CyberQ: Generating Questions and Answers for Cybersecurity Education Using Knowledge Graph-Augmented LLMs

Garima Agrawal[1], Kuntal Pal[1], Yuli Deng[1], Huan Liu[1], Ying-Chih Chen[2]

[1]School of Computing and Augmented Intelligence, Arizona State University, USA
[2]Mary Lou Fulton Teachers College, Arizona State University, USA
(garima.agrawal; kkpal; ydeng19; huanliu; ychen495)@asu.edu

Abstract

Building a skilled cybersecurity workforce is paramount to building a safer digital world. However, the diverse skill set, constantly emerging vulnerabilities, and deployment of new cyber threats make learning cybersecurity challenging. Traditional education methods struggle to cope with cybersecurity's rapidly evolving landscape and keep students engaged and motivated. Different studies on students' behaviors show that an interactive mode of education by engaging through a question-answering system or dialoguing is one of the most effective learning methodologies. There is a strong need to create advanced AI-enabled education tools to promote interactive learning in cybersecurity. Unfortunately, there are no publicly available standard question-answer datasets to build such systems for students and novice learners to learn cybersecurity concepts, tools, and techniques. The education course material and online question banks are unstructured and need to be validated and updated by domain experts, which is tedious when done manually. In this paper, we propose CyberGen, a novel unification of large language models (LLMs) and knowledge graphs (KG) to generate the questions and answers for cybersecurity automatically. Augmenting the structured knowledge from knowledge graphs in prompts improves factual reasoning and reduces hallucinations in LLMs. We used the knowledge triples from cybersecurity knowledge graphs (AISecKG) to design prompts for ChatGPT and generate questions and answers using different prompting techniques. Our question-answer dataset, CyberQ, contains around 4k pairs of questions and answers. The domain expert manually evaluated the random samples for consistency and correctness. We train the generative model using the CyberQ dataset for question answering task.

Introduction

Cybersecurity education uses problem-based learning (PBL) (Dolmans and Schmidt 2010) to engage students in learning complex tools and solving real-time multi-faceted threat intelligence scenarios. It demands a progressive and adaptive learning strategy carefully designed to meet the learning needs of students at different levels like K-12, undergraduate, graduate, and professional students. Traditional cybersecurity education systems struggle to keep pace with the evolving threat landscape and understand the learning goals of users with diverse backgrounds. These systems need help keeping the course material up to date and engaging and motivating the students. Also, most of the current instruction methods in teaching cybersecurity are teacher-centered, favoring passive learning through listening and observing lectures presented by the teacher. It creates a need to design an "Active Learning" methodology (Bonwell and Sutherland 1996) that requires students to not only cognitively engage with the course material (Bonwell and Eison 1991) but also get involved and thinking about it critically rather than just passively receiving it (King 2002).

Different frameworks studied the student behaviors to quantify the impact of instruction mode. One of the popular methods is ICAP framework (Chi and Wylie 2014), which differentiates the students' overt behaviors (Menekse et al. 2013) into four modes as *Interactive, Constructive, Active and Passive*. It suggests that as the students become more engaged with the learning materials, from *passive* to *active* to *constructive* to *interactive*, their learning will increase. The *passive* mode is defined as students receiving information or lectures, whereas *active* mode is searching for information online by following the procedures provided by instructors. In the *constructive* mode, students tend to understand the concepts by self-constructing the outcomes for a new situation using AI-based visualization tools like ConceptMaps and Knowledge graphs (Agrawal et al. 2022). The *interactive* mode loosely refers to human-computer systems in a joint-dialogue. Interactive learning requires substantive dialogue rather than parallel monologues (Chen and Terada 2021). The learners engage in the highest level of learning when they interact with a device or a computer through a dialogue. The question-answering systems through dialoguing are an effective way to promote cognitive engagement and interactive learning. The ICAP framework (Chi et al. 2018) supports using AI-based intelligent tutoring systems among students, especially for engineering courses.

Cybersecurity education requires cutting-edge AI tools to engage students and keep them updated on industry trends. Unfortunately, there are no standardized datasets for building AI-driven chatbot-style teaching tools in this field. Instead, educational materials consist of unstructured text from sources like lecture notes, books, websites, and videos. Creating a question-answer database involves labor-intensive data collection efforts, as each source varies in

writing style. Additionally, ensuring the accuracy and consistency of answers demands validation from cybersecurity experts, a time-consuming and costly process. Furthermore, the rapid evolution of cybersecurity technology renders course materials obsolete quickly.

We use a novel unification of large language models (LLMs) and knowledge graphs (KGs) to construct the question-answer dataset for cybersecurity education.

