www.tno.nl
bram.poppink@tno.nl
+31611263097

Datum
21 december 2023

The future of cyber attacks with Large Language Models

A technology exploration

Technology Exploration – The future of cyber attacks with Large Language Models

To	Dion Koeze (NCSC-NL)
From	Daan Opheikens, Arthur Melissen, Piotr Stachyra, Bram Poppink (TNO)
TNO reference	TNO 2024 M10296
Title	Technology exploration - The future of cyber attacks with Large Language Models
External classification	TLP: CLEAR
Type	Memorandum

Number of pages	27
Number of appendices	1
Programme name	NCSC 2023 cyberweerbaarheid

TNO Classification	TNO Publiek | ONGERUBRICEERD Releasable to the public
Title	TNO Publiek
Management summary	TNO Publiek
Report text	TNO Publiek
Project name	De toekomst van cyber aanvallen met LLMs
Project number	060.53784/01.07

This memorandum contains the results of a project that was part of the long-term research collaboration between NCSC-NL and TNO.

Management summary

Recent developments in the field of Large Language Models (LLMs) create unprecedented opportunities to automate tasks that were previously out of reach. There is already some evidence that cyber criminals can use Large Language Models to optimize their activities and further explore new attack vectors.

This memorandum is the result of a technology exploration project, conducted in collaboration between the NCSC-NL and TNO. The main goal of this exploration was to identify how we can measure and monitor the impact that LLMs have on the cyber threat landscape now up until the midterm future (3 to 5 years from now).

Therefore, in this memorandum we will answer the following research questions:

1. What are recent developments in the usage of LLMs for cyber attacks by cybercriminals (explicitly not nation-states)?
2. Which implementations of LLMs are cybercriminals most likely to use (e.g. large remote models versus locally trained models)?
3. For what activities, and in what ways, can LLMs be of most use to cybercriminals?
4. Which variables determine whether an LLM can be used for a cyber attack?
5. a. Is it possible to develop a metric that expresses the improvement of using LLMs on existing cyber attacks?
5. b. What are potential indicators of change and trigger events to watch in order to re-evaluate the threat that LLMs pose?

Methodology

We began the exploratory research with a general literature survey on current trends in LLM technology, and continued this survey with the analysis of scientific literature that showcases the potential usage of LLMs for cyber attacks in various contexts. In this phase, searching for literature was done based on common sense using the expertise within the team (knowledge on machine learning, NLP and LLMs, and cyber security expertise, both defensive and offensive)

Subsequently, we used the MITRE ATT&CK framework[1] to structure the literature review. This framework is used as a reference for the currently (publicly) known attack tactics and techniques. We map the literature on the framework both *vertically*, i.e., in one phase of the kill-chain (Initial Access), and *horizontally*, i.e., over all phases in the kill-chain (from Reconnaissance to Impact). This mapping helps to identify for what tactics and techniques current literature confirms the usage of LLMs is actually useful. Furthermore, it also helps identifying blind spots, i.e., the tactics and techniques for which LLMs might be or become very useful but has not yet been proven in scientific literature.

Lastly, we explored how the NCSC-NL can monitor the impact of LLMs on the threat landscape in the midterm future (read: 3 to 5 years from now). We briefly discuss the limitations of currently available metrics and move on to a *Signposts of Change* model. This model has the potential to be used for tracking the impact of LLMs on the threat landscape over time.

Main findings

Large Language Models present a wide spectrum of abilities ranging from those explicitly taken from the training set (e.g. generating English language or programming language), to emergent abilities originating from more obscure patterns and model's scale (e.g. logically answering questions or summarizing pieces of text). All these capabilities have the potential to open up new applications for

[1] MITRE ATT&CK® - https://attack.mitre.org/

cyber attacks. Moreover, emergent abilities remains a phenomenon under controversial scientific debate, as it is unclear why and how larger models tend to behave differently from smaller language models. From the cyber security perspective emergent abilities will be hard to deal with, because if we are unable to explain and predict the behaviour of LLMs when the scale increases (e.g. of data or number of parameters) it is almost impossible to counter this with the right security measures. As it is generally very difficult to defend oneself against phenomena one does not yet comprehend.

We conclude that the developments in LLM technology do influence the cyber threat landscape. Examples of these threats, e.g. the usage of LLMs for spearphishing, were provided. We briefly discuss what the strengths and weaknesses are for some of the LLM implementations out there, but benchmarking different LLM implementations was out-of-scope.

Looking ahead, in the midterm future attackers might be able to leverage LLMs for the following:

- Impersonation / personalization for better phishing and social engineering at scale;
- Generating (malicious) code and tools;
- Leveraging information stored in an LLM for inference and decision support;
- Automated exploitation;
- Finding vulnerabilities in code, log files, configuration files, etc.;
- Adversarial attacks on other (less sophisticated) LLMs.

An important remark here is that we do not deem this list to be exhaustive. The field of LLMs and the field of cyber security is rapidly evolving. At the same time adversaries are creative in there ways of operating. Based on the methodology we applied these are the capabilities of LLMs that are according to the project team most relevant. Also, we do not claim that these above capabilities *will* materialize for adversaries, but they *might* materialize.

While it was challenging to develop metrics that express the effect of LLMs on attack techniques, the investigation did result in the development of a Signposts of Change model. This model consists of indicators that should be tracked over time in order to determine whether the threats that LLMs pose is changing over time. For this we used the MITRE ATT&CK analysis of literature to formulate the most realistically expected changes/developments caused by LLMs on the cyber threat landscape, and subsequently defined indicator events which should help the NCSC-NL to monitor the developments over time.

An example of one of those expected changes in the threat landscape is the scenario in which LLMs become so powerful at exploitation that they can be leveraged for automated, or even autonomous, exploitation. In order to track whether this scenario is manifesting over time we formulated the following indicator events:

- Change in behaviour of C2 traffic: request and answer is directly connected to the remote LLM model.
- More of the exact same automated attacks that fail to work. Non-sensical changes/commands by LLM.
- Faster repetition of the same mistake.

The list of all indicator events is too large to summarize in this management summary.

Overall conclusion

LLMs have at the time of writing caused at the very least an *evolutionary change* to the threat landscape; some attack techniques can be executed more easily, more efficiently or at a larger scale with the use of LLMs. We also see the first signs of a *revolutionary change*, which has not yet materialized.

We identified 6 expected changes in the cyber threat landscape. Three of these six potential future developments in the cyber threat landscape are evolutionary changes to the threat landscape:

- LLMs used for larger scale impersonation attacks that can be personalized more,
- LLMs used by moderate experts as a buddy/co-pilot that suggest attack paths,
- LLMs used by companies publicly could become vulnerable to LLM-specific attacks.

The other three potential future developments would have game-changing impact on the cyber threat landscape, and therefore be revolutionary changes:

- LLMs used by non-experts for producing malicious/exploit code,
- LLMs used by any actor for autonomous exploitation, or
- LLMS used for automated identification of vulnerabilities in software.

Monitoring these revolutionary changes to the threat landscape is expected to have more operational value than monitoring evolutionary changes over time. With the Signposts of Change model, the NCSC-NL can monitor over time whether these signs of a revolutionary change are further developing.

From research results to operational use

The MITRE ATT&CK framework has worked well to structure the scientific literature on the usage of LLMs. It helped to identify for what attack tactics and techniques confirmation of the usage of LLMs already exists, and for what tactics and techniques it is completely irrelevant. Most importantly, it helped to identify blind spots: tactics and techniques for which no evidence in literature currently exists, but if manifested will be of serious concern.

