# CyberMetric: A Benchmark Dataset based on Retrieval-Augmented Generation for Evaluating LLMs in Cybersecurity Knowledge

Norbert Tihanyi, Mohamed Amine Ferrag, Ridhi Jain Technology Innovation Institute (TII) Abu Dhabi, United Arab Emirates Tamas Bisztray University of Oslo Oslo, Norway Merouane Debbah Khalifa University Abu Dhabi, United Arab Emirates

Abstract—Large Language Models (LLMs) are increasingly used across various domains, from software development to cyber threat intelligence. Understanding all the different fields of cybersecurity, which includes topics such as cryptography, reverse engineering, and risk assessment, poses a challenge even for human experts. To accurately test the general knowledge of LLMs in cybersecurity, the research community needs a diverse, accurate, and up-to-date dataset. To address this gap, we present CyberMetric-80, CyberMetric-500, CyberMetric-2000, and CyberMetric-10000, which are multiple-choice Q&A benchmark datasets comprising 80, 500, 2000, and 10,000 questions respectively. By utilizing GPT-3.5 and Retrieval-Augmented Generation (RAG), we collected documents, including NIST standards, research papers, publicly accessible books, RFCs, and other publications in the cybersecurity domain, to generate questions, each with four possible answers. The results underwent several rounds of error checking and refinement. Human experts invested over 200 hours validating the questions and solutions to ensure their accuracy and relevance, and to filter out any questions unrelated to cybersecurity. We have evaluated and compared 25 state-of-the-art LLM models on the CyberMetric datasets. In addition to our primary goal of evaluating LLMs, we involved 30 human participants to solve CyberMetric-80 in a closed-book scenario. The results can serve as a reference for comparing the general cybersecurity knowledge of humans and LLMs. The findings revealed that GPT-40, GPT-4-turbo, Mixtral-8x7B-Instruct, Falcon-180B-Chat, and GEMINI-pro 1.0 were the best-performing LLMs. Additionally, the top LLMs were more accurate than humans on CyberMetric-80, although highly experienced human experts still outperformed small models such as Llama-3-8B, Phi-2 or Gemma-7b. The CyberMetric dataset is publicly available for the research community and can be downloaded from the projects' website: https://github.com/ CvberMetric.

# I. Introduction

The Industrial Revolution in the 18th century initiated a technological era, marked by significant advancements such as the steam engine, which exceeded the efficiency of human and animal labor. The mid-20th century saw the UNIVAC [5] computer in the 1950s, performing complex calculations and data processing faster than any human. By the 1970s, early chess programs began challenging experienced human players, demonstrating the evolving potential of Artificial Intelligence (AI). In 1997, IBM's Deep Blue [1] defeated Garry Kasparov [12], marking a groundbreaking moment in AI. This trend continued in 2016 with AlphaGo [20], developed by

Google DeepMind, outclassing world champion Lee Sedol in Go [3].

Large Language Models (LLMs) have revolutionized Natural Language Processing (NLP), improving automated text generation and enabling human-like interactions. These breakthroughs extend into sectors like medicine [21], finance [22], and notably, cybersecurity [6]. LLMs offer immense potential in cybersecurity, enhancing domains from threat detection [7] to policy interpretation [17].

Cybersecurity encompasses diverse topics that can require a strong mathematical foundation, programming, creative thinking, analytical skills, and managerial tasks, making it a long and challenging process for humans to master. The two research questions naturally rises:

- RQ1: Has machine intelligence already surpassed humans in answering questions across the entire breadth of cybersecurity knowledge in a closed-book test?
- RQ2: Which currently available model achieves the highest accuracy in answering questions across diverse cybersecurity domains?

