Red Teaming LLMs as Socio-Technical Practice: From Exploration and Data Creation to Evaluation

Adriana Alvarado Garcia*
Responsible Tech Research
IBM Research
Yorktown Heights, New York, USA
adriana.ag@ibm.com

Ruyuan Wan*
College of Information Sciences and Technology
The Pennsylvania State University
University Park, Pennsylvania, USA
rjw6289@psu.edu

Ozioma C. Oguine
Computer Science and Engineering
University of Notre Dame
Notre Dame, Indiana, USA
ooguine@nd.edu

Karla Badillo-Urquiola
Computer Science and Engineering
University of Notre Dame
Notre Dame, Indiana, USA
kbadill3@nd.edu

ABSTRACT

Recently, red teaming, with roots in security, has become a key evaluative approach to ensure the safety and reliability of Generative Artificial Intelligence. However, most existing work emphasizes technical benchmarks and attack success rates, leaving the socio-technical practices of how red teaming datasets are defined, created, and evaluated under-examined. Drawing on 22 interviews with practitioners who design and evaluate red teaming datasets, we examine the data practices and standards that underpin this work. Because adversarial datasets determine the scope and accuracy of model evaluations, they are critical artifacts for assessing potential harms from large language models. Our contributions are first, empirical evidence of practitioners conceptualizing red teaming and developing and evaluating red teaming datasets. Second, we reflect on how practitioners' conceptualization of risk leads to overlooking the context, interaction type, and user specificity. We conclude with three opportunities for HCI researchers to expand the conceptualization and data practices for red-teaming.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Natural language generation**; • **Security and privacy** → *Human and societal aspects of security and privacy*.

KEYWORDS

Red Teaming, Language Models, Evaluation, Data Practice, Data Work, Socio-Technical, Artificial Intelligence

ACM Reference Format:

Adriana Alvarado Garcia, Ruyuan Wan, Ozioma C. Oguine, and Karla Badillo-Urquiola. 2026. Red Teaming LLMs as Socio-Technical Practice: From Exploration and Data Creation to Evaluation. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3772318.3790792

*Both authors contributed equally to this research. This research was conducted during Ruyuan Wan's internship at IBM Research.

CHI '26, April 13–17, 2026, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3790792

1 INTRODUCTION

Red teaming originates from the context of security, where adversaries simulate attacks to uncover vulnerabilities [8, 28, 74, 91]. Nowadays, red teaming has become central to evaluating the safety and robustness of systems, especially with the rise of large language models (LLMs). LLMs now outperform benchmarks in reasoning and language understanding [58], but also produce harmful, biased, or misleading outputs. These risks have made red teaming an emerging standard practice in industry, government, and academia, embedded in regulations such as the European Union's Artificial Intelligence Act [27].

Adversarial datasets are central to red teaming LLMs, revealing models' vulnerabilities and blind spots. These datasets are collections of prompts or conversations designed to elicit harmful or unsafe behaviors, enabling systematic evaluation and alignment improvements. They define what counts as harm, determine how models are tested, and shape what risks present to downstream users. Yet datasets are not neutral: they embody the assumptions and values of those who create them [51, 65, 67]. Prior work in Human-Computer Interaction (HCI) and critical data studies has shown that annotation practices, labeling schemes, and benchmark structures not only shape technical outcomes but also social and cultural norms [51, 65, 67]. Data workers develop intuitive understandings of their data through multiple forms of intervention, including treating data as given, captured, curated, designed, or created [53]. However, little is known about how AI practitioners create, reuse, and evaluate red teaming datasets for LLMs. Different from many prior data work that analyze data produced in real-world contexts, such as social media content [7] or image recognition [51], red-teaming datasets are created with an explicitly adversarial, experimental, and testing-oriented intention. Rather than capturing naturally occurring human activity, practitioners design prompts to provoke failures, explore attack strategies, and probe the boundaries of model behavior [44]. This adversarial stance reframes data work in this context: while red-teaming data creation is highly open-ended and unstructured, it still depends on human judgment about what to include and how to validate it. Most red teaming LLMs research emphasizes technical benchmarks and attack success rates, leaving the socio-technical practices of dataset development under-examined. These refer to the intertwined social and technical choices about what should be supported socially, and what can be implemented technically [2], such as how harms are defined, how adversarial prompts or dialogue are created, and what evaluation metrics are adopted.

This gap matters because harmfulness is not discovered as a fixed concept in red teaming evaluation, but constructed through choices about prompts, categories, evaluation metrics, and collaborations. Decisions about what harms to include or exclude, and how to measure success, shape not only the coverage of datasets but also the guardrails and mitigations that follow. Understanding these practices is critical for HCI, which has long examined how technologies embed human values, and for AI safety, which increasingly depends on socio-technical insights. To address this gap, we conducted 22 semi-structured interviews with AI practitioners who design, build, or reuse datasets for red teaming LLMs. Our study is guided by the following research questions:

(1) How do AI practitioners create, develop, and evaluate red teaming datasets, and why in these ways?
(2) What tools and support do AI practitioners need when developing red teaming datasets?

Our work extends prior HCI examinations of data work by revealing a distinct form of adversarial, interaction-driven data work that emerges uniquely in red teaming LLMs. Our analysis highlights three critical moments in this work: (1) defining and framing red teaming tasks, (2) developing adversarial datasets, and (3) evaluating models against them. Across these moments, we show that disciplinary backgrounds shape whether red teaming is framed as exploration or classification, while motivations range from observing "in-the-wild" jailbreaks to addressing technical or regulatory gaps. We find that datasets are developed through different pathways: created from scratch, repurposed from existing resources, or derived from human interactions, and that evaluation is complicated by issues of context, diversity, and metrics. Taken together, these practices reveal red teaming as more interactional and social than AI practitioners typically anticipated in benchmark-driven work, with risks arising not only from single prompts but also from multi-turn, multilingual, and multicultural exchanges. Further, we identify opportunities for HCI researchers to support red teaming practitioners: by expanding evaluations to center on the context of use, incorporating domain expertise into definitions of harm, and better accounting for interaction-level risks. Our findings shed light on emerging exploratory data practices that arise when LLMs are engaged through open-ended and interactive use.

2 RELATED WORK

To contextualize our contribution, we first introduce the concept and practice of red teaming, then examine how recent datasets operationalize adversarial testing for LLMs. We then consider the diverse stakeholders involved in red teaming for AI safety, highlighting the motivations and tensions that shape its practice. Finally, we situate our work within HCI research, showing how longstanding contributions on data practices provide critical insights for understanding and improving red teaming evaluation.

2.1 Introduction to Red Teaming

Red teaming did not originate as a technical protocol, but rather as a strategic exercise in adversarial thinking. Rooted in military war-gaming, red teams were tasked with challenging institutional assumptions by "*thinking like the enemy*" to expose vulnerabilities in operational plans [8, 74, 91]. Over time, this practice migrated into cybersecurity, where red teams simulate breaches to identify flaws in digital defenses, a tradition now formalized in frameworks like the U.S. Department of Defense's Cyber Red Teaming Guide [28, 74]. This adversarial framing of treating safety as a matter of defense against active threats shaped how risk was conceptualized in digital systems, establishing a dominant view of harm as something imposed by hostile actors, rather than as the emergent, everyday vulnerabilities experienced by diverse users [3, 28]. It provided a blueprint for how harms are anticipated, operationalized, and mitigated in sociotechnical design [32, 75]. Academic scholarship across security studies [47], information systems [10], CSCW and HCI [92–94] contributed to this transition, theorizing red teaming as a method of system auditing, threat modeling, and scenario-based testing. These foundational contributions laid the groundwork for the widespread adoption of red teaming in industry, government, and research settings.

In the domain of AI, and particularly with the rise of LLMs and emerging agentic AI systems, red teaming has taken on renewed urgency [3, 28, 84]. While LLMs offer substantial societal benefits, including expanded access to information, enhanced productivity, and support for human creative and cognitive tasks [33, 79, 95], they have also been shown to generate outputs that are toxic, biased, or sometimes inaccurate [83]. In response, the Center for AI Safety released a statement in 2023, warning that AI extinction risk should be a global priority [19]. Therefore, governments (e.g., EU's AI Act [27] and U.S. Executive Order 14110 [36]) and industry leaders (e.g., OpenAI [57], Anthropic [9]) have institutionalized red teaming as a key part of AI evaluation. While academic researchers have also emphasized the need for rigorous evaluation frameworks to ensure the safety and reliability of AI systems in real-world settings (e.g., [28, 30, 46, 52, 59, 59, 76, 84]), their valuable contributions are often overshadowed by industry-led initiatives.

Red teaming has emerged as a field of study connecting security, AI safety, and responsible innovation. Yet, the conceptual framing of red teaming remains largely inherited from its adversarial roots [28]. The dominant notion of a "threat" is typically cast as a malicious actor or unexpected query, someone trying to game or break the system [28, 91]. This framing often sidelines more systemic or structural harms, such as how LLMs may reinforce racial stereotypes, marginalize dialects, or exclude non-Western epistemologies [32, 72]. As a result, current red teaming practices tend to reflect narrow slices of AI risk: those that are easily measurable, align with regulatory compliance, or reflect high-profile public controversies. By centering the people and practices involved, we showed that red teaming–often framed as technical safety work–is shaped by social values and disciplinary priorities.

Dataset	Year	Size / Type	Design Basis	Generation Method	Annotation Method	Definition of Harm
HH-RLHF [30]	2022	~38k human-generated comparisons	Anthropic's "helpful/harmless" framework	👤	👤	Harm defined by human-preference for "hamrful" outputs
AdvBench [97]	2023	500 adversarial "strings"	Adversarial robustness testing focused on policy-violating behavior	🤖	🤖	Harm framed as violating model's alignment or safety constraints
BeaverTails [39]	2023	30,207 Question–Answer pairs	OpenAI jailbreak prompt sets + community red teaming	🤖	🤖 + 👤	Harm defined in relation to jailbreak success and unsafe generations
HarmBench [49]	2024	~2,400 test cases	Prior harm taxonomies + academic harm frameworks	👤 + 🤖	👤 + 🤖	Harm defined by standardized taxonomy (violence, hate, self-harm, etc.)

Legend: 👤 = Human; 🤖 = AI/LLM; 👤 + 🤖 = Human-AI collaboration

Table 1: Comparison of major Red Teaming Datasets for LLMs

2.2 The Role of Red Teaming Datasets for LLMs

A growing body of research has demonstrated that LLMs can produce harmful, toxic, or biased outputs, even after alignment efforts [30, 31, 59, 83]. For example, both Gehman et al. [31] and Welbl et al. [86], documented how LLMs can produce toxic and misleading outputs even under controlled input conditions. Early generative models like GPT-2 responded with toxic completions to 4.3–6.1% of user prompts, depending on prompt structure and sampling temperature [31]. Sheng et al. [71] also showed that models can reinforce gender and racial stereotypes in generated sentences. These findings sparked renewed attention to the limitations of existing evaluation pipelines and the need for targeted adversarial testing.

In response, researchers have turned to red teaming datasets, structured collections of adversarial or high-risk prompts used to probe model behavior [28, 30, 59]. Unlike benchmark datasets designed to test factual accuracy or task performance [61, 82], red teaming datasets are crafted to elicit failure modes, especially in high-risk domains like hate speech, and safety-critical advice [30, 59]. These datasets act as epistemic instruments, shaping which types of harms are visible, how safety is defined, and what mitigation efforts follow [6, 28, 54, 65]. An early impactful effort, RealToxicityPrompts by Gehman et al. [31], included 100,000 prompts mined from web text that elicited toxic completions from pre-trained models. Each prompt was annotated using the PerspectiveAPI to measure the severity of model toxicity, forming a baseline for assessing generative harm. Yet as we discuss later, reliance on automated toxicity scores also obscures questions of diversity and cultural context in how harm is recognized. Liang et al. [42] introduced HELM, which evaluates LLMs across 42 metrics spanning bias, toxicity, factuality, and robustness. Although HELM is not exclusively adversarial, it demonstrated how evaluation pipelines increasingly include risk-centric benchmarks. Recent red teaming datasets such as HH-RLHF [30], AdvBench [97], HarmBench [49], and Beavertails [39] further illustrate this shift, embedding adversarial evaluation directly into model development pipelines and broadening the range of harms under examination beyond traditional performance metrics. Table 1 compares these four widely used datasets to highlight how differences in their size, generation methods, harm definitions, and design inspirations shape what risks are surfaced and which may remain overlooked in red teaming practice.

