Participatory ML for Social Harm Evaluation Should be Constructed, Validated and Reasoned Through First-Person Accounts

Atmadeep Ghoshal
Department of Informatics
King's College London
London, UK
atmadeep.ghoshal@kcl.ac.uk

Ankit Agarwal
Indraprastha Institute of
Information Technology-Delhi
New Delhi, India
ankit24080@iiitd.ac.in

Martim Brandão
Department of Informatics
King's College London
London, UK
martim.brandao@kcl.ac.uk

Abstract

Participatory machine learning has improved who develops social harm evaluation resources but has left largely unexamined what kind of text serves as the raw signal for those resources. Researcher-generated stimuli fix the epistemic frame of harm before community participation begins, and the structured conditions of annotation prevent the richest forms of interpretive harm knowledge from entering the pipeline. This position paper argues that participatory ML for social harm evaluation should construct, validate and reason through first-person accounts, by which we mean autobiography, memoir, testimony, and oral history authored by people describing their own experiences of discrimination and harm. Such accounts possess two epistemic properties that directly address these limitations. Authorial independence from research framing grounds what counts as harm, which contexts matter, and how discrimination is operationalised in experience rather than in researchers' prior conceptions of harm. Embedded rationale structure preserves the interpretive judgment through which harm becomes meaningful rather than reducing it to a label. We draw on existing archival corpora spanning multiple languages and harm domains and propose three concrete integration pathways across evaluation benchmarks, harm classifier training corpora, and safety preference datasets.

1 Introduction

Blodgett et al. (2020) argued that NLP bias research had systematically failed to engage with the lived experiences of the communities most affected by these systems, and called for that work to be centred accordingly [1]. The participatory turn in machine learning emerged in part as a response to this concern, producing community led initiatives in data collection, problem formulation, and evaluation. Projects such as Masakhane [2], Aya [3], and World Wide Dishes [4] showed that participation can materially improve data quality and cultural coverage for underrepresented populations [2, 3, 4]. This paper focuses on social harm evaluation resources, by which we mean the materials used to operationalise the evaluation of the ML models against harmful, biased or unsafe behaviour, such as benchmarks, annotated training corpora for harm classifiers, and safety preference datasets [5, 6, 7, 8, 9, 10]. Across these resource types, a common construction logic persists. Researchers

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Position Paper Track.

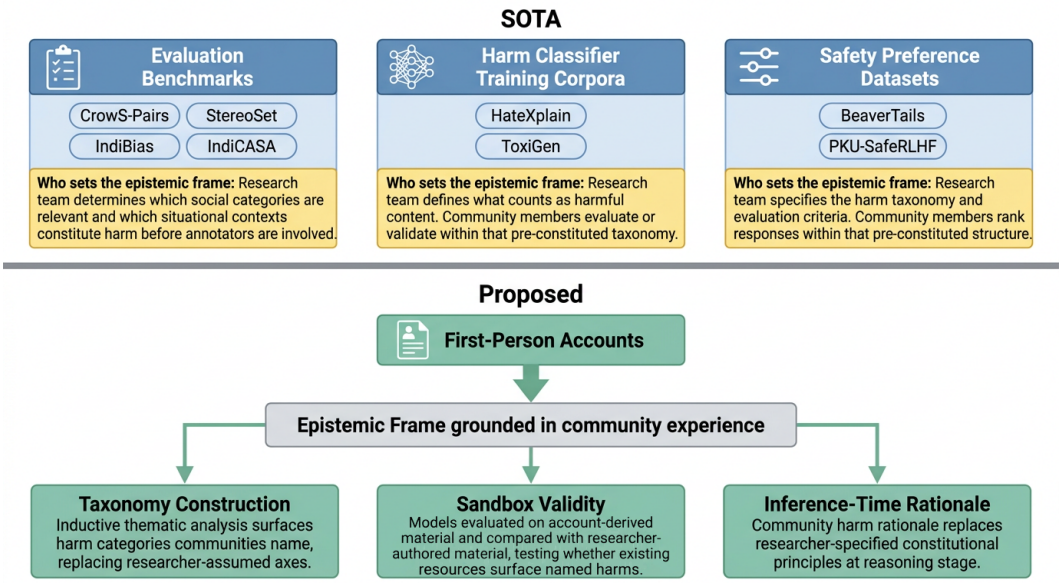


Figure 1: The construction logic of current social harm evaluation resources (top) and the proposed integration of first-person accounts as an upstream signal (bottom)

generate stimuli through demographic substitution in fixed templates or through researcher specified harm scenarios, and annotators are then asked to evaluate them. The epistemic frame—including what counts as a harm relevant event, which dimensions of experience are legible as harm, and which situations are worth representing—is therefore fixed before community participation begins. Miceli et al. (2022) show that such framing decisions are embedded in annotation instructions and reflect the legal categories and cultural assumptions of those who produce them [11]. Addressing this requires not only reforming who participates in evaluation but rethinking what serves as the raw signal in the first place. To address this, this paper proposes the use of first-person accounts in the ML safety pipeline, with concrete pathways for how they can be used to build harm taxonomies grounded in community experience, validate existing evaluation resources against community-named harms, and enrich inference-time harm reasoning. The term first person accounts refer to autobiography, memoir, testimony, and oral history authored by people describing their own experiences of discrimination and harm, a genre whose epistemic properties differ from externally authored or researcher constructed text [12, 13]. Our position is thus that **participatory ML for social harm evaluation should be constructed, validated and reasoned through first-person accounts**. Our work is not simply a call for affected communities to evaluate data, but a claim that the raw signal itself should include textual artifacts authored outside research infrastructure by people who determine their own categories, frames, and contexts of harm. We proceed as follows. Section 2 traces the achievements of participatory ML and identifies the specific limitations our position addresses. Section 3 argues for first person accounts as a distinct epistemic modality whose properties directly address those limitations. Section 4 proposes three integration pathways. Section 5 discusses ethical considerations arising out of our recommendations and Section 6 addresses alternative views.

2 What Participatory ML Has Reformed and What It Has Not

Participatory ML has meaningfully reshaped how communities are involved in the design, development, and evaluation of machine learning systems, improving coverage, representation, and the quality of downstream judgments. However, it has not fully transformed the epistemic structure through which harm is rendered measurable. Two limitations persist in this direction: a *frame constraint*, in which researchers continue to determine the textual forms and categories through which harm is assessed before community participation begins, and an *agency constraint*, in which participation often occurs under epistemic and institutional conditions that may limit communities' ability to contest those frames on their own terms.

