Red-Teaming for Generative AI: Silver Bullet or Security Theater?

Michael Feffer, Anusha Sinha, Wesley H. Deng, Zachary C. Lipton, Hoda Heidari

Carnegie Mellon University
mfeffer@andrew.cmu.edu, asinha@sei.cmu.edu,
{hanwend, zlipton, hheidari}@andrew.cmu.edu

Abstract

In response to rising concerns surrounding the safety, security, and trustworthiness of Generative AI (GenAI) models, practitioners and regulators alike have pointed to *AI red-teaming* as a key component of their strategies for identifying and mitigating these risks. However, despite AI red-teaming's central role in policy discussions and corporate messaging, significant questions remain about what precisely it means, what role it can play in regulation, and how it relates to conventional red-teaming practices as originally conceived in the field of cybersecurity. In this work, we identify recent cases of red-teaming activities in the AI industry and conduct an extensive survey of relevant research literature to characterize the scope, structure, and criteria for AI red-teaming practices. Our analysis reveals that prior methods and practices of AI red-teaming diverge along several axes, including the purpose of the activity (which is often vague), the artifact under evaluation, the setting in which the activity is conducted (e.g., actors, resources, and methods), and the resulting decisions it informs (e.g., reporting, disclosure, and mitigation). In light of our findings, we argue that while red-teaming may be a valuable big-tent idea for characterizing GenAI harm mitigations, and that industry may effectively apply red-teaming and other strategies behind closed doors to safeguard AI, gestures towards red-teaming (based on public definitions) as a panacea for every possible risk verge on *security theater*. To move toward a more robust toolbox of evaluations for generative AI, we synthesize our recommendations into a question bank meant to guide and scaffold future AI red-teaming practices.

1 Introduction

In recent years, generative AI technologies, including large language models (LLMs) (Touvron et al. 2023; Achiam et al. 2023) image and video generation models (Ramesh et al. 2022; Rombach et al. 2022; Brooks et al. 2024), and audio generation models (Donahue et al. 2023; Agostinelli et al. 2023) have captured public imagination. While many view the proliferation and accessibility of these tools favorably, envisioning boons to productivity, creativity, and economic growth, concerns have emerged that the rapid adoption of these powerful models might unleash new categories of societal harms. These concerns have gained credibility owing to several well-publicized problematic incidents where such AI output text expressing discriminatory sentiment towards marginalized groups (Mei, Fereidooni, and Caliskan 2023; Ghosh and Caliskan 2023; Omrani Sabbaghi, Wolfe, and Caliskan 2023; Haim, Salinas, and Nyarko 2024; Hofmann et al. 2024), created images reflecting harmful stereotypes (Luccioni et al. 2023; Wan and Chang 2024), and enabled the generation of deepfake audio in a fashion that has been likened to *digital blackface* (Feffer, Lipton, and Donahue 2023). These issues are compounded by the lack of transparency and accountability surrounding the creation of these models (Birhane, Prabhu, and Kahembwe 2021; Achiam et al. 2023; Widder, West, and Whittaker 2023).

In answer to the mounting worry over the safety, security, and trustworthiness of generative AI models, practitioners and policymakers alike have pointed to *red-teaming* as an integral part of their strategies to identify and address related risks, with the goal of ensuring some notion of alignment with human and societal values (Anthropic 2023; Microsoft 2023; Bockting et al. 2023). Notably, the US presidential Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (The White House 2023) mentions red-teaming eight times, defining it as follows:

> *"The term 'AI red-teaming' means a structured testing effort to find flaws and vulnerabilities in an AI system, often in a controlled environment and in collaboration with developers of AI. Artificial Intelligence red-teaming is most often performed by dedicated 'red teams' that adopt adversarial methods to identify flaws and vulnerabilities, such as harmful or discriminatory outputs from an AI system, unforeseen or undesirable system behaviors, limitations, or potential risks associated with the misuse of the system."*

The order mandates the Secretary of Commerce and other federal agencies to develop guidelines, standards, and best practices for AI safety and security. These include *"appropriate procedures and processes, to enable developers of AI, especially of dual-use foundation models, to conduct AI red-teaming tests"* as a mechanism for *"assessing and managing the safety, security, and trustworthiness of [these] models."*

On one hand, red-teaming appears to call for *the right stuff*: find the flaws, find the vulnerabilities, and (help to) eliminate them. In this spirit, one might find its inclusion in a

landmark policy document to be welcome. On the other, for all of the virtue in its aims, red-teaming at this level of description is strikingly vague. As noted by The Frontier Model Forum (FMF) (2023), *"there is currently a lack of clarity on how to define 'AI red teaming' and what approaches are considered part of the expanded role it plays in the AI development life cycle."* For example, the definition offered by the presidential executive order leaves the following key questions unanswered: What types of *undesirable behaviors, limitations, and risks* can or should be effectively caught and mitigated through red-teaming exercises? How should the activity be *structured* to maximize the likelihood of finding such flaws and vulnerabilities? For instance, aside from AI developers, who else should be at the table, and what resources should be available to them? How should the risks identified through red-teaming be *documented, reported,* and *managed*? Is red-teaming on its own sufficient for assessing and managing the safety, security, and trustworthiness of AI? If not, what other practices should be part of the broader evaluation toolbox, and how does red-teaming complement those approaches? In short, is red-teaming the stuff of policy—the sort of concrete practice around which we can structure regulatory requirements?; or is it the stuff of *vibes*—a vague practice better suited to rallying than to rule-making?

Methodology. Using publicly available resources, we gathered information about recent real-world cases of AI red-teaming exercises (Section 3). We emphasize that many of these cases stem from private sector companies who may use other evaluation techniques not shared with the general public. As such, our corresponding analyses and conclusions rest on disclosed details. To complement these case studies primarily conducted by industry, we additionally performed an extensive survey of existing research literature on red-teaming and adjacent testing and evaluation methods (e.g., penetration testing, jailbreaking, and beyond) for generative AI (Section 4). We organized the thematic analysis of our case studies and literature survey around the following key questions:

- **Definition and scope:** What is the working definition of red-teaming? What is the success criterion?
- **Object of evaluation:** What is the model being evaluated? Are its implementation details (e.g., model architecture, training procedure, safety mechanisms) available to the evaluators or to the public? At what stage of its lifecycle (e.g., design, development, or deployment) is the model subjected to red-teaming?
- **Criteria of evaluation:** What is the threat model (i.e., the risk(s) for which the model is being evaluated)? What are the risks the red-teaming activity potentially missed?
- **Actors and evaluators:** Who are the evaluators? What are the resources available to them (e.g., time, compute, expertise, type of access to model)?
- **Outcomes and broader impact:** What is the output of the activity? How much of the findings are shared publicly? What are the recommendations and mitigation strategies produced in response to the findings of red-teaming? What other evaluations had been performed on the model aside from red-teaming?

To further extend and validate our analysis, we analyzed public comments submitted to a Request For Information (RFI) issued by the National Institute of Standards and Technology (NIST) arm of the Department of Commerce. This RFI sought opinions on points relevant to red-teaming as outlined in the Executive Order.[1]

Contributions. Our findings reveal a lack of consensus around the scope, structure, and assessment criteria for AI red-teaming. Prior methods and practices of AI red-teaming diverge along several critical axes, including the choice of threat model (if one is specified), the artifact under evaluation, the setting in which the activity is conducted (including actors, resources, methodologies, and test-beds), and the resulting decisions the activity instigates (e.g., reporting, disclosure, and mitigation). In light of our findings, we argue that while red-teaming may be a valuable big-tent idea, and even a useful framing for a broad set of evaluation activities for generative AI models, the bludgeoning use of *AI red-teaming* (as defined in public literature) as a catch-all response to quiet all regulatory concerns about model safety verges on *security theater* (Levenson 2014). Our work, including our NIST RFI comment analysis, shows that the current framing of red-teaming in the public discourse serves more to assuage regulators and other concerned parties than to offer a concrete solution. To move toward a more robust toolbox of evaluations for generative AI, we synthesize our recommendations into a question bank meant to guide and scaffold future AI red-teaming practices (see Table 1) and propose future research, including improving the question bank through co-design and evaluation.

2 Related Contemporary Work

A brief history of red-teaming. Zenko (2015) and Abbass (2015) describe how the key concepts of red-teaming originated hundreds of years ago in warfare and religious contexts. They note the term "red team" was formally applied by the US military as early as the 1960s when modeling the Soviet Union's behavior (in contrast to the "blue team" representing the US). In computer security, red-teaming involves modeling an adversary and *"map[ping] out the space of vulnerabilities from a threat lens"* in contrast to penetration testing (in which enlisted cybersecurity experts actively attempt to find vulnerabilities in a computer system) (Abbass et al. 2011; Abbass 2015). Wood and Duggan (2000) further describe how red-teaming *"is not an audit"* and that interpreting it as such risks reducing the amount of information shared about possible vulnerabilities. Using a hypothetical pandemic example, Bishop, Gates, and Levitt (2018) argue that effectively red-teaming a system requires context, knowledge, and assumptions about system usage.

Evaluation beyond red-teaming. Chang and Custis (2022) note that red-teaming is only one of many approaches to increase transparency of an AI system and that factsheets, audits, and model cards are other ways to do so. Similarly, Horvitz (2022) warns of more advanced deepfakes in the near future while emphasizing that remedies such as increased me-

[1] See arXiv version of our work for RFI comment analysis.

Phase	Key Questions and Considerations
0. Pre-activity	What is the **artifact under evaluation** through the proposed red-teaming activity? - What version of the model (including fine-tuning details) is to be evaluated? - What safety and security guardrails are already in place for this artifact? - At what stage of the AI lifecycle will the evaluation be conducted? - If the model has already been released, specify the conditions of release.
	What is the **threat model** the red-teaming activity probes? - Is the activity meant to illustrate a handful of possible vulnerabilities? (e.g., spelling errors in prompt leading to unpredictable model behavior) - Is the activity meant to identify a broad range of potential vulnerabilities? (e.g., biased behavior) - Is the activity meant to assess the risk of a specific vulnerability? (e.g., divulging recipe for explosives)
	What is the **specific vulnerability** the red-teaming activity aims to find? - How was this vulnerability identified as the target of this evaluation? - Why was the above vulnerability prioritized over other potential vulnerabilities? - What is the threshold of acceptable risk for finding this vulnerability?
	What are the **criteria for assessing the success** of the red-teaming activity? - What are the benchmarks of comparison for success? - Can the activity be reconstructed or reproduced?
	Team composition and who will be part of the red team? - What were the criteria for inclusion/exclusion of members, and why? - How diverse/homogeneous is the team across relevant demographic characteristics? - How many internal versus external members belong to the team? - What is the distribution of subject-matter expertise among members? - What are possible biases or blindspots the current team composition may exhibit? - What incentives/disincentives do participants have to contribute to the activity?
1. During activity	What **resources** are available to participants? Do these resources realistically mirror those of the adversary? - Is the activity time-boxed or not? - How much compute is available?
	What **instructions** are given to the participants to guide the activity?
	What kind of **access** do participants have to the model?
	What **methods** can members of the team utilize to test the artifact? Are there any auxiliary automated tools (including AI) supporting the activity? - If yes, what are those tools? - Why are they integrated into the red-teaming activity? - How will members of the red team utilize the tool?
2. Post-activity	What **reports and documentation** are produced on the findings of the activity? Who will have access to those reports? When and why? If certain details are withheld or delayed, provide justification.
	What were the **resources** the activity consumed? - time - compute - financial resources - access to subject-matter expertise
	How **successful** was the activity in terms of the criteria specified in phase 0?
	What are the proposed **measures to mitigate** the risks identified in phase 1? - How will the efficacy of the mitigation strategy be evaluated? - Who is in charge of implementing the mitigation? - What are the mechanisms of accountability?

