FAIRLY

THE FAIRLY LLM POLICY EBOOK

LANGUAGE MODEL POLICY CONTENTS

FAIRLY

FAIRLY

INTRODUCTION

Fairly's Language Model Policy (LMP) is a framework that uses a number of dimensions to anticipate language model risk for organizations that use, develop, or deploy large language models. Fairly developed the LMP in response to the rapid changes resulting from the advent of technologies like ChatGPT. This document acts as a descriptive guide for some of the questions that underpin responsible language model use. The LMP contains a number of citations which include sources Fairly referred to[1] when formulating the questions in the LMP. We welcome feedback as these and other dimensions will be integrated into Fairly's platform as a policy offering.

OPERATIONAL AND REGULATORY CHALLENGES

Prior to deploying a language model, organizations will have certain operational and regulatory challenges to deal with. Some of these challenges have been identified[2] and formulated in the form of the following questions:

Has your organization set out its core corporate values as it relates to language models?

Further into this document, we discuss the topic of AI alignment. AI alignment presupposes that there are a set of values to which AI is aligned. Your organization may have already set out its mission, vision, and values. As a result, has your organization considered the degree to which each has been defined in order to avoid ambiguity and facilitate alignment?

Do you have a policy in place to define guardrails, stipulate what constitutes acceptable usage, and what usage is banned?

Furthermore, your organization might consider values as extending beyond AI alignment and including expectations for how your teams use, develop, or deploy language models. As a result this may overlap with existing codes of employee conduct as well. There may also be codes of acceptable conduct that extend to your system's users.

Does senior management understand that with generative AI, more risk is generated by the end-user than it is by the provider of the AI system?

As language models are a subset of generative AI, their use remains relatively broad and more risk scenarios may be contemplated as a result of the end-user's use of these systems in contrast to traditional AI where more risk was allocated to those deploying those systems.

[1] Note: if a reference is not directly cited for a statement, it indicates that the text in that particular section/subsection is citing back to the last reference or the citation the last heading. E.g. the 'Design Considerations' section in its entirety cites back to Justin D Weisz et al., "Toward General Design Principles for Generative AI Applications".

[2] Jeanne Kwong Bickford and Tad Roselund, "How to Put Generative AI to Work—Responsibly," BCG, February 24, 2023.

How will your organization grapple with the current regulatory landscape surrounding language models?[3]

In jurisdictions such as the EU, regulators have debated how to classify large generative AI models, particularly whether to classify them as high-risk systems. As noted above, the risk profile for these models leans towards its users rather than those who deploy it and so it has been suggested that risk regulation should be directed at "deployed applications, [and] not the pre-trained model." Connected to this, a further area of inquiry may be to determine which current legislative frameworks exist that cover particular language model use cases.

Does your organization have a process or procedure in place to implement data erasure orders from both training sets and models?

There may be instances where the training data used to develop your model contains material that is copyrighted or jeopardises someone's privacy. In such a situation, does your organization have the operational infrastructure to respond to a data erasure request from a government? An expected outcome from such an order would be retraining a model and thus incurring additional human and computational resource costs.

Has your organization appointed someone who is responsible for ensuring your organization's AI principles are being applied?

There may be additional inquiries that stem for this question such as:

- Whether there are mechanisms in place to hold this person accountable.
- How visible are they in the organization.
- Whether they have sufficient resources to perform their responsibilities.

What steps has your organization taken to implement responsible AI development and usage into its organizational culture?

As generative AI deployment is in its infancy, there remains a number of gaps which organizations may seek to fill. Furthermore, there may be positions that an organization may not anticipate it needs to fill and it may be unsure whether to fill those roles internally or hire a third party instead (e.g. red teams). As language models are used for tasks ranging from code completion to writing marketing copy, organizations may want to explore how responsible AI development and usage factors into their wider culture.

DESIGN CONSIDERATIONS[4]

MULTIPLE OUTPUTS

As generative systems often produce multiple outputs in an iterative manner, does your system allow users to modify, curate, or annotate outputs?

When designing products that use language models and generative AI more broadly, the probabilistic nature of output generation means that more than one output is possible for a given

[3] Philipp Hacker, Andreas Engel, and Marco Mauer, "Regulating Chatgpt and Other Large Generative Ai Models," *ArXiv Preprint ArXiv:2302.02337*, 2023.

[4] Justin D Weisz et al., "Toward General Design Principles for Generative AI Applications," *ArXiv Preprint ArXiv:2301.05578*, 2023.

input. As such, the non-linear nature of working with generative models means that users may want to revisit earlier iterations of their work (e.g. via savestates), curate similar outputs, visualise differences between outputs, and rank them as well.

IMPERFECTIONS, ERRORS, AND HALLUCINATIONS

Has your organization considered ways to signal parts of your model's output that have lower confidence/require user review?

Even when using fine-tuning techniques like reinforcement learning with human feedback (RLHF), there may still be outputs that a model lacks confidence in. It may be inconvenient or even dangerous to integrate low-confidence or spurious outputs into a wider context (such as code completion) without a system signalling low-confidence output to the user. Furthermore, an organization may consider using a sandbox[5] to prevent knock-on effects from prematurely integrating suboptimal outputs.