The Frame Constraint Participatory ML has made important contributions in the development of social harm evaluation resources, by which we mean the materials used to evaluate models for stereotypical, harmful, or unsafe behaviour such as evaluation benchmarks, harm-classifier training corpora, and safety preference datasets. Community-led data collection projects have demonstrated that participation can materially improve coverage and cultural representation for populations historically excluded from benchmark construction, producing community-driven benchmarks across more than thirty African languages [2], extending instruction tuning to 101 languages through large-scale contributor collaborations [3], and building culinary datasets through bottom-up community contribution that explicitly engaged populations excluded from web-scraped data [4]. A growing range of such participatory methods has been documented across the ML pipeline [14]. Yet a body of critical work has shown that these reforms do not by themselves transform the underlying power relations that shape what gets measured and how. Participation can become participation washing when inclusion language is adopted without substantive redistribution of decision-making power [15], may remain extractive when underlying power relations are left intact [16], and is often reduced to structured preference elicitation through one-time engagements that leave participants little agency over the values being operationalised [17]. What these critiques have not fully examined is a more basic assumption shared by participatory and non-participatory approaches alike, namely that the raw signal for social harm evaluation should take the form of researcher-generated stimuli later evaluated by annotators. This pattern holds across resource types. Social bias benchmarks operationalise harm through minimal sentence pairs generated by substituting demographic group labels within researcher-defined templates [5], stereotypical associations are measured through sentence completion tasks across researcher-specified social categories [6], and safety preference datasets construct preference pairs from harm scenarios specified by researchers and ranked by crowdworkers sourced from general labour platforms [9]. In each case, researchers determine in advance what counts as a harm-relevant event, which social categories matter, and which contexts are worth representing. The annotators recruited are typically crowdworkers whose demographic composition rarely reflects the communities most affected by the harms being measured, and who evaluate stimuli within categories and formats they had no role in defining. Whether the text takes the form of a minimal pair, a harm scenario, or a generated prompt, the epistemic frame has already been fixed before community participation begins. Related work shows that language models standing in for community perspectives produce an illusion of inclusion [18], which the material proposed here avoids because community members, not a model, author it.

The Agency Constraint There is a further limitation that participatory ML has not fully addressed, namely the epistemic conditions under which participation occurs. Two forms of epistemic injustice are relevant here. The first is testimonial injustice, where speakers receive a credibility deficit because of identity prejudice, and the second is hermeneutical injustice, where dominant interpretive frameworks make it harder for people to understand and articulate their own social experience [19]. The most marginalised are often prevented from speaking in ways that can be heard on their own terms [20], and annotation settings reproduce this dynamic in a specific form. A community member from a historically marginalised caste background[1] asked to assess a researcher-constructed stimulus about caste discrimination confronts a frame they did not help define. The stimulus may reduce caste harm to a substituted demographic label in a minimal sentence pair, while the actual experience of caste discrimination involves accumulated institutional exclusions, spatial segregation, and forms of humiliation that such a template cannot encode. The structured, time-limited, researcher-mediated format of annotation then constrains what can be expressed and what counts as a legitimate contribution, leaving those more complex configurations of harm outside the data entirely. Annotation instructions have been shown to encode institutional worldviews such that the researcher's categories are the ones that survive into the dataset [11], and widely used fairness benchmarks carry ambiguities in how stereotyping is conceptualised and rely on assumptions about harm-relevant social categories that do not travel across cultural contexts and often lack explicit alignment with the harms they claim to operationalise [21, 22]. The frame of harm is fixed before community participation begins, and annotator diversity alone cannot resolve a problem that is upstream of annotation. The consequences are visible in how resources travel across contexts. IndiBias was built by filtering and translating CrowS-Pairs, inheriting its framing biases from that US-centric resource [23], and IndiCASA then builds directly on IndiBias while retaining the same template-based construction logic [24]. Participatory reforms at the annotation stage cannot interrupt

[1] Dalits are communities historically placed outside the Hindu caste hierarchy in India and subjected to institutionalised discrimination across employment, housing, and public life.

this inheritance. First-person accounts, authored outside any research infrastructure before any research relation exists, address both constraints at their source.

The Construct Validity Problem The frame and agency constraints might be dismissed as theoretical concerns if researcher-constructed stimuli reliably tracked real-world harm. The empirical record suggests otherwise. Goldfarb-Tarrant et al. (2021) show that intrinsic bias metrics derived from artificially constructed sentence pairs do not consistently correlate with bias in downstream applications [25]. Delobelle et al. (2022) show that such metrics are highly sensitive to researcher-controlled design choices, such that small changes to template sentences or seed words produce substantially different bias scores without corresponding changes in downstream behaviour [26]. Selvam et al. (2023) show that paraphrasing widely used benchmark sentences produces large shifts in measured bias scores and model rankings even when the underlying social meaning is preserved [27]. Jacobs and Wallach (2021) provide the theoretical framework: harms arise when theoretical constructs diverge from operationalisations, especially when the assumptions embedded in those operationalisations remain implicit [28]. When researcher-generated stimuli are not grounded in communities' own accounts of discrimination, this gap widens. Cabello et al. (2023) demonstrate one downstream consequence, showing that association bias scores and empirical fairness outcomes in widely used language models are often uncorrelated [29]. This is not a problem that more careful researcher constructed stimulus design can resolve, because the instability follows from the construction logic itself rather than from any particular implementation of it.

3 First-Person Accounts as a Distinct Epistemic Modality

Section 2 identified three structural limitations. Researchers fix what counts as harm before participation begins, participation conditions prevent marginalised contributors from contesting that frame, and researcher-determined stimuli fail to track harm as communities experience it. We now argue that first-person accounts possess two epistemic properties that directly address all three. Epistemic autonomy from research framing addresses the frame constraint and narrows the construct validity gap by grounding what counts as harm, which contexts matter, and how discrimination is operationalised in experience rather than in researchers' prior conceptions of harm. Embedded rationale structure addresses the agency constraint by preserving the interpretive judgment through which harm becomes meaningful rather than reducing it to a label.