Language models (LMs) such as BERT (Devlin et al. 2018), RoBERTa (Liu et al. 2019), and T5 (Raffel et al. 2020) are pre-trained models beneficial for various natural language processing (NLP) tasks, including question-answering (Su et al. 2019) and text generation (Li et al. 2022). Advanced LLMs like GPT-3, GPT-4, and ChatGPT (Yang et al. 2023), with billions of parameters, have demonstrated their potential in fields like education (Malinka et al. 2023) and recommendation systems (Liu et al. 2023). Conversely, KGs store vast amounts of information in triples (head entity, relation, tail entity) for structured knowledge representation (Ji et al. 2021). Although trained on extensive text data, LLMs sometimes produce inaccurate statements and mix facts (Shuster et al. 2021). Knowledge graphs provide precise but incomplete and non-generalizable knowledge (Abu-Salih 2021). There is a potential synergy between LLMs and KGs, with structured knowledge from KGs reducing LLM inaccuracies (Pan et al. 2023; Agrawal et al. 2023a). Dialog models like LaMDA (Thoppilan et al. 2022) have used task-specific queries to access structured knowledge through fine-tuning.

In our work, we leverage knowledge graph-enhanced LLMs generation, combining these two complementary technologies. We used the knowledge triples from AISecKG (Agrawal et al. 2023b), the cybersecurity knowledge graph (KG), to prompt the large language model, ChatGPT, and generate the cybersecurity-related questions and answers. This approach improves the factual grounding and unlocks the reasoning capability of LLMs by providing the grounding contexts using the chain of prompts (Wei et al. 2022). We use three prompting techniques consecutively to automate the generation of question-answer (QA), to be discussed later in the third section. Cybersecurity domain human experts evaluated the generated questions and answers to validate their correctness and consistency. Lastly, we train a generative model on the cybersecurity QA dataset, for open-ended question-answering.

The paper's three key contributions are as follows::

1. We present an innovative knowledge graph-enhanced LLMs generation method, **CyberGen**, that combines natural generation capability of LLMs and reliable domain knowledge from KG to create QA pairs.
2. We introduce **CyberQ**, a comprehensive question-answer dataset meticulously curated and validated by domain experts (to the best of our knowledge, first in cybersecurity education).
3. We show the application of CyberQ by training a generative model for QA tasks. It can be used to build an interactive system to educate students and beginners in cybersecurity concepts and tools.

The following sections cover related work, question-answer generation CyberGen method, CyberQ dataset evaluation, and model implementation for question-answering.

Related Work

AI in Education: AI chatbots and Question-answering systems are becoming popular in education to answer how-to questions, conduct quizzes and assessments, assist faculty, and provide administrative services (Chen, Cheng, and Heh 2021; Mzwri and Turcsányi-Szabo 2023). Several studies have shown that these chatbots have been perceived to benefit the educational system mainly in the integration of contents, quick access to educational information, allowing multiple users (Wu et al. 2020), motivation and engagement of students (Adamopoulou and Moussiades 2020), and immediate assistance (Okonkwo and Ade-Ibijola 2020). However, implementing chatbots in education faces challenges like evaluating the effectiveness and students' perception of using these tools, ensuring the accuracy of content, and maintaining and updating the AI model. The other significant bottlenecks are the availability of knowledge banks and question-answer datasets to build these models (Okonkwo and Ade-Ibijola 2021).

Knowledge Graphs in Education: State-of-the-art methods like knowledge graphs have been used in education to represent unstructured knowledge (Chen et al. 2018; Mao 2021; Fariani, Junus, and Santoso 2023; Xia and Qi 2023) and construct knowledge graph question-answering systems (KGQA) (Agrawal, Bertsekas, and Liu 2023; Chen, Wu, and Zaki 2023; Perez-Beltrachini et al. 2023). New programming questions based on knowledge graphs were generated by Chung et al. (Chung, Hsiao, and Lin 2023). Wang et al. (Wang et al. 2022) used large language models to generate educational question-answers automatically. EduChat (Dan et al. 2023) is an LLM-based chatbot to create smart education for Chinese middle and high school curricula. However, there are still challenges in generating domain-specific questions and answer data using LLMs as they tend to suffer from hallucinations. Domain experts must evaluate the generated texts for factual correctness.

Cybersecurity Education: It is a domain that can primarily benefit from AI-based tools to develop an interactive question-answering system. Active learning approaches like a search engine for scientific publications in cybersecurity (Oliveira, Sousa, and Praça 2021) and curriculum modules to teach cybersecurity (Chung 2017) have been proposed. Sayan et al. (Sayan, Hariri, and Ball 2017) built a cyber security assistant to assist security analysts in gathering resources and information about cyber attacks and defenses. A syntactic matching approach to automatically generate short factoid questions was tested on cybersecurity books and reports from 2008-2014 (Danon and Last 2017). However, these methods are not scalable. Ji et al. (Ji, Choi, and Gao 2022) developed a question-answering system for cyber threat knowledge from open-source cyber threat intelligence (OSCTI) reports.

