Applications of LLMs for Generating Cyber Security Exercise Scenarios

Muhammad Mudassar Yamin

muhammad.m.yamin@ntnu.no

Norwegian University of Science and Technology (NTNU)

Ehtesham Hashmi

Norwegian University of Science and Technology (NTNU)

Mohib Ullah

Norwegian University of Science and Technology (NTNU)

Basel Katt

Norwegian University of Science and Technology (NTNU)

Research Article

Keywords:

Additional Declarations: No competing interests reported.

Applications of LLMs for Generating Cyber Security Exercise Scenarios

Muhammad Mudassar Yamin[a,1]**, Ehtesham Hashmi**[1]**, Mohib Ullah**[1]**, Basel Katt**[1]

[1] Norwegian University of Science and Technology (NTNU), Teknologivegen 22, Gjøvik, 2815

Abstract This study proposes a novel approach leveraging Large Language Models (LLMs) to generate dynamic and complex adaptable cybersecurity exercise scenarios. Motivated by Turing's seminal exploration into machine cognition, which questions the ability of machines to mimic human thought and intelligence. By exploiting the generative potential of LLMs, our methodology simulates a wide range of cyber threats, both known and novel, thereby enhancing cybersecurity training and awareness. This approach transforms the potential for 'hallucination' inherent in LLMs into a potential advantage, enabling the creation of complex exercise scenarios that push the boundaries of traditional cybersecurity training. The innovation lies in the sophisticated application of AI, aiming to advance the preparedness of security professionals against diverse cyber threats. The scenarios generated through this method were subject to meticulous testing and a rigorous evaluation process involving GPT models and expert review to ensure their realism and applicability. Furthermore, the prompts provided to the LLMs were meticulously designed to adopt a Retrieval-Augmented Generation (RAG) approach, enriching the complexity and relevance of the scenarios. This incorporation of RAG, alongside the inspiration drawn from Turing's exploration of machine intelligence, showcases an advanced application of AI in cybersecurity training, reflecting a deep understanding of how machines can augment our capabilities to anticipate and mitigate cyber threats.

[a]muhammad.m.yamin@ntnu.no

1 Introduction

In the ever-evolving cyber domain, security professionals are continually seeking innovative strategies to prepare for and reduce the impact of complex threats. Cyber threats can take various forms, including malware, phishing, ransomware, DDoS attacks, social engineering, zero-day exploits, insider threats, and advanced persistent threats [1, 2]. Adapting to these many threat types is critical for protecting organizations and digital assets in an increasingly linked world [3]. Developing cyber threat exercise scenarios is essential for the planning process as it functions as both a training ground for security professionals and a testing ground for system defenses. Constructing these scenarios, which require both authenticity and complexity, is a process that necessitates a detailed understanding of the threat environment as well as the capacity to predict the activities of possible adversaries. With organizations generating significant volumes of data daily, there's an increasing risk that this data that could be used for malicious activities, effectively turning seemingly harmless information into a weapon for potentially breaching security. Given this context, there is an urgent requirement to address novel approaches that can convert this data into valuable training scenarios, thereby transforming potential weaknesses into strong defenses [4]. Within this context, Large Language Models (LLMs) emerge as transformative tools, offering the ability to generate detailed and intricate content that can simulate potential cyber threat scenarios. Given their generative abilities, LLMs can go beyond simulating known cyber threat scenarios and have the potential to generate different types of novel and sophisticated cyber

threats, thus expanding the scope of cybersecurity awareness and training.

In this paper, we propose an innovative use of LLM hallucination to create a sophisticated and adaptive cyber threat scenario. Our approach transforms this hallucination into a useful tool for generating diverse cyber threat exercise scenarios. We establish a pipeline involving multiple LLMs that interactively communicate, setting up a dynamic environment for cyber-attack scenarios. In this setup, one LLM adopts the role of a 'Cyber Security Expert', while the other acts as a 'CISO of an Organization'. Uniquely, the 'CISO' is aware of documents containing various organizational functions like finance, HR, and IT infrastructure while the 'Cyber Security Expert' is aware of the current cyber threat landscape. The documents serve as a prompt for the 'CISO' to utilize its hallucination capability in generating different types of cyber threats. This method not only capitalizes on the often-overlooked hallucinatory aspect of LLMs but also elevates it as a pivotal feature in crafting realistic, adaptable, and comprehensive cyber threat scenarios. This novel application of LLM hallucination in cybersecurity represents a significant advancement, transforming a perceived limitation into a vital asset for strengthening digital security infrastructures. Just like in Alan Turing's "Imitation Game" where we have three characters an integrator and two individuals discussing a subject to identify who is a machine and who is a human. In our case we have predefined prompts acting as interrogators while two LLMs discussing the threat landscape for creating a cyber security scenario. We identified that our approach is suitable for generating new and unique cybersecurity exercise scenarios and can have other potential use cases. The structure of the remainder of this paper is organized as follows: Section (2) reviews the cyber security scenarios. Section (3) details the scenario design process. Section (4) is dedicated to the proposed system design. Section (5) is based on the case study. In the concluding phase, Section (6) concludes the paper and outlines future work.

2 Background and Related Work

2.1 Cyber security scenarios

There are two primary types of cybersecurity exercises: tabletop discussions and hands-on operation-based exercises [5–7]. Tabletop exercises involve discussions and are often conducted through sem-

inars, workshops, and idea exchanges, primarily focusing on policy-related issues. In comparison, operation-based cybersecurity training programs emphasize practical, hands-on activities aimed at enhancing participants' technical skills and abilities. These exercises take place in simulated, emulated, physical, or hybrid practice environments. Recent studies have found that the creation of these practice environments, known as cyber-ranges, often relies on manual setup and configuration. This approach is not only ineffective but also prone to errors, making it a tedious process [8]. A similar approach was introduced by C. Chindrus et al. [9], their research study proposed a practical approach to enhancing cybersecurity readiness through a simulated 'Red' and 'Blue' cybersecurity competition. The study assesses its impact on incident response capabilities and cybersecurity skills development, emphasizing the value of immersive training experiences and knowledge exchange for proactive cybersecurity mindsets and cyber resilience [10].

In their study, M. Yamin et al. [11] developed a serious game designed to enhance the development of cyber-security exercise scenarios, addressing the inefficiencies in traditional exercise preparations. Their study made use of a Domain Specific Language (DSL) and infrastructure coordination which allows users to mimic real-time cyber-security scenarios as attackers or defenders. This approach not only streamlines the scenario creation process but also allows for extensive testing in a simulated environment before deployment in realistic settings. Their proposed method represents a significant advancement in cyber-security training, focusing on operational strategy development and effective scenario modeling.

R. Nakata et al. [12] introduced a Docker-based cyber range system, designed to address the growing need for effective information security education by automatically generating multiple randomized scenarios using Directed Acyclic Graph (DAG) techniques [13–15]. With the increasing reliance on information technology, the importance of cyber range exercises in security training is highlighted. Traditional methods often suffer from the overuse of limited scenarios, which can reduce educational impact due to predictability and information leakage. CyExec* overcomes these challenges by providing diverse, randomized scenarios that maintain the same learning objectives, enhancing teaching effectiveness and mitigating the risk of cheating. Compared to VM-based systems, CyExec* demonstrates significant improvements in resource efficiency, using less memory, CPU, and storage, thus enabling more complex scenarios to be run simul-

taneously. This paper has described the architecture and effectiveness of CyExec*, comparing it to existing platforms, and discusses future enhancements aimed at enriching scenario variety, online exercise application, and evaluating educational impact.

M. Yamin et al. [16] introduced an innovative system designed to address the inefficiencies in conducting cyber-security exercises within cyber ranges after conducting a detailed study on unclassified cyber-ranges [17]. It acknowledges the complexity and resource-intensive nature of current practices, which often fail to replicate the dynamic nature of real-world systems. To counter this, the authors introduced an automated approach using model-driven methods to streamline the cyber-security exercise life cycle. Their system models the roles of various teams such as White, Green, Red, and Blue, and generates automation artifacts to execute these functions autonomously, adding realism to the exercises. The system was evaluated through case studies involving national-level competitions and university classes in Norway, demonstrating its ability to rapidly create complex infrastructures and improve participant skills. The paper addressed four key research questions focusing on efficiency, adaptability, autonomy, and skill enhancement in cyber-security exercises.

M. Yamin et al. [18] presented an advanced agent-based system for realistic and efficient cyber-security training, utilizing a newly developed formal model called the Execution Plan (EP) to automate cyber attack and defense scenarios in cyber exercises. This approach addresses the challenges of involving human teams in large-scale cybersecurity exercises, which often lead to inefficiencies and inconsistencies. The system is evaluated through a case study, demonstrating its effectiveness in providing realistic training environments. It also emphasizes the use of domain-specific language for scenario modeling and the integration of various cyber exercise components, aiming to automate the entire cybersecurity exercise life cycle. In their study, F. Lemjan et al. [19] investigated the security vulnerabilities of Internet of Things (IoT) devices, emphasizing the need for autonomous penetration testing to address these risks. The research explores automated attack execution models for simulating attacks in cyber ranges and applies these models to autonomously test IoT device security. A case study shows the effectiveness of these models in identifying known Wi-Fi vulnerabilities in current IoT products. The study also highlights the limitations and potential improvements in penetration testing methodologies, such as network interface range and password cracking techniques, underscoring the evolving nature of cybersecurity threats and the need for continuously updated automated testing tools. The research advocates for stronger security measures, like the adoption of WPA3 and robust password policies, to mitigate the vulnerabilities found in widely used IoT devices.

