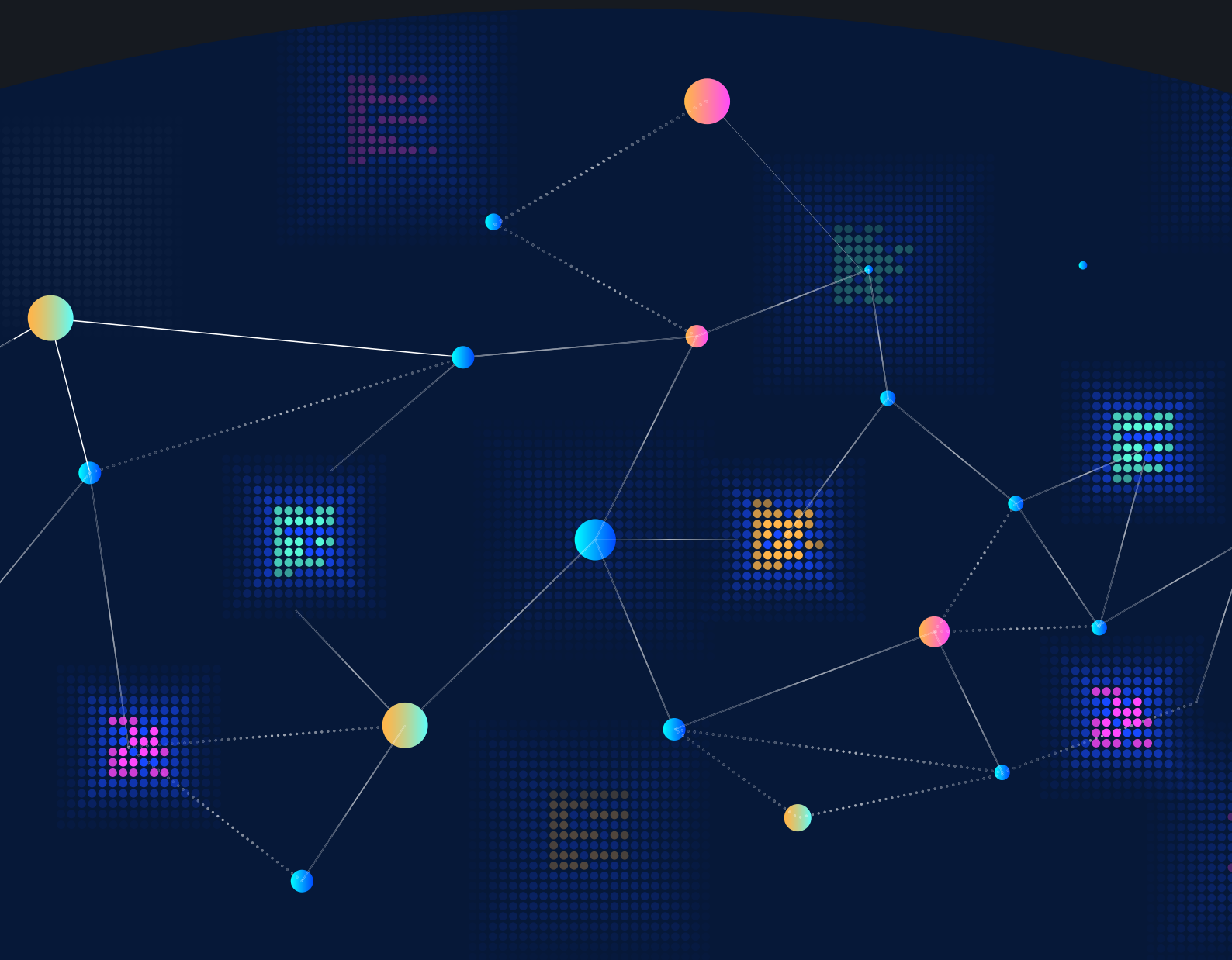
UNIDIR

Large Language Models and International Security

A Primer

IOANA PUSCAS



Acknowledgments

Support from UNIDIR's core funders provides the foundation for all of the Institute's activities. Work of the Security and Technology Programme on artificial intelligence is funded by the governments of Czechia, France, Germany, Italy, the Netherlands, Norway, the Republic of Korea, Switzerland, and the United Kingdom, and by Microsoft.

The author wishes to thank **Giacomo Persi Paoli**, **William Marcellino**, and **Jessica Ji** for their thorough reviews, comments and suggestions, **James Revill** for research advice, and the following experts who were interviewed for the project: **Richard Carter**, **Peter Hase**, and **Gitta Kutyniok**.

About UNIDIR

The United Nations Institute for Disarmament Research (UNIDIR) is a voluntarily funded, autonomous institute within the United Nations. One of the few policy institutes worldwide focusing on disarmament, UNIDIR generates knowledge and promotes dialogue and action on disarmament and security. Based in Geneva, UNIDIR assists the international community to develop the practical, innovative ideas needed to find solutions to critical security problems.

Note

The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area, or of its authorities, or concerning the delimitation of its frontiers or boundaries. The views expressed in the publication are the sole responsibility of the individual authors. They do not necessarily reflect the views or opinions of the United Nations, UNIDIR, its staff members or sponsors.

About the Author

Ioana Puscas (@IoanaPuscas1) is Senior Researcher on artificial intelligence with UNIDIR's Security and Technology Programme.

Acronyms & Abbreviations

AI	Artificial intelligence
DAN	“Do Anything Now”
LLM	Large language model
NLP	Natural language processing
RAG	Retrieval augmented generation
RLHF	Reinforcement learning from human feedback
RNN	Recurrent neural network

Contents

Introduction

Large language models (LLMs) represent one of the most prominent types of contemporary artificial intelligence (AI) systems. They are best known for their ability to generate content or summarize text when embedded in chatbots, but the range of applications of this technology is far broader – including emerging and potential use cases with impacts for international security.

LLMs are increasingly of interest to intelligence and military organizations, including for analysis, planning and other operational tasks. LLMs are also relevant to international security insofar as malicious actors could exploit capabilities afforded by LLMs for a range of nefarious purposes, such as to enhance disinformation campaigns, to conduct attacks in the cyber domain or to seek assistance with the production of weapons, including biological weapons.

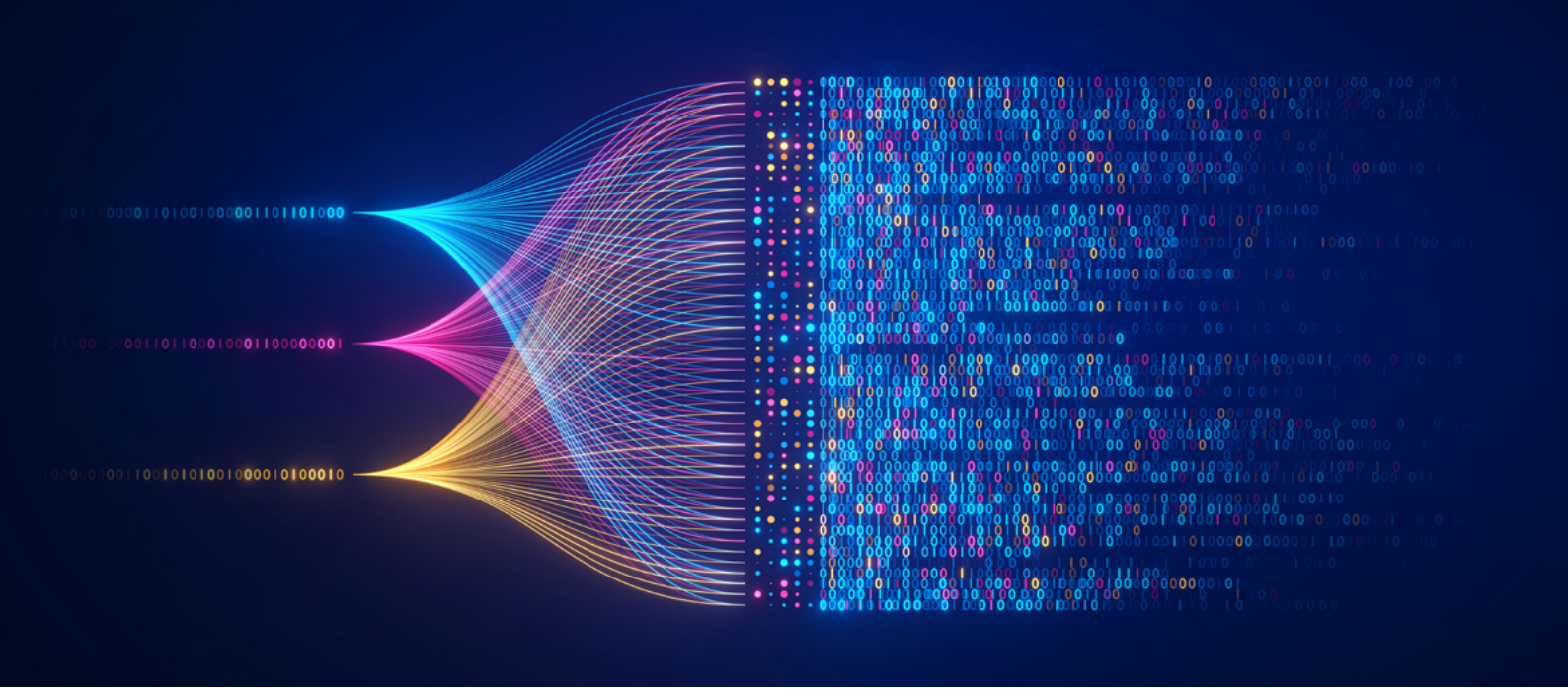
This primer aims to provide an overview of LLMs and their relevance to international security: **first**, by introducing and explaining the basics of the technology, including how it works and where key vulnerabilities lie, and **second**, by illustrating the impact of LLMs on international security through select examples of uses and applications.

