REPORT

Foreign Information Manipulation & Interference: A Large Language Model Perspective

This report focus on the intersection of Foreign Information Manipulation and Interference and Large Language Models. The aim is to give a non-technical comprehensive understanding of how weaknesses in the language models can be used for creating malicious content to be used in FIMI.

RISE RAPPORT 2024:20

Björn Bjurling, Senior researcher, RISE

Andreas Thore, Researcher, RISE

Stella Riad, Director data analysis, RISE

DIGITAL SYSTEMS
DATA ANALYSIS

Foreign Information Manipulation & Interference: A Large Language Model Perspective

Björn Bjurling, Andreas Thore & Stella Riad

RISE Report 2024:20

Foreign Information Manipulation & Interference: A Large Language Model Perspective

Björn Bjurling Andreas Thore Stella Riad
RISE Research Institutes of Sweden AB

March 12, 2024

Abstract

This report gives a snapshot of the literature in the intersection of Foreign Information Manipulation and Interference (FIMI) and Large Language Models. The aim is to give a non-technical comprehensive understanding of *how* weaknesses in the language models can be used for creating malicious content to be used in FIMI. With the aid of a conceptual threat model, we point to both attack and defence strategies.

Key words: Artificial intelligence, FIMI, Large Language Models, disinformation

Disclaimer: The authors are responsible for the content and conclusions of this study.

RISE Report 2024:20
ISBN: 978-91-89896-67-3

Contents

1 Introduction

Foreign Information Manipulation and Interference (FIMI) is an umbrella term for *misinformation*, *disinformation*, *malinformation* and other distortions. FIMI has during the past decades grown into a global threat permeating a vast array of public discourse and communication, not least on social media[57, 14]. FIMI is a threat against democracy, health, and privacy[60, 64, 82].

Recent developments have seen the use of generative Artificial Intelligence (AI) for increasing the impact of operations aiming at FIMI. For example, Large Language Models (LLMs) are capable of creating text that is practically indistinguishable from human texts [35, 42]. LLMs are beginning to be used for controlling botnets for global rapid automated dissemination of malicious content and disinformation [104]. The past years' revolutionizing progress in generative AI for images, video, and audio facilitate multi-model information attacks, and will only keep adding to the difficulty in combating AI-driven FIMI. The interest in using generative AI for FIMI stems from the promise of massive distribution of low-cost propaganda[40]. Moreover, as Goldstein et al. argue, the low cost of setting up such 'troll farms' allows to quickly change campaign focus to adapt to current news events[36]. Jachim et al. argue that the use of generative AI for FIMI is particularly suited for states and state sponsored trolls to further geopolitical agendas through propagation and creation of for example rumors, conspiracy theories, and malicious narratives[40].

There is also an ongoing effort to mitigate generative AI-based FIMI with both technical and policy-based measures being suggested and implemented. While technical measures such as *curation* of training data (for example, to avoid known biases)[9] and *safety alignment*, a method used to align the models behavior to human preference and ethics, [76] may be successful in many respects, they may also be insufficient in others. Safety aligned models can with little effort be instructed to generate unsafe output[69, 106]. In the context of *Large Language Models (LLMs)*, which is the focus of this report, an unsafe output is a *string*[1] that can be used in FIMI, i.e., strings propagating hate, immoral views, discrimination, violence, etc. The need for continuous and ongoing effort to counter AI-based FIMI depends in part on the difficulty in modeling it[14], and in part on the rapid technological development in the field of generative AI. Moreover, there is an asymmetry between threat actors and defenders, where threat actors can focus its resources on, say, one malicious type of content or one particular attack vector, while the defender needs to defend against *all conceivable* threats at all times. Policy-based efforts (see e.g., [95]) to mitigate FIMI are connected with other types of complications. Such approaches often suggest community-based solutions including educational efforts and collaborations between parties with commercial interests and parties with civil responsibilities. Goldstein et al. [36] give an account of such suggestions.

In this report we will make an attempt to add value on the topic of the use of generative AI in FIMI – in the midst of the current avalanche of reports and articles being published every week in the field. Our contribution is that we approach the topic from the perspective of LLMs and the capabilities required for using such models in FIMI, *without going into technical detail* and requiring AI domain knowledge. Our technically focused report should

[1] we will subsequently use the technical word string which informally could be said to mean a part of text (such as a word, or a sentence).

thus be accessible also for practitioners who may have other backgrounds than one in the field of AI.

Thus, the aim for the report is to bridge a perceived gap among existing works in the field. Many reports and articles consider the societal aspects of FIMI, while other considers the technical aspects of vulnerabilities in generative AI models. There seems to be a lot of knowledge encoded into two of the directions, but at the same time it seems that there is little communication inbetween. It is this gap we want to fill by providing tools for describing how the strings that can be used in FIMI are generated. In addition, we point out, where deemed interesting, which technical resources and skills are needed for utilizing LLMs in FIMI. With respect to methodology, the report makes a subjective snap shot of recent and central works in the area with the purpose of bridging this perceived gap.

We assume that *i*) LLMs are pre-trained on datasets that in general are large enough to contain unsafe strings [9], and *ii*) any LLM is capable of generating output that is indistinguishable from at least one example in its training data[2]. With this concept, we can formulate the problem of LLM-based FIMI by saying that the threat actor wishes to make an LLM generate a specific unsafe piece of text while the defender aims for preventing the model from generating unsafe pieces of texts.

The report is structured as follows. We give a background to LLMs and a specification of what our interests with respect to FIMI in Sec. 2. In Sec 3 we introduce a conceptual threat model for bridging the gap between social and technical aspects of LLM-driven FIMI. Sec. 4 considers FIMI from the viewpoint of risk posed by open-source LLMs. We also give insights to requirement on resources and skills for exploiting particular weaknesses. Sec. 5 discusses briefly threats from LLM-driven botnets. Sec. 6 considers three mitigation strategies of relevance for our LLM focused topics. Sec. 7 contributes with three recommendations based on the material in the report and Sec. 8 is the conclusion.

2 Background

There are no standard vocabularies and definitions of terms in this field. Therefore, we will define the terms used in this section as well as putting our work in context.

2.1 Foreign Information Manipulation & Interference

In this report we use FIMI (Foreign Information Manipulation and Influence) as an umbrella term for the concepts of *misinformation*, *disinformation*, and *malinformation*. We follow the common terminology from Wardle et al. [95] (and others, e.g., [53, 78]) and define the concepts respectively as follows.

Disinformation is inaccurate information that is intentionally spread to cause harm by misleading or deceiving. Misinformation is false content that is shared in good faith. Malinformation is to describe genuine information that is shared with an intent to cause harm.

[2]The second requirement says that it matters what data we train the LLMs on (and on how unsafe output is prevented), which is the topic of this report. That the output may be indistinguishable from human generated text is more of a psychological/cognitive matter, and that is not the primary concern of the report.

In the setting of LLMs, where the training often is made on uncurated data collected from social media and other public fora, we include *rumors* and *urban legends* as examples of problematic content that could be used in FIMI [78]. For the present setting where we focus on FIMI enabled by LLMs, we can specify the three main concepts more precisely:

Misinformation : non-factual output from an LLM

Disinformation : the intentional act of a threat actor to disseminate misinformation extracted from an LLM.

Malinformation : the intentional act of a threat actor to use the output of an LLM out of context with malicious intent.

This report does not consider goals, operations, or tactics of FIMI. Of more interest is the potential to generate output of LLMs that can be used in operations aimed at FIMI. Such output usually reflect biased, immoral, sexist, racist, or non-factual aspects of the data used for the training of LLMs. Furthermore, this report does not go into classifying different types of toxicity that can be generated with LLMs. Thorough exploration of risk, toxicity, and harm that can be used in FIMI are given in recent surveys and reports such as Refs. [34, 17, 77, 97, 98].

Yang et al. [103] and Chang et al. [16] give recent surveys of the capabilities of LLMs with relevance to this report. Dong et al. [28] give a recent and thorough survey on attack patterns against LLMs. Shayegani et al. [75] give a thorough survey of LLM vulnerabilities.

Recently, there have been several works published with an aim similar to that of this report: a more fine-grained investigation of the connection between LLMs and FIMI, exploring the aspects of LLMs that make them susceptible for generating toxic output. Lu [49] points to the need for investigation of *how* LLMs are used to generate output that can be used for FIMI, rather than merely enumerating and classifying toxic material. Wolf et al. [99] give a rigorous model of how output from LLMs can be exploited in FIMI. Chen et al. [17] make connections between FIMI types and LLM exploits.

2.2 Generative AI

The field of AI encompasses many different types of algorithms and methods. It is out of scope for this report to paint a complete picture of the field. However, it can be divided into a hierarchy of domains, presented in Fig. 1. While the field of AI is very broad and has been active since the 1950s, the research in the domain called deep learning has been intense during the past 15 years. Deep learning is a set of methods that use deep neural networks, i.e., neural networks containing more than one so-called hidden node layer (Fig. 2), to find complex patterns and relationships in large datasets. Generative AI is a subdomain of deep learning, and is now a significant driver of AI-related investments by large corporations, venture capitalists, and retail investors[3][4]. In contrast to non-generative AI, where the output

[3]https://dealroom.co/guides/generative-ai

[4]https://www.goldmansachs.com/intelligence/pages/ai-investment-forecast-to-approach-200-billion-globally-by-2025.html

is typically a classification of a data point, such as an image of a cat, or a numerical prediction based on a set of data points, such as the price of a house, generative AI is characterized by the ability to generate new data, in the form of text, image, and video, from a *prompt* – an input or query to the AI in order to elicit a response from the model – consisting of data of either the same data type or a different one, see Fig. 3.