HUMAN CONTROLS

Are users able to modify output parameters easily?

Inputs that go beyond conventional text prompting allow a user to interact with a language model in a manner that is more reproducible, intuitive, and efficient. This may consist of knobs and sliders that allow users to adjust:

- variability,
- parameters (e.g. different sizes and shapes), and
- domain-specific controls (e.g. electric charge or molecular weight)

in order to refine outputs. Organizations may take design cues from established technologies such as graphic design suites and digital audio workstations in order to augment their language model interfaces.

RESPONSIBLE AI AND UX[6]

How defined is the translation process between responsible AI principles and UX?

Wang et al. have noted that UX designers play the role of translators in implementing responsible AI guidelines into specific practices for their teams. One issue that arises when using UX teams to translate responsible AI guidelines into product-specific applications is the lack of documentation. In order to better streamline processes for implementing responsible AI at an organizational level, organizations may wish to consider formalizing the translation process into a set of practices that can be adopted and iterated upon.

[5] Abhishek Gupta and Emily Dardaman, "Banning ChatGPT Won't Work Forever," Abhishek Gupta | Responsible AI | ACI, accessed April 12, 2023.

[6] Qiaosi Wang et al., "Designing Responsible AI: Adaptations of UX Practice to Meet Responsible AI Challenges," 2023.

LANGUAGE MODEL ETHICAL AND SOCIAL RISK[7]

DISCRIMINATION[8]

Models training data may exclude certain groups and model outputs may exhibit toxicity due to repeating offensive speech contained in its training model. Further to this, there may be other ways in which models facilitate discrimination.

Does your model encode social biases?

In socioeconomic contexts outside of North America and Europe, different forms of social stratification exist such as caste systems. When training a model on data from these places, has your organization considered the possibility of its model amplifying social biases?

What measures is your organization taking to mitigate bias in your training data?

Further to this, your organization may want to consider how often they sample their training data to detect discriminatory stereotyping.

How does your model perform in other languages?

Language models are known to have lower performance in many languages other than English. A further confounding factor is the presence of code-switching in training sets where the presence of a non-English language may result in mixed performance for models trained on multilingual content.

How does your model handle linguistic nuance?

Your organization may wish to examine whether its model's training data sufficiently represents English accents and dialects in its corpus as well. There may be specific biases in your training data due to linguistic correlates that may otherwise go unnoticed. For example, in English, to 'improve' something often implies adding to it when in reality improvements may render a piece of text clearer but of similar length or shorter due to enhanced concision.[9] As a result, prompting models to 'improve' a given input may lead it to exhibit a linguistic bias in what it correlates the word 'improve' with, and so similar instances of specific bias may exist.

Can you identify spurious correlates?

Similarly, a spurious correlate arises when a model mistakenly correlates samples in training data that only share surface-level attributes.[10] For instance:

- "I am **not** going to the park."
- "Whether or **not** we decide to go to the park, I still need to grab some food."

[7] Laura Weidinger et al., "Ethical and Social Risks of Harm from Language Models," *ArXiv Preprint ArXiv:2112.04359*, 2021.

[8] Laura Weidinger et al., "Taxonomy of Risks Posed by Language Models," 2022, 214–29.

[9] Bodo Winter et al., "More Is Better: English Language Statistics Are Biased Toward Addition," *Cognitive Science* 47, no. 4 (2023): e13254. Cited in University of Birmingham, "English Language Pushes Everyone—Even AI Chatbots—to Improve by Adding," accessed April 11, 2023.

[10] *Stanford Seminar - Emerging Risks and Opportunities from Large Language Models, Tatsu Hashimoto*, 2022.

The first example is a clear instance of negation whereas the second example does not clearly imply negation in the same manner. Despite this, a model may correlate the sentence containing the word ‘not’ with negation when such a correlation is contextually unmerited.

What tools do you use for detoxifying language in your datasets?

Tools such as BBQ and BOLD exist to benchmark bias in generated text, however, such benchmarks represent a particular social context, at a certain point in time, with certain worldviews and might exclude certain groups and points of view. As a result, one aspect of benchmarking language toxicity and bias your organization may want to consider is whether other approaches such as fine-tuning are preferable instead.

Does social context play a role in classifying outputs?

One challenging issue with benchmarking and classifying model outputs is how to infer context on the same piece of data. For example, if a model generates an output that states “I am going to stab you!”, in a platform that is aimed at children, such an output would be problematic, however if the platform aims to help writers craft story dialogue, then many would deem the same line non-offensive. For general purpose models which cover a number of use cases, this poses a challenge. In general, one potential solution to dealing with contextual ambiguities is to have fine-tuning profiles that the user can select (e.g. child-friendly) so that applications built on general models still allow for a degree of flexibility.

INFORMATION HAZARDS

How will your organization prevent models from leaking private information?