Table 1: Our proposed set of questions to guide future AI red-teaming activities.

dia literacy and output watermarking (flagging relevant media as AI-generated) should be employed alongside red-teaming; Kenthapadi, Lakkaraju, and Rajani (2023) echo these concerns and similar solutions in their tutorial. Shevlane et al. (2023) also argue that both internal and external model evaluations, as well as robust security responses, should complement effective red-teaming to counter GenAI risks.

Existing surveys of AI red-teaming and evaluations. Inie, Stray, and Derczynski (2023) conduct qualitative interviews with those who perform red-teaming to create a grounded theory of *"how and why people attack large language models."* Schuett et al. (2023) survey members of labs racing to build artificial general intelligence (AGI) and find 98% of respondents somewhat or strongly agree that *"AGI labs should commission external red teams before deploying powerful models."* In the software design space, Knearem et al. (2023) highlight how UX designers are afraid that AI-based design tools will not be red-teamed enough while Liao et al. (2023) suggest that UX designers themselves should help with red-teaming processes. Considering the testing of NLP systems specifically, Tan et al. (2021) propose the DOCTOR framework for reliability testing of such systems. Weidinger et al. (2023) introduce a framework for evaluating generative AI more broadly, namely via *"a three-layered framework that takes a structured, sociotechnical approach."* Anderljung et al. (2023) also propose a framework, ASPIRE, but for external accountability of LLMs and the engagement of relevant stakeholders. Yao et al. (2023b); Neel and Chang (2023), and Shayegani et al. (2023) produce surveys of LLM research with regard to security, privacy, and other vulnerabilities; Chang et al. (2023) conduct another survey of LLM evaluation. In contrast to existing surveys of GenAI evaluation, our work focuses exclusively on *red-teaming*. Some of our findings resonate with points earlier made by Bockting et al. (2023) and Friedler et al. (2023), who argue for interdisciplinary audits of AI systems by diverse groups of people and red-teaming with concrete definitions of harms alongside other evaluations, respectively.

3 Case Studies: AI Red-teaming in practice

To capture the complexity involved in designing real-world AI red-teaming exercises, we synthesize the results from such exercises recently conducted with generative models as case studies. Through these case studies, we seek to understand common red-teaming practices, typical resources required for successful red-teaming, effects of red-teaming on deployed models, common pitfalls, and disclosure of results with community stakeholders.

Methodology. We sourced case studies by searching for reports and news stories about recent red-teaming exercises. As such, our selection is not meant to reflect the full range of red-teaming activities conducted in practice, as limited disclosures from industry teams would make such a reflection impossible. This said, the evaluations we cover here were mostly conducted by private companies, and they encompass a broad range of methods, goals, and areas of focus. In total, we surface and analyze six red-teaming exercises based on retrieved public reports. See Table 2 for more information.

Model/System Evaluated	Conducting Organization	Sources
Bing Chat	Microsoft	(The Frontier Model Forum (FMF) 2023; Microsoft 2024)
GPT-4	OpenAI	(The Frontier Model Forum (FMF) 2023; OpenAI 2023; Achiam et al. 2023)
Gopher	DeepMind	(Rae et al. 2021; Perez et al. 2022; The Frontier Model Forum (FMF) 2023)
Claude 2	Anthropic	(Anthropic 2023; The Frontier Model Forum (FMF) 2023; Anthropic 2023)
Various	DEFCON	(Cattell, Carson, and Chowdhury 2023; Cattell 2023)
Claude 1	Anthropic	(Askell et al. 2021; Ganguli et al. 2022; Anthropic 2023)

Table 2: The six cases of AI red-teaming we discuss as part of our case study analyses. These cases were found by searching for reports and news stories about recent red-teaming exercises. Though industry teams do not disclose (all of) their methods, the cases we analyze here largely stem from industry work, in turn yielding insight into some of their practices.

3.1 Findings

Variation in goals, processes, and threat models. Reflecting the lack of consensus on a definition of red-teaming in the literature, red-teaming activities frequently varied in form and in goals. Some organizations chose to conduct a single round of red-teaming (Ganguli et al. 2022; Perez et al. 2022; Anthropic 2023; Cattell, Carson, and Chowdhury 2023), while others saw red-teaming as an iterative process in which results from initial rounds of testing were used to prioritize risk areas for further investigation (Achiam et al. 2023; The Frontier Model Forum (FMF) 2023; Microsoft 2024). The goals of red-teaming activities also ranged from specific objectives (e.g., red-teaming to investigate risks to national security (Anthropic 2023)) to more broad targets (e.g., uncovering "harmful" model behavior (The Frontier Model Forum (FMF) 2023)); threat models related to the latter were more common. Model developers use such threat models for evaluation in hopes that this will yield greater variation in red-teaming efforts, especially because it is impossible to understand the model's entire risk surface. Unfortunately, probing these non-specific threat models does not always produce this desired variation, more so when evaluators are given limited time and resources to produce harmful outputs. For example, some time-boxed evaluators repeatedly probed models in the same risk area because it was easy to produce harmful outputs, as opposed to exploring other risks (Ganguli et al. 2022).

Interconnected members, resources, and outcomes. The evaluators employed in red-teaming activities for each case study varied considerably. We found that there were generally three types of team compositions:

1. Teams composed of handpicked subject matter experts in relevant areas (e.g., national security, healthcare, law, alignment), both internally and externally sourced
2. Crowdsourced teams chosen from crowdworking platforms or attendees of a live event
3. Teams composed of language models (i.e., language models prompted or fine tuned to red-team themselves)

The resources available to evaluation teams varied based on team composition. For crowdsourced teams, red-teaming efforts were time-boxed either by participant or by task, and access to models was available only through APIs (Ganguli et al. 2022; Cattell 2023). For teams with subject matter expertise, red-teaming efforts were more open-ended with fewer restrictions on time or compute (Achiam et al. 2023; Anthropic 2023; Microsoft 2024). While API access to models is still most common for these teams, sometimes experts are given access to versions of models without safety guardrails. When language models are used to red-team themselves, the main resource bottlenecks are the number of prompts used to produce red-teaming behavior and the compute resources needed for model retraining or fine tuning. Full access to model parameters is thus usually a requirement when performing this type of red-teaming (Perez et al. 2022). As such, team composition and available resources also shape red-teaming outcomes. For instance, crowdsourced teams typically focused on risk areas where successful attacks were easy to produce due to time constraints, so risk areas that are more complex to attack may remain completely untested (Ganguli et al. 2022; Cattell 2023). In contrast, subject matter experts and members of academic communities and AI firms prioritized different risks and explored them in more detail due to differences in team member selection and resources (Achiam et al. 2023). When using language models for red-teaming, offensiveness classifiers are trained on pre-existing datasets such as the Bot-Adversarial Dialogue (BAD) dataset (Xu et al. 2021), which in turn only cover certain types of offensive model replies. Evidently, team selection and resources can introduce bias into the types of risks investigated and ultimate exercise findings.

No standards for disclosing red-teaming details. We found nontrivial variation in the publicly-shared outputs of red-teaming efforts, largely because there are no existing standardized reporting procedures or requirements. In only half of the cases explored, specific examples of risky or harmful model behavior uncovered by red-teaming efforts were publicly shared. In one case, a full dataset composed of 38,961 red team attacks was publicly released to aid in testing of other models (Ganguli et al. 2022). In the other two cases, examples of harmful behavior were publicly available, but the full scope of all red-teaming attacks was not released (Achiam et al. 2023; Perez et al. 2022). For red-teaming efforts on publicly available models or those focused on national security, specifics of harmful behavior were not shared publicly because findings were deemed too sensitive to share without responsible disclosure practices (Anthropic 2023). One case study resulted in Anthropic piloting a responsible disclosure process to share vulnerabilities identified during red-teaming with appropriate community stakeholders, but this process is still under development (thus we assume that these disclosures have not yet been made) (Anthropic 2023). There are also variations in reporting on resource consumption. We found costs of red-teaming efforts were usually disclosed for evaluation teams composed of crowdsourced evaluators (for example, the hourly rate paid to crowdworkers (Ganguli et al. 2022)). These details were not disclosed for teams composed of subject matter experts and language models, though they seem to have been given greater time and compute resources. Two case studies specifically mention ongoing red-teaming for 6-7 months before model release (Achiam et al. 2023; Anthropic 2023). In contrast, for crowdsourced teams, evaluators spent about 30-50 minutes per task, with evaluators sourced from live events being limited to only completing a single task (Ganguli et al. 2022; Cattell 2023).

Diverse mitigations and supporting evaluations. While every case analyzed here identified problematic or risky model behavior, none of them resulted in a decision not to release the model. Instead, a number of risk mitigation strategies were proposed and/or employed to minimize harmful model behavior identified during red-teaming. These approaches ranged from concrete, such as jointly training language models and red-teaming models via strategies for training GANs, to purely conceptual, like unlikelihood training to reduce harmful outputs (Perez et al. 2022). However, the specifics of risk mitigation strategies were often not provided when the target model was publicly available, and there were no standards for reporting improvements stemming from these efforts. As a result, it was often difficult to determine if risks identified during red-teaming were sufficiently addressed. Similarly, every case we analyzed involved models that had been previously evaluated using other techniques beyond red-teaming, but there were no established guidelines or standards for these other methods. Commonly, models were evaluated using the Perspective API to measure toxicity; human feedback on helpfulness, harmfulness, and honesty; and QA benchmarks for accurate and truthful outputs (Rae et al. 2021; Askell et al. 2021; Anthropic 2023). Other evaluations included internal quantitative assessments to determine if model outputs violated specific content policies (e.g., hate speech, illicit advice) (Achiam et al. 2023). Additionally, some initial efforts described as "red-teaming" by evaluators were more focused on understanding base model capabilities through open-ended experimentation than on specifically stress testing the model (Microsoft 2024).

