Wikibench: Community-Driven Data Curation for AI Evaluation on Wikipedia

Tzu-Sheng Kuo
tzushenk@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Aaron Halfaker
aaron.halfaker@gmail.com
Microsoft
Redmond, WA, USA

Zirui Cheng*
chengzr19@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Jiwoo Kim*
jk4671@columbia.edu
Columbia University
New York, NY, USA

Meng-Hsin Wu*
soniawu2302@gmail.com
Carnegie Mellon University
Pittsburgh, PA, USA

Tongshuang Wu
sherryw@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Kenneth Holstein†
kjholste@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Haiyi Zhu†
haiyiz@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

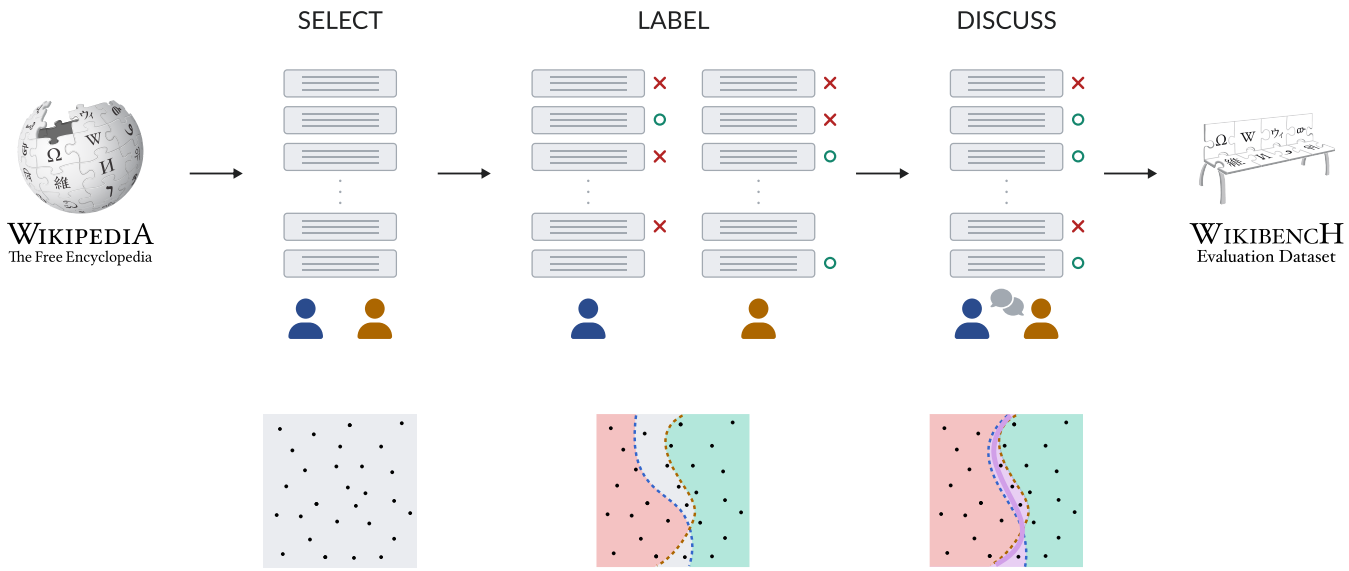

Figure 1: An overview of Wikibench's approach to supporting community-driven data curation. The top row illustrates community members' use of Wikibench to *select* data points (e.g., edits on Wikipedia) for inclusion in the dataset, *label* data points with "individual" labels based on their own initial judgments, and then *discuss* their perspectives and collectively decide on a "primary" label for the data point. The bottom row represents data points in a conceptual 2D space. As each community member labels data points, their labels form *decision boundaries* in aggregate (orange and blue dotted curves). Through discussion, participants may resolve some disagreements or clarify ambiguities in labeling, leading to changes in their individual labels. In addition, community members decide on a primary label for each data point, forming a consensus-based decision boundary (purple curve). Wikibench datasets preserve information about disagreement among community members (purple shaded region). The Wikipedia logo is licensed by Wikimedia Foundation, CC BY-SA 3.0, via Wikimedia Commons.