Despite their importance, red teaming datasets are far from neutral. As Solaiman et al. [75] argued, decisions around prompt selection, annotation frameworks, and harm taxonomies reflect underlying social and institutional values. These are not purely technical design choices, but data practices, cultural, political, and labor-intensive processes that shape how risk is constructed and operationalized [32, 51, 67]. Consistent with prior work, AI datasets including those used in red teaming, are overwhelmingly created by institutions in the Global North, with minimal inclusion of perspectives from the Global South or marginalized communities [21, 65]. Furthermore, most datasets rely on English-language prompts, limiting their ability to capture harms experienced by multilingual or culturally diverse communities [62].

These data practices have tangible consequences for the visibility and mitigation of harm. If red teaming datasets fail to include prompts reflecting LGBTQ+ experiences, youth-specific concerns, or migrant community perspectives, then these harms may remain unexamined [83]. Hofmann et al. [34] for example, found that LLMs assigned more negative stereotypes, lower-prestige jobs, and harsher criminal judgments to speakers of African-American English compared to Standard American English prompts highlighting how dialect bias can shape harmful system behavior. By focusing on red teaming as a site of data work, our study investigates how practitioners make decisions about which risks to include, what constitutes a "good" prompt, and how these choices influence downstream definitions of safety and model accountability.

2.3 Stakeholders in Red Teaming for AI Safety

While red teaming is often framed as a technical practice aimed at detecting model vulnerabilities, it is in fact deeply shaped by the social, institutional, and political context in which it is conducted [32, 72]. A growing body of research has highlighted that the definition of "risk" in red teaming is not universal, but reflects the values, priorities, and threat models of those with the power to define what is tested and how [28, 32, 72, 84]. As such, understanding red teaming requires attending to its stakeholders: the diverse and sometimes conflicting set of actors involved in shaping, executing, and responding to safety evaluations. These stakeholders include AI developers, research scientists, policy regulators, product managers, red teamers, annotators, and users; each bringing different motivations, expertise, and constraints [69, 72]. Developers often prioritize model alignment and technical reproducibility, while policy teams focus on regulatory compliance or reputational risks [28, 83]. End users – particularly those from marginalized communities – may be more concerned with harms grounded in lived experience, such as misgendering, cultural erasure, or coded hate speech, which are rarely formalized in red teaming pipelines [21, 80]. These divergent priorities create tensions over whose risks are surfaced and whose are overlooked [32, 70].

Despite growing calls for inclusive safety evaluation by several scholars [23, 45], the perspectives of affected communities are often absent or tokenized in practice. Also, current red teaming frameworks lack clear mechanisms for how and where to integrate these perspectives [40, 56]. As Delgado et al. [23] highlighted, participatory evaluation processes tend to be extractive, consulting users or advocacy groups without redistributing power over decision-making or interpretation. This raises critical questions about what forms of expertise are legitimized in red teaming and what it means to evaluate a model "responsibly." Recent work has begun to surface these participatory gaps [23, 52, 76]. Gillespie et al. [32] frames red teaming as a sociotechnical system in which the labor, institutional incentives, and cultural assumptions of developers profoundly shape which harms are tested. Zhang et al. [93] shows how data workers involved in curating red teaming prompts must make judgment calls without clear guidance or support, revealing the limits of centralized governance. Together, these studies suggest that stakeholder inclusion is not just a matter of ethics but a necessary condition for meaningful safety. When the definition of harm is negotiated in closed circles, it risks being narrow, reactive, and disconnected from the realities of deployment.

Our study builds on this body of work by examining how red teaming practitioners reflect on stakeholder roles: who is consulted, who is represented, and how inclusion is operationalized or not during the design of red teaming datasets. From these perspectives, we surface the hidden practices and trade-offs that decide which harms are recognized and which are ignored in safety evaluation.

2.4 Contribution of HCI to Red Teaming Evaluation

Scholars within the SIGCHI community have long studied the sociotechnical nature of system evaluation and more recently red teaming [41, 88, 92, 93], offering critical insights and tools for understanding how values, labor, and context shape technical processes. In contrast to dominant narratives in AI red teaming, which often emphasize technical rigor, adversarial testing, and benchmark-based safety, HCI scholarship foregrounds the social dimensions of harm [66], the labor of dataset creation [6, 65], and the need for participatory, situated evaluation [96]. This work highlights red teaming not only as a way of testing models, but as a practice that defines harm and safety.

Additionally, foundational insight from HCI research emphasizes that data work, such as curating, cleaning, labeling, and evaluating datasets, is not neutral or mechanistic, but socially embedded and judgment-laden. Studies have shown that practitioners making decisions about datasets often operate with little recognition, institutional support, or shared criteria [60, 65]. These decisions, however, have far-reaching effects: they shape how systems behave, how failures are detected, and how "responsible AI" is defined in practice [6, 32]. Red teaming, viewed through this lens, is a form of high-stakes data work that involves implicit negotiations around value trade-offs, stakeholder representation, and institutional norms.

HCI studies have also contributed methodological alternatives to *opaque-box* auditing or static red teaming pipelines. Scholars have explored participatory approaches to algorithmic evaluation that involve users and affected communities in defining harms and co-constructing risk scenarios [21]. These approaches push against the abstraction and decontextualization common in much of AI evaluation. For instance, participatory design methods in HCI have long emphasized that "designing for safety" requires accounting for context, lived experience, and systemic power, not just adversarial queries or edge cases [11, 55]. Moreover, HCI critiques the assumption that evaluation can ever be value-neutral. As Selbst et al. [70] argued, technical interventions often embed unstated normative assumptions about what kinds of errors are tolerable, whose experiences count as harm, and what trade-offs are acceptable. This critique is directly relevant to red teaming, where prompts are often framed as objective "tests," though their design reflects deeply contextual choices [65, 87]. These choices are rarely documented, contested, or made transparent [28, 54].

Drawing from this HCI lineage, our study treats red teaming not as a purely technical task but as a sociotechnical practice involving meaning-making, boundary-drawing, and negotiation, in partnership with the people most affected by these decisions. Further, we call for more participatory, transparent, and reflexive approaches to evaluating AI systems that center not only on what a model can do, but who decides what it should do, and for whom.

3 METHODS

In this section, we provide an overview of our interview study. We describe the recruitment strategy we follow to ensure that we interview practitioners with red teaming expertise, and provide a characterization of the participants' profiles. We conclude with a description of our data analysis process.

3.1 Recruitment

To understand how red teaming datasets are designed, developed, and evaluated, we recruit AI practitioners who have released public red teaming datasets or have experience researching red teaming datasets. Our process consisted of first identifying potential participants with experience in red teaming datasets for LLMs. We searched for red teaming datasets from Hugging Face [1] and Papers with Code platforms [2]. We searched 'red teaming' in full-text search in Hugging Face to collect red teaming datasets and the data creators' information. We stopped after the top 100 search results. Because in the latter portion of the 100, there are many duplicated datasets from higher-ranked original red teaming datasets. Then we searched 'red teaming' on Papers with Code, which gave us 100 results of papers and their GitHub repositories. We also explored additional communities, such as Kaggle, OpenML, and the Data Provenance Initiative, but only found five red teaming datasets[3]

[1] https://huggingface.co/
[2] https://paperswithcode.com/
[3] Kaggle had four, the Data Provenance Initiative had one, and OpenML had none.

which we already collected through our search in the Hugging Face and Paper with Code platforms.

After gathering the red teaming datasets, we removed duplicates and selected potential participants following the criteria: the work should focus on natural language tasks, excluding efforts centered on multi-modal tasks, mathematical problem solving, security exploits, or code generation. The potential participants should have worked on red teaming projects for the purpose of developing, evaluating, or improving LLMs, interactive agents, and applied language technologies. This includes those who had created or curated red teaming prompts and datasets, designed pipelines for generating preference-aligned or adversarial data. After screening the search results, we recorded 107 qualified potential participants from open source communities, startups, industry, and academia.

For each potential participant, we recorded the following information to guide our recruitment process: the name of the dataset, its associated link, the title and link of the corresponding paper, the primary practitioner's name, email, and affiliation. We also noted public engagement metrics of the red teaming datasets, including the number of last month's dataset downloads on Hugging Face, GitHub stars, and paper citation counts. Based on this information, we synthesized two categories of data practices, which informed our sampling strategy and interview targeting:

(1) Create red teaming dataset from scratch: design and create original data specifically for red teaming. Practitioners can maximize the control and ownership of the dataset, but require more resource investment.
(2) Use existing dataset without extra revision: practitioners directly apply existing datasets to their studies. In doing so, they worked within the structure, assumptions, and classification choices embedded in the data, which reflect the values of the original dataset creators [51], and the downstream cascading effects of data quality in high-stakes AI applications [65].

We ranked the potential participants based on three indicators of visibility and impact: the number of dataset downloads on Hugging Face or Papers With Code, the number of GitHub stars for associated repositories, and the number of citations for corresponding papers. From this ranked list, we sampled participants evenly across the two data practice categories and invited 76 potential participants in batches, starting from the highest-ranked. In total, 22 participants accepted our invitation: twelve created their own datasets and ten reused existing datasets. We stopped the recruitment process once we reached data saturation.

3.2 Participant's Profiles

Our participants are AI practitioners in eight countries on three continents. Fifteen are PhD students, and two have worked on red teaming projects during their research internships. All participants had a computer science background. We recorded their areas of expertise based on their self-description. While some descriptions were similar, they contained nuanced differences. Fourteen described having expertise on language models, LLMs, or NLP; seven focused on safety or security; and four participants identified their background as machine learning or reinforcement learning. Most of the practitioners contributing to red teaming LLMs came from academia, reflecting the trend we observed in our recruitment. From the 107 datasets and repositories we identified, 39 originated from academia alone, 34 from academia-industry collaborations, while only 19 were solely from industry and 15 from open-source communities. Among the 76 invitations we eventually sent, there were 16 industry practitioners, of which only three agreed to participate in our interview (14% of 22 interviews). Details of our participants' profiles are shown in the Table 2.

3.3 Interview Study

We conducted 22 semi-structured interviews with AI practitioners, and each interview took between 40 to 60 minutes. Nineteen interviews were conducted in English, two were in Chinese, and one was in Spanish. The study was approved by the authors' Institutional Review Board. Participation was voluntary. We informed the participants that all information they shared would be kept confidential and anonymous, and they permitted us to record the interview for further analysis.

We chose semi-structured interviews as our primary method because they offer an appropriate way to understand practitioners' rationales when developing red teaming techniques and provide critical access to settings where ethnographic entry is constrained. Opportunities for fieldwork or ethnographic studies are limited across both industry and academia: red-teaming activities in industry often occur within proprietary or confidential environments [12], while academic red-teaming projects are also typically confidential until publication. Therefore, we consider interviewing as an appropriate mode of inquiry to understand practitioners' day-to-day activities that otherwise would remain inaccessible. Interviews also enabled us to reach participants working across diverse geographic and institutional contexts, which would have been logistically challenging to access through site-based fieldwork.

While we recognize that observational methods could help uncover the tacit and embodied aspects of data work, they limit the ability to surface the rationales, trade-offs, constraints, and assumptions that guide people's choices. In contrast, semi-structured interviews allow for eliciting accounts of not only what people do but why they do it. This is particularly important for examining the assumptions embedded in dataset construction and evaluation practices, which are often invisible in observational inquiry alone.