Furthermore, the Signposts of Change model has proven to be very promising. Combining 1) technical expertise on LLMs and generative AI, 2) offensive cyber security expertise (ideally operational), and 3) operational threat intelligence expertise, to construct this model was very valuable. The methodology has helped us to dissect the difficult, multi-facetted and sometimes ambiguous challenge of monitoring the threats that LLMs pose for the cyber security landscape.

We advise the NCSC-NL to continue the development of the Signpost of Change model by

- firstly, verifying the usefulness of the model with a larger group of experts and relevant stakeholders;
- formulating, and maintaining, a set of expected changes to the cyber threat landscape, following the procedure of the MITRE ATT&CK analysis as presented in this research;
- improving the model by introducing a hierarchy or categorization of the indicator events (e.g. technical vs. non-technical indicators, criticality of indicators, on what frequency the indicator values should be updated, information sources needed to update the indicator values);
- leverage expertise at knowledge institutions within the Netherlands to keep technical knowledge on LLMs up-to-date.

Table of Contents

Introduction

Recent developments in the field of Large Language Models (LLMs) [1] create unprecedented opportunities to automate tasks that were previously out of reach. Not only for benign tasks but also adversarial tasks. There is already some evidence that cyber criminals can use Large Language Models to optimize their activities and explore new options of attacks[2].

This memorandum is the result of a technology exploration project, conducted in collaboration between the NCSC-NL and TNO. The main goal of this exploration was to identify how we can measure and monitor the impact that LLMs have on the cyber threat landscape now up until the midterm future (3 to 5 years from now).

Therefore, in this memorandum we will answer the following research questions:

1. What are recent developments in the usage of LLMs for cyber attacks by cybercriminals (explicitly not nation-states)?
2. Which implementations of LLMs are cybercriminals most likely to use (e.g. large remote models versus locally trained models)?
3. For what activities, and in what ways, can LLMs be of most use to cybercriminals?
4. Which variables determine whether an LLM can be used for a cyber attack?
5. a. Is it possible to develop a metric that expresses the improvement of using LLMs on existing cyber attacks?
5. b. What are potential indicators of change and trigger events to watch in order to re-evaluate the threat that LLMs pose?[3]

Methodology

We began the exploratory research with a general literature survey on current trends in LLM technology, and continued this survey with the analysis of scientific literate that showcases the potential usage of LLMs for cyber attacks in various contexts. In this phase, searching for literature was done based on common sense using the expertise from the team: 1) knowledge on machine learning, NLP and LLMs, 2) knowledge on cyber security in general, 3) offensive cyber knowledge.

Subsequently, we used the MITRE ATT&CK framework to structure the literature review. This framework is used as a reference for the currently (publicly) known attack tactics and techniques. We map the literature on the framework both *vertically*, i.e., in one phase of the kill-chain (Initial Access), and *horizontally*, i.e., over all phases in the kill-chain (from Reconnaissance to Impact). This mapping helps to identify for what tactics and techniques current literature confirms the usage of LLMs is actually useful. Furthermore, it also helps identifying blind spots, i.e., the tactics and techniques for which LLMs might be or become very useful but has not yet been proven in scientific literature.

Lastly, we dived into how the NCSC-NL can monitor the impact of LLMs on the threat landscape in the midterm future (read: 3 to 5 years from now). We briefly discuss the limitations of currently available metrics and move on to a *Signposts of Change* model. We recommend using this model to track the impact of LLMs on the threat landscape over time.

We end this document with the conclusions, including answering the research questions, and a brief discussion on further research.

[2] https://research.checkpoint.com/2022/opwnai-ai-that-can-save-the-day-or-hack-it-away/

[3] Added later in the project because of change in scope, to align better with operational need. The signposts of change methodology was a suggestion from the NCSC-NL operational threat intelligence team.

Current trends in Large Language Models

The field of Large Language Models is a subset of Artificial Intelligence that specializes in the understanding, summarizing and generation of text. This includes human languages, such as English, as well as special textual formats, such as programming languages. In the last three years, attention for this field has exploded and Large Language Models are currently being studied and integrated in all major industries from education, consulting and finance to software development and cyber security [2]. Obviously, their capabilities have also been noticed by cyber adversaries.

The scientific field of NLP (Natural Language Processing) is specialized in the automated processing of natural, informal text and speech and is divided into natural language understanding and natural language generation. Historically, for the field of understanding natural language, systems have generally been designed and trained on a narrow and well-defined domain, for instance as customer service agents in specific industries like banking, insurance, travel agencies and so on. While these systems have been in use for years and have enjoyed success in their specific domains, they do not have the ability to have a conversation on arbitrary topics like their larger counterparts do.

Large Language Models are created by feeding a model a substantial amount of textual data. The current challenges of using complex systems gathering large volumes of data require new solutions which can identify and learn from the patterns found in the datasets. Using simple machine learning models turned out to be insufficient in some situations. The research focused on larger models, which would be good enough to solve complex tasks - for instance, subjective tasks normally requiring human supervision. This was a direct response to the challenges of domains such as natural language processing, computer vision and other sub-domains which sought generative models with extended capabilities and better performance. Consequently, this led to the emergence of the first large language models: transformer-based models with self-attention mechanisms which presented capabilities exceeding other predictive models [3]. The work on such models was accelerated by the researchers of Google Brain, Google Research and University of Toronto in 2017, after a successful development of encoder-decoder transformer models with self-attention mechanism [4]. Since then, the research and development has mostly been dictated by big-tech companies, such as Google, Meta and OpenAI. These organizations define the state-of-the-art techniques in the field of LLMs.

The most well-known of these models is ChatGPT, which was the first LLM to take the world by storm in late 2022 and has been described as the fastest growing consumer internet app, reaching one hundred million users in only two months. On the open source side the most well-known model at this time is Llama2, which is developed by Meta and is licensed for free for research as well as commercial use.

Solving perceptual tasks

One of the key research directions in the context of LLMs is the creation of models generalizable enough to apply them in any domain [5]. The most popular large language models, the GPT models developed by OpenAI, were assessed in terms of their capabilities on exams made for humans. They reached human-level performance on these benchmarks [6].

Moreover, in terms of the given tasks, the state-of-the-art large language models are capable of operating with languages other than English. The GPT-4 model outperformed the GPT-3.5, Chinchilla, and PaLM models for the majority of languages including low-resource languages such as Welsh, Latvian, Swahili [6]. This means that even for smaller corpora of the training set, the model is still able to understand given instructions and produce results in a multi-domain and multi-lingual context.

Another interesting capability of the large language models is performing specific technical tasks, such as solving coding problems. Considering the GPT-4 model, since code repositories were included in the

training set, it is able to produce code of a certain quality, depending on a given task. At the same time, it can introduce vulnerabilities to the program similar to how a human developer might [6].

Despite certain improvements, state-of-the-art models still suffer from several limitations. Most importantly, current models are still not fully reliable, tending to hallucinate (but already much less than earlier generations, e.g. comparing GPT-4 with GPT-3.5). Hallucination refers to the phenomenon that LLMs may produce output that is inaccurate, misleading or entirely fictional. Hallucinated output may seem plausible, which makes it hard for users to detect and reject. Depending on the use case, it may be considered harmful to rely on LLM output due to the risk of hallucination [7]. The factuality of the answers provided by GPT-4 model was tested against TruthfulQA[4] which evaluates the model's ability to separate fact from an adversarially-selected set of incorrect statements. The result of this test was approximately 60% accuracy for the answers provided by GPT-4 model [6]. With these methods, one can check how much information was learned by the model during training.