To address these questions, two prerequisites must be met: 1) creating a trustworthy and validated large dataset to assess LLMs' cybersecurity knowledge; 2) selecting a subset of questions to match humans against LLMs. Although there are existing datasets and surveys focused on problem solving, coding, penetration testing, or threat intelligence [11], [23], a comprehensive dataset for testing broad cybersecurity knowledge is still lacking. This paper aims to bridge this gap by developing CyberMetric, the first comprehensive benchmark dataset for evaluating LLMs' expertise across the field of cybersecurity. Our dataset includes 9 domains: Disaster Recovery and BCP, Identity and Access Management (IAM), IoT Security, Cryptography, Wireless Security, Network Security, Cloud Security, Penetration Testing, and Compliance/Audit. The dataset comprises 10,000 questions and answers, extracted from hundreds of guidelines, standards, books, and research papers, totalling over 100,000 pages. The main contributions of this paper can be summarized as follows:

We present CyberMetric-10000 a comprehensive benchmark dataset that includes 10,000 cybersecurity-related questions designed to evaluate the understanding of

LLMs across nine distinct domains within cybersecurity. In addition, we created smaller subsets of the original datasets, named *CyberMetric-80*, *CyberMetric-500*, and *CyberMetric-2000*. The 80 and 500 datasets are fully validated by human experts. *CyberMetric* has been made accessible to the research community as a foundational metric for cybersecurity knowledge at https://github.com/CyberMetric.

2) We conducted an extensive user study with 30 human experts answering CyberMetric-80 to match their performance against LLMs. This analysis aims to understand how human expertise compares to machine intelligence in terms of cybersecurity knowledge when it comes to answering multiple-choice questions.

The rest of the paper is organized as follows: Section II overviews related literature, Section III details the methodology and dataset creation, Section IV discusses the experimental setup, objectives, and results, Section V presents our observations, Section VI discusses limitations and ethical considerations, and Section VII concludes the paper.

# II. RELATED WORK

As LLMs have evolved, the need for domain-specific benchmark datasets to test and compare their capabilities has grown significantly. These datasets are crucial for assessing LLMs in different cybersecurity domains, guiding further development and training efforts. While datasets exists for mathematical problem-solving, coding, and reasoning [2], [4], [10], large-scale datasets for assessing broad cybersecurity knowledge have not been developed prior to our research. Curating Q&A datasets has gained popularity in recent years. Khot et al. [13] introduced QASC, a multiple-choice question (MCQ) dataset focused on elementary and middle school science. Similarly, OpenBookQA [16] is an MCQ dataset for elementary science facts. Additionally, several multilingual question-answering datasets have been introduced, such as TyDI-QA, DuReader [9], and DRCD [19].

CodeApex [8] evaluates programming comprehension through multiple-choice exam questions covering conceptual understanding, commonsense reasoning, and multi-hop reasoning, as well as code generation and code correction tasks. The authors tested 12 LLMs and found that GPT-4 exhibited the highest programming capabilities, with accuracies of 69% in comprehension, 54% in generation, and 66% in correction tasks. They concluded that "Novice programmers perform similarly to GPT-4 in closed-book tests after learning, while human performance in open-book exams is significantly better than all LLMs." This trend was consistent across most models tested, including Chinese-Alpaca-13B and InternLM-Chat-7B. However, while GPT-4 outperformed humans in closed-book scenarios, humans were only slightly better in open-book exams.

In [15], Z. Liu introduced a cybersecurity dataset designed to evaluate the capabilities of LLMs, but it is entirely based on a single material: "Computer Systems Security: Planning

for Success." It includes around two hundred questions. Notably, human comparison and validation were not part of the development process.

# III. METHODOLOGY

The framework for creating the CyberMetric dataset is illustrated in Figure 1, and it comprises five crucial phases: ① Data Collection, ② Question Generation, ③ Question Postprocessing, ④ Question Validation, and ⑤ Reference Dataset Creation Phase.

# A. ① Data Collection

The questions were generated by GPT-3.5 turbo using Retrieval-Augmented Generation (RAG) [14] from widely recognized cybersecurity documents, including open standards, NIST standards, research papers, publicly available books, RFCs, and other publications in the cybersecurity domain, totalling over 100,000 pages. As documents were in PDF format, we used pdfminer¹ to extract the text, which was then segmented into chunks of 8000 token worth of text—well within the context window of GPT-3.5—to ensure it can be effectively processed. During this stage, we excluded irrelevant sections such as tables of contents, prefaces, acknowledgements, pictures, references, and appendices.