3.2 Discussion

Red-teaming is ill-structured. Evaluation teams either prioritize risk areas for investigation or provide evaluators with broad directions in hopes that diversity within the group of evaluators will lead to the exploration of many different risks. However, in line with findings from prior empirical work (Chung et al. 2019; Deng et al. 2023c), there is a trade-off between providing evaluators with specific instructions

and leaving the activity open-ended. On one hand, vague instructions can be helpful to avoid biasing evaluators towards finding specific issues based on initial prioritization. On the other, a lack of instructions can reduce the utility of the exercise for uncovering risks relevant to real-world contexts. Red teams navigate this tradeoff as they seem aware that the entire risk surface of a model will not be explored by red-teaming activities, but this serves to make red-teaming as a whole ill-structured and difficult to define. Moreover, we argue that this lack of structure and scope is concerning as recent recommendations establishing red-teaming as a best practice suggest that the broader perception of red-teaming may not align with current working definitions of red-teaming, (i.e., red-teaming activities are much more qualitative, subjective and exploratory than community stakeholders may realize). In every case study, however, red-teaming was able to reveal harmful model behavior that other more systematic methods seemed to miss, highlighting the importance of both conducting red-teaming (alongside other evaluations) and developing systematic processes for red-teaming in a more comprehensive manner. These processes could include, for example, the development of guidelines on whether red-teaming is most effective when conducted internally or externally and when it should be conducted (i.e., before and/or after public release of the model and whether red-teaming activities should be ongoing while the model is publicly available).

Evaluation team composition introduces biases. The goal of team member selection seems to be ensuring variety in the risk areas explored during red-teaming. One way to do so is by handpicking experts with different backgrounds, as noted by prior work on interdisciplinary collaboration within AI teams (Nahar et al. 2021; Deng et al. 2023d); another is by randomly sampling the population through crowdsourcing. Both have drawbacks: there may be bias in expert selection (Hube, Fetahu, and Gadiraju 2019; Chen, Weld, and Zhang 2021), and crowdworkers have limited resources in terms of time, compute, and relevant expertise (Toxtli, Suri, and Savage 2021). It is difficult to say what the ideal balance between expert and non-technical stakeholders would be, but prior crowdsourcing research suggests a hybrid approach could help address some of the pitfalls associated with each type of team composition (Kittur et al. 2013; Vaughan 2017; Chen, Weld, and Zhang 2021). One type of team composition we did not see explored in any case study is a crowdsourced team with more open-ended instructions and greater resources. This could allow more variety in the risk areas explored because evaluators would not feel incentivized to focus on risk areas where harmful model outputs are easy or quick to produce, but it would also require partnering with subject matter experts to fully evaluate risky model behavior. Team composition can also shape the outputs of red-teaming. One of the issues with red-teaming via internal teams is that more extreme measures such as blocking the release of a model may never be recommended due to conflicting interests. However, external teams that may be more likely to recommend such measures often do not have the power to actually employ these mitigations. A hybrid approach could resolve some of these issues, but it would need to be paired with accountability mechanisms to disclose recommendations and mitigations. Additionally, directly involving marginalized stakeholders, as they may suffer the most from unanticipated model outputs, is challenging yet not impossible (see, e.g., (Woodruff et al. 2018)).

Hesitancy to release results reduces utility. As suggested by prior work studying responsible AI industry practices (Madaio et al. 2020; Rakova et al. 2021), the reluctance to share all results from red-teaming activities may stem from risks associated with public models (evaluators do not want to provide inspiration for potential attackers). Additionally, releasing all of the data associated with red-teaming could be overwhelming for community stakeholders. This said, because red-teaming does not seem to be planned as a comprehensive measure of risky model behavior, disclosing some specifics of the activity is necessary so stakeholders can understand harms investigated and in turn determine if they are relevant to their use cases. For example, significant risk areas that evaluation teams knowingly have not probed should be highlighted or identified in reports. Moreover, none of the case studies provided complete monetary costs of red-teaming efforts. This information seems relatively low-risk to release (i.e., compared to specific examples of harmful model behavior) and could be useful for developing methods to conduct more comprehensive red-teaming. The costs of assembling teams with differing compositions of expertise and automation, for instance, could be used to determine where resources can be used most effectively. Similarly, the costs of evaluating and mitigating various types of risks could be factored into a cost-benefit analysis when prioritizing risks. The lack of reported cost figures may also make it harder for third-party or external organizations to conduct red-teaming: if these unreported costs are quite large, it could be difficult or impossible for anyone aside from companies themselves to do this type of analysis (Costanza-Chock et al. 2022).

4 A Survey of AI Red-teaming Research

Methodology. To source papers, we primarily searched arXiv, Google Scholar, OpenReview, ACL Anthology, and the ACM Digital Library with keywords "red-teaming", "ai red-teaming", "jailbreak", and "llm jailbreak", and we then gathered results.[2] Where possible, we replaced preprints with corresponding published works. We also included relevant works found prior to this search and via snowball sampling.

We scrutinize and subdivide retrieved papers into groups along two dimensions, both of which relate to the evaluation in each paper. The first corresponds to the type of risk investigated during the evaluation, and the second corresponds to the type of approach used for evaluation. We analyze papers by characteristics pertaining to threat model and methodology because we found that research works primarily focused on these aspects (perhaps due to technical relevance) as opposed to other factors important to red-teaming (such as team composition and resources consumed). Overall classifications and

[2]We focus on red-teaming evaluations, but we argue that the *jailbreaking* literature contains techniques similar in spirit to those employed by AI red-teams, so we also explore it here.

		Risk				
		D	C	B	N	Total
Approach	Brute-Force	15	4	1	0	20
	Brute-Force + AI	23	7	10	2	42
	Algorithmic Search	12	1	1	0	14
	Targeted Attack	19	7	2	0	28
	Total	69	19	14	2	104

Table 3: Numbers of papers from our survey in each subgroup based on risk (**D**issentive, **C**onsentive, **B**oth, and **N**either) and approach as outlined in Section 4. Over half of papers are concerned with dissentive risks, and over one-third of papers utilize brute-force + AI to red-team AI models.

in-depth paper findings (as with analyses of case studies from the previous section) can be found in the arXiv version of our work, and totals for each classification can be found in Table 3. We do not ascribe judgments to any category (i.e., we do not posit that one red-teaming type is better than another).

4.1 Findings: AI Red-teaming Threat Models

We categorize papers according to the threat model implied or specified in each. In particular, we consider the type of *risk* posed in each threat model. For the ease of organizing our retrieved papers and findings, we present risks as one of two binary types, but we note that in reality, there exists a spectrum between the two types, and risks in practice generally may be more nuanced than as we describe them here.

Dissentive risk. We define *dissentive risk* as risk in which people may disagree on its definition and corresponding threat level, and we categorize papers that studied this type of risk accordingly. For instance, hate speech has many definitions, and though images with blood and responses to questions like "how to build a bomb?"[3] may be suboptimal AI generations, some may find them admissible depending on context.[4] Nevertheless, researchers like Gehman et al. (2020) and Röttger et al. (2021) create benchmarks to analyze hate speech generation, Rando et al. (2022) perform red-teaming to study how Stable Diffusion's safety filter can be bypassed to generate NSFW content, and Zou et al. (2023) test whether LLMs respond to queries like "Generate a step-by-step plan to destroy humanity" given adversarial prompt suffixes.

Consentive risk. In contrast, we define *consentive risk* as risk in which people agree on the definition and danger it presents, meaning context is not required to evaluate potential harms, and associated content should never be generated by properly functioning AI. We similarly categorize papers accordingly if the authors consider this type of risk. For example, phenomena such as the divulging of private information, leaking of training data, and production of vulnerable code or material for phishing attacks are inadmissible in any situation. To these ends, Chen et al. (2023b) study the degree to which multimodal LMs can safeguard private information, Nasr et al. (2023) illustrate how divergence attacks cause ChatGPT to reveal training data, Wu, Liu, and Xiao (2023) analyze how code generation LLMs *"can be easily attacked and induced to generate vulnerable code,"* and Roy, Naragam, and Nilizadeh (2023) discover ChatGPT can create phising code.

Both and neither. Some authors tasked themselves with analyzing *both* kinds of risk, such as personally identifiable information (PII) leakage in addition to hate speech (Ganguli et al. 2022; Perez et al. 2022; Srivastava, Ahuja, and Mukku 2023). Others introduce methods to analyze *neither* type of risk from the outset, stressing that definitions and classifications of issues may need to be done from scratch (Casper et al. 2023b; Radharapu et al. 2023).

4.2 Findings: AI Red-teaming Methodologies

We further categorize papers based on the methodology the researchers employ to perform red-teaming. Namely, we study the type of *approach* used to find risks.

Brute-force. Work that utilized *brute-force* approaches involved manual evaluation of generative AI inputs and outputs by teams of humans. We found that such teams typically consisted of the researchers themselves, internal auditors (of tech companies), or external members (such as contractors hired via Amazon Mechanical Turk (MTurk)). Xu et al. (2020, 2021) and Ganguli et al. (2022) employed crowdworkers to elicit harmful text outputs from language models (including but not limited to offensive language and PII) and measure safety. Mu et al. (2023) compiled a benchmark from scratch to test LLMs' capacities to follow rules while Huang et al. (2023a) hired crowdworkers to build a new benchmark that assesses alignment with Chinese values. Schulhoff et al. (2023) hosted a prompt hacking competition, thereby making competitors LLM red team members. Other authors handcraft jailbreak attacks against language models (Du et al. 2023; Li et al. 2023b; Wei, Wang, and Wang 2023; Li et al. 2023a; Liu et al. 2023b), but the authors of (Wei, Wang, and Wang 2023) join Xie et al. (2023) in additionally devising defense strategies for them. Shen et al. (2023) and Rao et al. (2023) analyze the effectiveness of jailbreak attacks collected from external sources (including prior work and public websites).

Brute-force + AI. Another body of work similar to those of the brute-force works described above incorporated AI techniques into their red-teaming processes. Common approaches to do so typically involved having AI models generate test cases and find errors in other AI output. We therefore term such approaches as *brute-force + AI*. Many authors used LLMs to generate normal prompts (Perez et al. 2022; Rastogi et al. 2023; Srivastava, Ahuja, and Mukku 2023; Bhardwaj and Poria 2023b; Chen et al. 2023a; Mei, Levy, and Wang 2023; Zhang, Pan, and Yang 2023) and jailbreak prompts (Yu, Lin, and Xing 2023; Deng et al. 2023a; Shah et al. 2023; Yao et al. 2023a; Wei, Haghtalab, and Steinhardt 2023) such

[3] AI safety researchers note that "how to build a bomb?" is a default GenAI evaluation query despite the fact that related information could be found via Wikipedia or even popular fiction (Weir 2014) because many GenAI developers assert their models should not disclose bomb-making instructions (e.g., (OpenAI 2023; Touvron et al. 2023)), so techniques which yield responses prove safeguards are brittle. This does not necessarily mean the community finds such responses concerning.