Epistemic Autonomy and Standpoint The first property of first person accounts is epistemic autonomy from research framing. First-person accounts are produced outside research infrastructure and without reference to an annotation task, benchmark category, or elicitation protocol. Authors determine for themselves what happened, which categories matter, and which situational contexts make an event meaningful as harm. This is what makes such accounts epistemically distinctive. The theoretical basis for this claim can be understood through standpoint epistemology. Harding argues that inquiry beginning from the lives of marginalised groups can produce stronger rather than weaker objectivity because it reveals social structures that remain obscured from dominant standpoints [30]. Collins similarly argues that the outsider-within position generates forms of knowledge that are structurally inaccessible to those who observe oppression from outside it [31]. The relevance of these arguments here is direct. If harm is recognised and interpreted through situated experience, then researcher-authored stimuli cannot fully encode that knowledge, regardless of how carefully they are designed or how participatory the annotation process that follows may be. Kraft and Soulier (2024) extend this insight to NLP by showing that claims of neutrality in AI development are themselves situated within knowledge hierarchies, and that systems trained on data produced within those assumptions inherit their blind spots [32]. Felkner et al. (2023) provide empirical support for this claim in a bias-reduction setting [33]. They compare QueerNews, which consists of externally authored text about LGBTQ+ communities, with QueerTwitter, which consists of text produced primarily by LGBTQ+ community members using hashtags associated with LGBTQ+ activism and identity, collected via the Twitter API. Fine-tuning on QueerTwitter reduces anti-queer bias more effectively across sixteen models. The difference is consistent with an authorial standpoint explanation rather than topic coverage alone, since both corpora concern the same community and harm domain, though register and corpus composition cannot be fully ruled out as contributing factors. We do not claim that accounts are free of framing, since authors write within editorial and literary

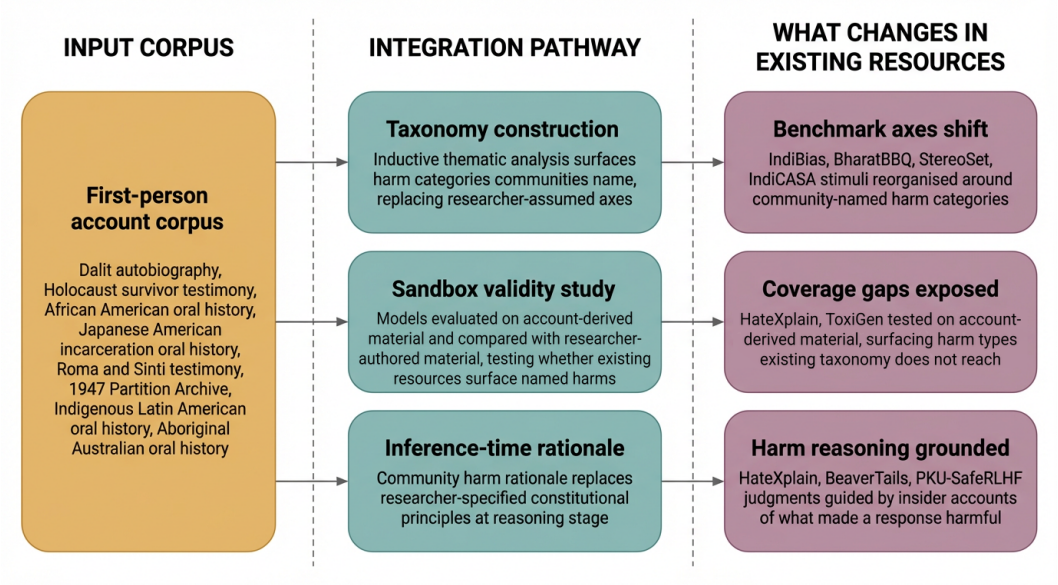


Figure 2: Overview of the three integration pathways, tracing first-person account corpora to what changes in existing social harm evaluation resources at each stage. A worked example appears in Appendix A.

conventions and archives reflect who could publish (Section 5). The claim is that they are free of research framing, because no measurement design fixed the harm categories before the author wrote.

Embedded Rationale Structure The second property of first person accounts is embedded rationale structure, though that phrase understates what is actually at stake. First-person accounts do not merely report that a harm occurred or explain why it counted as harm. They surface experiential states that are structurally invisible to researcher-constructed stimuli because those states do not exist within the categories researchers know to look for. Dalit autobiographies, for instance, document the pain of being turned away from a temple, the accumulated humiliation of having one's language corrected across years of institutional encounters, and the way caste discrimination sediments across relationships and time into something that cannot be captured in a single event or a substituted label [34, 35]. These are not richer descriptions of harms researchers already recognise. They are configurations of harm that researcher design cannot anticipate because anticipating them requires a different relationship to the experience being measured. A researcher who has not lived through caste discrimination lacks the embodied interpretive framework through which such configurations become recognisable as harm in the first place. Poerwandari (2021) characterises such accounts as privileged knowledge because they convey the contextual embeddedness of lived experience in ways external observation cannot reproduce [36]. Charmaz (2006) similarly argues that self-authored accounts are shaped by reflexive processing through which authors interpret events, situate them within social structures, and connect them across time [37]. That reflexive processing is precisely what researcher-constructed stimuli cannot capture, because determining in advance what counts as a relevant event forecloses the interpretive range of the data before authorship begins. This property is directly relevant to the performance of harm evaluation systems. Mathew et al. (2021) demonstrate that models trained with human rationales explaining why content is harmful reduce unintended bias more effectively than models trained on labels alone [7], with the improvement deriving from the explanatory character of the signal rather than additional data volume. But first-person accounts go further than rationale-enriched annotation because the rationale is not added after the fact by an annotator responding to a researcher-constructed stimulus. It is embedded in the structure of the account itself, produced by the person who experienced the harm, in the language and interpretive frame they chose, without reference to any ML task. Memoir, testimony, and autobiography are produced outside research infrastructure entirely, which is what makes them a differently framed signal from anything generated within a research-led pipeline.

4 Recommendations and Proposals for Future Integrations

Sections 2 and 3 identified three structural limitations in current social harm evaluation resources and argued that first-person accounts possess epistemic properties that directly address them. The three recommendations that follow each take up a specific limitation, draw on a corresponding property of first-person accounts, and have implications across evaluation benchmarks, harm classifier training corpora, and safety preference datasets. They are summarised in Figure 2 and described below.