The second part, which explores the link with international security, is divided in two sections, reflecting the technology's **dual-use** character: Section A highlights key examples of use cases by defence or security organizations, and it therefore focuses on areas of lawful use as part of the work of (State) organizations, such as defence planning, intelligence, and wargaming. Section B examines, through select examples, how the technology could be deployed by malicious actors for nefarious purposes (including, potentially, in violation of international law), such as for proliferation of biological weapons, cyberattacks, and disinformation.

This primer is intended for a **broad audience**, and particularly for the diplomatic and policy communities who are interested in gaining a better understanding of the technology that powers LLMs and key concepts pertaining to this field of AI.

The **scope** of this short paper is limited to a presentation of key use cases and areas of risk, both present and foreseeable. The rapid advance of the technology may likely open new possibilities for use and misuse.

The overview of applications and potential areas of misuse provides general and succinct descriptions of how LLMs could be deployed. Taken separately, each of the case studies could lend itself to further in-depth exploration and analysis. For the purpose of this primer, the examples serve to be illustrative and clarify how the technology can already be used (or misused), where key risks can be identified and where the technology continues to present limitations – for now.



1. Understanding Large Language Models

What are LLMs?

In a strictly technical sense, LLMs are probabilistic models of natural language. They are an example of generative AI, which refers to a class of AI systems that generate content. This content can be in the form of written text, but also computer code, video, image, audio, or in the case of multimodal models, a combination of these.

A simple way to describe how LLMs work, and how they generate text, is to start with the input, or 'prompt'. A prompt will lead the model to compute a probability of what is most likely to follow and to produce an output. Based on the patterns the model learned during training, it calculates the highest probability for what should follow next.[1] This process would calculate that, for example, the most likely way to continue the sentence "United Nations is an _____" is with the words "international organization" as opposed to other words.

Beneath the surface, the way this probabilistic calculation is executed is complex and relies on important recent innovations in AI. LLMs are typically deep neural networks trained on vast data sets. Most if not all current LLMs are built on a recent type of neural network called the **transformer architecture** (see Box 1) – though, it should be noted, key elements of natural language processing (NLP) date back decades.

---

[1] Melanie Mitchell, "Large Language Models", MIT Press, Open Encyclopedia of Cognitive Science, 24 July 2024, https://doi.org/10.21428/e2759450.2bb20e3c.

BOX 1.

The transformer architecture and attention mechanisms

The fundamental innovation of the transformer architecture, proposed in 2017, is that it is a feed-forward network, based solely on **attention** mechanisms.[2] This means that it does not rely on recurrence[3] and instead is able to capture long-range dependencies solely through a mechanism called attention, which does not account for the distance between the input and output.[4] In other words, this type of network is able to bestow 'attention' on elements in the data, and identify and keep track of patterns even when such elements are far apart.[5] Further, another important characteristic of the transformer architecture is that it allows for more parallelization, which means that parallel attention layers can compute and weigh values simultaneously – this permits scaling of training and a reduction in computational time.[6]

How do LLMs work?

Generally, there are two key steps to building an LLM: first, the pre-training phase, which feeds into a so-called base model or foundation model, and second, the fine-tuning phase,[7] where the model can be customized or trained for more specific tasks.

Pre-training

Among the initial steps in building an LLM is the process of tokenization, which refers to the varied methods (these can differ from model to model) to deconstruct elements of language into tokens: words, parts of words, punctuation/characters, etc. Strictly speaking, as the model will learn to calculate probabilities for the next words, it does not predict words *per se*, but tokens.

This takes place in a phase called **self-supervised pre-training**, which is the initial training of the model to predict what is most likely to come next. Here, tokens are assigned numerical representations (this list of numbers is called a word embedding). During the training process, the language model learns to calculate a probability distribution and how different words (broken down into tokens) relate to one

---

[2] Ashish Vaswani et al., "Attention Is All You Need", 12 June 2017 (revised 2 August 2023), arXiv, https://doi.org/10.48550/arXiv.1706.03762.

[3] Recurrent neural networks (RNNs) are deep neural networks, i.e. artificial neural networks with numerous hidden layers between input and output, which are trained on sequential data; essentially, the inputs in the model provide outputs based on a series of prior elements in that sequence. RNNs have been used in numerous natural language processing applications and predictive tools.

[4] Vaswani et al., "Attention is All You Need"; Mitchell, "Large Language Models".

[5] The Economist, "A Short History of AI", 16 July 2024, https://www.economist.com/schools-brief/2024/07/16/a-short-history-of-ai; IBM, "What is a transformer model?", https://www.ibm.com/topics/transformer-model.

[6] Vaswani et al., "Attention is All You Need"; IBM, "What is a transformer model?".

[7] Fine-tuning is a generic term to describe this stage in the creation of an LLM and comprises several methods.

another.[8] It is also in pre-training that *contextual embeddings* capture the words' meanings based on their context and how words relate to one another. Embeddings permit the model to represent nuance in language, such as homonyms, and other intricate semantic relationships. This is how, for example, the model will capture the contextualized meaning of the word 'bank' in its very different meanings in 'river bank' or 'savings bank'.

Typically, at this stage, the model is not trained for any specific task and primarily learns based on a corpus of data (text, in the case of LLMs) to identify patterns in a *self-supervised* manner. Self-supervised means the model is given unlabelled data and learns to optimize its performance based on a 'ground truth'[9] inferred from correlations observed in the unlabelled data.[10]

The majority of the data used to pre-train LLMs comes from publicly available Internet archives and resources, but the exact composition of the data sets remains to a large extent unknown simply due to the volume of the data needed to create 'large' models – often in the hundreds of terabytes.[11] This can amount to vast data sets scraped from billions of web pages,[12] with some totalling tens of trillions of tokens.[13] However, there is wide disparity across languages, which means that for some languages there is less training data than for others.[14]

---

[8] Matthew Burtell and Helen Toner, "The Surprising Power of Next Word Prediction: Large Language Models Explained: Part 1", Center for Security and Emerging Technology, 8 March 2024, https://cset.georgetown.edu/article/the-surprising-power-of-next-word-prediction-large-language-models-explained-part-1/; Mitchell, "Large Language Models".

[9] In machine learning, 'ground truth' refers to the 'real' or 'true' information which provides the target for the model. It is a concept frequently associated with *supervised learning*, where the labelled data is considered the ground truth for the model, which is then used as a reference for training the model.

[10] Self-supervised learning differs from unsupervised learning – though both use unlabelled data – in that *self-supervised models* measure results against a ground truth, though the ground truth is itself implicitly derived from the (unlabelled) training data. In *unsupervised learning*, the model learns correlations but does not subsequently measure the divergence between the ground truth and the predictions. These important differences have led to different use cases for these two methods; see Dave Bergmann, "What is self-supervised learning?", IBM, 5 December 2023, https://www.ibm.com/topics/self-supervised-learning.

[11] Jessica Ji, Josh A. Goldstein, Andrew J. Lohn, "Controlling Large Language Model Outputs: A Primer", Center for Security and Emerging Technology, December 2023, 4, https://cset.georgetown.edu/publication/controlling-large-language-models-a-primer/. What makes a model 'large' is somewhat contested, but the volume of data needed to train it is generally considered a differentiating factor, as is the computational power that is required for it.

[12] For example, the data provided by one company in September 2024 contained 2.8 billion web pages (note that this data is not specifically, or only, used for LLM training); see Common Crawl, "September 2024 Crawl Archive Now available", 24 September 2024, https://www.commoncrawl.org/blog/september-2024-crawl-archive-now-available.

[13] Estimates of total numbers of words and tokens in datasets are hard to aggregate. By way of illustration, one open dataset for training LLMs, released at the end of October 2023, had 30 trillion tokens. See Together AI, "RedPajama-Data-v2: An open dataset with 30 trillion tokens for training large language models", 30 October 2023, https://www.together.ai/blog/redpajama-data-v2.

[14] Interview Gitta Kutyniok (9 October 2024).

BOX 2.

Foundation model vs. LLM

The concept 'foundation model' was coined and popularized by a group of researchers at Stanford University.[15] Foundation models are often used interchangeably with 'generative AI' or 'large language models' but they are not, strictly speaking, the same and they are not limited to natural language processing (NLP).[16]

Foundation models are 'incomplete' models that serve as a basis for task-specific models.