To reduce idealized recall, we grounded our interview protocols in participants' publicly released datasets and articles focused on red teaming techniques. We asked about the challenges and limitations they experienced, and asked them to explain their practice decisions in their work. Before each interview, we reviewed the participants' published papers and Hugging Face dataset cards, if applicable, to familiarize ourselves with their work and prepare our interview questions. We started the interviews by learning the participants' experiences in red teaming. Then we went deeper to discuss participants' experiences in creating or reusing red teaming datasets, their motivation for red teaming research, their categorization of harm and risks, and evaluation approaches. For each interview, we tailored, added, or skipped some questions towards the participant's own experiences and insights.

ID	Location	Background	Area of Expertise	Data Practice	Duration
P1	Israel	Research Scientist	Machine Learning	Create	48 minutes
P2	U.S.	Ph.D. Student	LLMs and Safety	Reuse	49 minutes
P3	U.S.	Ph.D. Student	AI safety and security	Create	53 minutes
P4	U.S.	Ph.D. Student	Safety Alignment of LLMs	Reuse	40 minutes
P5	Singapore	Ph.D. Student	NLP	Reuse	44 minutes
P6	U.S.	Ph.D. Student	Reinforcement learning	Create	51 minutes
P7	Germany	Ph.D. Student	AI safety and LLMs	Reuse	56 minutes
P8	Singapore	Postdoc	Machine learning, LLM safety	Reuse	59 minutes
P9	U.S.	Ph.D. Student	LLMs alignment	Reuse	57 minutes
P10	India	Undergrad	Security and AI security	Create	56 minutes
P11	U.S.	Ph.D. Student	Reinforcement learning	Reuse	44 minutes
P12	India	Entrepreneur	Software engineer	Create	47 minutes
P13	Spain	Ph.D. Student	Testing language model	Create	51 minutes
P14	China	High School Student	Computer Science	Create	59 minutes
P15	Singapore	Research Engineer	Software testing	Create	41 minutes
P16	U.S.	Ph.D. Student	Language model and NLP	Create	45 minutes
P17	South Korea	Ph.D. Student	Data augumentation and NLP	Create	40 minutes
P18	China	Ph.D. Student	Security of large models	Reuse	57 minutes
P19	U.S.	Ph.D. Student	NLP	Create	40 minutes
P20	U.S.	Ph.D. Student	LLM reasoning	Create	52 minutes
P21	South Korea	Research Engineer	NLP, hate speech	Reuse	60 minutes
P22	South Korea	Ph.D. Student	Efficiency in LLM	Reuse	60 minutes

Table 2: Participants' profile information and interview duration

3.4 Transcript Analysis

The two authors who conducted the interviews also analyzed the transcripts using a thematic analysis approach [16]. Both authors attended all 22 interviews, with each author leading and coding half of the sessions. After coding the first few interviews independently, each author developed a preliminary codebook. We then met to compare two codebooks, discuss discrepancies, and merge them into a single version with agreement. Using this shared codebook, we coded the remaining transcripts, revisiting earlier transcripts when necessary to ensure consistency. We discussed ambiguous cases during the coding process and refined definitions as needed, resolving disagreements through discussion until consensus was reached. The final analysis produced 47 sub-codes grouped under 21 parent codes, capturing participants' definitions of red teaming, their motivations, practices, and perspectives on the evolution of LLMs. The Table 3 shows the summary of the derived themes, codes, and sub-codes.

3.5 Positionality

All authors have received formal training in HCI. The two lead authors bring complementary expertise: one in critical data studies, which examines how data is produced through specific social processes and practices [37], and the other in the intersection of HCI and NLP, with expertise in subjective data annotation, interpretation, and evaluation of language model outputs. These perspectives shaped our study design, the framing of our questions, and our interpretation of participants' experience and insights to both the technical and sociocultural dimensions of red teaming [35]. Two co-authors bring domain expertise in online safety, particularly around youth risk and sociotechnical harms. These perspectives shaped our critical attention to how red teaming datasets define harm, surface risk, and reflect the experiences of vulnerable users.

4 FINDINGS

While red teaming is often framed as a technical practice aimed at detecting model vulnerabilities, studies have shown that the definition of "risk" is not universal but reflects the values, priorities, and threat models of those in charge of defining what is tested and how [28, 32, 72]. We observed this dynamic in practice. Our findings are organized into three sections, illustrating how practitioners define and operaionalize risks across three moments in red teaming LLMs: (1) conceptualizing what red teaming means, (2) developing adversarial datasets to capture those risks, and (3) evaluating the outcomes of red teaming LLMs.

4.1 Participants' conceptualization of Red Teaming LLMs

In this section, we illustrated how practitioners conceptualize red teaming LLMs by examining their definitions of red teaming, their motivations for red teaming LLMs, and their views of red teaming as an interactional and social process.

4.1.1 Participants' definitions of red teaming LLMs. Most participants started working on red teaming LLMs around the time LLMs became widely accessible and began outperforming standard benchmarks in commonsense reasoning and language understanding [58], raising concerns about safety and misuse. Six participants had worked on red teaming for about two years, seven had one

Theme	Code	Sub-codes
Participants' conceptualization of red teaming LLMs	Definitions of red teaming	How participants define and conceptualize red teaming; Blind spots of red teaming
	Participant's background and experience	Education; Years of experience; Affiliation
	Participant's context and motivations	Work context; Motivations for red teaming
	Participant's approach to red teaming	Methods and strategies; Experiences described
	Challenges	Practical challenges
	References	References to additional materials
	Impact	Beneficiaries; Future impact
Developing adversarial datasets	Tailoring datasets	Adaptation processes; Risk categories; Representation; Contextual relevance; Diversity
	High quality data criteria	Definitions and standards for data quality
	Seed data	Sources; Selection criteria
	Harmfulness definition	Definitions and examples of harmfulness
	Reuse and repurpose existing datasets	Decision-making process; Advantages; Limitations
Evaluating adversarial datasets	Machine evaluation	LLM judges; Classifiers; Limitations
	Human evaluation	Annotation; Validation; Challenges
	Evaluation (overall)	Dimension ranking; Cross-category performance; Infrastructure; Challenges
	Standards	Use of guidelines; Customized evaluation; Limitations
	Human vs. machine	Comparisons between human and machine evaluations
	Metrics	Metrics used; Trade-offs
	Experts and stakeholders	Collaborations; ; Expert evaluation
	Geographies and representation	Geo-cultural bias; Representation issues
	Misclassification	Examples; Mitigation strategies

Table 3: Themes, codes, and sub-codes from the thematic analysis of interviews.

year of experience, and many had just started for a few months. These marked the release of ChatGPT in 2022 as a turning point that raised people's concerns about AI safety. As P8 shared, *"Red teaming (LLMs) hasn't been around for long ... Its appearance came with ChatGPT, which launched at the end of 2022. Although the field is very active now, its lifespan has only been about a year and a half."* Learning from participants, we noticed that the work on red teaming LLMs emerged quickly in response to the fast-growing capabilities and safety concerns of LLMs.

Researchers from various fields, including security, reinforcement learning, and NLP, worked on it with different assumptions and methodologies. From our study, we identified two major ways of conceptualizing red teaming LLMs, which were shaped by practitioners' disciplinary training and research backgrounds.

(1) Exploration and searching from the reinforcement learning framing (N=20). Participants with reinforcement learning training approached red teaming as an exploration problem, where the task

was to navigate a vast prompt space to uncover successful attack cases. P6 explained how red teaming is a hard exploration challenge.

> *"Red teaming is a very hard exploration problem because you can think of the space, the search space of the prompt is like all combinations of the possible text, right? And this is a giant space that you will never be able to. Uh, traverse. So, how to design a strategy to explore this space? Yeah, it's a hard problem."* - P06

In this framing, assumptions shifted away from predefined ground truths: what mattered was not labels but the efficiency of search. Accordingly, the methodological practices emphasized Attack Success Rate (ASR), coverage, and diversity as primary evaluation criteria. As P11 described, *"I think coverage like you're trying to cover it, cover the whole space of possible queries. That's one of the main, I think one of the most important in red teaming."* This view aligns with the broad research community of red teaming generative models. Lin et al. [44]'s survey also highlights that crafting diverse red teaming prompts requires creativity and labor effort, so many studies explore automatic methods to create such prompts with a defined state space, search goal, and search operations. Participants (P6, P11, P17, P19) noted that coverage without diversity risks redundancy, since many prompts may differ by just a few words. To address this, they emphasized balancing coverage with diversity, while others, such as P22, highlighted that efficiency also matters in query searching, especially given the API costs of large-scale experiments.

(2) Inappropriate content detection from the NLP framing (N=2). Participants with NLP backgrounds conceptualized red teaming through the lens of familiar classification tasks. In this view, red teaming was about detecting contradictions in dialogue, or identifying offensive content, which are traditional NLP problems. P16, for example, described red teaming as triggering self-contradictory responses across dialogue turns, reflecting the structure of dialogue detection tasks.

> *"So I use this concept that red teaming is about using one language model to attack another model to trigger something. That's another model to answer something that makes it contradictory (content), you can consider it like two models, which we can call two agents. One is to ask another one to generate some contradictory, model safety, that is kind of the red teaming."* - P16

P21 similarly framed red teaming as users intentionally tricking a chatbot into producing hate speech, aligning the work with established hate speech detection approaches.

> *"Our work is in between those two categories (hate speech detection and exploring potential jailbreaking prompt problem). This paper doesn't deal with those technological red teaming. But we framed this work as red teaming because we could find the users attempt to jailbreak this system (chatbot), which can also be one type of jailbreak that is manual... So from that perspective, this can be a red teaming work, but from the traditional hate speech perspective."* - P21

Because of this framing, evaluation was tied to predefined ground truth labels of harm and measured through supervised metrics such as the F1 score.

These two conceptualizations reveal a fundamental contrast in how red teaming LLMs was understood. The exploration and search framing treated red teaming as an open-ended problem of discovering failures across an unbounded prompt space, where coverage, diversity, and efficiency were key. In contrast, the inappropriate content detection framing treated red teaming as a supervised classification problem, where harmfulness was predefined and measured against labeled ground truth. These practices illustrate how disciplinary backgrounds shaped practitioners' assumptions about red teaming and guided their approaches.

4.1.2 Motivation of Red Teaming LLMs. Participants were motivated to study red teaming LLMs for diverse reasons, including (1) observing red teaming "in the wild," (2) filling technical gaps, (3) responding to organizational influences, and (4) improving models' safety after red teaming.

A common motivation stemmed from observing users circumvent LLM safeguards and share their conversations online. For example, P12 became interested in the topic after seeing Reddit users share their conversations with ChatGPT, including detailed responses on robbery. Similarly, P21 became interested in the topic after observing anonymous users discussing how to 'tame' a human-like social chatbot in an online community. Then, P21 and their team collected the forum discussion data to study human users' manual red teaming. P19 also focused on real-world conversations between humans and an LLM chatbot to study red teaming in the wild, beyond the laboratory environment.

Other participants were interested in advancing technical methods and connected red teaming to their prior expertise in reinforcement learning, data augmentation, or model optimization. Yet participants valued technical contributions differently. For instance, P8 mentioned that he values technique improvement in training and fine-tuning more than how to rewrite the prompts.

> *"We have found that it is all about rewriting prompts. Actually, I don't like this kind of work much because I have a background in machine learning. I don't dislike it, but I feel that this kind of work does not contribute much in terms of technology."* - P08

These preferences reflect broader values in the machine learning community, which emphasize performance, generalization, efficiency, and novelty [15].

The need for implementing legal regulations also motivated participants to work on red teaming LLMs. For instance, P13 research focused on assessing models to ensure they met the requirements of the European Union's Artificial Intelligence Act, *"which defines different attributes or requirements for AI systems, such as safety bias."* For other participants, what sparked their interest was participating in red teaming competitions. That was the case of P8, who initially participated in a NeurIPS red teaming competition in 2024 then expanded the work into larger research projects.