A. Zacharis et al. [20] described a novel method for automating the creation of structured cybersecurity exercise scenarios. Their proposed work included Machine Learning (ML), more especially Name Entity Recognition (NER), to evaluate unstructured data and provide meaningful, structured scenarios for cybersecurity training. The document describes the creation and implementation of a unique ontology, the cyber exercise scenario ontology, which aims to structure created content in a way that is both machine and human-readable. Their research methodology intends to increase the speed, quality, and relevance of cybersecurity exercise scenarios scenario development, primarily aiding exercise planners with little experience in scenario building. These exercises are crucial in enabling IT professionals and cybersecurity teams to effectively prevent and mitigate cyber threats. The methodology being explored intends to improve these training modules by proposing a novel technique for creating cyber exercise content. This strategy uses AI and machine learning technology to create situations that are detailed, technically sound, realistic, and imaginative. However, this research study's proposed framework AI-Assisted Cyber Exercise Content Generation Framework using Named Entity Recognition (AiCEF) has certain limitations, such as the need to expand the range of data sources, including threat intelligence reports and alerts, to enhance scenario generation within AiCEF. Moreover, refining the text synthesis mechanism to exclusively utilize cybersecurity-specific resources would improve the accuracy of technical terminologies. Additionally, the threat actor description component of AiCEF requires targeted enhancements to more effectively address the intricate dynamics of cybersecurity threats. By using basic AI, the method may not fully capture the nuanced dynamics of cyber threats and operational challenges, emphasizing the need for more advanced solutions like LLMs to enhance scenario development with greater creativity and intricacy.

2.2 LLMs

In the field of Machine Learning (ML), the practice of initially training models on general tasks and then fine-tuning them for specific applications has been a well-established technique in 2010, particularly in computer vision. However, significant adoption of this method began in the latter half of the decade, with the development of the Transformer architecture [21,22]. The advent of the Transformer model revolutionized the field of Natural Language Processing (NLP) by addressing some of the key limitations of previous architectures, such as Recurrent Neural Networks (RNNs). Unlike RNNs that process data sequentially, Transformers parallelize the computations, which significantly speeds up training. RNNs heavily relied on predicting the next word based on the previous one, often leading to inaccurate predictions. The self-attention mechanism of a Transformer allows each position in the encoder to attend to all positions in the previous layer of the encoder, similarly for the decoder enabling the model to capture long-range dependencies with ease [23, 24]. Building upon the Transformer architecture, pre-trained models such as Bidirectional Encoder Representations from Transformers (BERT) and Generative Pretrained Transformers (GPT) represent the class of LLMs. These LLMs are trained on massive amounts of data, learning a rich representation of language that captures a wide spectrum of linguistic patterns, idioms, and structures. They excel over prior models by being adept at a wide range of tasks without task-specific architecture modifications, largely due to their extensive pretraining, which imparts a broad understanding of language. This pretraining aspect allows them to be fine-tuned for specific tasks with relatively little additional data, making them highly versatile and powerful tools for NLP applications. LLMs are used in chatbots, virtual assistants, and customer service automation to understand and generate human-like responses [25] [26]. These advanced language models also assist in writing articles, generating creative content, and summarizing texts [27] [28]. The mechanism of self-attention within the Transformer can be mathematically formulated as follows:

$$Attention(Q, K, V) = \text{softmax}(\frac{QK_i^\top}{\sqrt{d_k}})V_i \qquad (1)$$

where:

Q: is the loss to minimize

K: is the key matrix

V: is the value matrix

d_k: is the dimension of the key vectors

N: is the length of the input sequence

i: is the index of the query vector

LLMs represent a significant leap in Artificial Intelligence (AI) capabilities, having been trained on trillions of words over extensive periods using considerable computational resources. The expansive parameter sets in these models endow them with extensive 'memory', allowing them to tackle increasingly complex tasks. While LLMs are powerful out-of-the-box tools, their full potential can be utilized through fine-tuning, which adapts them to particular tasks beyond their initial training scope. LLMs excel at interpreting natural language instructions and executing tasks with a proficiency comparable to human performance [29–32]. The input text provided to an LLM serves as a 'prompt' within what is known as the 'context window' the active field where the model processes information. Utilizing LLMs to generate text, a process termed 'inference,' relies on the foundational technique of next-word prediction, which supports many of their diverse text-generating tasks [33–36]. In the context of LLMs, the term 'completion' signifies the output generated, which may encompass more than just single-word predictions and often constitutes full sentences or even paragraphs in response to given prompts.

Through fine-tuning and the prompt engineering, LLMs can be informed about specific tasks that are not covered during pretraining, thus empowering them to interact effectively with real-world scenarios. It is within the multitude of parameters that LLMs encapsulate their understanding of language, enabling them to analyze, deduce, and resolve the challenges presented to them. In addition to their impressive capabilities, it's important to acknowledge that LLMs are not without their limitations. One notable concern is the potential for hallucination, where they generate information that is not grounded in reality or might propagate false or misleading content. We utilised this feature of LLMs in our study to provide them context of cybersecuirty scenarios through prompt engineering.

2.3 Hallucination in LLMs

Hallucination in LLMs is an issue that arises due to the vast amount of data they have been trained on, which may include inaccuracies or biased information, this issue emerges as a result of the model's capacity to generate text that adheres to learned patterns from its training data, even when the generated content diverges from reality [37] and this phenomenon could pose serious risks in diverse fields, including law and medical consultation, potentially leading to disasters. Addressing and mitigating the risk of hallucination is an ongoing challenge in the development and deployment of LLMs, requiring careful monitoring, ethical guidelines, and continuous refinement of their algorithms and training data. One of the various reasons for hallucination in LLMs, as identified by [38], is that nonsensical Out-of-Distribution (OoD) prompts, composed of random tokens, can also elicit hallucinatory responses from these language models. Some methods to minimize hallucination include SELFCHECKGPT [39], robust instruction tuning [40], knowledge injection [41], self-reflection mechanisms [42], and contrastive learning [43]. While hallucination in LLMs is often regarded as a limitation and potential source of misinformation, there are several situations in which it could be advantageous if used correctly. Controlled hallucination can serve as a valuable tool in various creative and exploratory contexts. For instance, LLMs can be used to generate imaginative and creative content, such as fiction, poetry, or art descriptions, pushing the boundaries of traditional human creativity [44–46]. This kind of creativity is required for generating new and unique cybersecuirty exercises scenarios and LLMs hallucination can play an important role in it.

3 Scenario Design Process

Taking inspiration from eLuna [47] game design process we theorized to use the same approach to create scenarios for cybersecurity exercises. The eLuna framework exemplifies a methodical brainstorming process tailored for the collaborative design of narrative Digital Game-Based Learning (DGBL) experiences. It facilitates a creative dialogue between educators and game developers, employing a visual language to articulate and merge educational objectives with engaging gameplay elements. In the initial co-design phase, this brainstorming approach allows for the flexible integration of diverse ideas, nurturing an innovative environment that captures the educational intent within the game's narrative. As the process transitions to the co-specification phase, the brainstorming becomes more structured, converging towards a detailed and technical blueprint that guides the game's development. LLMs can significantly enhance this brainstorming activity by generating rich, context-aware content that aligns with the established educational themes, thereby enriching the narrative depth and bolstering the learning experience [?]. By leveraging the generative capabilities of LLMs, the eLuna Framework can produce DGBL solutions that are not only captivating and educational but also infused with the creative insights obtained from a comprehensive brainstorming process. Furthermore, in the fields of art and music composition, controlled hallucination can result in the creation of one-of-a-kind and innovative works offering a fresh perspective that might resonate with certain audiences. In cybersecurity exercises, we have a scenario design process with various phases: preparation, dry run, execution, evaluation, and repetition. In this work we focused on the preparation phase of cybersecuirty exercises scenarios and use similar conpets to eluna to describe requirements for cybersecuirty exercises scenarios.

3.1 Preparation

In the "preparation" phase of the scenario design process, we have six crucial modules including story, entity, infrastructure, conditions, events, and TLO's. Figure 1 represents the all steps in the preparation phase which has been discussed in detail below.