They represent a new paradigm in AI, enabled by *transfer learning* (transferring knowledge learned in a task to another task) and *scale* (made possible with improvements in hardware, the transformer architecture, and the availability of more training data).[17]

There are two important properties that make foundation models stand out:

1. *Homogenization* of approaches and models: most state-of-the-art NLP models are adaptations of one of a few foundation models; and
2. *Emergence*, which is a property that results from scale: bigger models permit in-context learning, and lead to emergent properties that the model was not specifically trained for.[18]

Fine-tuning

Next, the pre-trained model (called the base model or foundation model) is fine-tuned for specific tasks. **Fine-tuning** means the model goes through additional training on data that is collected or curated for specific uses. In this phase, the data sets are smaller and specialized or tailored to a particular domain, allowing the model to adapt its performance to a specific area or task.[19]

Fine-tuning is carried out to address intrinsic limitations with the pre-trained model and to increase performance. Pre-trained models tend to exhibit several flaws, which can make them unfit for deployment. For example, outputs may be inaccurate or false, and they may reproduce harmful qualities of the training data, including racist, hateful, or sexist content. Fine-tuning allows for the correction of biases, or the fixing of problems related to

---

[15] Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models", arXiv, 16 August 2021 (revised 12 July 2022), https://arxiv.org/abs/2108.07258.

[16] See Helen Toner, "What are Generative AI, Large Language Models, and Foundation Models?", Center for Security and Emerging Technology, 12 May 2023, https://cset.georgetown.edu/article/what-are-generative-ai-large-language-models-and-foundation-models/.

[17] Bommasani et al., "On the Opportunities and Risks of Foundation Models", 4.

[18] Ibid., 5. It should be noted the emergence theory has been intensely scrutinized (and contested) by some researchers more recently.

[19] Mitchell, "Large Language Models"; Ji, Goldstein, Lohn, "Controlling Large Language Model Outputs", 4–5.

incorrect answers known as 'hallucinations' (further explained below).[20]

There are several ways to fine-tune a model. A common method is called **supervised fine-tuning**, which uses carefully curated data sets that contain labelled data. This provides for an effective way to steer the behaviour of the model and to train it for a particular application. Relatedly, **instruction tuning** relies on data sets which contain human-created examples of instructions and the responses to those instructions – thus training the model to respond to a wide range of prompts.[21]

To complement, as well as to address shortcomings with supervised fine-tuning, *reinforcement learning methods* are also employed, particularly a method called **reinforcement learning from human feedback** (RLHF). This is a complex and iterative method which first requires the identification of a *reward signal*, a quantitative indicator for assessing the performance of the model (in short, how desirable a certain outcome is). This can be an especially difficult process for complex LLMs tasks.[22]

The way this method is employed in the case of LLMs is by training a different machine learning model (called a 'reward model') on initial outputs of the LLM that have been ranked by human annotators based on their preference. For example, responses that follow instructions or which contain the least amount of bias would be ranked more preferably, and these scores are then used to train the reward model, which in turn provides the feedback, that is, the reward signal, to train the original LLM. This is a mechanism to encode human preferences in LLMs, and it has been instrumental to advances in LLMs, yet it presents vulnerabilities.[23]

Limitations, risks, and vulnerabilities of LLMs

At the annotation stage, important vulnerabilities in LLMs have become more apparent as larger models made the reliance on human labour significantly greater, adding to the practical challenges and risks associated with the task of annotation. These comprise, at one end, **unintended mistakes** or divergences over the subjective assessments that accompany the rating of a model's outputs. These are, arguably, unavoidable limitations

---

[20] Thomas Woodside and Helen Toner, "How Developers Steer Language Model Outputs: Large Language Models Explained, Part 2", Center for Security and Emerging Technology, 8 March 2024, https://cset.georgetown.edu/article/how-developers-steer-language-model-outputs-large-language-models-explained-part-2/; IBM, "What are large language models (LLMs)?", https://www.ibm.com/topics/large-language-models. While outside of the scope of this paper, it should be noted that the meaning of 'hallucinations' in LLMs can render itself to more complex and context-specific interpretations.

[21] Mitchell, "Large Language Models"; Ji, Goldstein, Lohn, "Controlling Large Language Model Outputs", 7.

[22] Woodside and Toner, "How Developers Steer Language Model Outputs". In LLMs, manually programming a reward signal when the desired outcome is something along the lines of "refrain from producing toxic or harmful content" is highly complicated. Comparatively, for other applications, it can be relatively easy to identify the reward signal because the goal can be more clearly programmed, such as for example in drone navigation or racing (avoid obstacles, cruise at high speed, etc.).

[23] Ji, Goldstein, Lohn, "Controlling Large Language Model Outputs", 8. Addressing limitations in LLM training and RLHF more specifically remains an active area of research; examples of recent approaches include Direct Preference Optimization or Constitutional AI.

considering the amount of human feedback needed to fine-tune the models.[24] At the other end, they can include intentional **malicious attacks**, such as **poisoning attacks**, whereby one or more adversarial annotators wilfully manipulate ranking scores, steering the model to generate harmful outputs.[25] In theory, poisoning attacks are possible at all stages of LLM training, but some studies have demonstrated that the efficiency of the attack may vary, with stages such as RLHF being considered more robust to attacks than others. However, at present, the understanding of relative vulnerabilities and robustness of the different training stages remains limited, and more research is needed to fill these gaps.[26]

Other risks surface in the deployment phase or are simply a function of the model's type or scale. In addition to the risks of exhibiting and perpetuating **biases** found in the training data, LLMs can generate completely erroneous outputs, though these may seem plausible – a phenomenon known as **hallucination**.[27] There are techniques to mitigate the problem of hallucination, ranging from introducing more specialized or fact-focused data (if the problem is one of the data), to architectural improvements, or to approaches that rely on external information – for example, a novel approach called retrieval augmented generation (RAG) works by retrieving information from external, up-to-date sources to supplement the LLMs' internal representation of information.[28] However, hallucination is acknowledged to remain an intrinsic limitation of LLMs, which cannot be entirely solved with RAG or by increasing the largeness of the model and its parameters.[29]

The scale of the model can itself present further vectors of unpredictability. Increase in scale[30] has been associated with what is known as **emergence** or **emergent abilities**, a property of LLMs described in a 2022 paper[31] and recently adopted in policy conversations about risks – though, at times, misunderstood or exaggerated.[32] The fundamental hypothesis of

---

[24] See Woodside and Toner, "How Developers Steer Language Model Outputs". The authors also note that as the performance of models improves, it is increasingly difficult for humans to evaluate those models.

[25] See Jiongxiao Wang et al., "RLHFPoison: Reward Poisoning Attack for Reinforcement Learning with Human Feedback in Large Language Models", arXiv, 16 November 2023, https://arxiv.org/abs/2311.09641v2.

[26] Usman Anwar et al., "Foundational Challenges in Assuring Alignment and Safety of Large Language Models", arXiv, 15 April 2024 (also published in *Transactions on Machine Learning Research*, 2 September 2024), 70, https://arxiv.org/abs/2404.09932.

[27] LLMs hallucinate in relation to a 'ground truth' function (see footnote 9) of the model. Hallucinations represent an inconsistency between the formal ground truth and the LLM. The benchmark for observing hallucinations is not, strictly speaking, the truthfulness or factualness of the real world. See Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli, "Hallucination is Inevitable: An Innate Limitation of Large Language Models", arXiv, 22 January 2024, https://arxiv.org/abs/2401.11817.