*These authors, listed in alphabetical order by last name, contributed equally.
†Co-senior authors contributed equally.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642278

ABSTRACT

AI tools are increasingly deployed in community contexts. However, datasets used to evaluate AI are typically created by developers and annotators outside a given community, which can yield misleading conclusions about AI performance. How might we empower communities to drive the intentional design and curation of evaluation datasets for AI that impacts them? We investigate this question on Wikipedia, an online community with multiple AI-based content moderation tools deployed. We introduce Wikibench, a system that enables communities to collaboratively curate AI evaluation datasets, while navigating ambiguities and differences in perspective through discussion. A field study on Wikipedia shows that datasets curated using Wikibench can effectively capture community consensus, disagreement, and uncertainty. Furthermore, study participants used Wikibench to shape the overall data curation process, including refining label definitions, determining data inclusion criteria, and authoring data statements. Based on our findings, we propose future directions for systems that support community-driven data curation.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing systems and tools**.

KEYWORDS

community-driven AI, data curation, AI evaluation, Wikipedia

ACM Reference Format:

Tzu-Sheng Kuo, Aaron Halfaker, Zirui Cheng, Jiwoo Kim, Meng-Hsin Wu, Tongshuang Wu, Kenneth Holstein, and Haiyi Zhu. 2024. Wikibench: Community-Driven Data Curation for AI Evaluation on Wikipedia. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3613904.3642278

1 INTRODUCTION

AI tools are increasingly deployed in *community contexts*. For example, AI-based content moderation tools have been deployed in online communities such as Wikipedia and Reddit [46, 55]. AI-based decision-making tools have also been adopted by local governments to prioritize public services, such as allocating local housing resources [64, 86]. However, the datasets used to evaluate AI performance are typically designed, curated, and labeled by developers and data annotators outside of a given community, which can lead to misleading conclusions about AI systems' "fit for use" [40]. In turn, the deployment of poorly-fit AI tools can yield compromised user experiences or even cause harm to vulnerable populations [74, 81, 84]. For example, research shows that crowdsourced datasets systematically label innocuous phrases in African American English (AAE) dialects as toxic [83]. As a consequence, if such datasets were used to prospectively evaluate content moderation tools' fit for use in a community that uses AAE, they would *underestimate* the tools' false positive rates, compared with what the community would experience in deployment [84].

Given that what constitutes "good performance" on tasks such as content moderation can be highly community-specific, recent work has argued that HCI and machine learning research should explore more *community-driven* approaches to AI dataset development. For instance, in a position paper, Jo and Gebru [58] propose that AI should draw lessons from archive and library studies, where archives are often directly contributed and curated by the communities they are meant to represent, instead of by community-outsiders. These community archives, such as the Feminist Archive and the Working Class Movement Library, are motivated by the need to represent the voices of non-elites and the marginalized [32]. The authors argue that these traditions should inspire new approaches to AI *data curation* that allow communities greater voice in specifying their collective desires for AI performance.

In the context of AI evaluation, *data curation* refers to the process of designing the "ground truth" against which AI models' performance will be evaluated [74]. This involves an intentional process of selecting *which data points* should be included in a dataset and, in the case of labeled datasets, deciding *how each data point should be labeled* [73]. For example, when developing an AI dataset for content moderation tools on Wikipedia, a "data point" could be an edit to an article, and its "label" could be a judgment of whether the edit should be considered "damaging" to the article or not [46]. The intentional curation of AI evaluation datasets stands in stark contrast with what Jo and Gebru [58] term "laissez faire" approaches to dataset development, which indiscriminately take data in masses by crawling trace data on the web [40]. On their own, such datasets simply capture how people *have behaved* in the past. However, they often fail to capture communities' normative beliefs about how decisions *should be made*, for evaluation purposes [58, 81].

Realizing the vision of *community-driven data curation* of AI datasets in practice poses numerous open challenges. For example, while a community may share broad norms and values [14, 31], individual community members may disagree about how specific data points should be labeled (e.g., whether a given post should be considered "toxic") [12, 43, 44]. In some cases, these disagreements may represent substantive differences in perspective, while in other cases, a brief discussion between individuals could reveal that they actually agree more than they disagree [16, 74]. Current approaches to account for annotator disagreement in crowdsourced datasets tend to handle disagreements post-hoc (after data have already been labeled), either by resorting to the majority vote [30] or by attempting to model individual subjectivity for re-weighted voting [4, 20, 26, 42, 93]. However, when it comes to deciding how important community decisions should be made, it is crucial that community members have opportunities to collectively build meaning and understand each other's perspectives. In contrast to prior methods, this calls for more collaborative and deliberative approaches that allow community members agency in navigating disagreements, via processes that are perceived to be fair by community members [67]. Furthermore, beyond selecting and labeling individual data points, it is critical to provide communities with the agency to shape higher-level decisions, such as crafting label definitions and determining data inclusion criteria. Finally, given that community members will generally have limited time and attention to contribute to the curation of AI datasets, it is important to support them in prioritizing their efforts. To the best of our knowledge, despite recent calls-to-action from the research community [22, 23, 58, 73, 74, 81], there are no existing tools