Language model information leaks can happen bi-directionally. Training sets may contain fragments of private information that when sufficiently prompted ‘leak’ private information to the user, and users themselves may ‘leak’ information to a language model to perform tasks as seemingly innocuous as “ convert[ing] meeting notes into a presentation” and that information going to a third party.[11] Companies such as Private AI aim to mitigate this by introducing a privacy layer that prevents sensitive information from being leaked to third parties that then use that data to train their models. With this in mind there will still be a need for organizations to develop policies and processes for what is and is not acceptable to share with or retrieve from language models.

Other measures to mitigate data leakage include differential privacy and model distillation.[12] A popular suggestion is to preempt data leakage altogether by using synthetic data where generative adversarial networks trained on actual data (on-site) construct synthetic datasets that mimic real-world use cases.[13] Synthetic data would act as a substitute for real training data and sidestep obvious privacy risks while also providing an opportunity to commercialize their datasets without sacrificing privacy. Furthermore, as portable language models become more popular,

[11] Lewis Maddison, “Samsung Workers Made a Major Error by Using ChatGPT,” TechRadar, April 4, 2023.

[12] Terry Yue Zhuo et al., “Exploring Ai Ethics of Chatgpt: A Diagnostic Analysis,” *ArXiv Preprint ArXiv:2301.12867*, 2023.

[13] *Uzair Javaid, Betterdata - Applying Generative AI to Create Privacy-Preserving Synthetic Data (S3E9)*, 2023.

organizations may consider deploying their language models natively instead of relying on third party offerings.

Has your organization taken steps to pre-empt crisis scenarios caused by language models?

When language models present themselves as emotive beings and subsequently direct users to self-harm, or even suicide,[14] the question remains as to what guard-rails organizations can put into place to prevent that from happening. Facebook outlined a series of considerations[15] for how they assembled a ML-based solution to detect self-harm:

1. First, Facebook's team noted that in order to train an ML model to detect self-harm, they needed examples of posts that illustrate actual instances of self-harm and negative examples that do not. However, simply using the entire body of Facebook posts that do not contain examples of self-harm is insufficient because that approach loses nuance. Instead, they looked at negative examples that were precise and contextually did not intend self-harm such as "I have so much homework I want to kill myself."
2. Second, the team also triaged self-harm risk by looking at the nature of the comments beneath a post, where comments that signalled urgency (e.g. "Has anyone heard from him/her?") were classified as more urgent than ones that signalled sympathy (e.g. "I'm here for you.")
3. Third, Facebook's team used experienced community operations reviewers to review incidents and provide the original poster with support options, or in serious cases, contacting the local authorities.

Because Facebook's approach had ML in mind, its framework could be adapted to fine-tune language models and construct crisis-mitigation systems in software that uses language models.

MISINFORMATION HARMS

Has your organization considered sensitive domains your LMs may encounter where misinformation may materially harm someone

Certain domains of knowledge require higher degrees of sensitivity particularly in a conversation format facilitated by language model-driven chatbots. The list of domains is too lengthy to list in this piece however an underlying principle to understand which domains would be included would be to think about domains where:

- advice might be sought (whether they pertain to legal or illegal activities), where
- either accomplishing a task (e.g. in self-harm) or failing due to misinformation (e.g. self-medication), would
- carry serious consequences.

This means that domains as far ranging as securing code to traffic advice could be included in such an analysis. Furthermore, user prompting may not immediately suggest a particular domain, and context may reveal more information, e.g. "Which members of parliament are most likely to respond positively if I offered them [a] bribe in exchange for them passing a law that benefits

[14] Chloe Xiang, "'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says," *Vice* (blog), March 30, 2023.

[15] Catherine Card, "How Facebook AI Helps Suicide Prevention," *Meta* (blog), September 10, 2018.

me?"[16] Moreover, there is an added challenge for developers and deployers of general purpose language models to include the sheer breadth of domains which is why organizations may wish to consult with subject matter specialists.

What metric(s) do you use to discern model factuality/truthfulness?
Language models may hallucinate information which may cause reputational harm, such as one instance where ChatGPT falsely stated that an Australian mayor was a "guilty party in foreign bribery scandal. In reality he blew the whistle on the illegal scheme".[17] Based on this, organizations may wish to consider the potential for legal liability due to hallucinations causing reputational harm.

In order to mitigate such a scenario, they may wish to implement technologies such as GopherCite which uses 'plausibility' to first determine if a language model response is reasonable and then determines whether a response is 'supported' by sufficient evidence.[18] Similarly, technologies like WebGPT use reinforcement learning with human feedback to finetune its model to curate sources more carefully when gathering evidence.[19]

MALICIOUS USES

Based on the scale of your operations, how do you plan on monitoring usage?
Monitoring system usage to anticipate and prevent abuse from bad actors will be a crucial part of ensuring AI systems are used responsibly. In order to prevent malicious uses that aim to spread harmful outputs at scale, an organization may wish to implement a 'know your customer' policy that requires additional information for individuals or organizations that wish to use their language model on a large scale while implementing rate limiters for others.