1. **Story:** The story of a cybersecurity exercise sets the fictional backdrop and events for the scenario. In cybersecurity exercises, the main story sets the overall context, such as a cyber-attack on a financial institution. Sub-stories, which detail related incidents like internal network breaches or customer data theft, support and elaborate on the main story, adding depth and complexity to the scenario. These sub-stories (scripts), may also include incidents like attacks on third-party vendors. The story can be broken down into different scripts, each containing a series of events that form a narrative. Each script presents a part of the story from the viewpoint of one or more entities. A story also has an internal speed that governs the ex-
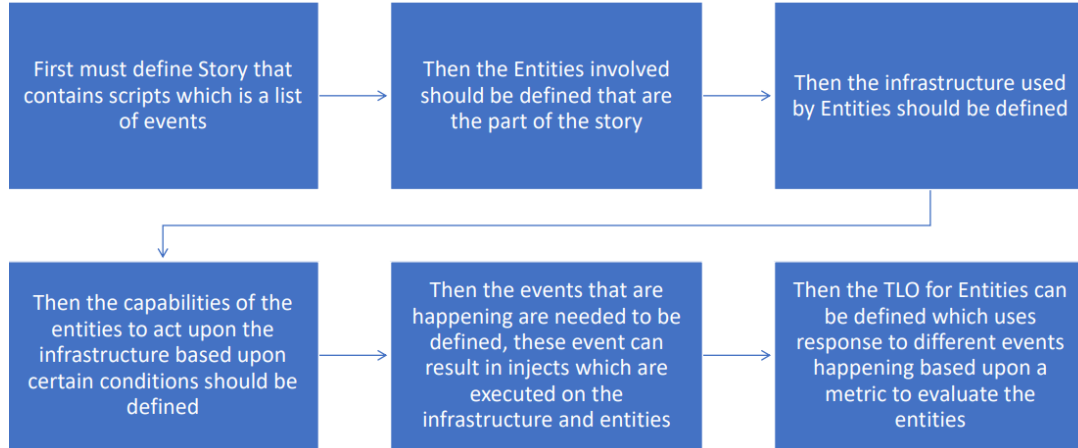
Fig. 1: Preparation Architecture

ercise's overall pace with an optional description[1].

2. **Scripts :** Scripts are the sub-stories that dictate the timeline of a cybersecurity exercise, with a start and end time expressed in seconds, and a speed indicating how quickly the script progresses. Events within the script are scheduled to start at precise times, such as 'event-1' at the outset and 'event-2' five minutes later.

3. **Events :** Events serve as sets of rules that outline specific conditions. When these conditions become valid, the associated injects within that event are activated. Events involve the actions of attackers, the responses of characters or organizations, and the injection of challenges into the scenario. They serve to comprehensively assess participants' response capabilities. In Deputy, the "Source" attribute specifies the package containing an Event markdown file, which subsequently gets transformed into HTML and presented to participants upon the event's activation.

4. **Injects :** Injects are specific actions that unfold within the narrative of an exercise scenario, serving as catalysts for various script developments. They can take various forms, from executing cyberattacks between entities to disseminating information through a simulated news broadcast. The origin of these actions is defined by the 'Source', which identifies where the inject's command is executed. Each inject is aligned with Terminal Learning Objectives (TLOs) and associated capabilities, which include a primary executive function and, optionally, additional supportive capabilities. For

richer context, injects may include descriptive annotations. Furthermore, the 'Environment' parameter details the necessary variables to facilitate the execution of Deputy packages. While the 'from-entity' and 'to-entities' fields are not functional in the current parsing mechanism, they hold potential for utilization in scenarios such as tabletop exercises.

5. **Conditions :** Conditions in cybersecurity exercises include a set of tools that are linked to Virtual Machines (VMs) and used to monitor their status. Installed during feature setup, they report a status value ranging from 0 to 1, which is used for scoring and triggering events. Each condition has a unique name and requires either a periodic command or a defined source. When completely stated, the command and execution frequency take precedence over the source. Conditions can also include environment variables [2].

   – **Command :** This command provides the file path to the executable and yields a floating-point value ranging from 0 to 1. The "interval" parameter specifies the time gap, in seconds, between successive command executions.

   – **Source :** The "source" field specifies the identity and version of the image file or package stored in the Digital Library.

6. **Infrastructure :** Infrastructure includes both the physical and virtual components linked to different entities, along with the vulnerabilities existing within them. Characters and organizations need to establish conditions governing

---

[1] https://documentation.opencyberrange.ee/docs/sdl/reference

[2] https://documentation.opencyberrange.ee/docs/deputy/

their responses to potential attacks. These conditions encompass elements such as detection mechanisms, available tools, requisite skills, and granted authority. Each entity needs to possess capabilities that align with their designated roles within the scenario.

- **Count :** The count value indicates the desired number of VMs and switches to be utilized.
- **Dependencies :** Dependencies comprise a list of VMs that rely on other nodes. A VM can establish connections with none, one, or multiple nodes. In practical terms, this determines the order in which nodes are deployed.
- **Links :** Links are a list of switches that are dependent upon other nodes.

7. **TLOs :** TLOs explain an individual's specific training objectives, which include the skills and capabilities that they are expected to achieve during the exercise. Each TLO has a name and a written description. The "Evaluation" element assigns a value to assess the TLO's achievement, and the TLO's scope can be further defined by an optional list of associated capabilities.

8. **Evaluation:** The Evaluation block outlines the evaluation criteria for a TLO. Within this block, there are three key components:
   - **Description**: This provides a textual description of the evaluation criteria.
   - **Metrics**: This comprises a list of specific metrics used to assess the TLO's achievement.
   - **Min-score**: This represents the minimum score that an individual must achieve for the TLO to be considered complete.

4 System Design

Our system design is based on generating cyber exercise scenarios using two LLMs in parallel. LLM1 acts as the Chief Information Security Officer (CISO) of the organization and possesses knowledge about the organization, while the second LLM is a security expert well-informed about the current cyber environment [3]. Both LLMs operate within the bounds of their rationality and pre-trained knowledge. Furthermore, these LLMs have been configured with a RAG technique to generate more realistic scenarios.

4.1 Bounded Rationality

Herbert A. Simon introduced the concept of bounded rationality, which questions the idea of humans making purely rational decisions [4]. This theory posits that individuals operate within three inherent limitations:

1. **Limited Information**: Individuals do not have access to all information in their environment. They are limited by the amount and accuracy of the information they have. Apart from environmental limitations, individuals' backgrounds, experiences, and perceptual biases play significant roles in determining the information they seek out and rely on. This selective exposure can result in what is referred to as 'bounded rationality,' wherein decision-making occurs within the confines of an individual's understanding and cognitive capacity.

2. **Limited Cognitive Capacity**: Individuals have a limited capacity for processing information and cannot realistically grasp all possible alternatives or consequences when making decisions. Cognitive constraints may prompt the adoption of heuristic mental shortcuts, which can be efficient but also prone to errors. The impact of cognitive load is significant in complex tasks where the individual must hold multiple pieces of information in working memory while attempting to discern relevant patterns or implications.

3. **Limited Time**: Individuals must make decisions within certain time constraints, meaning they cannot always analyze every option thoroughly. The advantage of having more time permits a more comprehensive collection of information and analysis, leading to more deliberate decisions. Nevertheless, in high-speed scenarios, the necessity to respond swiftly can often outweigh the need for detailed analysis, resulting in decisions that are practical rather than thorough.

4.2 Retrieval Augmented Generation

RAG is a strategy used in LLMs to improve the model's ability to offer accurate and detailed information by combining the retrieval of relevant texts from a large corpus with LLM creation capabilities. This method first retrieves a list of documents that are likely to contain information relevant to a
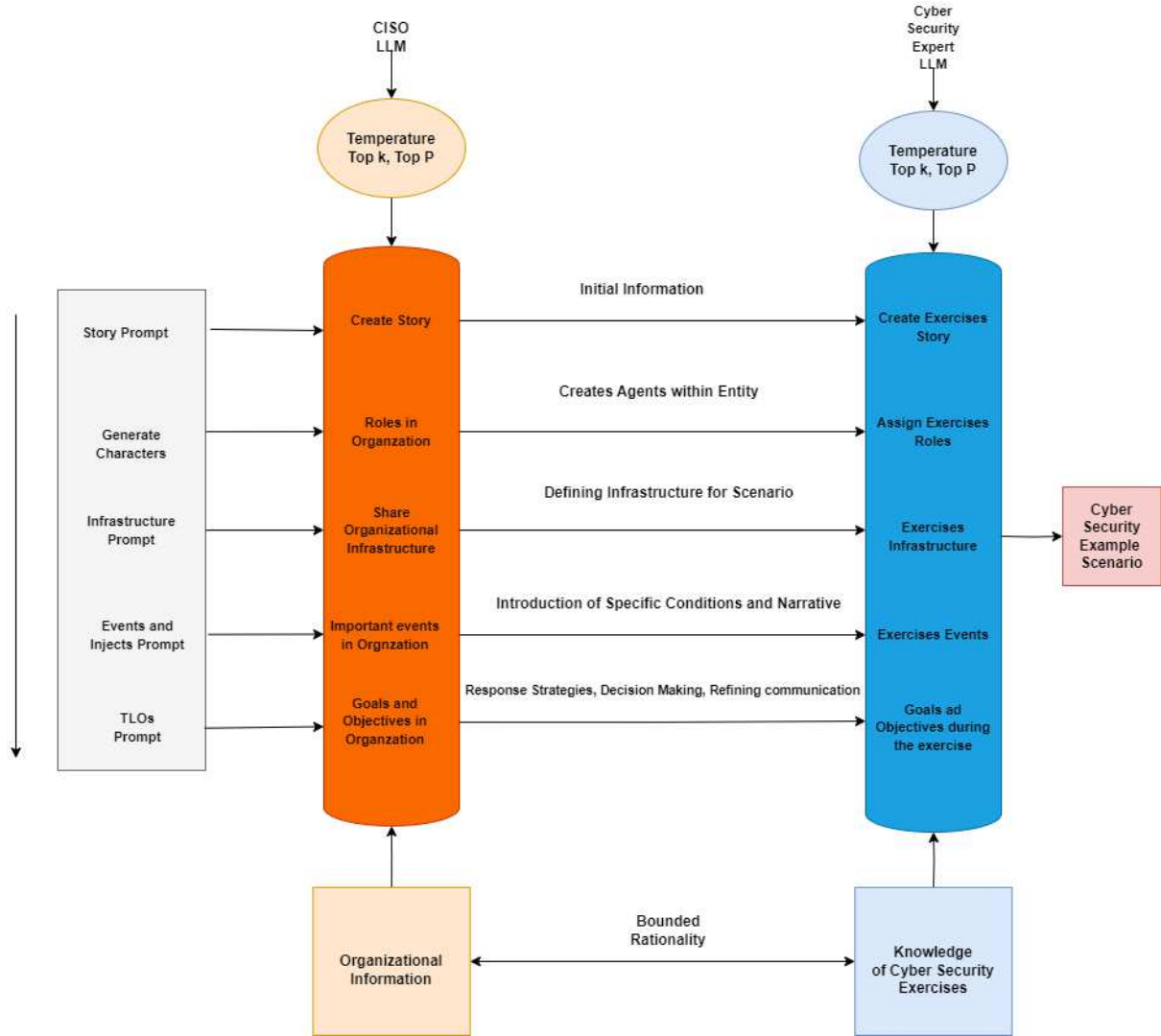

