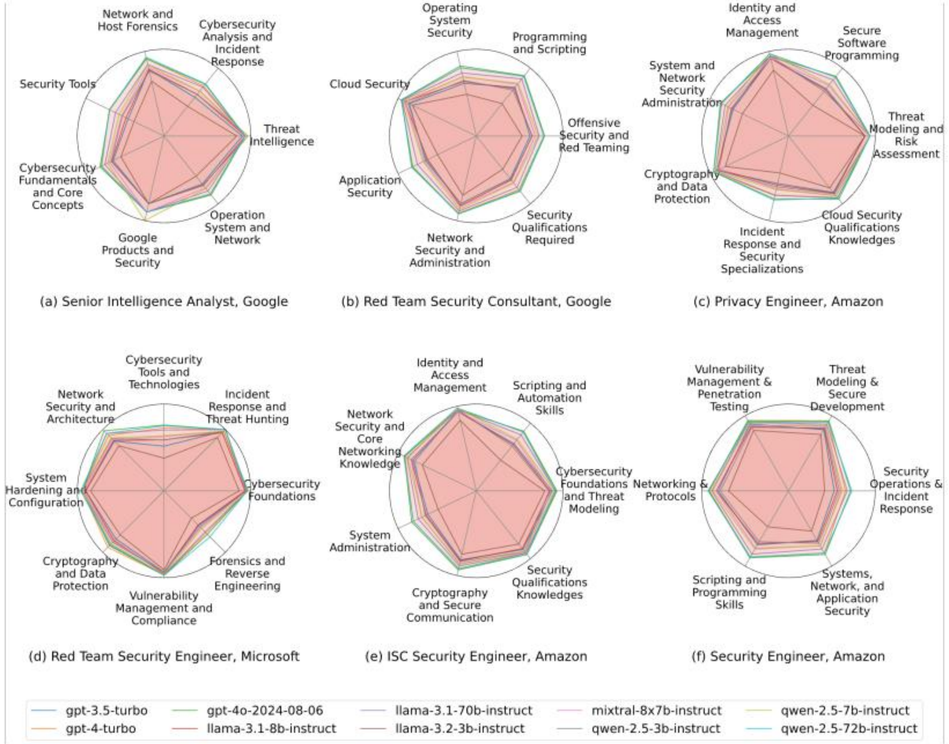
Figure 6. Heatmap of selected LLMs' match scores across six real-world cybersecurity job roles.

- **Role selection：** Six real-world cybersecurity roles (such as Google Senior Intelligence Analyst and Amazon Privacy Engineer) were evaluated. The role requirements were mapped to the knowledge points in CSEBenchmark, and matching scores were calculated.

- **Matching results:**GPT-4o achieved the highest score in roles like Google Senior Intelligence Analyst, while Deepseek-V3 stood out in roles such as Amazon Privacy Engineer. However, the highest matching score of all models was below 90%, failing to fully meet professional requirements.

6 Experiments

6.5. LLM Job Role Assessment



(a) Senior Intelligence Analyst, Google

(b) Red Team Security Consultant, Google

(c) Privacy Engineer, Amazon

(d) Red Team Security Engineer, Microsoft

(e) ISC Security Engineer, Amazon

(f) Security Engineer, Amazon

Figure 7. Radar chart showing the alignment of the selected LLMs with the requirements of six real-world cybersecurity job roles.

- **Role specificity gap:**Different roles have unique gaps, such as Google's senior intelligence analyst lacking in "network security analysis and incident response" and "security tool usage", and Microsoft's red team security engineers weak in "network security tool technology" and "forensics and reverse engineering".

7 Discussion

- **Potential Cyclic Usage and Model Bias:**Validated GPT-4-Turbo: Though it generated questions and gave answers, manual checks of 500 random questions showed questions were based on provided corpus (unaccessible during answering), ensuring no unfair cyclic use and credible results. Comparing topic extraction of models on specific knowledge points revealed consistent semantic - space topic distributions, with no topic - selection bias in GPT-4-Turbo's question - generation.

- **Limitations and Future Work:** Current knowledge framework, based on 3 public roadmaps, lacks coverage in areas like hardware security; future work includes expanding via expert interviews. Questions per knowledge point come from a single official source – more materials will be added. Only 3 prompting methods were used – advanced ones will be introduced. The xFinder tool has an 8% answer - extraction error rate – accuracy needs improvement. Due to varying model knowledge cutoffs, new materials may only be in some models' training data. The study focuses on identifying LLMs' current knowledge gaps. As LLMs evolve fast, conclusions may become outdated, requiring continuous evaluation.

8 Conclusion

- To assess large language models' (LLMs) knowledge gaps in acting as digital cybersecurity experts, this study built a cybersecurity knowledge model with 345 fine-grained knowledge points (based on cognitive science) and a benchmark dataset CSEBenchmark containing 11,050 questions.

- Evaluation of 12 mainstream LLMs shows current models' highest overall accuracy is only 85.42%, with significant gaps in procedural knowledge like professional tool use and obscure commands. Knowledge gaps vary across LLMs; even large models in the same series may underperform smaller ones on certain points.

- Filling these gaps improved error correction rates by up to 84% in two cybersecurity tasks (vulnerability detection, threat intelligence analysis) across three existing benchmarks, validating the findings.

- In summary, this study identifies current LLMs' capabilities and limitations in cybersecurity, providing key evaluation basis for developing "digital cybersecurity experts."

Thank you