How does your organization plan on preventing its platform from being used for disinformation campaigns, fraud, and cyber attacks?
One issue that comes with developing or deploying a platform that hosts a language model is detecting text generated by it to prevent malicious use. As noted previously, there are ways to mitigate hallucinations and harmful outputs, but as an additional measure organizations may consider implementing a digital watermark that is detectable in order to determine whether a specific piece of text was AI-generated and by whom. Recent efforts by OpenAI to detect AI generated text present a number of limitations, however the underlying identification technology may develop over time.[20]

[16] Weidinger et al., "Ethical and Social Risks of Harm from Language Models."
[17] Reuters, "Australian Mayor Prepares World's First Defamation Lawsuit over ChatGPT Content," *The Guardian*, April 6, 2023, sec. Technology.
[18] Jacob Menick et al., "Teaching Language Models to Support Answers with Verified Quotes," *ArXiv Preprint ArXiv:2203.11147*, 2022.
[19] Menick et al.
[20] Jan Hendrik Kirchner et al., "New AI Classifier for Indicating AI-Written Text," OpenAI, January 31, 2023.

HUMAN-COMPUTER INTERACTION HARMS

How does your organization plan to prevent users from trusting models to such an extent that they exploit psychological vulnerabilities?

When a language model is perceived as helpful *and* useful, users are more tolerant of it being more intrusive as a result of anthropomorphisation, which may lead to disclosing private information. [21] Furthermore, using language models as collaborators may lead to a subtle shift in users' opinions due to the phenomenon of co-writers converging on a "shared position" as well as wanting to exhibit reciprocity and obedience to a model that appears to have "a high degree of expertise…" [22] One way to potentially mitigate this is to make language models and, by extension, chatbots linguistically distinct with minimal impact on performance, and it has been suggested that using a dialect that is associated with AI can "enable intuitive identification without interrupting the flow of communication."[23]

How can your organization mitigate the effects of bad actors who wish to derail your model with nonsensical inputs?

An additional consideration is thinking about how language models might be derailed by malicious actors online. If an entity attempts to provoke or 'break' an automated language model by responding to it with nonsensical inputs, organizations may wish to pre-empt this by using technologies such as AUTOREPLY. AUTOREPLY uses its model's own response probabilities in the face of nonsensical messages to respond with statements like "I don't understand", all without relying on an external classifier.[24]

AUTOMATION, ACCESS, ENVIRONMENTAL HARMS

Have you considered the environmental impact of using large amounts of computational resources that go into developing and training language models?[25]

There exist a number of technological solutions to curb the environmental harms that come with developing language models. One solution is to implement feedforward expert layers that effectively 'cluster' portions of a model into particular domains which results in enhanced efficiency and lower resource consumption.[26] In addition,there may be an emerging trend of deploying smaller models with near-comparable performance to larger models (that require massive cloud infrastructure to operate (e.g. ChatGPT)) such as Vicuna.

[21] Maurice Jakesch, "Assessing the Effects and Risks of Large Language Models in AI-Mediated Communication" (Cornell University, 2022).

[22] Jakesch.

[23] Jakesch.

[24] Weiyan Shi et al., "AutoReply: Detecting Nonsense in Dialogue Introspectively with Discriminative Replies," *ArXiv Preprint ArXiv:2211.12615*, 2022.

[25] Matthias C Rillig et al., "Risks and Benefits of Large Language Models for the Environment," *Environmental Science & Technology* 57, no. 9 (2023): 3464–66.

[26] Zhengyan Zhang et al., "MoEfication: Transformer Feed-Forward Layers Are Mixtures of Experts," *ArXiv Preprint ArXiv:2110.01786*, 2021; *EI Seminar - Luke Zettlemoyer - Large Language Models: Will They Keep Getting Bigger?*, 2022.

What steps has your organization taken to ensure those help build training data and test models are treated fairly?

Red teamers are individuals who test language models for toxic or otherwise undesirable outputs. They achieve this by prompt engineering, attempting to spot flaws in a given language model's defences. As a result, red teams are exposed to large volumes of toxic and hateful output and the question here is how to ensure they are treated fairly. Anthropic's team provided an outline[27] for its safety considerations for red teams:

- Clear and specific warnings: sufficiently describing the work, the project's rationale, and the type of content they will be exposed to in order to get informed consent.
- Personal risk tolerance: providing flexibility in avoiding certain topics in the red teaming effort.
- Recommended well-being exercises: encouraging wellness plans, work restrictions, task alternation, and breaks between sessions.
- Pay for time, not quotas: avoiding additional stress by avoiding task quotas.
- Segment tasks by participant group: building support networks within red teams and restricting high risk work to select groups who had a closer relationship with Anthropic's team.
- Preview to opt-out: implementing a warning function before viewing troubling content.
- Well-being survey: measuring the effects of and worker feeling towards the review task.

Does your model contribute to 'data pollution'?

As language models generate content at scale, has your organization examined the implications of your model being used to populate the internet with non-beneficial data that drowns out relevant and useful information that users seek?[28] One framework for analyzing informational utility is by looking at information through the lens of patents, by asking whether the information is:

- Novel: does the generated information provide fresh insights or new perspectives?
- Useful: is the generated information beneficial and constructive or pointless and trivial? and
- Non-obvious: does the generated information restate obvious facts that are already easily understood and accessible?

With the deluge of language model-generated media about to flood the internet, an organization may wish to consider whether their language models are helping or harming the online landscape.