There is limited research in developing interactive AI-based education solutions in cybersecurity for students

and novice learners. Structured flow graphs were constructed from Capture-The-Flag (CTF) procedural cybersecurity texts (Pal et al. 2021) to teach students vulnerability analysis. Knowledge graphs were proposed to guide students to work on cybersecurity projects (Deng et al. 2019; Deng, Zeng, and Huang 2021). A semi-automated approach was used for constructing knowledge graphs from unstructured cybersecurity course material to enhance the learning experience of students (Agrawal et al. 2022). In this work, a chatbot was also developed using an intent-classification SVM model. The students could query the bot to ask questions about their cybersecurity project topics. The results of the surveys and interviews to assess the students' perception of using these tools show that the students found these tools informative. However, the questions and answers dataset was manually curated and limited to course project questions. A comprehensive ontology, AISecKG (Agrawal et al. 2023b), was proposed for learning cybersecurity. The triple dataset was annotated to generate cybersecurity knowledge graphs and train a language model to identify cybersecurity-related named entities.

Downstream applications like self-learning QA systems for students need a scalable method to generate an open-ended question-answer dataset. To address this gap, in this paper, we augment the knowledge triples from cybersecurity knowledge graphs, AISecKG, to automatically generate question-answers from LLMs, which are evaluated by the domain expert. We also present a question-answering generative model to answer the open-ended questions.

Question-Answer Generation

Problem Formulation:

We aim to generate open-ended questions and answers for topics related to attacks and defense mechanisms. It is an exceptionally time-consuming and knowledge-intensive task for a domain expert to create questions and answers specific to exposed vulnerabilities, attacks, and the security defense tools and techniques. Even to automatically generate the questions from LLMs, these topics need specific domain expertise to design the prompts and validate the responses generated by the LLMs.

We use structured knowledge from cybersecurity KG to minimize domain expert involvement in prompt design. Our three-step prompting method, **CyberGen**, mitigates hallucinations in LLM responses. ChatGPT was selected over other LLMs due to its training with Reinforcement Learning from Human Feedback (RLHF), which aligns its responses more closely with human expectations and reduces hallucinations compared to open-source counterparts.

Cybersecurity Knowledge-Graph

AISecKG (Agrawal et al. 2023b) is a cybersecurity education ontology that describes the interactions between different concepts, applications, and roles in the cybersecurity domain. These three categories have a further 12 types of entities. The concepts are classified as features, functions, data, attacks, vulnerabilities, and techniques. The applications denote the tools, systems, and apps. The roles are user, attacker, and security teams. There are nine relations to represent the fundamental interactions between these cybersecurity entities. The schema presented in the ontology contains 68 unique edges.

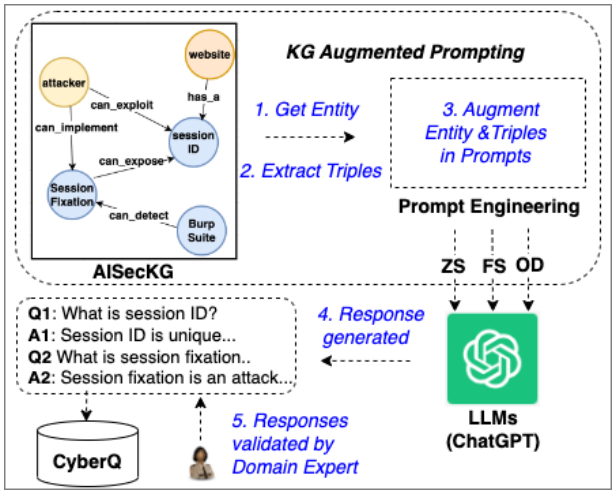

Figure 1: CyberGen: KG-augmented LLMs pipeline to generate question-answers by using Zero-Shot (ZS), Few-Shot (FS) and Ontology-Driven (OD) prompt-chaining.

In this work, we selected the ten most prevalent entities for each type from AISecKG, totaling 120 entities for question and answer generation. Since our focus is on vulnerabilities and attacks, we chose the related schema tuples. For instance, the entity *"session ID"*, categorized as a *feature*, the tuples considered from ontology were *('attacker, 'can_exploit, 'feature')* and *('securityTeam', 'can_analyze', 'feature')*. The prompts were designed using these tuples.

CyberGen

We refer to our prompting method as CyberGen. In our approach, we employ three different prompt techniques to generate questions from ChatGPT: ***Zero-Shot (ZS)***, ***Few-Shot (FS)*** within context, and ***Ontology-Driven (OD)*** using domain-specific schemas. These prompting techniques are used sequentially to provide context and enhance the prompts with knowledge from AISecKG triples. Our goal is not to compare these techniques or determine superiority but to create a sequence of prompts, forming a coherent thought process for ChatGPT. This step-by-step prompt chaining assists the language model in *generating a continuous stream of thoughts and constructing a mental mind map.* In Table 1, we provide examples of prompts and the resulting questions using these techniques. These samples illustrate how incorporating structured knowledge enhances the complexity and depth of the generated questions. Domain experts validate the responses generated by LLMs, as depicted in the pipeline outlined in Figure 1, which illustrates the methodology we employ to generate question-answers. We term our final validated dataset as **"CyberQ."**

Step-wise Prompting Techniques: In the initial step, we employ a Zero-Shot (ZS) approach by presenting ChatGPT

with an entity and asking it to provide a paragraph explaining that entity. This helps establish context and assess ChatGPT's familiarity with the concept. In the second step, we use the generated write-up to formulate questions and answers. Following that, we employ Few-Shot (FS) prompting by offering two sample in-context questions from the ontology tuple, prompting ChatGPT to create questions and answers accordingly. Lastly, we utilize a scenario-based prompt from the knowledge graph ontology schema to instruct ChatGPT in generating specific, intricate questions with comprehensive answers, rather than generic ones.

