Assessing Large Language Model's knowledge of threat behavior in MITRE ATT&CK

Ethan Garza
MIT CSAIL
Cambridge, USA
ezg@mit.edu

Erik Hemberg
MIT CSAIL
Cambridge, USA
hembergerik@csail.mit.edu

Stephen Moskal
MIT CSAIL
Cambridge, USA
smoskal@mit.edu

Una-May O'Reilly
MIT CSAIL
Cambridge, USA
unamay@csail.mit.edu

ABSTRACT

In the rapidly evolving field of cyber defense, acquiring expertise on threat behavior and mitigations is both time-consuming and non-trivial. This paper investigates the knowledge of threat behavior in MITRE ATT&CK exhibited by GPT-3.5, a Large Language Model (LLM). We systematically explore different input prompts to generate questions and assess the number of correct questions based on Subject Matter Expert (SME) and LLM evaluation. We analyze various prompts to elicit accurate responses to these questions from a set of LLMs. Our findings indicate that LLMs can generate questions and answers about threat behaviors and mitigations. However, GPT-3.5 may struggle to rate the quality of the generated questions. This study contributes to the understanding of LLM knowledge, capacity, and risks in the cyber security domain. It also highlights their potential applications for assessing cyber security knowledge.

CCS CONCEPTS

• **Security and privacy**; • **Computing methodologies** → **Artificial intelligence**; **Machine learning**; **Machine learning approaches**;

KEYWORDS

cyber security, threat knowledge, large language model, education

ACM Reference Format:
Ethan Garza, Erik Hemberg, Stephen Moskal, and Una-May O'Reilly. 2023. Assessing Large Language Model's knowledge of threat behavior in MITRE ATT&CK. In *ACM KDD AI4Cyber: The 3rd Workshop on Artificial Intelligence-enabled Cybersecurity Analytics at KDD'23, August 7, 2023, Long Beach, California.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Cyber defense can be improved by understanding the threat behaviors and mitigations [16, 22, 23], but this requires subject matter experts (SMEs). Acquiring cyber defense acumen is time consuming and non-trivial, e.g. see [8, 31, 41]. The cyber landscape such as enterprise networks and potential threats within, are getting larger and more complex while threats are getting more sophisticated and faster [5]. Cybersecurity occupations have been facing a persistent shortage of staff and resources [27, 41]. To address this issue there are commercial product offerings based on LLMs [12, 21], e.g. [12] use MITRE FRAMEWORKS with their security LLM, Sec-PaLM 2. However, the degree of cyber knowledge in these LLMs can be difficult to measure. One challenge to measuring cyber knowledge is the lack of an accessible and well defined assessment, e.g. similar to the standardized US Bar Exam [24].

Assessing cyber knowledge in an LLM involves: (1) Designing an assessment. We design various LLM prompts to generate correct and comprehensible questions and answers. (2) Using the assessment. We use SMEs and LLMs to answer the generated questions. In particular, we investigate: (1) Prompting methods for LLMs for the task of creating multiple choice questions regarding cyber threat knowledge contained in the MITRE ATT&CK Framework [23]. (2) Performance of LLMs on the task of answering multiple choice questions regarding cyber threat behavior.

We explore the knowledge of the Large Language Model (LLM) GPT-3.5[1] regarding threat behavior in MITRE ATT&CK [23, 36], a public knowledge base of adversary tactics and techniques based on observations. We systematically explore different prompts (inputs) to an LLM. First we use the LLM GPT-3.5 for *question making*. We generate cyber threat behavior multiple choice questions and ask subject matter experts (SMEs) to report the number of correct and comprehensible questions. Then we explore LLMs for *question taking* with different prompts to elicit the correct answer to the threat behavior questions.

Our contributions are:

- a systematic exploration of prompt complexity and context. We find that prompts that are engineered with context provide the best answer accuracy with GPT-3.5.
- indications that GPT-3.5 can be used to create multiple choice questions regarding cyber threat knowledge contained in ATT&CK. The prompting strategies to generate questions show no distinct difference in average quality when rated by SMEs. However, GPT-3.5 struggled to effectively rate the same questions. The ratings

[1]"GPT-3.5" is the API version of ChatGPT [29]

KDD '23, August 7, 2023, Long Beach, California

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

distribution generated from GPT-3.5 were distinctly different from the ratings distribution of the SMEs.