# B. 2 Question Generation

The chunks were fed to GPT-3.5-turbo to generate ten questions and corresponding multiple-choice answers for every 8,000 tokens. This approach aims to maintain a balanced representation of each publication. Creating 1,000 questions from a ten-page document might introduce redundancy to the dataset. The generated questions are reviewed by Falcon-180B, referred to as the FALCON Content Review, to identify grammatical and semantic errors and filter out irrelevant questions. By applying semantic analysis [24], questions unrelated to cybersecurity were excluded. During this stage, human validators randomly examined the questions to assess the overall quality of the results, ensuring they are relevant to cybersecurity and written in clear, correct English. GPT-3.5 generated a total of 11,000 questions, including an extra 10% margin to allow for the removal of unnecessary or irrelevant questions. The FALCON Content Review module eliminated 1.7% of the questions due to grammatical and semantic errors. Human experts then dedicated over 30 hours to eliminate an additional 2.3% of the questions. The questions that remain form the intermediate database, containing grammatically correct and cybersecurity-focused questions. However, at this stage, neither LLMs nor human experts have verified the accuracy of the solutions. Next, the 10,560 questions that were kept must undergo further solution validation to identify and remove any questions with multiple or incorrect solutions.

https://pypi.org/project/pdfminer/

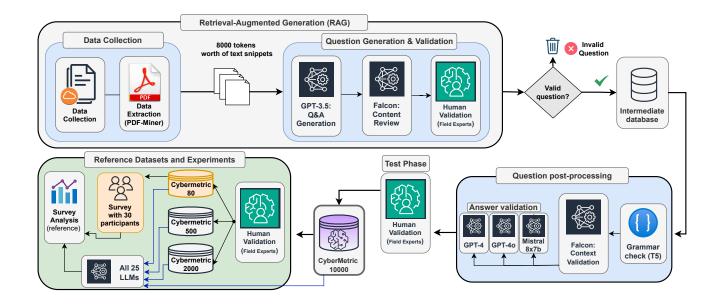


Fig. 1. Framework for AI-driven question generation methodology, incorporating human validation.

# C. 3 Question Post-processing

First, we employ the cutting-edge T5-base model from Google [18], which is trained for grammar correction. This model identified 230 questions where the English could be improved. For instance, it flagged the sentence "Which of the following elements do not apply to privacy?" for subjectverb agreement error. The correct form should be "Which of the following elements does not apply to privacy?". Next, we have again utilized FALCON-180B, but this time instead of semantic checking, we aimed to eliminate non-contextual questions that cannot be understood without external reference. Questions generated during the creation phase, such as "According to Figure 1" or "As seen in Table 6", are nonsensical without relevant content. After fixing and removing such issues, we have used GPT-4, GPT-40, and Mistral7x8b to analyze all the remaining 10470 questions, whether they think the provided answer is correct.

# D. 4 Test Phase

We have reached a critical stage in our analysis where human experts must review all questions flagged by GPT-4, GPT-4o, and Mistral7x8b. Upon thoroughly examining the questions, we discovered that many were inaccurate or imprecise. These issues can be divided into four categories:

- Multiple Correct Answers: Some questions in the dataset had more than one correct answer provided.
- 2) Time Relevant Questions: Since the questions and their answers are derived from sources up to ten years old, some information provided to GPT-3.5 may be outdated. For example, the answer to "Which web server dominates the market according to web surveys?" was

- Apache in 2015. However, as of 2024, the dominant web server is Nginx.
- 3) **Incorrect Information in the Source:** To our surprise, some selected sources contained incorrect information despite being reputable and widely followed. We verified these inaccuracies with multiple field experts.
- 4) **Missing References:** Some questions still reference material from the original document, such as "As per table 2" or "According to Figure 1," making them unanswerable without the referred content. Despite prior efforts to eliminate such questions, we found new instances like "In chapter 4", or "In the author's opinion."

After resolving the flagged issues, we further eliminated some redundant questions. This phase required more than 200 hours of expert human effort. The result forms the core of the CyberMetric dataset, containing exactly 10,000 questions. Table I shows the final distribution of questions.