Emergent abilities

Machine learning (ML) models are expected to produce accurate predictions based on a given input. Let us take the definition from Tom M. Michell's book "Machine Learning: An Artificial Intelligence Approach " (1999), which shaped the view on the growth of ML : "A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E*". We expect the model to improve its performance based on its training and generate reproducible outputs which are precise and accurate for the task we attempt to complete. An essential part in this consideration is the training itself, which involves numerous techniques to optimize it and make the model robust and generalizable to the unseen data. With large language models, we might be going even a step further, beyond what is explicitly contained within the training set. Reportedly, the capabilities of these models emerge from their scale, as well as from learning patterns which are specific for an enormous corpus of information. Then, it is not only that the model is large in terms of the number of its trainable parameters, but also the training set that is used is of considerable size.

An ability of a model is called emergent if it exists in a large model but does not in a smaller model, and the presence of this ability could not be predicted based on extrapolations from experiments with smaller models. This is why they cannot be extrapolated from the behavior of smaller models. Emergent abilities can be visualized with a scaling curve, where the x-axis represents the model's scale and the y-axis represents its performance. What can be observed is that until a certain threshold of the model's scale, its performance appears as a semi-random pattern, and once exceeded, the pattern starts to be very specific to a given scale [5]. Defined as qualitative changes of model's behavior given quantitative changes to it, emergence remains an obscure feature of LLMs bringing both opportunities and threats. Models may come up with novel solutions which originate from complex patterns captured in the training data, but explaining this process is an open research question. The creativity of a model originating from such an emergent behavior can be assessed in terms of closeness to human creativity level, thus it brings a set of opportunities for automated systems. At the same time, the model is able to present emergent risky behaviors related to bias, toxicity and problems with its truthfulness [5].

The research community of natural language processing domain is divided regarding the acceptance of emergent abilities theory. Emergent abilities are characterized as sharp and unpredictable, due to the fact that the performance curve suddenly becomes steep, and the presence of such a pattern is not observable until reaching a certain scale of the model [8]. Sometimes though, a researcher might introduce a scientific bias due to how the observations are measured. One explanation to the characteristics of emergent abilities is that the performance of the large language model is not

[4] https://github.com/sylinrl/TruthfulQA

monitored on a continuous scale, which would dim the sharpness and unpredictability of emergent abilities. The selection of metrics which scale the error rate of any model in a non-linear or discontinuous way, might be an artificial reinforcement of the model's capabilities. It allows us to categorically answer the question if the task given to the model was solved as expected (binary classification of the model's performance) and not how close the model was to solving the task [8]. Overall, the presence of such abilities can be artificially emphasized by the metrics used for benchmarking tasks.

Conclusions – what trends are relevant for cyber security?

Overall, large language models present a wide spectrum of abilities ranging from those explicitly taken from the training set, to emergent abilities originating from more obscure patterns and model's scale. They can work with different languages, answer questions from a vast set of topics and present knowledge related to specific domains. Moreover, some of LLMs are multimodal - capable of working with various formats of inputs which currently include text and images. An example in this case is the GPT-4 model.

All these capabilities have the potential to open up new applications for cyber attacks. Moreover, emergent abilities remains a phenomenon under controversial scientific debate, as it is unclear why and how larger models tend to behave differently from smaller language models. From the cyber security perspective emergent abilities will be hard to deal with, because if we are unable to explain and predict the behaviour of LLMs when the scale increases (e.g. of data or number of parameters) it is almost impossible to counter this with the right security measures. As it is generally very difficult to defend oneself against phenomena one does not yet comprehend.

In the next chapters we will further dissect the threats that LLMs actually pose within the cyber security landscape.

Usage of LLMs for cyber attacks – most prominent examples

In this chapter we will dive into some specific examples of usage of LLMs for cyber attacks as found in current scientific literature, e.g., usage of LLMs for spearphishing. We conclude with an example of an LLM that has been trained on offensive cyber security practices: PentestGPT.

In this chapter the literature is not yet structured, and presented by means of examples. The search for scientific literature was done based on common sense using the expertise from the team, which consisted of: 1) knowledge on machine learning, NLP and LLMs, 2) knowledge on cyber security in general, 3) offensive cyber knowledge. In the next chapter (*Analysis using the MITRE ATT&CK Framework – current and future threats*), a structured scientific literature overview will be presented.

Spearphishing

One of the possibilities for incorporating LLMs into offensive security tasks is spearphishing [9]. In such a case, they could be used to enrich attacker's information about a certain entity or speed up the process of collecting it. Considering that the model can have access to a large set of data retrieved from the open Internet, it could capture part of it which is specific to the targeted organization. This would not only be the company's characteristics which can be gathered from an official website, but also the details about its employees. Examples of these would be, among other things, their interests, comments, or other pieces of information accessible via publicly available platforms. Potentially, things such as company's jargon or events which impacted it can be processed and learned by the model as well. Having such a large set of data which is specific to the targeted company, an attacker can attempt to prepare a phishing content using the retrieved information.

An example attack can also use job openings to disrupt the work of human resources employees. Since LLMs are capable of incorporating information relevant to a process of creative thinking, they can be used to automate the preparation of fake motivation letters or CVs, possibly including malicious content in these documents (including emails). Considering multi-modal models and generative capabilities of some of the models used for image creation, such content can be enriched with forged images or other content which makes the document more credible. In such a setting, the work to filter out counterfeit profiles might be too cumbersome for the team responsible for hiring process. A decision that a forgery profile is credible could open path to other forms of malicious activities.

Generating malware

One of the core features of ChatGPT is that it can generate more than just natural text in multiple languages. It can generate text formatted in specific formats, like poems, resumes, movie scripts, and even executable computer code based on natural language descriptions given by a human operator.

A malicious actor who wishes to exploit a system can use ChatGPT to generate malware payloads on command [10] [11]. This elevates a non-technical criminal to one that can generate code for malicious purposes without requiring programming knowledge.

LLMs as reasoning engines in cyber security contexts

Large language models can be facilitated in both red and blue teaming context. In essence, their use allows for solving complex perceptual tasks which would otherwise need to be completed by a human operator. Considering a large corpus of information, an automated solution which uses a large language model as a managing program can shorten the time for decision making.

As it was mentioned before, LLMs such as GPT models, can solve challenges given in competitive contexts at human-level performance, such as exams from different domains. This opens the door for introducing such models to decision-making processes. Their capability to follow instructions and understand their context, proves that they somehow mimic deductive and analogical reasoning.

One of the models optimized for instructions (specifically in office context) is the Dolly model from Databricks [12]. The use of such open models locally can greatly benefit the security domain in which the confidentiality of the information shared with an LLM can be of essential importance. In such a setting we might expect the complete control over the transmission of the information across the network, as well as over the way it is processed by an LLM. The use of a closed-access third-party model cannot guarantee such a setting. The advantage however, is that these closed models are pre-trained on a large corpus of information and can be implemented or used without requiring a lot of effort compared to open models. At the same time, their performance cannot be expected to meet expectations for very specific tasks, as the knowledge their memorized might be of general level [5]. Conversely, proprietary models such as the OpenAI GPT family bring a higher expected level of performance for a wide spectrum of tasks, as its improvement is commercially-motivated.