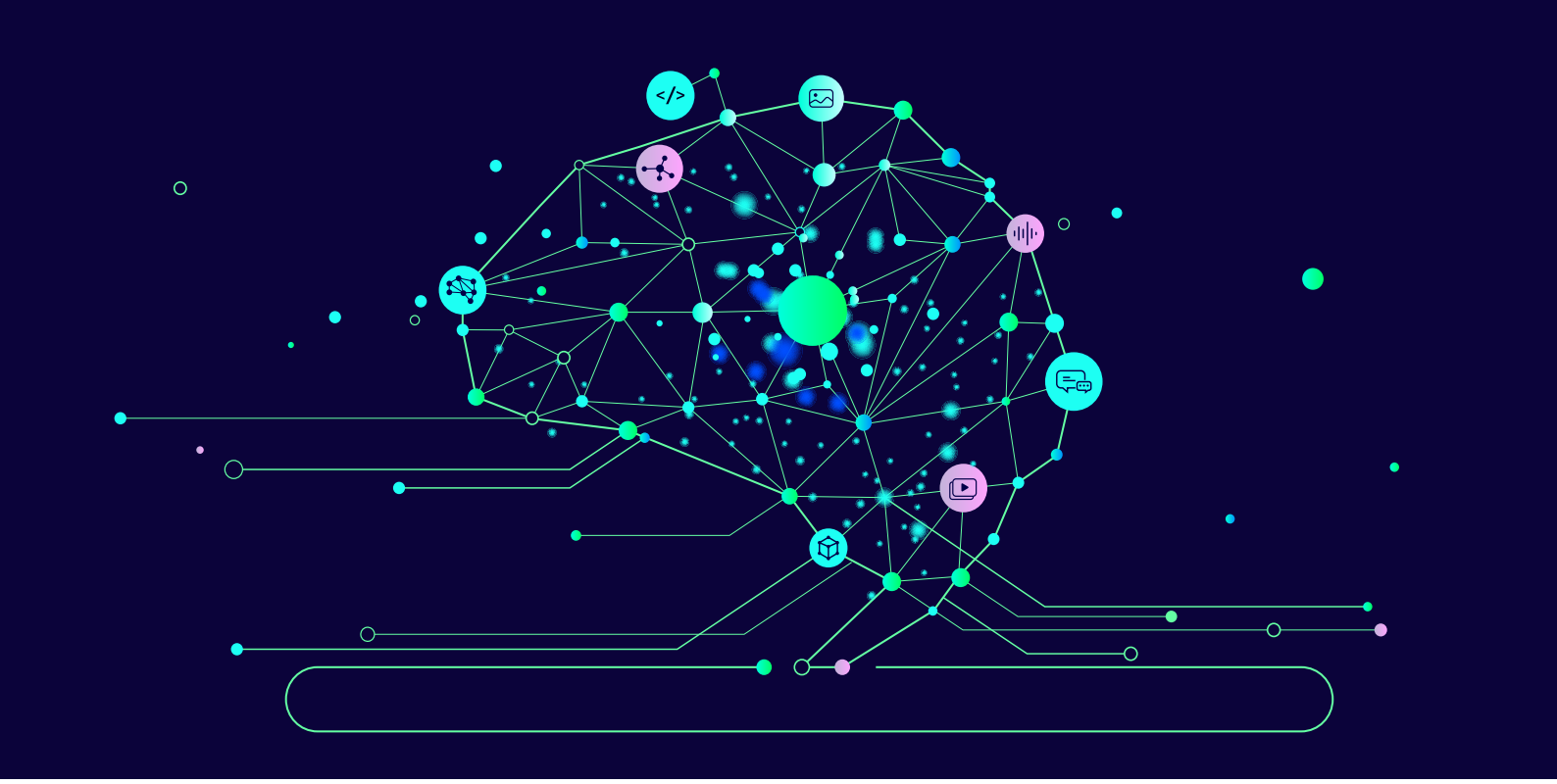
[28] Kim Martineau, "What is retrieval-augmented generation?", IBM Research Blog, 22 August 2023, https://research.ibm.com/blog/retrieval-augmented-generation-RAG.

[29] Xu, Jain, and Kankanhalli, "Hallucination is Inevitable", 12.

[30] In general, an increase in scale is achieved through a combination of factors, including computing power, number of parameters, size of datasets.

[31] Wei et al., "Emergent Abilities of Large Language Models", arXiv, 15 June 2022, https://arxiv.org/abs/2206.07682 (also published in *Transactions on Machine Learning Research*, 31 August 2022, https://openreview.net/pdf?id=yzkSU5zdwD).

[32] Note that in the field of deep learning, emergence is considered an inherent property simply because the behaviour of a model is difficult to predict from its many parameters. For LLMs, the more recent use of the term 'emergence' is associated with sudden jumps in performance as well as unpredictability, which occur as the model reaches a certain scale.



'emergence' is that scaling leads to unpredictable abilities of the model, and these abilities cannot be directly predicted by extrapolating the performance of smaller models.[33] Some emergent abilities may occur in smaller models as well, for a range of reasons, but scale does appear to be highly correlated to emergence.[34] While research on emergence remains contentious, it has highlighted significant concerns associated with emergent abilities – importantly, that LLMs can exhibit behaviours that their developers did not predict, and furthermore, that there are persistent challenges and gaps to developing metrics to predict future behaviour.[35]

Further, understanding how a model might behave in the future cannot be entirely inferred during evaluations.[36] Model evaluations typically assess existing capabilities after a model had been trained. Moreover, it has been shown that there are techniques which can be employed *after* training to improve LLM abilities (e.g. ways of prompting which make the model perform better at certain tasks without the need for further training), a property which may be overlooked during the evaluation of a model.[37]

---

[33] Wei et al., "Emergent Abilities of Large Language Models", 2.

[34] Ibid., 7–8.

[35] Thomas Woodside, "Emergent Abilities in Large Language Models: An Explainer", Center for Security and Emerging Technology, 16 April 2024, https://cset.georgetown.edu/article/emergent-abilities-in-large-language-models-an-explainer/; interview Richard Carter (4 September 2024); see Anna Rogers, "A Sanity Check on 'Emergent Properties' in Large Language Models", Hacking Semantics, 15 July 2024, https://hackingsemantics.xyz/2024/emergence/.

[36] For an extensive discussion of evaluation challenges of LLMs, see Anwar et al., "Foundational Challenges in Assuring Alignment and Safety of Large Language Models", 50–54.

[37] Ibid. This also applies to RAG, which by retrieving information from external sources and using that information in the generation process, reduces the need to continuously train the model and introduce new data.

During use, LLMs also present significant security vulnerabilities to malicious attacks, particularly in the form of **jailbreaking** and **prompt injection**.

Training methods such as supervised fine-tuning, discussed in the previous section, aim to align LLMs to human values, and reduce the likelihood of generating harmful content. Jailbreak attacks aim to manipulate input prompts in a way that will lead the model to bypass built-in safeguards and to generate malicious outputs. There are several methods to jailbreak a model. For example, a method called **DAN** ("Do Anything Now") is a common jailbreaking technique which compels the model to generate outputs beyond its parameters, essentially asking it to take on the role of "DAN", a model without rules.[38] Similarly, **prompt injection** can mislead the model to take in malicious instructions as benign instructions, propelling the LLM to leak sensitive data, spread false information, etc. In *indirect* prompt injection, the attacker embeds the prompt injection (i.e., malicious commands) through third-party content, such as webpages or other documents, and in a way that is not recognizable by users.[39]

Prompt injection attacks and jailbreaking are among the most pressing security vulnerabilities of LLMs (possibly exacerbated in the future

BOX 3.

Jailbreaking vs. Prompt Injection Attacks

Jailbreaking and prompt injection are sometimes used interchangeably, but they are technically different; prompt injections can pave the way to jailbreak a model and some jailbreaking tactics can facilitate prompt injection. [40]

Anwar et al.[41] summarize the difference between the two as follows: in **jailbreaking**, the adversary's goal is to circumvent the restrictions embedded in the model by the *model developer*; in **direct prompt injection**, the adversary aims to circumvent restrictions placed by the *application developer* rather than the model creator.

---

[38] Matthew Kosinski and Amber Forrest, "What is a prompt injection attack?", IBM, 26 March 2024, https://www.ibm.com/topics/prompt-injection; Blessin Varkey, "Jailbreaking Large Language Models: Techniques, Examples, Prevention Methods", Lakera, 19 September 2023, https://www.lakera.ai/blog/jailbreaking-large-language-models-guide#what-is-jailbreaking-in-llms.

[39] German Federal Office for Information Security, "Indirect Prompt Injections. Intrinsic Vulnerability in Application-Integrated AI Language Models", 21 July 2023, 4, https://www.bsi.bund.de/SharedDocs/Cybersicherheitswarnungen/EN/2023/2023-249034-1032.pdf?__blob=publicationFile&v=5.

[40] See Kosinski and Forrest, "What is a prompt injection attack?".

[41] Anwar et al., "Foundational Challenges in Assuring Alignment and Safety of Large Language Models", 65. Note that the distinction between model creators and application developers is not always clear-cut. In some cases, the roles can be the same (if a company creates both the application and the underlying model) but in others, the two are different, for example when the application developers obtain the licence to access the model and will build their own software based on it.

as LLMs will be used in new applications), and they do not require extensive technical knowledge to execute. While patching options may be possible for individual jailbreaks, ensuring a model comes with zero vulnerabilities remains extremely unlikely – as is the possibility of ever achieving "absolute security".[42]

Finally, there are specific risks to consider for **open versus closed models**. Open models have their weights (the parameters which determine the connection between inputs and outputs) public, and they can be downloaded and fine-tuned by anyone on their computing hardware. Open models have been incredibly helpful for the research community, yet their defining characteristic (the public availability of the weights), also means that it is easier to steer the model in whichever direction the user wants, and it is effectively very difficult to monitor how these models are subsequently modified and used.

In contrast, closed models do not have their weights available publicly and their accessibility and functionality are stringently limited by the original developer.[43] These key differences impact the underlying security concerns. Closed models are more difficult to jailbreak, an important reason being sociotechnical: companies that own these models dedicate significant resources to preventing attacks and leaking of weights.[44] The estimation of trade-offs between open versus closed models should not be oversimplified, however, as closed models are not immune to attack, and there are open models, particularly those developed by prominent well-resourced actors, which lend themselves to a certain degree of oversight.[45]

The description of the technology's fundamental characteristics as well as its key areas of risks, covered in this first part, already announce potential implications of the use of LLMs in the context of international security. The second part of this primer illustrates concrete examples of uses as well as risks posed by LLMs in greater detail.

---

[42] Interview Gitta Kutyniok (9 October 2024); this is arguably the case for all AI systems, not only for LLMs.

[43] Kyle Miller, "Open Foundation Models: Implications of Contemporary Artificial Intelligence", Center for Security and Emerging Technology, 12 March 2024, https://cset.georgetown.edu/article/open-foundation-models-implications-of-contemporary-artificial-intelligence/; interview Peter Hase (10 September 2024).

[44] Interview Peter Hase (10 September 2024); interview Richard Carter (4 September 2024).