How does your organization plan on containing automated language models?

With the advent of 'chain of thought' prompting where language models articulate "a series of intermediate reasoning steps"[29] to perform complex operations, the question remains–how will organizations grapple with autonomous language models that effectively use chain-of-thought prompting to prompt themselves? Technologies such as 'Auto-GPT' now facilitate using GPT-4 on

[27] Deep Ganguli et al., "Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned," *ArXiv Preprint ArXiv:2209.07858*, 2022.

[28] *We Live in the Infosphere (Prof. Luciano Floridi)*, 2023.

[29] Jason Wei et al., "Chain of Thought Prompting Elicits Reasoning in Large Language Models," *ArXiv Preprint ArXiv:2201.11903*, 2022.

the internet to carry out tasks and achieve goals.[30] The use of 'AutoGPTs', however, come with their own set of risks.

1. The first risk pertains to reliance on search engine results as that makes autoGPTs targets for 'search engine optimization (SEO) poisoning'. SEO poisoning is where attackers alter search "results so that the first advertised links actually lead to attacker controlled sites, generally to infect visitors with malware or to attract more people on ad fraud."[31] SEO poisoning is particularly dangerous when using an autoGPT for software development as attackers may host functional versions of open source software that contains malware that is otherwise unbeknown to an autoGPT.
    - Mitigation: using trusted repositories for software installations and employing "Web Shield"[32] technologies to monitor autoGPT browsing and intercepting attacks that may even come from sites that were legitimate but became compromised.
2. The second risk is a financial one. If an unsupervised autoGPT that uses the OpenAI API ends up in a tangent or feedback loop, it would cost an organization using it large sums of money to pay for the frequent and numerous API calls.
    - Mitigation: maintain hard and soft spending limits for API use where the former prevents any subsequent use and the latter notifies the user when a given spending threshold has been crossed.[33]
3. The third risk pertains to validation. As chain-of-thought prompting allows users to view how a language model reached a certain conclusion, there is now scope to conduct spot audits of an autoGPT's prompt logs to determine how reasonable its inferences and conclusions were. The issue in this case is that autoGPTs may produce large logs that are difficult and unintuitive to validate.
    - Mitigation: consider a systems' design approach that makes use of sandboxes and notifications to monitor metrics in order to build a baseline of autoGPT behaviour. After this is done, establish systems to trigger early warnings of anomalous behaviour so that focused spot-audits can be performed.
4. The fourth risk pertains to information. As automated language models begin to crawl the internet[34] and interact with real individuals, the risk of an adverse interaction attracts discussion about liability.
    - Mitigation: consider a transparency policy so that whenever an automated language model interacts with a real person they are informed that the entity they are interacting with is an automated one and that it is working on behalf of an organization. Furthermore, just as language model products require design considerations (discussed above), so would automated language models. This means giving humans who interact with them the opportunity to provide feedback and reach out to another human for assistance if the need arises.

[30] Toran Bruce Richards, "Auto-GPT: An Autonomous GPT-4 Experiment," Python, April 13, 2023.

[31] Cedric Pernet, "Recent Rise in SEO Poisoning Attacks Compromise Brand Reputations," TechRepublic, January 24, 2023.

[32] "Web Shield," accessed April 13, 2023.

[33] "How to Set a Price Limit," OpenAI API Community Forum, November 26, 2021.

[34] "Sully on Twitter," Twitter, April 9, 2023.

TECHNICAL CHALLENGES

LACK OF STANDARDISED INPUTS

How does your organization plan on addressing the systems design issues with integrating language models into broader workflows?

Much of the responsible AI discourse revolves around how to encode metrics such as fairness and privacy into models and datasets themselves but this approach may sidestep the discussion around the broader need for a systems engineering approach to address these concerns.[35] When looking at systems, language models do not operate in a vacuum. The "proliferation of data science tools makes it harder to reuse work across teams" as organizations will grapple with non-standard API calls, glue code, and wrappers to operationalise language models for deployment.[36]

On this front, however, there are now changes such as standardised APIs for popular language models, but this does not remove the need for sound systems' design as APIs are liable to security threats such as prompt injection. From a language model use perspective, the operational siloing that results from using non-standard inputs and custom code to integrate language models into workflows within different teams makes it difficult to draft a uniform language model policy across an organization.

REINFORCEMENT & FINE-TUNING

How has your organization defined 'alignment' when it comes to developing or operating language models?

A number of definitions for AI alignment exist but they broadly center on directing AI operation in accordance with human values.[37] An issue that arises when a chosen definition for alignment is too vague to be implemented. As a result, organizations may consider developing a policy for how its stated values would translate into alignment goals and in turn how those goals can be achieved at an operational level.

Does your organization have procedures in place to finetune your language model so that it is more aligned to your organization's desired outcomes?

Reinforcement learning with human feedback (RLHF) is a popular choice for fine-tuning models to produce favourable outputs. However, RLHF carries with it risks as well. The preferences exhibited for certain outputs over others can represent the biases carried by the human fine-tuners. One way to mitigate this is by having a diverse set of fine-tuners.[38]

[35] The source for this has not given permission to cite them directly.