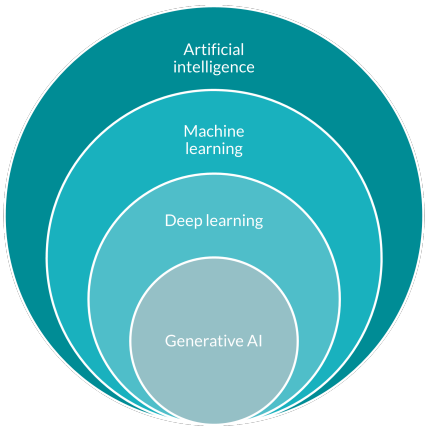

Figure 1: The different types of AI.

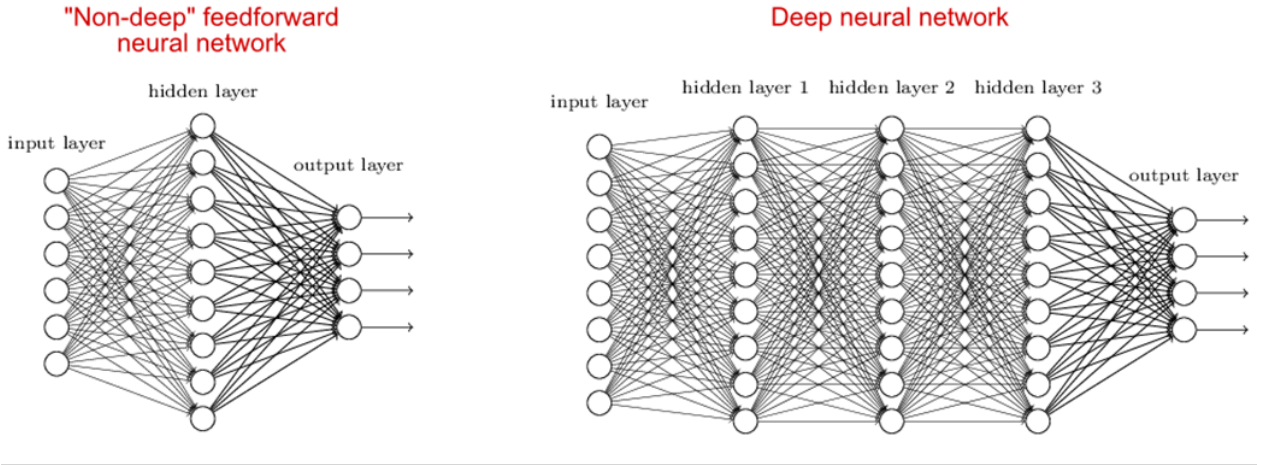

Figure 2: A "non-deep" (left) vs. a deep neural network (right).

Generative AI comes in many shapes and sizes. The state-of-the-art *image generators* are rapidly improving, and it is becoming increasingly difficult for humans to distinguish AI-generated images from real ones. While models such as Midjourney[5], Stable Diffusion[6] and DALL-E[7] had difficulties generating images of humans without considerable artifacts one year

[5]https://www.midjourney.com/home
[6]https://stability.ai/news/stable-diffusion-3
[7]https://openai.com/dall-e-3

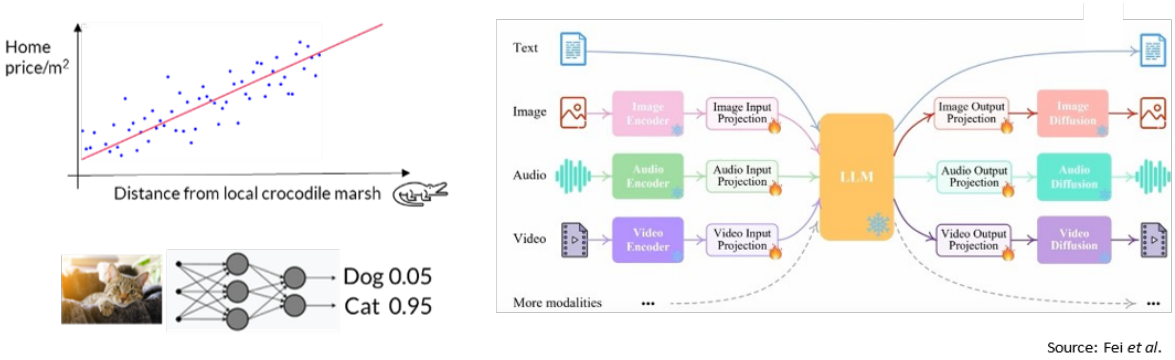

Figure 3: Non-generative (left) vs. generative AI (right).

back, the latest versions are in general capable of generating highly believable images. *Audio generators* in the form of speech generation models such as models from ElevenLabs[8] can generate realistic voices, with a very good grasp of speech modulation and breathing sounds. The same company also provides voice cloning models, which are becoming increasingly believable. *Video generators* seem to have come one step closer to realistic text-to-video generation with Sora by Open AI[9], which was announced and demonstrated (although not publicly released) in February 2024. *Multimodal models* or Large multimodal models (LMMs) – generative AI models that can both ingest and generate more than one data type – are considered a natural development of the field of generative AI. As an example GPT-4V[10] is multimodal and can provide a textual analysis of both images and text. The most advanced versions of Gemini[11], Gemini 1.0 Ultra and Gemini 1.5 Pro, can ingest text, images, and video, and can do image analysis as well as generate images. Multimodal "any-to-any" models, depicted on the right in Fig. 3, are actively discussed in the literature and are eventually set to emerge[100]. Arguably, the most notable generative models are still the LLMs, which is the focus of this report and will be further introduced below.

Like many technologies, generative AI is a double-edged sword: while its potential to benefit humanity is enormous, it is clear that it can also be used for deeply destructive purposes. Perhaps the biggest threat that it currently poses is as a method to enhance, generate, as well as propagate FIMI. Digital FIMI can consist of many different data types, either alone or in combination, and while at present generative AI models perform best with respect to text output, the steady increase in performance that is observed for other data types makes it highly likely that generative AI will eventually dominate or completely replace conventional methods for FIMI.

[8]https://elevenlabs.io/
[9]https://openai.com/sora
[10]https://openai.com/research/gpt-4v-system-card
[11]https://deepmind.google/technologies/gemini

2.2.1 Text generators: Large Language models (LLMs)

Large language models have in recent years both revolutionized language technology and achieved enormous impact on society at large. Initially, types of neural networks such as Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) models, which process and capture relationships in text sequentially, were used. However, there has been a very rapid development following the emergence of *transformers*[87]. A key to their success is that they use a mechanism, so-called *attention*, that makes it possible to capture dependencies that extend across entire documents. Transformers are used, for example, in BERT (Bidirectional Encoder Representations from Transformers)[25] by Google, and more recently in GPT-4 (Generative Pre-trained Transformer) by Open AI[61] and Llama2 by Meta[83].

LLMs can be prompted to write everything from poems, dinner recipes, and short stories, to scientific text and computer code. Some claim that they possess the ability to reason, or at least to simulate reasoning, since they can generate answers to questions that would require deductive or inductive reasoning for a human, in contrast to simple fact retrieval. LLMs work by ingesting all prompts and the corresponding generated replies in a conversation – this is commonly referred to as the LLM's *context*. Based on patterns learned during training, the model then predicts the word, or strictly speaking *token* (commonly consisting of one or several characters, for example in the form of a word), that is most likely follow. At the time of writing (March 2024), the context window is about 128 000 tokens for the most common LLMs, such as OpenAI's GPT-4[12], which implies that it is limited to read, write and reason about texts or code no longer than a typical novel (300 pp). While growing context windows can be expected to enhance the ability of LLMs to be used for creation of FIMI, it should be stressed that even LLMs with very small context windows still can be very capable.

In the last 1.5 years, media and the public have mostly had their attention focused on ChatGPT and to some extent Google's catch-up efforts with Bard[13] and Gemini. These are closed-source[14] models that can only be accessed via online graphical user interfaces or APIs.