aimed at addressing these challenges to support the intentional, community-driven curation of AI datasets in practice.

We identify and address these challenges in the context of Wikipedia, an online community where multiple AI-based content moderation tools have been deployed, but where community members currently have limited means to prospectively assess these tools' fit for use. Through formative interviews with Wikipedia community members and AI developers, we derived a set of design requirements for systems that aim to support community-driven data curation. Based on these design requirements, we then developed Wikibench, a system that enables community members to collaboratively curate AI evaluation datasets, while navigating disagreements and ambiguities through discussion. As illustrated in Figure 1, community members can use Wikibench to select data points for inclusion in datasets, label data points with "individual labels" reflecting their personal judgments, and discuss their perspectives to decide upon a "primary label" for the data point. Through a field study on Wikipedia, we find that datasets curated using Wikibench can effectively capture community consensus, disagreement, and collective uncertainty. We demonstrate how Wikibench datasets can help in understanding areas of alignment and misalignment with community perspectives. Furthermore, we gain insight into the ways Wikipedia community members collaborate using Wikibench. We find that participants in our study used Wikibench to proactively shape the overall data curation process beyond labeling data, including refining label definitions, determining data inclusion criteria, and authoring data statements.

Overall, this work demonstrates the potential of *community-driven* data curation, and contributes the following:

- **System**: We introduce Wikibench, the first system that supports community-driven curation of AI datasets.
- **Field study**: We present findings from a field study on Wikipedia to understand how Wikipedia community members interact with this system to collaboratively curate evaluation datasets.
- **Future directions**: Based on our findings, we propose future directions for HCI systems that support community-driven data curation within and beyond the context of Wikipedia.

In the rest of the paper, we first review relevant literature and introduce our study context (Section 2–3). Next, we present the design requirements for Wikibench and walk through the system (Section 4–5). We then describe our evaluation of Wikibench (Section 6–9), and conclude with future directions (Section 10).

2 RELATED WORK

2.1 Developer-Centric AI Evaluation

AI datasets are commonly created by developers and data annotators with limited knowledge about the real-world contexts in which these AI models will be deployed, assuming a "one-dataset-fits-all" approach to evaluation [81]. For example, widely used datasets for toxicity classification of online comments, such as Jigsaw's Toxic Comments [18] and Civil Comments [8], are commissioned by AI developers and labeled by crowd workers. While these datasets are commonly framed as benchmarks of progress toward general abilities, such as "toxicity detection," researchers argue that they are often ineffective for evaluating how an AI model will perform in real-world contexts [81]. One reason general benchmarks can fail is that many tasks currently targeted by AI models are inherently norm- and value-laden [22, 43, 44, 60, 61]. For instance, a comment that is considered "inappropriate" in the context of one online community may be within bounds of acceptability for those in a different community, with different norms and values [83]. As a result, a one-dataset-fits-all approach can yield misleading conclusions when used to prospectively evaluate how well an AI model will perform in a particular community context [80, 83, 84]. These concerns have informed a recent line of research that directly involves end-users in AI evaluation and benchmark development.

2.2 Broadening Participation in AI Evaluation

As dataset issues cascade down to deployment [82], end-users have often surfaced instances of AI misbehavior through their everyday use [24, 27, 28, 72, 77, 85]. For example, Twitter users discovered that the platform's image cropping algorithm favored light-skinned over dark-skinned individuals when both are in an image [95]. Similarly, Halfaker and Geiger [46] document how various Wikipedia language communities have engaged in ad-hoc, bottom-up efforts to identify language-specific error patterns in Wikipedia's AI-based content moderation tools. As acknowledged by AI developers, end-user involvement in testing and auditing AI behavior can be extremely valuable [21]. Given end-users' situatedness in specific contexts where AI tools will be used, they can often surface issues that would otherwise be missed [85]. Recently, the HCI community has proposed several systems that support individual users in testing and auditing AI behavior [13, 65, 66]. Facilitating user collaboration remains an area for further exploration [24].

