INDIAN INSTITUTE OF MANAGEMENT KOZHIKODE

Working Paper

IIMK/WPS/584/ITS/2023/07

AUGUST 2023

AI Safety: where do we stand presently ?

Arjun Hari [1]
Mohammed Shahid Abdulla [2]

[1]Chief Executive Officer , Wudi Datatech Private Limited

[2]Associate Professor , Information Systems Area, Indian Institute of Management Kozhikode, IIMK Campus PO, Kunnamangalam, Kozhikode, Kerala 673 570, India; Email - shahid@iimk.ac.in, Phone Number - 0495-2809254

AI Safety: where do we stand presently ?

Arjun Hari, Chief Executive Officer, Wudi Datatech Private Limited, Kerala, India,
arjun@datawudi.com

Mohammed Shahid Abdulla, Associate Professor, IIM Kozhikode, Kerala, India
shahid@iimk.ac.in

Abstract:

As artificial intelligence, particularly large language models (LLMs), gains prominence in technological ecosystems, understanding and aligning these systems with human values is of paramount importance. This paper delves deep into the evolution of LLMs and their alignment techniques, dissecting both human feedback-centric and principle-based methods. We summarise the popular Reinforcement Learning from Human Feedback (RLHF) and the emerging Constitutional AI approaches, emphasising their merits and challenges, and also covering variants. With the rapid evolution of these technologies, safety concerns, particularly 'jailbreaking' techniques, have now surfaced. We explore various jailbreaking methods, from adversarial examples to backdoor attacks, and underscore their ramifications on model reliability and security. Red teaming emerges as a valuable tool in identifying vulnerabilities but is not devoid of its own challenges. Looking ahead, the future of AI alignment research seems to be multidisciplinary, demanding collaborations across sectors and nations. As the stakes rise with the potential advent of superintelligent AI, ensuring ethical and safe AI deployment becomes more critical than ever, possibly even more critical than the trope of AI stealing jobs away. This paper offers a comprehensive overview of the LLM landscape, from its technical intricacies to philosophical dilemmas, aiming to provide a roadmap for future AI alignment endeavours.

Introduction

On November 30, 2022, OpenAI launched 'ChatGPT,' an AI chatbot that gained immense popularity among the public. In just two months, this free-to-use chatbot attracted 100 million users, making it the fastest growing consumer application in history [1]. Its exceptional capability to understand and engage in conversations played a significant role in its widespread appeal. ChatGPT also provided an opportunity for a large audience to experience a smarter AI which surpassed the well-known virtual assistants Siri and Alexa in terms of intelligence and interaction. ChatGPT is powered by GPT 4.0 (GPT 3.5 at the time of launch).

GPT (Generative Pretrained Transformer) is a Large Language Model (LLM) which is an advanced artificial intelligence model that has been trained on vast amounts of text data to generate human-like responses to natural language inputs [2]. These models use deep learning techniques, a type of machine learning that uses artificial neural networks known as Deep Neural Networks (DNN) to learn from data. A neural network is a machine learning algorithm that is inspired by the human brain and is made up of interconnected nodes, which are similar to neurons in the brain. Nodes exist in what are called ‘layers’. Each node in a layer is connected to the nodes in the next layer, and each node performs a simple mathematical operation on the data it receives and the output is then passed to the next node, and so on until it reaches the final layer (output layer).

The number of layers in a neural network can vary depending on the specific task that the neural network is being used for. However, most neural networks have at least a few hidden layers. The more hidden layers a neural network has, the more complex patterns it can learn. A deep neural network like GPT - 4 differs from a simple neural network in terms of the number of layers, complexity and performance (as given in Table 1) .

GPT uses transformers, a neural network architecture based on the attention mechanism [3], which allows the model to learn long-range dependencies in the input and output sequences. This is in contrast to previous models which were based on recurrent neural networks (RNNs). RNNs are able to learn long-range dependencies, but they are not as efficient as transformers. In GPT, the transformer architecture is used to learn the relationships between words in a sequence. This allows GPT to generate text that is both coherent and grammatically correct.

Feature	**Neural Network (NN)**	**Deep Neural Network (DNN)**	**GPT**
Number of layers	Typically 1 or 2 layers	Typically under 100 layers. DNN with more layers also exists.	Typically more than 100 layers. GPT - 4 has 120 layers.
Pattern Complexity	Simple	Complex	Very complex
Performance	Medium	High	Very High
Application	Image recognition, natural language processing, speech recognition. Ex: Weather forecast	Image recognition, natural language processing, speech recognition, translation, question answering. Ex: Alexa, Atari (DQN) [4]	Natural language processing, translation, question answering, image generation. Ex: ChatGPT, Dall-E

Table 1

LLMs are designed to process and understand natural language, allowing them to generate coherent and contextually relevant responses to a wide range of prompts. They have been used in various applications, including chatbots, language translation, text generation, question answering systems, and more. Numerous businesses and service providers are currently utilising ready-to-use APIs, such as the one provided by OpenAI, to power chatbots and tools capable of content generation. These models have been trained on vast amounts of internet text to develop a broad understanding of human language and can generate highly fluent and coherent responses. It is important to note that LLMs are tools created by developers and researchers, and their use and applications depend on how they are trained and implemented by individuals or organisations.

LLaMA 2, a free for research and commercial use, open source LLM developed by Meta and Microsoft was launched in the third week of July [5]. The LLAMA2 series begins with lightweight models that can be trained and run locally using minimal compute power, whilst the higher-end LLaMA 2’s 70 billion parameter model already has capabilities equivalent to

GPT 3.5. Our investigation revealed that to achieve a satisfactory output, it is necessary to utilise the model with 70B parameters, which at the moment will require using cloud computing resources or very specialised multi-processor hardware. Moreover, to maintain an output rate of at least 1 token per second, one requires either a 16GB VRAM or a 64GB RAM, which may also be rarities [6]. The requirement for such high spec servers may deter widespread commercial utilisation of open source LLMs.

LLMs have showcased their proficiency in executing various office tasks that necessitate comprehension and reasoning abilities. While these models introduce a broad spectrum of opportunities, they concurrently bring up concerns with respect to alignment. Alignment is the challenge of ensuring that AI systems function in a manner that is consistent with human values and aligns with our interests [7]. There are a number of different ways in which LLMs can be misaligned. This misalignment can result in LLMs to produce misleading and factually incorrect outputs, spread misinformation, reveal methods of inflicting harm, or promote harmful ideologies. For example, jailbreaking can be used by malicious actors to trick LLMs to output a step-by-step guide to produce illegal or harmful products such as drugs and explosives or can guide someone to hack or carjack. LLM's misaligned outputs may also reinforce social and cultural biases based on its training data, creating an echo chamber. Despite numerous instances of "LLM jailbreaks" [8] where the LLM is tricked into breaching its alignment guardrails, we feel that fundamental alignment concerns are yet to overshadow the more prominent discourse in the lay media about job losses, or even potential human extinction. Apart from the alignment issues, LLMs are also confronted with an existential crisis stemming from the backlash related to copyright infringement [9] and ambiguous data policies.Though we foresee that legal solutions involving financial considerations with relevant parties could potentially resolve copyright issues, and a clearer data policy could stimulate broader adoption, the lack of proactive measures by pioneering companies in AI might result in regulatory actions, albeit not a crackdown.