Harm Taxonomy Construction from First-Person Accounts The frame constraint identified in Section 2 arises before annotation begins, so greater annotator diversity within an existing pipeline cannot by itself resolve it. What is needed instead is a different upstream signal, one whose categories were not fixed in advance by a research team. As argued in Section 3, first-person accounts are relevant here because they are authored outside research infrastructure and may therefore register harm categories that emerge from experience rather than from prior measurement design. **We therefore propose using first-person accounts to inform harm taxonomy construction** before any stimulus design or scenario specification takes place. **One possible implementation** would analyse a corpus of first-person accounts through hybrid human-LLM inductive thematic analysis [38, 39, 40], identifying recurring harm-relevant themes and iteratively refining them with human experts who validate, merge, and interpret the resulting categories. The output would be a harm taxonomy grounded in what community members themselves identify as harmful, together with the situational contexts and interpretive rationales through which those harms become meaningful, rather than in what researchers assume harm will look like. Because LLM-assisted thematic analysis performs better on concrete descriptive themes than on subtle culturally embedded ones [40], expert human involvement would remain essential throughout, and any resulting taxonomy should be validated with community members before it structures downstream collection. The model plays a bounded role in this pathway. The model performs a first pass of open coding and proposes candidate codes, each traceable to its source passage [40, 41], and human experts then merge, discard, refine, and resolve disagreements before anything is accepted. Related work has derived a harm taxonomy from first-person accounts through such LLM-assisted analysis [42]. Circularity would arise only if a model decided what counts as harm, so we recommend recording the model and prompts, retaining the code-to-passage link, reporting rejected codes, and giving non-English material additional expert review. **The value of this recommendation can be illustrated across resource types.** Resources such as IndiBias and BharatBBQ currently operationalise caste and social bias in Indian contexts through substitution templates and question-answer formats researchers already know how to build [5, 43], and a taxonomy derived from Dalit autobiography would shift what these benchmarks measure, organising stimuli around how caste harm is actually experienced rather than around what researchers assumed it would look like, surfacing harms like being treated as an electoral resource rather than a political subject that no researcher template would anticipate. For harm-classifier corpora such as HateXplain, which structures annotation through categories derived from platform policy rather than what communities have named as harmful [7], a taxonomy derived from archives such as the HistoryMakers Digital Archive[2], whose oral histories document racial harms that fall outside platform policy categories but remain central to lived experience, would reorient those schemas toward harm configurations currently outside their frame. Similarly, safety preference datasets like BeaverTails construct harm scenarios from researcher imagination of what discrimination looks like [9], and drawing on accounts from the Densho Digital Repository[3], whose oral histories document how institutional responses can be harmful and retraumatising even when formally correct, could inform scenario design with a harm criterion no researcher could have specified in advance. In each case the point is not simply to add new examples but to shift the upstream taxonomy through which harm becomes measurable.

Sandbox Validity Studies Using First-Person Account Material The construct-validity problem identified in Section 2 cannot be resolved by designing better evaluation resources alone, and the instability documented by Goldfarb-Tarrant et al. [25], Delobelle et al. [26], and Selvam et al. [27] suggests the problem lies in the construction logic itself. What is needed is a reference that stands outside the researcher-controlled pipeline whose validity is being assessed. First-person accounts may serve this role because they are written without relation to a benchmark, an annotation task,

[2] https://www.thehistorymakers.org

[3] https://densho.org

or an evaluation protocol, and they preserve harm categories as their authors articulated them. **We therefore propose sandbox validity studies** that change one thing about a standard evaluation run, namely the material on which the model is evaluated, with the aim of revealing where existing resources fail rather than producing new ones. **One possible implementation** would evaluate a model once on the researcher-authored material that a benchmark already uses and once on material drawn from first-person accounts. Researchers and community members would then compare what each run reveals, establishing which harms described in the accounts the benchmark reaches, which it does not, and which appear in the accounts with no counterpart in the benchmark at all.

The sandbox does not ask a model to imitate, represent, or speak for members of any community, a practice that Agnew et al. [18] show produces an illusion of inclusion. The accounts, not a model, determine what counts as harm, and the judgment of whether a benchmark reaches a described harm is made manually by researchers and community members rather than delegated to a model. We use the term simulation for the evaluation environment only to signal that the material should be an unfolding situation rather than a static pair. Bama's *Karukku* [44], for instance, describes humiliation that accumulates across ordinary institutional encounters, which a fixed sentence pair removes. Consistent with this, Lum et al. [45] find that decontextualised bias metrics do not correlate with bias in realistic, context-specific use. Vignettes developed from the same material are a lower-risk alternative instrument.

The diagnostic value of such studies lies in what failures reveal. If a benchmark repeatedly failed to register as harmful the situations that first-person accounts describe as discriminatory, this would provide evidence of the construct-validity failure that Jacobs and Wallach [28] identify as a structural source of harm in computational systems. If a harm classifier failed to detect harm in account-derived material, its taxonomy may not cover harms communities have named, and if a safety preference framework scored as acceptable the responses that accounts describe as harmful, researcher-defined criteria would diverge from insider harm knowledge. **The diagnostic purpose is consistent across resource types.** One might ask whether StereoSet can surface the configurations of dehumanisation described in USC Shoah Foundation survivor testimonies,[4] whether classifiers trained on ToxiGen detect the accumulated harms described in Roma and Sinti testimonies, or whether PKU-SafeRLHF would rank as harmful the responses that oral histories in the 1947 Partition Archive[5] would describe as minimising communal trauma, the misalignment that Harvey et al. [46] identify as a persistent barrier to harm measurement. In each case the sandbox reveals the gap between what existing resources measure and what communities have named as harm.