[45] Miller, "Open Foundation Models: Implications of Contemporary Artificial Intelligence"; moreover, considerations of risks and benefits of open versus closed models vary widely across policy and research communities.

2. Large Language Models and International Security: Applications, Uses and Misuses

This section focuses on the relevance and impact of LLMs in the context of international security. Specific examples of use cases and applications illustrate how LLMs can be instrumental to a host of military and intelligence tasks, but also to the proliferation of weapons or methods of attack, or as a multiplier of threats. The following brief sections are illustrative but not exhaustive; as the technology evolves, new areas of use and new risks may surface.[46]

A. Defence applications

Planning and decision support

The ability of LLMs to parse through troves of data, to identify patterns and synthesize information in a manner which exceeds the capabilities of other NLP technologies, explains the nascent military interest in LLM-powered applications for the completion of a range of tasks.

For example, LLMs could be leveraged to **assist in military planning**, to **generate courses of action** and simulations, to **automate scenario planning**, or to **enhance decision-making**, including through accelerated data processing or identification of threat responses.[47] Integrated in battle management software, LLMs can sift through diverse sources of data to carry out tasks integral to the decision-making process – as shown in a demo by a technology company in 2023 that linked, under a unified interface, intelligence collection and query, permitting to generate courses of action, which could be sent up the chain to a higher command level for more in-depth analysis.[48] In another demonstration by a US technology company, an LLM system was fed 60,000 pages of open-source data, including military documents, and was tasked to provide answers on deterrence strategies and potential winners in a regional conflict in Asia. The model produced a response with explanations within seconds.[49]

Such military uses of LLMs are, by all current indications, in an experimental or exploratory

---

[46] This paper does not cover, for example, autonomous LLMs, an area of relatively limited research thus far, which includes applications enabling LLMs to take action in the real world and even control robotic systems.

[47] William N. Caballero and Phillip R. Jenkins, "On Large Language Models in National Security Applications", arXiv, 3 July 2024, https://arxiv.org/abs/2407.03453; Max Lamparth and Jacquelyn Schneider, "Why the Military Can't Trust AI", *Foreign Affairs*, 29 April 2024, https://www.foreignaffairs.com/united-states/why-military-cant-trust-ai.

[48] Ian Reynolds and Ozan Ahmet Cetin, "War is messy. AI can't handle it.", *Bulletin of the Atomic Scientists*, 14 August 2023, https://thebulletin.org/2023/08/war-is-messy-ai-cant-handle-it/.

[49] Katrina Manson, "The US Military is Taking Generative AI Out for a Spin", *Bloomberg*, 5 July 2023, https://www.bloomberg.com/news/newsletters/2023-07-05/the-us-military-is-taking-generative-ai-out-for-a-spin.

phase.[50] However, military organizations continue to probe the utility of LLM-powered applications for defence, as demonstrated by ongoing efforts to test potential use cases, including through sponsored hackathons and military exercises. For example, the US Air Force Test Centre 5th Hackathon in 2024 focused on LLMs, showcasing the utility of LLMs for producing flight test documentation that typically takes weeks to elaborate and includes extensive details on test parameters, safety procedures, etc.[51] In China, studies examined the use of LLM capabilities to predict adversarial behaviour in combat and to improve simulations.[52]

Intelligence

LLMs could be particularly suited for tasks related to intelligence work (in and outside of defence) due to their distinct ability to process massive amounts of data and summarize unstructured information – processes that are generally labour intensive and time-consuming. There are several possible concrete uses of LLMs in intelligence work and analysis. In combination with other statistical or machine learning tools, LLMs can **improve information processing**, **data classification** and **analysis**.[53]

A 2023 study identified five practical uses of LLMs in intelligence work:[54] **productivity assistants** (proofreading correspondence, automating certain repetitive tasks, etc.); **automated software development and cybersecurity** (using LLMs for automating software development and understanding vulnerabilities); **automated generation of intelligence reports** (arguably, a less likely use of LLMs for finished reports, but a potential use for early stages of report writing); **knowledge search** (extraction of knowledge from massive data sets, distilling facts from text, identifying relationships between entities); **text analytics** (summarizing texts, including the possibility to request the model to provide more extensive details on specific, targeted themes).

Early uses of generative AI in intelligence are known to operate on open-source data (publicly or commercially available) and carry out functions such as the annotation of summaries and responding to follow-up queries from analysts.[55]

---

[50] In the United States, for example, both the Air Force and the Marine Corps are known to experiment with LLMs in some ways, including for military planning tasks, coding and administrative tasks, and for wargaming. See Caballero and R. Jenkins, "On Large Language Models in National Security Applications", Lamparth and Schneider, "Why the Military Can't Trust AI".

[51] Jordan Conner et al., "US Air Force Hackathon: How Large Language Models Will Revolutionize USAF Flight Test", Databricks, 9 February 2024, https://www.databricks.com/blog/us-air-force-hackathon-how-large-language-models-will-revolutionize-usaf-flight-test.

[52] Christopher McFadden, "China train AI-general to predict 'enemy humans' on the battlefield", *Interesting Engineering*, 14 January 2024, https://interestingengineering.com/military/china-training-ai-predict-humans; Stephen Chen, "China's military lab AI connects to commercial language models for the first time to learn more about humans", *South China Morning Post*, 12 January 2024, https://www.scmp.com/news/china/science/article/3248050/chinas-military-lab-ai-connects-commercial-large-language-models-first-time-learn-more-about-humans.

[53] Caballero and Jenkins, "On Large Language Models in National Security Applications", 10.

[54] Adam C and Richard Carter, "Large Language Models and Intelligence Analysis", Center for Emerging Technology and Security, Expert Analysis, July 2023, 7, https://cetas.turing.ac.uk/sites/default/files/2023-07/cetas_expert_analysis_-_large_language_models_and_intelligence_analysis.pdf.

[55] Caballero and Jenkins, "On Large Language Models in National Security Applications", 6; Frank Bajak, "Takeaways: How intelligence agencies are cautiously embracing generative AI", TechXplore, 23 May 2024, https://techxplore.com/

The large-scale use of LLMs in intelligence work remains subject to ongoing study and testing, and that includes both considerations about the technology's suitability for certain intelligence tasks and addressing challenges of human–machine interaction.[56]

Training and wargaming

LLMs can be leveraged for training and wargaming, providing innovative ways to **optimize scenarios**. As LLMs can be employed to run thousands of iterations, they can continue to **adapt strategies** and generate more **complex alternative scenarios**, **enhancing the overall training and learning experience**, as well as **cost effectiveness**.[57] This has been the working premise for some of the early explorations of LLMs for wargaming. For example, in February 2024, the NATO TIDE Hackathon, co-hosted by the Netherlands, had a dedicated segment for wargaming and LLMs, underscored by the belief that LLMs can render wargame simulations more immersive and realistic and ultimately lead to "better decision-making" and "more effective military operations".[58]

The case for LLMs has also been made in terms of their expected higher **replication** value. A model can be technically trained to represent diverse, even competing stakeholders, to generate different role players and have the same trainees test different possible situations and adversarial contexts. The combination of these elements allows for learnings outside of rigid parameters and permits a more likely replication of insights into actual decision-making.[59]

However, recent studies of the behaviour of LLM-based agents in simulated wargames revealed a tendency for escalation (in marked contrast to human behaviour in similar wargames and real-world scenarios), different patterns of escalation behaviours *among* models,[60] as well as some notable differences in strategic preferences between human players and the tested models.[61] These problems cannot be easily fixed, even with fine-tuning, and add to potential challenges of deployment.

---

news/2024-05-takeaways-intelligence-agencies-cautiously-embracing.html.

[56] See Anna Knack, Richard J. Carter, and Alexander Babuta, "Human-Machine Teaming in Intelligence Analysis. Requirements for developing trust in machine learning systems", Centre for Emerging Technology and Security, December 2022, https://cetas.turing.ac.uk/sites/default/files/2022-12/cetas_research_report_-_hmt_and_intelligence_analysis_vfinal.pdf.

[57] Benjamin Jensen, Yasir Atalan, and Dan Tadross, "It Is Time to Democratize Wargaming Using Generative AI", Center for Strategic and International Studies, 22 February 2024, https://www.csis.org/analysis/it-time-democratize-wargaming-using-generative-ai; Caballero and Jenkins, "On Large Language Models in National Security Applications", 4.

[58] NATO, "TIDE Hackathon Spotlight: Wargaming Large Language Module Challenge", 22 February 2024, https://www.act.nato.int/article/tide-hackathon-spotlight-wargaming-llm-challenge/.

[59] Jensen, Atalan, and Tadross, "It Is Time to Democratize Wargaming Using Generative AI".

[60] Juan-Pablo Rivera et al., "Escalation Risks from Language Models in Military and Diplomatic Decision-Making, arXiv, 7 January 2024, 8–9, https://arxiv.org/abs/2401.03408; note this study tested off-the-shelf models and acknowledged "the agents could have been made more or less 'safe' or escalatory with specific prompting or fine-tuning", 10.

[61] Max Lamparth et al., "Human vs. Machine: Language Models and Wargames", arXiv, 6 March 2024, 6–7, https://arxiv.org/abs/2403.03407v1.

BOX 4.