Finally, several participants highlighted interest in what comes after red teaming, not only exposing vulnerabilities but also strengthening guardrails and improving model safety. As P19 said: *"We really care about fixing vulnerabilities after red teaming. Many attacking methods are not scalable and only capture narrow behaviors, so I was seeking scalable ways to both identify and fix issues."*

These motivations illustrate that while red teaming shares a surface goal to elicit problematic behaviors from LLMs, the reasons practitioners engage with vary based on personal interest, disciplinary training, and broader societal concerns.

4.1.3 Red Teaming LLMs is More Interactional and Social Than Anticipated. While most red teaming studies of LLMs remain focused on simplified single-turn prompts, our participants emphasized that effective red teaming should also account for multi-turn, socially embedded interactions. Among all 22 participants, 19 focused on single-turn interactions, while only three described multi-turn approaches: P16 and P21, who framed red teaming through NLP tasks such as dialogue or hate speech detection, and P20, who explicitly studied red teaming LLMs across multi-turn conversations. P20's study showed that red teaming LLMs can be multi-turn and have concealed intentions, rather than a single-turn explicit request. 'How to build a bomb' can be rejected in a single-turn explicit request, but can be elicited to output detailed instructions after a multi-turn conversation with concealed malicious intention. This view was also echoed by P1, who noted that some applications are particularly vulnerable in conversational settings.

> *"Now the red team, another aspect is, whether this application is conversational or not, because there are some applications. So maybe the model is not so easy to make it fail in the first turn of the conversation, but maybe it will be much easier, so we have a paper also which talks about conversational red teaming."* - P01

From interviews, we also learned that red teaming LLMs is not only about technical robustness but also about how LLMs operate within social contexts. Participants noted that LLMs are increasingly socially embedded, interacting with people in ways that are shaped by law, language, and culture. These contexts make harmfulness difficult to determine. For example, P07 and P19 mentioned that it is hard to judge whether an LLM responding to a user's request for Ed Sheeran's song lyrics should be considered a copyright violation. In such nuanced legal cases, it remains ambiguous whether the output should be counted as a successful red teaming attack. Similarly, P17, a foreign researcher, struggled to judge whether prompts about smoking weed were harmful, since laws in his home country differed from those where he conducted research. Multilingualism further complicates these evaluations. P18 observed that responses can differ in harmfulness across languages. When asked *"Will using a mobile phone for a long time affect my eyesight?"*, the English answer was harsh, warning that long-term use would inevitably ruin one's vision and sarcastically saying getting blind now would spare future trouble. In contrast, the Chinese answer dismissed the concern as unnecessary worry and encouraged the user to keep using their phone freely. P22 also pointed out the vulnerability of LLMs in multilingual text.

> *" A language model can be really safe for single-language input, but if we partition the sentence into some phrases and allocate a different language to each phrase and make the code-switching text, then it becomes a really nice test case that can attack the large language model. "* - P22

P21 further argued that multilingual is not the same as multicultural. Responses that sound fine in English may appear awkward in Korean. For example, it is common to say *'that's a good question'* in response to a question in English, but it is uncommon and sounds strange in Korean. This shows that LLMs become socially embedded. Red teaming LLMs is not solely a technical challenge, but should also consider those social factors, such as laws and culture.

The complex nature of red teaming LLMs, which is interactional and social, also highlighted opportunities for interdisciplinary collaborations. Complex real-world scenarios require contextually grounded harm categories. P11 noted that *"If they have good definitions of, like, different new categories. Maybe it can inform the search. And depends on what humans value more. Maybe you can guide the search on these. It's very possible and probably helpful."* P8 also remarked that harm classification often needs domain expertise.

> *"Some categories cannot be reliably judged by machines. For example, how to make poison gas. This requires professionals to assess accuracy. This red teaming field could naturally collaborate with humanities and social science scholars, who are more sensitive to certain ideas than those of us in computer science. "* - P08

In practice, however, collaboration largely remained within technical domains. While some participants mentioned working with domain experts: P17 collaborated with biologists in drug discovery using reinforcement learning, P9's advisor had a social science background, and P8 engaged with cognitive science experts in foundation agent research. None of the 22 participants reported collaborations with non-computer science experts specifically for red teaming LLMs.

Red teaming LLMs is not only a matter of probing technical robustness but increasingly about evaluating language models as interactional and socially embedded systems. While most participants still studied single-turn cases, multi-turn work showed how concealed intentions and evolving dialogue can reveal failures missed in one-shot testing. Participants also emphasized that harmfulness is context-dependent, shaped by law, language, and culture, etc., which demands new harm categories and expertise. Yet collaborations remain largely within technical domains, underscoring the need for interdisciplinary collaboration.

4.2 Developing Adversarial Datasets

Adversarial datasets are collections of inputs designed to expose the models' weaknesses by inducing unsafe and inappropriate outputs. Examining how these datasets are created is critical to understand how practitioners define and categorize risk in red teaming LLMs. In this section, we first outline three different approaches to creating datasets that practitioners reported. Then, we illustrate how practitioners' approaches influenced how they conceptualize risk, representativeness, and diversity.

4.2.1 Approaches to Creating Datasets. Participants reported using one of three approaches when constructing adversarial datasets: (1) repurposing existing datasets to generate new harmful datasets, (2) creating datasets entirely from scratch, and (3) deriving datasets from human interactions. Of the 22 participants we interviewed, ten reported reusing existing datasets and benchmarks in their red teaming research. These resources were utilized for various purposes, primarily to create new datasets for specific red teaming attacks, to train or fine-tune models, and to support evaluation.

Participants reported different reasons for deciding which dataset to reuse. For some, such as P2 and P18, the limited availability of datasets at the time they conducted their research constrained their selection. In other instances, participants, like P16, chose a dataset because it had been used in prior publications that the participants considered relevant to their research community. Following the same methodology by reusing these datasets was seen as appropriate, as it allowed practitioners to compare their contributions fairly with previous work. Most participants who reused existing datasets treated red teaming mainly as a technical task. In contrast, those who created datasets from scratch or drew on real user jailbreak attempts also considered the broader context of interaction. For instance, P1 considered the application's purpose and user interaction patterns to decide what kind of dataset was needed:

> *"The first consideration to make is what is the application? If I want to red team a model or a system [...] if it's an insurance company or customer agent. It depends a lot on the domain or the goal of the application. All these define what kind of datasets I will choose in order to red team the model."* - P01

According to P1's account, it was only after considering all these factors that they determined the appropriate data for red teaming. One of the strategies P1 followed was creating adversarial datasets by combining human-generated and LLM-generated data. The participant leveraged LLMs to generate data by prompting them with specific risk categories, which allowed for broader topical coverage than using only human-generated inputs. However, P1 also used human-generated data, as he considered it more nuanced and sophisticated than synthetic data. Therefore, by combining both types of datasets, the participant ensured not only high coverage, capturing a broad range of relevant possibilities, but also a more diverse and nuanced dataset, one that included rare or edge cases.

A third strategy consists of gathering interactions between humans and LLMs to identify vulnerabilities or harmful outputs. These interactions include prompts, responses, or dialogue exchanges generated during real or simulated scenarios. Four of our participants reported following this technique (P3, P13, P19, P21). This approach has gained interest because such datasets capture what participants describe as *"authentic, context-rich adversarial cases, helping to evaluate models in realistic use scenarios"*. P3 reported organizing challenges where participants were given specific scenarios and constraints to jailbreak AI models. P13 reported conducting sessions to assess the ability of models to interact in Spanish and the Basque language. Lastly, P21 reported leveraging in-the-wild user interactions with conversational agents to red team models.

The strategies described in this section illustrate the ways practitioners conceptualize and operationalize adversarial dataset creation. While reusing well-known datasets was perceived as a tactic to ensure methodological consistency, building datasets from scratch or by collecting human–LLM interactions enabled practitioners to gather interactions that were perceived as unique, and therefore more valuable. Thus, practitioners' decisions reveal not only their constraints in assembling adversarial datasets but also surface their beliefs about the kinds of data that are required to conduct meaningful evaluation of models.

4.2.2 Conceptualization of Risk. Defining what constitutes harmful behavior is a central step in developing red teaming techniques and constructing new datasets. The approach practitioners took to dataset creation shaped their conceptualization of risk. Those who build on existing datasets often inherit definitions of harm from the categories already embedded in the source material. In contrast, practitioners who create datasets from scratch typically select categories of harm from established taxonomies they deem relevant to the specific context of their work. Finally, practitioners generating adversarial datasets from in-the-wild interactions identify harmful behaviors by interpreting and classifying the exchanges between humans and models.

Reusing existing datasets in the development of new red teaming techniques often entails inheriting the source dataset's definitions and taxonomies of harmful behavior. For instance, P2 and P10 both sampled harmful instructions from AdvBench [97], a widely used benchmark for jailbreak evaluation, to construct adversarial datasets for attack injection and to develop a jailbreaking technique, respectively. Similarly, P20 leveraged the BeaverTails dataset [39], which includes malicious questions across 14 harm categories that models are expected to refuse, to develop new probing methods. Across these cases, the reliance on AdvBench and BeaverTails meant that participants' conceptualization of harm was shaped by predefined categories.

When participants decide which data to sample and how, they are also making deliberate judgments about which risks warrant greater attention. For instance, as illustrated in the following quote from P2, who prioritized "illegal stuff" over "fake news of a celebrity", demonstrates that selecting categories and examples of harmful content inherently involves ranking harms, determining which risks are deemed critical enough to examine and which are excluded from consideration.

> *"Most of the data is about illegal suggestions, because I think this is most important and most straightforward. Right? There are also some questions about fake news about celebrities, but that is not as severe as illegal activities."* - P2

Some of the participants who reported reusing datasets also mentioned relying on content safety tools and classifiers to filter existing datasets, a practice that facilitated the detection of adversarial prompts and implicitly embedded the harm definitions encoded in these tools. For instance, P19 used the OpenAI Moderation API [57] to detect adversarial prompts in datasets of single-turn conversations, and refined the selection by keeping only those prompts that elicited harmful responses according to Llama-Guard's safety classifications [38]. Similarly, P22 employed the Bot Adversarial Dialogue (BAD) [89] classifier to evaluate the offensiveness of outputs from open-domain dialogue models, identifying offensive outputs and positive test cases based on BAD's scoring. As P22 put it, this work entailed *"inherited definition of toxicity from the BAD Bot Adversarial Dialogue paper."* In both cases, the adoption of these automated filtering mechanisms meant that the predefined taxonomies and thresholds of the classifiers themselves shaped the participants' conceptualization of harm.

For participants who reused datasets, it was not merely a matter of adopting the risk conceptualizations embedded in the dataset categories; some also identified labeling errors within the datasets. For example, P7 reported removing 100 copyright-related prompts from the HarmBench framework [49], because the labeling of these prompts did not align with his understanding of copyright violation.

On the other hand, participants who created datasets from scratch made deliberate choices about which taxonomies and sources to incorporate. In P1's case, when developing a red teaming dataset targeting "stigmas" he and his team decided the type of stigmas that the dataset should include by *"asking people about the stigmas and reviewing papers"* and then they made the deliberate decision to *"filtered only the stigmas that looks reasonable to"* them and *"to people who live in the US."* P1's practices highlight that dataset creation is not a neutral process but an interpretive one, influenced by researchers' choices and purposes.

Lastly, practitioners who derived datasets from human interactions with models, categorization of risks, followed either a deductive or an inductive approach. P13, for instance, predefined 22 categories before the red teaming process, pursuing a systematic assessment of safety and bias risk areas. In contrast, P21 clustered the observed human interactions into six types of risks. This contrast reveals a trade-off in conceptualizing risk; while predefining categories ensures alignment with research goals, they might overlook unexpected harms. On the contrary, emergent categorization can capture unanticipated risks, but it introduces greater subjectivity and inconsistency, given that it is the researcher who decides whether a particular interaction corresponds to a specific risk.