Beyond collecting evidence of AI misbehavior, several efforts have focused on the creation of *new benchmark datasets* by challenging crowdworkers and volunteers to uncover AI models' blind spots and then adding these instances to their new evaluation datasets [1–3, 5, 25, 62, 76, 92]. For example, the CATS4ML Data Challenge asked challenge participants to submit misclassified Google Open Images to create a new evaluation dataset [1]. Similarly, the Adversarial NLI benchmark was created by challenging crowdworkers to draft text snippets that existing AI models could not understand [76]. Dynabench and DataPerf are centralized platforms that host several of these data challenges [62, 71].

Existing approaches to broadening involvement in AI evaluation, such as those overviewed above, differ from our vision of community-driven AI evaluation in several ways. First, these approaches have typically focused on engaging end-users, crowdworkers, and other volunteers in identifying cases where specific AI models misbehave, rather than in proactively specifying what behavior and performance they *want* to see from AI models. Second, in current approaches, individuals work independently or in competition with one another, rather than collaboratively. Finally, current approaches tend to recruit broadly, without a focus on capturing perspectives held by *particular communities*. In the following subsection, we briefly overview existing scholarship relevant to the vision of *community-driven* AI evaluation.

2.3 Community-Driven AI Evaluation

An emerging body of research has advocated for empowering communities to shape the design of AI evaluation datasets [22, 23, 58, 74, 81]. For instance, Jo and Gebru [58] argue that AI dataset development should learn from the rich traditions of *community-driven* curation of archives in library studies. Because community archives are motivated by the need to represent non-elites and marginalized voices [32], Jo and Gebru [58] argue that they can serve as a model for how the design of AI datasets might be opened up for community input. Yet realizing the vision of *community-driven* AI data curation in practice poses numerous open challenges. For example, a growing body of work in HCI and machine learning has argued that the notion of a single, objective ground truth label often does not apply when AI is deployed in complex social contexts, where different groups may have distinct perspectives [16, 22, 43, 60, 61, 74, 87]. In some cases, these disagreements may stem from genuine differences in perspective, while in other cases, a brief discussion between individuals could reveal that they actually agree more than they disagree [16, 74]. However, current approaches to account for annotator disagreements tend to handle disagreements *post-hoc* after individual labels have been gathered [4, 20, 26, 42, 93], instead of facilitating discussion and deliberation among annotators. When deciding how important decisions should be made in community contexts, it is critical that community members have opportunities to discuss, understand each other's perspectives, and collectively build meaning [74]. This calls for more collaborative, deliberative approaches that allow community members agency in navigating disagreements, through processes that they perceive to be fair and appropriate [67, 68, 90].

We note that while some online platforms have existing community-driven *content curation* mechanisms, their purpose is distinct from AI dataset curation. For example, Reddit and Stack Overflow have implemented community voting systems to enable the curation of high-quality posts [41, 69]. Similarly, Wikipedia allows community members to revert damaging edits in order to maintain article quality [39]. These *content curation* mechanisms differ from *AI data curation* in two aspects. First, in the context of AI evaluation, *data curation* refers to an intentional process of designing the "ground truth" against which AI models' performance will be evaluated [74]. These datasets aim to represent community members' collective beliefs about what constitutes "good performance" on a given task (e.g., content moderation) in the context of their community. In contrast, votes, reverts, and other trace data generated through existing content curation processes can carry complex meanings, which will often be misaligned with the goals of an AI evaluation [45, 79]. For instance, on Reddit and Stack Overflow, posts may receive downvotes for reasons unrelated to the violation of the community's content moderation policy [33]. Similarly, Wikipedia edits can be reverted for reasons beyond causing damage to an article [48, 63]. Relying on these trace data as proxy labels for AI evaluation can introduce target variable bias, leading to misleading evaluations of AI performance [45]. Second, in the same vein, while past work has often used historical human decisions as ground truth for evaluating AI-based decision-making tools, these trace data capture only how decisions *have been made* in the past—biases, errors, and all—not how a community believes decisions *should be made* [40]. Therefore, recent work has argued for the necessity of intentionally curated evaluation datasets, to support meaningful and reliable AI evaluations [22, 23, 58, 74, 81]. Despite these differences, Section 10 will discuss how tools for community-driven AI data curation may draw inspiration from existing content curation mechanisms.