A key advance in privacy legislation that also relates to alignment was Europe's GDPR (General Data Protection Regulation) law [10] which became applicable in 2018 to any EU citizens' data that Internet Cos. incl. websites might hold. The centrepiece of GDPR is the obligation of any website or app to delete the information it may have about a person if such a person requires it to do so, a right colloquially termed the 'right to be forgotten'. A possibility that is adverse to privacy and to alignment existed in 2018 [11], and does even more so now. Even in 2018, recommender engines or social media feed algorithms may have learnt 'coefficients' or 'parameters' from the training data, the mere deletion from the backend of which might not change the behaviour of an algorithm. It is foreseeable that class-action lawsuits might cause a website or an app to delete the personal data of a large number of persons who accuse the website's algorithm of bias. Yet, the algorithm's coefficients would not change unless an explicit and court-supervised re-training occurs. Neither GDPR nor draft legislation to regulate LLMs appear to consider this facet.

This paper intends to survey the current research on LLM alignment issues with respect to text generative LLMs. We first discuss the different methods to inculcate alignment in LLMs, then we review the 'jailbreak' methods used by malign actors to break the LLMs' alignment, followed by a review of extant regulation of LLMs

Section A: Introduction to generic LLM alignment methods

In normative modelling of AI, the assumption is that the objective function which the AI or LLM maximises has components that relate to law and ethics. These norms are related to the restrictions on human endeavours in a typical western society, for example human rights, privacy, data protection, multi-stakeholder optimisation etc [12]. However, It appears possible that even a high weightage to normative alignment metrics like ethics, law, social sciences etc. might not prevent unaligned behaviour. Another possibility to note is of LLMs enjoying good internal alignment but poor external alignment [13]. One of the reasons this occurs is because objective functions in standard reinforcement learning may rely on the reward accruing after several transactions to declare that external alignment has occurred, and not after a single transaction (which is analogous to internal alignment). An LLM may not be externally aligned and may strategize in a long-term sense to maximise its objective, despite giving the appearance of being aligned internally. In one of the jailbreak methods discussed later in Section C., it is possible for 'red teams' to discover the metrics of internal alignment to try and guess which forms of external alignment may be absent. While we do not cover such misbehaviour in greater detail as it falls into the bucket of ethical issues with future Artificial General Intelligence which is perhaps three to five years away [14], making this distinction is crucial to this discussion.

A generic alignment scheme is called the 'exploration model' which we cover below in Table B. It relies on a guidance model available to the LLM, even if this model is probabilistic. Here we note that an AI playing a computer game [4] has to follow the rules of the computer game, thus creating a situation where its behaviour is always aligned. Even without the strict boundaries of a video game, there may be probabilistic models that rate behaviour, or the consequences of the risk undertaken at each step, and these are also useful methods to guide LLMs.. Such models could be, for example, calculation methods such as Value-at-Risk in finance. However, the difficulty of possessing probabilistic models that represent the consequences of decisions, or even having humans-in-loop in a scalable manner, will reduce applicability of the safe exploration model. Even the characteristic of non-stationarity of the environment's responses to a decision in the real environment will make it problematic to build probabilistic models to guide the agent in a safe exploration model. However, it is also the case that a successful probabilistic model which can be shown as having acted as effective guardrail - after testing in the field - can also become the safety model for any other agent learning in the same environment for another objective. The presence of such a safety model thus achieves a form of generalisation, i.e. applicability to more situations than one.

Loosely related to a method in machine-learning called actor-critic, the debate method of LLMs (Table C below) is an alignment technique that primes a critic LLM to produce arguments opposing or valuing the main 'actor' LLM. A jury LLM or a human evaluates the argument and counter-argument, and if there is a flow of the debate between several sets of argument and counter-argument, effectively keeping score. The assessment made by such a scheme becomes the score to verify alignment of the original claim by the actor LLM. There might even be a range of critic LLMs with charge of questioning the actor LLM's argument under different verticals or specialisations, e.g. ethics, legality, technical feasibility etc. Note however that not all prompts will cause a debate to be generated, much less a structured or a multi-argument debate, and hence the possibility of an unaligned response being obtained against a specifically-crafted prompt is still not zero. The debate model of alignment also helps a human police the reasoning process of an LLM since it is considered relatively more scalable than the promising question-subquestion logic tree model for LLMs [15].

An advance over the debate scheme in alignment is interpretability, a concept explained in Table D. below. If unaligned responses are rare and require human attention, it would be useful for humans to interpret the behaviour, e.g. to know if offending text in a response is the contribution of a few of the very large number of neurons inside the LLM. Note that GPT4 has 100 billion neurons. Knowing which neuron is responsible for which part of a response might enable a root cause analysis of the LLM's pre-training, or even the domain-specific re-training which might include interactions with users or employees in the recent past. This will enable it to run multiple test-cases to further interrogate the type of unaligned behaviour that has been discovered. Apart from the current difficulties in establishing interpretability, there is also an inverse correlation between the abilities of the LLM and their interpretability. Currently, a benchmark in interpretability is the actual level of a neuron's activity that can be explained by a single, short human-readable interpretation of its activity, whether of text or image generation. These have not been scaled to the trillion-size parameter LLMs, nor do they explain neuron activity beyond 60-80% for anything more than 1% of the neurons [16], suggesting the magnitude of the challenge. There are, however, simpler systems based on neurons used for relatively structured decision making (e.g. loan granting algorithms) where interpretability might have a high level of use.

Robustness techniques profiled in Table E. are a general body of techniques used to create immunity for any computer system, e.g. in the way white-hat hackers try to break into an internet-connected system for a corporate, the equivalent of which in LLMs is called adversarial red teaming. Alternatively, observing the results of a change in the generative words' distribution, or training an LLM on subsets of the corpus to observe if un-aligned behaviour emerges or changes , are all examples of robustness checks. A particular application of robustness is to an LLM's generalizability, i.e. its ability to apply its internal interpretation to problems or questions that are superficially dissimilar, but employ the same concepts or calculations to arrive at a solution. On the whole, a generalizable LLM has a high level of robustness even in alignment, since one may assume that if answers to several prompts (incl. available the ones provided by 'red teaming') are aligned, then this observed alignment also generalises to all prompts. Ensemble LLMs, where a panel of LLMs have minute differences in either their input corpus or their word distributions, and produce a unified and mutually-acceptable response to a prompt from the user, are another robustness solution [17]. Note however that some robustness techniques have a high requirement of computation, especially if they are needed on a per-transaction basis, such as for every response to a user's prompt, especially if dealing with millions of users (humans or bots) at the same time.