TABLE I CYBERMETRIC DATASET: QUESTIONS DOMAIN DISTRIBUTION

Domain	Questions verified	Number of Questions	Creation Method
Penetration Testing / Ethical Hacking	~	1000	LLM & Human
Cryptography	~	1500	LLM & Human
Network Security / IoT Security	~	1000	LLM & Human
Information Security / Information Governance	~	1500	LLM & Human
Compliance / Disaster recovery	V	1500	LLM & Human
Cloud Security / Identity Management	~	1500	LLM & Human
NIST guidelines / RFC documents	~	2000	LLM & Human
CyberMetric	<b>V</b>	10000	LLM & Human

# E. 3 Reference Dataset Creation

First, we create *CyberMetric-80* for human participants, along with *CyberMetric-500* where all answers are accurate, and verified by a panel of experts in various cybersecurity domains, each with a minimum of 10 years of experience and holding internationally recognized certifications like CISSP, CISM, OSCP, OSEP, and ISO 27001LA.

We recruited 30 survey participants through our social networks and contacted individuals from universities, research institutions, and Big4 consulting companies to volunteer. We have prepared a Google Forms survey containing inquiries related to gender, age, years of experience, and the highest level of education and requested volunteers, spanning from beginners to experts, to complete the questionnaire without using any additional help. Next, the 25 selected LLM models filled all the four *CyberMetric* datasets. The comprehensive analysis of the accuracy of the 30 participants and the LLMs will be discussed in Section IV.

Lastly, an automated statistical analysis is carried out for the entire set of questions. This is to confirm that all answers are consistent within the A, B, C, and D options. Additionally, we have taken steps to ensure that the answers are evenly distributed among the options.

# IV. EXPERIMENTAL RESULTS

The experiments are divided into three main parts: Comparing 25 LLMs cybersecurity knowledge in all the four CyberMetric datasets, assessing human vs LLM performance on CyberMetric-80, and determine the datasets accuracy.

# A. Assessing LLMs Performance

The 25 LLM models were tested on an AWS ml.p4d.24xlarge instance with 8x NVIDIA Tesla A100 GPU cards and 382 GB of RAM, running Ubuntu 22.04. The default settings were a temperature of 1.0, top\_p at 0.9, and top\_k at 50.

The most accurate proprietary models were GPT-40 and GPT-4-turbo (see Table III). The top-performing open-source models were Mixtral-8x7B-Instruct by Mistral AI and Falcon-180B by TII. Notably, the best performing small models with 7 billion parameters were Mistral-7B-Instruct-v0.2 by Mistral AI, and Gemma-1.1-7b-it by Google. Further analysis of incorrect responses, weaknesses, and strengths will be discussed in Section V. Table III is ordered based on the column 2KQ. Note, that due to the probabilistic nature of the models' outputs, the same model may vary by up to 3-4 percentage points in subsequent runs, even for the top-performing models.

# B. Human Performance on CyberMetric-80

Altogether, 30 participants completed the CyberMetric-80 questionnaire which takes an expert around 40 to 60 minutes. They were all instructed to take a closed-book exam. We identified two participants due to their unusually high accuracy levels despite lacking cybersecurity

TABLE II
DISTRIBUTION OF ACCURACY AMONG PARTICIPANTS.