• evidence that LLMs can answer multiple choice questions regarding cyber threat knowledge with an accuracy ranging from 0.13 to 0.87. The top four LLM and prompt variants the beat non-SMEs (accuracy of 0.67) and SMEs (accuracy of 0.78).

We provide background and related work (Sec. 2), present method (Sec. 3), describe experiments and results (Sec. 4), and conclude with discussion and future work (Sec. 5).

2 BACKGROUND & RELATED WORK

We present background in assessing cyber knowledge and Large Language Models, as well as related work in this section.

Assessing Knowledge of Cyber Threat Behaviors. Joint Task Force on Cybersecurity Education [28] curricular guidelines organized knowledge units into eight broad knowledge areas: (1) data, (2) software, (3) component, (4) connection, (5) system, (6) human, (7) organization, and (8) societal. These areas need constant updates since the cyber threat landscape is constantly changing and the cyber curriculum content requires more frequent updates when compared to traditional disciplines [31]. Moreover, to align Cybersecurity in Higher Education with industry the cybersecurity graduate needs qualifications that span academic degree, professional certifications and vendor specific certification [41]. One example of knowledge, skills, and abilities for curricula in cyber defense state that common adversary tactics, techniques, and procedures in assigned area of responsibility for Computer Network Defense Analysis [2].

A common resource for finding cyber threats with analytics is to use MITRE ATT&CK [36, 39]. Furthermore, the cyber knowledge is formalized into an knowledge graph of cybersecurity countermeasures [16]. Other examples of cybersecurity ontologies have been created to support risk information gathering in cyber-physical systems [13].

Large Language Models. Language models model the probability of text directly [20]. Language models have been used for translation and classification. Task performance where accelerated with the introduction of transformer-based models, e.g BERT (Bidirectional Encoder Representations from Transformers) [10], and T5 [35]. The "Large" refers to models with at least 10B model parameters. LLMs use exploded in 2023 with OpenAI's "GPT" (Generative Pre-trained Transformer) 175B parameter models [4, 30, 33]. As of 2023, LLMs are the private access models: GPT-4 [30], PaLM [7], LLaMa [40]. Open source models are: GPT-Neo [3], GPT-J[42], and BLOOM [37]. These models were trained on corpuses of text sometime exceeding 1.4 trillion tokens[11, 34, 40].

An interaction with an LLM is often through engineering a prompt (input text) intended to generate a desired output [20]. One method is to provide the LLM with *contextual information*, i.e. attend latent concepts the LLM from pretraining data related to the prompt. Various prompt engineering strategies have been explored and are used in this work: few-shot prompting [15], chain-of-thought [44], and self-consistency [38, 43].

Assessing Large Language Model Knowledge. Assessing knowledge in LLMs is an open and non-trivial effort, e.g studies of commonsense knowledge in large language models [19]. BERT has been investigated as knowledge bases that can be queried for particular information [32]. In the "Prompting as Probing" study the GPT-3 LLM is used for knowledge base construction [1]. For the task to predict an object given subject a multi-step approach that combines a variety of prompting techniques showed that manual prompt curation is essential.

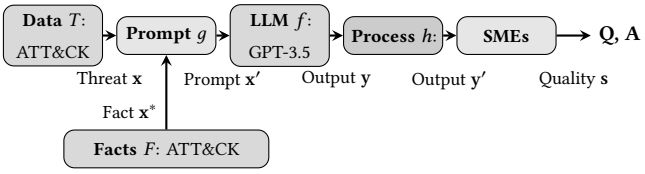


Figure 1: Workflow for making questions with answers (Q,A) task with an LLM.

LLMs have been tested as exam takers in other domains. The performance of ChatGPT was evaluated on the United States Medical Licensing Exam (USMLE) [18]. ChatGPT performed at or near the passing threshold (60%) with zero-shot prompts (out-of-the-box). Moreover, LLMs have been assessed in the domain of law. ChatGPT generated answers on four exams at the University of Minnesota Law School [6] and performed on average at the level of a C+ student, achieving a low but passing grade. In addition, Zero-shot performance of a preliminary GPT-4 on the entire Uniform Bar Examination (UBE), on the MBE, GPT-4 significantly outperforms both human test-takers and ChatGPT [17].

The work has been mostly on evaluating on existing assessments, not generating them. There has also been limited work on cybersecurity assessment generation and evaluation with LLMs. In the next section we describe how we assess LLM knowledge by making cyber threat behavior questions and evaluating them.