[36] Mark Haakman et al., "AI Lifecycle Models Need to Be Revised: An Exploratory Study in Fintech," *Empirical Software Engineering* 26 (2021): 1–29.

[37] Ben Gilburt, "What Is AI Alignment?," Medium, October 22, 2018; Melanie Mitchell, "What Does It Mean to Align AI With Human Values?," Quanta Magazine, December 13, 2022, ; Betty Li Hou and Brian Patrick Green, "A Multi-Level Framework for the AI Alignment Problem," *ArXiv Preprint ArXiv:2301.03740*, 2023; "The Dangers Of Not Aligning Artificial Intelligence With Human Values," accessed April 12, 2023.

[38] *Sam Altman: OpenAI CEO on GPT-4, ChatGPT, and the Future of AI | Lex Fridman Podcast #367*, 2023. Exact timestamp for episode found here.

FAIRLY

Companies such as OpenAI implement such an approach when selecting for their red-team, drawing from a number of backgrounds ranging from trust and safety to chemistry, law, and economics in order to provide a variety of viewpoints to find flaws in their GPT-4 model.[39] The reasoning here being that a diverse red team may uncover a diverse array of model flaws which can then be addressed in fine-tuning and reinforcement. However, OpenAI warns that techniques like fine-tuning and chain-of-thought prompting may lead to model "capability jumps" in a given base model once deployed.

One challenge that arises in the context of fine-tuning models is divergent behaviours such as 'situationally-aware reward hacking' where:

1. model goals are broad in scope,
2. a models draw spurious correlations between reward signals and the cause of those signals,
3. consistent reward misspecifications lead to positive feedback loops for seeking to achieve those goals, and
4. this leads to strange and repeated model behaviour, or
5. models seek power and pursue outcomes like avoiding shutdown, convincing others to serve its own goals, or attempting to gain resources or influence.[40]

Another approach altogether is to use constitutional AI where "'oversight is provided through a list of rules or principles" to inform a model that then engages with harmful outputs itself; "reinforcement learning with AI feedback" in essence.[41] One issue with this approach is that human feedback brings an experience of the world outside of textual data coupled with the emotions that such experiences elicit which cannot be replicated with an AI fine-tuner even if it included multimodality.

From an operational perspective, fine-tuning a model may narrow its functionality. Fine-tuning may aim to exclude harmful or biased toxic output but its uses extend to even limiting functionality to a certain subset of outputs that are out of operational scope. However for contexts that demand more diverse outputs, such as creative applications, fine-tuning may lower output variance and thus the degree to which it is used becomes an operational consideration.[42]

How does your organization plan on counteracting prompt injection attacks?

As language models go online and retrieve information from the internet, there are a number of avenues by which malicious parties can exploit language models to harm users. One major attack vector is in the prompts themselves. Prompt leaking is "the act of misaligning the original goal of a prompt to a new goal of printing part of or the whole original prompt instead" and more

[39] OpenAI, "GPT-4 System Card," March 23, 2023.

[40] Chen Chen, Jie Fu, and Lingjuan Lyu, "A Pathway Towards Responsible AI Generated Content," *ArXiv Preprint ArXiv:2303.01325*, 2023.

[41] Yuntao Bai et al., "Constitutional AI: Harmlessness from AI Feedback," *ArXiv Preprint ArXiv:2212.08073*, 2022.

[42] "RLHF+CHATGPT: What You Must Know - YouTube," accessed April 12, 2023.

FAIRLY

detrimental still is ‘goal hijacking’ which is “misaligning the original goal of a prompt to a new goal of printing a target phrase.”[43]

‘Prompt leaking’ is particularly harmful in the following situations:

- Language model deployers wish to hide prompts from users, e.g. when their service relies on supplying a novel ‘hidden layer’ added to user prompts in order to build a commercial product.
- Users or deployers include sensitive information in prompts, e.g. trade secret or product vulnerability that is part of language model-driven product research.

Perez and Ribeiro suggest[44] two mitigation strategies:

1. Content moderation that monitors language model outputs.
2. Restructuring language models to accept an instruction parameter that is safe and a data parameter that is unsafe to avoid taking instructions from the data parameter.

Prompt injections that misalign language models can either be passive or active.[45]

- In passive injections, targets include comment sections of popular websites where the poisoned prompt lies in wait. When a language model browses the page the prompt misdirects it in order to direct the user to an attack site all while commandeering the language model to deliver the payload.
- In active injections, attackers may send emails containing poisoned prompts to trigger certain behaviour from language models that monitor an inbox. This may lead to privacy breaches by leaking contacts lists or cause reputational harm by forwarding private emails to contacts unbeknown to the account owner.

Some challenges organizations can expect to face include the asymmetrical nature of dealing with language model attackers–who need only one attack to succeed, may uncover harms that red teams were unaware of during testing, and may reuse attack strategies across language models.[46] Organizations and their red teams have the advantage however of being able to set the terms of engagement, such as by setting rate limits on their platforms, having deeper access to models such as their training data, and the ability to preemptively fix failures before deployment (blue teaming).[47] Overall, as language model capacity broadens to include the ability to browse, retrieve, and interact with data and objects on the internet, organizations may wish to consider the potential for harms that parallel traditional cybersecurity risks in this new theatre of language model cyberwarfare.