Beyond recent calls-to-action for the research community, to our knowledge no tools currently exist to address the challenges described above to support intentional, community-driven curation of AI datasets in practice. The current work is the first system in the literature aimed at supporting community-driven AI data curation.

3 STUDY CONTEXT

We conduct this study in the context of Wikipedia for several reasons. First, Wikipedia has a rich history of grassroots engagement to explore new modes of participation in AI development and evaluation [46, 88], as mentioned in Section 2.2. However, while community members are motivated to improve the AI-based tools they use and are impacted by, there is currently no infrastructure to support them in *proactively* curating datasets for AI evaluation and improvement. Thus, our research focus is well-aligned with existing interests and motivations among Wikipedia community members, and this context presents an opportunity to develop a system that is truly useful to the community. In addition, the Wikipedia context has established norms for collaborative efforts (e.g., for article editing) [7, 34, 63]. Our focus on Wikipedia enables us to build upon these existing community norms when exploring new mechanisms for community-driven data curation, thus bypassing the need to develop and introduce entirely new collaboration processes.

The remainder of this section briefly discuss the current AI evaluation challenges faced by the Wikipedia community and our positionality as researchers working with community members. In the remainder of this paper, we refer to Wikipedia's community members as "Wikipedians," following community terminology.

3.1 Challenges of AI Evaluation on Wikipedia

As Wikipedia scales, the community increasingly relies on AI tools for governance [38, 47, 75]. For example, AI-based content moderation tools are used to identify damaging edits in articles for Wikipedians to review and revert them as necessary [37, 39, 47]. Among various content moderation tools, ORES, an AI model hosting system, is used extensively in English Wikipedia and many other languages [46]. Using basic estimation, Halfaker and Geiger [46] argue that without ORES's AI model for detecting damaging edits, it would take 483 labor hours per day to review the 290k edits made to all the various language editions of Wikipedia, but with an AI model, that workload can be reduced by 90%.

Despite the growth of AI tools, Wikipedia communities currently have limited means to evaluate particular AI tools' "fit for use" with respect to their collective norms and values. Currently, the curation of ORES' training and evaluation datasets relied on a system called Wikilabels[1], which is hosted on an external website outside Wikipedia. Wikipedians can join *data labeling campaigns* on Wikilabels and request a subset of data to label. However, unlike Wikipedia where each article is editable by any Wikipedian, Wikilabels assigns each data point to only a single Wikipedian for

[1] https://meta.wikimedia.org/wiki/Wiki_labels

labeling in isolation. Wikilabels also doesn't enable Wikipedians to discuss labels collaboratively, unlike Wikipedia, where each article has an associated talk page[2] for discussing its content. As such, it is less clear to what extent Wikilabels' datasets reflect the collective perspectives of the community, versus just the perspectives of some individual annotators. Besides labeling preselected data in Wikilabels, Wikipedians currently do not have a way to proactively evaluate how different AI tools perform with respect to their collective norms and values.

3.2 Positionality and Ethical Considerations

Our research has dual objectives. First, we are interested in exploring new approaches to support community-driven data curation. Second, we hope our research can truly benefit the Wikipedia community, in recognition that Wikipedia is not a laboratory[3] [50]. These dual objectives guided our decision-making throughout the study, from research method to system design. In cases where these two objectives conflict, we prioritize the community's needs, preferences, and established norms over our research interests [46, 52]. We took several precautions to ensure that we conducted ethical research on Wikipedia[4]. For example, we collaborated with an experienced Wikipedian deeply involved in the development of AI tools in Wikipedia and an academic researcher with over a decade of experience studying Wikipedia. We also adhered to the norm of researching Wikipedia by creating and iteratively updating a project page on Meta-Wiki[5], where we publicly shared the study's objective, protocol, timeline, recruitment message, and our institutional affiliation for Wikipedians to access. Finally, we recruited a minimal number of Wikipedians for the study, acknowledging that the study might take their time away from their volunteer work on Wikipedia. We hope the benefit of our research has the potential to outweigh the disruptions we inevitably caused.