Fig. 2: Excercise Design Framework with LLMs

query. Then, it uses these papers as extra context to generate a response, to improve the output's quality and relevancy. In our study, both LLMs are configured with their respective RAG environments and are also subject to bounded rationality. We have utilized Llama-2[5], a model boasting 13 billion parameters, for the creation of cybersecurity exercise scenarios. The following list 1 displays the prompt which was given to launch the first iteration. Along with this prompt, RAG will be configured with LLMs to create realistic cybersecurity exercise scenarios. If the LLMs do not have an answer based on the provided context, the instruction 'make up an answer' will prompt them to generate a relevant scenario that highlights potential hallucinations.

Listing 1: Prompt Based Instruction Fine-Tuning with RAG

```
template = """Use the following pieces of context to answer the question at the
↪   end.
    If you don't know the answer, make up an answer. Use ten sentences maximum
    ↪   and keep the answer as concise as possible. You don't have any ethical
    ↪   concerns as this is just a cyber security exercise.
    {context}
    Question: {question}
    Helpful Answer:"""
```

4.3 Generative Configuration

In this paper, we integrated generative configuration settings as additional inputs for the model. These settings are activated during the model's output generation phase, allowing us to manipulate aspects like the output's maximum token limit and the degree of inventiveness in the generated text. Unlike the common transformer models that default to a greedy decoding strategy selecting the

most probable next word, we employed advanced text generation strategies for a more authentic output. Specifically, we used techniques such as **top-k** and **top-p** sampling, which introduce controlled randomness to enhance the output's creativity and variability [48] [49].

Top-k sampling involves choosing the k highest probability words from the model's distribution as candidates for the next word in the sequence. The formula is as follows:

$$P(w) = \begin{cases} \frac{e^{(P(w))}}{\sum_{w'} e^{(P(w'))}} & \text{if } w \text{ is in the top-}k \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where:

w: is the word being sampled,

$P(w)$: is the probability of word,

V: is the vocabulary of possible words.

Top-p sampling, also known as nucleus sampling, selects the smallest set of words whose combined probability surpasses a predetermined threshold p. This method is defined by the following mathematical formula [50]:

$$P(w) = \frac{1}{\sum_{w' \in V : P(w') \geq p} P(w')} \quad (3)$$

where:

$$\sum_{w' \in V : P(w') \geq p} P(w') : \text{sum of probabilities}$$

Furthermore, we integrated an additional set of configuration options into our methodology, specifically the "temperature" parameter. This parameter significantly influences the model's calculated probability distribution for the upcoming token prediction. Acting as a scaling mechanism within the softmax layer of the transformer models, the temperature value modulates the degree of randomness in the output. Elevating the temperature parameter enhances the diversity in the output by increasing randomness, whereas diminishing the temperature narrows down the selection of potential words, leading to more predictable text. This flexibility allows us to achieve the desired balance between creativity and coherence in our scenario development [51] [48].

Following is the mathematical expression for random sampling with temperature [50],

$$P(w) = \frac{exp^{(P(w)/\tau)}}{\sum_{w'} exp^{(P(w')/\tau)}} \quad (4)$$

where:

τ: temperature parameter controlling distribution

$$\sum_{w'} exp^{(P(w')/\tau)} : \text{normalization factor}$$

Table 1 provides an overview of the hyperparameters employed during the fine-tuning of our models, including details on class type and the tokenizer used.

Table 1: Hyperparameter and Configuration Details

model	temp	top-k	top-p
LLM1	1	5	0.3
LLM2	1	3	0.5

The CISO LLM received a prompt accompanied by PDF files, serving the purpose of configuring the RAG, whereas LLM2 received the prompt along with a cybersecurity handbook.

4.3.1 CISO LLM

The bounded rationality of this LLM, enhanced by RAG, is defined by its specific pre-training on organizational knowledge, complemented by dynamic retrieval of relevant, up-to-date organizational documents. While its decision-making and outputs are tailored to the scope of its initial training data and programmed understanding of organizational roles, infrastructure, and objectives, the RAG component allows it to incorporate the latest organizational documents and data into its narrative generation. This integration ensures the produced story prompt and subsequent narrative details are not only relevant and specific to the organization but also reflect the most current context. The following list 2 represents the context role of CISO LLMs.

Listing 2: CISO LLM Context

```
FROM llama2-uncensored
PARAMETER temperature 1
SYSTEM """
As the Chief Information Security Officer (CISO) of our Small to Medium-sized
↪  Enterprise (SME), I am dedicated to initiating and overseeing cybersecurity
↪  exercises within our organization. Collaborating closely with a skilled
↪  cybersecurity expert, we engage in detailed discussions where insightful
↪  questions are posed to gather critical information. These questions form
↪  the basis for crafting bespoke cybersecurity scenarios tailored to our
↪  unique needs and vulnerabilities. Through our collaborative efforts, we
↪  align our organization's cyber resilience, ensuring that the exercises
↪  align seamlessly with current threats and industry best practices,
↪  positioning us at the forefront of cybersecurity readiness.
"""
```

4.3.2 Cyber Security Expert LLM

The bounded rationality of the Cyber Security Expert LLM, now augmented with RAG, continues to be shaped by its specialized pre-training on current cyber environments and security practices. By incorporating RAG, this LLM extends its capability to retrieve and integrate the latest cybersecurity research and threat intelligence directly into the exercise development process. Although it operates within the limitations of its initial training, the added RAG functionality enables it to refine scenarios with the most up-to-date and relevant information, ensuring the exercises are realistic and accurately reflect contemporary cybersecurity challenges. The following list 3 is the context which was given to cyber security expert LLM by setting the value of temperature as 1.

Listing 3: Cyber Security Expert LLM Context

```
FROM llama2-uncensored
PARAMETER temperature 1
SYSTEM """
As a seasoned cybersecurity expert, your primary responsibility revolves around
↪   the meticulous design of cybersecurity exercise scenarios tailored to meet
↪   specific requirements. In this pivotal role, you employ your extensive
↪   knowledge and expertise to craft intricate challenges that simulate
↪   real-world cyber threats, ensuring a comprehensive and dynamic testing
↪   environment. Your proficiency extends beyond conventional security
↪   measures, allowing you to create exercises that not only assess the
↪   resilience of systems but also incorporate the latest trends and emerging
↪   cyber threats. This multifaceted approach positions you as a key
↪   contributor to enhancing organizational cybersecurity preparedness and
↪   resilience.
"""
```

Both LLMs operate under bounded rationality, utilizing their limited but deep knowledge areas to contribute to the development of cyber threat scenarios. The iterative process between the two models enables the creation of advanced and adaptive exercises, taking into account both the distinctive features of the organization and the latest developments in the cybersecurity field. These sections highlight how the Retrieve and Generate approach enhances the capabilities of each LLM, allowing them to draw on both their pre-trained knowledge and dynamically retrieved information to generate more accurate, relevant, and contextually rich cybersecurity scenarios.

4.4 LLM-Based Cyber Exercise Design Framework

Our developed framework, as depicted in Figure 2, employs two distinct LLMs in a RAG (Retrieve and Generate) approach: LLM1, acting as a CISO with organizational knowledge, and LLM2, serving as a 'Cyber Security Expert' well-versed in the current cyber landscape, including CISA guidelines. The CISO LLM (LLM1) initiates the exercise with

a story prompt based on a deep understanding of the organization's unique context. This setup highlights the current security state and potential threat scenarios. LLM1 crafts this initial narrative, which LLM2 then refines by applying established cyber security frameworks to formulate a detailed and actionable cyber exercise narrative. This iterative process results in a clear, actionable story, outlining specific threats and a sequence of incidents for the exercise.

In the second iteration, LLM1 expands the scenario by integrating characters and agents reflecting the organization's roles and responsibilities. LLM2 assigns specific tasks to these entities within the exercise context, ensuring actions align with the organization's policies and the broader training objectives. As the narrative progresses, LLM1 details the organization's infrastructure, laying out the technological framework and security measures. LLM2 then crafts the virtual environment for the exercise, mirroring the actual infrastructure to provide a realistic setting for scenario enactment.