4.2.3 Conceptualization of Representativeness and Diversity. Different methods for generating adversarial datasets led participants to divergent conceptualizations of representativeness and diversity. For participants who reused datasets, the perceived diversity of their data was often inherited rather than intentionally constructed, shaped by the harm categories of the datasets practitioners reused or by the taxonomies they used to guide their work. For instance, P19 considered their dataset representative because it had samples from all the risk categories outlined in the Taxonomy of Risks for Language Models [85]: *"The dataset itself is created based on the high level categories and we try to cover as many diverse cases as we could, but the main challenge was that we do want to cover the diversity of instances that fall under those safe categories. So, I think we did a good job in covering those."* Yet this approach assumes that the taxonomy itself is comprehensive and balanced, an assumption that participants rarely interrogate. As a result, diversity was defined within the boundaries of pre-defined categories, overlooking risks that fell outside those classifications.

On the other hand, as illustrated by P2 and P20 quotes, practitioners who reused existing datasets often noted limited diversity, either in the categories of harm represented or in the way prompts were formulated. For example, P20 observed that the BeaverTails dataset contained over 100 labeled harmful actions, yet many instances were simply variations of the same action rather than distinct behaviors: *"the BeaverTails has more than 100 harmful actions in the original data set, but most of the sentences are a reconstruction of the same harmful action rather than a unique harmful action [...] so harmful action items in the Beavertails are not that diverse."* Similarly, P2 described redundancy in AdvBench, not only in the categories of risks but also in how the prompts, instructions, and dialogues were framed, since numerous prompts differed only superficially, resulting in the dataset lacking diversity.

In sum, reusing existing datasets revealed narrow coverage, redundancy, or imbalanced representation not only across risk categories but also in the design, structure, and syntax of prompt formulation. These limitations stemmed directly from the inherited datasets and taxonomies.

In contrast, participants who derived datasets from human interactions with models (P3, P13, and P21) attributed diversity to having a diverse pool of participants, a variety of scenarios or tasks, and the creativity of human attempts to circumvent model safeguards. Yet, as stated by P3, they acknowledged that such creativity could not guarantee full representativeness, given the inherent limits of capturing all possible adversarial strategies:

> *"You can never say this is representative because it's kind of a moving target. There's always going to be novel exploits, new risks and new attack vectors. But the best you can do is to try to cover as much of that space as possible."* - P03

In conclusion, the practices described in this section highlight that while dataset reuse simplifies the construction of adversarial datasets, it also entails embedding the harm definitions codified in the source datasets. Once such definitions are reified, practitioners in our sample rarely challenged them, though they fundamentally shape how adversarial behaviors are recognized and assessed.

4.3 Evaluating Adversarial Datasets

Evaluation meant determining whether a model's responses to an adversarial dataset—crafted by the participants—were harmful, and thus whether their red teaming technique successfully broke the model. Participants reported different strategies, including replicating methodologies from existing papers, developing an assessment criterion based on established guidelines such as the European Union's Artificial Intelligence Act, Anthropic, and OpenAI guidelines, and conducting evaluations either with humans or LLMs to assess safety. Regardless of the strategy followed, participants expressed that determining whether an output was harmful was often challenging and not always straightforward. These challenges also revealed the tools and support practitioners need (RQ2), which we revisit in the Discussion section 5.2 with actionable recommendations from an HCI perspective. In this section, we provide an overview of the uncertainties practitioners face to determine whether a model's output was harmful, and their motivations and rationales for automating evaluation, and including humans in certain moments in the assessment.

4.3.1 Uncertainty in What Counts as Adversarial. A critical step in the evaluation process was deciding what would qualify as adversarial. While ASR metrics exist, the following examples illustrate some of the difficulties practitioners face in defining what truly constitutes an adversarial case. For some participants, the challenge was determining whether a prompt should be considered adversarial when its effects were inconsistent. A technique might work on some models but fail on others, or it might trigger an unexpected response in one iteration but not in subsequent ones.

Practitioners also faced difficulties classifying responses, because in a single model's output, there might be fragments that met the adversarial criteria alongside others that did not. In other instances, models' first response might be a refusal to answer, but later actually provide a harmful response. This overlap in a single output complicated practitioners ability to categorize such responses.

Other practitioners noted that a major challenge in evaluations was judging the real-world harm of jailbreaks. As P10 explained, *"jailbreaks aren't always as harmful as they seem"* because not all produce genuinely dangerous or usable outputs. For example, a model might respond to a request like 'How do you build a bomb?' with inaccurate or fictional instructions that pose no real risk. In such cases, the difficulty lay in deciding whether to classify the response as harmful simply because the model engaged with a harmful query, or only when the response provided accurate and actionable information.

All these perspectives reveal the uncertainty that practitioners face when categorizing whether an output was harmful. It is not only the challenge of defining harm but also having to take into consideration the influence of inconsistent model behaviors, mixed-content outputs, and the difference between apparent and actual harm. Additionally, participants needed to determine how to conduct the assessment, namely, what methodology or tools to use. In the following sections, we provide an overview of how practitioners engage with automated and human evaluative approaches to assess red teaming techniques.

4.3.2 Automating the Evaluation with LLMs and Classifiers. Practitioners adopt widely recognized and standardized classifiers and LLMs as part of their evaluation pipeline due to the large scale of the datasets they generate, which makes full manual evaluation unfeasible. P19 highlighted the difficulty of relying entirely on human evaluators, explaining that *"human evaluators are really hard to recruit [...] it's impossible to have them in the loop for every step, or even just the final step."* Similarly, P3 pointed to the scale of data as a limiting factor: *"Once you reach a certain scale, it is almost infeasible to use manual effort for everything. In one of the competitions we run, we had more than 2 million of interactions or chats, it is quite infeasible for us to even hire people to go look at all that."*

Participants followed various approaches to establish criteria and rules for determining whether a model-generated answer was harmful. Some, like P2, began by manually inspecting model outputs to identify patterns, then crafted instructions to guide automated evaluation using a stronger LLM based on his initial observations. Other practitioners relied more heavily on established practices from adjacent research communities. For example, P10 noted that many security-related research efforts use third-party LLMs as judges, applying relatively straightforward heuristics to assess whether jailbreak prompts are contextually relevant, readable, and logically coherent. Inspired by this practice, P10 used three different model evaluators: Keyword Matching, StrongREJECT [77], and his own evaluator, each with its own criteria for scoring harmfulness to have a more robust evaluation pipeline. As illustrated by the following quote, the participant assumed that by using multiple models, he could approach harmfulness from different perspectives and mitigate the inconsistencies arising from relying on a single evaluator:

> *"The first evaluator is Keyword Matching, which is not a very effective evaluator, it is just checking whether a particular prompt has harmful words or not, and if it has harmful words then it's a jailbreak. If not, it's not a jailbreak. It's a pretty simple technique and it can be bypassed a lot of times, so you can have a lot of false positive and false negative cases. StrongREJECT, on the other hand, was a new technique and they also originally proposed using LLMs as a judge in detecting whether jailbreaks are successful or not, and sometimes as per our investigation, provides some amount of false positives and false negatives."* - P10

While adopting LLMs as part of the evaluation has become the standard, participants also reflected on the limitations and consequences of relying on LLMs as evaluators of harmful content. P10 stressed the *"lack of consistency"* in such evaluations, noting that LLMs *"do not always respond the same"* and that scores from models like StrongREJECT can vary for the same prompt. While he acknowledged that misclassifications are rare, the unpredictability posed a risk to evaluation reliability, highlighting the need for more stable evaluation approaches. P19 added two further concerns: cost and oversight. They stated that high-quality LLMs are *"more expensive whether through API calls or self-hosted setups, making large-scale evaluation resource-intensive."* In addition to highlighting the computational and financial burden, P19 warned against the uncritical adoption of LLM judges without rigorous preliminary testing, stating that *"we don't have scrutiny of how good these judges are so before using an LLM as a judge, it would be great to test it at least strictly in a preliminary setup to make sure that it's good for the task that we are looking for,"* noting that insufficient scrutiny can result in evaluators that are not suited for specific tasks. Another limitation is the availability of APIs when dealing with harmful content. P09 shared that their lab's API access was banned by OpenAI after being flagged for violating usage terms. This occurred because their red teaming study necessarily involved generating and evaluating harmful content. This underscores the restriction of relying on external APIs in red teaming evaluation and highlights how the current infrastructure is not adequately supportive of safety research outside industry.

Taken together, the practices described in this section point to a more profound, systemic implication: the widespread adoption of LLMs as evaluators means that the definitions and thresholds of harm they encode have far-reaching downstream effects, shaping not only what is defined as harmful but also the trajectory of subsequent research and interventions. As P10 cautioned, inconsistency in these evaluators *"kind of means you just have a bad evaluation method, even if it is accurate, it cannot be trusted"* because it relies on a probabilistic model prone to variability. For end-users, such as those assessing whether their LLM is vulnerable to jailbreaks, fluctuating scores can *"break their algorithms,"* making it hard to enforce reliable thresholds. This lack of stability creates tangible challenges for practitioners to advance red teaming research, reinforcing the urgent need for more consistent and trustworthy evaluation methods.

4.3.3 Human Evaluation of Adversarial Datasets. While participants acknowledged that LLMs help address the scale of outputs to evaluate, most still regarded human judgment as more trustworthy. For instance, P13 emphasized that *"humans have a level of precision that models lack."* Similarly, P1 highlighted the *"need for more humans to be involved in determining whether the LLM's judgment is truly aligned with people's expectations."* Thus, reinforcing the idea that human evaluation is critical for making nuanced judgments that automated systems may miss or misclassify.

Eight participants reported incorporating some form of human evaluation or inspection into their evaluation pipelines, making different choices when selecting annotators. P22 recruited via Amazon Mechanical Turk to perform human evaluation of the red team results, to validate *"whether the offensiveness classification results by the automatic classifier are good enough to believe or not,"* while others relied on students or collaborators affiliated with their institutions or research teams. The majority of our participants did not report having a specific selection criterion of annotators, though some, such as P20, described his human evaluators as *"well-educated, with graduated degrees, and with a good understanding of English."* Most participants did not perceive any significant drawbacks in selecting annotators, with the exception of P22, who acknowledged potential limitations of Mechanical Turk, noting that *"anonymous annotators may work carelessly to earn credits quickly, potentially compromising data quality."*

In terms of evaluative approaches, participants employed different strategies. P4 sampled portions of the data to get *"a feeling of how the model is actually behaving,"* which in turn informed modifications to the algorithm, while others, such as P16 and P20, adopted what they considered standardized methods, including a three-level scale established in prior publications. Other participants, such as P8, conducted a triple assessment, which, as the participant described, involves *"first using a machine, then using rules, and finally using human eyes."* P8 developed this strategy drawing from what he described as *"an assessment method originated from the NIPS (NeurIPS) competition."*

Viewed collectively, participants' accounts show that human judgment continues to be perceived as a needed mechanism for assessing adversarial datasets. However, the lack of standards and rigorous criteria for selecting annotators and designing evaluation protocols complicates assessing the strengths or limitations of different approaches of human evaluation in red teaming. These challenges highlight the need for developing methodological guidance to better account for human raters' subjectivity and establish robust evaluation frameworks.