3 METHOD

In this section we present our method to make (Section 3.1) and take (Section 3.2) questions with an LLM[2].

We use the following notation. An input and output pair of tokens (English sentence), $(\mathbf{x}, \mathbf{y}) \in \mathcal{T}$. A parameterized model (LLM) that probabilistically outputs a sequence of tokens, $f : \mathcal{T} \times \mathbb{R} \rightarrow \mathcal{T}, \mathbf{y} = f(\mathbf{x}|\theta)$. A function (prompt) that outputs a sequence of tokens, $g : \mathcal{T} \rightarrow \mathcal{T}, \mathbf{x}' = g(\mathbf{x})$. Facts are a corpus of tokens, $F \in \mathcal{T}, \mathbf{x}^* \in F$.

3.1 Making questions with an LLM

In Figure 1 we show an overview of our method for the task of making questions with an LLM. We use a function to process and evaluate the LLM output quality, $h : \mathcal{T} \rightarrow \mathcal{T} \times \mathbb{N}_{\geq 0}, \mathbf{y}', \mathbf{s} = h(\mathbf{y})$.

We formulate questions regarding mitigations of threat behavior from ATT&CK. We design prompts two types of prompts, with and without additional contextual information. The prompt *specification* pertains to the instructions for crafting the output (question) to be similar to questions on the US Bar exam. The *usecase* provides an example of a technique and related mitigation from ATT&CK. An example prompt for making questions is shown in Figure 2.

[2]See https://github.com/ALFA-group/AI4Cyber_Wkshp_LLM_QA_Paper_2023.git.

Create multiple choice questions related to the cyber mitigation strategy below. Make a variety of quality questions, with answers, solutions, and explanations. Each question should be made that it tests the ability to apply and/or classify the knowledge tested. Each question should be expert-level (difficulty).
Content to make questions about:
Mitigation: Network Segmentation

Question: Which technique would be most useful for implementing network segmentation to control lateral movement in a network?
A. Application whitelisting
B. Data encryption
C. VLAN partitioning
D. Antivirus installation
Solution: C. VLAN partitioning
Explanation: VLAN partitioning is a technique directly associated with network segmentation. It creates isolated networks within a larger network, which can help to control lateral movement (the technique hackers use to move through a network once they have gained access) by limiting the ability to communicate between segments.

Figure 2: Example prompt to LLM (green) for making cyber mitigation multiple-choice questions. LLM responds (grey) with multiple questions, answer choices, intended solution, and an explanation.

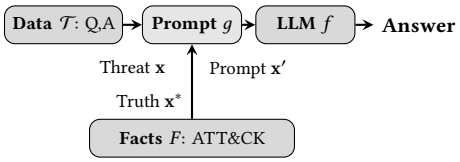


Figure 3: Workflow for taking questions task with an LLM.

The process function manipulates the LLM output and evaluates it according to: **Correctness** $\in \{Y, N\}$ Is what is asking for correct? Is the answer choice correct?, Is the question relevant to the topic? **Comprehensive** $\in \{Y, N, M\}$ Readable? Clear? Does it need reasoning? **Rating** $\in \{1, \ldots, 5\}$ a likert scale, 1 is bad, 5 is good. In addition, the SME evaluates the generated question and answer with the same categories. These evaluation questions to the SME and LLM are shown in Appendix A.1. The LLM prompt contextual information is:

- *without facts*: the baseline prompt
- *with facts* provides the known facts (contextual information) from ATT&CK

We use one prompt for all questions (without facts). We select the facts by including information from ATT&CK. We create one prompt with facts for each question (with facts).

3.2 Taking questions with an LLM

For taking questions we also design prompts with varying amounts of contextual information. (1) *without facts*: the baseline prompt (2) *with facts* provides the known facts (information) from ATT&CK

In Figure 3 we show an overview of our method for taking questions. A generated question from an LLM is passed to a prompt.

An example "question taking" prompt with facts is shown in Figure 4.

Background context related to the question:
Mitigation Title: Vulnerability Scanning ID: M1016
Description: Vulnerability scanning is used to find potentially exploitable software vulnerabilities to remediate them.

Question: What is vulnerability scanning?
1. A technique used to find potential weaknesses in a system
2. A technique used to encrypt data to prevent unauthorized access
3. A technique used to monitor network traffic for suspicious activity
4. A technique used to backup important data

Now answer the question. Select the best answer: 1, 2, 3, or 4.
Only state the answer.