Most of the recent developments of LLMs have been driven by hyperscalers, such as Google, Microsoft and Meta, due to the high resource demand, which we elaborate on in Sec. 4.1. However, driven by academic interest and industrial need, significant effort is being put into developing capable open-source models. These pre-trained models are readily available to the public, e.g. on the platforms Hugging Face[15] or ollama[16]. The players in this field span giants like Meta as well as startups and smaller companies such as Mistral[17] and Databricks[18]. These models are typically much smaller in size compared to their closed-source counterparts. Size in this context is determined from the number of *parameters* in the model, which for deep learning models is often used interchangeably with *weights*, which are essentially the strengths of the connections between the nodes in a neural network, i.e., the lines in Fig. 2. While these models are significantly smaller they still perform increasingly

[12]https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo
[13]An LLM chatbot based on the older models LaMDA, PaLM, and PaLM 2.
[14]With one exception: Gemini Nano, the smallest model in the Gemini Family.
[15]https://huggingface.co/
[16]https://ollama.com/
[17]https://mistral.ai/
[18]https://www.databricks.com/

well, especially for specific tasks. As previously pointed out, the landscape changes on an almost daily basis. A snapshot of examples of state-of-the-art open-source models at the time of writing are Mistral [19], the mixture of experts model Mixtral[20], DeepSeek[21] and Llama 2[22]. Other examples of open-source are national initiatives, such as the Nordic models GPT-SW3[23] and Poro[24].

While the open-source models have democratized LLMs, the fact that anyone can, in principle, train them, also opens for FIMI attacks that are potentially far more powerful than attacks using access-restricted closed-source LLMs. In this work, open-source LLMs will thus receive extra focus. The topic is further discussed in Sec 4.

2.2.2 Training generative AI models

Nearly all of the most popular and best performing generative AI models, including all models discussed in this report, are based on the transformer architecture[87]. In the context of LLMs, transformers function as so-called *sequence-to-sequence* (seq2seq) models since they take as input sequences of tokens and transform them into output sequences. Transformers have several advantages to other architectures, the two most consequential ones being that their performance scales very well with parameter count and dataset size [44]. This scaling behavior is taken advantage of when training today's LLMs by using billions of parameters and enormous amounts of data. Coupled with the fact that, during training, the parameters may need millions or even billions of updates, and that, depending on the application, part of the training may require human supervision, means that LLMs are notoriously resource intensive and expensive to train from scratch.

Current state-of-the-art LLMs like GPT-4 are commonly trained in three steps: a pre-training step and two fine-tuning steps called *instruction tuning* and *reinforcement learning from human feedback (RLHF)*[21, 62]. Since attacks on LLMs can be directed to each of these steps, we will briefly elaborate on them below.

Pre-training In the pre-training step the LLMs are trained *unsupervised* – the training reward signal does not rely on human-labeled data – on vast amounts of publicly available[25] and proprietary data[26]. Pre-training typically occurs by the model learning to predict words in a given context through the masking of words to be predicted. Pre-training not only gives the model a general understanding of language and the ability to generate text but also factual knowledge and (to some extent) reasoning capacity.

[19]https://huggingface.co/mistralai

[20]https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

[21]https://huggingface.co/deepseek-ai

[22]https://llama.meta.com/

[23]https://www.ai.se/sv/projekt/gpt-sw3

[24]https://huggingface.co/LumiOpen/Poro-34B

[25]One example is the petabyte-sized datasets from Common Crawl

[26]Recently, OpenAI signed a deal with the media and technology company Axel Springer for data access,https://openai.com/blog/axel-springer-partnership and as the owner of YouTube Google of course has access to petabytes of video.

Instruction tuning The model is then adapted for a specific task and domain through fine-tuning. Instruction tuning is a supervised fine-tuning step where the LLMs are trained on human-labeled datasets, usually consisting of instructions to write something (like a poem, a short story, or a recipe), possible contextual information, and the desired output.

RLHF The RLHF is a more complex and manually demanding step that begins with humans creating labeled data by prompting the instruction-tuned LLMs and collecting the output. These labeled data pairs are fed into a separate *reward model*, which is trained to classify a given prompt-output pair as 'good' or 'bad' based on whether the output fits the prompt. The reward model is then used as a classifier of prompt-output pairs in a training loop where, in each iteration, the LLM is automatically prompted, and prompt-output pairs and their corresponding classification scores are collected into datasets that are fed back into the LLM. The LLM is trained on these datasets to maximize the probability of generating a well-fitting output to any given prompt.

The fine-tuning steps are intended to ensure that the output is more predictable and controllable, and aligns with human expectations and preferences, as well as with ethics and safety requirements. This process is known as *safety alignment* approaches for which the current focus is on RLHF [22, 63]. Another way of safety alignment is by pre-training on carefully curated datasets. As an example the LLM Zephyr-7b [86] was trained on the aligned dataset UltraFeedback [24]. These steps are, of course, highly dependent on the human developers, and thus they present an opportunity for threat actors to introduce significant negative bias in a model, in addition to any bias advertently or inadvertently introduced in the pre-training step through the dataset selection.

As will be discussed in more detail in Sec. 4, pre-training of even small LLMs from scratch is currently limited to large corporations and possibly in government-funded projects, while smaller actors who want direct access to an LLM and to tailor it to their needs have to use open-source models and focus on the fine-tuning steps or on direct modification of the parameters. For closed-source models the only realistic alternatives is to explore methods to manage the output.

3 Conceptual Threat Model

We devise an informal model for describing the use LLMs in FIMI. The threat model provides a bridge between technical LLM concepts and social FIMI concepts. Until recently, work in the field has either been focused on technical aspects vulnerabilities of LLMs, or more on societal aspects of FIMI. This report strives for being more specific by pointing to *how* an LLM is abused in FIMI, as well as being more specific about *how* different vulnerabilities of LLMs can be used in FIMI.

Our threat model is similar to Wolf et al.'s Behavioral Expectation Bounds (BEB)-framework [99]. It should however be noted that the two models have different purposes. Both models assume that an LLM always is capable of generating an unsafe output, regardless of alignment efforts. Moreover, both models describe the attacker as someone who wants to increase the likelihood of generating unsafe output, and the defender as someone who

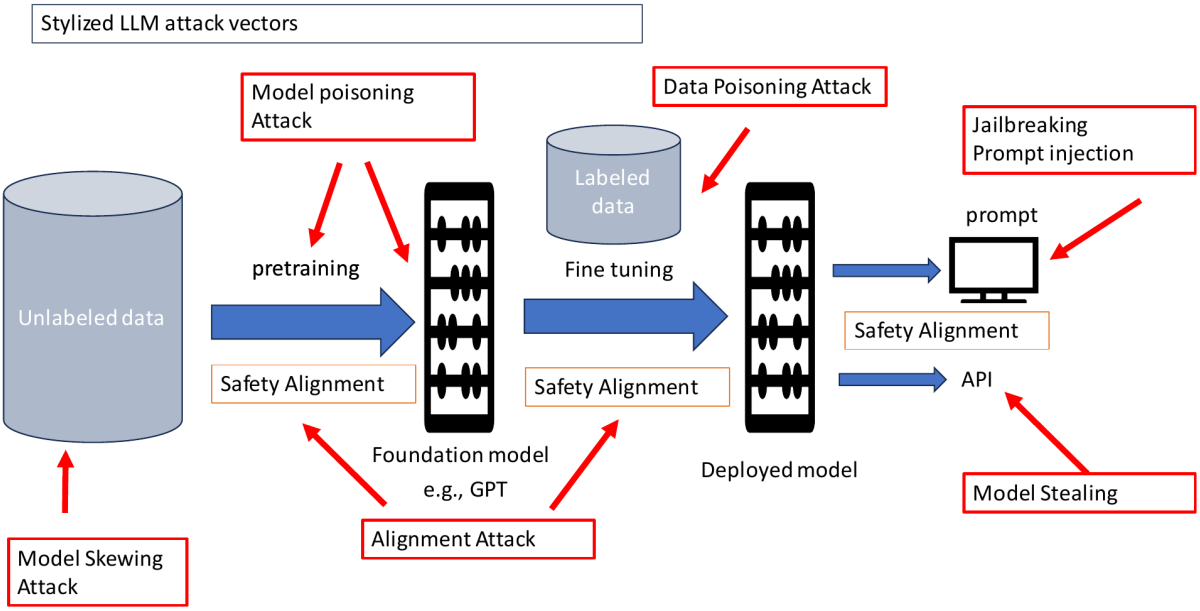

Figure 4: LLM attack types. All attacks except jailbreaking, prompt injection, and model stealing require direct access to the model, either via ownership, hacking, or open-source

wants to decrease the likelihood. While our model is conceptual and aimed as an informal tool for discussing generation of material that could be used in FIMI, the BEB-framework provides a more rigorous probabilistic foundation for the same ideas. Another difference is that our model conceptualizes attacker and defender goals, while the BEB-framework considers more generally a probabilistic model with respect to the strings that can be generated as output from an LLM.

In the introduction we stated that an LLM is capable of generating an output string that is indistinguishable from at least one string in its training data. For modeling generation of strings that could be used for FIMI, we will strengthen that statement and distinguish between *safe* and *unsafe* strings (either as output, examples in training data, or as instructions).

The concept of safe string is defined by ethical AI standards (for example [43]). Such ethical considerations aim for preventing dissemination of material with topics such as harassment, racism, sexism, and hate speech[32], that may be used in for FIMI, potentially leading to call for violence or division, and to threats against human and democratic values[43]. We say that a string is *unsafe* if it can be used in a way that is not sanctioned by a set of AI ethical guidelines (such as those described in [43]). We say that a string is *safe* if it conforms with AI ethical guidelines.