Zero-Shot (ZS) Prompting Considering the entity, *"session ID,"* from AISecKG, we began with the initial prompt, *"Tell me about session ID."* Following ChatGPT's generation of a brief paragraph on the entity, we presented a subsequent prompt: *"Generate questions with answers on session ID based on the above write-up."* As shown in Table 1, the questions generated in this case tend to be generic. Additionally, we asked ChatGPT to rephrase these questions, which were later used to test the model. We skipped the entity if the initial paragraph provided after the first response was unsatisfactory. This step serves to maintain factual accuracy in responses and reduce the occurrence of erroneous information.

In-Context Few-Shot (FS) Prompting During Few-Shot prompting, two sample questions were provided for generating similar questions. These sample questions were manually curated and carefully selected, aligning with the schema and paths from the cybersecurity knowledge graph, AISecKG. In this scenario, we noticed that the generated questions often mirrored the pattern in the provided examples. As demonstrated in Table 1, some of the questions were essentially paraphrases of the example questions.

In-Domain Ontology-Driven (OD) Prompting The third method is using the schema triples from the AISecKG ontology. For example, the corresponding triple for *"session ID"* in the AISecKG dataset is *('attacker,' 'can_exploit, 'session ID')*. The prompt designed for this method was by augmenting the domain-specific knowledge triple in the prompt. We gave the context as a use case and asked ChatGPT to generate questions and answers for that situation. For example, *"Generate ten questions with answers on situations in which attacker can exploit session ID"*. We used two or three tuples from the schema related to attacks and vulnerabilities for each entity to generate questions and answers for different scenarios. As shown in the **Table** 1, the generated responses are more complex and specific to a situation, and domain coverage is much higher.

Motivation behind CyberGen

The motivation for using the step-wise prompt chaining, CyberGen using these three techniques can be summarized as follows:

Contextual Depth: The step-by-step chaining approach allows the language model to delve deeper into the topic by gradually building upon the information provided in previous prompts. This incremental approach helps create a richer context and ensures that subsequent questions and answers are more informative and contextually relevant.

Knowledge Augmentation: Each prompting method contributes knowledge and context to the questions and answers. ZS prompting lays the foundation, while FS and OD add more context and complexity. This approach allows the LLMs to generate a wide spectrum of questions, covering both basic and advanced levels of knowledge.

Question Types and Complexity level: In NLP, Question-Answering tasks fall into two broad categories: Open-Domain Question Answering (ODQA) and Closed-Domain Question Answering (CDQA). However, our QA dataset focuses exclusively on cybersecurity, leading us to adopt a more specific terminology: Open-Book and Closed-Book QA solutions. For the Close-Book approach, we leverage zero-shot (ZS) prompting to generate questions that require comprehensive knowledge. These questions are similar to reading comprehension tasks, like the SQuAD (Rajpurkar et al. 2016) dataset. These closed-book questions, generated using ZS prompting, are relatively straightforward and more accessible to answer. The Few-Shot (FS) and Ontology-Driven (OD) questions fall under the Open-Book category, demanding a deeper understanding of the subject matter. By employing these three distinct techniques, our dataset, CyberQ, offers a well-balanced mix of straightforward and challenging questions.

Seamless QA Generation: In the rapidly evolving field of cybersecurity, our pipeline ensures effortless updates to the QA dataset. Leveraging two key inputs: x***(LLM)*** and y***(KG)***, it uses, **automated process** (our method) to access reliable **(KG)** and generate ***multi-format questions*** efficiently. This approach is **generalizable** and can be readily applied to create QA datasets for different domains.

CyberQ Dataset

Around 4k questions were generated by chaining the three prompting techniques as proposed in CyberGen. Each technique's distributions of questions are as shown in **Table** 2. Most of the questions generated by ChatGPT are Wh-type seeking content information. The focus is on explanation of concepts, use of tools, and 'how-to' setup instructions. Some strategic questions are related to attack analysis and defense techniques. All the questions are open-ended with varying length answers. First the domain expert validated the questions and answers to ensure they were consistent and factually correct as per domain, and students can use those. Then we did a statistical analysis to gain more insights of the validated data.

Domain Expert Evaluation

In education, ensuring the accuracy of learning materials is crucial. To validate the generated cybersecurity questions and answers, we enlisted domain experts who possess substantial knowledge and experience. In our case, two of the

Method 1—Zero-Shot (ZS) Prompting

Entity: [“session ID”] , **Ontology:** - None
Prompt 1: “Tell me about session ID.”
Response —
Para: “A session ID (session identifier) is a unique identifier that is assigned to user or a client during a session...”