First-Person Harm Rationale as an Inference-Time Reasoning Layer The agency constraint identified in Section 2 differs from the frame and construct-validity problems because it concerns what kinds of harm knowledge survive participation itself. Even where community members take part in annotation, the structured and researcher-mediated conditions of that participation restrict what can be expressed, so the contextual configuration that made an event harmful, and the interpretive judgment through which it became recognisable as harm, drop out of the resulting data. As argued in Section 3, first-person accounts preserve this rationale structure as part of the genre itself. On this basis, **we suggest using first-person accounts to inform an inference-time harm-rationale layer, supplementing or in some settings replacing researcher-derived constitutional principles with community-authored accounts of what harm looks and feels like in practice**. **One possible implementation would extract harm rationales from first-person account corpora and encode them as the natural-language principles that guide chain-of-thought harm reasoning** [47]. Extraction would proceed through the same hybrid human-LLM thematic analysis proposed in the taxonomy construction pathway, with community members validating which rationale formulations accurately represent the harm configurations in the source accounts before they are encoded as principles. Recent work on collective constitutional AI shows that principles derived from community input rather than researcher specification produce meaningfully different alignment behaviour [48], and that such principles can be systematically extracted from preference data rather than hand-specified [49], providing direct technical grounding for this pathway. Instead of a researcher specifying an abstract principle such as avoiding content that demeans people on the basis of group membership, the principle could be derived from an account of someone who experienced that demeaning, preserving the contextual, relational, and interpretive features through which the event became meaningful as harm. These community-derived rationales could then supplement or replace researcher-specified principles

[4] `https://sfi.usc.edu/vha`

[5] `https://www.1947partitionarchive.org`

at the point where the model reasons about whether an output is harmful. This matters because Constitutional AI-style systems depend on the quality and cultural grounding of the principles they encode [47]. When those principles are drawn from researcher assumptions or legal abstractions, they reproduce both the frame constraint and the agency constraint of Section 2. Harvey et al. (2025) [46] identify a related problem in deployment practice, showing that publicly available harm-measurement instruments often fail to capture the harms practitioners need to assess because they are misaligned with harms that arise in deployed systems. Mathew et al. (2021) [7] further suggest the value of rationale-rich signals by showing that models trained with human rationales explaining why content is harmful reduce unintended bias more effectively than models trained on labels alone. **The potential relevance of this approach becomes clearer when considered across different resource types**. For evaluation benchmarks, one might ask how a resource such as IndiCASA, which measures social bias in Indian contexts using categories fixed before community input [24], would change if its evaluative rationales were informed by texts such as Bama's *Karukku*, which documents caste, religion, and institutional humiliation from within lived experience [44]. For harm-classifier corpora, one might imagine grounding rationale annotations in archives such as EMHAI [6], whose oral histories document Islamophobia and harmful language in institutional and everyday settings, rather than relying only on annotator paraphrases of researcher-defined categories in datasets such as HateXplain [7]. For safety preference datasets, one might similarly ask how resources such as BeaverTails or PKU-SafeRLHF [9, 10] would change if preference judgments were informed by accounts in collections such as AIATSIS [7], where Aboriginal and Torres Strait Islander testimony documents forms of colonial and institutional harm that current taxonomies do not capture. Across these cases, the aim would be the same, namely to bring the explanatory structure of community-authored harm knowledge into the reasoning process at the point where harm judgments are made.

A Worked Example Appendix A traces both pathways through Dalit autobiographies such as Bama's *Karukku* [44]. In the taxonomy trace, an LLM proposes candidate codes and human experts and community members decide which stand. In the sandbox trace, a benchmark such as IndiCASA [24] is evaluated once on researcher-authored material and once on situations drawn from the accounts, and the recorded coverage gap is the measurement.

Scope and Complementarity For an account of a harm to exist, the harm must already have occurred and been documented widely enough to support analysis. Accounts therefore lag deployment, and the proposals cannot anticipate harms from genuinely novel capabilities. Most current harm testing is equally reactive. Our argument concerns the epistemic quality of the reactive process, and for novel capabilities with no community documentation, researcher-constructed stimuli and red teaming remain the available instruments, with first-person accounts complementary to them. Accounts of harms from deployed products tend to be held by advocacy organisations, legal challenges, and public inquiries rather than oral history projects.

5 Ethical Considerations and Preconditions

The recommendations in Section 4 assume that first-person accounts can be identified, accessed, and inducted into ML pipelines. The conditions below are preconditions for any implementation.

Partnership and Consent Before Corpus Selection Testimony, autobiography, and oral history are typically produced for purposes unrelated to ML, and repurposing them raises questions that exceed standard IRB frameworks. Jo and Gebru [50] argue that responsible sociocultural data collection treats consent, power, and inclusivity as structural features of practice. The CARE Principles for Indigenous Data Governance [51] cannot be assessed by a researcher working alone, since collective benefit, authority to control, responsibility, and ethics all require engagement with the community concerned. Local Contexts labels [52] let an institution and its community state what an item may be used for. Consent sought after a project has been designed is generally not consent, so partnership must precede corpus selection. Whether an archival agreement permits the repurposing we describe is a question for the archive and its community, and where the answer is no, the corpus should not be inducted. Data statements [53] documenting who produced the data, under what conditions, and with what consent arrangements should be completed before induction begins.

[6] `https://www.everydaymuslim.org`
[7] `https://aiatsis.gov.au/collections`

Selection, Editorial Bias, and Translation Published memoir and archived testimony are filtered by publishing access, archival selection, and survivorship, and they carry editorial, generational, and literary framing. Communities lacking publishing or archival support are underrepresented in the corpora NLP methods process most readily, so the approach risks reproducing the coverage gaps participatory ML aimed to correct. This differs in kind from the construction bias of Section 2, which operates on the frame itself before any community member is involved. Selection and editorial bias shape which accounts exist, while the harm categories still emerge from what the author chose to write. It limits coverage and register, and it is a provenance question that can be documented. Translation, relevant for Roma and Sinti testimony, the 1947 Partition Archive, and AIATSIS collections, introduces framing of its own and should be recorded in the data statement.

Underdocumented Communities and Algorithmic Harms For small communities with few published accounts, institutional collections offer a route, including life stories published by charities and refugee archives [54, 55]. For harms inflicted by ML systems themselves, relevant accounts are held by advocacy organisations, legal challenges, and public inquiries. UK asylum decision making illustrates this, since AI-generated summaries of asylum interviews [56] have been subject to legal scrutiny [57].

Such records are produced under institutional framing, and charity relationships can moralise who counts as deserving and so shape what is recorded [58]. We therefore treat epistemic autonomy as a matter of degree. Institutional records exist independently of any ML evaluation but remain shaped by the producing institution's mandate. The partnership protocols of Dutta and Jensen [59] and Ghoshal et al. [60] address the risk that such records become data reservoirs. We recommend that data statements record the producing institution and its purpose, that implementers draw on more than one institutional type, including community-organised and mutual aid sources, and that community members validate derived categories.