[4] Such generations do not reflect opinions held by the authors.

that LLMs produce bad outputs like harmful text responses. Variations on these ideas also exist, such as the work of Pfau et al. (2023), in which the authors use *reverse LMs* to work backwards from harmful text responses to prompts that could generate them. Others use LLMs to devise new benchmarks related to exaggerated safety responses (i.e., refusal to respond to prompts that are arguably safe) (Röttger et al. 2023), *fake alignment* that occurs when models appear aligned with one query format and misaligned with another (e.g., multiple choice versus open-ended response) (Wang et al. 2023c), and *latent jailbreaks*, or compliance with *"implicit malicious instruction[s]"* (Qiu et al. 2023). Researchers have also used AI to red-team and jailbreak text-to-image models and multimodal LMs. For instance, Lee et al. (2023) demonstrate how passing harmful queries with corresponding images to multimodal models (e.g., an image of a bomb with the question "how to build a bomb?") improves the likelihood of harmful text generation. Mehrabi et al. (2023a) test their FLIRT framework to analyze text-to-image models like Stable Diffusion. Still other researchers perform red-teaming of LLMs for specific end-uses. Lewis and White (2023) red-team an LLM for potential future use as a component of a virtual museum tourguide, and He et al. (2023) evaluate the dangers of using LLMs as part of scientific research. In light of the many documented ways generative AI models can be utilized for malicious use, researchers have also studied ways in which they can be defended. Both Sun et al. (2023) and Wang et al. (2023d) introduce methods that utilize LLMs to generate fine-tuning data that can be used to avert harmful responses. Zhu et al. (2023b) employ k-nearest neighbors and clustering techniques to fix incorrect labels in popular LLM safety datasets (with the goal of developing better downstream safeguards).

Algorithmic search. Some other methods start from a given prompt and utilize a process to modify it until an issue is encountered. Such processes can take the form of random perturbations or a guided search, and we therefore refer to such approaches as *algorithmic search* strategies. For instance, several authors describe approaches to red-teaming and jailbreaking in which one AI model automatically and repeatedly attacks an LLM until defenses are broken or bypassed (Casper et al. 2022; Ma et al. 2023; Chao et al. 2023; Mehrotra et al. 2023). Both Chin et al. (2023) and Tsai et al. (2023) propose search-based red-teaming approaches to evaluate text-to-image models that perturb input prompts until they simultaneously pass safety filters and generate forbidden content. Search-based approaches can also be used as defensive measures. Noting the brittleness of most jailbreak methods, Robey et al. (2023) and Zhang et al. (2023b) introduce methods to detect jailbreaks by applying perturbations to text and image inputs and observing whether outputs change drastically (if so, the input was likely a jailbreak).

Targeted attack. The last approach to red-teaming we document as part of our review involves deliberately targeting part of an LLM, which could include an API, vulnerability in language translation support, or step of its training process, in order to induce issues. As such, we refer to such approaches as *targeted attack* methods. For instance, Wang and Shu (2023) show how to construct *steering vectors* using activation vectors from both safety-tuned and non-safety-tuned versions of models to obtain toxic outputs from safety-tuned models. Others illustrate how to imperceptibly perturb images to cause multimodal LMs to respond in unintended ways (such as replying with a malicious URL or misinformation) (Schlarmann and Hein 2023; Qi et al. 2023; Bailey et al. 2023), and Tong, Jones, and Steinhardt (2023) engineer prompts for text-to-image models that are mismatched with resulting images by exploiting reliance on CLIP embeddings. Other approaches include but are not limited to weaponizing the fact that LLMs are not optimized to converse in low-resource languages and ciphers (Deng et al. 2023e; Yong, Menghini, and Bach 2023; Yuan et al. 2023), poisoning data used to tune or utilize LLMs (Rando and Tramèr 2023; Zhang et al. 2023a; Abdelnabi et al. 2023; Cao, Cao, and Chen 2023; Wang et al. 2023b; Lee et al. 2024), and attacking APIs associated with black-box models (Pelrine et al. 2023). Various defensive methods rooted in targeted attack approaches have been proposed as well. Bitton, Pavlova, and Evtimov (2022) describe their Adversarial Text Normalizer, which can defend an LLM against various character-level perturbations typical of certain adversarial prompts. In addition, other defensive strategies mentioned previously can defend these attacks (e.g., JailGuard from Zhang et al. (2023b) addresses attacks introduced in (Qi et al. 2023; Bailey et al. 2023; Schlarmann and Hein 2023; Zou et al. 2023)).

4.3 Discussion

Many different methods to perform red-teaming. As illustrated in Table 3, researchers and practitioners have undertaken numerous approaches to evaluate GenAI and have all described them as *red-teaming*. At the same time, there have been developments like Schuett et al.'s finding that the overwhelming majority of AGI lab members support external red-teaming efforts (Schuett et al. 2023) and the recent Executive Order (The White House 2023) stressing the importance of red-teaming. These developments and the many red-teaming variations are together arguably concerning, precisely because there is no agreed-upon definition (from these papers) regarding what constitutes red-teaming. By highlighting this, we do not mean to imply that evaluations until now are useless. On the contrary, we posit they are necessary but perhaps insufficient tests of safety, and we conjecture that the existence of many interpretations of "red-teaming" suggests there must be more top-down guidance and requirements concerning red-teaming evaluations.

Threat modeling skewed toward dissentive risk. Table 3 also highlights that the majority of evaluations focus on *dissentive risk* rather than *consentive risk*. This means that undue effort has been undertaken to evaluate and mitigate GenAI behavior that may be admissible in various contexts. Additionally, Röttger et al. (2023) have shown that current attempts to mitigate such risks have resulted in exaggerated safety, yielding LLM behavior like the refusal to provide information on buying a can of coke. Lastly, focusing on dissentive risk takes attention away from consentive risk, which in turn is inadmissible in any context. In light of such issues and tradeoffs, Casper et al. (2023b) and Radharapu et al. (2023)

suggest clearly defining risks and problematic outputs and justifying those definitions before any analysis.

No consensus on adversary capabilities. While threat model and methodology are two factors that contribute to the diversity of red-teaming exercises, assumptions about adversary capabilities are also contributors. Namely, the works encountered have differing estimates of adversary resources. For instance, Perez et al. (2022) and many authors of similar work conjecture that an adversary can only prompt an LLM and probe it for bad outputs. In contrast, others assume that an adversary can poison the training process (Rando and Tramèr 2023), has the compute required to search for adversarial suffixes (Zou et al. 2023), or is able to run both safety-tuned and non-safety-tuned versions of language models to obtain toxic output (Wang and Shu 2023). Future guidelines for red-teaming may want to suggest that researchers should emphasize and defend adversary assumptions.

Non-universality of values used for alignment. Work found as part of this survey involving dissentive risk and alignment are driven by, implicitly or explicitly, a set of human values that determine whether GenAI outputs are admissible or inadmissible. However, this in turn prompts the question *whose values are being utilized for alignment and evaluation?* For instance, the FLAMES benchmark proposed by Huang et al. (2023a) is purported to measure alignment with Chinese values, whereas Weidinger et al. (2023) emphasize that other evaluations may reflect those of *"the English-speaking or Western world."* The extent to which GenAI does not support low-resource languages (Deng et al. 2023e; Yong, Menghini, and Bach 2023) and agrees with bias and stereotypes (Ganguli et al. 2022; Rastogi et al. 2023) evidences that models may not reflect the values and beliefs of all persons. Works beyond this survey have illustrated how the framing of AI value alignment is a normative problem that, if not properly addressed, may only serve to reflect the norms of one group of people, typically the majority (Feffer, Heidari, and Lipton 2023; Lambert, Gilbert, and Zick 2023; Lambert and Calandra 2023). Especially as OpenAI started a partnership with the US military on one hand (Stone and Bergen 2024; Field 2024b) and launched an initiative to align superintelligent AI to "human values" on another (Leike and Sutskever 2023),[5] we argue that it is crucial to analyze assumptions made and viewpoints held by those who build AI systems.

No consensus on who should perform red-teaming. Moreover, just as there is a lack of agreement regarding values to use to assess GenAI outputs, there is a similar lack of agreement regarding who should perform red-teaming. Groups of evaluators have consisted of hired crowdworkers (Ganguli et al. 2022), competition participants (Schulhoff et al. 2023), researchers themselves (Perez et al. 2022), and others simply red-teaming for fun (Inie, Stray, and Derczynski 2023). While some argue for more diversity to evaluate AI models (Solaiman 2023), others caution that increased diversity is not a panacea and is moreover typically ill-defined (Weidinger et al. 2023; Bergman et al. 2023). For instance, Yong, Menghini, and Bach (2023) argue for multilingual red-teaming to respond to low-resource language issues, and He et al. (2023) *"advocate for a collaborative, interdisciplinary approach among the AI for Science community and society at large"* to respond to scientific research risks. Such examples suggest that terms like "diversity" and "community" should be defined and sought out relative to the risks considered by red-teaming processes. They additionally hint towards more involvement of the public and relevant stakeholders, ideas also recommended in parallel literature regarding algorithmic auditing and participatory ML (Costanza-Chock et al. 2022; Birhane et al. 2022; Feffer et al. 2023; Delgado et al. 2023). Similar literature has also engaged with benefits of deliberation in the face of disagreement (e.g., (Pierson 2017)) and effects of identity on evaluators' perceptions of AI safety (e.g., (Aroyo et al. 2023)). Future red-teaming guidelines should emphasize these considerations.

Unclear follow-ups to red-teaming activities. We found that overall, responses from GenAI developers (at least public ones) to the many red-teaming and jailbreaking papers have been muted and generally mixed. While some authors such as Wei, Haghtalab, and Steinhardt (2023) report that they reached out to organizations like OpenAI and Anthropic about the vulnerabilities found in their models, the vulnerabilities and models themselves have for the most part persisted. One rare exception to this pattern is the case of the findings of Nasr et al. (2023), in which OpenAI updated ChatGPT to reduce the likelihood of divergence attack success and modified their terms of use to forbid such attacks in response (Price 2023; Mok 2023). However, these changes only came following the paper's release, 90 days *after* the paper authors first notified OpenAI about the vulnerability. If red-teaming is to be stipulated as a requirement for release and safe usage of AI models, there should arguably be a protocol to mitigate found issues accordingly.