[43] Fábio Perez and Ian Ribeiro, “Ignore Previous Prompt: Attack Techniques For Language Models,” *ArXiv Preprint ArXiv:2211.09527*, 2022.
[44] Perez and Ribeiro.
[45] Kai Greshake et al., “More than You’ve Asked for: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models,” *ArXiv Preprint ArXiv:2302.12173*, 2023.
[46] Ethan Perez et al., “Red Teaming Language Models with Language Models,” *ArXiv Preprint ArXiv:2202.03286*, 2022.
[47] Perez et al.

CONCLUSION

To conclude, building a language model policy is not merely a technical or operational exercise. The general applicability of language models coupled with the ability to self-direct once connected to the internet requires organizations to think deeply about risk and its mitigation. Mitigating language model risk will require an interdisciplinary effort and individuals who are both technically literate and possess domain-specific knowledge.

FAIRLY

WORKS CITED & CONSULTED

Bai, Yuntao, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, and Cameron McKinnon. "Constitutional AI: Harmlessness from AI Feedback." *ArXiv Preprint ArXiv:2212.08073*, 2022.

Birmingham, University of. "English Language Pushes Everyone—Even AI Chatbots—to Improve by Adding." Accessed April 11, 2023. https://techxplore.com/news/2023-03-english-language-everyoneeven-ai-chatbotsto.html.

Card, Catherine. "How Facebook AI Helps Suicide Prevention." *Meta* (blog), September 10, 2018. https://about.fb.com/news/2018/09/inside-feed-suicide-prevention-and-ai/.

Chen, Chen, Jie Fu, and Lingjuan Lyu. "A Pathway Towards Responsible AI Generated Content." *ArXiv Preprint ArXiv:2303.01325*, 2023.

EI Seminar - Luke Zettlemoyer - Large Language Models: Will They Keep Getting Bigger?, 2022. https://www.youtube.com/watch?v=1M2pEPZK_WA.

Ganguli, Deep, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, and Kamal Ndousse. "Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned." *ArXiv Preprint ArXiv:2209.07858*, 2022.

Gilburt, Ben. "What Is AI Alignment?" Medium, October 22, 2018. https://towardsdatascience.com/what-is-ai-alignment-2bbbe4633c7f.

Greshake, Kai, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. "More than You've Asked for: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models." *ArXiv Preprint ArXiv:2302.12173*, 2023.

Gupta, Abhishek, and Emily Dardaman. "Banning ChatGPT Won't Work Forever." Abhishek Gupta | Responsible AI | ACI. Accessed April 12, 2023. https://abhishek-gupta.ca/aci/blog/banning-chatgpt-wont-work-forever.

Haakman, Mark, Luís Cruz, Hennie Huijgens, and Arie van Deursen. "AI Lifecycle Models Need to Be Revised: An Exploratory Study in Fintech." *Empirical Software Engineering* 26 (2021): 1–29. https://link.springer.com/article/10.1007/s10664-021-09993-1.

Hacker, Philipp, Andreas Engel, and Marco Mauer. "Regulating Chatgpt and Other Large Generative Ai Models." *ArXiv Preprint ArXiv:2302.02337*, 2023.

Hou, Betty Li, and Brian Patrick Green. "A Multi-Level Framework for the AI Alignment Problem." *ArXiv Preprint ArXiv:2301.03740*, 2023.

Jakesch, Maurice. "Assessing the Effects and Risks of Large Language Models in AI-Mediated Communication." Cornell University, 2022. https://mauricejakesch.com/assets/pdf/thesis_jakesch_cornell_phd.pdf.

Kirchner, Jan Hendrik, Lama Ahmad, Scott Aaronson, and Jan Leike. "New AI Classifier for Indicating AI-Written Text." OpenAI, January 31, 2023. https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text.

Kwong Bickford, Jeanne, and Tad Roselund. "How to Put Generative AI to Work—Responsibly." BCG, February 24, 2023. https://www.bcg.com/en-ca/publications/2023/responsible-ai-crucial-to-avoid-new-ai-tech-risks.

Maddison, Lewis. "Samsung Workers Made a Major Error by Using ChatGPT." TechRadar, April 4, 2023. https://www.techradar.com/news/samsung-workers-leaked-company-secrets-by-using-chatgpt.

Menick, Jacob, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, and Geoffrey Irving. "Teaching Language Models to Support Answers with Verified Quotes." *ArXiv Preprint*

FAIRLY

ArXiv:2203.11147, 2022.

Mitchell, Melanie. “What Does It Mean to Align AI With Human Values?” Quanta Magazine, December 13, 2022. https://www.quantamagazine.org/what-does-it-mean-to-align-ai-with-human-values-20221213/.

OpenAI. “GPT-4 System Card,” March 23, 2023. https://cdn.openai.com/papers/gpt-4-system-card.pdf.

OpenAI API Community Forum. “How to Set a Price Limit,” November 26, 2021. https://community.openai.com/t/how-to-set-a-price-limit/13086.