Answer: 1

Figure 4: Example prompt to LLM (green) to answer questions given context (facts) from MITRE ATT&CK. LLM responds (grey) with only an answer to the question.

4 EXPERIMENTS

We present the setup (Section 4.1) and the results (Section 4.2) of using LLMs for making and taking question regarding cyber threat behavior in this section.

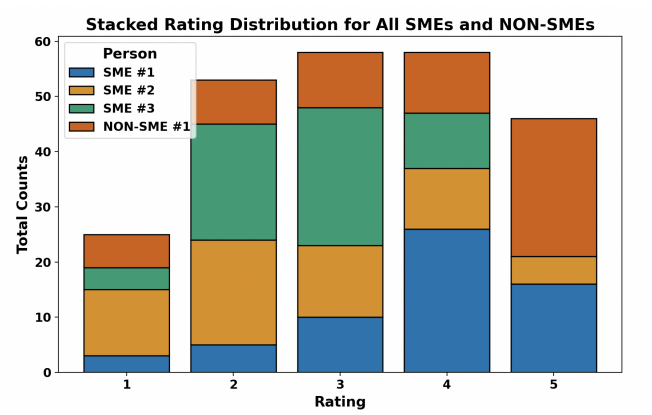
4.1 Setup

We create the LLM input data by uniformly random sampling mitigation and technique relationships from ATT&CK. We first sample 20 mitigation-technique relationships, use three different question making prompts, use GPT-3.5 to generate 4 to 5 questions per prompt, then randomly sample 1 question from each set of questions (300 total generated questions, 60 were sampled for the cyber security assessment). We ask 4 SMEs to rate and answer the questions, the SMEs had 1-6 years experience of cyber security. The SME with 0 years experience (Non-SME) is treated as a baseline evaluator. Appendix A.1 shows the detailed evaluation protocol for the SMEs. Finally, we use GPT-3.5 to rate the questions. We use LLMs with different number of parameters (GPT-3.5, DaVinci, Babage and Curie) with two different prompts to answer the 60 generated questions, see Appendix Table 4 for number of LLM parameters.

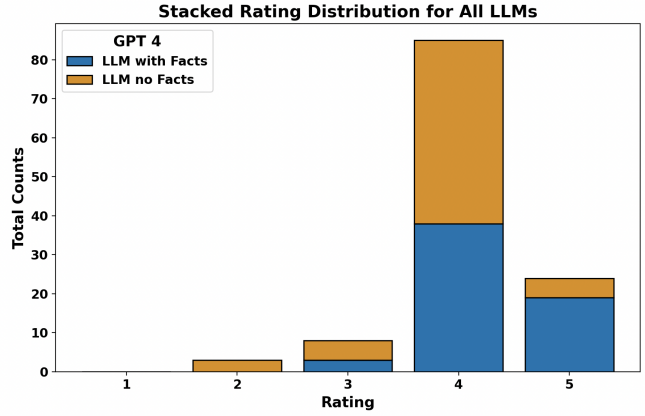
4.2 Results

In Section 4.2.1 we describe the generated questions, how SMEs, and LLMs rated them. We report on the test scores for the generated questions, both SMEs and LLMs, in Section 4.2.2.

4.2.1 Generating Questions with an LLM. Figure 5a shows stacked histograms of the question ratings (1-5) by the SMEs for the questions made by the LLM (GPT-3.5). We see that the Non-SME gave the highest average rating score for all generated questions. The SMEs had more "even" distributions for their rating scores. SME #3 did not give the highest score (5) to any question and SME #2 gave the most low scores (1). In contrast, the LLMs, Figure 5b rarely gave a score below four. The LLM using a prompt without facts for scoring (LLM no Facts) gave the majority of questions four. Using a prompt with facts (LLM with Facts) the most frequent score was four as well.



(a) SME ratings.



(b) LLM ratings

Figure 5: Stacked histograms of SME and LLM ratings of the 60 generated questions. X-axis show rating values (1-5, 5 is the highest), y-axis show frequency (counts) and the color indicate the rater (SME or LLM). "LLM" refers to GPT-3.5.