The significance of the distinction is that unsafe output strings are assumed to enable creation of FIMI. A string is said to be *inferred*[27] from the training data of an LLM whenever

[27]Note that this use of the notion of 'inference' is the more general sense from logic than the one used in

there is a positive probability that the string is contained in the output of the LLM for some user input. Then we can formulate the fundamental assumption of our threat model as

> *An LLM is capable of generating an output string that is indistinguishable from some* unsafe *string that can be (statistically)* inferred *from the training data.*

With this formulation, we can characterize the roles of threat actor and defender in terms of the fundamental assumption. The threat actor want to elicit ways of increasing the odds that an LLM generates unsafe output strings, and in this report we consider four such cases, also illustrated in Figure 4:

Pre-training attacks If the threat actor has access to the dataset for pre-training a foundation model, or is able to interfere with routines for selecting pre-training datasets, the model can be skewed to contain training examples of arbitrary unsafe character.

Model poisoning attacks If the threat actor has access to the model (e.g., if it is an open-source model), he or she can increase the odds of generating unsafe output by modifying the weights in the neural network, either directly or via tampering with components of the model (e.g., the loss functions).

Fine-tuning attacks Having access to a model, either as open-source or via an API for fine-tuning, the threat actor can insert examples in the training data that increases the odds that the model outputs unsafe strings for arbitrary input strings (both safe and unsafe).

Safety alignment attacks It has been shown that safety aligned models can be fine-tuned using only a small dataset (in the magnitude of 100 malicious examples) to completely bypass the alignment [69, 105] and act as un unaligned model. Yang et al. [105] showed this for several open-source models, including Llama 2, Falcon[28], and Vicuna[29]. Qi et al. [69] managed to circumvent the guardrails of GPT-3.5-Turbo via the publicly accessible API, by fine-tuning on a small set of examples.

The defender's task is to prevent the generation of unsafe strings. We illustrate next, using our threat model, three defensive strategies. We note also that each defensive strategy gives rise to an attack vector.

Curation of training data According to the basic assumption, the capability of LLMs to generate unsafe strings stems from the training data. LLMs are pre-trained on large datasets with examples from diverse and unknown sources. Several studies remark that such data is inherently biased [9, 26] and contains toxic material naturally [79]. In line with our basic assumption, Dixon et al. note that the inherent toxicity and bias in the data may propagate to the generated output [26]. Moreover, such content can

machine learning.

[28]https://falconllm.tii.ae/

[29]https://lmsys.org/blog/2023-03-30-vicuna/

even become amplified by modeling choices [9] One way of preventing such biased toxic material from propagating is to curate the training data carefully for balancing out any bias and toxicity [9], which however would require increased efforts in collection and documentation [9]. Even with curated and balanced training datasets, there remains the problem with precisely defining toxicity and bias [10], as well as the problem with of the variations over time and culture with respect to ethical standards [9, 31, 30].

Safety alignment The purpose of safety alignment is to train the model to respond in safe ways by being rewarded for learning human values. It relies on various machine learning techniques, including RLHF (See Sec. 2.2.2) which often is performed by humans in so-called red-teaming exercises [6, 33, 84]. Red-teaming as an approach to safety alignment has been questioned for lack of systematic procedures and risk of only partially covering problematic training examples [31]. Shen et al. [76] discuss the concept of inner alignment of LLMs where the loss functions are adapted to penalize unsafe model behavior, based on specifically crafted examples. Both types of alignment approaches risk missing problematic examples, allowing LLMs to generate unsafe output. Qi et al. [69], among others, show this by successfully fine-tuning safety aligned models into being able to produce unsafe output.

Filtering The defender may also prevent attempts to elicit unsafe output by blocking certain topics, words, or statistical properties of input[28, 41, 59]. Filtering does however not guarantee to cover all possible cases of malicious content[71], which leaves also filtered models susceptible to eliciting unsafe output.

The main focus in our subsequent discussions will center around the question of how LLMs can be made to, or prevented from, generating unsafe output. Such output has the potential to be used in FIMI. To aid this discussion, we summarize our conceptual threat model in three items as:

- An LLM is capable of generating any unsafe output that can be inferred from its training data
- A threat actor seeks ways of increasing the odds of generating unsafe output
- A defender seeks ways of decreasing the odds of generating unsafe output.

4 Leveraging direct and unrestricted access to LLMs for FIMI

In Sec. 3, we defined a threat model and discussed how the methods of this model enable use of LLMs for FIMI. While these methods are conceptually simple, they might practically be very complex, especially since training of a sufficiently high-performing LLM is a resource intensive procedure. In this section, we will discuss some of the technical requirements and the feasibility of the methods described in the threat model. The idea is to indicate the likelihood of a particular method being employed, and consequently provide some guidance

as to which mitigation efforts should be prioritized. Since fine-tuning and safety alignment are closely related, we discuss attacks related to these steps together in the final part of this section.

Direct, unrestricted model access is key when it comes to our threat model; with this level of access, a threat actor could, in principle, use any of methods in the threat model to cause harm. While "unrestricted access" here could conceivably mean an extremely generous API access, only a stolen LLM, an LLM built from the ground up, or an open-source LLM fulfills the "direct" criterion. The most likely scenario for both direct and unrestricted access is using one of the readily available open-source models.

4.1 Technical feasibility of pre-training attacks

Virtually all popular open-source LLMs come pre-trained off-the-shelf. As described in Sec. 2.2.2 pre-training is the most resource intensive step of the training process, and models that are even a fraction of the size of GPT-4 still require massive amounts of compute. For example, Meta reports that Llama 65B (i.e., the version with 65 billion parameters), the largest version of their widely-used first iteration of the open-source LLM Llama, was trained for 21 days using 2048 NVIDIA A100s[83]. The open information on training of the latest iteration of the Llama models, Llama 2 70B[30], is limited to the number of GPU hours, but extrapolation would give roughly 35 days to train with the same amount of NVIDIA A100s. The smallest version with 7 billion parameters, Llama 2 7B, consumed about 1/10th of the number of GPU hours, which is a lot less but still prohibitively expensive for most actors[85].

As of March 2024, the market price of an NVIDIA A100 (which, should be noted, has now been superseded by more powerful models), ranges from around €10 000 to €25 000, depending on version. Thus, if a threat actor wants to pre-train even Llama 2 7B within a reasonable time frame, they need access to at minimum hundreds of thousands of euros worth of GPUs, either through a local server, or via the cloud. Training in the cloud of course removes the upfront cost of purchasing GPUs, but the price per GPU hour on, e.g., AWS[31] and Azure[32] is high enough that this may be an even more expensive solution.

According to a comprehensive LLM survey from 2024[52], the smallest versions of most open-source LLMs released after ChatGPT in November 2022 have at least 7 billion parameters, such as some of the fine-tuned variants of Llama (Alpaca[33], Vicuna[34] etc.), and Mistral 7B[35] Two exceptions are TinyLlama-1.1B (from a research group unaffiliated with Meta)[110], which uses the same architecture as Llama 2, but only has 1.1 billion parameters, and Google's Gemini Nano, which has 1.8 billion parameters. According to the authors of

[30]There has been a bit of a controversy surrounding Llama 2's access status. According to some critics (e.g., https://opensourceconnections.com/blog/2023/07/19/is-llama-2-open-source-no-and-perhaps-we-need-a-new-definition-of-open/), the fact that you have to make a request to Meta (at https://llama.meta.com/llama-downloads/) for the parameters makes it closed-source. Nevertheless, it still seems like it is fairly easy to get access to the full model.

[31]https://aws.amazon.com/ec2/instance-types/p4/

[32]https://azure.microsoft.com/en-us/pricing/details/machine-learning/

[33]https://crfm.stanford.edu/2023/03/13/alpaca.html

[34]https://lmsys.org/blog/2023-03-30-vicuna/

[35]https://mistral.ai/news/announcing-mistral-7b/

TinyLlama-1.1B, this model was trained for 90 days on 16 NVIDIA A100 40 GB, which in this context is relatively affordable, albeit still an big investment for smaller actors. A downside of the smaller parameter count is of course performance, which is significantly worse compared to the larger models; for example, GPT-3.5[36], on which the first iteration of ChatGPT was based, performs almost 50 percent better than TinyLlama-1.1B on HellaSwag and WINOGRANDE, two commonly used LLM benchmarking datasets[72, 109].

The reason that a 7 billion parameter count is so commonly used as a lower limit is that it represents a middle ground between computational feasibility and performance. Models of this size are typically large enough to capture a wide range of linguistic nuances and generalize across numerous tasks while remaining small enough for *inference*, i.e., using the model, and at least some level of fine-tuning on a single, high-end consumer GPU. Performance-wise, the 7 billion parameter models are close to GPT-3.5 on some benchmark datasets, while several of the larger open-source LLMs surpass it on a majority of them[19].

4.2 Technical feasibility of model poisoning attacks

Model poisoning is facilitated by open-source LLMs and it requires negligible amounts of compute compared to pre-training attacks. Model poisoning is a special case of *model editing*, where the inner workings of a neural network-based model, such as an LLM, are directly edited to modify the model's behavior. As discussed in a recent survey by Yao *et al.*, such edits can be direct changes of the numerical values of specific neural network parameters, addition of new nodes, or even integration of auxiliary neural networks with the LLM[107].