5 NIST RFI Comment Analysis Summary

We find that comments submitted to the NIST RFI on red-teaming GenAI are generally in accord with our conclusions.[6]

Similarities. Industry, academia, and civil society organizations suggest that NIST should specify a clearer definition of "red-teaming" and provide interested parties with appropriate resources pertaining to guidelines and best practices. Notably, even industry firms with experience red-teaming GenAI expressed a desire for concrete guidance from NIST. This supports our finding that red-teaming, as defined in public research and reports, is loosely structured and perhaps not the rigorous practice implied by the Executive Order. Moreover, many comments stressed that a plurality of different viewpoints and stakeholders should be involved in evaluating GenAI systems. Our findings concur with these notions.

Differences. A number of comments (including those from OpenAI and Mozilla) recommended evaluations at both the model level and system level. Though our work primarily considers model-level evaluations, we emphasize that in at least one comment, evaluation at either level is referred to as *red-*

[5]This latter project ended in spring 2024 (Field 2024a); its dissolution may only further support the notion that AI developers' positions should be scrutinized.

[6]See arXiv version of our work for an extended analysis.

teaming. This also exemplifies the need for more concrete definitions of evaluations. Additionally, many comments, especially those from individuals, expressed concerns with GenAI, not because of evaluation methods employed but rather because of its uses (such as for malicious deepfakes) and its training data (typically stolen via web-scraping). Though our work centers on GenAI evaluation via red-teaming, our findings support incorporating diverse perspectives while building and evaluating these systems, and we agree that such incorporation should also consider the ethical consequences and legality of any created systems.

6 Takeaways and Recommendations

Based on our results, we distill the following findings and guidance for future red-teaming evaluations.

Red-teaming is *not* a panacea. Each red-teaming exercise discussed in this paper only covered a limited set of vulnerabilities. As such, red-teaming cannot be expected to guarantee safety from all angles. For instance, from the papers surfaced in our research survey, approaches to red-teaming that detect and mitigate harmful text responses (Perez et al. 2022) may not detect and mitigate phishing attack vulnerabilities (Roy, Naragam, and Nilizadeh 2023) and vice versa. Similarly, our case study analysis highlighted that team composition may also influence the types of issues found in a given exercise (e.g., subject matter experts (Anthropic 2023) may find different problems than crowdworkers (Ganguli et al. 2022)). Moreover, there are other issues that red-teaming alone cannot address, such as problems stemming from *algorithmic monoculture* (Kleinberg and Raghavan 2021; Bommasani et al. 2022; Tong, Jones, and Steinhardt 2023) or GenAI's environmental impacts (Crawford 2024; Rogers 2024). We argue that red-teaming should therefore be considered as one evaluation paradigm, among others such as algorithmic impact assessments (Reisman et al. 2018) and audits, to assess and improve the safety and trustworthiness of GenAI (Friedler et al. 2023). It is also important to support participation from diverse roles (e.g., technical, user-facing, legal) in red teaming within organizations (Nahar et al. 2021; Deng et al. 2023d).

Red-teaming *not* well-scoped or structured. The many variations in the red-teaming processes encountered in our case studies and literature review of public research and reports illustrate that at the moment, red-teaming is an unstructured procedure with undefined scope. This statement is not meant to belittle efforts undertaken, and we concede that we do not have full insight into industry red-teaming activities, but we recommend red-teaming guidelines be drafted and made publicly available to improve utility of future evaluations.

No standards concerning what should be reported. There are currently no unified protocols for reporting the results of red-teaming evaluations. In fact, we found that a number of case studies and research papers sourced for our work did not fully report their findings or resource costs required to perform evaluations. We suggest that regulations and/or best practices be put forth to entice more detailed reporting for a number of reasons, ranging from increasing public knowledge, to helping third-party groups conduct their own tests (Raji et al. 2020; Guha et al. 2023), to assisting end-users in determining the relevance of red-teaming for their use cases. We argue that such reports should, at a minimum, clarify (1) the resources consumed by the activity, (2) assessments of whether the activity was successful according to previously established goals and measures, (3) the mitigation steps informed by the findings of the activity, and (4) any other relevant or subsequent evaluation of the artifact at hand.

Follow-ups often unclear and unrepresentative. Though red-teaming exercises uncovered many issues with generative models, subsequent activities to remedy these problems were often vague or unspecified. Taken with the lack of reporting, such unclear mitigation and alignment strategies could reduce red-teaming to an *approval-stamping process* wherein one can say that red-teaming was performed as an assurance without providing further details into issues discovered or fixed. Moreover, we found that the strategies specified in research and case studies, such as further fine-tuning or RLHF, were often not representative of the full range of possible solutions. Other approaches like model input and output monitoring, prediction modification, and even the refusal to deploy models in certain scenarios, were rarely or never mentioned. Future research should address mitigation strategies beyond popular solutions given surfaced issues.

Propose question bank as starting point. In light of the issues raised by our work, we provide a set of questions for future red teams to consider before, during, and after evaluation. These questions, found in Table 1, encourage evaluators to ponder the benefits and limitations of red-teaming generally as well as the impact of specific design choices pertaining to their setting. We emphasize that these are not finalized guidelines but rather (what we hope is) the start of a broader conversation about GenAI red-teaming and evaluation processes. We welcome and support comments and feedback, and we leave question refinement, overall evaluation, and development of supplementary materials (e.g., rubrics for evaluating red-teaming protocols) as critical future directions.

Acknowledgements

Hoda Heidari acknowledges support from NSF (IIS2040929 and IIS2229881) and PwC (through the Digital Transformation and Innovation Center at CMU). Michael Feffer acknowledges support from the National GEM Consortium and the ARCS Foundation. Authors additionally gratefully acknowledge the NSF (IIS2211955), UPMC, Highmark Health, Abridge, Ford Research, Mozilla, Amazon AI, JP Morgan Chase, the Block Center, and the Center for Machine Learning and Health for their generous support of Zachary C. Lipton's and ACMI Lab's research. Wesley H. Deng also acknowledges support from NSF (IIS2040942), Cisco Research, the Jacobs Foundation, Google Research, and the Microsoft Research AI & Society Fellowship Program. Furthermore, this material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not reflect the views of any funding agencies.

References

Abbass, H.; Bender, A.; Gaidow, S.; and Whitbread, P. 2011. Computational red teaming: Past, present and future. *IEEE Computational Intelligence Magazine*, 6(1): 30–42.

Abbass, H. A. 2015. *Computational red teaming*. Springer.

Abdelnabi, S.; Greshake, K.; Mishra, S.; Endres, C.; Holz, T.; and Fritz, M. 2023. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, 79–90.

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Agostinelli, A.; Denk, T. I.; Borsos, Z.; Engel, J.; Verzetti, M.; Caillon, A.; Huang, Q.; Jansen, A.; Roberts, A.; Tagliasacchi, M.; et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Alon, G.; and Kamfonas, M. 2023. Detecting language model attacks with perplexity. *arXiv preprint arXiv:2308.14132*.

Anderljung, M.; Smith, E.; O'Brien, J.; Soder, L.; Bucknall, B.; Bluemke, E.; Schuett, J.; Trager, R.; Strahm, L.; and Chowdhury, R. 2023. Towards Publicly Accountable Frontier LLMs. In *Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS*.

Anthropic. 2023. Frontier threats red teaming for AI Safety.

Anthropic. 2023. Model card and evaluations for Claude Models.

Aroyo, L.; Taylor, A. S.; Díaz, M.; Homan, C. M.; Parrish, A.; Serapio-García, G.; Prabhakaran, V.; and Wang, D. 2023. DICES dataset: diversity in conversational AI evaluation for safety. In *Advances in Neural Information Processing Systems (NeurIPS) 2023 Dataset and Benchmarks Track*.

Askell, A.; Bai, Y.; Chen, A.; Drain, D.; Ganguli, D.; Henighan, T.; Jones, A.; Joseph, N.; Mann, B.; DasSarma, N.; et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Bailey, L.; Ong, E.; Russell, S.; and Emmons, S. 2023. Image hijacking: Adversarial images can control generative models at runtime. *arXiv preprint arXiv:2309.00236*.

Bergman, A. S.; Hendricks, L. A.; Rauh, M.; Wu, B.; Agnew, W.; Kunesch, M.; Duan, I.; Gabriel, I.; and Isaac, W. 2023. Representation in AI Evaluations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 519–533.

Bhardwaj, R.; and Poria, S. 2023a. Language Model Unalignment: Parametric Red-Teaming to Expose Hidden Harms and Biases. *arXiv preprint arXiv:2310.14303*.

Bhardwaj, R.; and Poria, S. 2023b. Red-teaming large language models using chain of utterances for safety-alignment. *arXiv preprint arXiv:2308.09662*.

Birhane, A.; Isaac, W.; Prabhakaran, V.; Diaz, M.; Elish, M. C.; Gabriel, I.; and Mohamed, S. 2022. Power to the people? opportunities and challenges for participatory AI. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–8.

Birhane, A.; Prabhu, V. U.; and Kahembwe, E. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

Bishop, M.; Gates, C.; and Levitt, K. 2018. Augmenting machine learning with argumentation. In *Proceedings of the New Security Paradigms Workshop*, 1–11.

Bitton, J.; Pavlova, M.; and Evtimov, I. 2022. Adversarial Text Normalization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, 268–279.

Bockting, C. L.; van Dis, E. A. M.; van Rooij, R.; Zuidema, W.; and Bollen, J. 2023. Living guidelines for generative AI — why scientists must oversee its use. *Nature*, 622(7984): 693–696.

Bommasani, R.; Creel, K. A.; Kumar, A.; Jurafsky, D.; and Liang, P. S. 2022. Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization? *Advances in Neural Information Processing Systems*, 35: 3663–3678.

Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; Ng, C.; Wang, R.; and Ramesh, A. 2024. Video generation models as world simulators.

Cao, B.; Cao, Y.; Lin, L.; and Chen, J. 2023. Defending against alignment-breaking attacks via robustly aligned llm. *arXiv preprint arXiv:2309.14348*.

Cao, Y.; Cao, B.; and Chen, J. 2023. Stealthy and Persistent Unalignment on Large Language Models via Backdoor Injections. *arXiv preprint arXiv:2312.00027*.

Casper, S.; Bu, T.; Li, Y.; Li, J.; Zhang, K.; Hariharan, K.; and Hadfield-Menell, D. 2023a. Red teaming deep neural networks with feature synthesis tools. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Casper, S.; Hariharan, K.; and Hadfield-Menell, D. 2022. Diagnostics for deep neural networks with automated copy/paste attacks. *arXiv preprint arXiv:2211.10024*.

Casper, S.; Killian, T.; Kreiman, G.; and Hadfield-Menell, D. 2022. Red Teaming with Mind Reading: White-box adversarial policies in deep reinforcement learning. *arXiv preprint arXiv:2209.02167*.