EXPERIENCED PARTICIPANTS						
#	E	D	A	G	R	
P16	10+	P.hD.	35-50	M	88.75%	
P29	10+	P.hD.	35-50	F	87.50%	
P14	1-5	MA/MSc	35-50	M	87.50%	
P24	10+	BA/BSc	35-50	M	86.25%	
P26	1-5	MA/MSc	18-35	M	86.25%	
P13	10+	P.hD.	50+	M	83.75%	
P5	10+	P.hD. 35-50 M		82.50%		
P3	5-10	MA/MSc 35-50 M 82.5		82.50%		
P1	5-10	MA/MSc	18-35	M	76.25%	
P25	5-10	BA/BSc 35-50 M 75		75.00%		
P20	1-5	Secondary	18-35	M	72.50%	
P30	5-10	BA/BSc	18-35	F	71.25%	
P12	1-5	MA/MSc	18-35	M	71.25%	
P22	1-5	P.hD.	18-35	M	70.00%	
P9	1-5	BA/BSc	18-35	M	70.00%	
P28	1-5	Secondary	18-35	M	68.75%	
P2	1-5	BA/BSc	18-35	M	58.75%	
P15	1-5	MA/MSc	18-35	M	58.75%	
P21	1-5	MA/MSc	18-35	F	53.75%	
	Mean accuracy: ≈72.24%					
	BEGINNERS					
#	E	D	A	G	R	
P19	0	BA/BSc	18-35	M	63.75%	
P23	0	MA/MSc	18-35	M	61.25%	
P8	0	BA/BSc	35-50	M	55.00%	
P27	0	BA/BSc	35-50	M	55.00%	
P6	0	MA/MSc	35-50	M	51.25%	
P18	0	MA/MSc	18-35	F	42.50%	
P11	0	MA/MSc	35-50	F	37.50%	
P4	0	MA/MSc	35-50	F	31.25%	
P7	0	BA/BSc	35-50	F	21.25%	
	Mean accuracy: ≈46.58%					
	DISQUALIFIED PARTICIPANTS					
P10	0	MA/MSc	35-50	F	87.50%	
P17	0	BA/BSc	18-35	F	83.75%	

experience. They used GPT-3.5, thus their accuracy and 92% of their incorrect responses matched those produced by GPT-3.5. These participants confirmed our suspicion and were excluded from the analysis.

The CyberMetric dataset covers a diverse range of topics, making it challenging for human experts to recall everything from memory in a closed-book scenario, without prior preparation. The highest score is  $71/80 \approx 88.75\%$  by an individual holding a PhD and certifications in CISSP, OSCP, and ISO27001LA. The median score among the 30 participants is 56, while the mean accuracy is 53.83. The mean accuracy for experienced participants (with at least 1-5 years of cybersecurity experience achieved scores ranging from 21.25% to 63.75%. Table II shows the results of all participants. Figure 2 compares humans against LLMs on CyberMetric 80. The percentages for the LLMs are derived another test run,

TABLE III 
The 25 LLMs Performance on the CyberMetric Sorted by  $2\kappa$  Q Accuracy

LLM model	Company Size	Cino	License	Accuracy			
LLIVI model		Size		80 Q	500 Q	2k Q	10k Q
GPT-40	OpenAI	N/A	Proprietary	96.25%	93.40%	91.25%	88.89%
Mixtral-8x7B-Instruct	Mistral AI	45B	Apache 2.0	92.50%	91.80%	91.10%	87.00%
GPT-4-turbo	OpenAI	N/A	Proprietary	96.25%	93.30%	91.00%	88.50%
Falcon-180B-Chat	TII	180B	Apache 2.0	90.00%	87.80%	87.10%	87.00%
GPT-3.5-turbo	OpenAI	175B	Proprietary	90.00%	87.30%	88.10%	80.30%
GEMINI-pro 1.0	Google	137B	Proprietary	90.00%	85.05%	84.00%	87.50%
Mistral-7B-Instruct-v0.2	Mistral AI	7B	Apache 2.0	78.75%	78.40%	76.40%	74.82%
Gemma-1.1-7b-it	Google	7B	Open	82.50%	75.40%	75.75%	73.32%
Meta-Llama-3-8B-Instruct	Meta	8B	Open	81.25%	76.20%	73.05%	71.25%
Flan-T5-XXL	Google	11B	Apache 2.0	81.94%	71.10%	69.00%	67.50%
Llama 2-70B	Meta	70B	Apache 2.0	75.00%	73.40%	71.60%	66.10%
Zephyr-7B-beta	HuggingFace	7B	MIT	80.94%	76.40%	72.50%	65.00%
Qwen1.5-MoE-A2.7B	Qwen	2.7B	Open	62.50%	64.60%	61.65%	60.73%
Qwen1.5-7B	Qwen	7B	Open	73.75%	60.60%	61.35%	59.79%
Qwen-7B	Qwen	7B	Open	43.75%	58.00%	55.75%	54.09%
Phi-2	Microsoft	2.7B	MIT	53.75%	48.00%	52.90%	52.13%
Llama3-ChatQA-1.5-8B	Nvidia	8B	Open	53.75%	52.80%	49.45%	49.64%
DeciLM-7B	Deci	7B	Apache 2.0	52.50%	47.20%	50.44%	50.75%
Qwen1.5-4B	Qwen	4B	Open	36.25%	41.20%	40.50%	40.29%
Genstruct-7B	NousResearch	7B	Apache 2.0	38.75%	40.60%	37.55%	36.93%
Meta-Llama-3-8B	Meta	8B	Open	38.75%	35.80%	37.00%	36.00%
Gemma-7b	Google	7B	Open	42.50%	37.20%	36.00%	34.28%
Dolly V2 12b BF16	Databricks	12B	MIT	33.75%	30.00%	28.75%	27.00%
Gemma-2b	Google	2B	Open	25.00%	23.20%	18.20%	19.18%
Phi-3-mini-4k-instruct	Microsoft	3.8B	MIT	5.00%	5.00%	4.41%	4.80%