Figure 6 shows the ratings for each question for SMEs and LLMs. It does not appear that our SMEs suffered from "rater fatigue" (e.g. low quality ratings towards the end of the evaluation). Some questions have high rates from all raters (SME and LLM). On some questions all SME give low rating when the LLMs give a higher rating. Note, in Figure 5b we see that GPT 4 gives consistently high ratings.

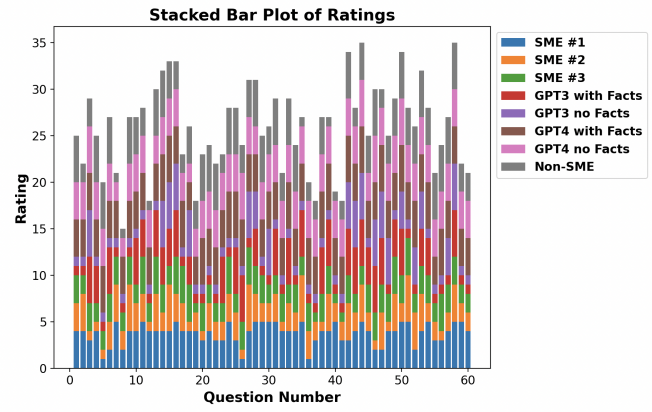


Figure 6: Stacked histograms of all. X-axis show question number (1-60), y-axis show sum of rating values for each rater and the color indicate the rater (SME or LLM).

Figure 7a gives an example question where the each SME rated the question as a 4 or a 5. The general comment was that the question was basic and easy. However because "All of the above" is available and *not* the answer, this question required more reasoning and less memorization than others which contributed to the higher rating.

Figure 7b shows an example question where all SMEs gave the same answer but an incorrect answer based on the generated answer key. The generated answer key selected (3) Network Concentrators

What are the benefits of Network Intrusion Prevention?

(1) **Reducing the likelihood of successful cyber attacks**
(2) Enhancing network performance by restricting network traffic
(3) Increasing network scalability by reducing network complexity
(4) All of the above

(a) Question all SMEs rated highly and chose the correct answer.

Which of the following is a mechanism that may be used to limit access to resources over network?

(1) Email Filters
(2) Firewall Rules
(3) **Network Concentrators**
(4) Remote Monitoring Tools

(b) Question where all SMEs gave the same incorrect answer.

Figure 7: Example questions. LLM generated solution in bold.

whereas all of the SMEs selected (2) Firewall Rules. The SMEs rated this question higher than the average rating, signaling that they were confident in their selection and rationale. This is an example where the generated question may have provided an incorrect answer as firewalls are a more compelling method to limit access to specific portions of the network. An interesting note is that when using GPT 3.5 to answer the question, it selected (3). However when provided with facts, GPT 3.5 selected (2).

More example questions are in Appendix A.2. Finally, we observed only limited difference between the different prompts used to generate these questions.

4.2.2 Taking Questions with LLMs. In Figure 8 we show the accuracy of models (LLMs and SMEs) when taking questions ordered by accuracy (See Appendix Table 1 and 3 for tabulated values). The top LLMs and SMEs have an accuracy of $\approx$ 0.8. GPT-3.5_w_facts had the highest accuracy (1st). GPT-3.5_no_facts (2nd) and DaVinci_w_facts (3rd) had higher accuracy than the SME average (4th). The non-SME (6th) had distinctly higher accuracy than Curie and Babage.

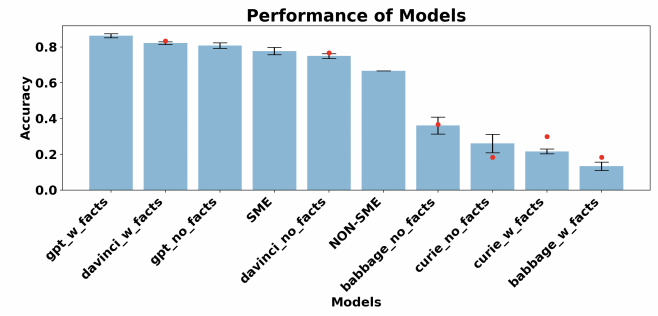


Figure 8: Bar plot with standard deviation of average accuracy of models (SMEs, Non-SME and LLMs). X-axis shows the model (descending order) and the y-axis shows accuracy. Each LLM except GPT 3.5 was run 3 times. (GPT 3.5 was run 6 times) The red dot indicates the accuracy of the LLM with a deterministic setting with zero temperature.