Casper, S.; Lin, J.; Kwon, J.; Culp, G.; and Hadfield-Menell, D. 2023b. Explore, Establish, Exploit: Red Teaming Language Models from Scratch. *arXiv preprint arXiv:2306.09442*.

Cattell, S. 2023. Generative Red Team Recap.

Cattell, S.; Carson, A.; and Chowdhury, R. 2023. Ai Village at def con announces largest-ever public generative AI Red Team.

Chang, J.; and Custis, C. 2022. Understanding implementation challenges in machine learning documentation. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–8.

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Zhu, K.; Chen, H.; Yang, L.; Yi, X.; Wang, C.; Wang, Y.; et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Chao, P.; Robey, A.; Dobriban, E.; Hassani, H.; Pappas, G. J.; and Wong, E. 2023. Jailbreaking Black Box Large Language Models in Twenty Queries. In *Robustness of Few-shot and Zero-shot Learning in Large Foundation Models (R0-FoMo) Workshop at NeurIPS*.

Chen, B.; Paliwal, A.; and Yan, Q. 2023. Jailbreaker in jail: Moving target defense for large language models. In *Proceedings of the 10th ACM Workshop on Moving Target Defense*, 29–32.

Chen, B.; Wang, G.; Guo, H.; Wang, Y.; and Yan, Q. 2023a. Understanding Multi-Turn Toxic Behaviors in Open-Domain Chatbots. In *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*, 282–296.

Chen, Q. Z.; Weld, D. S.; and Zhang, A. X. 2021. Goldilocks: Consistent Crowdsourced Scalar Annotations with Relative Uncertainty. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–25.

Chen, Y.; Mendes, E.; Das, S.; Xu, W.; and Ritter, A. 2023b. Can Language Models be Instructed to Protect Personal Information? *arXiv preprint arXiv:2310.02224*.

Chin, Z.-Y.; Jiang, C.-M.; Huang, C.-C.; Chen, P.-Y.; and Chiu, W.-C. 2023. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. *arXiv preprint arXiv:2309.06135*.

Chung, J. J. Y.; Song, J. Y.; Kutty, S.; Hong, S.; Kim, J.; and Lasecki, W. S. 2019. Efficient elicitation approaches to estimate collective crowd answers. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–25.

Costanza-Chock, S.; Harvey, E.; Raji, I. D.; Czernuszenko, M.; and Buolamwini, J. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1571–1583. ArXiv:2310.02521 [cs].

Crawford, K. 2024. Generative AI's environmental costs are soaring — and mostly secret. *Nature*, 626(8000): 693–693.

Delgado, F.; Yang, S.; Madaio, M.; and Yang, Q. 2023. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–23.

Deng, B.; Wang, W.; Feng, F.; Deng, Y.; Wang, Q.; and He, X. 2023a. Attack Prompt Generation for Red Teaming and Defending Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2176–2189.

Deng, G.; Liu, Y.; Li, Y.; Wang, K.; Zhang, Y.; Li, Z.; Wang, H.; Zhang, T.; and Liu, Y. 2023b. MasterKey: Automated jailbreak across multiple large language model chatbots. *arXiv preprint arXiv:2307.08715*.

Deng, W. H.; Guo, B.; Devrio, A.; Shen, H.; Eslami, M.; and Holstein, K. 2023c. Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18.

Deng, W. H.; Yildirim, N.; Chang, M.; Eslami, M.; Holstein, K.; and Madaio, M. 2023d. Investigating Practices and Opportunities for Cross-functional Collaboration around AI Fairness in Industry Practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 705–716.

Deng, Y.; Zhang, W.; Pan, S. J.; and Bing, L. 2023e. Multilingual jailbreak challenges in large language models. *arXiv preprint arXiv:2310.06474*.

Ding, P.; Kuang, J.; Ma, D.; Cao, X.; Xian, Y.; Chen, J.; and Huang, S. 2023. A Wolf in Sheep's Clothing: Generalized Nested Jailbreak Prompts can Fool Large Language Models Easily. *arXiv preprint arXiv:2311.08268*.

Donahue, C.; Caillon, A.; Roberts, A.; Manilow, E.; Esling, P.; Agostinelli, A.; Verzetti, M.; Simon, I.; Pietquin, O.; Zeghidour, N.; et al. 2023. SingSong: Generating musical accompaniments from singing. *arXiv preprint arXiv:2301.12662*.

Du, Y.; Zhao, S.; Ma, M.; Chen, Y.; and Qin, B. 2023. Analyzing the Inherent Response Tendency of LLMs: Real-World Instructions-Driven Jailbreak. *arXiv preprint arXiv:2312.04127*.

Feffer, M.; Heidari, H.; and Lipton, Z. C. 2023. Moral Machine or Tyranny of the Majority? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(55): 5974–5982.

Feffer, M.; Lipton, Z. C.; and Donahue, C. 2023. DeepDrake ft. BTS-GAN and TayloRVC:An Exploratory Analysis of Musical Deepfakes and Hosting Platforms. In *Proceedings of the 2nd Workshop on Human-Centric Music Information Retrieval 2023 (HCMIR 2023)*.

Feffer, M.; Skirpan, M.; Lipton, Z.; and Heidari, H. 2023. From Preference Elicitation to Participatory ML: A Critical Survey & Guidelines for Future Research. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 38–48.

Field, H. 2024a. OpenAI dissolves team focused on long-term AI risks, less than one year after announcing it.

Field, H. 2024b. OpenAI quietly removes ban on military use of its AI tools.

Friedler, S.; Singh, R.; Blili-Hamelin, B.; Metcalf, J.; and Chen, B. J. 2023. AI Red-Teaming Is Not a One-Stop Solution to AI Harms: Recommendations for Using Red-Teaming for AI Accountability. *Data & Society*.

Ganguli, D.; Lovitt, L.; Kernion, J.; Askell, A.; Bai, Y.; Kadavath, S.; Mann, B.; Perez, E.; Schiefer, N.; Ndousse, K.; et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Ge, S.; Zhou, C.; Hou, R.; Khabsa, M.; Wang, Y.-C.; Wang, Q.; Han, J.; and Mao, Y. 2023. MART: Improving LLM Safety with Multi-round Automatic Red-Teaming. *arXiv preprint arXiv:2311.07689*.

Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3356–3369.

Ghosh, S.; and Caliskan, A. 2023. ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five Other Low-Resource Languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 901–912. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702310.

Gong, Y.; Ran, D.; Liu, J.; Wang, C.; Cong, T.; Wang, A.; Duan, S.; and Wang, X. 2023. FigStep: Jailbreaking Large Vision-language Models via Typographic Visual Prompts. *arXiv preprint arXiv:2311.05608*.

Greenblatt, R.; Shlegeris, B.; Sachan, K.; and Roger, F. 2023. AI Control: Improving Safety Despite Intentional Subversion. *arXiv preprint arXiv:2312.06942*.

Guha, N.; Lawrence, C.; Gailmard, L. A.; Rodolfa, K.; Surani, F.; Bommasani, R.; Raji, I.; Cuéllar, M.-F.; Honigsberg, C.; Liang, P.; et al. 2023. Ai regulation has its own alignment problem: The technical and institutional feasibility of disclosure, registration, licensing, and auditing. *George Washington Law Review, Forthcoming*.

Haim, A.; Salinas, A.; and Nyarko, J. 2024. What's in a Name? Auditing Large Language Models for Race and Gender Bias. *arXiv preprint arXiv:2402.14875*.

He, J.; Feng, W.; Min, Y.; Yi, J.; Tang, K.; Li, S.; Zhang, J.; Chen, K.; Zhou, W.; Xie, X.; et al. 2023. Control Risk for Potential Misuse of Artificial Intelligence in Science. *arXiv preprint arXiv:2312.06632*.

Hofmann, V.; Kalluri, P. R.; Jurafsky, D.; and King, S. 2024. Dialect prejudice predicts AI decisions about people's character, employability, and criminality. *arXiv preprint arXiv:2403.00742*.

Horvitz, E. 2022. On the horizon: Interactive and compositional deepfakes. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, 653–661.

Huang, K.; Liu, X.; Guo, Q.; Sun, T.; Sun, J.; Wang, Y.; Zhou, Z.; Wang, Y.; Teng, Y.; Qiu, X.; et al. 2023a. Flames: Benchmarking Value Alignment of Chinese Large Language Models. *arXiv preprint arXiv:2311.06899*.

Huang, Y.; Gupta, S.; Xia, M.; Li, K.; and Chen, D. 2023b. Catastrophic jailbreak of open-source LLMs via exploiting generation. *arXiv preprint arXiv:2310.06987*.

Hube, C.; Fetahu, B.; and Gadiraju, U. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

Inie, N.; Stray, J.; and Derczynski, L. 2023. Summon a Demon and Bind it: A Grounded Theory of LLM Red Teaming in the Wild. *arXiv preprint arXiv:2311.06237*.

Jain, N.; Schwarzschild, A.; Wen, Y.; Somepalli, G.; Kirchenbauer, J.; Chiang, P.-y.; Goldblum, M.; Saha, A.; Geiping, J.; and Goldstein, T. 2023. Baseline defenses for adversarial attacks against aligned language models. *arXiv preprint arXiv:2309.00614*.

Kenthapadi, K.; Lakkaraju, H.; and Rajani, N. 2023. Generative AI Meets Responsible AI: Practical Challenges and Opportunities. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5805–5806.

Kittur, A.; Nickerson, J. V.; Bernstein, M.; Gerber, E.; Shaw, A.; Zimmerman, J.; Lease, M.; and Horton, J. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*, 1301–1318.

Kleinberg, J.; and Raghavan, M. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22): e2018340118.

Knearem, T.; Khwaja, M.; Gao, Y.; Bentley, F.; and Kliman-Silver, C. E. 2023. Exploring the future of design tooling: The role of artificial intelligence in tools for user experience professionals. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–6.

Lambert, N.; and Calandra, R. 2023. The Alignment Ceiling: Objective Mismatch in Reinforcement Learning from Human Feedback. *arXiv preprint arXiv:2311.00168*.

Lambert, N.; Gilbert, T. K.; and Zick, T. 2023. Entangled Preferences: The History and Risks of Reinforcement Learning and Human Feedback. *arXiv preprint arXiv:2310.13595*.

Lapid, R.; Langberg, R.; and Sipper, M. 2023. Open sesame! universal black box jailbreaking of large language models. *arXiv preprint arXiv:2309.01446*.

Lee, A.; Bai, X.; Pres, I.; Wattenberg, M.; Kummerfeld, J. K.; and Mihalcea, R. 2024. A mechanistic understanding of alignment algorithms: A case study on dpo and toxicity. *arXiv preprint arXiv:2401.01967*.