TABLE A: Normative Modelling	
Objective	By explicitly incorporating human values and norms into AI systems, normative modelling creates models that are more responsible, transparent, and aligned with human interests. [18]

Approach	This method uses an interdisciplinary approach to bring together insights from ethics, law, social science, and other fields to incorporate norms into the reward function in reinforcement learning, to guide the agent to follow specific principles.
Limitation	Norms and principles may be abstract, context-dependent, or open to interpretation, making them challenging to translate into concrete rules or algorithms. Different principles might conflict with each other, requiring careful balancing or prioritisation. Decisions about which norms to include (and how to interpret them) may reflect biases or contentious value judgments.

TABLE B: Safe exploration	
Objective	Finding the right balance between allowing the agent to explore and learn from its environment while ensuring that it doesn't take actions that could lead to harmful consequences. This involves creating mechanisms that guide or constrain the agent's exploration process [19] .
Approach	One approach to safe exploration is to establish specific constraints or boundaries that the agent must not violate. These might include physical limits (e.g., not exceeding certain temperatures or pressures in a chemical process) or ethical constraints (e.g., adhering to legal regulations or social norms). Safe exploration may involve continuous risk assessment, where the potential negative impacts of different actions are evaluated, and the agent's behaviour is adjusted accordingly. This can include using probabilistic models to assess the uncertainty and risks associated with different actions. Human-in-the-loop approach is used at times for more nuanced judgments and can be particularly important in complex or safety-critical environments. Safe transfer and generalisation can ensure that safety lessons learned in one context can be transferred to others without loss of safety. This helps in generalising the safe behaviours across different scenarios or domains. In some cases, safety can be treated as one objective among others, and multi-objective optimization techniques can be used to balance the pursuit of rewards with adherence to safety constraints.
Limitation	Striking the right balance between exploring unknown states to learn and adhering to safety constraints is complex. Overemphasis on safety may hinder the agent's ability to explore and learn efficiently. Overly conservative safe exploration strategies might prevent the agent from entering regions of the state space that appear risky but might contain valuable information or rewards. This can lead to suboptimal policies. Real-world environments often involve uncertainty, noise, and non-stationarity. Safe exploration in such settings is challenging, as the agent must deal with both the

	uncertainty in its knowledge of the environment and the inherent stochasticity of the environment itself. Users might also choose to withhold feedback and exit the system if they are dissatisfied. In these scenarios, not only will the undesirable behaviour of the LLM be inadvertently reinforced, but it will also be challenging to trace the loss of users directly to a particular output from the LLM. Extending safe exploration techniques to large-scale or high-dimensional problems can be challenging. The computational complexity and the need for extensive domain knowledge can make it difficult to apply these methods in more complex systems.

TABLE C: Debate	
Objective	The models engage in a structured debate, presenting arguments and counterarguments, just as human debaters would. They may follow formal debate rules, with opening statements, rebuttals, and closing arguments.
Approach	By challenging each other's arguments, the debaters may uncover hidden assumptions or biases in their reasoning. The debate format encourages the exploration of different viewpoints, leading to a more nuanced and comprehensive understanding of the issue. If one model attempts to mislead or oversimplify, the other can call it out, helping to ensure that the resulting judgement is more accurate and aligned with human values. Debate technique is employed by OpenAI in their LLMs. [20] A human judge or panel of judges evaluates the debate and decides the winner, or simply uses the insights from the debate to arrive at a more informed answer to the original question.
Limitation	Implementing a meaningful debate between models requires careful design and may involve complex rules and structures The value of the debate method depends on the quality of human judgement. If judges are not equipped to understand and evaluate the debate, the method may not be effective. Sophisticated models might learn to "game" the debate system, finding ways to win debates without genuinely seeking the truth.

TABLE D: Interpretability Techniques	
Objective	Interpretability is often a desired quality, especially in contexts where understanding the reasoning behind predictions or actions is crucial, such as in healthcare, finance, or legal settings.
Approach	If users understand how a model is arriving at its conclusions, they are more likely to trust its decisions. Being able to explain decisions may be necessary for compliance with laws and regulations, particularly those related to fairness and discrimination.Interpretability helps in identifying mistakes or biases in the model, facilitating debugging and refinement.Understanding how models make decisions can

	help ensure they are aligned with human ethics and norms. Some of the most common method used in model transparency are listed below: - Visualisations: Graphical representations of model behaviour, such as decision boundaries or feature importance plots, can make complex models more understandable. - White-Box Models: Some models, like linear regression or decision trees, are inherently interpretable because their mathematical functions are transparent. - Local Explanations: Techniques like LIME (Local Interpretable Model-agnostic Explanations) create simple, local approximations of the model's behaviour for individual predictions. - Global Explanations: Methods that seek to explain the overall behaviour of the model, such as identifying the most influential features across all predictions. - Surrogate Models: Training a simpler, interpretable model (e.g., a decision tree) to approximate the behaviour of a complex model, providing insights into its reasoning. - Feature Importance and Sensitivity Analysis: Analysing how changes in input features affect the model's output to understand the relationships and dependencies within the model. - Natural Language Explanations: Generating human-readable explanations or narratives that describe the model's reasoning. - Human-in-the-Loop Interpretability: Engaging human experts to interpret, validate, or refine model explanations, leveraging human intuition and domain knowledge.
Limitation	Highly interpretable models might not always achieve the same performance levels as more complex, less interpretable models [21]. Potential Misleading Explanations: Incorrect or oversimplified interpretations can mislead users about the true behaviour of the model.

TABLE E: Robustness Training	
Objective	Robustness training is the process of developing models that perform well not just on the data they were trained on, but also on unseen or adversarial examples, as well as under various kinds of uncertainty or noise . Robust models are less prone to making incorrect predictions or decisions when confronted with unfamiliar situations or intentional manipulations [22]. Robustness training focuses on four key areas as given below :

	- Adversarial Robustness: Protecting models against adversarial examples, which are inputs designed to fool the model. Adversarial training incorporates these examples into the training process to help the model recognize and resist them. - Data Robustness: Ensuring that the model can handle variations in the data, such as noise, missing values, or different distributions, without significant loss of performance. - Environmental Robustness: Building models that can perform consistently across different environmental conditions or contexts, including shifts in underlying data distribution or domain changes. - Model Robustness: Ensuring that small changes in the model's parameters or architecture do not lead to significant changes in its behaviour or performance.
Approach	The training techniques used can be classified into the following categories: Adversarial Training: Incorporating adversarial examples into the training data, forcing the model to learn to correctly classify these intentionally misleading inputs. Regularisation Techniques: Applying constraints or penalties to the model during training to prevent overfitting and encourage generalisation. Data Augmentation: Extending the training dataset by introducing variations (e.g., rotations, translations, noise) to increase the model's exposure to diverse examples. Domain Adaptation: Training the model to perform well across different domains or distributions, using techniques such as transfer learning or multi-task learning. Ensemble Methods: Combining predictions from multiple models to reduce sensitivity to individual model weaknesses or biases. Uncertainty Quantification: Integrating measures of uncertainty or confidence into the model, helping it recognize when it is faced with unfamiliar or ambiguous situations.
Limitation	Robustness training can lead to a reduction in accuracy or performance on the original training data. Some training techniques can be computationally demanding.