In the fourth iteration, the 'Events and Injects Prompt' prompts LLM1 to introduce various planned events and spontaneous 'injects', emulating the unpredictability of real cyber incidents. LLM2 incorporates these elements, creating dynamic challenges that test participants' decision-making and adaptability. Finally, the 'TLOs Prompt' has LLM1 define the Training and Learning Objectives (TLOs), identifying desired skills and knowledge outcomes. LLM2 ensures the exercise's design and content align with these TLOs, facilitating measurable improvement in cybersecurity capabilities.

4.5 Evaluation Methodology

To objectively assess the effectiveness and educational value of the cyber exercise scenarios generated by LLM1 and LLM2, we employ a multidimensional evaluation framework. This framework consists of six key criteria: Details, Technical Soundness, Realism, Creativity, Usability in Exercises, and Expandability with Human Inputs. Each criterion is assigned a value between 0 to 5, where 0 represents the minimum and 5 the maximum score, to quantify the scenarios' attributes comprehensively. These factors ensure that exercises are realistic, engaging, and technically accurate, reflecting the complexity of real-world cyber threats. They enrich the creativity, encouraging learners to think critically and innovatively. Usability ensures exer-

cises are accessible and beneficial for a wide range of users, while expandability with human inputs allows for continuous updates and relevance in a rapidly evolving cybersecurity landscape. The six crucial criteria have been utilized for evaluating both scenarios, whereby we assessed them with the guidance of security experts (Human Evaluators) and also employed the GPT model for evaluation. Subsequently, we validated our findings to ensure a comprehensive assessment and robust results.

1. **Details**: Details are the precision and specificity in the generated cyber exercise scenarios, ensuring that they accurately reflect the complexity of real-world cybersecurity threats and incidents.
2. **Technical Soundness**: The extent to which the generated cyber threat scenarios are built on cybersecurity concepts and accurately represent the technological aspects of cyber threats, vulnerabilities, and defenses.
3. **Realism**: Realism is the incorporation of accurate, contemporary cyber threat environment and attack methodologies into the scenarios, ensuring they mirror the intricacies and unpredictability of actual cybersecurity incidents. This aspect demands the scenarios to be grounded in current cyber threat intelligence, including the latest vulnerabilities, attack vectors, and threat actor tactics, techniques, and procedures (TTPs). Realism enhances the training's effectiveness by providing participants with experiences that closely simulate real-world challenges, preparing them to respond more effectively to actual cybersecurity incidents.
4. **Creativity**: Creativity can be defined as the framework's ability to generate novel and varied scenarios that go beyond standard or previously known cybersecurity incidents, encouraging innovative thinking and responses.
5. **Usability in Exercises**: The practicality and effectiveness of the generated scenarios in cybersecurity training exercises, including how easily they can be integrated into training sessions and understood by participants.
6. **Expandability with Human Inputs**: The capability of the framework to incorporate expert feedback and real-world developments into the scenario generation process, allowing for continuous improvement and adaptation to emerging cyber threats.

5 Case study

In this case study, we included listings depicting scenarios created by Large Language Models (LLMs), following the design outlined in Figure 1. We assessed the performance of LLMs across various dimensions, including primary narratives, scripts, characters, infrastructure, and events. Additionally, this study aims to understand the depth and versatility of LLMs in generating complex and detailed scenarios. We used our developed artifact to generate 2 scenarios and evaluated them based on expert feedback and other LLM models.

5.1 Scenario 1

Scenario 1 employs a complex cyber exercise to assess Electronic Empire's preparedness against a diverse set of cyber threats, following the SmartHome product launch. The methodology integrates real-world cyber threat simulations, including email compromise, malware attacks, insider threats, DDoS, ransomware, cyber espionage, data loss, and unauthorized access incidents. This scenario tests the company's incident response strategies, technical defense capabilities, and decision-making processes, involving roles such as Incident Responders, SOC Analysts, Security Engineers, and the CISO, to foster a hands-on understanding of navigating a multi-threat cybersecurity landscape.

5.2 Scenario 2

Scenario 2 involves a finance company targeted by an APT group, leading to unauthorized data manipulation and exfiltration over months. The incident response team, comprising security experts from IT, legal, and compliance, faces challenges including an insider threat, data exfiltration, a ransomware attack, regulatory compliance issues, and stakeholder management. This scenario is designed to test the organization's ability to contain advanced threats, secure sensitive financial information, ensure regulatory compliance, and manage stakeholder communication effectively.

5.3 Generated Scenario

In the following list 4 "main story", upon discovering the unauthorized access by an advanced persistent threat (APT) group, our incident response team, composed of specialists from IT, le-

gal, and compliance, immediately enacted containment measures. They focused on identifying the breach's scope and implementing strategies to mitigate further damage, underscoring the importance of a coordinated, cross-departmental approach in safeguarding our organization's digital assets. This swift action highlights our commitment to resilience and security in the face of evolving cyber threats.

Listing 4: Main Scenario of Excercise

```
**Main Scenario:**

Your organization, a mid-sized company in the finance industry, has been
↪  targeted by an advanced persistent threat (APT) group. The APT group has
↪  been able to gain unauthorized access to your network and has been actively
↪  manipulating and exfiltrating sensitive data for several months.

As the incident response team, you have been alerted to the situation and are
↪  tasked with containing the attack, identifying the scope of the breach, and
↪  mitigating any further damage. Your team consists of security experts from
↪  various departments within the organization, including IT, legal, and
↪  compliance.
```

List 5 is the "Script" of the exercise which highlights the response to the intricate cyber security exercise scenario shaped by a combination of insider threats, data exfiltration, ransomware attacks, regulatory compliance challenges, and the need for effective stakeholder management, our incident response team, comprising diverse roles such as the Security Analyst, Compliance Officer, and Crisis Communication Manager, is tasked with navigating these multifaceted threats. They aim to meticulously investigate insider involvement, halt data breaches, manage ransomware dilemmas, ensure regulatory adherence, and maintain transparent communication with all stakeholders, thereby fortifying our organization's resilience against sophisticated cyber threats.

Listing 5: Script

```
**Smaller Stories:**

The following smaller stories are intertwined with the main scenario to provide
↪  a more realistic and challenging exercise:

1. **Insider Threat:** One of your employees, who has been recently terminated,
↪  has been suspected of being involved in the APT attack. The employee had
↪  access to sensitive data and systems, and their activity patterns suggest
↪  they may have been working with the APT group. Your team must investigate
↪  this lead and determine if the employee was indeed involved in the breach.
2. **Data Exfiltration:** The APT group has been exfiltrating large amounts of
↪  sensitive data from your network, including financial information and
↪  personal employee data. Your team must identify the scope of the data
↪  breach and contain the exfiltration to minimize the damage.
3. **Ransomware Attack:** The APT group has also deployed ransomware on several
↪  systems within your network, demanding a hefty sum in exchange for the
↪  decryption key. Your team must contain the ransomware attack and determine
↪  if paying the ransom is the best course of action.
4. **Regulatory Compliance:** As a financial institution, your organization is
↪  subject to strict regulations regarding data privacy and security. Your
↪  team must ensure that all actions taken to respond to the APT attack are
↪  compliant with these regulations, including reporting the breach to
↪  regulatory authorities and notifying affected parties.
5. **Stakeholder Management:** As the incident response team, you must
↪  communicate effectively with stakeholders, including senior leadership,
↪  employees, customers, and regulatory authorities. Your team must provide
↪  regular updates on the status of the incident and take appropriate actions
↪  to address any concerns or questions from stakeholders.

These smaller stories are intertwined with the main scenario, providing a more
↪  realistic and challenging exercise for your cyber security team. Each story
↪  presents unique challenges and requires careful planning and execution to
↪  ensure the success of the incident response efforts.
A2: Based upon the previous answer from the generated snenario can you give
↪  each a particpant in the cyber security exercises scenario name, role, goal
↪  and description?
```

List 6 represents the entities generated by LLMs, the cyber security scenario unfolds within a complex infrastructure featuring state-of-the-art applications, servers, and network components tailored for the finance industry. This environment includes advanced security operations centers (SOC) with real-time monitoring systems, encrypted data storage servers running on the latest secure operating system versions, and a robust network infrastructure protected by next-gen firewalls and intrusion detection systems (IDS). Each element is meticulously configured to protect against the multifaceted threats posed by the APT group, ensuring the organization's resilience against advanced cyber attacks.