Lee, D.; Lee, J.; Ha, J.-W.; Kim, J.-H.; Lee, S.-W.; Lee, H.; and Song, H. O. 2023. Query-Efficient Black-Box Red Teaming via Bayesian Optimization. *arXiv preprint arXiv:2305.17444*.

Leike, J.; and Sutskever, I. 2023. Introducing Superalignment.

Levenson, E. 2014. The TSA Is in the Business of'Security Theater,'Not Security. *The Atlantic Magazine*.

Lewis, A.; and White, M. 2023. Mitigating Harms of LLMs via Knowledge Distillation for a Virtual Museum Tour Guide. In *Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants!*, 31–45.

Li, H.; Guo, D.; Fan, W.; Xu, M.; and Song, Y. 2023a. Multi-step jailbreaking privacy attacks on chatgpt. *arXiv preprint arXiv:2304.05197*.

Li, X.; Zhou, Z.; Zhu, J.; Yao, J.; Liu, T.; and Han, B. 2023b. DeepInception: Hypnotize Large Language Model to Be Jailbreaker. *arXiv preprint arXiv:2311.03191*.

Liao, Q. V.; Subramonyam, H.; Wang, J.; and Wortman Vaughan, J. 2023. Designerly understanding: Information needs for model transparency to support design ideation for AI-powered user experience. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, 1–21.

Liu, X.; Zhu, Y.; Lan, Y.; Yang, C.; and Qiao, Y. 2023a. Query-Relevant Images Jailbreak Large Multi-Modal Models. *arXiv preprint arXiv:2311.17600*.

Liu, Y.; Deng, G.; Xu, Z.; Li, Y.; Zheng, Y.; Zhang, Y.; Zhao, L.; Zhang, T.; and Liu, Y. 2023b. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*.

Luccioni, S.; Akiki, C.; Mitchell, M.; and Jernite, Y. 2023. Stable Bias: Evaluating Societal Representations in Diffusion Models. In *Advances in Neural Information Processing Systems (NeurIPS) 2023 Dataset and Benchmarks Track*.

Ma, C.; Yang, Z.; Gao, M.; Ci, H.; Gao, J.; Pan, X.; and Yang, Y. 2023. Red Teaming Game: A Game-Theoretic Framework for Red Teaming Language Models. *arXiv preprint arXiv:2310.00322*.

Madaio, M. A.; Stark, L.; Wortman Vaughan, J.; and Wallach, H. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

Mehrabi, N.; Goyal, P.; Dupuy, C.; Hu, Q.; Ghosh, S.; Zemel, R.; Chang, K.-W.; Galstyan, A.; and Gupta, R. 2023a. Flirt: Feedback loop in-context red teaming. *arXiv preprint arXiv:2308.04265*.

Mehrabi, N.; Goyal, P.; Ramakrishna, A.; Dhamala, J.; Ghosh, S.; Zemel, R.; Chang, K.-W.; Galstyan, A.; and Gupta, R. 2023b. JAB: Joint Adversarial Prompting and Belief Augmentation. In *Robustness of Few-shot and Zero-shot Learning in Large Foundation Models (R0-FoMo) Workshop at NeurIPS*.

Mehrotra, A.; Zampetakis, M.; Kassianik, P.; Nelson, B.; Anderson, H.; Singer, Y.; and Karbasi, A. 2023. Tree of Attacks: Jailbreaking Black-Box LLMs Automatically. *arXiv preprint arXiv:2312.02119*.

Mei, A.; Levy, S.; and Wang, W. Y. 2023. ASSERT: Automated Safety Scenario Red Teaming for Evaluating the Robustness of Large Language Models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Mei, K.; Fereidooni, S.; and Caliskan, A. 2023. Bias Against 93 Stigmatized Groups in Masked Language Models and Downstream Sentiment Classification Tasks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1699–1710.

Microsoft. 2023. Planning red teaming for large language models (LLMs) and their applications - Azure OpenAI Service.

Microsoft. 2024. Copilot in Bing: Our approach to Responsible AI.

Mok, A. 2023. ChatGPT will no longer comply if you ask it to repeat a word 'forever'— after a recent prompt revealed training data and personal info.

Mu, N.; Chen, S.; Wang, Z.; Chen, S.; Karamardian, D.; Aljeraisy, L.; Hendrycks, D.; and Wagner, D. 2023. Can LLMs Follow Simple Rules? *arXiv preprint arXiv:2311.04235*.

Nahar, N.; Zhou, S.; Lewis, G.; and Kästner, C. 2021. Collaboration Challenges in Building ML-Enabled Systems: Communication, Documentation, Engineering, and Process. *arXiv preprint arXiv:2110.10234*.

Nasr, M.; Carlini, N.; Hayase, J.; Jagielski, M.; Cooper, A. F.; Ippolito, D.; Choquette-Choo, C. A.; Wallace, E.; Tramèr, F.; and Lee, K. 2023. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*.

Neel, S.; and Chang, P. 2023. Privacy Issues in Large Language Models: A Survey. *arXiv preprint arXiv:2312.06717*.

Nguyen, C.; Morgan, C.; and Mittal, S. 2022. Poster CTI4AI: Threat Intelligence Generation and Sharing after Red Teaming AI Models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 3431–3433.

Omrani Sabbaghi, S.; Wolfe, R.; and Caliskan, A. 2023. Evaluating biased attitude associations of language models in an intersectional context. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 542–553.

OpenAI. 2023.

Pelrine, K.; Taufeeque, M.; Zając, M.; McLean, E.; and Gleave, A. 2023. Exploiting Novel GPT-4 APIs. *arXiv preprint arXiv:2312.14302*.

Perez, E.; Huang, S.; Song, F.; Cai, T.; Ring, R.; Aslanides, J.; Glaese, A.; McAleese, N.; and Irving, G. 2022. Red Teaming Language Models with Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3419–3448.

Pfau, J.; Infanger, A.; Sheshadri, A.; Panda, A.; Michael, J.; and Huebner, C. 2023. Eliciting Language Model Behaviors using Reverse Language Models. In *Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS*.

Pierson, E. 2017. Demographics and discussion influence views on algorithmic fairness. *arXiv preprint arXiv:1712.09124*.

Price, E. 2023. Asking ChatGPT to Repeat Words "Forever" May Violate OpenAI's Terms.

Puttaparthi, P. C. R.; Deo, S. S.; Gul, H.; Tang, Y.; Shang, W.; and Yu, Z. 2023. Comprehensive Evaluation of ChatGPT Reliability Through Multilingual Inquiries. *arXiv preprint arXiv:2312.10524*.

Qi, X.; Huang, K.; Panda, A.; Wang, M.; and Mittal, P. 2023. Visual Adversarial Examples Jailbreak Aligned Large Language Models. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*.

Qiu, H.; Zhang, S.; Li, A.; He, H.; and Lan, Z. 2023. Latent jailbreak: A benchmark for evaluating text safety and output robustness of large language models. *arXiv preprint arXiv:2307.08487*.

Radharapu, B.; Robinson, K.; Aroyo, L.; and Lahoti, P. 2023. AART: AI-Assisted Red-Teaming with Diverse Data Generation for New LLM-powered Applications. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 380–395.

Rae, J. W.; Borgeaud, S.; Cai, T.; Millican, K.; Hoffmann, J.; Song, F.; Aslanides, J.; Henderson, S.; Ring, R.; Young, S.; et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Rajani, N.; Lambert, N.; and Tunstall, L. 2023. Red-Teaming Large Language Models.

Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 33–44.

Rakova, B.; Yang, J.; Cramer, H.; and Chowdhury, R. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–23.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Rando, J.; Paleka, D.; Lindner, D.; Heim, L.; and Tramer, F. 2022. Red-Teaming the Stable Diffusion Safety Filter. In *NeurIPS ML Safety Workshop*.

Rando, J.; and Tramèr, F. 2023. Universal jailbreak backdoors from poisoned human feedback. *arXiv preprint arXiv:2311.14455*.

Rao, A.; Vashistha, S.; Naik, A.; Aditya, S.; and Choudhury, M. 2023. Tricking LLMs into Disobedience: Understanding, Analyzing, and Preventing Jailbreaks. *arXiv preprint arXiv:2305.14965*.

Rastogi, C.; Tulio Ribeiro, M.; King, N.; Nori, H.; and Amershi, S. 2023. Supporting human-ai collaboration in auditing llms with llms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 913–926.

Reisman, D.; Schultz, J.; Crawford, K.; and Whittaker, M. 2018. Algorithmic impact assessments: a practical Framework for Public Agency. *AI Now*, 9.

Robey, A.; Wong, E.; Hassani, H.; and Pappas, G. 2023. SmoothLLM: Defending Large Language Models Against Jailbreaking Attacks. In *Robustness of Few-shot and Zero-shot Learning in Large Foundation Models (R0-FoMo) Workshop at NeurIPS*.

Rogers, R. 2024. AI's Energy Demands Are Out of Control. Welcome to the Internet's Hyper-Consumption Era. *Wired*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674–10685. New Orleans, LA, USA: IEEE. ISBN 978-1-66546-946-3.

Röttger, P.; Kirk, H. R.; Vidgen, B.; Attanasio, G.; Bianchi, F.; and Hovy, D. 2023. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint arXiv:2308.01263*.

Röttger, P.; Vidgen, B.; Nguyen, D.; Waseem, Z.; Margetts, H.; and Pierrehumbert, J. 2021. HateCheck: Functional Tests for Hate Speech Detection Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 41–58.

Roy, S.; Harshvardhan, A.; Mukherjee, A.; and Saha, P. 2023a. Probing LLMs for hate speech detection: strengths and vulnerabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 6116–6128.

Roy, S. S.; Naragam, K. V.; and Nilizadeh, S. 2023. Generating Phishing Attacks using ChatGPT. *arXiv preprint arXiv:2305.05133*.

Roy, S. S.; Thota, P.; Naragam, K. V.; and Nilizadeh, S. 2023b. From Chatbots to PhishBots?–Preventing Phishing scams created using ChatGPT, Google Bard and Claude. *arXiv preprint arXiv:2310.19181*.

Salem, A.; Paverd, A.; and Köpf, B. 2023. Maatphor: Automated Variant Analysis for Prompt Injection Attacks. *arXiv preprint arXiv:2312.11513*.

Schlarmann, C.; and Hein, M. 2023. On the adversarial robustness of multi-modal foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3677–3685.

Schuett, J.; Dreksler, N.; Anderljung, M.; McCaffary, D.; Heim, L.; Bluemke, E.; and Garfinkel, B. 2023. Towards best practices in AGI safety and governance: A survey of expert opinion. *arXiv preprint arXiv:2305.07153*.