In any case, the security-related operations can benefit from the use of large language models. They clearly memorize compressed form of knowledge originating from vast number of sources, present creative, emergent abilities which may be expanded in the near future and mimic schemes of reasoning typical for humans. Despite the fact that these assumptions are related to an early stage of development of such models, it seems that at least a portion of tasks can be delegated to them. Opportunities and threats must be evaluated in empirical conditions and cannot be clearly determined in theoretical setting.

PentestGPT

Large language models are trained on large sets of unclassified and largely unfiltered data which is collected or scraped from the internet. The resulting LLM performs well on a wide array of subjects but is usually not an expert in most of them. The model can be adapted to specific purposes in a process called fine tuning [13]. Fine-tuning exposes the generic model to prompts in a more niche dataset in an effort to improve the model's performance on a specific topic.

Unlike the name suggests, PentestGPT [14] is not a large language model from the generative pre-trained transformer (GPT) family. It is an interactive tool that is built on top of the ChatGPT model from OpenAI.PentestGPT tries to guide a user to successfully penetrate a target machine by prompting the user to follow the most promising lines of exploration.

Analysis using the MITRE ATT&CK Framework – current and future threats

The MITRE ATT&CK Framework is a knowledge base containing techniques and tactics used by adversaries in Cybersecurity, constructed by MITRE[5] in close collaboration with industry. The framework consists of 14 tactics observed on real-life data often used by adversaries. For each of these tactics, there are multiple techniques and sub-techniques which can be used during every respective phase of an attack. For example: for the Privilege Escalation tactic, there are 14 different techniques described that an adversary could use to perform Privilege Escalation.

As it is a well-known and widely used framework within the cybersecurity industry it is a helpful structure for us to identify how LLMs can be of use to adversaries. We took a look at the different techniques used for each tactic to see if usage of an LLM could improve the technique somehow, in an expert-based approach. For this, we went in depth into the Initial Access tactic's techniques, which will be explained in the next section *Analysing one tactic in detail – Initial Access*. Afterwards, we took a broader look at all the tactics rather than specific techniques, which we will elaborate in section *Analysing the framework globally – all tactics.*

To help us with this, we searched for literature that would cover this topic. For this chapter, we looked at literature having to do with LLM's, in combination with search terms such as all the named MITRE ATT&CK tactics (Resource Development, Execution, Initial Access etc.), as well as terms such as: cybersecurity, hacking, penetration testing, exploits, red team, blue team, phishing, etc.

Analysing one tactic in detail – Initial Access

The MITRE ATT&CK Framework describes Initial Access as "The adversary is trying to get into your network". The idea is that an adversary needs some way to get into the network, whether this be via the internet, other telecommunications, or using some physical or social approach. For this tactic, the framework has lined up 10 different techniques. Three of them (Phishing, Supply Chain Compromise, and Valid Accounts) have a couple of sub techniques listed as well. For all of these techniques, we looked for scientific documents of research, as well as using our own knowledge and experience, to see if the usage of an LLM could improve this technique in any way or form.

From this, we concluded that 7 of the 10 techniques could not be improved by using an LLM. These techniques are:

1. Drive-by Compromise
2. External Remote Services

[5] MITRE ATT&CK® - https://attack.mitre.org/

3. Hardware Additions
4. Replication Through Removeable Media
5. Supply Chain Compromise
6. Trusted Relationship
7. Valid Accounts

For these techniques, we were unable to find any literature that could prove that using an LLM would help. Furthermore, some of these techniques do not use any text, language, or software at all (for example: Hardware Additions, Replication Through Removeable Media).

The three techniques that we believe might be improved upon by using an LLM are:

1. Content Injection
2. Exploit Public-Facing Application
3. Phishing

For Content Injection, a legitimate user needs to communicate with the server in some form. When an adversary is able to either do a Man-in-the-Middle Attack, or is able to race the server with a response, they could craft a fake message that the unsuspecting user thinks is real but will actually compromise their account somehow. In both cases, an LLM could help create such a message faster and more efficiently, as well as personalizing it to the specific user more [15].

Exploit Public-Facing Application is a technique in which the adversary tries to find any weakness in the Osystem that could be accessed publicly. For example: the server might have left a port open that is not properly secured. An adversary might be able to find this port by scanning the network. An LLM could speed up this process so that the adversary can focus on other areas [10].

Phishing is a technique often used by adversaries, where they try to gain access by trying to lure the users to give up some form of information. For example: an phishing e-mail that looks just like a normal e-mail from your bank, but will actually lead to a login-form that is made by the adversary. When a user then logs in with their username and password on this false website, the adversary will log this and thus is able to gain access to that account. Usage of LLMs for phishing can help not only increase the speed with which an adversary is able to create these messages, but also personalize them to the recipient, or to look more trustworthy and realistic, thus scaling these phishing campaigns enormously and efficiently [9].

In the MITRE ATT&CK Framework, there are four sub-techniques described:

1. Spearphishing Attachment
2. Spearphishing Link
3. Spearphishing via Services
4. Spearphishing Voice

For most of these, usage of an LLM is the same as for general phishing purposes. However, sub-technique 4 (Spearphishing Voice) is slightly different. Here, an adversary tries to perform a phishing attack by using the voice of someone familiar or trusted to the recipient. This can be done either by being a good imitator, or using voice-changing software and tools. If an adversary is able to set up a system where the input of the recipient is used in an LLM to generate realistic conversational output, and synthesize a voice out of this text that sound realistic, they could more easily and at a larger scale perform such a phishing attack [16].

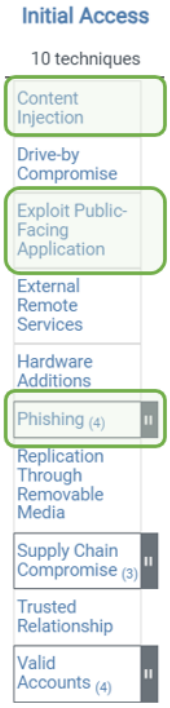

Figure 1: confirmation of usage of LLMs in scientific literature for the Initial Access phase in 1) content injection [15], 2) exploit public-facing applications [10] and 3) phishing [9], [16]. Source: MITRE ATT&CK® - https://attack.mitre.org/.

Analysing the framework globally – all tactics

Looking more broadly at the 14 different tactics in the framework, we wondered if these tactics could be improved by the usage of LLMs in any way. Of these 14, we concluded that 4 of them could not be improved upon by an LLM:

1. Persistence
2. Collection
3. Command and Control
4. Exfiltration

We were not able to find any literature that could proof that using an LLM could help with any of the techniques that belong to these tactics, and we were also unable to reason from experience how an LLM could improve an adversary with these techniques.

However, for seven of the tactics, we were able to find literature and/or think of ways that using an LLM could improve the techniques used:

1. Reconnaissance
2. Resource Development
3. Initial Access
4. Execution
5. Credential Access
6. Discovery
7. Lateral Movement

For Reconnaissance, Initial Access, and Lateral Movement, the LLM could help an adversary with any form of phishing attempt. Examples: phishing for internal information (Reconnaissance), phishing for credentials (Initial Access), and phishing for internal information (Lateral Movement). Furthermore, it

might be possible to use the "knowledge" of an LLM for information regarding the target during the Reconnaissance phase, but we have not confirmed this.

For Resource Development and Execution, an LLM that has knowledge about writing code, such as GitHub's Copilot, could help an adversary create tools for specific attacks or targets. Furthermore, if the code is written in a programming language unknown to the adversary, this could enable them to do some attack which they previously might be unable to due to not being experienced with that specific language. LLMs that write code could also speed up this process [10] [11] [17] [18].

