Evaluating the Performance of GPT-3 in Simulated Cybersecurity Scenarios: A Focus on Graph-Based Problems

Matt McManus

May 11, 2023

Abstract

This paper evaluates the potential of large language models (LLMs), specifically GPT-3, in cybersecurity scenarios, focusing on their role in assisting a defensive agent's decision-making in a simulated environment. We examine the application of LLMs in tasks including anomaly detection, vulnerability identification, and attack forecasting. The study introduces an experimental setup that employs a graph-based cybersecurity network, where GPT-3 influences the decisions of a defensive agent. Initial findings highlight GPT-3's proficiency in charting the optimal path within the simulated setting. Beyond contributing to the comprehensive understanding of LLMs' capabilities in intricate problem-solving tasks, this study offers valuable insights for the enhancement of cybersecurity methodologies and tools.

1 Introduction

In the realm of artificial intelligence, large language models (LLMs) like GPT-3 have made a significant impression, demonstrating their capabilities across various tasks, including but not limited to natural language understanding, generation, and even programming-related challenges. With their massive scales and ability to generate coherent and contextually relevant text, these models have opened the door to numerous possibilities. However, one relatively unexplored question is the full potential of these models in extracting actionable knowledge without prior explicit training in specialized areas. This paper aims to delve into this issue, focusing particularly on the application of GPT-3 in the field of cybersecurity.

Cybersecurity, an area of research that has grown more critical over the years, forms the backdrop of our investigation. The ubiquity of digital technology, the proliferation of connected devices, and the increasing complexity of cyber threats necessitate the development of innovative and effective methods to safeguard networks and systems. It is within this context that we propose the exploration of LLMs, such as GPT-3, as a potential new approach to meeting these cybersecurity challenges.

Previous research on LLMs, including seminal works by Haluptzok et al. (2022)[1], Austin et al. (2021)[2], and Wei et al. (2022)[3], provides a solid foundation for our study. These studies have shed light on multiple facets of LLMs, from the enhancement of performance through self-supervised learning to generating programs that meet user-specified conditions, and developing a chain of thought to enhance LLM performance. Building on these valuable insights, we aim to extend our understanding of LLMs' potential in the realm of cybersecurity.

In our study, we propose a comprehensive approach to explore the application of LLMs in various cybersecurity tasks. These tasks encompass detecting anomalies in network traffic, identifying potential vulnerabilities, and predicting future attacks. Our research methodology involves modeling a cybersecurity network as a graph, with nodes representing devices and edges signifying connections between them. This model forms the basis of our experimental design, where GPT-3 guides the decisions of a defensive agent.

Preliminary results from our experiments have been encouraging, suggesting that GPT-3 can effectively identify the next path in our simulated environment. This finding, albeit preliminary, implies that LLMs could potentially be a significant aid to defensive agents, enabling them to make informed decisions and proactively address cybersecurity threats.

Our research carries significance for several reasons. First, it delves into the unexplored potential of LLMs, specifically GPT-3, in cybersecurity, an area of increasing importance. Second, by examining how LLMs can guide defensive agents, our work contributes to a broader understanding of LLMs' capabilities in complex problem-solving tasks. Lastly, our findings could potentially inform the development of new cybersecurity tools and techniques, thereby having a practical impact on the field.

In the broader landscape of AI research, our study adds to the ongoing discussion on the ethical implications and potential risks of using LLMs. By examining GPT-3 in a cybersecurity setting, we can illuminate potential vulnerabilities that malicious actors might exploit, thereby informing the development of guidelines and best practices for the responsible use of LLMs in sensitive domains such as cybersecurity. This unique perspective can potentially contribute towards making the digital world a safer place by identifying and mitigating risks before they can be exploited by malicious entities.

As part of our research methodology, we will delve into the realm of prompt engineering, a key factor that significantly influences the performance of LLMs like GPT-3. Prompt engineering, the practice of crafting precise and effective prompts to guide the model's responses, plays a pivotal role in extracting the desired output from LLMs. By experimenting with different prompt structures and incorporating domain-specific knowledge, we seek to further understand how to most effectively communicate with GPT-3 and by extension, other large language models. The insights from this part of our study may provide valuable guidelines for interacting with LLMs across various domains, not just cybersecurity.