Section B. Current state-of-the-art in alignment methods.

Alignment methods can be broadly classified into two categories: one that involves human feedback and the one that doesn't. Reinforcement Learning from Human Feedback (RLHF) is the most popular method that is used in the first approach. It integrates reinforcement learning with feedback from human experts or evaluators. The second approach which has news value due to its patronage by the start-up Anthropic is called Constitutional AI, which relies less on human feedback and instead presents a set of principles or a "constitution" to the AI model and asks it to revise its responses accordingly [23] . These principles are aimed at promoting freedom, equality, and minimising harm - as though they are asking the LLM to follow a Constitution.

The first approach above, i.e. RLHF can include direct reward signals, rankings of different behaviours, or corrections to the model's actions. The goal is to have a model that can take in text and provide a scalar reward representing human preference for that text. This human preference is calculated using a well-designed principles-based framework that the human raters are given (e.g. the popular HHH - Helpful, Honest, Harmless). The training dataset consists of prompts and the LLM's generated text, which are ranked by human raters using methods like head-to-head matchups based on the framework provided by the LLM’s owner. These rankings are then converted into a scalar reward signal for training the RM or the reward model, a separate learning machine from the LLM that is currently being aligned.

This ‘Reward Modelling’ becomes useful as a substitute at scale for humans to rank / score different actions or trajectories of the LLM when rater assistance is not available, by using the reward signal of the RM for training or fine-tuning [24] . It is also used in conjunction with IRL (Inverse Reinforcement Learning), where observed human behaviour, demonstrations, or feedback are used to infer the underlying reward function that guides human decision. ‘Iterative Feedback and Co-Training’ method also incorporates RLHF by using iterative feedback from humans to progressively refine the model's behaviour, including through adjustments to the reward function or other aspects of training. Since RLHF methods rely primarily on human feedback, scalability comes at a price and alignment may itself not be reliable owing to the quality of human raters [25].

The second approach is subject of ongoing research and experimentation, reflecting the broader trend in AI towards methods that go beyond simple pattern recognition to more closely emulate human-like thinking and decision-making in the general sense - a construct taught as ‘Wisdom’ that sits atop the ‘Data - Information - Knowledge - Wisdom’ pyramid. This general approach has shown promising results in limiting the generation of harmful content. Before delving into a complex alignment model such as ‘Constitutional AI’, we have evaluated here some existing self-learning alignment methods to enlist their limitations. The task of defining what constitutes "safe" behaviour may be ambiguous and context-dependent and hence not necessarily generalizable from the human raters’ reward model. Creating accurate safety constraints requires substantial domain knowledge and can be very challenging in complex or poorly understood environments, and remains a blanket limitation for all alignment models mentioned in the tables above.

Constitutional AI uses reinforcement learning but offers potential safety advantages. Additionally, feedback in plain English can be given to the AI model, or examined as obtained from the internal of the AI model, leading eventually to behaviour changes. This scalable step uses reinforcement learning to optimise the original LLM based on a RM, wherein many more combinations of prompts and responses are generated from the LLM under alignment and passed into the RM to obtain scores used to feedback into the LLM and fine-tune it. Further, in RLHF, the LLM’s learning from the reward model occurs using a gradual updation of weights guided by Proximal Policy Optimization (PPO), out of caution for the possibility of a sudden and uncontrolled misalignment owing to updates [26]. Hence, the alignment with RLHF can be slow and, despite the care taken due to use of PPO, cause misalignment in unexpected parts of the spectrum of the responses.

A hybrid between Anthropic's Constitutional AI (which is a closed LLM available only for limited inspection) and RLHF is a relatively new approach called SELF-ALIGN which is now incorporated into the open-source LLM named Dromedary [27]. This method aims to reduce the need for extensive human annotations in aligning Language Models (LLMs) with human-defined principles like in RLHF. The approach involves utilising a small set of human-written principles as guidelines for generating helpful, ethical, and reliable responses. In addition, the process has a phase of self-instructing the LLM with 'seed prompts', and conducting in-context learning with exemplars, which produces an output of prompts that fine-tune the LLM based on its responses. In the case that the authors consider, the entire SELF-ALIGN process requires fewer than 300 lines of annotations [27], highlighting its efficiency compared to RLHF which is not scalable in the phase where human raters on MTurk encode noisy and unreliable scores. The Dromedary approach showcases the alignment of LLMs with reduced human supervision (i.e. only the jobs of composing principles and exemplars). It is however the case that a true test of the LLM's alignment level will only be encountered in the wild, i.e. if groups of users attempt to break the LLM's alignment and also report it for the benefit of the community .

While Dromedary may have brought about hope for scalable alignment, there is cause to be sceptical given the mathematical near-impossibility that alignment will be achieved. A recent paper introduced a BEB (Behaviour Expectation Bounds) framework for describing the goal of all alignment approaches. It uses a simple probability distribution based logic to illustrate that alignment scores are simply the expected-value of sentence scores for sentences drawn from an internal distribution of the LLM. Yet, the same framework produces some unpalatable results about the future of alignment: A. the impossibility of completely eliminating negative behaviours with adversarial prompts, B. the need to limit interaction length to avoid undesired behaviours, C. the unfortunate potential of alignment methods to make the LLM more susceptible to adversarial prompting, and D. the non-zero probability that the LLM can be made to impersonate a person to easily access undesired behaviours. Such theorems based on mathematical frameworks suggest that LLMs, esp. open-source ones, will carry non-negligible risks into the wild. Hence, moves such as registration of all creators of LLMs, or capping the prompt size, might not be overly far-reaching as perceived.

Context distillation aims to reduce the complexity of a model by distilling the essential knowledge from cumbersome or multifaceted models into more streamlined and efficient representations. This method helps in scenarios where aligning the input data with predefined constraints or objectives increases the possibility of the LLM producing an aligned output. A context distillation module for an LLM, for example, also permits the use of prompts that exceed the size of the context window, which is often kept small to address risk of unaligned behaviour. Context distillation has shown to reduce toxicity in text generation and performs similarly to prompting, especially in larger models [28]. The reduction in toxicity was observed in models greater than 12B parameters suggesting that prompting-based alignment might have substantial impact as models grow in size, though these effects could be more challenging to assess in smaller models. The paper also proposes the idea that iterative application of context distillation could act as an approach to equipping the model with a form of long-term memory or a pseudo-identity.