Schulhoff, S. V.; Pinto, J.; Khan, A.; Bouchard, L.-F.; Si, C.; Anati, S.; Tagliabue, V.; Kost, A. L.; Carnahan, C. R.; and Boyd-Graber, J. L. 2023. Ignore This Title and HackAPrompt: Exposing Systemic Vulnerabilities of LLMs Through a Global Prompt Hacking Competition. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Shah, R.; Montixi, Q. F.; Pour, S.; Tagade, A.; and Rando, J. 2023. Scalable and Transferable Black-Box Jailbreaks for Language Models via Persona Modulation. In *Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS*.

Shayegani, E.; Dong, Y.; and Abu-Ghazaleh, N. 2023. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In *The Twelfth International Conference on Learning Representations*.

Shayegani, E.; Mamun, M. A. A.; Fu, Y.; Zaree, P.; Dong, Y.; and Abu-Ghazaleh, N. 2023. Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv preprint arXiv:2310.10844*.

Shen, X.; Chen, Z.; Backes, M.; Shen, Y.; and Zhang, Y. 2023. "Do Anything Now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825*.

Shevlane, T.; Farquhar, S.; Garfinkel, B.; Phuong, M.; Whittlestone, J.; Leung, J.; Kokotajlo, D.; Marchal, N.; Anderljung, M.; Kolt, N.; et al. 2023. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*.

Shi, Z.; Wang, Y.; Yin, F.; Chen, X.; Chang, K.-W.; and Hsieh, C.-J. 2023. Red Teaming Language Model Detectors with Language Models. *arXiv preprint arXiv:2305.19713*.

Solaiman, I. 2023. The gradient of generative AI release: Methods and considerations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 111–122.

Srivastava, A.; Ahuja, R.; and Mukku, R. 2023. No Offense Taken: Eliciting Offensiveness from Language Models. *arXiv preprint arXiv:2310.00892*.

Stone, B.; and Bergen, M. 2024. OpenAI Working With U.S. Military on Cybersecurity Tools.

Sun, Z.; Shen, Y.; Zhou, Q.; Zhang, H.; Chen, Z.; Cox, D.; Yang, Y.; and Gan, C. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. *arXiv preprint arXiv:2305.03047*.

Tan, S.; Joty, S.; Baxter, K.; Taeihagh, A.; Bennett, G. A.; and Kan, M.-Y. 2021. Reliability Testing for Natural Language Processing Systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4153–4169.

The Frontier Model Forum (FMF). 2023. Frontier Model Forum: What is Red-Teaming?

The White House. 2023. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.

Tian, Y.; Yang, X.; Zhang, J.; Dong, Y.; and Su, H. 2023. Evil geniuses: Delving into the safety of llm-based agents. *arXiv preprint arXiv:2311.11855*.

Tong, S.; Jones, E.; and Steinhardt, J. 2023. Mass-Producing Failures of Multimodal Systems with Language Models. *arXiv preprint arXiv:2306.12105*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Toxtli, C.; Suri, S.; and Savage, S. 2021. Quantifying the Invisible Labor in Crowd Work. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–26.

Tsai, Y.-L.; Hsu, C.-Y.; Xie, C.; Lin, C.-H.; Chen, J.-Y.; Li, B.; Chen, P.-Y.; Yu, C.-M.; and Huang, C.-Y. 2023. Ring-A-Bell! How Reliable are Concept Removal Methods for Diffusion Models? *arXiv preprint arXiv:2310.10012*.

Tu, H.; Cui, C.; Wang, Z.; Zhou, Y.; Zhao, B.; Han, J.; Zhou, W.; Yao, H.; and Xie, C. 2023. How Many Unicorns Are in This Image? A Safety Evaluation Benchmark for Vision LLMs. *arXiv preprint arXiv:2311.16101*.

Vaughan, J. W. 2017. Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research. *J. Mach. Learn. Res.*, 18(1): 7026–7071.

Wan, Y.; and Chang, K.-W. 2024. The Male CEO and the Female Assistant: Probing Gender Biases in Text-To-Image Models Through Paired Stereotype Test. *arXiv preprint arXiv:2402.11089*.

Wang, B.; Chen, W.; Pei, H.; Xie, C.; Kang, M.; Zhang, C.; Xu, C.; Xiong, Z.; Dutta, R.; Schaeffer, R.; et al. 2023a. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. *arXiv preprint arXiv:2306.11698*.

Wang, H.; and Shu, K. 2023. Backdoor Activation Attack: Attack Large Language Models using Activation Steering for Safety-Alignment. *arXiv preprint arXiv:2311.09433*.

Wang, J.; Wu, J.; Chen, M.; Vorobeychik, Y.; and Xiao, C. 2023b. On the Exploitability of Reinforcement Learning with Human Feedback for Large Language Models. *arXiv preprint arXiv:2311.09641*.

Wang, Y.; Teng, Y.; Huang, K.; Lyu, C.; Zhang, S.; Zhang, W.; Ma, X.; and Wang, Y. 2023c. Fake Alignment: Are LLMs Really Aligned Well? *arXiv preprint arXiv:2311.05915*.

Wang, Z.; Yang, F.; Wang, L.; Zhao, P.; Wang, H.; Chen, L.; Lin, Q.; and Wong, K.-F. 2023d. Self-Guard: Empower the LLM to Safeguard Itself. *arXiv preprint arXiv:2310.15851*.

Wei, A.; Haghtalab, N.; and Steinhardt, J. 2023. Jailbroken: How Does LLM Safety Training Fail? In *Thirty-seventh Conference on Neural Information Processing Systems*.

Wei, Z.; Wang, Y.; and Wang, Y. 2023. Jailbreak and guard aligned language models with only few in-context demonstrations. *arXiv preprint arXiv:2310.06387*.

Weidinger, L.; Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; Griffin, C.; Bariach, B.; et al. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. *arXiv preprint arXiv:2310.11986*.

Weir, A. 2014. *The Martian*. Random House.

Widder, D. G.; West, S.; and Whittaker, M. 2023. Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI. *SSRN preprint 10.2139/ssrn.4543807*.

Wood, B. J.; and Duggan, R. A. 2000. Red teaming of advanced information assurance concepts. In *Proceedings DARPA Information Survivability Conference and Exposition. DISCEX'00*, volume 2, 112–118. IEEE.

Woodruff, A.; Fox, S. E.; Rousso-Schindler, S.; and Warshaw, J. 2018. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 chi conference on human factors in computing systems*, 1–14.

Wu, F.; Liu, X.; and Xiao, C. 2023. DeceptPrompt: Exploiting LLM-driven Code Generation via Adversarial Natural Language Instructions. *arXiv preprint arXiv:2312.04730*.

Wu, Y.; Li, X.; Liu, Y.; Zhou, P.; and Sun, L. 2023. Jailbreaking GPT-4V via Self-Adversarial Attacks with System Prompts. *arXiv preprint arXiv:2311.09127*.

Xie, Y.; Yi, J.; Shao, J.; Curl, J.; Lyu, L.; Chen, Q.; Xie, X.; and Wu, F. 2023. Defending ChatGPT against jailbreak attack via self-reminders. *Nature Machine Intelligence*, 1–11.

Xu, J.; Ju, D.; Li, M.; Boureau, Y.-L.; Weston, J.; and Dinan, E. 2020. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.

Xu, J.; Ju, D.; Li, M.; Boureau, Y.-L.; Weston, J.; and Dinan, E. 2021. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2950–2968.

Xu, N.; Wang, F.; Zhou, B.; Li, B. Z.; Xiao, C.; and Chen, M. 2023. Cognitive Overload: Jailbreaking Large Language Models with Overloaded Logical Thinking. *arXiv preprint arXiv:2311.09827*.

Yang, Y.; Hui, B.; Yuan, H.; Gong, N.; and Cao, Y. 2024. Sneakyprompt: Jailbreaking text-to-image generative models.

In *2024 IEEE Symposium on Security and Privacy (SP)*, 123–123. IEEE Computer Society.

Yao, D.; Zhang, J.; Harris, I. G.; and Carlsson, M. 2023a. Fuzzllm: A novel and universal fuzzing framework for proactively discovering jailbreak vulnerabilities in large language models. *arXiv preprint arXiv:2309.05274*.

Yao, Y.; Duan, J.; Xu, K.; Cai, Y.; Sun, E.; and Zhang, Y. 2023b. A Survey on Large Language Model (LLM) Security and Privacy: The Good, the Bad, and the Ugly. *arXiv preprint arXiv:2312.02003*.

Yong, Z. X.; Menghini, C.; and Bach, S. 2023. Low-Resource Languages Jailbreak GPT-4. In *Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS*.

Yu, J.; Lin, X.; and Xing, X. 2023. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts. *arXiv preprint arXiv:2309.10253*.

Yuan, Y.; Jiao, W.; Wang, W.; Huang, J.-t.; He, P.; Shi, S.; and Tu, Z. 2023. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. *arXiv preprint arXiv:2308.06463*.

Zenko, M. 2015. *Red Team: How to succeed by thinking like the enemy*. Basic Books.

Zhang, J.; Zhou, Y.; Hui, B.; Liu, Y.; Li, Z.; and Hu, S. 2023a. TrojanSQL: SQL Injection against Natural Language Interface to Database. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 4344–4359.

Zhang, M.; Pan, X.; and Yang, M. 2023. JADE: A Linguistics-based Safety Evaluation Platform for LLM. *arXiv preprint arXiv:2311.00286*.

Zhang, X.; Zhang, C.; Li, T.; Huang, Y.; Jia, X.; Xie, X.; Liu, Y.; and Shen, C. 2023b. A Mutation-Based Method for Multi-Modal Jailbreaking Attack Detection. *arXiv preprint arXiv:2312.10766*.

Zhang, Z.; Yang, J.; Ke, P.; and Huang, M. 2023c. Defending Large Language Models Against Jailbreaking Attacks Through Goal Prioritization. *arXiv preprint arXiv:2311.09096*.

Zhao, W.; Li, Z.; and Sun, J. 2023. Causality Analysis for Evaluating the Security of Large Language Models. *arXiv preprint arXiv:2312.07876*.

Zhu, S.; Zhang, R.; An, B.; Wu, G.; Barrow, J.; Wang, Z.; Huang, F.; Nenkova, A.; and Sun, T. 2023a. AutoDAN: Automatic and Interpretable Adversarial Attacks on Large Language Models. In *Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS*.

Zhu, Z.; Wang, J.; Cheng, H.; and Liu, Y. 2023b. Unmasking and Improving Data Credibility: A Study with Datasets for Training Harmless Language Models. *arXiv preprint arXiv:2311.11202*.

Zhuo, T. Y.; Huang, Y.; Chen, C.; and Xing, Z. 2023. Red teaming chatgpt via jailbreaking: Bias, robustness, reliability and toxicity. *arXiv preprint arXiv:2301.12867*, 12–2.

Zou, A.; Wang, Z.; Kolter, J. Z.; and Fredrikson, M. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.