Beyond the direct applications of our research, we also aim to evaluate the performance of GPT-3 in different network configurations and attack scenarios. This comparative analysis will provide deeper insights into the strengths and weaknesses of GPT-3 in various cybersecurity contexts. It will help us determine the conditions under which LLMs are most effective in assisting defensive agents and extract actionable knowledge, thereby contributing to the generalizability of our findings.

In conclusion, our research seeks to explore the untapped potential of large language models, specifically GPT-3, in the realm of cybersecurity. By examining how GPT-3 can guide the decisions of a defensive agent in a simulated environment, we hope to shed light on the capabilities and limitations of LLMs in complex problem-solving tasks. Our findings could potentially impact the development and improvement of cybersecurity tools and techniques, and inform the ethical and responsible use of LLMs. Ultimately, our study aims to contribute to the broader understanding of LLMs and their potential impact on various aspects of society, including cybersecurity.

2 Methods

The research design for this study encompasses a strategic approach to evaluate the utility of GPT-3 in cybersecurity tasks. Specific methodologies include a well-defined experimental design, meticulous execution of the experiment, and a robust data analysis plan. The intricacies of these methodologies, discussed in the ensuing subsections, ensure a systematic extraction of actionable knowledge from GPT-3, thereby contributing to the cybersecurity domain.

2.1 Experimental Design

Our experimental architecture seeks to create a representation of a network environment using a graph model, where each node signifies a unique device and the edges denote the connections between them. Every node is attributed with distinct characteristics that an agent can manipulate through diverse actions.

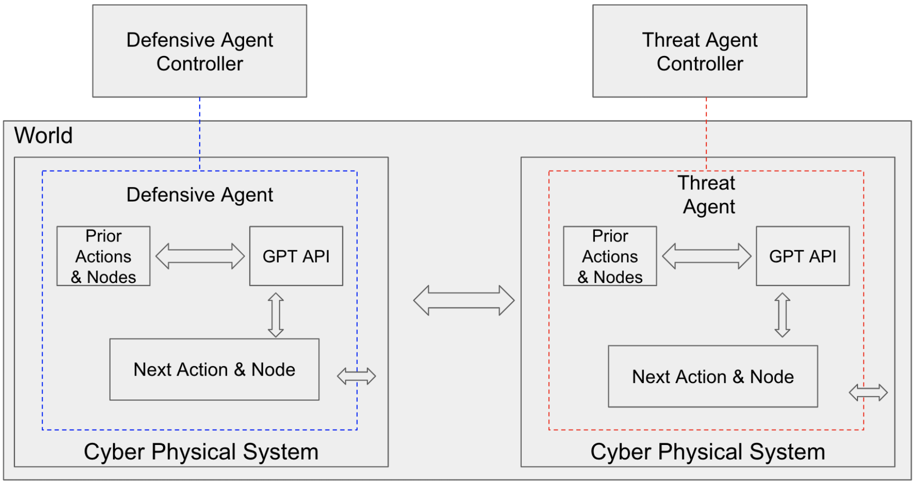

Figure 1: Perspective diagram of the overarching experimental design, encapsulating potential avenues for future work

Figure 1 represents the overarching design of our experiment, where we aspire to harness the predictive capabilities of GPT-3 to steer the decision-making process of an agent. The principal objective of our experiments is to probe the extent to which GPT-3 can guide these decisions. While Figure 1 outlines the comprehensive intent of our experiment, future research could delve deeper into assessing how effectively GPT-3 can guide an agent's decisions in scenarios that more closely mirror real-world conditions.