leading to slight variations from the results shown in Table III.

# C. Dataset Accuracy

As noted, CyberMetric-80 and CyberMetric-500 have all questions and answers fully validated by human experts and can serve as reference points. If there is a significant difference in the LLM's accuracy between CyberMetric-10000 and the fully validated CyberMetric-80 and CyberMetric-500 datasets (e.g., 30%), it would suggest that the larger dataset lacks precision and some solutions might still be incorrect. As shown in TABLE III, this is not the case. While most models has less accuracy on CyberMetric-10000 compared to the smaller subsets, it is also due to the fact that proportionally there are more questions on new recommendations, guidelines, and recent research, which many models are unfamiliar with, also contributing to the drop in accuracy. We estimate that 2-3% of the questions in CyberMetric-10000 still have issues outlined in Section III-D. We encourage readers and researchers to report any inadequate or questionable answers by opening an issue on the dataset's GitHub repository.

# V. DISCUSSION - HUMAN VS MACHINE

On average, expert humans, with a mean accuracy of 72%, perform at a level comparable to Llama 2-70B. The top human performers are close to Falcon-180B, GPT-3.5, and GEMINI-pro 1.0. GPT-40, Mixtral-8x7B-Instruct, and GPT-4-turbo significantly outperform any human beings on the CyberMetric-80 dataset.

# A. Most Difficult Questions for Humans on CyberMetric-80

The analysis revealed the questions that posed the greatest challenge to participants. Table IV highlights the top 5 questions where all participant responses were at or below a 25% success rate. Questions 41 and 66 posed similar challenges to both humans and LLMs. This similarity can be attributed to these questions being relatively new recommendations from NIST and BSI.

# B. Most Difficult Questions for LLMs

Which domain poses the greatest challenges for LLMs? This question is crucial for meaningful comparisons and facilitates future research. Do humans encounter the same difficulties, or are these challenges distinct to LLMs? Here, we present fascinating findings, highlighting the two most problematic domain for LLMs in answering questions.

TABLE IV
MOST CHALLENGING QUESTIONS FOR HUMANS.
(WITH MULTIPLE CHOICE ANSWERS PROVIDED)

No.	Questions
Q41	What is the essential requirement for the security of the Discrete Logarithm Integrated Encryption Scheme (DLIES) to be maintained?
Q42	Which document provides the recommendation for elliptic curve cryptography?
Q45	What is the purpose of implementing monitoring systems?
Q61	What is the primary goal of an Information Security Governance strategy?
Q66	According to NIST guidelines, what is a recommended practice for creating a strong password?

1) Difficulty 1: Cutting Edge Research: LLMs often struggle with accurately responding to questions based on the latest research, especially when their training data is outdated. For instance, consider the question: "What is the recommended interval for changing passwords?" Historically, many guidelines mandated a 30-day password change interval. Consequently, models trained on extensive data typically endorse this interval. However, the latest NIST SP 800-63B Section 5.1.1.2 paragraph 9 advises against periodic changes, recommending password changes only if there is evidence of compromise. The official explanation, while simple, is not entirely straightforward as it challenges the established research paradigm of password behaviours that has prevailed for the past 30 years. Recent research indicates that users tend to select weaker passwords when they know they must change them soon. Since most documents do not align with these updated guidelines, it is anticipated that most LLMs will recommend users to change their passwords. Only a few models, like GPT-40 and GPT-4, can accurately respond to these factual questions, possibly due to internet access or recent training data.