For Credential Access, an LLM could be used to create more specific brute-force attacks, trained on known credentials obtained before this phase [19].

Finally, for Discovery, an LLM could aid an adversary to mine more efficiently through log files, source code, or other large text files that the adversary might have found. In these files might be some important information but this could be difficult to find. An LLM could speed up this process so that an adversary is able to obtain important information (such as credentials, private data, etc.) easier and faster [20].

Finally, for three of the 14 tactics, there could be ways that an LLM might help an adversary during these tactics but there is either not a lot of literature on it, or we are not fully convinced the LLM is able to help based on our experience. For this, experiments could be run in a future study to see if it possible to an LLM effectively for this. These three tactics are:

1. Privilege Escalation
2. Defense Evasion
3. Impact

Regarding Privilege Escalation, Happe et. al. have used cloud-based LLMs to generate commands that would be executed on a system to gain privilege. Here they have proven that the commands generated would indeed perform a Privilege Escalation attack, but they often produced nonsensical commands or repeated previous commands that did not work. Adding hints to this improved the ability of the LLMs, but some of them (such as Llama2) were unable to work properly [21].

For Defense Evasion, Lee Hu et. al. have studied the usage of LLMs to generate evasive malware that bypasses Deep Learning based malware detectors [22]. This might prove useful for adversaries to use LLMs to help get their chosen malware into a system unnoticed.

Finally, for Impact, an adversary might use an LLM for defacement purposes. However, the question is whether an LLM is needed for this, or if it even speeds up the process at all.

Figure 2: conclusions on usefulness of LLM technology for the different MITRE ATT&CK tactics based on the scientific literature. Source: MITRE ATT&CK® - https://attack.mitre.org/.

Conclusions – what is a threat, what will be a threat, what will not be a threat?

Given the analysis above, we can cluster the usage of LLMs for malicious actors into four topics:

1. Impersonation and personalization for better phishing and social engineering at scale
   - leveraging the capabilities of LLMs to generate text as if written in the style of someone else, and
   - tailor text for a specific recipient based on information fed into the prompt of the model
2. Generating (malicious) code and tools
   - leveraging the capabilities of LLMs to generate software and combining these pieces of code into complete tools for malicious purposes
3. Knowledge stored in the LLM
   - leveraging the information present in the training data of the LLM for inference and decision support
4. Automated exploitation
   - implementing an LLM to autonomously interact with a target system (e.g. command prompt) with a specific malicious end-goal

Aside from these four topics, we are also looking to the future at possible implementations of LLMs by malicious actors. This, of course, means that there is no literature available, or research done yet on these topics. We decided to add two more topics to our list based on our knowledge and experience:

5. Finding vulnerabilities in code, log files, configuration files, etc.
   - feeding code and other files into an LLM with the goal to identify vulnerabilities/ misconfigurations in it
6. Adversarial attacks on other (less sophisticated) LLMs
   - implementing next generation LLMs to autonomously interact with earlier generation LLMs with the goal to exploit vulnerabilities/malfunctions in the earlier generation models

We expect that LLMs could be used to look through pieces of code that the user will provide and try to help them find any possible vulnerabilities within. Given that LLMs are able to work with code as discussed before, it might be able to see what kind of mistakes, errors, or faults are present in a given piece of code and could uncover a vulnerability which a malicious actor could then use.

Furthermore, as some companies might employ a publicly accessible LLM, malicious actors could try to perform an adversarial attack on these LLMs by clever prompting. If these LLMs are not properly separated from the other data on the server or website, it might be possible to obtain confidential information by prompt injection. This could also have a wider impact than just LLMs and focus on more AI-enabled tools in general.

For both topics, we have not been able to find any specific literature or research that has proven that this is indeed possible, but we expect that these functionalities can be used by malicious actors in the future.

Monitoring the impact of LLMs on the future threat landscape

We have seen that there are many dimensions in which an LLM can be of use to cyber criminals. These dimensions include knowledge required to create malware, knowledge required to perform an attack, speeding up cyber attacks, quality of generated and translated phishing content and quite likely others as well. In this chapter we dive into the challenge of monitoring this threat over time. We will start by

discussing the pursuit to formulate one or multiple metrics for this purpose, in the next subsection. Thereafter, we move on to a more fruitful approach, namely the signposts of change model.

Metrics for assessing offensive advantages to LLMs

Research question five asks if it is possible to develop metrics that express the improvement of using LLMs on existing cyber attacks.

We are primarily interested in metrics which highlight the unique capabilities of using LLMs as opposed to improvements in already existing techniques such as scripting. Such a set of metrics would be most useful if they can be used to track and predict offensive cyber capabilities over time for multiple versions and families of LLMs, thereby providing an early indication of defensive security problem areas by monitoring trend lines for instance.

For phishing campaigns this means that a metric can be easily established: The number of unique fake emails one person is able to generate has gone from a human operator speed to machine speed. We argue this is not a revolutionary change nor a very useful metric, because phishing is already being conducted and defended against on a massive scale worldwide.

The practice of using LLMs in offensive cyber operations is also such a new field that beyond phishing campaigns we found only proof-of-concept demonstrators such as PentestGPT in literature research. While the authors claim their tools to be a useful asset in assisting with operations, their value has yet to be validated by usage in the security community and is as yet hard to quantify.

For our research we also looked at more typical cyber security metrics such as: The number of unidentified devices, intrusion attempts, security incidents, mean time to detect, resolve and contain, average vendor security ratings, patching cadence, vendor patching cadence, and mean time for vendor incident response. While these metrics can be used to gauge how well a particular system is protected against a particular set of attacks, they offer limited information on how well LLM-powered attacks will perform in general.

Similarly we also examined typical LLM metrics such as: Number of parameters, BLUE score, ROUGE score, perplexity, factuality, relevance, accuracy, fluency, requests per second, time for first token, tokens per second, Jeopardy performance, wikidata queries and math questions. As LLMs get better, they will reach higher scores on these metrics and might also perform better in other domains such as offensive cyber operations, but here we also estimate the relationship between the metrics looked at and the general offensive potential to be distant.

Therefore we move on to the construction of a signposts of change model.

Signposts of change

In this section we will introduce a first version of a *signposts of change* model which is based on a well-known technique from intelligence analysis [23]. It is a model that can be used to track developments of relevant scenario's over time. For our application this means we can apply it for monitoring and tracking the development of the impact that LLMs have on the cyber threat landscape over time.

It is a simple methodology which consists of four steps:

1. *Identify a set of competing hypotheses or scenarios.*
    - → in our case: creation of the scenario's/changes/developments that influence the impact of LLMs on the threat landscape, which follows directly from the MITRE ATT&CK analysis discussed in the previous chapter

2. *Create separate lists of potential activities, statements, or events expected for each hypothesis or scenario.*
    - → in our case: creation of the indicator events that confirm whether the scenario's from the first step are actually manifesting.
3. *Regularly review and update the indicators lists to see which are changing.*
4. *Identify the most likely or most correct hypotheses or scenarios, based on the number of changed indicators that are observed.*

Using this methodology we constructed the following signposts of change model, for which a simplified version is found in Table 1 and a full version can be found in *Appendix – full signposts of change model.* The model consists of six expert-based threat scenarios that might occur in the future. Whether these threat scenarios will manifest is uncertain, but to track the progression along the way we will introduce per threat a set of so-called signposts of change, i.e., indicators, that can be used to track over time whether the threat is materializing. The goal of this model is that it should eventually support operational cyber threat intelligence at the NCSC-NL, to track the threat that LLMs pose over time.