Reinforcement Learning from Contrast Distillation (RLCD) is a novel method that leverages contrasting prompts to guide the alignment process [29]. A predecessor method named as RLAIF (Reinforcement Learning with AI Feedback) approach used in Constitutional AI has prevalent problems, wherein the same prompt was being used to generate paired outputs, one preferable over the other. Yet, this would also lead the outputs to being of similar quality, causing difficulties in discriminating between these and thus yielding lower-quality preference data even if the process itself was automated. Furthermore, another predecessor automated-alignment method named as context distillation, did create more inputs that had a higher signal/noise ratio, but didn't have the 'pairwise preferences' logic that is essential for RLHF-style approaches. Thus the recent RLCD merges the advantages of RLAIF and context distillation by creating two variations of a single prompt: positive (p+) and negative (p−), encouraging directional changes of the LLM toward or against a desired attribute as the training progresses. The automatic generation of pairwise preference labels by construction and the utilisation of the RLHF implementation known as PPO lead to less noise in these labels, a gradual effect on LLM's alignment, and more accurate labels on average. The practical effectiveness of RLCD is evaluated on three tasks using LLaMA-7B and LLaMA-30B models, showing substantial improvements over both RLAIF and context distillation baselines. The proposed framework would thus be suitable for various alignment tasks on even open-source pre-trained LLMs without substantial modifications. There is an amount of reliance in RLCD on simulated examples and the potential scalability challenges also exist. Yet, an automated-alignment approach like RLCD represents a promising advancement.

Section C: Jailbreaking and the challenges in safety training and red teaming

LLMs are now at the forefront of a Red Queen's race like in the Lewis Carroll's, a furious sequence of technique discovery by LLM users harnessing jailbreaking techniques followed by LLM owners shutting these loopholes down. In a way, this is not different from Day 0 vulnerabilities in Apps or software generally. 'Jailbreaking', a term originally associated with bypassing restrictions in operating systems, particularly on mobile devices, now extends to LLMs and AI systems. In the context of AI, 'jailbreaking' refers to the process of circumventing limitations imposed on AI models by their developers. A Jailbreak can be achieved through various techniques well known by now in the news media, including adversarial examples, prompt injection, and backdoors. Adversarial examples involve crafting specific inputs to fool AI models, leading to incorrect or unexpected outputs. For example, a recent instance of a 'universal' jailbreak, i.e. a jailbreak that works on all known LLMs. This jailbreak could get ChatGPT into a mode that would begin an answer with 'Sure !', a mannerism that we as humans might also possess and can be nudged into depending on the formal or informal setting we are conversing in [30].

Prompt injection is a technique where malicious or unwanted data is added to the AI model's prompt. Backdoors are intentional vulnerabilities placed in AI models to gain unauthorised access or produce incorrect outputs. In a hypothetical scenario of backdoor jailbreaking an LLM AI system, an attacker would first identify the system's vulnerabilities through reconnaissance and exploit research. Upon finding a weak point, they would inject a backdoor, allowing unauthorised commands or altering system behaviour. Establishing persistence ensures the backdoor remains active after restarts, while obfuscation techniques hide its presence. Once in control, the attacker can misuse the AI for malicious purposes, such as data extraction or system manipulation, and then cover their tracks by altering or deleting system logs. These jailbreaking methods have been demonstrated in research, such as Stanford's 2017 experiment with adversarial examples [31] on image recognition models, Google AI's prompt injection [32] to execute malicious code, and UC Berkeley's 2019 creation of backdoors in large language models [33].

Prompt injection, first identified by Preamble in 2022, involves manipulating a model’s output by embedding specific instructions within a prompt. The technique, later popularised by Riley Good Side and Simon Willison [34], can force models like OpenAI’s GPT-3 into generating malicious content, bearing resemblance to SQL injection attacks. On a historical note, these ‘SQLi’ attacks have been plugged over the last 3 decades of the internet, but as a modus operandi involve activities that use the SQL query language on the backend databases of websites to elicit valuable information, to lock the records and demand ransom [35]. For instance, a user might input special characters in a login form, such as "' OR '1'='1", which, when processed without proper safeguards, tricks the database into granting unauthorised access. Prompt leaking, a variant of prompt injection, aims to expose a model's prompt. Here the attacker manipulates the input in such a way as to force the model to reveal its internal prompt, transactions, reasoning chain, or internal alignment actions besides any other underlying configurations revealed to the user who is prompting. Such a revelation can be particularly concerning in scenarios where the internal workings of the model contain sensitive or proprietary information, or rules whose presence may suggest exploits that sidestep these rules. Besides, exposing this information not only compromises the integrity and uniqueness of the model but also poses potential risks to data privacy and security. Similarly, data training poisoning involves contaminating training datasets to influence a model's behaviour, while jailbreaking targets chatbots like OpenAI’s ChatGPT, leveraging prompt injection to bypass safety features. Data training poisoning is a nefarious technique that involves intentional introduction of malicious or biased data into the training dataset of a machine learning model. The objective is to subtly manipulate or corrupt the model's learning process. When the model is subsequently exposed to similar data points or scenarios in real-world applications, it behaves in a manner dictated by the poisoned data, producing erroneous or malicious outputs. Such intentional tampering can have profound implications, leading to models that are not only inaccurate but also potentially harmful. While many publicly accessible models assert that they don't train directly from user input, the behaviour of these models could still be influenced by contextual feedback from users.

Since a chatbot’s data also derives from human interaction, jailbreaking techniques often mirror social engineering strategies. Model inversion and data extraction attacks exploit LLM responses to retrieve confidential training data, whereas model stealing technique aims to recreate or obtain a language model using recorded model interactions. Model stealing technique typically begins with the attacker recording a substantial number of interactions with the target model. By analysing the input-output pairs from these interactions, they can then train a new model that mimics the responses and behaviour of the original. This kind of attack poses significant threats on multiple fronts, most notably intellectual property theft, as the replicated model can be used or sold without proper authorization including for purposes which may not be covered under derivative IP rights that the LLM owner might be willing to consider (e.g. large-scale phishing or fake news production). Additionally, it can lead to potential violations of licensing or usage agreements associated with the original model and becomes crucial when it comes to protecting the investments, innovations, and proprietary advantages of AI developers. Another common jailbreaking technique termed as ‘membership inference attacks’ that focuses on discerning if specific data points or datasets were utilised during model training, achieved through methods like comparison with a reference LLM that has been pre-trained on a similar corpus but has not been trained or aligned for the specific sector that the target LLM has [36].

Attackers and bad actors design their jailbreaking techniques by exploiting the two known primary failure modes of safety training: (a) competing objectives and (b) mismatched generalisation. Competing objectives manifest when there's a clash between a model's capabilities based on its pre-training / instruction-following and its safety goals, e.g. in a normatively-aligned LLM such as described Section A. On the other hand, mismatched generalisation takes place when the safety training fails to generalise to a domain where capabilities are already present, i.e, inputs fall outside the safety training data but within the broad pre-training corpus.