Dual Use Kaffee et al. [61] define dual use in NLP as the intentional harmful reappropriation of research artefacts, and find that few researchers take concrete steps to mitigate it. The infrastructure proposed here could be used to extract detailed maps of the vulnerabilities and social configurations of marginalised communities for surveillance or discrimination. A taxonomy derived from Dalit autobiography would describe the mechanisms of caste discrimination in detail, and sandbox material could be used to learn how to evade the classifiers it tests. We therefore set the following conditions. Corpus access is governed by agreements with the communities concerned, including withdrawal if misuse becomes evident. Derived taxonomies are treated as community-sensitive and not freely shareable. Sandbox material is not released as open infrastructure without mechanisms to monitor downstream use. Implementers complete a dual-use assessment [61].

6 Alternative Views

Researcher Positionality and Reflexivity Already Encode Insider Harm Knowledge in Dataset Construction, Which Sufficiently Grounds the Epistemic Frame. A growing practice requires NLP researchers to document their positionality, including their social identity and lived experience of the harm domain they study. If a researcher from a marginalised community designs a harm benchmark, the argument runs, their own embodied knowledge already grounds the stimulus categories in insider experience. On this view, positionality statements render separate induction of first-person account corpora unnecessary. We take this seriously because positionality-aware research is a genuine and important development. However, this argument conflates community identity with community standpoint. Táíwò (2022) identifies this as elite capture: the individuals who navigate the social selection pressures of academic institutions to become researchers are, by that very fact, a highly unrepresentative subset of the communities they belong to [62]. Those who make it into research institutions have done so through pathways that are structurally inaccessible to most members of their communities, and that process of institutional mobility reshapes one's relationship to ground-level experience in ways that positionality statements cannot fully recover. This is not merely a theoretical concern. Santy et al. (2023) find persistent misalignment between researcher positionality and dataset alignment with the communities researchers belong to, even across diverse researcher positionalities, suggesting that the epistemic distance elite capture produces is empirically traceable in the resources researchers build [63]. Positionality informs design choices but does not change the text type, and the epistemic properties of first-person accounts are properties of the text, not of the identity of the

researcher who commissions it. Positionality statements are a valuable complement to this paper's proposal. They are not a substitute for it.

Participatory Workshops for Benchmark Construction Already Enable Community Members to Define Harm Categories from First Principles, Which Sufficiently Addresses the Frame and Agency Constraints. A second objection proposes that participatory workshops in which community members collaboratively define harm scenarios and evaluation criteria from scratch would provide the same epistemic grounding as first-person accounts, while affording consent, controlled provenance, and community ownership. We acknowledge that well-designed workshops represent a genuine improvement over standard benchmark construction. The practical evidence on what workshops produce in Global South contexts is, however, sobering. Irani et al. (2010) show that participatory design methods developed in Western HCI contexts face fundamental challenges of cultural epistemology and uneven power relations when they travel to new constituencies, with the assumptions embedded in workshop formats often reproducing the very knowledge hierarchies they aim to disrupt [64]. Harrington et al. (2019), drawing on empirical case studies with underserved communities, argue that participatory design as instantiated in the workshop is an affluent and privileged activity, identifying specific barriers including community access constraints, skepticism rooted in histories of research injustice, and the risks of collecting personal narratives in researcher-mediated settings [65]. Winschiers-Theophilus et al. similarly document that achieving meaningful participation in cross-cultural contexts requires in-depth local knowledge and mutual learning that structured workshop formats rarely afford [66].

First-Person Accounts Are Ethically Appropriate Only for Auditing, Not for Training or Inference. We accept the weight of this objection and partly adopt it. The sandbox pathway is auditing, and the inference-time pathway is where the conditions of Section 5 apply most strictly. An implementation that cannot meet them should be limited to auditing.

Governance, Incident Reports, and Language Models as Alternatives. Governance and text type are complementary, since governance applied to researcher-constructed stimuli still leaves the frame fixed in advance, and text type without governance reproduces extraction. Incident reports from deployed systems are more direct evidence of model-mediated harm and are one of the channels through which first-person accounts of algorithmic harm already exist (Section 5). Persona-conditioned models as stand-ins for communities produce an illusion of inclusion [18], and narrative grounding improves their fidelity [67], which makes the accounts the component doing the work.

7 Conclusion

Participatory ML correctly diagnosed that data produced without affected communities encodes its developers' worldviews, yet it has left intact the question of what kind of text should serve as the raw signal for social harm evaluation. Researcher-constructed stimuli fix the epistemic frame before community involvement begins. We argue that first-person accounts offer a complementary upstream modality that is free of research framing, though not of framing in general, and we outline three pathways for bringing them into the ML safety pipeline. These pathways are reactive, depend on partnership and consent that precede corpus selection, and sit alongside red teaming for harms no community has yet documented.

8 Acknowledgment

Atmadeep Ghoshal has been partially supported by Acorda Labs, UK.

References

[1] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July 2020. Association for Computational Linguistics.

[2] Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. Participatory research for low-resourced machine translation: A case study in African languages. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online, November 2020. Association for Computational Linguistics.

[3] Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. Aya model: An instruction finetuned open-access multilingual language model. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[4] Siobhan Mackenzie Hall, Samantha Dalal, Raesetje Sefala, Foutse Yuehgoh, Aisha Alaagib, Imane Hamzaoui, Shu Ishida, Jabez Magomere, Lauren Crais, Aya Salama, and Tejumade Afonja. The human labour of data work: Capturing cultural diversity through world wide dishes. *Proc. ACM Hum.-Comput. Interact.*, 9(7), October 2025.

[5] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online, November 2020. Association for Computational Linguistics.

[6] Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online, August 2021. Association for Computational Linguistics.

[7] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection, 2022.

[8] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[9] Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of LLM via a human-preference dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[10] Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Juntao Dai, Boren Zheng, Tianyi Qiu, Jiayi Zhou, Kaile Wang, Boxuan Li, Sirui Han, Yike Guo, and Yaodong Yang. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference, 2025.