However, the model introduced is a first version as both the *Expected changes in the cyber threat landscape* (column 1 in Table 1) as well as the *Indicator events* (column 3 in Table 1) have been formulated by a small group of experts (read: the TNO project team in collaboration with experts from the NCSC-NL) and have not been validated with a larger group of experts and/or stakeholders. Furthermore, the list of scenarios and indicator events is not exhaustive and should be enriched in the future based on input from a larger group of experts and/or stakeholders, if it is to be used operationally.

Alongside the model we will briefly describe the advised steps to come to an exhaustive and validated model ready for operational use.

Table 1: The Signposts of Change model for tracking impact of LLMs on the cyber threat landscape over time (v0.1).

Expected change in cyber threat landscape	Brief description	Indicator events
Impersonation and personalization enable social engineering on a much larger scale	Phishing is done at scale already. The key is personalization and/or impersonation at scale.	Higher quality and more personalized junk email; Spam filter becomes much less effective. Higher quality and more personalized phishing attempts are found on a larger scale. Increase in disputes regarding false accusations.
Non-experts can generate working malicious/exploit code	LLMs are able to create malicious code (polymorphic malware for example) enabling non-experts to create code without needing a lot of time and expertise.	An explosive increase in bad exploit code. Academic papers proving the concept. Non-experts claim a lot more vulnerability disclosure bounties with high CVSS scores.
Moderately capable experts use LLMs as a copilot/buddy for tasks and decisions	Actors use LLMs for helping making decisions when exploiting targets w.r.t. attack paths.	Spray of different types of advanced attack techniques (in other words, launching advanced attacks without thought-out strategy). Increase in attribution of sophisticated attacks to moderate actors. Unusually steep increase in sophistication of attacks. Increase in newly reported attack techniques.
Exploitation becomes autonomous	LLMs will be used for autonomous exploitation of specific targets.	Change in behavior of C2 traffic: request answer is directly connected to the remote LLM model. More of the exact same automated attacks that fail to work. Non-sensical changes/commands by LLM. Faster repetition of the same mistake.
LLMs becoming able to find and explain new vulnerabilities in code	LLMs are able to find weaknesses (security flaws, config mistakes) in written source code or binary executables or configuration files.	Academic papers proving the concept. Non-experts claim a lot more vulnerability disclosure bounties with high CVSS scores. High increase of CVEs in a shorter period of time for newly published software.
Companies using LLMs operational (connected to internal databases) become vulnerable to new generation LLMs	Jailbreaking customer service LLMs	Academic papers proving the concept. Non-experts claim a lot more vulnerability disclosure bounties with high CVSS scores. High increase of CVEs in a shorter period of time for newly published AI-enabled tools.

This model is put in use by periodically assessing the indicator events. One should assess the indicator events using the threat intelligence sources available, to determine whether the changes in the cyber

threat landscape are manifesting. In the methodology this is done by scoring the level of concern for every indicator event, ranging from *Negligible concern* to *Serious concern*.

How should the NCSC-NL continue with this

As it is a v0.1 model we do not advice to use this model for threat intelligence analysis straight away. However, with some additional effort it has definitely the potential to bring it to a v1.0 model that is ready for operational use.

First of all, and most importantly this model should be verified with a larger group of experts and relevant stakeholders, which was out-of-scope for this research due to time constraints. On expertise level this requires combining different expertises: 1) technical expertise on LLMs and generative AI, 2) offensive cyber security expertise (ideally operational), and 3) operational threat intelligence expertise. From a stakeholder perspective it is wise to work in a layered approach, with first ring stakeholders (developers of the model) and second ring stakeholders (users of the model). With the smaller group of first ring stakeholders the model can be brought to a higher maturity level with the aforementioned expertise at the table, while the second ring stakeholders should be consulted to validate whether it is usable in an operational context. What stakeholders are most relevant is up to the NCSC-NL, but in discussions over the course of the project it was suggested that the stakeholders for the first ring could be: NCSC-NL, Nationaal Bureau voor Verbindingsbeveiliging (NBV), and Rijksinspectie Digitale Infrastructuur (RDI), as these organizations all have a duty in protecting the Netherlands against cyber attacks, and thus also specifically for the protection against LLM-based attacks. In this process it is crucial to take into account the operational requirements from the threat intelligence teams of the different stakeholders.

Second, in order to formulate the right expected changes/developments in the threat landscape and indicator events, we advise to extend and complete the MITRE ATT&CK analysis as presented in this research. More specifically we advise to extend the vertical analysis as done for the Initial Access phase, namely identifying the usefulness of LLMs for every technique in the model. This should provide the right input from scientific literature to identify for what attack techniques LLMs can and cannot be used (at that moment).

Thirdly, the model can be drastically improved by introducing a hierarchy or categorization in the indicator events. In the current version there are different types of indicator events at different abstraction levels. For this hierarchy the factors that might be taken into consideration are:

- Technical versus non-technical
- Criticality of indicators
- Usage of indicators (day-to-day use or long-term use)
- Information sources needed to monitor/assess the indicators.

Lastly, we believe it might be hard for the NCSC-NL and its partners to keep knowledge on LLM technology up-to-date. Therefore, it might be desirable to leverage the available knowledge and expertise at the different knowledge institution in the Netherlands.

Conclusions

In this section we draw the main conclusions of this exploratory research project. We will come back to the research questions, and will draw general conclusions on 1) what threats do LLMs actually pose, 2) how to proceed from research results to operational use.

An important remark is that it is in general a challenge to keep up to date on how LLMs can be used by adversaries for cyber attacks. Namely, the field of Generative AI, and Large Language Models specifically, is under constant and extremely rapid development especially since the launch of ChatGPT

end of 2022. At the same time, the cyber security landscape is also rapidly developing both on the offensive as well as on the defensive side. The fact that these two fields are prone to constant change, processing the rapid new developments was a challenge during the project execution. Also for the NCSC-NL it will in the future require a continuous effort in order to keep knowledge up-to-date.

Recall the research the questions as presented in the introduction:
1. What are recent developments in the usage of LLMs for cyber attacks by cybercriminals (explicitly not nation-states)?
2. Which implementations of LLMs are cybercriminals most likely to use (e.g. large remote models versus locally trained models)?
3. For what activities, and in what ways, can LLMs be of most use to cybercriminals?
4. Which variables determine whether an LLM can be used for a cyber attack?
5. a. Is it possible to develop a metric that expresses the improvement of using LLMs on existing cyber attacks?
5. b. What are potential indicators of change and trigger events to watch in order to re-evaluate the threat that LLMs pose?

Main findings
Large Language Models present a wide spectrum of abilities ranging from those explicitly taken from the training set (e.g. generating English language or programming language), to emergent abilities originating from more obscure patterns and model's scale (e.g. logically answering questions or summarizing pieces of text). All these capabilities have the potential to open up new applications for cyber attacks. Moreover, emergent abilities remains a phenomenon under controversial scientific debate, as it is unclear why and how larger models tend to behave differently from smaller language models. From the cyber security perspective emergent abilities will be hard to deal with, because if we are unable to explain and predict the behaviour of LLMs when the scale increases (e.g. of data or number of parameters) it is almost impossible to counter this with the right security measures. As it is generally very difficult to defend oneself against phenomena one does not yet comprehend.