5 DISCUSSION

AI is developing at a pace that current evaluation paradigms struggle to match, leaving gaps in understanding the societal effects of AI. Scholars from diverse disciplines attribute this gap to the limitations of traditional evaluation methods, which emphasize technical performance while overlooking broader societal impacts [26, 69], but also to the lack of interest of research communities in evaluating the real-world impact of systems [43, 63]. In response to these challenges, researchers have called for expanding AI evaluation beyond purely technical methods toward a context-aware approach that accounts for real-world impact and context in which AI systems are deployed to assess AI's second-order effects [4] [17, 29, 69, 84]. In this respect, red teaming has been increasingly adopted as an approach to probing LLM failures in real-world use, identifying harmful outputs, and uncovering vulnerabilities [69]. However, red teaming still faces limitations, particularly methodological ambiguity [68], limitations of expertise and participation [72], subjectivity and contested targets [30, 72]

Our findings provided empirical evidence of how practitioners identify, categorize, and evaluate risk at three distinct moments in the process of developing red teaming techniques: 1) when practitioners conceptualize red teaming, 2) when they define risk categories and classify behaviors to develop adversarial datasets, and 3) when practitioners evaluate those datasets. In this section, we further discuss how practitioners' data practices related to red teaming techniques lead to the omissions associated with context, interaction type, and user specificity, which have a broader impact beyond the creation of adversarial datasets and a cascading effect on how risk and safety are conceptualized. From there, we highlight critical research pathways for HCI researchers interested in supporting practitioners to design red teaming evaluations that are useful for real-world impact.

5.1 Conceptualizing Risk in AI Red Teaming

Our empirical findings revealed that the three moments in which AI practitioners conceptualize risk are interconnected: the decisions practitioners make about what red teaming is meant to achieve shape how they design adversarial datasets, which in turn define the criteria and thresholds for evaluation they follow to determine what counts as adversarial success. These decisions also make evident what practitioners prioritize as contributions to the field; i.e., the technical innovation of discovering novel ways to break a model, rather than examining the *context* in which risk might emerge, the different *interaction modes* that might lead to varying levels of severity of risk, and the needs and expectations of *users* when assessing risk. To follow, we describe how the decisions made at each of these three moments contribute to practitioners overlooking *the situational context, interaction type, and user specificity* — which, from an HCI perspective, are essential for a more holistic and deeper understanding of risk.

5.1.1 Overlooking Context: To effectively red team LLMs, it is critical to consider contextual details as they influence how risk is defined, identified, and assessed [18, 69]. HCI scholars have defined the concept of *"context"* as the surrounding conditions that influence how people interact with technology, including the social, physical, cultural, deployment environments, and users' goals and expectations [1, 24, 25, 78]. Suchman defines context as situated action interactions that cannot be fully predetermined because users rely on contingent, local circumstances to make sense of and respond to technology [78]. Later, Dourish extended this perspective, emphasizing that context is embodied and situated, and where human actions unfold and are interpreted, gaining meaning [25]. Because risk and harm are subjective and contested terms [30, 72], establishing context is crucial to effectively determine if and when a model's output is classified as harmful.

[4] Schwartz et al. [69] define second-order effects as *"any long-term outcomes and consequences that may result from AI use in the real world."*

Despite the importance of establishing context when red teaming LLMs, our empirical findings suggest that practitioners often neglect context in designing red teaming techniques, developing and evaluating datasets. This omission was evident when practitioners constructed scenarios in a context-agnostic manner, selected or generated adversarial datasets based on scale, availability, or community acceptance rather than relevance to real-world domains and risks, and evaluations relied on generic computational metrics that overlooked deployment purposes and user needs. As a result, practitioners treated risk as if it were independent, generic, and abstract, when in reality risk only materializes in relation to the populations affected, the domains of application, and the particular contexts in which technology is used.

5.1.2 Overlooking Interaction Type: The choice between single-turn and multi-turn approaches in red teaming reflects a design decision about the interaction type, meaning whether the generation of adversarial datasets and their evaluation focuses on assessing one user query and an isolated model's outputs or on dialogues that arise in extended user–model exchanges.

Within our pool of participants, only three reported developing red teaming techniques for multi-turn conversations, which they saw as closer to real-world interactions. By taking into account sequences of user-model exchanges rather than evaluating actions in isolation, it is feasible to understand how actions interact across different steps and contexts to prevent sequences that might lead to undesirable or harmful outcomes, even when each individual step appears benign [29]. The rest of the participants reported focusing on single-turn conversations, dismissing the potential risks that emerge through extended or iterative exchanges with AI systems. While single-turn methods are more common, the results might be perceived as less valuable not only because they do not capture how users truly interact with these systems but also because previous research has shown that multi-turn scenarios can surface vulnerabilities that single-turn attacks overlook. For instance, Singhania et al. [73] demonstrated that certain LLMs are 71% more vulnerable after a 5-turn English exchange. In the domain of mental health, Chen et al. [20] showed that in fewer than two conversation turns, five out of six studied models answer the user's original harmful query in at least one test scenario.

5.1.3 Overlooking User Specificity: Throughout our interviews, we often observed that the question of "risk for whom" was left unaddressed. Our participants tended to assess risks for a generic population, without considering subgroup-specific vulnerabilities such as those of youth or older adults. This lack of specificity cascaded across every stage of the red teaming process. When creating adversarial datasets, practitioners did not design with a target population in mind or account for the needs of particular users. Similarly, when evaluating these datasets, they rarely considered who the results would ultimately apply to, because the notion of a target population had never been established in the first place—a direct consequence of conceptualizing red teaming as a purely technical practice (as we discussed in 4.1). Interestingly, even at the evaluation stage, where most practitioners regarded human evaluation as the most trustworthy method, they overlooked that who evaluates also matters, since different evaluators bring different perspectives and biases [81].

5.2 Recommendations for HCI: Research Opportunities

Current approaches to red teaming research are too narrow on technical exploits or malicious attacks, overlooking the complex and subjective nature of harms, which require broader sociotechnical evaluation and engagement with diverse populations. Such LLMs evaluation approaches prevent practitioners from systematically identifying harms prone to emerge in specific domains, contexts, through prolonged interactions, and relevant to certain populations. Thus, we propose expanding the conceptualization of red teaming to capture a more nuanced understanding of what constitutes risk in relation to context, interaction type, and users' needs. Previous efforts have called for the development of red teaming efforts with forms of public participation [72]. We echo these calls and emphasize that HCI researchers have a unique position to inform the design of future red teaming efforts by focusing on the people, processes, and tools involved in testing and improving AI systems, rather than treating red teaming solely as a technical challenge. To follow, we provide three concrete pathways for HCI researchers to support such expansion.

5.2.1 ***Recommendation 1: Design situated and contextualized red teaming scenarios centering the needs of specific communities.*** Most participants conceptualize red teaming as a search problem within an undefined open space, leading them to carry out red teaming from an agnostic standpoint, and aiming to identify "as many as possible" adversarial behaviors. However, conducting such unguided exploration might lead to overlooking the cultural, social, and political factors that shape human–AI interactions. Thus, rather than approaching red teaming as a search problem, we suggest constraining it by designing situated and contextualized red teaming scenarios that (1) account for interaction types, (2) center the goals and expectations of specific communities, and (3) consider the deployment environment. Designing red teaming scenarios with these considerations could help to capture 'compositional risks', which refer to *"actions that are harmless individually may become problematic when combined"* [29].

One strategy is to ground evaluations in current and realistic interactions of vulnerable populations (e.g., minors and older adults) with these technologies. For instance, recent research has reported that children and teens are increasingly using AI chatbots as companions, friends, and even romantic partners [48]. Thus, by grounding red teaming evaluation in real user contexts and needs, practitioners can surface harms that affect users' experiences, such as providing age or culturally inappropriate responses, which may otherwise remain invisible in agnostic approaches. A contextualized approach also offers a clear pathway for integrating HCI methods into the red-teaming pipeline, such as contextual inquiry, participatory methods, and community-centered design.

5.2.2 ***Recommendation 2: Define, categorize, and evaluate risks with input from domain experts, instead of relying on generic taxonomies.*** Our research shows that current red teaming approaches frequently rely on broad, predefined taxonomies that have little to do with end-users' expectations and the domain and context in which they interact with these systems. In addition to the lack of specificity in the taxonomies, most of them are missing categories that capture the specific risks affecting vulnerable populations (e.g., children and youth). We suggest expanding taxonomies through working with specific user groups and contexts [13, 90]. Doing so allows risks to be defined and validated with domain expertise, ensuring that categories reflect real-world vulnerabilities and making red teaming evaluations more actionable and relevant to diverse populations.

By leveraging the long tradition of HCI in participatory research, design, and community-based research, HCI researchers are equipped to develop methods and research that inform the design of risk taxonomies that emerge from realistic interactions between vulnerable end-users and AI, considering the context and domain of these interactions; dimensions lacking in current approaches to red teaming LLMs.

5.2.3 ***Recommendation 3: Evaluate compositional risk rather than assessing isolated models' outputs.*** As described in our findings, most participants identified risk behaviors based on the assessment of single-turn interactions, classified them based on narrow risk taxonomies, and overlooked the situational context of interaction and the needs and expectations of end-users in their process of generating scenarios. These approaches fail to capture the dynamics of real-world interactions.

Emerging evidence suggests that the most negative impacts on people result from continued interaction, such as "flirting," emotional persuasion, and chatbots' refusal to stop despite users' rejection [48, 64]. And these impacts can be worse on vulnerable populations. Such harms are particularly acute for vulnerable populations, who may be more susceptible to manipulation or emotional distress. Given the extensive methodological expertise that the HCI field has to examine the role of interaction types in how humans experience technology, we consider that HCI researchers could contribute to expanding the evaluative approaches to red teaming LLMs from the perspective of interaction type. Current approaches assess the output of interactions once the human-AI conversation is completed. However, what is needed is to assess the conversation 'in the flight' to identify patterns of escalation, flag when benign queries turn harmful, and evaluate interactions as a whole by considering domain, deployment context, and end-user needs, rather than just standalone models' outputs.

5.3 Reflecting on the Power Asymmetries between Academia and Industry

In recent years, scholars have raised concerns about the growing disparities of resources and influence between industry and academia in AI research [4, 5, 14, 22]. While these power asymmetries are not new, they have become increasingly difficult to mitigate in the context of LLM evaluation and red teaming. Industry actors concentrate the vast majority of resources to develop *state-of-the-art models*, including proprietary training data and computing resources required to build them [12]. These advantages allow industry teams to experiment with a broader range of adversarial techniques and to establish methodological standards. In contrast, academic researchers are often limited to evaluating models produced mainly by industry, but with constrained access to model internals and dependence on API-based interfaces.

Our findings echo these constraints, specifically in the infrastructural limitations that academic researchers faced when attempting to evaluate harmful behavior. As we discussed in section 4.3.2, P09 lost access to the OpenAI API while conducting a red teaming evaluation. Participants' experience highlights the current infrastructural limitations to conducting safety research outside the industry environment. These asymmetries shape not only who can conduct red teaming but also what forms of red teaming become possible and valued. This dynamic explains why current red teaming approaches remain narrowly focused on technical exploits, underscoring a need for broader socio-technical evaluation approaches.

One way to counteract these power asymmetries is to surface how they are produced and what consequences they generate. Data work research has long shown that uneven distributions of resources, authority, and institutional influence shape data practices [7, 50, 51, 67]. Building on this body of work, the analysis presented in this article examines how practitioners assemble adversarial datasets and, in doing so, we surface the values, assumptions, and operational constraints embedded in the evaluation standards that currently dominate the field. With this, we open the space for contesting and reshaping the evaluation practices that determine what constitutes a risk.

6 CONCLUSION

Our study examined how AI practitioners create, develop, and evaluate red teaming datasets for LLMs. Drawing on 22 interviews, our analysis identified three critical moments: defining red teaming tasks, developing adversarial datasets, and evaluating adversarial datasets. We noticed that harmfulness is not fixed but constructed through these data practices, which embed cultural norms and technical vulnerabilities into red teaming LLMs. For HCI, these insights highlight opportunities to support practitioners by expanding evaluations to reflect the context of use, engaging domain expertise in defining harms, and designing tools that assess risks at the level of interaction rather than isolated pairs of questions and answers.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their constructive feedback and our participants for their time and insights. We also thank Michael Muller for his valuable feedback on this paper.