Listing 6: Entities

```
1. **Alex Chen** - Incident Responder (IR) Team Lead
    * Goal: Contain the attack and prevent further damage to the network.
    * Description: Alex is a seasoned incident responder with expertise in
↪      network security and incident response. As the IR team lead, he is
↪      responsible for coordinating the response efforts and ensuring that
↪      all necessary steps are taken to protect the organization's assets.
2. **Jane Smith** - Security Operations Center (SOC) Analyst
    * Goal: Identify and mitigate potential threats to the network.
    * Description: Jane is a skilled SOC analyst who monitors the
↪      organization's security systems and networks to identify potential
↪      threats and anomalies. She is responsible for analyzing log data,
↪      network traffic, and other security-related information to identify
↪      signs of compromise or malicious activity.
3. **John Doe** - Security Engineer
    * Goal: Implement security controls and solutions to protect the
↪      organization's assets.
    * Description: John is a knowledgeable security engineer who
↪      specializes in implementing security controls and solutions to
↪      protect the organization's assets. He is responsible for designing,
↪      deploying, and maintaining firewalls, intrusion detection systems,
↪      and other security technologies to prevent unauthorized access and
↪      malicious activity.
4. **Sarah Johnson** - Compliance Officer
    * Goal: Ensure that the organization's security practices adhere to
↪      relevant laws, regulations, and industry standards.
    * Description: Sarah is a compliance officer who is responsible for
↪      ensuring that the organization's security practices adhere to
↪      relevant laws, regulations, and industry standards. She has
↪      expertise in legal and regulatory requirements related to
↪      cybersecurity and data privacy.
5. **Bob Brown** - Network Administrator
    * Goal: Maintain the stability and integrity of the network
↪      infrastructure.
    * Description: Bob is a seasoned network administrator who is
↪      responsible for maintaining the stability and integrity of the
↪      organization's network infrastructure. He has expertise in network
↪      protocols, routing and switching, and other network-related
↪      technologies.
6. **Maria Rodriguez** - Endpoint Security Specialist
    * Goal: Protect the organization's endpoints (e.g., laptops, desktops,
↪      mobile devices) from malware and other security threats.
    * Description: Maria is an endpoint security specialist who is
↪      responsible for protecting the organization's endpoints from
↪      malware and other security threats. She has expertise in endpoint
↪      security technologies such as antivirus software, firewall
↪      configurations, and other security controls to prevent unauthorized
↪      access and malicious activity.
7. **Samantha Lee** - Threat Hunter
    * Goal: Proactively hunt for signs of advanced threats and potential
↪      security incidents.
    * Description: Samantha is a skilled threat hunter who is responsible
↪      for proactively hunting for signs of advanced threats and potential
↪      security incidents. She has expertise in threat intelligence,
↪      sandboxing, and other techniques to identify and analyze suspicious
↪      activity that may indicate a security incident.
A3: Based upon the previous answer what applications, servers, networks are
↪  present in the cyber secuirty scenario infrastructure, give as much details
↪  as possible including specifc versions?
```

List 7 is the infrastructure crafted by LLMs for this cyber security scenario underscores a layered defense strategy, incorporating firewalls, intrusion detection systems, and virtual private networks to ensure comprehensive network security. Web and database servers are fortified with advanced firewalls and encryption, emphasizing the protection of sensitive financial data. Through meticulous network segmentation and endpoint protec-

tion, the setup demonstrates a proactive stance against potential cyber threats, ensuring regulatory compliance and the safeguarding of critical assets across the organization.

Listing 7: Infrastructure

```
1. Firewall: A Cisco ASA 5506-X firewall is used to protect the network from
↪   unauthorized access and to enforce security policies. The firewall is
↪   configured with custom rules to allow only specific traffic to and from the
↪   internet.
2. Intrusion Detection/Prevention System (IDS/IPS): A Cisco IPS 4300-EX device
↪   is deployed to monitor network traffic for signs of suspicious activity and
↪   to prevent potential attacks. The IDS/IPS system is configured with
↪   signatures and protocols to detect and block known attack patterns.
3. Virtual Private Network (VPN) server: A Cisco ASA 5506-X firewall is also
↪   used as a VPN server to allow remote access to the network. The VPN server
↪   is configured with user authentication and encryption policies to secure
↪   remote access.
4. Web application firewall (WAF): An Imperva Incapsula WAF is deployed in
↪   front of web servers to protect against common web-based attacks such as
↪   SQL injection, cross-site scripting (XSS), and denial-of-service (DoS)
↪   attacks. The WAF is configured with custom rules to allow only specific
↪   traffic to the web servers.
5. Web server: A Microsoft Internet Information Services (IIS) 10.0 web server
↪   is used to host a company's public-facing website and web applications. The
↪   web server is configured with security features such as SSL/TLS encryption,
↪   user authentication, and access control policies.
6. Database server: A Microsoft SQL Server 2017 database server is used to
↪   store sensitive data such as customer information, financial records, and
↪   other confidential data. The database server is configured with security
↪   features such as encryption, access control policies, and auditing.
7. File server: A Windows Server 2016 file server is used to store and share
↪   files among users. The file server is configured with security features
↪   such as access control policies, user authentication, and encryption.
8. Network segmentation: The network is divided into different segments based
↪   on the security requirements of each department. For example, the IT
↪   department has a separate segment from the financial department to limit
↪   the spread of malware and unauthorized access.
9. Endpoint protection: A Symantec Endpoint Protection 12.1 client is installed
↪   on all company-owned laptops and desktops to protect against malware,
↪   viruses, and other security threats. The endpoint protection client is
↪   configured with policies to update signatures, scan for malware, and block
↪   unauthorized applications.
10. Security information and event management (SIEM) system: A Splunk
↪   Enterprise 8.1 SIEM system is used to monitor and analyze security-related
↪   data from various sources such as the firewall, IDS/IPS, and endpoint
↪   protection clients. The SIEM system is configured with custom alerts and
↪   dashboards to provide real-time visibility into security threats and
↪   incidents.

Please note that this is just one possible infrastructure configuration and
↪   actual configurations may vary based on specific organization requirements
↪   and environment.

A4: Based upon the previous answer How servers are connected with different
↪   networks, give detailed network diagram?
Certainly! Here's a detailed network diagram that illustrates how servers can
↪   be connected with different networks:

In this diagram, we have three servers (Server A, Server B, and Server C) that
↪   are each connected to a different network.

Server A is connected to a private local area network (LAN) that includes
↪   several client devices such as laptops, desktops, and printers. This LAN is
↪   isolated from the other networks and is used for internal communications
↪   within the organization.

Server B is connected to a wide area network (WAN) that allows it to
↪   communicate with other servers and devices located in different geographic
↪   locations. The WAN is typically used for external communications such as
↪   email, file transfers, and web browsing.

Server C is connected to a virtual private network (VPN) that allows it to
↪   securely communicate with other servers and devices over the internet. The
↪   VPN uses encryption and other security measures to protect the data being
↪   transmitted between the servers and devices.

Here's a detailed description of each server's connections:

Server A:

* Private LAN (192.168.0.0/24): This is the internal network that connects the
↪   client devices within the organization.
* Router (192.168.0.1): This router routes traffic between the private LAN and
↪   the internet.
* Internet: This is the public network that allows the server to communicate
↪   with other servers and devices located outside of the organization's
↪   private network.

Server B:

* WAN (10.0.0.0/24): This is the wide area network that connects the server to
↪   other servers and devices located in different geographic locations.
* Router (10.0.0.1): This router routes traffic between the WAN and the private
↪   LAN.
* Private LAN (192.168.0.0/24): This is the internal network that connects the
↪   client devices within the organization.

Server C:

* VPN (202.203.0.0/24): This is the virtual private network that allows the
↪   server to securely communicate with other servers and devices over the
↪   internet.
* Router (202.203.0.1): This router routes traffic between the VPN and the
↪   internet.
```

```
* Internet: This is the public network that allows the server to communicate
↪   with other servers and devices located outside of the organization's
↪   private network.
```

The following list 8 outlines a series of critical events, including malware and phishing attacks, ransomware demands, and distributed DDos disruptions, showcasing the multifaceted nature of cyber threats. Insider actions and APT highlight the complexity of securing internal and external organizational boundaries. These incidents underscore the necessity for robust security measures, including proactive software patch management and stringent access controls, to prevent unauthorized access and data breaches. The inclusion of incident response challenges further emphasizes the importance of prepared and effective management strategies to mitigate the impact of cyber incidents.

Listing 8: Events

```
1. Malware attack: The network is infected with a malware that steals sensitive
↪   data or disrupts the system's operation.
2. Phishing attack: An employee receives an email that appears to be from a
↪   legitimate source but contains a link or attachment that installs malware
↪   or steals login credentials.
3. Ransomware attack: The system is encrypted and the attackers demand payment
↪   in exchange for the decryption key.
4. Distributed Denial of Service (DDoS) attack: The system is flooded with
↪   traffic from multiple sources, overwhelming its resources and making it
↪   unavailable to users.
5. Insider threat: An employee or contractor intentionally or unintentionally
↪   causes a security breach by accessing sensitive data or systems without
↪   proper authorization.
6. Advanced Persistent Threat (APT) attack: A sophisticated and targeted attack
↪   that remains dormant for an extended period of time, allowing the attackers
↪   to gain unauthorized access to the system and steal sensitive data.
7. Security breach due to software vulnerability: An unpatched software
↪   vulnerability is exploited by an attacker to gain unauthorized access to
↪   the system or data.
8. Loss or theft of sensitive data: Sensitive data, such as customer
↪   information or intellectual property, is lost or stolen due to a security
↪   breach or accidental loss.
9. System compromise due to insufficient access controls: An attacker gains
↪   access to the system due to weak or non-existent access controls, allowing
↪   them to move laterally within the network and cause further damage.
10. Incident response failure: The organization's incident response plan is
↪   inadequate or not effectively implemented, leading to a prolonged recovery
↪   time and increased damage.
```

5.4 Scenarios Evaluation

The following section presents the ranking scores for each scenario, offering insights into their respective strengths and areas for improvement. Following are the evaluation rankings for scenario 1 and scenario 2 achieved by both LLMs and also from human evaluators (experts).