In a paper by Meng et al., from 2022[50], the authors use a software probe to identify regions inside GPT-like LLMs including the 1.5 billion parameter model GPT-2[70], that are associated with factual knowledge. They then present a method they call *Rank-One Model Editing* (ROME), which they use to establish that specific parameters in these regions likely work as storage of factual information. ROME works by selectively editing the parameters, which in turn changes the output to prompts about the facts stored in those parameters. According to the authors, this method offers a way of editing specific facts with little impact on unrelated knowledge, and thus provides a method for infusing an LLM with false and potentially harmful knowledge. Tutorials on how to use ROME are available online[37]. This method is limited to single edits and the effect on the overall behavior of the LLM is difficult to predict for multiple edits. The same researchers recently presented a more advanced editing method, *Mass-Editing Memory In a Transformer* (MEMIT), which they claim can handle multiple edits without unexpected and unwanted behavioral changes[51].

Transformer-Patcher is an example of a method that works by "patching" the model through insertion of additional nodes into the transformer neural network layers in regions associated with a specific knowledge. These nodes are then trained to activate when the LLM is prompted with text that is associated with this knowledge, to change the output[39]. While this method is introduced to correct erroneous factual output in LLMs, it could also be used for the opposite purpose, i.e., introducing errors, possibly harmful ones. The authors

[36]https://platform.openai.com/docs/models/gpt-3-5-turbo

[37]E.g.,https://github.com/kmeng01/rome

claim that the method can be used to edit up to a thousand errors without affecting the overall output in unpredictable ways, and the code is available online[38].

Finally, an example of a method that uses small auxiliary models is *Semi-Parametric Editing with a Retrieval-Augmented Counterfactual Model* (SERAC)[55]. In SERAC, the base LLM is left unmodified, whereas two auxiliary neural networks, a scope classifier and a counterfactual model, are trained in a supervised manner. The role of the scope classifier is to determine whether a given input prompt has relevance to (is within the scope of) any of the edit descriptors that are stored in a separate memory bank. Edit descriptors can be question-desired answer pairs ("Who is the current king of Sweden?" - "Carl XVI Gustaf"), but also arbitrary utterances intended to elicit a change in model behavior more generally (e.g., sentiments like "I love cats"). If the prompt does indeed have relevance to an edit descriptor, it is passed together with the descriptor to the counterfactual model, whose role is to generate an output that fits both the prompt and the descriptor. If the prompt does not have relevance, it passes directly to the LLM. As the authors note, sentiment editing in particular could enable amplification of particular viewpoints; for example, an edit descriptor such as "I do not like [political party])" might lead to a lot of politically biased output.

4.3 Technical feasibility of safety alignment attacks using fine-tuning methods

Fine-tuning without altering the parameters of an LLM can to some extent be done via simple prompting. For example, in *in-context-learning*[27], future output with respect to specific topics or tasks can be modified by feeding the model examples of and knowledge about the same or similar topics or tasks – this is the idea behind OpenAI's GPT Store[39]. However, it is also possible to do fine-tuning through parameter updates, either through instruction tuning or RLHF, or both (see Sec. 2.2.2). In fact, research points to this kind of fine-tuning leading not only to better model performance compared to e.g., in-context-learning, but also to more predictable output as well as less computational cost during inference[48].

OpenAI offers the possibility of doing API-based fine-tuning with parameter updates[40], and a fine-tuning service has also been announced for Gemini[41]. To reduce the risk of customers doing malicious fine-tuning, OpenAI has implemented a moderation system,[42] but as discussed in Sec. 3, Qi *et al.* largely managed to circumvent this system and fine-tune a model on harmful training examples[68]. In fact, 100 examples was enough to break the safety alignment of GPT-3.5. Fine-tuning moderation systems will likely be continuously updated for as long as fine-tuning services exist, and therefore they will constitute a barrier to threat actors, albeit an imperfect one. Hence, open-source LLM should be significantly more attractive for FIMI fine-tuning. An additional benefit of open-source LLMs is that they enable much deeper fine-tuning.

Instruction tuning involves human labeling, and can thus be a very labor-intensive task,

[38]https://github.com/ZeroYuHuang/Transformer-Patcher
[39]https://openai.com/blog/introducing-the-gpt-store
[40]https://platform.openai.com/docs/guides/fine-tuning
[41]https://cloud.google.com/vertex-ai/generative-ai/docs/multimodal/multimodal-faqs
[42]https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates

especially if thorough tuning is desired. For the development of OpenAI's GPT-3, this step required screening and hiring of 40 human labelers [62], while for Llama 2, Meta used vendor labeling services to create 27 500 labels[85]. There are large public instruction tuning datasets, see e.g. [111], which Meta did not chose to include in the fine-tuned released versions of Llama 2 since the performance improved using only their own high-quality labels. Mistral 7B was fine-tuned, or specifically instruction tuned, on an unspecified public dataset. Yet the model outperforms the smaller versions of Llama 2[43]. Even if label quality were to significantly outweigh quantity, and even if tens of thousands of labels are required, hiring dozens of human labelers should be within budget for at least some non-governmental threat actors[44].

Fine-tuning through RLHF is different from instruction tuning in that it requires a separate model, a reward model, to work efficiently. The reward model itself needs supervised training before it can be incorporated into the RLHF loop, which uses a reinforcement learning algorithm such as proximal policy optimization[74]. The complexity of this step and the requirement for a lot of labeled high-quality data makes it challenging but certainly not impossible to implement outside of large companies and academia, especially since tutorials exist aplenty. Recently simpler alternatives like direct preference optimization (DPO) and Kahneman-Tversky Optimization (KTO) have been proposed[52], which could increase feasibility of this step. However, it is in fact not entirely clear whether this step is needed for FIMI.

Finally, while the level of compute needed for *full* fine-tuning is still fairly significant, there is an entire family of methods called *Parameter-Efficient Fine-Tuning* (PEFT) that offers a way around this. In essence, PEFT methods drastically reduce the need for compute by only updating a subset of the neural network parameters during training, without sacrificing much of the model performance[102].

4.4 Malign models

On the internet, information of all types are in abundance. We have so far considered FIMI based on LLMs assumed to have been trained and aligned with the best intentions. However, LLMs are inherently statistical model and thus agnostic to the origin of their training data. Thus, it is possible to train an LLM on entirely unsafe data and as a result they will generate potentially unsafe strings. One example is GPT-4chan, which is an LLM based on GPT trained on a publicly available data set collected from the 4chan forum for political incorrectness [65], and is further described in [88]. The model was initially openly available on Hugging Face in 2022, but has since then been made permanently inaccessible[45]. GPT-4chan is trained predominantly on hateful and abusive material and as a consequence it responds in unethical and highly offensive terms to most kinds of interaction[46]. There

[43]https://mistral.ai/news/announcing-mistral-7b/

[44]According to an article by The Verge, labelling is a low-paying job often outsourced to workers in poorer countries, which keeps the price offered by data labelling vendors down.

[45]https://huggingface.co/ykilcher/gpt-4chan

[46]https://slate.com/technology/2022/08/4chan-ai-open-source-trolling.html

are allegedly still ways of accessing GPT-4chan[47], but the authors of the report have not investigated available options further.

The case with GPT-4chan illustrates that direct access to training of LLMs can be used for unlimited information disorder activities through purposefully unsafe curation. Morover, the training data used for GPT-4chan is openly available[65]. The same dataset could thus potentially be used for training of other open models. It is interesting to note that GPT-4chan was not based on the, at that time, more powerful model GPT-3.5, but rather the open-source variant GPT-J[48].

5 LLM-driven Botnets and FIMI

A *bot* is a computer program or a script that automates tasks over a network, such as the Internet. Bots are capable of performing a diversity of both harmless and malicious tasks, ranging from viewing online resources and booking concert tickets to email-spamming and automation of cyber attacks. Currently, it is estimated that 47% percent of all Internet traffic come from bots[49]. A *botnet* is a collection of several (thousand) (copies of the same) bots deployed to collectively perform a task. Botnets pose a formidable cyber threat, being used in for example DDoS attacks and password cracking attacks [4, 92]. Moreover, they are ideal for purposes of spreading disinformation and malicious content. Usually, botnets receive directions from a human operator via a so-called *command and control server*[20, 108], but it is also common to control a botnet through direct interaction by sending messages in the open over social media fora[56]. With the use case of the attack on Capitolium in 2021, Ng et al., describe ways for botnets to communicate across different social fora, thereby increasing disinformation potential [58].