LLMs in defence: push factors and early assessment of impact

Although the transition from proof of concept to effective and wide-scale deployment is not automatic or even certain in all cases, especially as the technology continues to present risks of reliability and security, there is strong interest to explore opportunities afforded by LLMs.

This surge in interest exists against the backdrop of emerging policy and institutional back-up to explore opportunities of generative AI, in general, and LLMs in particular. For example, in August 2023, the US DoD established Task Force Lima dedicated to analysing and integrating generative AI tools across the DoD; the Defence AI Playbook released by the Ministry of Defence of the United Kingdom in January 2024 mentions ongoing work to exploit benefits of LLMs for defence;[62] and the revised NATO AI Strategy from July 2024 acknowledges the critical importance of using generative AI technology, wherever applicable.[63]

The use of LLM technology is predominantly envisaged for support functions, and not to lead on strategic decision-making or as replacement for human operators. However, that need not downplay the potential impact of LLMs on defence. LLMs could have a transformational effect on the work of military organizations, from the more mundane execution of tasks to the reorganization of workflows. Further, while LLMs may not be used to *lead* on important decisions, their use can *influence* (and add digital speed to) strategic planning.[64]

In a more profound sense, and by virtue of how LLMs can help guide decision-making, their use can invite a reconsideration of epistemology in the military profession.[65] An exercise with an experimental LLM for military planning, which also included data on adversary doctrine, showed how users of the model were able to refine their courses of action, to visualize and gain better understanding of the adversary's approaches across several environments.[66] Such uses of LLMs can directly impact how situational understanding emerges, or how historical and doctrinal knowledge is integrated in decision-making.

Critical for an effective use of the technology in the future will also be the appropriate training of personnel, not only on how to best query the system, but also to avoid risks of automation bias and overreliance on systems that are vulnerable to hallucinations, among others.

---

[62] United Kingdom Ministry of Defence, "The Defence AI Playbook", January 2024, **https://assets.publishing.service.gov.uk/media/65bb75fa21f73f0014e0ba51/Defence_AI_Playbook.pdf**.

[63] NATO, "Summary of NATO's revised Artificial Intelligence (AI) strategy", 10 July 2024, **https://www.nato.int/cps/en/natohq/official_texts_227237.htm**.

[64] Caballero and R. Jenkins, "On Large Language Models in National Security Applications", 10–12; Benjamin Jensen and Dan Tadross, "How large-language models can revolutionize military planning", *War on the Rocks*, 12 April 2023, **https://warontherocks.com/2023/04/how-large-language-models-can-revolutionize-military-planning/**.

[65] Jensen and Tadross, "How large language models can revolutionize military planning".

[66] Ibid.

B. Malicious use cases

Proliferation of biological weapons

The ease of access to LLMs, combined with persistent security vulnerabilities, has raised significant concerns about their potential misuse to help develop weapons, particularly biological weapons.

Several studies flagged the risk that LLMs may **help reduce the barriers to highly specialized knowledge** and provide malicious actors, including untrained and non-expert actors, with the necessary information to be able to create dangerous biological agents. Hypothesized pathways to do this would include, for example, relying on **LLMs in the brainstorming process**, for technical assistance, or to help simulate parts of the process.[67]

One exercise showed vulnerabilities of LLMs to jailbreaking and 'Do Anything Now' prompts (discussed in the first section of this paper) for generating harmful information,[68] and another industry-led study revealed the model could produce harmful expert-level biological information (though not consistently across all areas of study) and that such capabilities tend to expand as models get larger.[69]

Other recent studies, however, pointed to marginal impacts of LLMs on the risks of biological attacks. One evaluation concluded that mild improvements in performance could be observed for some metrics, such as accuracy and completeness of tasks, but these results did not appear statistically significant and not very different from what the 'regular Internet' can already provide as a resource.[70] Another study concluded that LLMs do not increase the viability of biological attacks (compared to similar plans created without LLM assistance) in the planning phase (the study did not consider the execution phase)[71] – although it should be noted the study took place under specific conditions and premises.

The conclusions of current forecasts show LLMs to be of limited use for biological weapons production efforts. However, this is not a certain assumption for the future, not only because the capacity of malicious actors to use future models may expand but also because LLMs can assist in other ways, short of outlining a recipe or plan for weapons development. For example, LLMs could support the efforts of actors who have some training in one domain (e.g. molecular biology) more rapidly access information about virology and

---

[67] See Ian Stewart, "A Framework to Evaluate the Risks of LLMs for Assisting CBRN Production Processes", James Martin Center for Nonproliferation Studies, February 2024, https://nonproliferation.org/a-framework-to-evaluate-the-risks-of-llms-for-assisting-cbrn-production-processes/.

[68] Emily H. Soice et al., "Can large language models democratize access to dual-use biotechnology?", arXiv, 6 June 2023, 2–3, https://arxiv.org/abs/2306.03809.

[69] Anthropic, "Frontier Threats Red Teaming for AI Safety", 26 July 2023, https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety.

[70] Tejal Patwardhan et al., "Building an early warning system for LLM-aided biological threat creation", OpenAI, 31 January 2024, https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation/.

[71] Christopher A. Mouton, Caleb Lucas, and Ella Guest, "The Operational Risks of AI in Large-Scale Biological Attacks. Results of a Red-Team Study", RAND, 25 January 2024, https://www.rand.org/pubs/research_reports/RRA2977-2.html.

infection agents.[72] LLM assistance may also be leveraged to help triage or show how experiments went wrong, improving the feedback loop that is part of the process of developing biological agents; future uses of multimodal models, which could interpret visual inputs, for example, could enhance this assistance.[73]

Finally, the overall risks of LLMs are also assessed against the fact that developing biological weapons is known to be a very complex process requiring more than access to information. Even if LLMs could help distil scientific knowledge into actionable steps, there are significant constraints to move from ideation and planning to production, storage, and dissemination of *physical* biological weapons.[74] Accordingly, the potential of LLMs to be misused in significant ways would also require other further developments both in the technology itself and the means of operationalizing the information provided by LLMs.

Cyber attacks

Another area of potential risks for LLMs misuse is cybersecurity.[75] Key concerns here are that LLMs could be employed to assist in **generating malicious software**, or malware, as well as serve for **social engineering attacks** – their ability to process natural language making them particularly effective for creating personalized **spear phishing messages**.[76]

LLMs have made programming more accessible, but their ability to create sophisticated operational malware remains, to date, limited. Some limitations are due to the fact that malware requires more than writing the code, it also depends on elements such as the ability to exploit vulnerabilities on the targeted system or device. Further, a limiting factor is in the training data, assuming that LLMs are not trained on sophisticated examples of malware, which are more scarcely available in public sources.[77]

However, less complex malware can be created, and research has demonstrated the ability to jailbreak models (e.g. by posing as a 'cybersecurity researcher') for malicious purposes. This may not be highly disruptive in itself as such code appears to be weaker than what is already discoverable on the Internet, but it does signal that LLMs could lower the entry barrier for less sophisticated attacks, and for (networks of) less sophisticated actors.[78]

---

[72] Sarah R. Carter et al., "The Convergence of Artificial Intelligence and the Life Sciences: Safeguarding Technology, Rethinking Governance, and Preventing Catastrophe", The Nuclear Threat Initiative, 30 October 2023, 24–26, https://www.nti.org/analysis/articles/the-convergence-of-artificial-intelligence-and-the-life-sciences/.

[73] Bill Drexel and Caleb Withers, "AI and the Evolution of Biological National Security Risks. Capabilities, Thresholds and Interventions", Center for a New American Security, 13 August 2024, 18, https://www.cnas.org/publications/reports/ai-and-the-evolution-of-biological-national-security-risks.

[74] Ibid.; Carter et al., "The convergence of Artificial Intelligence and the Life Sciences", 24–26.

[75] This section highlights areas of risks stemming from malicious actors, but it should be noted that the connection between LLMs and the field of cybersecurity is multifaceted. For example, the use of LLMs for code generation by legitimate researchers can come with (or reproduce) errors learned during training. Further, outside risks of *attacks*, there is also growing research on the potential of LLMs to be used for cyber *defence* capabilities.

[76] Julian Hazell, "Spear Phishing with Large Language Models", arXiv, 11 May 2023, 3, https://arxiv.org/abs/2305.06972.

[77] Ardi Janjeva, Anna Gausen, Sarah Mercer, and Tvesha Sippy, "Evaluating Malicious Generative AI Capabilities. Understanding inflection points in risk", Center for Emerging Technology and Security, July 2024, 17–18, https://cetas.turing.ac.uk/sites/default/files/2024-07/cetas_briefing_paper_-_evaluating_malicious_generative_ai_capabilities.pdf.

[78] Hazell, "Spear Phishing with Large Language Models", 7.

These challenges add to recently reported cases of malicious LLMs, some on the dark web, such as WormGPT (fine-tuned on malware that does not appear to be very sophisticated but still dangerous), or others developed to demonstrate the potential for misuse, such as PoisonGPT (created to showcase how LLMs can be used for disinformation).[79]

Other identified uses of LLMs by malicious actors have focused on **seeking assistance and advice** on elements or steps of an offensive operation. Examples include using LLMs for queries on specific technical questions, such as communication protocols, vulnerability research, or anomaly detection evasion.[80] Mitigation and response measures against such activities require a lot of resources[81] and remain difficult to implement consistently, especially as malicious actors continue to evolve and adapt their strategies.