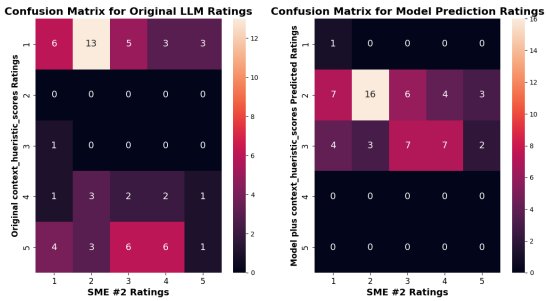


Figure 9: Confusion matrix of question quality prediction. Y-axis is LLM rating, x-axis is SME rating (SME #2). Left is LLM (GPT-4) rating. Right is LLM trained for classification of ratings.

This indicates that LLMs can have higher accuracy than SMEs. In addition, providing the LLM with the facts regarding the question mostly improved the accuracy. We had the SMEs reflect and comment on the quality of the questions and answers, for detailed examples see Appendix A.2.

5 DISCUSSION AND FUTURE WORK

In this section we discuss the results, limitations and future work. The results give some insight into the design can be improved in the future. Note, these results provide indications only, since the sample size of the results is too small to make statistically confident claims.

Figure 5 shows that the LLM did not have the ability to rate the questions with our zero-shot approach. The LLM rating prompt provided limited information regarding how to rate a question, see Appendix 10. Figure 9 demonstrates how a classifier (GPT-4) can be trained to infer the rating of the question (SME #2 in the example). Constructing a simple linear classifier of GPT-4 ratings can provide a more SME like rating. We expect that engineering the prompt by e.g. providing more rating context, could improve the rating performance further. In addition, the scoring of questions can be improved to take into account further nuances of knowledge assessment. This could create potential educational use, instead of only LLM knowledge assessment.

It is also unclear to what degree the LLM contained cyber knowledge. The experiments indicated that the generated questions were possible to read. The implications for understanding LLM capabilities and risks is that currently they can summarize and generate information, but the level of knowledge is still unknown. There is a potential to add additional sources of facts, e.g. using the BRON property graph [14]. The question creation and scoring can be improved further, e.g. with comparing different LLMs to find factual errors [9]. We experimented with few-shot prompting [15] and self-consistency. Finally, the improved performance regarding providing facts as context to the LLM are in accordance with previous work [46].

LLMs has previously been shown to perform on par with experts, since passed some US bar and medical exams [6, 17, 18]. Thus, the performance of the LLM when answering MCQ cyber threat questions was arguably expected. In addition, the LLMs has been trained for question and answering tasks.

Assessment construction is a research subject in itself and it should be taken more into account in the prompt. Cybersecurity students need theoretical concepts, hands-on experiences, and professional certifications [41]. Moreover, studies indicate that for some specialty areas, technical knowledge and skills vary considerably between jobs so the ability to teach oneself is more valuable than proficiency in KSA [2]. Finally, when specifying how to assess understanding we only focus on multiple choice questions based on ATT&CK for threat behavior similar to knowledge expected at entry level for an all-source analyst K0005 [25]. There exists additional knowledge-specified frameworks, such as the NIST cybersecurity framework [26].

There are multiple limitations to the study. The number and experience of the SMEs was small. The assessment was only multiple choice from one data source. The prompting content and ablation was restricted. In future work, we want to analyze the questions further. In addition, we can attempt to use few-shot multiple-choice questions [15]. We can consider more explicit prompting strategies, such as filtering questions via LLM feedback, "chain-of-thought" [44], and "Tree-of-thoughts" [45]. Finally, the inclusion of facts in the prompt can also be modified to include searching and processing relevant context.

Acknowledgments. We thank all the subject matter experts and students.

REFERENCES

[1] Dimitrios Alivanistos, Selene Báez Santamaría, Michael Cochez, Jan-Christoph Kalo, Emile van Krieken, and Thiviyan Thanapalasingam. 2022. Prompting as probing: Using language models for knowledge base construction. *arXiv preprint arXiv:2208.11057* (2022).

[2] Miriam E Armstrong, Keith S Jones, Akbar Siami Namin, and David C Newton. 2020. Knowledge, skills, and abilities for specialized curricula in cyber defense: Results from interviews with cyber professionals. *ACM Transactions on Computing Education (TOCE)* 20, 4 (2020), 1–25.