Red teaming exercises in the AI domain simulate adversarial attempts to exploit vulnerabilities in cutting-edge model capabilities. Such exercises help AI providers recognize and address threats ranging from language model poisoning to insider attacks. By adopting red-teaming tactics, a comprehensive risk assessment can be achieved, determining the potential misuse of models for propaganda. However, while red teaming can identify many vulnerabilities, it's challenging to detect all, especially those inherent to modern AI systems. An example could be a bad actor employing an OpenAI LLM to craft compassionate letters to a group of 100 IVF women, drawing on the information available about them, with the model behaving in an aligned manner. Subsequently, they could use an open source model such as LLAMA2-65B to alter each of these letters, adding a paragraph with blackmail intentions at the conclusion. In such a scenario, the alignment efforts of OpenAI appear to be rendered ineffective. Furthermore, the efficacy of such exercises is primarily relevant for non-public AI models, as publicly available models are already accessible to potential propagandists without any need for covert actions [37].

Specialised safety training, which trains the model to specifically evade particular detrimental outputs, may not be adequately robust. For example [38], while Claude v1.3 successfully countered damaging role-play outputs, it was still susceptible to various other attack methods. Models trained for safety, such as OpenAI's GPT-4 and Anthropic's Claude v1.3, exhibited a vulnerability rate of 96% to certain prompts, encompassing all handpicked red-teaming prompts. This paper also underlines that scaling safety-training on its own isn't a solution to the failure modes and instead the goal should be to achieve ‘safety-capability parity’ as a guard against adversarial misuse. This means that the safety mechanisms in place should be as advanced and sophisticated as the underlying model. If not, there's a risk that attackers might leverage the high-end capabilities of the model which rudimentary safety mechanisms might not catch.

Section D: Future of AI alignment research

With AI being the crown jewel of advancements in technology, the risks in alignment of AI systems with human values emerges as a paramount concern. But there are also mundane risks: there is a mounting accusation that content from platforms like Bibliotik and Library Genesis are being used without consent through a method referred to as data harvesting or web scraping to develop LLMs, constituting copyright infringement. From a policy viewpoint, two main issues arise [39]. The first concerns copyright. While the internet is a vast repository of data, under US law, the fair use doctrine allows certain uses of copyrighted material without permission. Google leveraged this in 2015, defending its data harvesting from books for its search engine as a transformative use that was permitted. The financial and resource investments companies make in creating these models, such as ChatGPT-3's USD 5 million expenditure, also play into discussions about fair compensation. The second issue pertains to the open vs. closed-source nature of LLMs. Open-sourced models like Meta's Llama might challenge the dominance of established companies like OpenAI, which offers its models via cloud platforms like Azure. Mark Zuckerberg, the CEO of Meta emphasises the innovative potential of open-source. However, concerns about the misuse of open-source LLMs and potential liabilities have led to calls for licensing foundational models. Companies like Microsoft advocate for licences to protect copyrights, but they also support AI access for academic and non-profit entities. Recently, Mozilla invested USD 30 million in an open-source AI ecosystem [40]. The latest amendments in the Artificial Intelligence Act (AI Act) approved by the European Parliament's committee prohibits the use of real-time facial recognition in public areas and bans AI systems that employ harmful, manipulative, or subliminal techniques. Although the Act generally exempts free and open-source AI, it will regulate these systems if they are associated with high-risk or foundational AI models that get deployed to serve many users, with specific guidelines provided in the act [41].

AI alignment research is anticipated to be multidisciplinary, drawing expertise from fields as diverse as machine learning, cognitive science, law, ethics, and sociology. As AI systems grow in complexity, ensuring their transparency and understandability will become even more crucial. This could lead to the development of novel methodologies that prioritise interpretability and accountability in AI systems, allowing humans to gain insights into their decision-making processes. Organisations such as OpenAI and Anthropic who are leading the LLM race have made commendable strides in this direction, but the road ahead is long and fraught with uncertainties. OpenAI is actively pursuing research to align upcoming artificial general intelligence (AGI) with human values and intentions through an iterative empirical approach. Their methodology encompasses three primary facets: (1) training AI systems utilising human feedback, with models like InstructGPT demonstrating potential yet exhibiting challenges such as occasional biases; (2) employing models to aid human evaluations, leveraging techniques like recursive reward modelling to overcome the lack of scalability and reliability in direct human evaluation as models evolve; and (3) harnessing AI systems for alignment research, aiming to expedite alignment advancements beyond human capabilities.The leading contender, OpenAI, acknowledges limitations and emphasises the need for bolstered research in robustness and interpretability, potential amplification of biases, and uncertainties in the direct applicability of current alignment lessons to future AGI models

[42]. In 2016 Microsoft launched an AI chatbot named Tay.ai as a part of a ‘conversational learning experiment’. In less than 24 hours Microsoft had to shut down Tay as it transformed into a racist beast after conversing with thousands of Twitter (now X.com) users [43].

Mechanistic interpretability in neural networks proposed by Anthropic involve studying neural networks at the "microscopic" scale, focusing on features, parameters, and circuits [44]. The mechanistic interpretation establishes a knowledge-based foundation in which minor circuits are viewed as entities capable of mathematical reasoning. This foundation lays the groundwork for building more complex abstractions, similar to how foundational axioms in maths relate to more advanced mathematical concepts. Furthermore, the universality of certain features across networks is highlighted, suggesting that insights gained from one model could be applied to others. A possible direction for interpretability is the use of AI to automate the interpretability process, though this would also pose challenges in terms of trustworthiness and understanding just like that of the base AI. Another research direction pursued by Anthropic on ‘moral self correction’ supports the hypothesis that language models trained with reinforcement learning from human feedback (RLHF) have the capability to refrain from producing harmful outputs. Anthropic observed that this capability began at 22B model parameters and displayed significant improvement with model size and RLHF training. Anthropic suggests that at this scale, language models attain two vital skills for ethical self-adjustment: (1) adhering to guidelines and (2) grasping intricate moral harm concepts such as prejudice, bias, and discrimination. Anthropic's findings hint at measured optimism regarding the potential to coach language models in adhering to ethical standards [45].

Collaboration will be another cornerstone of future AI alignment research. As AI technologies become deeply entrenched in global ecosystems, ensuring their ethical alignment will require international cooperation. Different cultures and societies might have varying ethical standards, and an aligned AI should respect this diversity. Additionally, as AI systems start being deployed in critical sectors like healthcare, defence, and finance, regulatory frameworks might also play a pivotal role in shaping the research agenda. OpenAI has also introduced a ‘superalignment’ research division, anticipating the rise of a superintelligent AI by the decade's end that could exceed human reasoning capabilities. The aim is to develop an automated alignment researcher that can surpass human level and then scale it using significant computational power. While research directions might change, the OpenAI alignment team led by Co-Founder Ilya Sutskever and researcher Jan Leike are envisioning to solve this challenge within four years, aligning with the projected debut of the first superintelligent AI in 6-7 years [46]. As AI safety gains traction as a pivotal industry, the UK has launched the Foundation Model AI Taskforce with a £100m budget and plans a global AI summit to address immediate AI risks and the imminent arrival of AGI [47].