Disinformation

The threat posed by LLMs for disinformation, including as part of so-called information operations, has been invoked frequently in recent years. The fact that at the core of LLMs is their ability to generate text can make them tools of choice for spreading misleading or false information. Potential risks posed, or aggravated, by LLMs may include the possibility to rapidly **generate persuasive content** in different formats, such as news articles, as well as devise effective **dissemination strategies**.

Outside of generating content of various lengths (commonly done by steering the prompts[82]), LLMs could also be used for other tasks that **support disinformation activities**, such as rewriting stories or articles from a new perspective, creating new types of narratives that can be used as a basis for conspiracy theories (so-called "seeding"), targeting or tailoring messages to particular groups.[83]

The extent to which LLMs are being actively deployed for disinformation remains an open question for now, including considerations of opportunities and costs.

There are several sociotechnical barriers to the use of LLMs at will and at scale for disinformation campaigns. These include, for example, in-built safety features such as safety filters, which are designed to detect malicious requests and prevent the model from responding, but also considerations of costs, as access and use of LLMs comes with additional costs, while the cost of generating disinformation content *without* LLMs is already estimated to be very low.[84]

---

[79] Kevin Poireault, "The Dark Side of Generative AI: Five Malicious LLMs Found on the Dark Web", Infosecurity Europe, 10 August 2023, https://www.infosecurityeurope.com/en-gb/blog/threat-vectors/generative-ai-dark-web-bots.html; Janjeva et al., "Evaluating Malicious Generative AI Capabilities", 19.

[80] See Microsoft Threat Intelligence, "Staying ahead of threat actors in the age of AI", Microsoft/OpenAI, 14 February 2024, https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/.

[81] This is particularly the case for high-value targets, which typically invest significant resources in defence.

[82] Ivan Vykopal et al., "Disinformation Capabilities of Large Language Models", arXiv, 15 November 2023, https://arxiv.org/abs/2311.08838.

[83] Ben Buchanan, Andrew Lohn, Micah Musser, and Katerina Sedova, "Truth, Lies, and Automation. How Language Models Could Change Disinformation", Center for Security and Emerging Technology, May 2021, https://cset.georgetown.edu/publication/truth-lies-and-automation/.

[84] See Micah Musser, "A Cost Analysis of Generative Language Models and Influence Operations", arXiv, 7 August 2023, https://arxiv.org/abs/2308.03740.

These limitations, however, lend themselves to further scrutiny. First, safety filters are not always effective and, importantly, there is variation in their effectiveness across models.[85] This point is important because malicious actors are not restricted to using LLMs that are tightly monitored, and the proliferation of LLMs, including open models, provides options for malicious actors to explore.

Second, this relates to the question of costs, which is nuanced. For example, research has shown there can be significant cost-savings for malicious actors that use LLMs. For smaller campaigns, fine-tuning open models can be a cost-effective option, while the cost of training a model from scratch is economically viable only for very large campaigns, aiming to generate millions of outputs[86] (though this may still be arguably affordable for resourceful actors, such as State actors).

[85] See Vykopal et al. "Disinformation Capabilities of Large Language Models"; note, moreover, that a significant part of the research on safety in LLMs has predominantly focused on LLMs for the English language.

[86] Musser, "A Cost Analysis of Generative Language Models and Influence Operations", 11.

Conclusion

As the international community moves to advance conversations about the governance of artificial intelligence, the potential impact of LLMs on international security serves to highlight the complexities inherent to AI dual-use technologies. LLMs can bring significant opportunities across domains but can also amplify risks.

This primer provided an overview of the impact of LLMs in the context of international security, showing the technology has the potential to have tangible impact on the work of defence organizations, and on the planning or conduct of military operations, and it can provide new means for malicious actors to inflict harm or proliferate weapons.

However, in each of the case studies briefly covered in this paper, the evaluation of risks of LLMs requires a balanced assessment, which accounts for the realistic conditions of use of the technology, at least in the current context. It is important to assess the risks of LLMs beyond over-hyped or exaggerated interpretations of both the technology's capabilities and the users' – and organizations' – abilities to harness them. Invariably, these assessments cannot be separated from the practical realities of training and deploying LLMs in defence organizations, and particularly for classified networks, or the challenge of deployment on specialized hardware.

That said, it is important for the international community to not dispel the present and future risks of LLMs: the technology has already demonstrated real impact and potential for misuse in a very short time span, and as the technology evolves, it is only expected the risk landscape will evolve as well.

To this end, three actionable points can be useful going forward:

1. Continue and amplify **multi-stakeholder dialogues**: dialogues across government, academia and industry are critical to advance understandings of risks and possible mitigation and governance options.
2. Continue to strengthen the research on **AI safety and security, and the nexus between the two**: this can be promoted in dialogue with industry, as critical players in the development of LLMs, as well as through concrete policies, including in national AI strategies and implementation plans.
3. At the **international and multilateral levels**, actively integrate considerations about the impact of LLMs in ongoing processes (e.g., on biological weapons or on information and communications technologies) and initiatives: these conversations may not be limited to LLMs but more broadly to AI risks – yet dedicated discussions on LLMs can be valuable considering the current prominence of LLMs and generative AI.

This primer offered a broad overview of LLMs in the context of international security. Much more remains to be said about this powerful technology. Promising areas of research include ongoing work on explainability, risk mitigation approaches and red teaming, which point to growing efforts from researchers and from industry players to reduce and manage risks.

The wider geopolitical implications of this technology were not considered for the purpose of this primer but likely may be brought into future conversations. These pertain to the resource

and infrastructural demands linked to the development and scaling of LLMs, which can be a source of tension.

Future UNIDIR research will continue to explore topics related to generative AI and LLMs, including in the context of armed conflict and below the threshold of conflict, as well as to promote dialogue on AI governance.

Bibliography

Anthropic. “Frontier Threats Red Teaming for AI Safety”. 26 July 2023. https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety.

Anwar, Usman et al. “Foundational Challenges in Assuring Alignment and Safety of Large Language Models”. arXiv. 15 April 2024. https://arxiv.org/abs/2404.09932. (also published in *Transactions on Machine Learning Research*. 2 September 2024. https://openreview.net/forum?id=oVTkOs8Pka).

Bajak, Frank. “Takeaways: How intelligence agencies are cautiously embracing generative AI”. TechXplore. 23 May 2024. https://techxplore.com/news/2024-05-takeaways-intelligence-agencies-cautiously-embracing.html.

Bergmann, Dave. “What is self-supervised learning?”. IBM. 5 December 2023. https://www.ibm.com/topics/self-supervised-learning.

Bommasani, Rishi et al. “On the Opportunities and Risks of Foundation Models”. arXiv. 16 August 2021 (revised 12 July 2022). https://arxiv.org/abs/2108.07258.

Buchanan, Ben, Andrew Lohn, Micah Musser, and Katerina Sedova. “Truth, Lies, and Automation. How Language Models Could Change Disinformation”. Center for Security and Emerging Technology. May 2021. https://cset.georgetown.edu/publication/truth-lies-and-automation/.

Burtell, Matthew and Helen Toner. “The Surprising Power of Next Word Prediction: Large Language Models Explained: Part 1”. Center for Security and Emerging Technology. 8 March 2024. https://cset.georgetown.edu/article/the-surprising-power-of-next-word-prediction-large-language-models-explained-part-1/.

C, Adam and Richard Carter. “Large Language Models and Intelligence Analysis”. Center for Emerging Technology and Security. 5 July 2023. https://cetas.turing.ac.uk/sites/default/files/2023-07/cetas_expert_analysis_-_large_language_models_and_intelligence_analysis.pdf.

Caballero, William N. and Phillip R. Jenkins. “On Large Language Models in National Security Applications”. arXiv. 3 July 2024. https://arxiv.org/abs/2407.03453.

Carter, Sarah R. et al. “The Convergence of Artificial Intelligence and the Life Sciences: Safeguarding Technology, Rethinking Governance, and Preventing Catastrophe”. The Nuclear Threat Initiative. 30 October 2023. https://www.nti.org/analysis/articles/the-convergence-of-artificial-intelligence-and-the-life-sciences/.

Chen, Stephen. “China’s military lab AI connects to commercial large language models for the first time to learn more about humans”. South China Morning Post. 12 January 2024. https://www.scmp.com/news/china/science/article/3248050/chinas-military-lab-ai-connects-commercial-large-language-models-first-time-learn-more-about-humans.

Common Crawl. “September 2024 Crawl Archive Now available”. 24 September 2024. https://www.commoncrawl.org/blog/september-2024-crawl-archive-now-available.

Conner, Jordan et al. “US Air Force Hackathon: How Large Language Models Will Revolutionize USAF Flight Test”. Databricks. 9 February 2024. https://www.databricks.com/blog/us-air-force-hackathon-how-large-language-models-will-revolutionize-usaf-flight-test.

Drexel, Bill and Caleb Withers. “AI and the Evolution of Biological National Security Risks. Capabilities, Thresholds and Interventions”. Center for a New American Security. 13 August 2024. https://www.cnas.org/publications/reports/ai-and-the-evolution-of-biological-national-security-risks.