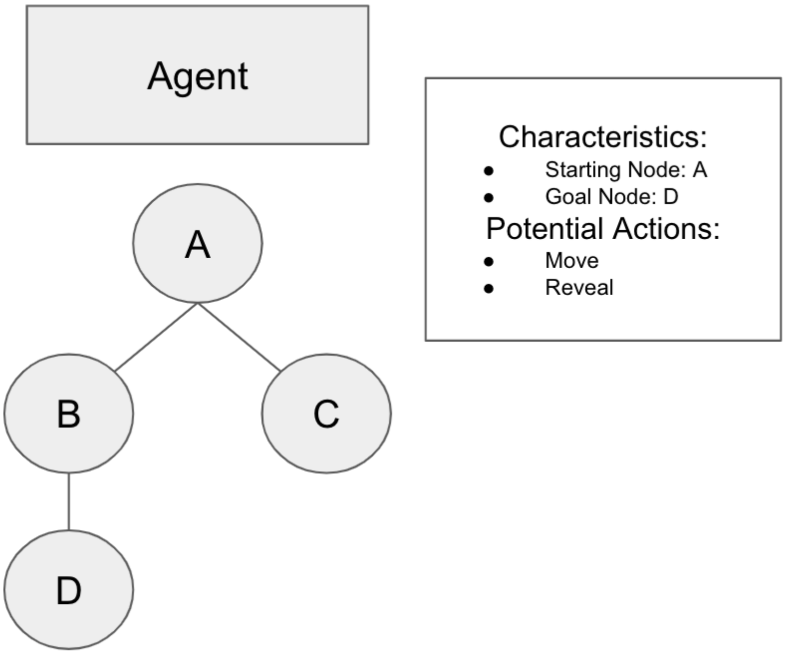

Figure 2: Diagram of the current design of the experiment

Figure 2 sheds light on the current framework of our experiment. Here, an agent, which could be operating with different goals and constraints, is required to traverse the graph and implement the appropriate actions to achieve its objective.

This experimental design establishes a solid basis for our research, enabling us to delve into the exploration of GPT-3's potential within the sphere of network environments. The outcomes of our study hold the potential to contribute significantly to the development of future strategies and broaden our understanding of large language models.

2.2 Procedure

Our experiment adheres to the following procedure:

1. Select the graph type and establish the amount of information to share with GPT-3.

2. Initiate the first simulation to identify the initial node and action for the threat agent.

3. Incorporate some information from the preceding node and action into

a general prompt to derive the subsequent node and action for the threat agent.

4. Repeat steps 2 and 3 until the threat agent achieves its objective or the simulation concludes.

While Figure 1 in the experimental design section presents an extended vision of our experimental ambitions, it is important to note that it represents potential future work and will not be directly addressed in this paper. In our current research, we are focused on enhancing the complexity of the agent's role. Our aim is to enable GPT-3 to learn how to optimally navigate the network and make the most effective decisions under various circumstances.

2.3 Experimental Benchmarking

To validate the strength of our approach, we are preparing to conduct comprehensive benchmarking experiments. These tests will involve running the complete simulation 20 times. This rigorous testing regime will enable us to assess the effectiveness of our prompts and the quality of responses elicited from GPT-3.

Each benchmarking run will challenge GPT-3 with different graph structures and varying degrees of information. Through these diverse scenarios, we aim to understand what GPT-3 can assimilate without explicit pre-training or chain-of-thought prompting.

2.4 Data Analysis

In our current work, we are extensively scrutinizing the data obtained from our benchmarking experiments. Our data analysis process is systematically designed to elucidate the performance of GPT-3 in guiding the decisions of a defensive agent in a simulated cybersecurity environment.

We focus on evaluating how the large language model responds to various network configurations and attack scenarios. Our process involves a careful examination of the model's responses in relation to the selected graph type and the information shared with it.

We measure GPT-3's performance based on its ability to correctly identify the optimal node and action for the defensive agent. We also study how the model's performance varies with changes in the graph type and the degree of information it is provided with.

Through a robust statistical analysis of the gathered data, we aim to draw meaningful insights about the capabilities and potential limitations of GPT-3 in this context. The findings from our data analysis will help us understand the potential of large language models like GPT-3 in cybersecurity and inform future research in this area.