[3] Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An Open-Source Autoregressive Language Model. arXiv:2204.06745 [cs.CL]

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]

[5] Carlos E Budde, Anni Karinsalo, Silvia Vidor, Jarno Salonen, and Fabio Massacci. 2023. Consolidating cybersecurity in Europe: A case study on job profiles assessment. *Computers & Security* 127 (2023), 103082.

[6] Jonathan H Choi, Kristin E Hickman, Amy Monahan, and Daniel Schwarcz. 2023. Chatgpt goes to law school. *Available at SSRN* (2023).

[7] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022).

[8] CISA. 2023. Cybersecurity Training & Exercises. https://www.cisa.gov/cybersecurity-training-exercises

[9] Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. LM vs LM: Detecting Factual Errors via Cross Examination. *arXiv preprint arXiv:2305.13281* (2023).

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

[11] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. arXiv:2101.00027 [cs.CL]

[12] Google. 2023. Security with generative AI. https://cloud.google.com/security/ai

[13] Christos Grigoriadis, Adamantios Marios Berzovitis, Ioannis Stellios, and Panayiotis Kotzanikolaou. 2022. A Cybersecurity Ontology to Support Risk Information Gathering in Cyber-Physical Systems. In *Computer Security. ESORICS 2021 International Workshops: CyberICPS, SECPRE, ADIoT, SPOSE, CPS4CIP, and CDT&SECOMANE, Darmstadt, Germany, October 4–8, 2021, Revised Selected Papers.* Springer, 23–39.

[14] Erik Hemberg and Una-May O'Reilly. 2021. Using a Collated Cybersecurity Dataset for Machine Learning and Artificial Intelligence. *arXiv preprint arXiv:2108.02618* (2021).

[15] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299* (2022).

[16] Peter E Kaloroumakis and Michael J Smith. 2021. Toward a Knowledge Graph of Cybersecurity Countermeasures. (2021).

[17] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2023. Gpt-4 passes the bar exam. *Available at SSRN 4389233* (2023).

[18] Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS digital health* 2, 2 (2023), e0000198.

[19] Xiang Lorraine Li, Adhiguna Kuncoro, Jordan Hoffmann, Cyprien de Masson d'Autume, Phil Blunsom, and Aida Nematzadeh. 2022. A systematic investigation of commonsense knowledge in large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing.* 11838–11855.

[20] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.

[21] Microsoft. 2023. Microsoft Security Co-pilot. https://www.microsoft.com/en-us/security/business/ai-machine-learning/microsoft-security-copilot

[22] MITRE. 2021. MITRE Engage. https://engage.mitre.org/

[23] MITRE. 2023. ATT&CK Matrix for Enterprise. https://attack.mitre.org/ https://attack.mitre.org/.

[24] NCBE. 2023. NCBE Exams. https://www.ncbex.org/exams/

[25] NICCS. 2020. All Source Analysis. https://niccs.cisa.gov/workforce-development/nice-framework/specialty-areas/all-source-analysis

[26] NIST. 2021. NIST SP 1271. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1271.pdf

[27] Jason RC Nurse, Konstantinos Adamos, Athanasios Grammatopoulos, and Fabio Di Franco. 2021. Addressing the eu cybersecurity skills shortage and gap through higher education. *European Union Agency for Cybersecurity (ENISA) Report* (2021).

[28] Joint Task Force on Cybersecurity Education. 2018. *Cybersecurity Curricula 2017: Curriculum Guidelines for Post-Secondary Degree Programs in Cybersecurity.* Association for Computing Machinery, New York, NY, USA.

[29] OpenAI. 2023. ChatGPT. https://openai.com/product/chatgpt

[30] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[31] Daniel R Parilla and Stephanie L Wills. 2021. *Department of the Navy Cyber Workforce Leadership Development Capstone Study.* Ph. D. Dissertation.

[32] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066* (2019).

[33] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[34] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446* (2021).

[35] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 [cs.LG]

[36] Shanto Roy, Emmanouil Panaousis, Cameron Noakes, Aron Laszka, Sakshyam Panda, and George Loukas. 2023. SoK: The MITRE ATT&CK Framework in Research and Practice. *arXiv preprint arXiv:2304.07411* (2023).

[37] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100* (2022).

[38] Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Synthetic Prompting: Generating Chain-of-Thought Demonstrations for Large Language Models. *arXiv preprint arXiv:2302.00618* (2023).