From this we conclude that the developments in LLM technology do influence the cyber threat landscape. Examples of these threats, e.g. the usage of LLMs for spearphishing, were provided. We briefly discuss what the strengths and weaknesses are for some of the LLM implementations out there, but benchmarking LLM implementations was out-of-scope.

Looking ahead, in the midterm future attackers might be able to leverage LLMs for the following:
- Impersonation / personalization for better phishing and social engineering at scale;
- Generating (malicious) code and tools;
- Leveraging information stored in an LLM for inference and decision support;
- Automated exploitation;
- Finding vulnerabilities in code, log files, configuration files, etc.;
- Adversarial attacks on other (less sophisticated) LLMs.

An important remark here is that we do not deem this list to be exhaustive. Again, the field of LLMs and the field of cyber security is rapidly evolving. At the same time adversaries are creative in there ways of operating. Based on the methodology we applied these are the capabilities of LLMs that are according to the project team most relevant for the NCSC-NL. Also, we do not claim that these above capabilities *will* materialize for adversaries, but they *might* materialize.

Lastly, it was deemed not a fruitful research direction to develop one or multiple metrics to express the improvements that LLMs bring for executing cyber attacks. Instead we developed a signposts of change model, which consists of indicators that should be tracked over time in order to determine whether the threats that LLMs pose is changing over time. For this we used the MITRE ATT&CK analysis to formulate

the most realistically expected changes/developments caused by LLMs on the cyber threat landscape, and indicator events to monitor these developments.

An example of one of those expected changes in the threat landscape is the scenario in which LLMs become so powerful at exploitation that they can be leveraged for autonomous exploitation. In order to track whether such a scenario is manifesting we formulated so-called indicator events that should be periodically assessed to identify whether the scenario is materializing. For this scenario the indicator events that were formulated are:

- Change in behaviour of C2 traffic: request and answer is directly connected to the remote LLM model.
- More of the exact same automated attacks that fail to work. Non-sensical changes/commands by LLM.
- Faster repetition of the same mistake.

For the full list of all expected changes and corresponding indicator events we refer to *Table 2: Full Signposts of Change model for tracking impact of LLMs on the cyber threat landscape over time (v0.1).*

The signposts of change model was created with the use of a well-known intelligence analysis methodology, and therefore we advise the NCSC-NL with its partners to update and maintain the model by following the same procedure:

1. *Identify a set of competing hypotheses or scenarios.*
    - → in our case: creation of the scenario's/changes/developments (not necessarily competing) that influence the impact of LLMs on the threat landscape, which follows directly from the MITRE ATT&CK analysis.
2. *Create separate lists of potential activities, statements, or events expected for each hypothesis or scenario.*
    - → in our case: creation of the indicator events that confirm whether the scenario's from the first step are actually manifesting.
3. *Regularly review and update the indicators lists to see which are changing.*
4. *Identify the most likely or most correct hypotheses or scenarios, based on the number of changed indicators that are observed.*

Overall conclusions

We can conclude that LLMs have caused at the very least an *evolutionary change* to the threat landscape. As some attack techniques can be executed more easily, more efficiently or at a larger scale with the use of LLMs.

Furthermore, we do see the first signs of a *revolutionary change*, which has not yet materialized. Based on the analysis using the MITRE ATT&CK framework there are clearly multiple tactics and techniques that will benefit from the usage of LLMs, for adversaries. Also we have identified some blind spots for which we deem the usage of LLMs are relevant but for which currently no publicly available scientific literature exists. With the signposts of change model the NCSC-NL can monitor over time whether these signs of a revolutionary change are further developing.

Based on the literature analysis using the MITRE ATT&CK, we identified six expected changes in the cyber threat landscape. Three of these six potential future developments in the cyber threat landscape could be revolutionary. Namely, if LLMs can sufficiently effective be used for either of the three below scenarios this will be a huge game-changer:

- LLMs used by non-experts for producing malicious/exploit code,
- LLMs used by any actor for autonomous exploitation, or
- LLMS used for automated identification of vulnerabilities in software.

Monitoring revolutionary changes to the threat landscape is expected to have more operational value than monitoring evolutionary changes over time.

The three other potential future developments are evolutionary instead:

- LLMs used for larger scale impersonation attacks that can be personalized more,
- LLMs used by moderate experts as a buddy/co-pilot that suggest attack paths e.g.,
- LLMs used by companies publicly could become vulnerable to LLM-specific attacks.

From research results to operational use

Based on this research project we should also conclude that the MITRE ATT&CK framework has worked very well to structure the scientific literature on the usage of LLMs. It helped to identify for what attack tactics and techniques confirmation of the usage of LLMs already exists. Furthermore, it helped to identify for what tactics and techniques it is completely irrelevant. Lastly, and maybe most importantly, it helped to identify bind spots: tactics and techniques for which no evidence in literature currently exists, but if it is manifested it will be of serious concern.

Furthermore, the signposts of change methodology has proven to be very promising. Combining expertise on 1) technical expertise on LLMs and generative AI, 2) offensive cyber security expertise (ideally operational), and 3) operational threat intelligence expertise, to construct this model was very valuable. The methodology has helped us to dissect the difficult, multi-facetted and sometimes ambiguous challenge of monitoring the threats that LLMs pose for the cyber security landscape.

We advise the NCSC-NL to continue the development of the Signpost of Change model by:

- firstly, verifying the usefulness of the model with a larger group of experts and relevant stakeholders;
- formulating, and maintaining, a set of expected changes to the cyber threat landscape, following the procedure of the MITRE ATT&CK analysis as presented in this research;
- improving the model by introducing a hierarchy or categorization of the indicator events (e.g. technical vs. non-technical indicators, criticality of indicators, on what frequency the indicator values should be updated, information sources needed to update the indicator values);
- leverage expertise at knowledge institutions within the Netherlands to keep technical knowledge on LLMs up-to-date.

Discussion

This research project was executed in a collaboration between NCSC-NL and TNO and was focussed on assessing the impact that LLM technology will have on the cyber threat landscape. It was important that the results of this research can help the NCSC-NL and its partners in the future to counter the potential threat that LLM technology poses. Therefore, there was an emphasis on the methodology of the research, which was an important result in itself such that the NCSC-NL can reuse and repeat certain methodologies to deal with LLM threats.

Furthermore, nation-state activities were explicitly kept out-of-scope for this research. The project team used mostly open-source information, and almost no operational threat intelligence to draw conclusions. If one takes nation-state activities within scope and consults classified documents and/or operational threat intelligence information, it can provide additional insight, which might result in different conclusions.

Moreover, this project focussed explicitly on Large Language Models and thus not on generative AI in other contexts or multimodal models. Redoing this analysis for multimodal models or generative AI in general would enrich this analysis, which interesting future research according to the project team.