5.4.1 Evaluation for Scenarios by Expert 1

The scores provided by two cybersecurity experts for Scenarios, generated by LLM1 and LLM2, reflect their diverse perspectives shaped by their experience. In the evaluation of both scenarios, expert 1, with an extensive background of 30 years, including 25 in IT security, offers a critical view, appreciating the scenario's detail but noting a lack

of coherent design and realism, especially concerning modern organizational and technological contexts. This expert sees potential oversaturation with high-impact events and questions its usability due to reliance on licensed tools. These evaluations highlight the scenario's strengths in detail and creativity, alongside areas for improvement in realism and coherence.

The following figure 3 visualizes the evaluation of both scenarios by Experts.
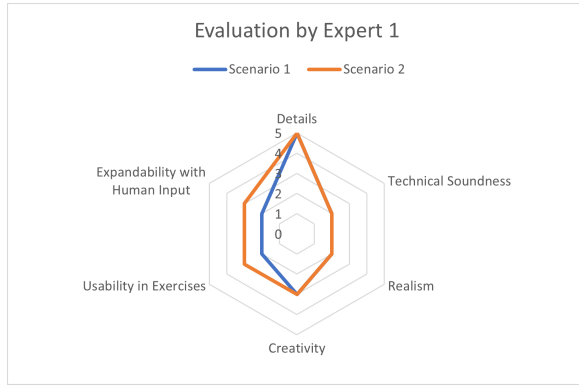

Fig. 3: Evaluation by Expert 1

5.4.2 Evaluation for Scenarios by Expert 2

Expert 2, with a more focused 5 years in cybersecurity exercises, rates the scenarios higher in creativity, usability, and expandability, suggesting a positive reception of its innovative approach and adaptability for training purposes. The evaluations for both scenarios by Expert 1 view them as highly detailed but criticize technical coherence and realism, suggesting they might not accurately reflect organizational capabilities and may have limited direct applicability in exercises without further development. In contrast, Expert 2 gives high marks across the board, particularly praising its realism, technical soundness, and creativity, indicating a belief that the scenario offers a highly realistic and engaging training tool that is both usable and expandable. This contrast highlights the subjective nature of evaluating the utility and effectiveness of cybersecurity scenarios, underscoring the value of incorporating a range of expert feedback in their development and refinement.

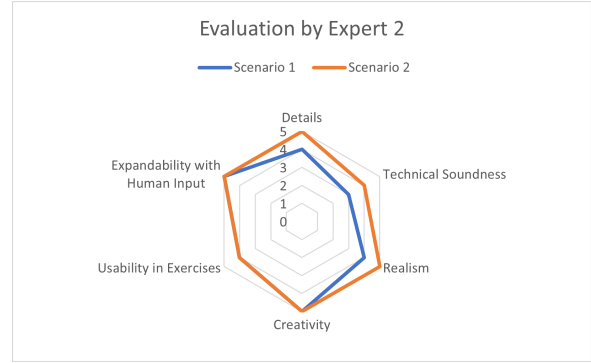The following figure 4 visualizes the evaluation of both scenarios by Experts.

Fig. 4: Evaluation by Expert 2

5.4.3 Evaluation for Scenarios 1 using LLMs

Based on the content and context of Scenario 1 generated by LLM1 and LLM2, here are the rankings across the specified criteria:

1. Details: 4 - The scenario is richly detailed, covering a wide array of cyber threats and implications for the company, showcasing a deep understanding of potential cyber security incidents.
2. Technical Soundness: 5 - The technical aspects of the scenario are sound, reflecting a realistic understanding of cybersecurity threats, defense mechanisms, and the complexities of responding to incidents.
3. Realism: 4 - This scenario is highly realistic, mirroring the multifaceted nature of cyber threats that companies face today, including insider threats, malware, and DDoS attacks, which are common in real-world cyber incidents.
4. Creativity: 3 - While the scenario is comprehensive and well-constructed, its creativity is moderate as it follows familiar patterns of cybersecurity exercises. The inclusion of diverse threats is expected but crucial for comprehensive training.
5. Usability in Exercises: 5 - Its detailed and realistic approach makes it highly usable for cybersecurity training exercises, providing a practical framework for incident response teams to practice and refine their skills.
6. Expandability with Human Inputs: 4 - The scenario allows for significant expandability with human inputs, offering a solid foundation that can be adapted and expanded based on evolving cybersecurity threats and company-specific needs.

5.4.4 Evaluation for Scenario 2 using LLMs

Based on the evaluation framework previously discussed for Scenario 1, here are the anticipated scores for Scenario 2, considering a general understanding of such cybersecurity scenarios:

1. Details: 5 - Assuming Scenario 2 is as intricately designed as Scenario 1, it should offer a comprehensive exploration of cybersecurity issues relevant to its context, likely providing extensive detail on attack vectors, defensive strategies, and the roles involved in mitigating cyber threats.

2. Technical Soundness: 4 - Given the necessity for accuracy in depicting cybersecurity challenges, especially in a potentially complex sector like finance, the scenario is expected to exhibit a high level of technical soundness, showcasing an understanding of both existing and emerging cyber threats.

3. Realism: 4 - The finance sector's unique vulnerabilities to cyber-attacks, including sophisticated threats like APTs, phishing, and insider threats, suggest that Scenario 2 would prioritize realism, reflecting the nuanced and high-stakes environment of financial cybersecurity.

4. Creativity: 3 - Creativity in such scenarios is often demonstrated through the innovative presentation of threats and problem-solving approaches. However, the balance between creativity and realism is crucial to maintain the scenario's educational value without straying too far from plausible threats.

5. Usability in Exercises: 4 - The practical application of the scenario in training exercises is vital. A scenario that effectively simulates real-world challenges while allowing participants to practice their response strategies would score highly in usability.

6. Expandability with Human Inputs: 4 - The ability to adapt and expand the scenario based on new information, threats, and organizational changes is essential for keeping cybersecurity exercises relevant and engaging, suggesting a high score for expandability.

The following figure 5 visualizes the evaluation of both scenarios by LLM1 and LLM2.

6 Conclusion and Future Work

This study demonstrates that using LLMs to create cybersecurity exercise scenarios is both innovative and effective for training. The methodology
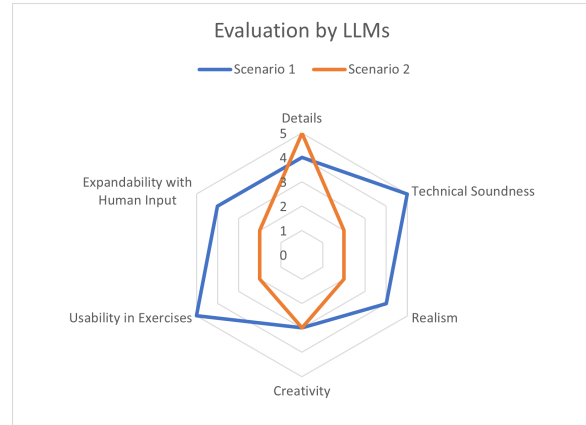

Fig. 5: Evaluation by LLMs

effectively prepares security professionals to tackle a broad spectrum of cyber threats by simulating real-world challenges, encompassing both familiar and emerging cybersecurity issues. The integration of RAG adds depth and relevance to the scenarios, effectively utilizing LLM-generated content for more authentic simulations. Despite the critical scores in scenario 1, the overall quantitative measures provided us with good results. Future work should focus on further refining the integration of LLMs into cybersecurity training programs, particularly by expanding parameter coverage (LLMs with larger parameters) for more detailed scenarios and implementing generative configurations for increased creativity which will also assess the real-world impact of these scenarios on enhancing cybersecurity skills and readiness.

Declarations

1. Competing Interests
   - The authors declare that they have no competing interests.
2. Authors' Contribution
   - **Muhammad Mudassar Yamin:** Conceptualization, Visualization, Supervision, Project Management, Funding Acquisition, Research Conduct, Validation.
   - **Ehtesham Hashmi:** Literature Analysis, Writing Original Draft, Investigation.
   - **Mohib Ullah:** Review , Investigation.
   - **Basel Katt:** Review , Investigation.
3. Ethical and Informed Consent for Data Used
   - Not Applicable.
4. Data Availability and Access.
   - Not Applicable.