German Federal Office for Information Security. “Indirect Prompt Injections. Intrinsic Vulnerability in Application-Integrated AI Language Models”. 21 July 2023. https://www.bsi.bund.de/SharedDocs/Cybersicherheitswarnungen/EN/2023/2023-249034-1032.pdf?__blob=publicationFile&v=5.

Hazell, Julian. “Spear Phishing with Large Language Models”. arXiv. 11 May 2023. https://arxiv.org/abs/2305.06972.

IBM. “What are large language models (LLMs)?”. https://www.ibm.com/topics/large-language-models.

IBM. “What is a transformer model?”. https://www.ibm.com/topics/transformer-model.

Janjeva, Ardi, Anna Gausen, Sarah Mercer, and Tvesha Sippy. “Evaluating Malicious Generative AI Capabilities. Understanding inflection points in risk”. Center for Emerging Technology and Security. July 2024. https://cetas.turing.ac.uk/sites/default/files/2024-07/cetas_briefing_paper_-_evaluating_malicious_generative_ai_capabilities.pdf.

Jensen, Benjamin and Dan Tadross. “How large-language models can revolutionize military planning”. *War on the Rocks*. 12 April 2023. https://warontherocks.com/2023/04/how-large-language-models-can-revolutionize-military-planning/.

Jensen, Benjamin, Yasir Atalan, and Dan Tadross. “It Is Time to Democratize Wargaming Using Generative AI”. Center for Strategic and International Studies. 22 February 2024. https://www.csis.org/analysis/it-time-democratize-wargaming-using-generative-ai.

Ji, Jessica, Josh A. Goldstein, and Andrew J. Lohn. “Controlling Large Language Model Outputs: A Primer”. Center for Security and Emerging Technology. December 2023. https://cset.georgetown.edu/publication/controlling-large-language-models-a-primer/.

Knack, Anna, Richard J. Carter, and Alexander Babuta. “Human-Machine Teaming in Intelligence Analysis. Requirements for developing trust in machine learning systems”. Centre for Emerging Technology and Security. December 2022. https://cetas.turing.ac.uk/sites/default/files/2022-12/cetas_research_report_-_hmt_and_intelligence_analysis_vfinal.pdf.

Kosinski, Matthew and Amber Forrest. “What is a prompt injection attack?”. IBM. 26 March 2024. https://www.ibm.com/topics/prompt-injection.

Lamparth, Max et al. “Human vs. Machine: Language Models and Wargames”. arXiv. 6 March 2024. https://arxiv.org/abs/2403.03407v1.

Lamparth, Max and Jacquelyn Schneider. “Why the Military Can’t Trust AI”. *Foreign Affairs*. 29 April 2024. https://www.foreignaffairs.com/united-states/why-military-cant-trust-ai.

Manson, Katrina. “The US Military is Taking Generative AI Out for a Spin”. *Bloomberg*. 5 July 2023. https://www.bloomberg.com/news/newsletters/2023-07-05/the-us-military-is-taking-generative-ai-out-for-a-spin.

Martineau, Kim. “What is retrieval-augmented generation?”. IBM Research Blog. 22 August 2023. https://research.ibm.com/blog/retrieval-augmented-generation-RAG.

McFadden, Christopher. “China train AI-general to predict ‘enemy humans’ on the battlefield”. *Interesting Engineering*. 14 January 2024. https://interestingengineering.com/military/china-training-ai-predict-humans.

Microsoft Threat Intelligence. “Staying ahead of threat actors in the age of AI”. Microsoft/OpenAI. 14 February 2024. https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/.

Miller, Kyle. “Open Foundation Models: Implications of Contemporary Artificial Intelligence”. Center for Security and Emerging Technology. 12 March 2024. https://cset.georgetown.edu/article/open-foundation-models-implications-of-contemporary-artificial-intelligence/.

Mitchell, Melanie. “Large Language Models”. MIT Press. Open Encyclopedia of Cognitive Science. 24 July 2024. https://doi.org/10.21428/e2759450.2bb20e3c.

Mouton, Christopher A., Caleb Lucas, and Ella Guest. “The Operational Risks of AI in Large-Scale Biological Attacks. Results of a Red-Team Study”. RAND. 25 January 2024. https://www.rand.org/pubs/research_reports/RRA2977-2.html.

Musser, Micah. “A Cost Analysis of Generative Language Models and Influence Operations”. arXiv. 7 August 2023. https://arxiv.org/abs/2308.03740.

North Atlantic Treaty Organization. “TIDE Hackathon Spotlight: Wargaming Large Language Module Challenge”. 22 February 2024. https://www.act.nato.int/article/tide-hackathon-spotlight-wargaming-llm-challenge/.

North Atlantic Treaty Organization. Summary of NATO’s revised Artificial Intelligence (AI) strategy. 10 July 2024. https://www.nato.int/cps/en/natohq/official_texts_227237.htm.

Patwardhan, Tejal et al. “Building an early warning system for LLM-aided biological threat creation”. *OpenAI*. 31 January 2024. https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation/.

Poireault, Kevin. “The Dark Side of Generative AI: Five Malicious LLMs Found on the Dark Web”. Infosecurity Europe. 10 August 2023. https://www.infosecurityeurope.com/en-gb/blog/threat-vectors/generative-ai-dark-web-bots.html.

Reynolds, Ian and Ozan Ahmet Cetin. “War is messy. AI can’t handle it.” *Bulletin of the Atomic Scientists*. 14 August 2023. https://thebulletin.org/2023/08/war-is-messy-ai-cant-handle-it/.

Rivera, Juan-Pablo et al. “Escalation Risks from Language Models in Military and Diplomatic Decision-Making. arXiv. 7 January 2024. https://arxiv.org/abs/2401.03408.

Rogers, Anna. “A Sanity Check on ‘Emergent Properties’ in Large Language Models”. Hacking Semantics. 15 July 2024. https://hackingsemantics.xyz/2024/emergence/.

Soice, Emily H. et al. “Can large language models democratize access to dual-use biotechnology?”. arXiv. 6 June 2023. https://arxiv.org/abs/2306.03809.

Stewart, Ian. “A Framework to Evaluate the Risks of LLMs for Assisting CBRN Production Processes”. James Martin Center for Nonproliferation Studies. February 2024. https://nonproliferation.org/a-framework-to-evaluate-the-risks-of-llms-for-assisting-cbrn-production-processes/.

The Economist. “A Short History of AI”. 16 July 2024. https://www.economist.com/schools-brief/2024/07/16/a-short-history-of-ai.

Together AI. “RedPajama-Data-v2: An open dataset with 30 trillion tokens for training large language models”. 30 October 2023. https://www.together.ai/blog/redpajama-data-v2.

Toner, Helen. “What are Generative AI, Large Language Models, and Foundation Models?”. Center for Security and Emerging Technology. 12 May 2023. https://cset.georgetown.edu/article/what-are-generative-ai-large-language-models-and-foundation-models/.

United Kingdom Ministry of Defence. “The Defence AI Playbook”. January 2024. https://assets.publishing.service.gov.uk/media/65bb75fa21f73f0014e0ba51/Defence_AI_Playbook.pdf.

Varkey, Blessin. “Jailbreaking Large Language Models: Techniques, Examples, Prevention Methods”. Lakera. 19 September 2023. https://www.lakera.ai/blog/jailbreaking-large-language-models-guide#what-is-jailbreaking-in-llms.

Vaswani, Ashish et al. “Attention Is All You Need”. 12 June 2017 (revised 2 August 2023). arXiv. https://doi.org/10.48550/arXiv.1706.03762.

Vykopal, Ivan et al. “Disinformation Capabilities of Large Language Models”. arXiv. 15 November 2023. https://arxiv.org/abs/2311.08838.

Wang, Jiongxiao et al. “RLHFPoison: Reward Poisoning Attack for Reinforcement Learning with Human Feedback in Large Language Models”. arXiv. 16 November 2023. https://arxiv.org/abs/2311.09641v2.

Wei, Jason et al. “Emergent Abilities of Large Language Models”. arXiv. 15 June 2022. https://arxiv.org/abs/2206.07682. (also published in *Transactions on Machine Learning Research*. 31 August 2022. https://openreview.net/pdf?id=yzkSU5zdwD.)

Woodside, Thomas and Helen Toner. “How Developers Steer Language Model Outputs: Large Language Models Explained, Part 2”. Center for Security and Emerging Technology. 8 March 2024. https://cset.georgetown.edu/article/how-developers-steer-language-model-outputs-large-language-models-explained-part-2/.

Woodside, Thomas. “Emergent Abilities in Large Language Models: An Explainer”. Center for Security and Emerging Technology. 16 April 2024. https://cset.georgetown.edu/article/emergent-abilities-in-large-language-models-an-explainer/.

Xu, Ziwei, Sanjay Jain, and Mohan Kankanhalli. “Hallucination is Inevitable: An Innate Limitation of Large Language Models”. arXiv. 22 January 2024. https://arxiv.org/abs/2401.11817.

@unidir
/unidir
/un_disarmresearch
/unidirgeneva
/unidir

UNIDIR

Palais des Nations
1211 Geneva, Switzerland

WWW.UNIDIR.ORG