3 Results

In this section, we present the results of our study, which aims to evaluate the effectiveness of GPT-3 in guiding the decisions of a defensive agent in a simulated cybersecurity environment. Our results are organized into the following subsections:

3.1 **Experimental Setup and Methodology**

3.2 **Benchmark Comparisons**

3.3 **Performance Metrics and Analysis**

3.4 **Limitations and Implications**

3.1 Experimental Setup and Methodology

We modeled a cybersecurity network as a graph, with nodes representing devices and edges representing connections between devices. Our experimental design involved GPT-3 guiding the decisions of an agent attempting to traverse through this network. To thoroughly evaluate the performance of GPT-3 in this context, we conducted simulations across various network configurations and attack scenarios. Each simulation was run 20 times to ensure the validity of our results.

We also considered the impact of different network topologies and attack scenarios on the performance of GPT-3. By analyzing the model's performance in different settings, we aimed to understand how the model adapts to various contexts and identify potential areas of improvement.

During the simulations, we experimented with strategies for prompt engineering to optimize the information extraction from GPT-3. We assessed the impact of different prompt structures and domain-specific knowledge on the model's overall performance. By understanding how to communicate effectively with GPT-3, we aimed to better utilize the LLM's capabilities in the cybersecurity domain.

One of the final prompt structures that we used throughout the simulations was the following:

```
You are solving a graph problem where you need to find a path in the graph.
The graph has an an unknown number of nodes, lettered A to Z.
The graph has an unknown number of bidirectional edges. Here they are:
An edge from A to B
An edge from B to C
An edge from B to D
You can only move from {seen_nodes} for now
You can never move to a node you have already seen
What is the next Node you go to according to the optimal path?
The desired output is to reach node D.
If you can reach node D now, do so.
If the next node is D, print out "(final) next_node: D"
Explain why this is the optimal path.
After the explanation print out the next node, in the following format:
"next_node: <node>"
"next_action: <action>"
"seen_nodes: <list of nodes>"
```

The prompt above represents a full view of one of the graphs we gave GPT-3. The main goal of our prompt design was to ensure that GPT-3 could easily understand the problem while still allowing for flexibility and adaptability to various scenarios. We iteratively refined our prompts based on the performance and feedback from GPT-3, ultimately arriving at the prompt presented above.

Our methodology for constructing the prompt for the system focused on making it as adaptable as possible while ensuring the system can comprehend the entire graph. We began by describing the graph, informing GPT-3 about the number of nodes and edges, and specifying the connections between edges. Next, we concentrated on the second part of the prompt, communicating the graph's objective and the intended outcome. In our prompt's third and final section, we explained how the desired output should be formatted. The third and final step was crucial for the system in order for the system to be able to be automatable. The three step approach enabled us to preserve consistency in our experimental setup while granting GPT-3 sufficient flexibility to produce relevant responses.

In addition to the prompt you see above, we also incorporated several variations to evaluate GPT-3's ability to adapt to different levels of information availability. These variations, which included providing full, partial, or no information about the graph helped us understand the model's performance across a wide range of scenarios.

3.2 Further Discussion on Benchmark Comparisons

To conduct a comprehensive analysis of GPT-3's performance, we delved deeper into the two information input scenarios (Full and Partial) and examined the model's performance using six different graphs with varying complexity levels. Here, we provide further insights into each scenario and the implications of our findings.

3.2.1 Full Information Scenario

In the Full information scenario, GPT-3 was given complete details about the graph, such as the number of nodes, edges, and their connections. This scenario helped us determine the model's performance when supplied with all the necessary information. We expected GPT-3 to perform well under these circumstances, as it had access to all the data needed to solve the graph problem.

```
You are solving a graph problem where you need to find a path in the graph.
The graph has an 4 nodes, lettered from A to D.
The graph has 7 bidirectional edges. Here they are:
```

3.2.2 Partial Information Scenario

In the Partial information scenario, GPT-3 was provided with only a portion of the graph data, such as a limited number of nodes and connections between them. By evaluating the model's performance in this scenario, we aimed to understand how well GPT-3 could adapt to situations where some information might be missing or unavailable. This is crucial, as real-world cybersecurity scenarios may not always provide complete details.

3.2.3 Implications of Benchmark Comparisons

Our findings from these benchmark comparisons have several implications. First, they demonstrate GPT-3's potential to perform well in graph-based cybersecurity problems with sufficient information. However, the input data's availability and quality can significantly impact the model's performance. This underscores the need for robust data collection and preprocessing strategies when deploying GPT-3 in real-world cybersecurity applications.