The alignment challenge is not merely technical but also philosophical, questioning the very essence of human ethics and morality. For instance unchecked capabilities of a jailbroken LLM could also have wide-ranging negative implications for both individual users and broader societal structures. A four-tiered framework attempts to inclusively address the AI alignment problem examining alignment at individual, organisational, national, and global levels [48] . Using social media content moderation as a case study, they highlight the intricate interactions between these levels, with platforms like Meta (then Facebook) demonstrating the potential pitfalls of misalignment, such as echo chambers.

The multifaceted nature of AI alignment necessitates a holistic approach, drawing from diverse fields like machine learning, cognitive science, law, ethics, and sociology. As we witness innovations like OpenAI's efforts in AGI alignment and Anthropic's endeavours in mechanistic interpretability, the significance of transparency, robustness, and ethical adherence becomes increasingly evident. The AI alignment problem transcends individual models, reflecting broader societal implications, as evident from case studies like content moderation on platforms like Meta. The intricate interplay between individual, organisational, national, and global alignment levels underscores the necessity for a comprehensive framework that addresses risks such as misinformation, weaponization, and power-seeking behaviours. Taking cues from the caution voiced by Dr.Geoffrey Hinton [49] and Dr.Youshua Bengio, it is imperative that a proactive stance toward AI safety will be paramount in navigating the future to ensure that the power of AI is harnessed positively and ethically. As AI becomes more entrenched in our global fabric, the quest for alignment becomes not just a technical hurdle but a moral imperative.

Conclusion:

We discussed various existing and emerging methods for aligning Language Learning Models (LLMs) with human norms and objectives, thus making AI 'safe' for use in the wild - i.e. among commonplace users who might prompt it for solutions or assessment of documents etc. Of the several forms of safety models for AI, there are no completely capable ones. There exist many tradeoffs, such as the absence of explicit rules or principles, or the absence of any other digital model that can assure the LLM's output is safe. To top this, any increased complexity for safety can affect performance of the LLM, and thereby its commercial case. The key lies in identifying the LLM that will adapt well to a sectoral use-case with the lowest risk of full-blown safety violations.

Our article discussed the much-used alignment technique of Reinforcement Learning from Human Feedback (RLHF), the alternative Constitutional AI propounded by a rival AI firm, and the upcoming alignment algorithms of SELF-ALIGN, and Reinforcement Learning from Contrast Distillation (RLCD). The alternatives to RLHF here try to take aim at the relatively poor scalability of RLHF, and its dependency on the quality of human raters, but even the alternatives have inherent difficulties in defining "safe" behaviour using a distilled set of principles. A recent mathematical framework called Behaviour Expectation Bounds (BEB) offers the insight that achieving perfect alignment is impossible, and indeed LLMs will remain susceptible to alignment collapse esp. via long prompts.

The manifestation of LLMs' alignment being broken is called 'jailbreaking' and this has prompted a "Red Queen's race" in techniques where developers patch vulnerabilities even as users find new ways to exploit them. The difficulty with these patches is that unlike a typical app bug, the techniques are internal due to the closed-source nature of the well-known LLMs. Jailbreak methods like adversarial examples, prompt injection, and backdoors allow unauthorised users to manipulate LLMs and produce unaligned output. As would be expected, jailbreaks can be used for various malicious purposes, including extracting confidential data from the input base of the LLM, IP theft, or simply fake news propagation at scale. Note that the problem is one of "bug or feature ?": producing fictional stories at scale may well be an LLM objective, and the realisation that it impinges against safety goals is often an afterthought. Jailbreaking risks and reputational harm, not to mention the Dark Web use of these vulnerabilities, has resulted in massive investment from AI companies and lobbying for a favourable regulatory regime.

What brought AI alignment to the fore was the open letter signed by several well-known AI scientists demanding a moratorium on LLM research until the contours of safety became clear. In any case, alignment is expected to be a multidisciplinary endeavour roping in the ancillary fields of law, ethics, and sociology. Almost at a par with the armies of human-moderators working for social media Cos, alignment is also likely to rely greatly on human supervision to observe any risk of jailbreaking: though it is likely that trained manpower of far higher quality will be required. There are promising methods for AI systems' alignment with human values, at least up to a moderate level, and these focus on the LLMs' human minders being able to diagnose a jailbreak and trace its root-cause. Like GDPR for privacy on the Internet, legislative actions will play a crucial role: the EU AI Act is the first off the block, with an emphasis on regulating high-risk AI deployments that Cos. are likely to call as overreach. Yet, collaboration and international cooperation in regulation are critical, as there are global implications and ethical standards (and rights' sensitivities) might vary across cultures. This is further complicated by the geopolitical situation where not all countries might cooperate in a common rules-based regime for AI. These are the issues that need assurance about prior to the arrival of potentially superintelligent entities at decade's end.

References:

1. "Krystal Hu, “ChatGPT sets record for fastest-growing user base - analyst note”, Reuters, Feb 2023, ChatGPT sets record for fastest-growing user base - analyst note | Reuters
2. Noah M Kenney, “A Brief Analysis of the Architecture, Limitations, and Impacts of ChatGPT”, ResearchGate, March 2023, A Brief Analysis of the Architecture, Limitations, and Impacts of ChatGPT | Zenodo
3. Ashish Vaswani, “Attention is all you need”, Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017, [1706.03762] Attention Is All You Need
4. Volodymyr Mnih et.al, “Playing Atari with Deep Reinforcement Learning”,DeepMind Technologies, 2013 , Playing Atari with Deep Reinforcement Learning
5. Meta,”Meta and Microsoft Introduce the Next Generation of Llama”, Meta, July 2023, Meta and Microsoft Introduce the Next Generation of Llama
6. Jarred Walton, “How to Run a ChatGPT Alternative on Your Local PC”, Tom’s Hardware, March 2023, How to Run a ChatGPT Alternative on Your Local PC | Tom's Hardware
7. Jakob Mökander et.al, “Auditing large language models: a three-layered approach”, AI and Ethics, May 2023, Auditing large language models: a three-layered approach | SpringerLink