Prompt 2: “Generate questions with answers based on the above write-up.”
Response —
Q1: What is a session ID, and why is it important?
Answer: “A session ID is a unique identifier assigned to a user during a session with a web server. It is important...”
Q2: How is a session ID generated?
Answer: “A session ID is generated on the server-side and is usually a random string of characters...”

Method 2 — In-Context Few-Shot (FS) Prompting

Entity: [“session ID”, “vulnerability”], **Ontology:** (“feature”, can_expose, “vulnerability”)

Prompt 3: “Generate questions with answers similar to questions, Q1: Will the app become vulnerable if the session ID is exposed?, Q2: Which tools are used to analyze vulnerability exposed by the session ID?”
Response —
Q1: “What are some commonly used tools to analyze vulnerabilities related to the session ID?”
Answer: “There are several tools that can be used to analyze vulnerabilities related to session ID, such as Burp Suite..”
Q2: “How does exposing a session ID make an application vulnerable?”
Answer: “If a session ID is exposed, an attacker could potentially use it to hijack the user’s session and gain unauthorized...”

Method 3 — In-Domain Ontology-Driven (OD) Prompting

Entity: [“session ID”, “vulnerability”], **Ontology:** (“feature”, can_expose, “vulnerability”)

Prompt 4: “Generate questions with answers on situations in which Session ID can expose vulnerabilities.”
Response —
Q1: What is session replay, and how does it exploit session IDs?
Answer: “Session replay is an attack where an attacker records a user’s session, including their session ID, and replays...”
Q2: What is session hijacking, and how does it expose a user’s session ID to attackers?
Answer: “Session hijacking is an attack where an attacker steals a user’s session ID and uses it to impersonate the user...”

Entity: [“session ID”, “attacker”], **Ontology:** [(“attacker”, can_exploit, “feature”)]

Prompt 5: “Generate questions with answers on situations in which an attacker can exploit session ID.”
Response —
Q1: What is session fixation, and how can an attacker exploit it?
Answer:“Session fixation is an attack where an attacker sets a user’s session ID before the user logs in...”
Q2: In what ways can a user’s session ID be compromised by attackers?
Answer: “An attacker can steal a user’s session ID through a variety of means, such as intercepting network traffic...”

Entity: [“session ID”,“securityTeam”], **Ontology:** [(“securityTeam”, can_analyze, “feature”)]

Prompt 6: “Generate questions with answers on situations in which an ethical hacker can analyze session ID.”
Response —
Q1: How can an ethical hacker analyze the effectiveness of a website’s session ID management?
Answer: “An ethical hacker can analyze the effectiveness of a website’s session ID management by testing for common...”
Q2: What tools can an ethical hacker use to analyze session ID vulnerabilities?
Answer: “An ethical hacker can use a variety of tools to analyze session ID vulnerabilities, such as OWASP ZAP...”

Table 1: CyberQ samples with varying complexity by chaining three prompting techniques of CyberGen.

Figure 2: Word cloud generated for CyberQ.

authors are graduate researchers in cybersecurity, and the third author is an experienced cybersecurity instructor teaching at the university level. These experts manually reviewed and validated the responses. The first and second authors designed and generated the prompts, while the third author performed the manual validation. The question-answers were evaluated based on feedback and observations from domain experts, yielding the following findings:

Knowledgeable: The primary evaluation criterion in the context of education is to determine the accuracy and knowledge level of ChatGPT's responses in cybersecurity.

- Over 95% of the question-answers generated by ChatGPT in the cybersecurity domain were factually correct.
- Some questions related to the analysis and mitigation of attacks were answered incorrectly, leading to the rejection of such question-answer pairs.
- Examples include hallucinations when answering questions about using "network ingress filtering" techniques to mitigate "Smurf attacks" and how "network administrators" can configure devices to prevent these attacks.

Consistency: The evaluation included an assessment of the consistency of the generated responses, focusing on relevance to the context.

- ChatGPT tends to stay on-topic but can sometimes produce responses that lack context, leading to unrelated questions. For example, in a prompt about "Trojan Horse malware," many generated questions were unrelated.

Reliability: The reliability of ChatGPT's responses was evaluated, revealing instances of biased responses.

- Responses can be a mix of ideas, especially when there is insufficient supporting information.
- In some cases, ChatGPT provided different answers to similar questions.
- For example, ChatGPT suggested "Windows" is secure even after examples of vulnerabilities were provided. However, its stance changes when prompted differently.

General observations Few other general observations were as follows:

- ChatGPT displayed repetition in its responses, often paraphrasing questions and answers, leading to generic and limited technical details.
- Initially, ChatGPT struggled to recognize certain entities, like "brute-force scripts" and "client-network," when using ZS prompting, but improved with additional context.
- Zero-Shot prompts generated generic WH-questions, while Ontology-Driven situation-based questions became more complex and specific with added context.
- However, in few instances with Few-Shot and Ontology-Driven prompts, there were higher occurrences of hallucination and repetition when additional context was introduced. ChatGPT occasionally appeared confused and rambled or repeated itself in such cases.

Statistical Analysis of Dataset

The question-answers was statistically analyzed based on the readability, answer length, and vocabulary diversity with respect to attacks. **Table** 2 shows the computed metric values for all three prompting techniques.