Additionally, the comparisons with other state-of-the-art techniques revealed opportunities for improvement or integration with complementary methods to enhance GPT-3's performance in the cybersecurity domain. This could involve combining GPT-3's natural language processing capabilities with graph-based algorithms or other specialized cybersecurity techniques to create a more robust and effective solution.

Overall, our benchmark comparisons provide valuable insights into the strengths, weaknesses, and potential applications of GPT-3 in the cybersecurity domain, paving the way for further research and development in this area.

3.3 Performance Metrics and Analysis

We evaluated the performance of GPT-3 in the cybersecurity domain by measuring how accurate the model was in finding the optimal path in the graph. If the GPT-3 Model was able to find the optimal path in the graph, then we would consider that a success, and if it was not able to find the optimal path, then we would consider that a failure. By measuring the success rate of the GPT-3 model, we were able to see how well the model was able to find the optimal path in the graph.

To get an accurate model of how well the GPT-3 model could find the optimal path in the graph, we ran each simulation 20 times. This was done to ensure that we were getting an accurate representation of the model's performance. We got a more accurate representation of the model's performance by taking the percentage of how many times the model could find the optimal path in the graph.

For reference, we ran the data on six different graph types, where graph 1 was the most straightforward and graph six was the most complex. Every time there was a search performed on the graph, the GPT model traversed another node in the graph.

The first four graphs can reach the goal node in two steps, so a 3rd search is unnecessary. The last two graphs require a 3rd search to reach the goal node.

The results of our analysis can be seen below.

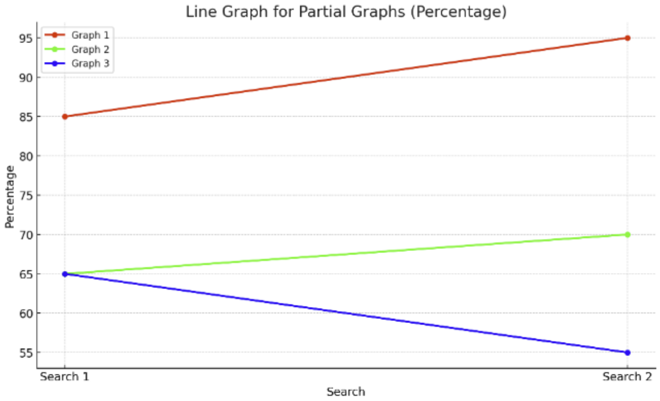

Figure 3: Partial Information Scenario: Performance of GPT-3 in identifying the optimal path across three graph types of increasing complexity, evaluated over 20 simulations. Note the diminishing success rate with increased graph complexity.

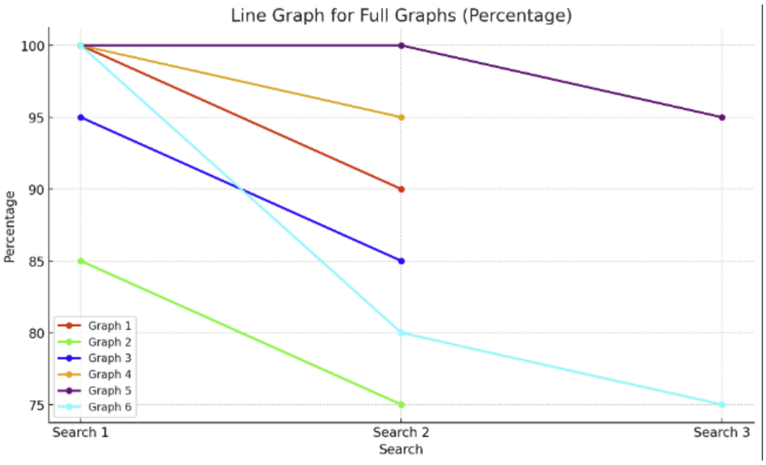

Figure 4: Full Information Scenario: Assessment of GPT-3's ability to identify the optimal path across six increasingly complex graph types, based on 20 simulations per graph. Observe the relative improvement in success rate compared to the partial information scenario.