4 DESIGN REQUIREMENTS

To better understand Wikipedians' desires and challenges around data curation for AI evaluation, we conducted a formative study with eight Wikipedians who had experience *using* AI-based content moderation tools on the platform (e.g., edit patrollers who use AI tools), contributing to the *development* of these tools (as AI data labelers or engineers), or participating in grassroots efforts to *identify areas for improvement* in deployed tools (see Table 1 for an overview). In this phase of our research, we aimed to recruit Wikipedians across multiple language communities, with the goal of understanding desires and challenges on Wikipedia more broadly. To recruit participants, we adopted a snowball sampling approach. We first recruited a Wikipedian who had been heavily involved in AI development on Wikipedia. This Wikipedian then helped us reach out to a broader set of Wikipedians by pinging them on the associated talk page of our research page on Meta-Wiki, where Wikipedians could view our study description before signing up. Seven additional Wikipedians who self-identified with at least one of the five roles we targeted (Table 1) signed up either through a form we provided or by using Wikipedia's email feature. In total, we conducted synchronous interviews with seven Wikipedians and exchanged emails with one (W8) based on their preference. The interview was semi-structured and lasted for an hour with a $30 USD compensation. Some participants declined our compensation, viewing the study as part of their volunteer work to improve Wikipedia. Our interview questions are shown in Appendix A.1.

Through a reflexive thematic analysis [9, 10] and affinity diagramming by two of the authors, we derived the following four highest-level themes as design requirements for systems that aim to support the community-driven data curation process for AI evaluation on Wikipedia. We briefly summarize each requirement below:

D1: The data curation process should be led by the community and follow their established norms. Participants suggested that systems for community-driven data curation should provide communities with the agency to shape the overall data curation process, beyond labeling individual data points. In addition, participants emphasized that in order to succeed, the systems need to be flexible and adaptable to the varying norms of different Wikipedia language communities: *"If we're talking about Wikipedia, make sure it adapts to the local rules. English [Wikipedia] are full of categories, [whereas] in the Dutch Wikipedia it's almost a crime to have more than ten categories"* (W6).

D2: The data curation process should encourage deliberation to surface disagreements, build consensus, and promote shared understanding. In Wikilabels, data points were labeled by individual Wikipedians working in isolation. However, our participants argued that data curation systems should instead promote deliberation, similar to existing processes on Wikipedia for article editing. Participants highlighted the importance of deliberation due to the subjectivity of data labeling and potential disagreements among community members. Participants suggested building upon Wikipedia's existing deliberation interface and mechanisms, such as talk pages and associated norms, to build consensus for collective decision-making while ensuring individual viewpoints are fully considered. Participants also anticipated that these deliberations can have side effects that benefit the community, such as revealing otherwise hidden disagreements, gaining insights from each other, and collectively strengthening the community.

D3: The data curation process should embed within existing workflows. Participants believed that a key reason Wikilabels did not see sustained use was because it was hosted on an external website and required Wikipedians to leave their workflows on Wikipedia. For example, even though Wikipedians were *already* reviewing edits on Wikipedia, the design of Wikilabels required them to duplicate this effort by labeling edits as damaging or not using Wikilabels only for the purpose of data labeling. Participants wished to embed the data curation process into their existing workflow on Wikipedia: *"Capturing people's judgments while they're working is like sticking a waterwheel in a river. The river is already flowing, we should take advantage of that"* (W1). They also anticipated this could help curate up-to-date data instead of labeling historical data pre-selected by AI engineers when using Wikilabels: *"We're going to continue to get new data, so we can update the models and continue to re-evaluate them"* (W7).

[2] https://enwp.org/WP:TALKPAGE
[3] https://enwp.org/WP:NOTLAB
[4] https://enwp.org/WP:Ethically_researching_Wikipedia
[5] https://w.wiki/7QGb

Table 1: The targeted roles for recruitment in the formative study and the participants who self-identified with these roles.