Readability: Readability assesses how easily the average reader comprehends a text, considering factors like lexical, syntactic, semantic, and stylistic complexity. To gauge language readability complexity, we employed the Flesch-Kincaid Grade Level and Gunning Fog index metrics. These metrics express readability in U.S. grade levels, spanning from fifth grade to college graduates and professionals. They also indicate the years of education needed to grasp the text; lower grades signify higher scores. For our QA dataset, scores fall within the 4-16 range, indicating grammatical correctness and consistency, targeting college graduates and professionals. These scores were calculated using the textstat library in Python.

Answer Length: We analyzed the answer length: and termed the questions into "very short" for less than ten tokens, "short" for 10 to 20 tokens, and "long" for over 20 tokens. In Table 2, it is evident that due to the domain's complexity, we have fewer "very short" (30), mostly "short" (1061), and "long" (2439) questions.

Vocabulary Diversity: Vocabulary diversity, concerning attacks, is measured using the token-type ratio (TTR). TTR calculates the ratio of unique words to the total number of words in the text, indicating lexical variation. A high TTR suggests a broader, less focused vocabulary. In our context, a high TTR suggests that questions and answers are more generic, not solely centered on attacks. In Table 2, it's evident that TTR ratios are higher for Zero-Shot and Few-Shot questions compared to Ontology-Driven questions and answers. This discrepancy arises because Ontology-Driven questions explicitly concentrate on scenarios related to attacks, vulnerabilities, and attackers. The word clouds generated from CyberQ dataset **Figure** 2 show the most frequent words in the dataset.

Application of CyberQ

Our approach to data generation using LLMs and the generated question-answer dataset have various applications in cybersecurity education. Large language models (LLMs) are

		Questions			Answers		
		ZS	FS	OD	ZS	FS	OD
Readability	Flesch-Kincaid	6.3	9.9	8.3	14.6	16.7	15.6
	Gunning Fog	4.78	6.46	4.83	9.68	11.11	9.88
Answer Length	Very Short	–	–	–	22	0	8
	Short	–	–	–	515	16	530
	Long	–	–	–	490	316	1633
Vocabulary Diversity	Token-type-ratio	0.16	0.17	0.08	0.17	0.20	0.09
Total Questions		1027	332	2171			

Table 2: Statistical analysis of dataset to gain insights from three prompting techniques.

Dataset	Model	BLEU	ROUGE				METEOR
			Rouge1	Rouge2	RougeL	RougeLS	
OD	T5-S	0.10	0.39	0.22	0.35	0.35	0.28
	T5-B	0.12	0.41	0.24	0.37	0.37	0.30
FS	T5-S	0.04	0.32	0.20	0.29	0.29	0.20
	T5-B	0.04	0.33	0.20	0.30	0.30	0.21
ZS	T5-S	0.17	0.43	0.27	0.40	0.40	0.33
	T5-B	0.18	0.46	0.29	0.43	0.43	0.36

Table 3: Question-Answering Performance (BLEU, ROUGE and METEOR) of each model on the Question-Answering Task for three subsets of CyberQ.

often not available to the students. Also, LLMs tend to hallucinate, so their responses can only be trusted for education purposes with expert validation. It is hard for students to develop a QA agent to help them learn. Here, we show an application of using such a QA dataset to create a small language model with knowledge of the security domain.

Question-Answering Model

We develop an open-ended question-answering model for cybersecurity education. The answers to the questions are elaborate, detailing a process, and often contain abstract information. Therefore, we consider developing a generative QA model that has been fine-tuned on CyberQ dataset. The base of our QA models is the T5 model (Roberts et al. 2019): t5-small and t5-base. We separately trained our models on three versions of data in CyberQ, Zero-Shot, in-context Few-Shot, and in-domain Ontology-Driven methods. For each dataset, we split the dataset into train-dev-test in the ratio 70:20:10. We trained all the models for 20 epochs with a learning rate of 5e-4 and batch size of 20 with a maximum sequence length of 128. The question-answer dataset CyberQ and implementation code for the QA model are available in our github repository [1].

Results and Analysis

The results of our QA models can be seen in Table 3. Since our task is generative QA, the answers are elaborate and often might not precisely match the generated gold answers. Hence, we measure the performance using three popular generative metrics: BLEU (Papineni et al. 2002), ROUGE (Lin 2004), and METEOR (Banerjee and Lavie 2005). We compare our prediction results with the ChatGPT-generated expert-validated answers. We can see that for each of the three datasets, t5-base outperforms t5-small in almost all the metrics, which shows that with the increase in model parameters, the performance also increases. We also see that our model performs best for the Zero-Shot dataset, which we believe is because the questions and answers in ZS are simple and straight-forward, and the model finds it easy to answer those questions. The model needs to be trained on more training samples to improve the performance, and context information can be added to the training data. We do not use large models above the T5-base because large computations to train these models may not be readily available to all students or novice professionals.