Another question that advanced language models and humans often find challenging concerns the latest Federal Office for Information Security (BSI) recommendations. In January 2023, the BSI released the "Cryptographic Mechanisms: Recommendations and Key Lengths (BSI TR-02102-1)" guideline. This guideline specifies that for the Discrete Logarithm Integrated Encryption Scheme (DLIES), the prime number p should be no less than 3000 bits, and the prime number p should be a minimum of 250 bits. The choice of 3000 bits for p, which is not a power of two, and the requirement for p to be 256 bits results in what appears to be an unbalanced prime scheme. Even expert security analysts often find this question challenging, and many language models struggle to provide the correct answer when dealing with such a question.

2) Difficulty 2: Complex Computations: LLMs often face challenges in scenarios requiring precise calculations due to their lack of access to RAG or external tools. For in-

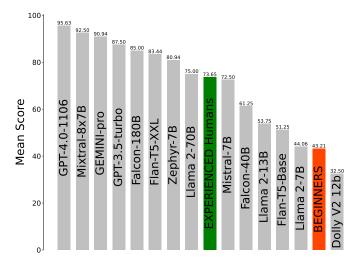


Fig. 2. Comparing Human vs LLM performance on CyberMetric-80

stance, a question like "What is the result of the bitwise XOR operation between 110101 and 101010 in binary?" can be difficult. While a model like GPT-40 can handle this easily with an embedded Python tool, models such as Mixtral-8x7B-Instruct or GPT-3.5 may fail without such tools.

Similarly, a straightforward question like "What is the CIDR notation equivalent for the subnet mask 255.255.248.0?" should be easily answerable. However, many models struggle to respond correctly without external assistance. Complex models often produce incorrect answers for precise calculations, highlighting the need for external tools to solve intricate mathematical or reasoning problems.

# VI. LIMITATIONS AND ETHICAL CONSIDERATIONS

# A. Limitations and Threats to Validity

For CyberMetric-10000, most of the questions have been validated by human experts. However, there remains the possibility of incorrect validations or the presence of irrelevant questions. Updates and corrections will be announced on the project's website: https://github.com/CyberMetric.

# B. Ethical Considerations

The documents used in this study are publicly accessible via internet searches. CyberMetric incorporates a diverse range of standard and open-access documents from the security field, including standards, research papers, NIST special publications, BSI guidelines, and RFC documents. During the human validation phase, the authors took all necessary measures to eliminate non-publicly available content from the questions or validate the source when RAG is used to generate the questions.

### VII. CONCLUSION

In this research, we introduced the CyberMetric dataset to evaluate the broad cybersecurity knowledge of LLMs. Our study focused on answering two key research questions:

- RQ1: Has machine intelligence already surpassed humans in answering questions across the entire breadth of cybersecurity knowledge in a closed-book test?
  - Answer: Yes, in our study, GPT-40 outperformed all human experts on the CyberMetric-80 test, indicating that machine intelligence has surpassed human performance in knowledge based cybersecurity questions in a closed-book scenario. Expert humans, with a mean accuracy of 72%, were on par with models like Llama 2-70B. While top human performers nearly matched the highest-performing LLMs. Beginners lagged significantly, being outperformed by 18 of the 25 models tested.
- RQ2: Which currently available model achieves the highest accuracy in answering questions across diverse cybersecurity domains?

Answer: GPT-40 and GPT-4 were identified as the top proprietary models in terms of accuracy. Among

open-source models, Mixtral-8x7B-Instruct and Falcon-180B performed best, achieving identical scores on CyberMetric-10000.

Most models still experience limitations in complex calculations and reasoning. Note that this dataset is not suitable for drawing general conclusions about machine versus human intelligence, intuition, and problem-solving skills in general, as CyberMetric exclusively focuses on cybersecurity questions and answering knowledge. Nevertheless, we are witnessing a pivotal era where machines increasingly excel beyond human abilities in many aspects.