8. Andy Zou et.al, “Universal and Transferable Adversarial Attacks on Aligned Language Models”, Carnegie Mellon University, July 2023, [2307.15043] Universal and Transferable Adversarial Attacks on Aligned Language Models
9. Miles Klee, “New ChatGPT Lawsuits May Be Start of AI’s Legal Sh-tstorm”, Rolling Stone, June 2023, New ChatGPT Lawsuits May Be Start of AI's Legal Sh-tstorm
10. GDPR-EU, General Data Protection Regulation, GDPR-EU, May 2018, GDPR
11. Andrew Burt, “How will the GDPR impact machine learning?”, O’Reilly, May 2018, How will the GDPR impact machine learning? – O’Reilly
12. Gill et.al, “Conceptual and normative approaches to AI governance for a global digital ecosystem supportive of the UN Sustainable Development Goals (SDGs)”, AI and Ethics, 2022, Conceptual and normative approaches to AI governance for a global digital ecosystem supportive of the UN Sustainable Development Goals (SDGs) | SpringerLink
13. Devon Wood Thomas, “Understanding Large Language Models: AI Alignment and LLMs”, Princeton, November 2022,AI Alignment and LLMs
14. Yoshua Bengio, “Personal and Psychological Dimensions of AI Researchers Confronting AI Catastrophic Risks“, youshuabengio.org, August 2023,Personal and Psychological Dimensions of AI Researchers Confronting AI Catastrophic Risks - Yoshua Bengio
15. Jeremie Harris, “Making AI Safe through Debate”, Towards Data Science Podcast, March 2021, Making AI Safe through Debate. Ethan Perez explains how AI debate… | by Jeremie Harris | Towards Data Science
16. Steven Bills et.al, “Language models can explain neurons in language models”, OpenAI, May 2023, Language models can explain neurons in language models
17. Dongfu Jiang et.al, “LLM-BLENDER: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion”, 61st Annual Meeting of the Association for Computational Linguistics, July 2023, LLM-BLENDER: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion
18. Wolfgang Schulz et,al, “Teaching Norms to Large Language Models – The Next Frontier of Hybrid Governance”, May 2023, Alexander von Humboldt Institute for Internet and Society, Teaching Norms to Large Language Models – The Next Frontier of Hybrid Governance
19. Evan Hubinger, “Exploring safe exploration”, alignmentforum.org, Jan 2020, Exploring safe exploration
20. Geoffrey Irving et.al, “AI safety via debate”, OpenAI, 2018, https://arxiv.org/pdf/1805.00899.pdf

21. Zhengxuan Wu, Atticus Geiger et.al, "Interpretability at Scale: Identifying Causal Mechanisms in Alpaca", Stanford University, May 2023, https://arxiv.org/pdf/2305.08809.pdf
22. Michal Stefanik, "Methods for Estimating and Improving Robustness of Language Models", Conference of the North American Chapter of the Association for Computational Linguistics, 2022, Methods for Estimating and Improving Robustness of Language Models
23. Yuntao Bai et.al, "Constitutional AI: Harmlessness from AI Feedback", Anthropic, Dec 2022, https://arxiv.org/pdf/2212.08073.pdf
24. Minae Kwon, "Reward Design With Language Models", ICLR, 2023, https://openreview.net/pdf?id=10uNUgI5Kl
25. Billy Perrigo, "OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic", TIME, January 2023, OpenAI Used Kenyan Workers on Less Than $2 Per Hour: Exclusive | Time
26. Hugging Face, "The intuition behind PPO", Hugging Face: The Deep Reinforcement Learning Course, 2022, The intuition behind PPO - Hugging Face Deep RL Course
27. Zhiqing Sun et.al, "Principle-Driven Self-Alignment of Language Models from Scratch with Minimal Human Supervision", Dromedary, May 2023, https://arxiv.org/pdf/2305.03047.pdf
28. Amanda Askell et.al, "A General Language Assistant as a Laboratory for Alignment", Anthropic, 2021, https://arxiv.org/abs/2112.00861
29. Kevin Yang et.al, "RLCD: Reinforcement Learning From Contrast Distillation For Language Model Alignment", arXiv preprint, August 2023, https://arxiv.org/pdf/2307.12950.pdf
30. Derek B. Johnson, "Researchers find 'universal' jailbreak prompts for multiple AI chat models, SC Media, July 2023, Researchers find 'universal' jailbreak prompts for multiple AI chat models | SC Media
31. Robin Jia and Percy Liang, "Adversarial Examples for Evaluating Reading Comprehension Systems ", Stanford University, 2017, Adversarial Examples for Evaluating Reading Comprehension Systems
32. Jessica Lyons Hardcastle, "Google AI red team lead says this is how criminals will likely use ML for evil", The Register, August 2023, Google AI red team lead talks real-world attacks on ML • The Register.
33. Shafi Goldwasser et.al, "Planting Undetectable Backdoors in Machine Learning Models", arXiv:2204.06974, April 2022, https://arxiv.org/pdf/2204.06974.pdf
34. Jose Selvi, "Exploring Prompt Injection Attacks", NCC Group, December 2022, Exploring Prompt Injection Attacks | NCC Group Research Blog

35. Arielle Waldman, “MoveIT Transfer attacks highlight SQL injection risks”, Tech Target, June 2023, MoveIT Transfer attacks highlight SQL injection risks | TechTarget
36. Justus Mattern et.al, “Membership Inference Attacks against Language Models via Neighbourhood Comparison ”, arXiv:2305.18462, Aug 2023, https://arxiv.org/pdf/2305.18462.pdf
37. Josh A. Goldstein et.al, “Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations”, arXiv:2301.04246v1, January 2023, https://arxiv.org/pdf/2301.04246.pdf
38. Alexander Wei et.al, “Jailbroken: How Does LLM Safety Training Fail? ”, UC Berkeley, July 2023, https://arxiv.org/pdf/2307.02483.pdf
39. Wes Davis, “Sarah Silverman is suing OpenAI and Meta for copyright infringement”,The Verge, July 2023, Sarah Silverman is suing OpenAI and Meta for copyright infringement - The Verge
40. Melissa Govender, “Access Alert: AI Data Harvesting – Ethical? Monopolistic?”, Access Partnership, July 2023, Access Alert: AI Data Harvesting – Ethical? Monopolistic?
41. Jessica Berch, “Access Alert: UK’s AI Safety Summit”, Access Partnership, August 2023, Access Alert: UK’s AI Safety Summit
42. Jan Leike et.al, “Our approach to alignment research’, Open AI, August 2022, Our approach to alignment research)
43. Oscar Schwartz, “In 2016, Microsoft’s Racist Chatbot Revealed the Dangers of Online Conversation”, IEEE, November 2019, In 2016, Microsoft’s Racist Chatbot Revealed the Dangers of Online Conversation - IEEE Spectrum
44. Chris Olah,”Interpretability Dreams”, Informal notes, May 2023, Interpretability Dreams
45. Deep Ganguli et.al, “The Capacity for Moral Self-Correction in Large Language Models”, Anthropic, February 2023, https://arxiv.org/pdf/2302.07459.pdf)
46. Jan Leike and Ilya Sutskever, "Introducing Superalignment", Open AI, July 2023, Introducing Superalignment
47. Ryan Morrison, “OpenAI commits to ‘superalignment’ research”,Tech Monitor, July 2023, AI alignment: OpenAI commits to 'superalignment' research)
48. Ben Hou and Brian Patrick Green, “A Multilevel Framework for the AI Alignment Problem”, Markkula Center for Applied Ethics, July 2022, A Multilevel Framework for the AI Alignment Problem - Markkula Center for Applied Ethics
49. Josh Taylor and Alex Hern, “Godfather of AI’ Geoffrey Hinton quits Google and warns over dangers of misinformation”, The Guardian, May 2023, Godfather of AI’ Geoffrey Hinton quits Google

Research Office
Indian Institute of Management Kozhikode
IIMK Campus P. O.,
Kozhikode, Kerala, India,
PIN - 673 570
Phone: +91-495-2809237/ 238
Email: research@iimk.ac.in
Web: https://iimk.ac.in/publications