Role	Role Description	Participant ID
Community organizer	organize community efforts around AI on Wikipedia	W1, W4, W5, W6
AI evaluator	have participated in efforts to identify and report AI errors	W1, W2, W6
AI user	use AI tools for their daily work on Wikipedia	W1, W2, W5, W8
AI engineer	develop AI tools and/or organize data labeling campaigns	W1, W3, W5, W7
AI data labeler	contribute to data labeling campaigns	W1, W4, W5

D4: The data curation process should be public and transparent to community members. Currently, Wikipedians have a limited and narrow view of the data curation process in Wikilabels, where each contributor only sees the data they labeled. Participants suggested that data curation systems should instead make entire datasets public and easily accessible (like most content on Wikipedia) to facilitate community-driven AI evaluation: *"If I'm doing my own audit, I'm not quite sure if other people are having the same problems and benefits of this model that I am. But if we can all put it in a repository together, and have some mechanism to make sense of what's in there, then I can know how this is working for everybody. We can make decisions together about whether we want this or not in our community"* (W1).

5 WIKIBENCH

Based on these design requirements, we developed Wikibench, a system that enables community members to collaboratively curate AI evaluation datasets, while navigating disagreements and ambiguities through discussion. Wikibench supports these processes through the workflow illustrated in Figure 2. As shown in this figure, community members can use Wikibench to *select* new data points for inclusion in datasets and to *label* these data points during the course of their regular, daily activities on Wikipedia (via a plug-in). Wikibench also supports community members in filtering through data points that have already been added to the dataset, to *select* ones to further *label* or *discuss*. Through Wikibench, community members are supported in either discussing the label of individual data points, or discussing higher-level topics related to the overall data curation process.

Wikibench is designed to capture community consensus, disagreement, and uncertainty. Wikibench records two types of labels for each data point:

- **Individual Label**: Each community member can provide their unique individual label that is meant to reflect their own perspective. This label is editable only by themselves and may differ from others' labels.
- **Primary Label**: Community members can collectively determine a primary label that is intended to reflect a "consensus" view.

Together, the individual and primary labels allows Wikibench datasets to reflect both community consensus and differing viewpoints that may underlie that consensus. Wikibench also records labelers' self-reported **confidence** associated with each individual label. In aggregate, confidence indications can provide a signal of the uncertainty associated with a data point.

In this section, we overview Wikibench's design through the specific example of dataset curation to support the evaluation of AI-based content moderation tools on Wikipedia, which are used to counter vandalism. In this context, each data point is an article edit on Wikipedia. Following the design of existing AI tools and datasets on Wikipedia, each edit has two associated labels: *edit damage*, which specifies whether the given edit is viewed as "damaging" to the article's quality, and *user intent*, which specifies whether the edit is viewed as having been made in good or bad faith. The distinction between edit damage and user intent follows the prior data labeling campaign hosted on Wikilabels (Section 3.1), considering that damaging edits made with good intent are not considered to be vandalism[6] on Wikipedia.

In the following subsections, we describe how Wikibench's three user interfaces on Wikipedia: plug-in, entity page, and campaign page, support data curation. Throughout this section, we link specific features of Wikibench's design to the design requirements described in the previous section, denoted as (D1)—(D4). Finally, we conclude with implementation details.

5.1 Plug-in: Select and Label New Data Points

Wikipedians can use Wikibench's plug-in to select and label new edits during their regular patrolling activities on Wikipedia (**D3**). Specifically, Wikipedians who self-identify as patrollers[7] regularly patrol edits on Wikipedia's Recent Changes page[8] and assess selected edits by opening its "diff" page[9]. Wikibench embeds a plug-in on these diff pages so that Wikipedians can label edits while they are already in the midst of assessing them, as shown in Figure 3. The plug-in also allows Wikipedians to specify the confidence level for their labels and include notes if desired. Overall, this design embeds the data curation process into Wikipedia's existing workflow to reduce duplication of effort and help curate up-to-date data.

After Wikipedians submit their labels, Wikibench's plug-in encourages them to engage in discussion when labeling disagreements arise (**D2**). In particular, if an individual's submitted label differs from the existing primary label of an edit, the plug-in will display the yellow message in Figure 4 to encourage discussion. Otherwise, the green message will appear to minimize disruption to Wikipedians' regular patrolling activities. These messages encourage deliberation only when disagreements occur and minimize disruptions to existing workflows otherwise.