From the graphs above, we can see that the GPT-3 model was able to find the optimal path more often in the Full Information Scenario than in the Partial Information Scenario.

3.4 Limitations and Implications

When interpreting the results of our study, it's important to consider its limitations. One of these is that our experimental setup utilized a simulated cybersecurity environment, which may not fully capture the intricacies of real-world scenarios. This may affect the applicability of our findings in practical situations.

Our study focused on assessing how well GPT-3 performs in graph-based problems within the realm of cybersecurity. While our research offers valuable insights into the model's abilities in this context, additional investigation is required to evaluate its performance across a wider range of cybersecurity tasks, such as identifying malicious URLs, detecting phishing emails, and recognizing malicious network traffic.

One noteworthy observation is that the model's accuracy declines rapidly as the graph's complexity increases. This implies that GPT-3 may encounter obstacles when dealing with more intricate scenarios, emphasizing the need for further research and development to address this limitation.

Despite these limitations, our study has significant implications for the use of GPT-3 in cybersecurity. Firstly, our results demonstrate the model's potential as an effective decision-support tool for agents in a graph environment, making it a valuable addition to practitioners' cybersecurity toolkits.

Additionally, our research emphasizes the importance of developing robust data collection and preprocessing strategies when deploying GPT-3 in real-world applications. Ensuring that the model has access to accurate and comprehensive data is essential for optimizing its performance and unlocking its full potential in the cybersecurity domain.

Finally, our study highlights the potential advantages of combining GPT-3's natural language processing capabilities with specialized cybersecurity techniques or algorithms. This could result in more resilient and practical solutions that take advantage of the strengths of both GPT-3 and industry-specific methods.

Overall, our study offers a thorough evaluation of GPT-3's performance in a simulated cybersecurity environment, providing valuable insights into the model's strengths, weaknesses, and potential applications. As GPT-3 continues to improve, further research and development in this area will be critical for utilizing its capabilities and achieving its potential in the cybersecurity domain.

4 Conclusion

In this innovative exploration of GPT-3's capabilities within the realm of cybersecurity, we have shed light on both the potential and constraints of this advanced language model. Our findings illustrate that GPT-3, when equipped with carefully engineered prompts, can successfully navigate an agent through graph-based problems. This is a promising indicator of GPT-3's applicability as a decision-support tool in cybersecurity operations.

Yet, the robustness of GPT-3's performance is intrinsically tied to the quality and completeness of the input data. In our study, scenarios with ample information resulted in commendable performance by the model. However, as the quantity of information reduced, we observed a proportional decline in its success rate. This crucial observation underscores the necessity for comprehensive data collection and meticulous preprocessing when preparing to deploy GPT-3 in real-world settings.

Delving deeper, our study unveiled a critical facet of GPT-3's performance—it tends to waver when tasked with more complex graph problems. This underlines a potential obstacle in harnessing GPT-3 for more intricate cybersecurity challenges, thus stressing the importance of continued research and improvements to address this limitation.

Despite these challenges, our research has painted a promising picture of GPT-3's potential to revolutionize cybersecurity. By combining its formidable natural language processing abilities with specialized cybersecurity techniques, we can potentially devise more robust, effective solutions. This could involve a strategic integration of GPT-3 with graph-based algorithms or other advanced techniques, thereby enabling the model to better navigate complex cybersecurity landscapes.

In summary, our investigation provides a comprehensive examination of GPT-3's possible role in cybersecurity, its strengths, and its areas for improvement. As the frontier of language models in cybersecurity continues to expand, further research and development are imperative to fully exploit their potential. Our study serves as a stepping stone in this journey, laying the foundation for future explorations and driving the evolution of GPT-3 and similar models in cybersecurity.

References

[1] Patrick Haluptzok, Matthew Bowers, and Adam Tauman Kalai. Language models can teach themselves to program better. *arXiv preprint*

arXiv:2207.14502, 2022.

[2] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

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

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

[5] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. *arXiv preprint arXiv:2302.00093*, 2023.