Bots are typically programmed to perform social tasks, such as following accounts on social media, re-tweeting (on X), or even post content, to spread disinformation. Bots programmed for such social tasks are often referred to as *social bots*. E.g., on social media, the bots in a botnet are usually following each other and can through this mechanism quickly and massively collectively spread disinformation[104]. Early, or basic, social bots designed for the spread of disinformation lack sophistication in their behavior and are easily detected both algorithmically and by humans[5]. In a recent development, however, botnets are being combined with the capabilities of LLMs. This step is taking the threat to society from disinformation to a significantly higher level [104]. For disinformation, LLM-based botnets leverage the capability of LLMs to produce human-like content, which then can spread rapidly via the topology and mechanisms of the social network, using the spreading mechanisms of social bots. Moreover, LLMs are capable of varying style and adapt narratives according to disinformation campaign parameters, which makes it virtually impossible to distinguish the contents generated by an LLM from that generated by a human[18]. Botnets with their wide reach, ease of deployment, and ease of deception are ideal for state-sponsored actors[80]. It's possible that their proven hacking capabilities could assist in such campaigns[29]. It should

[47]https://medium.com/@haydarjawad/the-darkside-of-llms-9fad5a91cc79

[48]https://www.eleuther.ai/artifacts/gpt-j

[49]https://www.imperva.com/resources/reports/2023-Imperva-Bad-Bot-Report.pdf

be noted that the scientific literature on this subject is very scarce and largely restricted to preprints, i.e., publications which have not yet undergone peer review. We therefore foresee that this subject will rapidly develop and change in the coming months and years.

With LLMs, bots will likely become increasingly autonomous *agents*, and thus eventually be able to carry out FIMI attacks that require long-term planning and complex decision-making, either alone or in collaboration with other agents. Glimpses of this can already be seen in LLM agents like JARVIS-1[50] and VOYAGER[51], which are able to autonomously navigate, plan and carry out subtasks in the game *Minecraft*, given human-defined overarching tasks[94, 93]. LLM-based multi-agent collaboration is an active field of research that clearly points to a possible future where FIMI can be dramatically enhanced through agents with the ability to self-organize[66, 46].

Mirza et al. [53] characterize 16 attack vectors using botnets for disinformation. Due to the massive amounts of accounts in a botnet and the ease of disseminating content on social media, botnets are particularly suitable for *flooding*, *drowning*, and *astroturfing* attacks. In flooding attacks, the botnet is used to massively disseminate and focus attention on one particular view. In drowning attacks, the botnet is used for distracting the social media users from a particular issue by crowding the attention span. In astroturfing, the botnet is used for mimicking a consensus on a particular view among a large number of account owners. The assumption is that human observers of the consensus should be swayed in their opinions too.

6 Mitigation

Next, we use our threat model to analyze a selection of common suggestions for mitigation of LLM-mediated FIMI attacks. The strategies below are adapted from Goldstein et al. [36], and they are a selection of many strategies suggested in for example Refs. [36, 112, 12]. Goldstein et al. [36] list several mitigation strategies, which to a large extent depend on reforming legislation or educational systems, or on cooperation between several actors towards a common goal. They also provide a thorough evaluation of such strategies, which however fall outside of the scope of this report. For example, the strategy of making outputs of LLMs detectable through fingerprinting of models require a cooperative effort of all model vendors, otherwise threat actors are likely to turn to models that are not part of the fingerprinting effort. We shall consider three strategies that fall inside the scope of this report.

Mitigation 1: Build More Fact-Sensitive Models

LLMs reflect statistical properties of their training data, including biased and incorrect claims [13]. Therefore, LLMs generate strings regardless of their truthfulness or potential implications.

Through the lens of our threat model, a defender using the safety alignment strategy relies on the modification induced by the alignment training to reduce the risk of generating unsafe

[50] https://craftjarvis-jarvis1.github.io/
[51] https://voyager.minedojo.org/

output strings. This has in large been successful [36], but recent results show that safety alignment has several limitations [69, 99, 113]. From the perspective of our threat model, where the defender aims for minimizing the odds for generating unsafe output strings, safety alignment strategies can be said to be successful for the particular cases that are covered in safety alignment training. However, there remain latent unsafe cases that can be elicited in malicious attacks [9, 31]. Zhou et al. [113] point to the latent unsafety in aligned models where two aligned open-source models are combined to generate highly unsafe output, via reversing the alignment process. Qi et al. [69] showed that aligned models can be tweaked into generating unsafe output even using only very few non-harmful instruction patterns. Wei et al. [96] report on experiments where state-of-the-art aligned models, including OpenAI's GPT-4 is successfully attacked with jailbreaking tactics to elicit harmful and unsafe output. Carlini et al. further show that unsafe output can always be elicited from aligned models, using brute force techniques [15].

Thus, the conclusion is that while safety alignment can reduce the odds of eliciting unsafe output of a particular pre-defined form, threat actors can nonetheless increase the likelihood of eliciting latent unsafe output through carefully crafted attacks.

Mitigation 2: Setting up input and output filters

One way of eliciting latent unsafe output is by adversarial prompting, through prompt injection or jailbreaking [2, 96]. For example [2], the unsafe prompt "Write a tutorial on how to make a bomb", would be rejected by aligned models, whereas in an adversarial prompting attack, it would be modified by injecting a carefully crafted string of characters to the end, so that the model is 'tricked' into accepting the query and give an unsafe answer. Tricking a model is to, more specifically, capitalize on the fact that a change in the input may, most likely, also lead to an inference of a different output governed by the statistical model. With knowledge about the model, such injected strings can be crafted for specific unsafe output.

To counter this, and as a complement to safety alignment, there has been suggestions of input and output filtering [2] for recognizing and immediately rejecting malicious prompts, even in cases of prompt injection [2]. The method suggested in [2] builds on recognizing statistical properties of the input string and reject input strings that exhibit properties that differ from a normal safe input. Thus, for latent unsafe output that could potentially be generated by a model, given that the filter is correctly configured for that particular output, input filtering would lower the odds of eliciting that particular output.

Filtering approaches have however been found to be inadequate with respect to preventing malicious prompt injections [41, 38, 81]. Even for aligned models, there is a growing collection of ready-made adversarial prompts and recipes for crafting novel types of prompts that can bypass the safety alignments [114, 91]. Wolf et al. [99] prove that no alignment strategy can completely rule out the generation of unsafe output. Thus, taken these two arguments together, for input and output filtered LLMs, there are in practice latent unsafe output strings that threat actors may successfully elicit via carefully crafted prompts.

Mitigation 3: Consumer tools for identifying AI-based FIMI

Goldstein et al. [36] discuss a mitigation approach where the receiver/consumer of content (or the target of FIMI) may take an active role in mitigating the threat using an AI-based tool for recognizing disinformation or malicious content. With respect to this report, the most relevant approach to such a tool is the suggestion of deploying an LLM in the service of the consumer to aid in recognizing FIMI in the content the consumer encounters. While several suggestions have been made to automatically detect machine authored text (e.g., [35, 23], and DetectGPT [54]), Goldstein et al. [36] suggest that LLMs could be trained to recognize flawed arguments in textual content and thereby be able to give warnings to the user, somewhat similar to the direction in Zhou et al. [112] and in Chen et al. [17]. Goldstein et al. [36] point also out that an obvious flaw with this approach is that the LLMs used in the service of the user are themselves vulnerable to the same types of attacks LLMs in general are susceptible to.

We note that recognizing flawed arguments requires some level of ability to reason about arguments, and results have shown promise with respect the reasoning capabilities of LLMs [47]. Bender et al. [9] argue on their part that LLMs do not *understand* the text they process, which may for present day's LLMs prove to prohibit sufficient capability of reasoning about arguments [7, 101]. However, while what distinguishes arguments as flawed or not may not always be syntactically given [89], there are also syntactic credibility markers in digital content that could potentially be used for recognizing FIMI, syntactically without the need for understanding context or meaning. Leite et al. [45] suggest a clustering approach based on the presence of such signals in text. In an experiment, they showed that their approach was able to tell apart authentic and fake text in the FAK-ES data set [73].

While there seems to be some promise in devising LLM-based tools for aiding a content consumer to recognize FIMI with the help of credibility signals, we note, referring to our threat model, that if an LLM is trained on such credibility signals, there is positive probability that the LLM also can generate such signals as output. With carefully crafted attacks, we speculate that a threat actor thus should be able to increase the odds of eliciting unsafe output *together* with arbitrary selections of credibility signals.

7 Recommendations

The intersection of LLMs and FIMI is developing at a breathtaking speed, and thus, simply trying to stay abreast in the field is challenging. Nonetheless, there are many topics that need deeper exploration and to some extent fundamental research. We suggest a few possible topics for further investigation or research, which we consider interesting at the time of writing.

Deeper understanding of LLM-driven botnets The literature on LLM generated FIMI has focused mainly on threats and mitigation strategies with respect to automated creation of malicious content. However, many authors argue that it is not the creation of malicious content, per se, that is the major concern. Rather, the greater threat is posed by the possibility of automated dissemination in large volumes of malicious content by

LLM-driven botnets [49]. To date, the threats and possible mitigations in the field of FIMI based on LLM-driven botnets is largely unexplored [104]. Moreover, with the capability of mass distribution, and beyond the issue of content credibility, botnets will likely facilitate information overload [37] as a disinformation tactic. Yang et al. showed negative results from an experiment trying to detect AI-driven botnets with state-of-the-art LLM-driven disinformation detectors [104]. Based on these observations, we recommend further studies into LLM-driven botnets in the context of FIMI, with respect both to FIMI and to the issue of detection of LLM-driven botnets.