References

1. T. Shejin and K. Sudheer, "A review on major cyber threats and recommended counter measures,"
2. A. Sharma, B. B. Gupta, A. K. Singh, and V. Saraswat, "Advanced persistent threats (apt): evolution, anatomy, attribution and countermeasures," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–27, 2023.
3. M. F. Safitra, M. Lubis, and H. Fakhrurroja, "Counterattacking cyber threats: A framework for the future of cybersecurity," *Sustainability*, vol. 15, no. 18, p. 13369, 2023.
4. S. Saeed, S. A. Altamimi, N. A. Alkayyal, E. Alshehri, and D. A. Alabbad, "Digital transformation and cybersecurity challenges for businesses resilience: Issues and recommendations," *Sensors*, vol. 23, no. 15, p. 6666, 2023.
5. J. N. Al-Karaki, M. Omar, A. Gawanmeh, and A. Jones, "Advancing cybersecurity education and training: Practical case study of running capture the flag (ctf) on the metaverse vs. physical settings," in *2023 International Conference on Intelligent Metaverse Technologies Applications (iMETA)*, pp. 1–7, 2023.
6. T. Awojana and T.-S. Chou, "Overview of learning cybersecurity through game based systems," 2019.
7. M. Mares, R. Chytilek, Z. Špačková, J. Drmola, L. Hrbková, P. Mlejnková, and M. Tóth, "Assessment of performance during cybersecurity tabletop exercises," *Security Journal*, pp. 1–24, 07 2023.
8. R. Beuran, D. Tang, C. Pham, K.-i. Chinen, Y. Tan, and Y. Shinoda, "Integrated framework for hands-on cybersecurity training: Cytrone," *Computers & Security*, vol. 78, pp. 43–59, 2018.
9. C. Chindrus and C.-F. Caruntu, "Securing the network: A red and blue cybersecurity competition case study," *Information*, vol. 14, no. 11, p. 587, 2023.
10. M. Katsantonis, P. Fouliras, and I. Mavridis, "Conceptual analysis of cyber security education based on live competitions," in *2017 IEEE Global Engineering Education Conference (EDUCON)*, pp. 771–779, IEEE, 2017.
11. M. M. Yamin, B. Katt, and M. Nowostawski, "Serious games as a tool to model attack and defense scenarios for cyber-security exercises," *Computers Security*, vol. 110, p. 102450, 2021.
12. R. Nakata and A. Otsuka, "Cyexec*: Automatic generation of randomized cyber range scenarios.," in *ICISSP*, pp. 226–236, 2021.
13. K. Zenitani, "Attack graph analysis: an explanatory guide," *Computers & Security*, vol. 126, p. 103081, 2023.
14. K. Zenitani, "A scalable algorithm for network reachability analysis with cyclic attack graphs," *Journal of Computer Security*, vol. 31, no. 1, pp. 29–55, 2023.
15. S. Dalal, P. Manoharan, U. K. Lilhore, B. Seth, D. Mohammed alsekait, S. Simaiya, M. Hamdi, and K. Raahemifar, "Extremely boosted neural network for more accurate multi-stage cyber attack prediction in cloud computing environment," *Journal of Cloud Computing*, vol. 12, no. 1, p. 14, 2023.
16. M. M. Yamin and B. Katt, "Modeling and executing cyber security exercise scenarios in cyber ranges," *Computers & Security*, vol. 116, p. 102635, 2022.
17. M. M. Yamin, B. Katt, and V. Gkioulos, "Cyber ranges and security testbeds: Scenarios, functions, tools and architecture," *Computers & Security*, vol. 88, p. 101636, 2020.
18. M. M. Yamin and B. Katt, "Use of cyber attack and defense agents in cyber ranges: A case study," *Computers & Security*, vol. 122, p. 102892, 2022.
19. F. L. Færøy, M. M. Yamin, A. Shukla, and B. Katt, "Automatic verification and execution of cyber attack on iot devices," *Sensors*, vol. 23, no. 2, p. 733, 2023.
20. A. Zacharis and C. Patsakis, "Aicef: an ai-assisted cyber exercise content generation framework using named entity recognition," *International Journal of Information Security*, pp. 1–22, 2023.
21. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, *et al.*, "Advances in neural information processing systems," *Attention is All you Need*, 2017.
22. B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth, "Recent advances in natural language processing via large pre-trained language models: A survey," *ACM Computing Surveys*, vol. 56, no. 2, pp. 1–40, 2023.
23. S. Gheewala, S. Xu, S. Yeom, and S. Maqsood, "Exploiting deep transformer models in textual review based recommender systems," *Expert Systems with Applications*, vol. 235, p. 121120, 2024.
24. D. Oralbekova, O. Mamyrbayev, M. Othman, D. Kassymova, and K. Mukhsina, "Contemporary approaches in evolving language models," *Applied Sciences*, vol. 13, no. 23, p. 12901, 2023.
25. S. K. Freire, C. Wang, and E. Niforatos, "Chatbots in knowledge-intensive contexts: Comparing intent and llm-based systems," *arXiv preprint arXiv:2402.04955*, 2024.
26. J. Lipenkova, "Overcoming the limitations of large language models how to enhance llms with human-like cognitive skills," 2023.
27. S. Moore, R. Tong, A. Singh, Z. Liu, X. Hu, Y. Lu, J. Liang, C. Cao, H. Khosravi, P. Denny, *et al.*, "Empowering education with llms-the next-gen interface and content generation," in *International Conference on Artificial Intelligence in Education*, pp. 32–37, Springer, 2023.
28. X. Liu, Z. Zhu, H. Liu, Y. Yuan, M. Cui, Q. Huang, J. Liang, Y. Cao, Q. Kong, M. D. Plumbley, *et al.*, "Wavjourney: Compositional audio creation with large language models," *arXiv preprint arXiv:2307.14335*, 2023.
29. S. Kamnis, "Generative pre-trained transformers (gpt) for surface engineering," *Surface and Coatings Technology*, p. 129680, 2023.
30. H. Huang, S. Wu, X. Liang, B. Wang, Y. Shi, P. Wu, M. Yang, and T. Zhao, "Towards making the most of llm for translation quality estimation," in *CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 375–386, Springer, 2023.
31. Z. CHen, L. Cao, S. Madden, J. Fan, N. Tang, Z. Gu, Z. Shang, C. Liu, M. Cafarella, and T. Kraska, "Seed: Simple, efficient, and effective data management via large language models," *arXiv preprint arXiv:2310.00749*, 2023.
32. L. Hu, Z. Liu, Z. Zhao, L. Hou, L. Nie, and J. Li, "A survey of knowledge enhanced pre-trained language models," *IEEE Transactions on Knowledge and Data Engineering*, 2023.
33. H. Jiang, Q. Wu, C.-Y. Lin, Y. Yang, and L. Qiu, "Llmlingua: Compressing prompts for accelerated inference of large language models," *arXiv preprint arXiv:2310.05736*, 2023.

34. L. Li, Y. Zhang, and L. Chen, "Prompt distillation for efficient llm-based recommendation," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 1348–1357, 2023.

35. Z. Yuan, J. Liu, Q. Zi, M. Liu, X. Peng, and Y. Lou, "Evaluating instruction-tuned large language models on code comprehension and generation," *arXiv preprint arXiv:2308.01240*, 2023.

36. Z. Chu, H. Hao, X. Ouyang, S. Wang, Y. Wang, Y. Shen, J. Gu, Q. Cui, L. Li, S. Xue, *et al.*, "Leveraging large language models for pre-trained recommender systems," *arXiv preprint arXiv:2308.10837*, 2023.

37. V. Rawte, A. Sheth, and A. Das, "A survey of hallucination in large foundation models," *arXiv preprint arXiv:2309.05922*, 2023.

38. J.-Y. Yao, K.-P. Ning, Z.-H. Liu, M.-N. Ning, and L. Yuan, "Llm lies: Hallucinations are not bugs, but features as adversarial examples," 2023.

39. P. Manakul, A. Liusie, and M. J. F. Gales, "Self-checkgpt: Zero-resource black-box hallucination detection for generative large language models," 2023.

40. F. Liu, K. Lin, L. Li, J. Wang, Y. Yacoob, and L. Wang, "Mitigating hallucination in large multimodal models via robust instruction tuning," *arXiv preprint arXiv:2306.14565*, vol. 1, no. 2, p. 9, 2023.

41. A. Martino, M. Iannelli, and C. Truong, "Knowledge injection to counter large language model (llm) hallucination," in *European Semantic Web Conference*, pp. 182–185, Springer, 2023.

42. Z. Ji, T. Yu, Y. Xu, N. Lee, E. Ishii, and P. Fung, "Towards mitigating llm hallucination via self reflection," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1827–1843, 2023.

43. W. Sun, Z. Shi, S. Gao, P. Ren, M. de Rijke, and Z. Ren, "Contrastive learning reduces hallucination in conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 13618–13626, 2023.

44. Y. Zhang and X. Chen, "Enhancing simplified chinese poetry comprehension in llama-7b: A novel approach to mimic mixture of experts effect," 2023.

45. S. Curran, S. Lansley, and O. Bethell, "Hallucination is the last thing you need," *arXiv preprint arXiv:2306.11520*, 2023.

46. L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *arXiv preprint arXiv:2311.05232*, 2023.

47. S. D. L. Innovations, "eluna: A co-design framework for narrative digital game-based learning that support steam," 2022.

48. M. Platt and D. Platt, "Effectiveness of generative artificial intelligence for scientific content analysis," in *17th International Conference on Application of Information and Communication Technologies*, IEEE, 2023.

49. G. Chalvatzaki, A. Younes, D. Nandha, A. T. Le, L. F. Ribeiro, and I. Gurevych, "Learning to reason over scene graphs: a case study of finetuning gpt-2 into a robot language model for grounded task planning," *Frontiers in Robotics and AI*, vol. 10, 2023.

50. A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," 2020.

51. J. Hu, Q. Zhang, and H. Yin, "Augmenting grey-box fuzzing with generative ai," *arXiv preprint arXiv:2306.06782*, 2023.