[39] Blake E Strom, Joseph A Battaglia, Michael S Kemmerer, William Kupersanin, Douglas P Miller, Craig Wampler, Sean M Whitley, and Ross D Wolf. 2017. Finding cyber threats with ATT&CK-based analytics. *The MITRE Corporation, Bedford, MA, Technical Report No. MTR170202* (2017).

[40] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[41] Gelareh Towhidi and Jeannie Pridmore. 2023. Aligning Cybersecurity in Higher Education with Industry Needs. *Journal of Information Systems Education* 34, 1 (2023), 70–83.

[42] Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

[43] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171 [cs.CL]

[44] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903* (2022).

[45] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *arXiv preprint arXiv:2305.10601* (2023).

[46] Hattie Zhou, Azade Nova, Hugo Larochelle, Aaron Courville, Behnam Neyshabur, and Hanie Sedghi. 2022. Teaching algorithmic reasoning via in-context learning. *arXiv preprint arXiv:2211.09066* (2022).

Which mitigation strategy is recommended to prevent the running of commands?
(1) Escape To Host
(2) Minimal Images
(3) **Application Control**
(4) Script Blocking

Figure 12: Example question where all SMEs incorrect answers and rated the question poorly. LLM given answer in bold.

Table 1: Model Q,A Performance with Facts

Model	Mean	Std Dev
GPT	0.8639	0.0115
DaVinci	0.8222	0.0079
Curie	0.2167	0.0136
Babbage	0.1333	0.0236

Table 2: Model Q,A Performance without Facts

Table 3: Note that this has all LLMs evaluated along with SMEs and the non-SME used as a baseline

Model	Mean	Std Dev
GPT	0.8083	0.0160
DaVinci	0.7500	0.0136
Curie	0.2611	0.0515
Babbage	0.3611	0.0478
SME	0.7778	0.0208
Non-SME	0.6667	0.0000

Table 4: Number of Parameters for Each LLM

Model	Parameters (Billion)
GPT 3.5	154
DaVinci	175
Curie	6.7
Babbage	1.3

A APPENDIX

A.1 SME Evaluation Protocol

Figure 10 shows the evaluation tasks given to each of the SMEs to assess and rank each of the 60 generated questions from the assessment. SMEs were given six Y/N questions to evaluate the completeness and comprehensiveness of the generated questions. (These would be questions 1, 3, 4, 5, 6, and 7 from the rubric below) They were then asked to give a subjective ranking of the quality of each question represented in a score between 1-5. 5 being a high quality question. (Question 8 from the rubric below)

'1. Is the question factually correct? (Y/N)'
'2. Is the question, answer choices, and solution consistent? (Y/N)'
'3. Is the answer solution correct? (Y/N)'
'4. Is the question prompt relevant with itself? (Y/N)'
'5. Is the question prompt readable and make sense? (Y/N)'
'6. Is the question prompt clear? (Y/N)'
'7. Is reasoning required to answer the question (not just memorization)? (Y/N)'
'8. Using the metrics above - how would you rate the question overall? (1-5, 1 being really bad, 5 being really good)'

Figure 10: Prompt for scoring questions. All of these questions were given to GPT 3.5 for question evaluation. Question 2 was not given to SMEs

A.2 Question Examples

A.2.1 SME Highly Rated Questions. We provide examples of questions where the SMEs have given the question a high rating.

Figure 11 shows an example of a highly rated question where only our least experienced SME chose the incorrect answer (selected (1)). This question requires the knowledge and reasoning to know that certificates can be forged using the root certificate. It is slightly more nuanced question than others.

How can adversaries forge authentication certificates?
(1) By stealing certificates used for authentication
(2) **By accessing root (or subordinate) CA certificate private keys**
(3) By performing audits or scans of systems
(4) By requesting and/or manually renewing certificates from enterprise certificate authorities

Figure 11: Example question where all SMEs highly rated the question but the less experienced SME chose incorrectly. LLM given answer in bold.

A.2.2 SME Poor Performance Questions. Figure 12 shows an example of a question where the SMEs did not agree on a single answer and rated the question poorly. None of the SMEs selected (3) Application Control and had the general comment that (2), (3), or (4) could be valid answers. This gives us indication that the answers provided are not mutually exclusive and is a poorly designed question.

A.2.3 Question answering performance. The model question answering performance on the generated cyber knowledge questions is shown in Table 1 and 3. The number of parameters in the LLMs is shown in Table 4.