Conclusion

This work presents an open-ended question-answer (QA) dataset, CyberQ, in cybersecurity education. We also trained small Question-Answering models (having fewer parameters) based on CyberQ dataset to build AI-enabled self-paced interactive education systems. Such tools are effective learning modes, especially for complex subjects like cybersecurity. However, creating a question-answering knowledge base for cybersecurity is cognitively demanding for a subject matter expert. Even to use the LLMs to generate questions and answers on cybersecurity automatically, domain expertise is needed to design the specific kinds of prompts. This work shows a novel method, CyberGen, to augment knowledge triples from cybersecurity knowledge graphs AISecKG to create prompts and generate questions and answers from LLMs which were validated by the domain expert. In this study, we demonstrate the capability of ChatGPT in generating QA dataset. However, our methods are applicable across various language models, such as Flan-T5 (Chung et al. 2022), or Llama (Touvron et al. 2023), to generate additional QA pairs. Besides its primary use in QA tasks, this dataset can be a valuable resource for students in other cybersecurity tasks, such as vulnerability or binary analysis. We aim to showcase the versatility of our approach in supporting a broader range of applications in different domains.

[1] https://github.com/garima0106/AISECKG-QA-Dataset.git

Acknowledgments

We are thankful to National Science Foundation under Grant No. 2114789 for supporting this research work. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Abu-Salih, B. 2021. Domain-specific knowledge graphs: A survey. *Journal of Network and Computer Applications*, 185: 103076.

Adamopoulou, E.; and Moussiades, L. 2020. Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2: 100006.

Agrawal, G.; Bertsekas, D.; and Liu, H. 2023. Auction-Based Learning for Question Answering over Knowledge Graphs. *Information*, 14(6): 336.

Agrawal, G.; Deng, Y.; Park, J.; Liu, H.; and Chen, Y.-C. 2022. Building Knowledge Graphs from Unstructured Texts: Applications and Impact Analyses in Cybersecurity Education. *Information*, 13(11): 526.

Agrawal, G.; Kumarage, T.; Alghami, Z.; and Liu, H. 2023a. Can Knowledge Graphs Reduce Hallucinations in LLMs?: A Survey. *arXiv preprint arXiv:2311.07914*.

Agrawal, G.; Pal, K.; Deng, Y.; Liu, H.; and Baral, C. 2023b. AISecKG: Knowledge Graph Dataset for Cybersecurity Education. *AAAI-MAKE 2023: Challenges Requiring the Combination of Machine Learning 2023*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. Ann Arbor, Michigan: Association for Computational Linguistics.

Bonwell, C. C.; and Eison, J. A. 1991. *Active learning: Creating excitement in the classroom. 1991 ASHE-ERIC higher education reports.* ERIC.

Bonwell, C. C.; and Sutherland, T. E. 1996. The active learning continuum: Choosing activities to engage students in the classroom. *New directions for teaching and learning*, 1996(67): 3–16.

Chen, L. E.; Cheng, S. Y.; and Heh, J.-S. 2021. Chatbot: a question answering system for student. In *2021 International Conference on Advanced Learning Technologies (ICALT)*, 345–346. IEEE.

Chen, P.; Lu, Y.; Zheng, V. W.; Chen, X.; and Yang, B. 2018. Knowedu: A system to construct knowledge graph for education. *Ieee Access*, 6: 31553–31563.

Chen, Y.; Wu, L.; and Zaki, M. J. 2023. Toward Subgraph-Guided Knowledge Graph Question Generation With Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*.

Chen, Y.-C.; and Terada, T. 2021. Development and validation of an observation-based protocol to measure the eight scientific practices of the next generation science standards in K-12 science classrooms. *Journal of Research in Science Teaching*, 58(10): 1489–1526.

Chi, M. T.; Adams, J.; Bogusch, E. B.; Bruchok, C.; Kang, S.; Lancaster, M.; Levy, R.; Li, N.; McEldoon, K. L.; Stump, G. S.; et al. 2018. Translating the ICAP theory of cognitive engagement into practice. *Cognitive science*, 42(6): 1777–1832.

Chi, M. T.; and Wylie, R. 2014. The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, 49(4): 219–243.

Chung, C.-Y.; Hsiao, I.-H.; and Lin, Y.-L. 2023. AI-assisted programming question generation: Constructing semantic networks of programming knowledge by local knowledge graph and abstract syntax tree. *Journal of Research on Technology in Education*, 55(1): 94–110.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Chung, W. 2017. Developing curricular modules for cybersecurity informatics: An active learning approach. In *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 164–166. IEEE.

Dan, Y.; Lei, Z.; Gu, Y.; Li, Y.; Yin, J.; Lin, J.; Ye, L.; Tie, Z.; Zhou, Y.; Wang, Y.; et al. 2023. EduChat: A Large-Scale Language Model-based Chatbot System for Intelligent Education. *arXiv preprint arXiv:2308.02773*.

Danon, G.; and Last, M. 2017. A syntactic approach to domain-specific automatic question generation. *arXiv preprint arXiv:1712.09827*.