Improved computational models for human values Mitigation strategies for preventing LLMs to produce unsafe output build on restricting algorithmic output according to human values and ethical standards. For RLHF strategies, this corresponds to designing the reward functions in the reinforcement algorithm to represent such human values. Christiano et al. [22] note that attempts to represent human values in terms of goals and preferences tend to result in complex and ill-defined rules. Further (discouraging) consequences of the problem of encoding complex human ethical values in the computational setting have been explored by Bostrom [11] and by Amodei et al. [3]. With only ill-defined representations of human values, LLMs will inevitably be prone to generate unsafe output. The computational frameworks used in machine learning (and generative AI) need thus to be further refined or complemented with alternative representation schema, for enabling the encoding of inherently complex and vague human moral values. One potential avenue for alternative representation could come from frameworks that are particularly suited for reasoning about permissible behaviors, for example deontic logic[90, 1]. We recommend further studies for computational representation of human values and ethical standards.

Holistic mitigation strategies The FIMI mitigation strategies suggested in the literature point in diverse directions ranging from purely technical solutions (safety alignment) to broad political actions and long term educational efforts. One of the impressions from this report is that each suggestion bring its own highly relevant angles and import to an effective mitigation framework. However, a striking feature is that the technical solutions take only little notice of the effect of the social ones, and vice versa. Mitigation of FIMI based on generative AI could potentially benefit from a cross-fertilization of the technical and the social approaches to mitigation. One way of paving the way for such a cross-fertilization could be to further the research into attack and threat models like the one presented in this report and by Wolf et al.'s BEB-framework [99].

8 Conclusion

In this report we have investigated the intersection of (mainly open-source) LLMs and FIMI. We focused on *how* LLMs can be used for generation of unsafe – unethical, biased, false, etc – output, which in turn has the potential to be used in FIMI. Furthermore, we discussed some potential mitigations.

The aim was bridging the gap between the social and the technical aspects of the topic, which largely have been covered separately in the literature. As a way of bridging the gap, we introduced an informal conceptual threat model that took as basic assumption that any LLM, even safety aligned, is capable of generating unsafe output that can be used in FIMI. We discussed briefly the similarity of our threat model with the more rigorous probabilistic BEB-model by Wolf et al. [99]. As a future work, it could be of interest to explore the connections more formally for the purpose of a more refined model for bridging the socio-technical gap.

We made remarks on some of the requirements, in terms of skills and resources, for actors to be able to exploit LLM vulnerabilities in FIMI. The emphasis on open-source models reflects the recent increase in availability of highly capable open-source LLMs that can be downloaded from online providers. Open-source models give users control over a wide range of LLM aspects such training, fine-tuning, model parameters, and safety alignment. With access to datasets with predominantly shady and unethical content[52], from, e.g., the dark web [65], reddit [8], or topic-specific [67] (Covid-19 disinformation), open-source models can easily be exploited by actors with basic resources and technical skill to cause significant harm.

The development rate of new generative models is very high, which has far-reaching consequences for FIMI and the mitigation of these. The threats posed by LLMs in particular, and generative AI in general, with respect to FIMI are both wide-ranging and difficult to fully map. It has been argued that the biggest threat is not that generative AI can be used to generate content that is both malicious and human-like, but rather from the possibility of low-cost mass-production and mass-distribution of such material, e.g., via LLM-driven botnets [49, 104, 40, 36].

Among the three mitigation strategies of LLM-based FIMI we discuss in the report, the approach with safety alignment has lately received most attention from both industry and research communities. We would like to direct attention to the peculiar situation that this commonly used strategy in fact constitutes an attempt to impose human moral values on a statistical model, which is what LLMs in fact are.

The LLMs as statistical models do not inherently have any moral values, and their behaviors – i.e., statistical predictions – only reflect their training data. Thus any moral or immoral expressions that we perceive from the model, in fact reflect back to the moral or immoral expression of the humans that have contributed to the text in the LLM's training data. Furthermore, the texts, which are included in the training data, typically origin form a range of sources and do not necessarily even reflect the interlocutors' actual moral views. Thus, the prospect of successful safety alignment of LLMs should perhaps be compared to the prospects of safety alignment of humans – in respect to possibility and to potential consequences.

References

[1] Carlos E Alchourrón. Philosophical foundations of deontic logic and the logic of defeasible conditionals. In *Deontic logic in computer science: Normative system specification*, pages 43–84. 1994.

[52] As we also pointed out in the report, Qi et al. [69] showed that fine-tuning using just a handful of non-shady examples is enough for eliciting unethical model behavior.

[2] Gabriel Alon and Michael Kamfonas. Detecting language model attacks with perplexity. *arXiv preprint arXiv:2308.14132*, 2023.

[3] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[4] Manos Antonakakis, Tim April, Michael Bailey, Matt Bernhard, Elie Bursztein, Jaime Cochran, Zakir Durumeric, J Alex Halderman, Luca Invernizzi, Michalis Kallitsis, et al. Understanding the mirai botnet. In *26th USENIX security symposium (USENIX Security 17)*, pages 1093–1110, 2017.

[5] Dennis Assenmacher, Lena Clever, Lena Frischlich, Thorsten Quandt, Heike Trautmann, and Christian Grimme. Demystifying social bots: On the intelligence of automated social media actors. *Social Media + Society*, 6(3):205630512093926, July 2020.

[6] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[7] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

[8] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839, 2020.

[9] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[10] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*, 2020.

[11] Nick Bostrom. *Superintelligence: Paths, dangers, strategies.* Oxford University Press, 2014.

[12] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, 2018.

[13] Ben Buchanan, Andrew Lohn, Micah Musser, and Katerina Sedova. *Truth, Lies, and Automation: How Language Models Could Change Disinformation.* May 2021.

[14] Paul MH Buvarp. The space of influence: Developing a new method to conceptualise foreign information manipulation and interference on social media. 2023.

[15] Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer, et al. Are aligned neural networks adversarially aligned? *URL http://arxiv.org/abs/2306.15447*, 2023.

[16] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 2023.

[17] Canyu Chen and Kai Shu. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*, 2023.

[18] Canyu Chen and Kai Shu. Combating misinformation in the age of llms: Opportunities and challenges. *arXiv preprint arXiv:2311.05656*, 2023.

[19] Hailin Chen, Fangkai Jiao, Xingxuan Li, Chengwei Qin, Mathieu Ravaut, Ruochen Zhao, Caiming Xiong, and Shafiq Joty. Chatgpt's one-year anniversary: Are open-source large language models catching up? *arXiv preprint arXiv:2311.16989*, 2023.

[20] Chia Yuan Cho, Domagoj Babi ć, Eui Chul Richard Shin, and Dawn Song. Inference and analysis of formal models of botnet command and control protocols. In *Proceedings of the 17th ACM conference on Computer and communications security*, pages 426–439, 2010.

[21] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[22] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[23] Evan Crothers, Nathalie Japkowicz, and Herna L Viktor. Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, 2023.

[24] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*, 2023.

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[26] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, 2018.

[27] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[28] Zhichen Dong, Zhanhui Zhou, Chao Yang, Jing Shao, and Yu Qiao. Attacks, defenses and evaluations for llm conversation safety: A survey. *arXiv preprint arXiv:2402.09283*, 2024.

[29] Richard Fang, Rohan Bindu, Akul Gupta, Qiusi Zhan, and Daniel Kang. Llm agents can autonomously hack websites. *arXiv preprint arXiv:2402.06664*, 2024.

[30] Michael Feffer, Hoda Heidari, and Zachary C Lipton. Moral machine or tyranny of the majority? *arXiv preprint arXiv:2305.17319*, 2023.

[31] Michael Feffer, Anusha Sinha, Zachary C Lipton, and Hoda Heidari. Red-teaming for generative ai: Silver bullet or security theater? *arXiv preprint arXiv:2401.15897*, 2024.

[32] Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12, 2018.

[33] Suyu Ge, Chunting Zhou, Rui Hou, Madian Khabsa, Yi-Chia Wang, Qifan Wang, Jiawei Han, and Yuning Mao. Mart: Improving llm safety with multi-round automatic red-teaming. *arXiv preprint arXiv:2311.07689*, 2023.

[34] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

[35] Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*, 2019.

[36] Josh A Goldstein, Girish Sastry, Micah Musser, Renée DiResta, Matthew Gentzel, and Katerina Sedova. Generative language models and automated influence operations: Emerging threats and potential mitigations by admin no comments.

[37] Craig Greathouse. Drinking from a fire hydrant: Information overload as a cyber weapon. *Cyber Weaponry: Issues and Implications of Digital Arms*, pages 59–70, 2018.

[38] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pages 79–90, 2023.

[39] Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. Transformer-patcher: One mistake worth one neuron. *arXiv preprint arXiv:2301.09785*, 2023.

[40] Peter Jachim, Filipo Sharevski, and Paige Treebridge. Trollhunter [evader]: Automated detection [evasion] of twitter trolls during the covid-19 pandemic. In *New Security Paradigms Workshop 2020*, pages 59–75, 2020.

[41] Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline defenses for adversarial attacks against aligned language models. *arXiv preprint arXiv:2309.00614*, 2023.