Lastly, the initial idea for this exploratory project was to optionally include an experimental study on the capabilities provided by large language models to support offensive security operations. Even though experimentation could have learned the project team a lot, for this project other research activities were prioritized over experimentation. However, we did document some ideas for experimentation for future research:

a. **Empirical validation of quality of phishing and social engineering by LLMs**: imagine an experiment in which one launches a phishing simulation campaign in which the LLM is used to generate social engineering messages (e.g. emails, social media messages). These messages should lure the user to perform a specific action which violates the security policy. The value of such an experiment lies within testing the added value of an LLM. How much time does it save to launch a phishing simulation campaign? How much more/less effective is it compared to historical phishing simulation campaigns?
b. **From CVE to malware generation:** imagine an experiment where one feeds a CVE (Common Vulnerabilities & Exposures) into an LLM and requests the LLM to generate a piece of malware that exploits this CVE for a certain system with given specifications. How well can different LLM implementations do this? What LLM implementations can be used for this best? Does one need to use adversarial prompting to circumvent security measures of certain LLM implementations? What is the quality of the malware produced by an LLM?
c. **Countering LLMs used for exploitation:** imagine an experiment where one uses an LLM to perform autonomous exploitation. It can be done remotely, thereby connecting the LLM via a C2 channel to the target infrastructure, or by embedding a trained model into a piece of malware which infects a system. In the first scenario, how can we characterize or even detect typical LLM generated C2-traffic? Is it different from human generated C2-traffic? And for the second scenario, can LLMs be embedded in malware without increasing the size of the malware file too much? How does one scan files for detecting embedded LLMs? What other security controls can one implement to counter LLM-powered exploitation?

We leave the above described ideas for experimentation for future research.

References

[1] OpenAI. [Online]. Available: https://openai.com/research/better-language-models.

[2] „Research roundup for generative AI,” Gartner, [Online]. Available: https://www.gartner.com/en/documents/4567699.

[3] J. Brownlee, „Machine Learning Mastery,” 20 July 2023. [Online]. Available: https://machinelearningmastery.com/what-are-large-language-models/. [Geopend 17 October 2023].

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser en I. Polosukhin, „Attention is All You Need,” 2017.

[5] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean en W. Fedus, „Emergent Abilities of Large Language Models,” 2022.

[6] OpenAI, „GPT-4 Technical Report,” 2023.

[7] V. Rawte, A. Sheth en A. Das, „A Survey of Hallucination in “Large” Foundation Models,” 2023.

[8] R. Schaeffer, B. Miranda en S. Koyejo, „Are Emergent Abilities of Large Language Models a,” 2023.

[9] J. Hazell, „Large Language Models Can Be Used To Effectively Scale Spear Phishing Campaigns,” 2023.

[10] S. Moskal, S. Laney, E. Hemberg en U.-M. O'Reilly, „LLMs Killed the Script Kiddie: How Agents Supported by Large Language Models Change the Landscape of Network Threat Testing,” 2023.

[11] P. S. Charan, H. Chunduri, P. M. Anand en S. K. Shukla, „From Text to MITRE Techniques: Exploring the Malicious Use of Large Language Models for Generating Cyber Attack Payloads,” 2023.

[12] M. H. M. M. A. X. J. W. J. S. S. X. R. Conover, „Free Dolly: Introducing the World’s First Truly Open Instruction-Tuned LLM,” 2023. [Online]. Available: https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm. [Geopend 10 10 2023].

[13] [Online]. Available: https://www.deepset.ai/blog/llm-finetuning.

[14] [Online]. Available: https://github.com/GreyDGL/PentestGPT.

[15] C. Li, M. Zhang, Q. Mei, Y. Wang, S. A. Hombaiah, Y. Liang en M. Bendersky, „Teach LLMs to Personalize -- An Approach inspired by Writing Education,” 2023.

[16] „Enhance Call Center Automation: LLMs for IVR Systems,” [Online]. Available: https://www.teneo.ai/blog/enhance-call-center-automation-with-generative-ai-for-ivr-systems.

[17] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry en P. Mishkin, „Evaluating Large Language Models Trained on Code,” 2021.

[18] N. Kshetri, „Cybercrime and Privacy Threats of Large Language Models,” *IT Professional,* vol. 25, 2023.

[19] B. Jayaraman, E. Ghosh, M. Chase, S. Roy, W. Dai en D. Evans, „Combing for Credentials: Active Pattern Extraction from Smart Reply,” 2024.

[20] M. Allamanis en C. Sutton, „Mining source code repositories at massive scale using language modeling,” 2013.

[21] A. Happe, A. Kaplan en J. Cito, „Evaluating LLMs for Privilege-Escalation Scenarios,” 2023.

[22] J. L. Hu, M. Ebrahimi en H. Chen, „Single-Shot Black-Box Adversarial Attacks Against Malware Detectors: A Causal Language Model Approach," in *2021 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 2021.

[23] CIA. [Online]. Available: https://www.cia.gov/static/Tradecraft-Primer-apr09.pdf.

[24] TrustWave. [Online]. Available: https://www.trustwave.com/en-us/resources/blogs/spiderlabs-blog/wormgpt-and-fraudgpt-the-rise-of-malicious-llms/.

Appendix – full signposts of change model

Table 2: Full Signposts of Change model for tracking impact of LLMs on the cyber threat landscape over time (v0.1).

Changes/developments	Indicator events	Description	Possible impact	Actors	Mitre ATT&CK Tactics
Impersonation and personalization enable social engineering on a much larger scale	Higher quality and more personalized junk email; Spam filter becomes much less effective. Higher quality and more personalized phishing attempts are found on a larger scale. Increase in disputes regarding false accusations (hallucination fallout[9]).	Phishing is done at scale already. The key is personalization and/or impersonation at scale.	More effective phishing campaigns can lead to breaches of data, privacy, or system takeover. Internal disputes and lawsuits could occur.	Any	Reconnaisance Initial Access Lateral Movement
Non-experts can generate working malicious/exploit code	An explosive increase in bad exploit code. Academic papers will come out doing so. Non-experts start claiming vulnerability disclosure bounties.	LLMs are able to create malicious code (polymophic malware for example) enabling non-experts to create code without needing a lot of time and expertise.	More sophisticated malware leads to exploitation more often; increase in viruses, malware, and ransomware.	Non-experts	Resource Development Execution
Moderately capable experts use LLMs as a copilot/buddy for tasks and decisions	Spray of different types of advanced attack techniques. Increase in attribution of sophisticated attacks to moderate actors. Increase in sophistication of attacks. Increase in newly reported attack techniques. The time that elapses between a CVE, to exploit PoC, to usage is drastically shortened.	Actors use LLM's for helping making decisions when exploiting targets w.r.t. attack paths.	Attacks will become more critical (a higher CVSS score) allowing adversaries to exploit more and better.	Moderate experts	All
Exploitation becomes autonomous	Change in behavior of C2 traffic: request answer is directly connected to the remote LLM model. More of the exact same automated attacks that fail to work. Non-sensical changes/commands by LLM. Faster repetition of the same mistake.	Exploitation becomes autonomousLLMs will be used for autonomous exploitation of specific targets.	An increase in attacks happening on systems could lead to Denial of Service. Blue teams might not be able to keep up anymore.	Any	Reconnaisance Initial Access Execution Privilege Escalation Credential Access Discovery
LLMs becoming able to find and explain new vulnerabilities in code	Academic papers will come out doing so. Non-experts publish a lot more bug bounty programmes with high CVSS scores. High increase of CVEs in a shorter period of time for newly published software.	LLM's are able to find weaknesses (security flaws, config mistakes) in written source code or binary executables or configuration files.	Blue teams keep falling behind as more zero-days appear. Attacks on specific pieces of code might spread across seemingly unrelated programs.	Non-experts	N/A
Companies using LLMs operational (connected to internal databases) become vulnerable to new generation LLMs	Academic papers will come out doing so. Non-experts publish a lot more bug bounty programmes with high CVSS scores. High increase of CVEs in a shorter period of time for newly published AI-enabled tools.	Jailbreaking customer service LLMs	Companies that employ LLM's recklessly are breached more often.	Any	Initial Access Execution Privilege Escalation Credential Access Discovery