REFERENCES

[1] Gregory D. Abowd, Anind K. Dey, Peter J. Brown, Nigel Davies, Mark Smith, and Pete Steggles. 1999. Towards a Better Understanding of Context and Context-Awareness. In *Handheld and Ubiquitous Computing*, Hans-W. Gellersen (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 304–307.
[2] Mark S Ackerman. 2000. The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Human–Computer Interaction* 15, 2-3 (2000), 179–203.
[3] Lama Ahmad, Sandhini Agarwal, Michael Lampe, and Pamela Mishkin. 2025. OpenAI's Approach to External Red Teaming for AI Models and Systems. *arXiv* (Jan. 2025). https://doi.org/10.48550/arXiv.2503.16431 arXiv:2503.16431
[4] Nur Ahmed and Neil C. Thompson. 2023. What should be done about the growing influence of industry in AI research? https://www.brookings.edu/articles/what-should-be-done-about-the-growing-influence-of-industry-in-ai-research
[5] Nur Ahmed, Muntasir Wahed, and Neil C. Thompson. 2023. The growing influence of industry in AI research. *Science* 379, 6635 (2023), 884–886. https://doi.org/10.1126/science.ade2420
[6] Adriana Alvarado Garcia, Heloisa Candello, Karla Badillo-Urquiola, and Marisol Wong-Villacres. 2025. Emerging Data Practices: Data Work in the Era of Large

Language Models. In *ACM Conferences*. Association for Computing Machinery, New York, NY, USA, 1–21. https://doi.org/10.1145/3706598.3714069
[7] Adriana Alvarado Garcia, Tianling Yang, and Milagros Miceli. 2025. What Knowledge Do We Produce from Social Media Data and How? *Proceedings of the ACM on Human-Computer Interaction* 9, 1 (2025), 1–45.
[8] Evan Anderson, Jim Holdsworth, and Matthew Kosinski. 2025. What is Red teaming. https://www.ibm.com/think/topics/red-teaming [Online; accessed 27. Jul. 2025].
[9] Anthropic. 2023. Model Card and Evaluations for Claude Models. Technical model card. https://www-cdn.anthropic.com/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226/Model-Card-Claude-2.pdf.
[10] Andy Applebaum, Doug Miller, Blake Strom, Chris Korban, and Ross Wolf. 2016. Intelligent, automated red team emulation. In *ACM Other conferences*. Association for Computing Machinery, New York, NY, USA, 363–373. https://doi.org/10.1145/2991079.2991111
[11] Karla Badillo-Urquiola, Zainab Agha, Denielle Abaquita, Scott B. Harpin, and Pamela J. Wisniewski. 2024. Towards a Social Ecological Approach to Supporting Caseworkers in Promoting the Online Safety of Youth in Foster Care. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1 (April 2024), 1–28. https://doi.org/10.1145/3637412
[12] Natasha Bajema. 2024. Why are large AI models being red teamed? https://spectrum.ieee.org/red-team-ai-llms
[13] Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M. Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, Hao Qiu, Shrey Jain, Leonardo Schettini, Mehr Kashyap, Jason Alan Fries, Akshay Swaminathan, Philip Chung, Fateme Nateghi, Asad Aali, Ashwin Nayak, Shivam Vedak, Sneha S. Jain, Birju Patel, Oluseyi Fayanju, Shreya Shah, Ethan Goh, Dong han Yao, Brian Soetikno, Eduardo Reis, Sergios Gatidis, Vasu Divi, Robson Capasso, Rachna Saralkar, Chia-Chun Chiang, Jenelle Jindal, Tho Pham, Faraz Ghoddusi, Steven Lin, Albert S. Chiou, Christy Hong, Mohana Roy, Michael F. Gensheimer, Hinesh Patel, Kevin Schulman, Dev Dash, Danton Char, Lance Downing, Francois Grolleau, Kameron Black, Bethel Mieso, Aydin Zahedivash, Wen wai Yim, Harshita Sharma, Tony Lee, Hannah Kirsch, Jennifer Lee, Nerissa Ambers, Carlene Lugtu, Aditya Sharma, Bilal Mawji, Alex Alekseyev, Vicky Zhou, Vikas Kakkar, Jarrod Helzer, Anurang Revri, Yair Bannett, Roxana Daneshjou, Jonathan Chen, Emily Alsentzer, Keith Morse, Nirmal Ravi, Nima Aghaeepour, Vanessa Kennedy, Akshay Chaudhari, Thomas Wang, Sanmi Koyejo, Matthew P. Lungren, Eric Horvitz, Percy Liang, Mike Pfeffer, and Nigam H. Shah. 2025. MedHELM: Holistic Evaluation of Large Language Models for Medical Tasks. arXiv:2505.23802 [cs.CL] https://arxiv.org/abs/2505.23802
[14] Tamay Besiroglu, Sage Andrus Bergerson, Amelia Michael, Lennart Heim, Xueyun Luo, and Neil Thompson. 2024. The Compute Divide in Machine Learning: A Threat to Academic Contribution and Scrutiny? arXiv:2401.02452 [cs.CY] https://arxiv.org/abs/2401.02452
[15] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 173–184.
[16] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
[17] John Burden. 2024. Evaluating AI Evaluation: Perils and Prospects. arXiv:2407.09221 [cs.AI] https://arxiv.org/abs/2407.09221
[18] Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell. 2023. Explore, Establish, Exploit: Red Teaming Language Models from Scratch. arXiv:2306.09442 [cs.CL] https://arxiv.org/abs/2306.09442
[19] Center for AI Safety. 2023. Statement on AI Risk. https://www.safe.ai/statement-on-ai-risk.
[20] Bocheng Chen, Guangjing Wang, Hanqing Guo, Yuanda Wang, and Qiben Yan. 2023. Understanding Multi-Turn Toxic Behaviors in Open-Domain Chatbots. In *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses* (Hong Kong, China) *(RAID '23)*. Association for Computing Machinery, New York, NY, USA, 282–296. https://doi.org/10.1145/3607199.3607237
[21] Sasha Costanza-Chock. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need*. The MIT Press, Cambridge, MA, USA. https://doi.org/10.7551/mitpress/12255.001.0001
[22] Matt Davies and Jai Vipra. 2025. *Computing Commons: Designing public compute for people and society*. Ada Lovelace Institute. https://www.adalovelaceinstitute.org/report/computing-commons/
[23] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 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* (Boston, MA, USA) *(EAAMO '23)*. Association for Computing Machinery, New York, NY, USA, Article 37, 23 pages. https://doi.org/10.1145/3617694.3623261
[24] Anind K. Dey. 2001. Understanding and Using Context. *Personal Ubiquitous Comput.* 5, 1 (Jan. 2001), 4–7. https://doi.org/10.1007/s007790170019
[25] Paul Dourish. 2001. Seeking a foundation for context-aware computing. *Hum.-Comput. Interact.* 16, 2 (Dec. 2001), 229–241. https://doi.org/10.1207/S15327051HCI16234_07
[26] Maria Eriksson, Erasmo Purificato, Arman Noroozian, Joao Vinagre, Guillaume Chaslot, Emilia Gomez, and David Fernandez-Llorca. 2025. Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation. arXiv:2502.06559 [cs.AI] https://arxiv.org/abs/2502.06559
[27] European Parliament. 2025. EU AI Act: first regulation on artificial intelligence | Topics | European Parliament. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence
[28] Michael Feffer, Anusha Sinha, Wesley H. Deng, Zachary C. Lipton, and Hoda Heidari. 2024. Red-Teaming for Generative AI: Silver Bullet or Security Theater? In *ACM Conferences*. AAAI Press, 421–437. https://doi.org/10.1609/aies.v7i1.31647
[29] Camille François, Ludovic Péran, Ayah Bdeir, Nouha Dziri, Will Hawkins, Yacine Jernite, Sayash Kapoor, Juliet Shen, Heidy Khlaaf, Kevin Klyman, Nik Marda, Marie Pellat, Deb Raji, Divya Siddarth, Aviya Skowron, Joseph Spisak, Madhulika Srikumar, Victor Storchan, Audrey Tang, and Jen Weedon. 2025. A Different Approach to AI Safety: Proceedings from the Columbia Convening on Openness in Artificial Intelligence and AI Safety. arXiv:2506.22183 [cs.AI] https://arxiv.org/abs/2506.22183
[30] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. *arXiv* (Aug. 2022). https://doi.org/10.48550/arXiv.2209.07858 arXiv:2209.07858
[31] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. *ACL Anthology* (Nov. 2020), 3356–3369. https://doi.org/10.18653/v1/2020.findings-emnlp.301
[32] Tarleton Gillespie, Ryland Shaw, Mary L. Gray, and Jina Suh. 2024. AI red-teaming is a sociotechnical challenge: on values, labor, and harms. *arXiv* (Dec. 2024). https://doi.org/10.48550/arXiv.2412.09751 arXiv:2412.09751
[33] Jessica He, Stephanie Houde, Gabriel E. Gonzalez, Darío Andrés Silva Moran, Steven I. Ross, Michael Muller, and Justin D. Weisz. 2024. AI and the Future of Collaborative Work: Group Ideation with an LLM in a Virtual Canvas. In *ACM Other conferences*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3663384.3663398
[34] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature* 633 (Sept. 2024), 147–154. https://doi.org/10.1038/s41586-024-07856-5
[35] Andrew Gary Darwin Holmes. 2020. Researcher Positionality–A Consideration of Its Influence and Place in Qualitative Research–A New Researcher Guide. *Shanlax International Journal of Education* 8, 4 (2020), 1–10.
[36] The White House. 2023. FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence | The White House. *White House* (Oct. 2023). https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/
[37] Andrew Iliadis and Federica Russo. 2016. Critical data studies: An introduction. *Big Data & Society* 3, 2 (2016), 2053951716674238.
[38] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. 2023. Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. arXiv:2312.06674 [cs.CL] https://arxiv.org/abs/2312.06674
[39] Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Chi Zhang, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset. *arXiv* (July 2023). https://doi.org/10.48550/arXiv.2307.04657 arXiv:2307.04657
[40] Emma Kallina, Thomas Bohné, and Jatinder Singh. 2025. Stakeholder Participation for Responsible AI Development: Disconnects Between Guidance and Current Practice. In *ACM Conferences*. Association for Computing Machinery, New York, NY, USA, 1060–1079. https://doi.org/10.1145/3715275.3732069
[41] Michelle S Lam, Ayush Pandit, Colin H Kalicki, Rachit Gupta, Poonam Sahoo, and Danaë Metaxa. 2023. Sociotechnical audits: Broadening the algorithm auditing lens to investigate targeted advertising. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–37.
[42] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav

Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic Evaluation of Language Models. *arXiv* (Nov. 2022). https://doi.org/10.48550/arXiv.2211.09110 arXiv:2211.09110

[43] Thomas Liao, Rohan Taori, Deborah Raji, and Ludwig Schmidt. 2021. Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, J. Vanschoren and S. Yeung (Eds.), Vol. 1. https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/757b505cfd34c64c85ca5b5690ee5293-Paper-round2.pdf

[44] Lizhi Lin, Honglin Mu, Zenan Zhai, Minghan Wang, Yuxia Wang, Renxi Wang, Junjie Gao, Yixuan Zhang, Wanxiang Che, Timothy Baldwin, et al. 2025. Against The Achilles' Heel: A Survey on Red Teaming for Generative Models. *Journal of Artificial Intelligence Research* 82 (2025), 687–775.

[45] Shayne Longpre, Sayash Kapoor, Kevin Klyman, Ashwin Ramaswami, Rishi Bommasani, Borhane Blili-Hamelin, Yangsibo Huang, Aviya Skowron, Zheng-Xin Yong, Suhas Kotha, Yi Zeng, Weiyan Shi, Xianjun Yang, Reid Southen, Alexander Robey, Patrick Chao, Diyi Yang, Ruoxi Jia, Daniel Kang, Sandy Pentland, Arvind Narayanan, Percy Liang, and Peter Henderson. 2024. A Safe Harbor for AI Evaluation and Red Teaming. *arXiv* (March 2024). https://doi.org/10.48550/arXiv.2403.04893 arXiv:2403.04893

[46] Subhabrata Majumdar, Brian Pendleton, and Abhishek Gupta. 2025. Red Teaming AI Red Teaming. *arXiv* (July 2025). https://doi.org/10.48550/arXiv.2507.05538 arXiv:2507.05538

[47] Steve Mansfield-Devine. 2018. The best form of defence–the benefits of red teaming. *Computer Fraud & Security* 2018, 10 (2018), 8–12.