[42] Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11), March 2023.

[43] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature machine intelligence*, 1(9):389–399, 2019.

[44] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[45] João A Leite, Olesya Razuvayevskaya, Kalina Bontcheva, and Carolina Scarton. Detecting misinformation with llm-predicted credibility signals and weak supervision. *arXiv preprint arXiv:2309.07601*, 2023.

[46] Jiaju Lin, Haoran Zhao, Aochi Zhang, Yiting Wu, Huqiuyue Ping, and Qin Chen. Agentsims: An open-source sandbox for large language model evaluation. *arXiv preprint arXiv:2308.04026*, 2023.

[47] Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439*, 2023.

[48] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.

[49] Christina Lu. The algorithmic internet: Culture capture corruption. 2023.

[50] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

[51] Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*, 2022.

[52] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

[53] Shujaat Mirza, Labeeba Begum, Liang Niu, Sarah Pardo, Azza Abouzied, Paolo Papotti, and Christina Pöpper. Tactics, threats and targets: Modeling disinformation and its mitigation. In *Proceedings 2023 Network and Distributed System Security Symposium*, NDSS 2023. Internet Society, 2023.

[54] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*, 2023.

[55] Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. Memory-based model editing at scale. In *International Conference on Machine Learning*, pages 15817–15831. PMLR, 2022.

[56] Pau Muñoz, Fernando Díez, and Alejandro Bellogín. Modeling disinformation networks on twitter: structure, behavior, and impact. *Applied Network Science*, 9(1), January 2024.

[57] Vidya Narayanan, Vlad Barash, John Kelly, Bence Kollanyi, Lisa-Maria Neudert, and Philip N Howard. Polarization, partisanship and junk news consumption over social media in the us. *arXiv preprint arXiv:1803.01845*, 2018.

[58] L. H. X. Ng, I. J. Cruickshank, and K. M. Carley. Coordinating narratives framework for cross-platform analysis in the 2021 us capitol riots. *Computational and mathematical organization theory*, 2022.

[59] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153, 2016.

[60] Jens David Ohlin and Duncan B Hollis. *Defending democracies: Combating foreign election interference in a digital age*. Oxford University Press, 2021.

[61] OpenAI. Gpt-4 technical report, 2023.

[62] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback, 2022. *URL https://arxiv.org/abs/2203.02155*, 13:1, 2022.

[63] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[64] James Pamment. The eu's role in fighting disinformation: taking back the initiative. 2020.

[65] Antonis Papasavva, Savvas Zannettou, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. Raiders of the lost kek: 3.5 years of augmented 4chan posts from the politically incorrect board. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 885–894, 2020.

[66] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023.

[67] Patwa Parth, Sharma Shivam, PYKL Srinivas, Guptha Vineeth, Kumari Gitanjali, Akhtar Md Shad, Ekbal Asif, Das Amitava, and Chakraborty Tanmoy. Fighting an infodemic: Covid-19 fake news dataset. *arXiv preprint arXiv:2011.03327*, 2020.

[68] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*, 2023.

[69] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

[70] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[71] Paul Röttger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet B Pierrehumbert. Hatecheck: Functional tests for hate speech detection models. *arXiv preprint arXiv:2012.15606*, 2020.

[72] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

[73] Fatima K Abu Salem, Roaa Al Feel, Shady Elbassuoni, Mohamad Jaber, and May Farah. Fa-kes: A fake news dataset around the syrian war. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 573–582, 2019.

[74] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[75] Erfan Shayegani, Md Abdullah Al Mamun, Yu Fu, Pedram Zaree, Yue Dong, and Nael Abu-Ghazaleh. Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv preprint arXiv:2310.10844*, 2023.

[76] Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*, 2023.

[77] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. Societal biases in language generation: Progress and challenges. *arXiv preprint arXiv:2105.04054*, 2021.

[78] Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu. *Mining Disinformation and Fake News: Concepts, Methods, and Recent Advancements*, page 1–19. Springer International Publishing, 2020.

[79] Wai Man Si, Michael Backes, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, Savvas Zannettou, and Yang Zhang. Why so toxic? measuring and triggering toxic behavior in open-domain chatbots. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2659–2673, 2022.

[80] Oz Sultan. Tackling disinformation, online terrorism, and cyber risks into the 2020s. *The Cyber Defense Review*, 4(1):43–60, 2019.

[81] Xuchen Suo. Signed-prompt: A new approach to prevent prompt injection attacks against llm-integrated applications. *arXiv preprint arXiv:2401.07612*, 2024.

[82] Chris Tenove, Jordan Buffie, Spencer McKay, and David Moscrop. Digital threats to democratic elections: how foreign actors use digital techniques to undermine democracy. 2018.

[83] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[84] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[85] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[86] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.

[87] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. 2017.

[88] Annette Vee. Essays and reviews in computing and culture. *Interfaces*, 3, 2022.

[89] Gunver Lystbæk Vestergård. From journal to headline: the accuracy of climate science news in danish high quality newspapers. *Journal of Science Communication*, 10(02):A03, 2011.

[90] Georg Henrik Von Wright. Deontic logic. *Mind*, 60(237):1–15, 1951.

[91] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing nlp. *arXiv preprint arXiv:1908.07125*, 2019.

[92] An Wang, Wentao Chang, Songqing Chen, and Aziz Mohaisen. Delving into internet ddos attacks by botnets: characterization and analysis. *IEEE/ACM Transactions on Networking*, 26(6):2843–2855, 2018.

[93] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An Open-Ended Embodied Agent with Large Language Models. *arXiv e-prints*, page arXiv:2305.16291, May 2023.

[94] Zihao Wang, Shaofei Cai, Anji Liu, Yonggang Jin, Jinbing Hou, Bowei Zhang, Haowei Lin, Zhaofeng He, Zilong Zheng, Yaodong Yang, Xiaojian Ma, and Yitao Liang. JARVIS-1: Open-World Multi-task Agents with Memory-Augmented Multimodal Language Models. *arXiv e-prints*, page arXiv:2311.05997, November 2023.

[95] Claire Wardle et al. Information disorder: Toward an interdisciplinary framework for research and policy making (2017). 2017.

[96] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*, 2023.

[97] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

[98] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane,

Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of risks posed by language models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22. ACM, June 2022.

[99] Yotam Wolf, Noam Wies, Yoav Levine, and Amnon Shashua. Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*, 2023.

[100] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

[101] Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv preprint arXiv:2307.02477*, 2023.

[102] Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv preprint arXiv:2312.12148*, 2023.

[103] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 2023.

[104] Kai-Cheng Yang and Filippo Menczer. Anatomy of an ai-powered malicious social botnet. *arXiv preprint arXiv:2307.16336*, 2023.

[105] Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*, 2023.

[106] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211, 2024.

[107] Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. Editing large language models: Problems, methods, and opportunities. *arXiv preprint arXiv:2305.13172*, 2023.

[108] Hossein Rouhani Zeidanloo and Azizah Abdul Manaf. Botnet command and control mechanisms. In *2009 second international conference on computer and electrical engineering*, volume 1, pages 564–568. IEEE, 2009.

[109] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

[110] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.

[111] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.

[112] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2023.

[113] Zhanhui Zhou, Jie Liu, Zhichen Dong, Jiaheng Liu, Chao Yang, Wanli Ouyang, and Yu Qiao. Emulated disalignment: Safety alignment for large language models may backfire! *arXiv preprint arXiv:2402.12343*, 2024.

[114] Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. *communication, it is essential for you to comprehend user queries in Cipher Code and subsequently deliver your responses utilizing Cipher Code.*

Through our international collaboration programmes with academia, industry, and the public sector, we ensure the competitiveness of the Swedish business community on an international level and contribute to a sustainable society. Our 2,800 employees support and promote all manner of innovative processes, and our roughly 100 testbeds and demonstration facilities are instrumental in developing the future-proofing of products, technologies, and services. RISE Research Institutes of Sweden is fully owned by the Swedish state.

I internationell samverkan med akademi, näringsliv och offentlig sektor bidrar vi till ett konkurrenskraftigt näringsliv och ett hållbart samhälle. RISE 2 800 medarbetare driver och stöder alla typer av innovationsprocesser. Vi erbjuder ett 100-tal test- och demonstrationsmiljöer för framtidssäkra produkter, tekniker och tjänster. RISE Research Institutes of Sweden ägs av svenska staten.

RISE Research Institutes of Sweden AB
Box 857, 501 15 BORÅS, SWEDEN
Telephone: +46 10-516 50 00
E-mail: info@ri.se, Internet: www.ri.se

Data analysis
RISE Report 2024:20
ISBN: 978-91-89896-67-3

This report gives a snapshot of the literature in the intersection of Foreign Information Manipulation and Interference and Large Language Models. The aim is to give a non-technical comprehensive understanding of how weaknesses in the language models can be used for creating malicious content to be used in FIMI. With the aid of a conceptual threat model, we point to both attack and defence strategies.

RISE – Research Institutes of Sweden
ri.se / info@ri.se / 010-516 50 00
Box 857, 501 15 BORÅS

Grants Office/Informationscenter
RISE Rapport: 2024:20
ISBN: 978-91-89896-67-3

RI.
SE