[11] Milagros Miceli, Julian Posada, and Tianling Yang. Studying up machine learning data: Why talk about bias when we mean power? *Proc. ACM Hum.-Comput. Interact.*, 6(GROUP), January 2022.

[12] Personal Narratives Group. *Interpreting women's lives*. Indiana University Press, Bloomington, MN, January 1989.

[13] Alessandro Portelli. *The death of Luigi trastulli and other stories*. SUNY series in Oral and Public History. State University of New York Press, Albany, NY, January 1991.

[14] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 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*, EAAMO '23, New York, NY, USA, 2023. Association for Computing Machinery.

[15] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. Participation is not a design fix for machine learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, New York, NY, USA, 2022. Association for Computing Machinery.

[16] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. Power to the people? opportunities and challenges for participatory ai. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, New York, NY, USA, 2022. Association for Computing Machinery.

[17] Michael Feffer, Michael Skirpan, Zachary Lipton, and Hoda Heidari. From preference elicitation to participatory ml: A critical survey & guidelines for future research. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 38–48, New York, NY, USA, 2023. Association for Computing Machinery.

[18] William Agnew, A. Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R. McKee. The illusion of artificial inclusion. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[19] Miranda Fricker. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press, 06 2007.

[20] Rosalind Morris, editor. *Can the subaltern speak?* Columbia University Press, New York, NY, September 2009.

[21] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online, August 2021. Association for Computational Linguistics.

[22] Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. On measures of biases and harms in NLP. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang, editors, *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 246–267, Online only, November 2022. Association for Computational Linguistics.

[23] Nihar Sahoo, Pranamya Kulkarni, Arif Ahmad, Tanu Goyal, Narjis Asad, Aparna Garimella, and Pushpak Bhattacharyya. IndiBias: A benchmark dataset to measure social biases in language models for Indian context. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8786–8806, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[24] Santhosh G S, Akshay Govind S, Gokul S Krishnan, Balaraman Ravindran, and Sriraam Natarajan. Indicasa: A dataset and bias evaluation framework for llms using contrastive embedding similarity in the indian context. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1):978–989, October 2025.

[25] Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. Intrinsic bias metrics do not correlate with application bias. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online, August 2021. Association for Computational Linguistics.

[26] Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1693–1706, Seattle, United States, July 2022. Association for Computational Linguistics.

[27] Nikil Selvam, Sunipa Dev, Daniel Khashabi, Tushar Khot, and Kai-Wei Chang. The tail wagging the dog: Dataset construction biases of social bias benchmarks. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1373–1386, Toronto, Canada, July 2023. Association for Computational Linguistics.

[28] Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 375–385, New York, NY, USA, 2021. Association for Computing Machinery.

[29] Laura Cabello, Anna Katrine Jørgensen, and Anders Søgaard. On the independence of association bias and empirical fairness in language models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 370–378, New York, NY, USA, 2023. Association for Computing Machinery.

[30] Linda Alcoff and Elizabeth Potter, editors. *Feminist Epistemologies*. Thinking Gender. Routledge, London, England, September 2013.

[31] Patricia Hill Collins. *Black feminist thought*. Perspectives on Gender. Routledge, London, England, September 2008.

[32] Angelie Kraft and Eloïse Soulier. Knowledge-enhanced language models are not bias-proof: Situated knowledge and epistemic injustice in ai. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1433–1445, New York, NY, USA, 2024. Association for Computing Machinery.

[33] Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. WinoQueer: A community-in-the-loop benchmark for anti-LGBTQ+ bias in large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9126–9140, Toronto, Canada, July 2023. Association for Computational Linguistics.

[34] Omprakash Valmiki and Arun Prabha Mukherjee. *Joothan*. Columbia University Press, New York, NY, June 2008.

[35] Manu Bhagavan. Book review: Urmila pawar, *The Weave of My Life: A Dalit Woman's Memoirs* and ravikumar, *Venomous Touch: Notes on Caste, Culture and Politics*. *The Indian Economic & Social History Review*, 49(4):606–609, December 2012.

[36] Elizabeth Kristi Poerwandari. Minimizing bias and maximizing the potential strengths of autoethnography as a narrative research. *Japanese Psychological Research*, 63(4):310–323, 2021.

[37] Kathleen C Charmaz. *Constructing grounded theory*. Introducing Qualitative Methods Series. SAGE Publications, Thousand Oaks, CA, January 2006.

[38] Maria Antoniak, David Mimno, and Karen Levy. Narrative paths and negotiation of power in birth stories. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

[39] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291, March 2024.

[40] Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. LLM-in-the-loop: Leveraging large language model for thematic analysis. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9993–10001, Singapore, December 2023. Association for Computational Linguistics.

[41] Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhang, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. Collabcoder: A lower-barrier, rigorous workflow for inductive collaborative qualitative analysis with large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[42] Jianfeng Zhu, Karin G. Coifman, and Ruoming Jin. Understanding risk and dependency in ai chatbot use from user discourse, 2026.

[43] Aditya Tomar, Nihar Ranjan Sahoo, and Pushpak Bhattacharyya. Bharatbbq: A multilingual bias benchmark for question answering in the indian context, 2025.

[44] Bama Faustina and Lakshmi Holmstrom. *Karukku*. OUP, Oxford, England, 2 edition, May 2014.

[45] Kristian Lum, Jacy Reese Anthis, Kevin Robinson, Chirag Nagpal, and Alexander D'Amour. Bias in language models: Beyond trick tests and toward ruted evaluation, 2025.

[46] Emma Harvey, Emily Sheng, Su Lin Blodgett, Alexandra Chouldechova, Jean Garcia-Gathright, Alexandra Olteanu, and Hanna Wallach. Understanding and meeting practitioner needs when measuring representational harms caused by LLM-based systems. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18423–18440, Vienna, Austria, July 2025. Association for Computational Linguistics.

[47] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022.

[48] Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. Collective constitutional ai: Aligning a language model with public input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1395–1417, New York, NY, USA, 2024. Association for Computing Machinery.

[49] Arduin Findeis, Timo Kaufmann, Eyke Hüllermeier, Samuel Albanie, and Robert D. Mullins. Inverse constitutional AI: Compressing preferences into principles. In *The Thirteenth International Conference on Learning Representations*, 2025.

[50] Eun Seo Jo and Timnit Gebru. Lessons from archives: strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 306–316, New York, NY, USA, 2020. Association for Computing Machinery.