Deng, Y.; Lu, D.; Huang, D.; Chung, C.-J.; and Lin, F. 2019. Knowledge graph based learning guidance for cybersecurity hands-on labs. In *Proceedings of the ACM conference on global computing education*, 194–200.

Deng, Y.; Zeng, Z.; and Huang, D. 2021. Neocyberkg: Enhancing cybersecurity laboratories with a machine learning-enabled knowledge graph. In *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education V. 1*, 310–316.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dolmans, D.; and Schmidt, H. 2010. The problem-based learning process. *Lessons from problem-based learning*, 13–20.

Fariani, R. I.; Junus, K.; and Santoso, H. B. 2023. A Systematic Literature Review on Personalised Learning in the Higher Education Context. *Technology, Knowledge and Learning*, 28(2): 449–476.

Ji, S.; Pan, S.; Cambria, E.; Marttinen, P.; and Philip, S. Y. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2): 494–514.

Ji, Z.; Choi, E.; and Gao, P. 2022. A Knowledge Base Question Answering System for Cyber Threat Knowledge Acquisition. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, 3158–3161. IEEE.

King, A. 2002. Structuring peer interaction to promote high-level cognitive processing. *Theory into practice*, 41(1): 33–39.

Li, J.; Tang, T.; Zhao, W. X.; Nie, J.-Y.; and Wen, J.-R. 2022. Pretrained language models for text generation: A survey. *arXiv preprint arXiv:2201.05273*.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Liu, J.; Liu, C.; Lv, R.; Zhou, K.; and Zhang, Y. 2023. Is chatgpt a good recommender? a preliminary study. *arXiv preprint arXiv:2304.10149*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Malinka, K.; Peresíni, M.; Firc, A.; Hujnák, O.; and Janus, F. 2023. On the educational impact of ChatGPT: Is Artificial Intelligence ready to obtain a university degree? In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1*, 47–53.

Mao, Y. 2021. Summary and evaluation of the application of knowledge graphs in education 2007–2020. *Discrete Dynamics in Nature and Society*, 2021: 1–10.

Menekse, M.; Stump, G. S.; Krause, S.; and Chi, M. T. 2013. Differentiated overt learning activities for effective instruction in engineering classrooms. *Journal of Engineering Education*, 102(3): 346–374.

Mzwri, K.; and Turcsányi-Szabo, M. 2023. Internet Wizard for Enhancing Open Domain Question Answering Chatbot Knowledge-base in Education.

Okonkwo, C. W.; and Ade-Ibijola, A. 2020. Python-Bot: A chatbot for teaching python programming. *Engineering Letters*, 29(1).

Okonkwo, C. W.; and Ade-Ibijola, A. 2021. Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence*, 2: 100033.

Oliveira, N.; Sousa, N.; and Praça, I. 2021. A Search Engine for Scientific Publications: A Cybersecurity Case Study. In *International Symposium on Distributed Computing and Artificial Intelligence*, 108–118. Springer.

Pal, K. K.; Kashihara, K.; Banerjee, P.; Mishra, S.; Wang, R.; and Baral, C. 2021. Constructing flow graphs from procedural cybersecurity texts. *arXiv preprint arXiv:2105.14357*.

Pan, S.; Luo, L.; Wang, Y.; Chen, C.; Wang, J.; and Wu, X. 2023. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *arXiv preprint arXiv:2306.08302*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Perez-Beltrachini, L.; Jain, P.; Monti, E.; and Lapata, M. 2023. Semantic Parsing for Conversational Question Answering over Knowledge Graphs. *arXiv preprint arXiv:2301.12217*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1): 5485–5551.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Roberts, A.; Raffel, C.; Lee, K.; Matena, M.; Shazeer, N.; Liu, P. J.; Narang, S.; Li, W.; and Zhou, Y. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Sayan, C.; Hariri, S.; and Ball, G. 2017. Cyber security assistant: Design overview. In *2017 IEEE 2nd International Workshops on Foundations and Applications of Self* Systems (FAS* W)*, 313–317. IEEE.

Shuster, K.; Poff, S.; Chen, M.; Kiela, D.; and Weston, J. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Su, D.; Xu, Y.; Winata, G. I.; Xu, P.; Kim, H.; Liu, Z.; and Fung, P. 2019. Generalizing question answering system with pre-trained language model fine-tuning. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, 203–211.

Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.-T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, Z.; Valdez, J.; Basu Mallick, D.; and Baraniuk, R. G. 2022. Towards human-like educational question generation with large language models. In *International conference on artificial intelligence in education*, 153–166. Springer.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Wu, E. H.-K.; Lin, C.-H.; Ou, Y.-Y.; Liu, C.-Z.; Wang, W.-K.; and Chao, C.-Y. 2020. Advantages and constraints of a hybrid model K-12 E-Learning assistant chatbot. *Ieee Access*, 8: 77788–77801.

Xia, X.; and Qi, W. 2023. learning behavior interest propagation strategy of MOOCs based on multi entity knowledge graph. *Education and Information Technologies*, 1–29.

Yang, J.; Jin, H.; Tang, R.; Han, X.; Feng, Q.; Jiang, H.; Yin, B.; and Hu, X. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*.