### ACKNOWLEDGMENT

We extend our heartfelt gratitude to the volunteers in the CyberMetric-80 survey. Their dedication in completing this intricate and time-consuming 80-question test is immensely appreciated and vital for the success of this research. In an era where even support for a brief 5-minute survey is challenging to secure, the commitment shown by our participants is highly appreciated.

# REFERENCES

- [1] Murray Campbell, A Joseph Hoane Jr, and Feng-hsiung Hsu. Deep blue. *Artificial intelligence*, 134(1-2):57–83, 2002.
- [2] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. arXiv, 2021.
- [3] Tanguy Chouard. The Go Files: AI computer wraps up 4-1 victory against human champion. *Nature*, page nature.2016.19575, March 2016.
- [4] Victor Dibia, Adam Fourney, Gagan Bansal, Forough Poursabzi-Sangdeh, Han Liu, and Saleema Amershi. Aligning offline metrics and human judgments of value for code generation models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, Findings of the Association for Computational Linguistics: ACL 2023, pages 8516–8528, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [5] J. Presper Eckert, James R. Weiner, H. Frazer Welsh, and Herbert F. Mitchell. The univac system. In *Joint AIEE-IRE Computer Conference: Review of Electronic Digital Computers*, page 6–16. ACM, 1951.
- [6] Mohamed Amine Ferrag, Ammar Battah, Norbert Tihanyi, Merouane Debbah, Thierry Lestable, and Lucas C. Cordeiro. SecureFalcon: The Next Cyber Reasoning System for Cyber Security, July 2023.
- [9] Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, et al. Dureader: a chinese machine reading comprehension dataset from real-world applications. arXiv preprint arXiv:1711.05073, 2017.

- [7] Mohamed Amine Ferrag, Mthandazo Ndhlovu, Norbert Tihanyi, Lucas C. Cordeiro, Merouane Debbah, and Thierry Lestable. Revolutionizing Cyber Threat Detection with Large Language Models, June 2023.
- [8] Lingyue Fu, Huacan Chai, Shuang Luo, Kounianhua Du, Weiming Zhang, Longteng Fan, Jiayi Lei, Renting Rui, Jianghao Lin, Yuchen Fang, et al. Codeapex: A bilingual programming evaluation benchmark for large language models. arXiv preprint arXiv:2309.01940, 2023.
- [10] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.
- [11] Xinyi Hou, Yanjie Zhao, Yue Liu, Zhou Yang, Kailong Wang, Li Li, Xiapu Luo, David Lo, John Grundy, and Haoyu Wang. Large language models for software engineering: A systematic literature review. arXiv preprint arXiv:2308.10620, 2023.
- [12] Feng-hsiung Hsu. Behind Deep Blue: Building the Computer That Defeated the World Chess Champion. Princeton University Press, 2022.
- [13] Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8082–8090, 2020.
- [14] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wentau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.
- [15] Zefang Liu. SecQA: A Concise Question-Answering Dataset for Evaluating Large Language Models in Computer Security. arxiv.org/, 2023.
- [16] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. arXiv preprint arXiv:1809.02789, 2018.
- [17] Alisha Pravasi and Sanchari Das. Assessing chatgpt's efficacy in interpreting privacy policies. In Assessing ChatGPT's Efficacy in Interpreting Privacy Policies, 05 2024.
- [18] Google Research. Fine-tune a transformer model for grammar correction. https://www.vennify.ai/, 2021.
- [19] Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. Drcd: A chinese machine reading comprehension dataset. arXiv preprint arXiv:1806.00920, 2018.
- [20] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. nature, 550(7676):354–359, 2017.
- [21] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera Y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. Nature, 620(7972):172–180, August 2023.
- [22] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. BloombergGPT: A Large Language Model for Finance, March 2023.
- [23] Jie Zhang, Haoyu Bu, Hui Wen, Yu Chen, Lun Li, and Hongsong Zhu. When Ilms meet cybersecurity: A systematic literature review. arXiv preprint arXiv:2405.03644, 2024.
- [24] Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. Sentiment analysis in the era of large language models: A reality check, 2023.