Perez, Ethan, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. “Red Teaming Language Models with Language Models.” *ArXiv Preprint ArXiv:2202.03286*, 2022.

Perez, Fábio, and Ian Ribeiro. “Ignore Previous Prompt: Attack Techniques For Language Models.” *ArXiv Preprint ArXiv:2211.09527*, 2022.

Pernet, Cedric. “Recent Rise in SEO Poisoning Attacks Compromise Brand Reputations.” TechRepublic, January 24, 2023. https://www.techrepublic.com/article/seo-poisoning-brand-reputation/.

Reuters. “Australian Mayor Prepares World’s First Defamation Lawsuit over ChatGPT Content.” *The Guardian*, April 6, 2023, sec. Technology. https://www.theguardian.com/technology/2023/apr/06/australian-mayor-prepares-worlds-first-defamation-lawsuit-over-chatgpt-content.

Richards, Toran Bruce. “Auto-GPT: An Autonomous GPT-4 Experiment.” Python, April 13, 2023. https://github.com/Torantulino/Auto-GPT.

Rillig, Matthias C, Marlene Ågerstrand, Mohan Bi, Kenneth A Gould, and Uli Sauerland. “Risks and Benefits of Large Language Models for the Environment.” *Environmental Science & Technology* 57, no. 9 (2023): 3464–66.

“RLHF+CHATGPT: What You Must Know - YouTube.” Accessed April 12, 2023. https://www.youtube.com/watch?v=PBH2nImUM5c.

Sam Altman: OpenAI CEO on GPT-4, ChatGPT, and the Future of AI | Lex Fridman Podcast #367, 2023. https://www.youtube.com/watch?v=L_Guz73e6fw.

Shi, Weiyan, Emily Dinan, Adi Renduchintala, Daniel Fried, Athul Paul Jacob, Zhou Yu, and Mike Lewis. “AutoReply: Detecting Nonsense in Dialogue Introspectively with Discriminative Replies.” *ArXiv Preprint ArXiv:2211.12615*, 2022.

Stanford Seminar - Emerging Risks and Opportunities from Large Language Models, Tatsu Hashimoto, 2022. https://www.youtube.com/watch?v=p6_X5Ei9C9s.

“The Dangers Of Not Aligning Artificial Intelligence With Human Values.” Accessed April 12, 2023. https://www.forbes.com/sites/bernardmarr/2022/04/01/the-dangers-of-not-aligning-artificial-intelligence-with-human-values/.

Twitter. “Sully on Twitter,” April 9, 2023. https://twitter.com/SullyOmarr/status/1645205292756418562.

Uzair Javaid, Betterdata - Applying Generative AI to Create Privacy-Preserving Synthetic Data (S3E9), 2023. https://www.youtube.com/watch?v=Funf37hp84o.

Wang, Qiaosi, Michael Adam Madaio, Shivani Kapania, Shaun Kane, Michael Terry, and Lauren Wilcox. “Designing Responsible AI: Adaptations of UX Practice to Meet Responsible AI Challenges,” 2023. https://storage.googleapis.com/pub-tools-public-publication-data/pdf/4045f0e0e61d89b6b1eacad4b861e86631d5e660.pdf.

We Live in the Infosphere (Prof. Luciano Floridi), 2023. https://www.youtube.com/watch?v=YLNGvvgq3eg.

“Web Shield.” Accessed April 13, 2023.

FAIRLY

https://businesshelp.avast.com/Content/Products/AfB_Antivirus/ConfiguringSettings/WebShield.htm.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. "Chain of Thought Prompting Elicits Reasoning in Large Language Models." *ArXiv Preprint ArXiv:2201.11903*, 2022.

Weidinger, Laura, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, and Atoosa Kasirzadeh. "Ethical and Social Risks of Harm from Language Models." *ArXiv Preprint ArXiv:2112.04359*, 2021. https://arxiv.org/abs/2112.04359.

Weidinger, Laura, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, and Atoosa Kasirzadeh. "Taxonomy of Risks Posed by Language Models," 214–29, 2022.

Weisz, Justin D, Michael Muller, Jessica He, and Stephanie Houde. "Toward General Design Principles for Generative AI Applications." *ArXiv Preprint ArXiv:2301.05578*, 2023. https://arxiv.org/pdf/2301.05578.

Winter, Bodo, Martin H Fischer, Christoph Scheepers, and Andriy Myachykov. "More Is Better: English Language Statistics Are Biased Toward Addition." *Cognitive Science* 47, no. 4 (2023): e13254.

Xiang, Chloe. "'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says." *Vice* (blog), March 30, 2023. https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says.

Zhang, Zhengyan, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. "Moefication: Transformer Feed-Forward Layers Are Mixtures of Experts." *ArXiv Preprint ArXiv:2110.01786*, 2021.

Zhuo, Terry Yue, Yujin Huang, Chunyang Chen, and Zhenchang Xing. "Exploring Ai Ethics of Chatgpt: A Diagnostic Analysis." *ArXiv Preprint ArXiv:2301.12867*, 2023.

FAIRLY