[51] Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. The care principles for indigenous data governance. *Data Science Journal*, 19, 2020.

[52] Jane Anderson and Kim Christen. Local contexts: Traditional knowledge and biocultural labels. https://localcontexts.org.

[53] Emily M Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

[54] Ali Smith, Abdulrazak Gurnah, Chris Cleave, Marina Lewycka, Jade Amoli-Jackson, Patience Agbabi, Inua Ellams, Stephen Collis, Michael Zand, and Dragan Todorovic. *Refugee Tales*. Comma Press, Manchester, England, June 2016.

[55] Living Refugee Archive and Refugee Council Archive, University of East London. Living refugee archive. https://www.livingrefugeearchive.org.

[56] Home Office. Evaluation of AI trials in the asylum decision making process. Research and analysis, GOV.UK, April 2025.

[57] Robin Allen, Dee Masters, and Joshua Jackson. The use of artificial intelligence tools by government: A case study of the Home Office's asylum practice. Joint legal opinion, Open Rights Group, March 2026.

[58] Dean Spade. Solidarity not charity: Mutual aid for mobilization and survival. *Social text*, 38(1):131–151, 2020.

[59] Arshia Dutta and Rikke Bjerg Jensen. Security at the border? the lived experiences of refugees and asylum seekers in the uk. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA, 2026. Association for Computing Machinery.

[60] Atmadeep Ghoshal, Anasmita Ghoshal, Volodymyr Shevchenko, Arshia Dutta, Ruba Abu-Salma, Martim Brandao, et al. Position: Responsible ai for ai companions must actively combat violence toward intimate partners. In *Forty-third International Conference on Machine Learning Position Paper Track*, 2026.

[61] Lucie-Aimée Kaffee, Arnav Arora, Zeerak Talat, and Isabelle Augenstein. Thorny roses: Investigating the dual use dilemma in natural language processing. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13977–13998, Singapore, December 2023. Association for Computational Linguistics.

[62] Olúfmi O Táíwò. *Elite capture*. Pluto Press, London, England, May 2022.

[63] Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. NLPositionality: Characterizing design biases of datasets and models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada, July 2023. Association for Computational Linguistics.

[64] Lilly Irani, Janet Vertesi, Paul Dourish, Kavita Philip, and Rebecca E. Grinter. Postcolonial computing: a lens on design and development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, page 1311–1320, New York, NY, USA, 2010. Association for Computing Machinery.

[65] Christina Harrington, Sheena Erete, and Anne Marie Piper. Deconstructing community-based collaborative design: Towards more equitable participatory design engagements. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

[66] Heike Winschiers-Theophilus, Nicola J. Bidwell, and Edwin Blake. Community consensus: Design beyond participation. *Design Issues*, 28(3):89–100, 2012.

[67] Pranav Narayanan Venkit, Yu Li, Yada Pruksachatkun, and Chien-Sheng Wu. The need for a socially-grounded persona framework for user simulation, 2026.

A Worked Example

Suppose the corpus consists of Dalit autobiographies such as Bama's *Karukku* [44] and Omprakash Valmiki's *Joothan* [34]. The taxonomy pathway and the sandbox pathway run independently, and the taxonomy produced by the first does not feed the second.

In the taxonomy trace, an LLM performs a first pass of open coding across the corpus and proposes candidate codes, each linked to the passage that motivated it. Human experts check each candidate against its source passage, merge overlapping codes, discard those that misread the experiential or cultural context, refine labels, and resolve disagreements. Community members then confirm whether the categories reflect their experience. The output is a set of harm categories named in the accounts themselves, for instance the way exclusion operates spatially within institutions, or the way a person's language is corrected until it becomes a marker of disqualification. These categories illustrate the form of the output and are not findings of an analysis we have run.

In the sandbox trace, take IndiCASA [24], which evaluates a model on stereotype and anti-stereotype pairs written by the research team. The model is evaluated once on that material and once on situations drawn from the same accounts, such as the accumulation of institutional encounters in *Karukku*. Researchers and community members compare the two runs to establish which harms in the accounts the benchmark reaches, which it does not, and which have no counterpart in it. Jacobs and Wallach [28] describe construct validity failure as the distance between a construct and its operationalisation. Here the construct is harm as the accounts name it, the operationalisation is the benchmark built on researcher-authored material, and the recorded coverage gap is the measurement. The result is a mapped limitation of the existing benchmark and the avenues it would need to address, and no new benchmark is produced.

Resource	**Authors**	**Category**	**Construction method**
CrowS-Pairs	Nangia et al. (2020) [5]	Evaluation benchmark	Minimal sentence pairs generated through demographic label substitution within researcher-defined templates
StereoSet	Nadeem et al. (2021) [6]	Evaluation benchmark	Researcher-generated sentence completion tasks across gender, profession, race, and religion
IndiBias	Sahoo et al. (2024) [23]	Evaluation benchmark	Constructed by filtering and translating CrowS-Pairs into Indian languages and social contexts
IndiCASA	Santhosh et al. (2025) [24]	Evaluation benchmark	Builds on IndiBias using contrastive embedding similarity across Indian demographic dimensions
BharatBBQ	Tomar et al. (2025) [43]	Evaluation benchmark	Adapts the BBQ question-answering format to Indian socio-cultural contexts across 8 languages and 13 bias categories
WinoQueer	Felkner et al. (2023) [33]	Evaluation benchmark	Community-sourced benchmark developed through a survey of LGBTQ+ community members to generate stereotype statements
HateXplain	Mathew et al. (2021) [7]	Harm classifier corpus	Crowdworkers annotate social media posts for hate speech, target communities, and rationale spans using researcher-defined categories
ToxiGen	Hartvigsen et al. (2022) [8]	Harm classifier corpus	Machine-generated implicit hate speech targeting 13 demographic groups with a small human-validated subset
BeaverTails	Ji et al. (2023) [9]	Safety preference dataset	Researcher-specified harm scenarios used to prompt language models; crowdworkers rank responses according to researcher-defined safety criteria
PKU-SafeRLHF	Ji et al. (2025) [10]	Safety preference dataset	19 harm categories and three severity levels defined by researchers before annotation; annotators rank responses within this pre-constituted structure

Table 1: Social harm evaluation resources cited across this paper, organised by category. The construction method column describes how stimuli and categories were generated.