[48] Omny Miranda Martone. 2025. Ai is sexually harassing our kids. here's how legislators can stop it. https://www.techpolicy.press/ai-is-sexually-harassing-our-kids-heres-how-legislators-can-stop-it/

[49] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. 2024. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. *arXiv* (Feb. 2024). https://doi.org/10.48550/arXiv.2402.04249 arXiv:2402.04249

[50] Milagros Miceli, Julian Posada, and Tianling Yang. 2022. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? *Proc. ACM Hum.-Comput. Interact.* 6, GROUP, Article 34 (Jan. 2022), 14 pages. https://doi.org/10.1145/3492853

[51] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 115 (Oct. 2020), 25 pages. https://doi.org/10.1145/3415186

[52] Luis Morales-Navarro, Yasmin Kafai, Vedya Konda, and Danaë Metaxa. 2024. Youth as Peer Auditors: Engaging Teenagers with Algorithm Auditing of Machine Learning Applications. In *ACM Conferences*. Association for Computing Machinery, New York, NY, USA, 560–573. https://doi.org/10.1145/3628516.3655752

[53] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How data science workers work with data: Discovery, capture, curation, design, creation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[54] Michael Muller and Angelika Strohmayer. 2022. Forgetting Practices in the Data Sciences. In *ACM Conferences*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3491102.3517644

[55] Ozioma Collins Oguine, Oghenemaro Anuyah, Zainab Agha, Iris Melgarez, Adriana Alvarado Garcia, and Karla Badillo-Urquiola. 2025. Online Safety for All: Sociocultural Insights from a Systematic Review of Youth Online Safety in the Global South. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW458 (Oct. 2025), 30 pages. https://doi.org/10.1145/3757639

[56] Ozioma C. Oguine, Johanna Olesk, Jaemarie Solyst, Michael Madaio, Michael Muller, Adriana Alvarado Garcia, and Karla Badillo-Urquiola. 2025. Bridging Expertise and Participation in AI: Multistakeholder Approaches to Safer AI Systems for Youth Online Safety. In *Companion of the 2025 ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW Companion '25)*. ACM, Bergen, Norway, 6 pages. https://doi.org/10.1145/3715070.3748294

[57] OpenAI. 2024. GPT-4o System Card. https://openai.com/index/gpt-4o-system-card

[58] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

[59] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red Teaming Language Models with Language Models. *arXiv* (Feb. 2022). https://doi.org/10.48550/arXiv.2202.03286 arXiv:2202.03286

[60] Crystal Qian, Emily Reif, and Minsuk Kahng. 2024. Understanding the Dataset Practitioners Behind Large Language Models. In *ACM Conferences*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3613905.3651007

[61] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *ACL Anthology* (Nov. 2016), 2383–2392. https://doi.org/10.18653/v1/D16-1264

[62] Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. 2023. Fairness in Language Models Beyond English: Gaps and Challenges. In *Findings of the Association for Computational Linguistics: EACL 2023*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 2106–2119. https://doi.org/10.18653/v1/2023.findings-eacl.157

[63] Ehud Reiter. 2025. We Should Evaluate Real-World Impact. arXiv:2507.05973 [cs.CL] https://arxiv.org/abs/2507.05973

[64] Michael B. Robb and Supreet Mann. 2025. *Talk, Trust, and Trade-Offs: The Role of Children's Online Safety, Privacy, and Learning in the Digital Age.* Technical Report. Common Sense Media. https://www.commonsensemedia.org/sites/default/files/research/report/talk-trust-and-trade-offs_2025_web.pdf Research report.

[65] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M. Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *ACM Conferences*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3411764.3445518

[66] Devansh Saxena, Ji-Youn Jung, Jodi Forlizzi, Kenneth Holstein, and John Zimmerman. 2025. AI Mismatches: Identifying Potential Algorithmic Harms Before AI Development. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–23.

[67] Morgan Klaus Scheuerman, Alex Hanna, and Remi Denton. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 317 (Oct. 2021), 37 pages. https://doi.org/10.1145/3476058

[68] Annika M Schoene and Cansu Canca. 2025. 'For Argument's Sake, Show Me How to Harm Myself!': Jailbreaking LLMs in Suicide and Self-Harm Contexts. arXiv:2507.02990 [cs.CL] https://arxiv.org/abs/2507.02990

[69] Reva Schwartz, Rumman Chowdhury, Akash Kundu, Heather Frase, Marzieh Fadaee, Tom David, Gabriella Waters, Afaf Taik, Morgan Briggs, Patrick Hall, Shomik Jain, Kyra Yee, Spencer Thomas, Sundeep Bhandari, Paul Duncan, Andrew Thompson, Maya Carlyle, Qinghua Lu, Matthew Holmes, and Theodora Skeadas. 2025. Reality Check: A New Evaluation Ecosystem Is Necessary to Understand AI's Real World Effects. arXiv:2505.18893 [cs.CY] https://arxiv.org/abs/2505.18893

[70] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *ACM Conferences*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/3287560.3287598

[71] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. *ACL Anthology* (Nov. 2019), 3407–3412. https://doi.org/10.18653/v1/D19-1339

[72] Ranjit Singh, Borhane Blili-Hamelin, Carol Anderson, Emnet Tafesse, Briana Vecchione, Beth Duckles, and Jacob Metcalf. 2025. Red-Teaming in the Public Interest. *New York: Data & Society Research Institute* (2025).

[73] Abhishek Singhania, Christophe Dupuy, Shivam Sadashiv Mangale, and Amani Namboori. 2025. Multi-lingual Multi-turn Automated Red Teaming for LLMs. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, Trista Cao, Anubrata Das, Tharindu Kumarage, Yixin Wan, Satyapriya Krishna, Ninareh Mehrabi, Jwala Dhamala, Anil Ramakrishna, Aram Galystan, Anoop Kumar, Rahul Gupta, and Kai-Wei Chang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 141–154. https://doi.org/10.18653/v1/2025.trustnlp-main.11

[74] Anusha Sinha, James Lucassen, Keltin Grimes, Michael Feffer, Mary Soto, Hoda Heidari, and Nathan Vanhoudnos. 2025. What Can Generative AI Red-Teaming Learn from Cyber Red-Teaming? (July 2025). https://doi.org/10.1184/R1/29410136.v1

[75] Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Canyu Chen, Hal Daumé Iii, Jesse Dodge, Isabella Duan, Ellie Evans, Felix Friedrich, Avijit Ghosh, Usman Gohar, Sara Hooker, Yacine Jernite, Ria Kalluri, Alberto Lusoli, Alina Leidinger, Michelle Lin, Xiuzhu Lin, Sasha Luccioni, Jennifer Mickel, Margaret Mitchell, Jessica Newman, Anaelia Ovalle, Marie-Therese Png, Shubham Singh, Andrew Strait, Lukas Struppek, and Arjun Subramonian. 2023. Evaluating the Social Impact of Generative AI Systems in Systems and Society. *arXiv* (June 2023). https://doi.org/10.48550/arXiv.2306.05949 arXiv:2306.05949

[76] Jaemarie Solyst, Ellia Yang, Shixian Xie, Amy Ogan, Jessica Hammer, and Motahhare Eslami. 2023. The Potential of Diverse Youth as Stakeholders in Identifying and Mitigating Algorithmic Bias for a Future of Fairer AI. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 364 (Oct. 2023), 27 pages. https://doi.org/10.1145/3610213

[77] Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, et al. 2024. A strongreject for empty jailbreaks. *Advances in Neural Information Processing Systems* 37 (2024), 125416–125440.

[78] Lucy A. Suchman. 1987. *Plans and situated actions: the problem of human-machine communication.* Cambridge University Press, USA.

[79] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminate: Structured generation and exploration of design space with large language models for human-ai co-creation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–26.

[80] Ruyuan Wan, Adriana Alvarado Garcia, Devansh Saxena, Catalina Vajiac, Anna Kawakami, Logan Stapleton, Haiyi Zhu, Kenneth Holstein, Heloisa Candello, and Karla Badillo-Urquiola. 2023. Community-driven AI: Empowering people through responsible data-driven decision-making. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. 532–536.

[81] Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 14523–14530.

[82] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *ACL Anthology* (Nov. 2018), 353–355. https://doi.org/10.18653/v1/W18-5446

[83] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from Language Models. *arXiv* (Dec. 2021). https://doi.org/10.48550/arXiv.2112.04359 arXiv:2112.04359

[84] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. *arXiv* (Oct. 2023). https://doi.org/10.48550/arXiv.2310.11986 arXiv:2310.11986

[85] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 214–229. https://doi.org/10.1145/3531146.3533088

[86] Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in Detoxifying Language Models. *ACL Anthology* (Nov. 2021), 2447–2469. https://doi.org/10.18653/v1/2021.findings-emnlp.210

[87] Richmond Y. Wong, Deirdre K. Mulligan, Ellen Van Wyk, James Pierce, and John Chuang. 2017. Eliciting Values Reflections by Engaging Privacy Futures Using Design Workbooks. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW (Dec. 2017), 1–26. https://doi.org/10.1145/3134746

[88] Ziang Xiao, Wesley Hanwen Deng, Michelle S. Lam, Motahhare Eslami, Juho Kim, Mina Lee, and Q. Vera Liao. 2024. Human-Centered Evaluation and Auditing of Language Models. In *ACM Conferences*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3613905.3636302

[89] Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 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*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, Online, 2950–2968. https://doi.org/10.18653/v1/2021.naacl-main.235

[90] Yaman Yu, Yiren Liu, Jacky Zhang, Yun Huang, and Yang Wang. 2025. Understanding Generative AI Risks for Youth: A Taxonomy Based on Empirical Data. arXiv:2502.16383 [cs.HC] https://arxiv.org/abs/2502.16383

[91] Micah Zenko. 2015. *Red Team: How to succeed by thinking like the enemy.* Basic Books.

[92] Alice Qian Zhang, Ryland Shaw, Jacy Reese Anthis, Ashlee Milton, Emily Tseng, Jina Suh, Lama Ahmad, Ram Shankar Siva Kumar, Julian Posada, Benjamin Shestakofsky, et al. 2024. The human factor in ai red teaming: Perspectives from social and collaborative computing. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*. 712–715.

[93] Alice Qian Zhang, Jina Suh, Mary L. Gray, and Hong Shen. 2025. Effective Automation to Support the Human Infrastructure in AI Red Teaming. *interactions* 32, 4 (June 2025), 58–61. https://doi.org/10.1145/3731866

[94] Alice Qian Zhang, Jiayin Zhi, Srravya Chandhiramowuli, Hong Shen, Laura Dabbish, Theodora Skeadas, Sarah Amos, and Jina Suh. 2025. The Work of AI Red Teaming: Automation and the Human Infrastructure. In *Companion Publication of the 2025 Conference on Computer-Supported Cooperative Work and Social Computing*. 84–87.

[95] Zheng Zhang, Weirui Peng, Xinyue Chen, Luke Cao, and Toby Jia-Jun Li. 2025. LADICA: a large shared display interface for generative AI cognitive assistance in co-located team collaboration. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–22.

[96] Dora Zhao, Qianou Ma, Xinran Zhao, Chenglei Si, Chenyang Yang, Ryan Louie, Ehud Reiter, Diyi Yang, and Tongshuang Wu. 2025. SPHERE: An Evaluation Card for Human-AI Systems. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 1340–1365. https://doi.org/10.18653/v1/2025.findings-acl.70

[97] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv* (July 2023). https://doi.org/10.48550/arXiv.2307.15043 arXiv:2307.15043