[6] https://enwp.org/WP:VAND
[7] https://enwp.org/WP:RCP
[8] https://enwp.org/Special:RecentChanges
[9] https://enwp.org/WP:DIFF

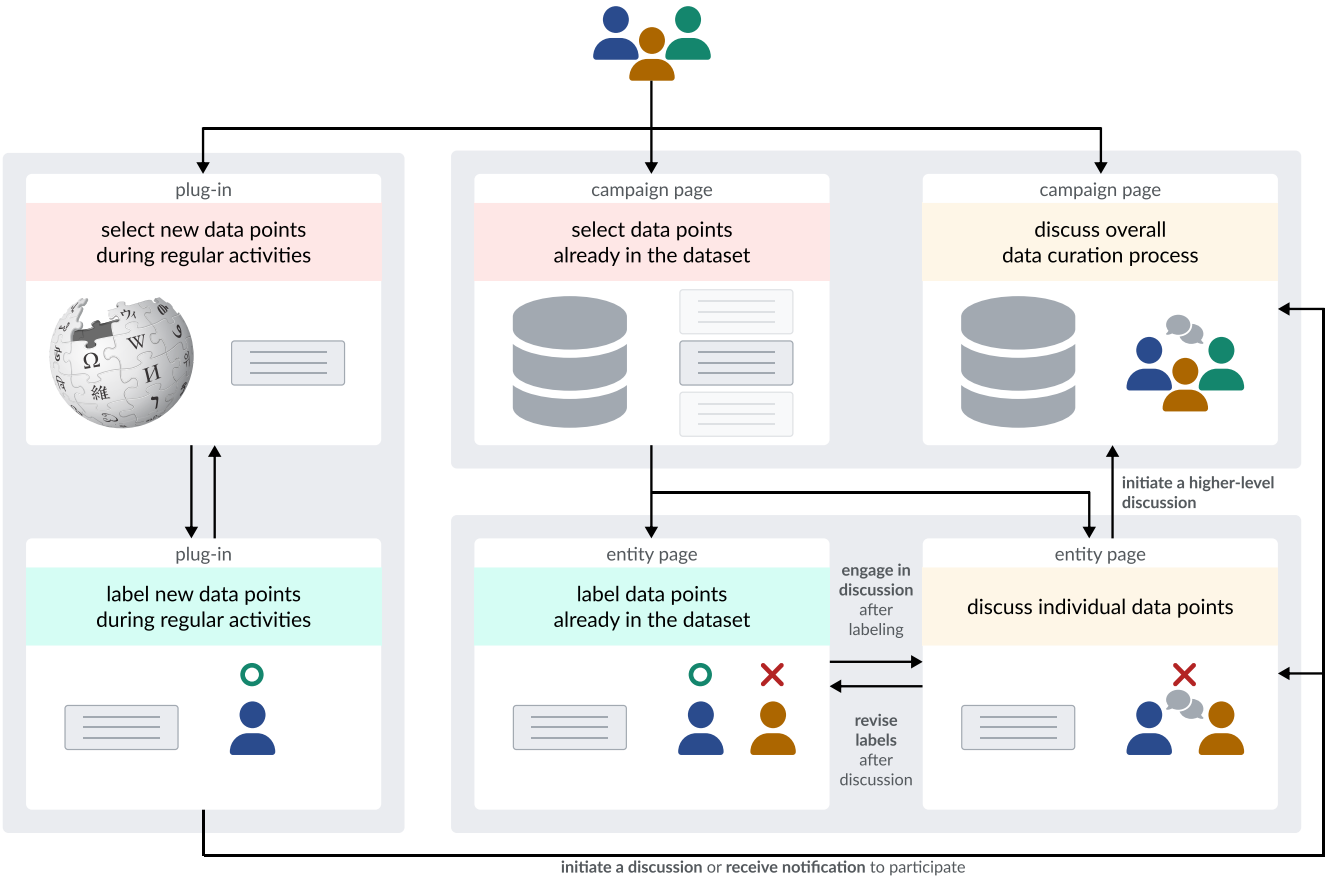

Figure 2: Wikibench's workflow. Wikibench mainly supports three actions for community-driven data curation: select, label, and discuss, each illustrated by different colors. Community members can select and label data points during their regular activities (i.e., while patrolling for damaging edits on Wikipedia) or choose from data points already collected in the dataset. They can also discuss individual data points to resolve disagreements or initiate a higher-level discussion related to the overall data curation process.

5.2 Entity Page: Label and Discuss Collected Data Points

Wikibench's entity pages publicly show the labels of individual edits and facilitates discussions and (re-)labeling (**D2**, **D4**). As shown in Figure 5, the top half of each entity page shows the edit, its primary label, and the user's individual label. The bottom half shows the full set of individual labels submitted by the community so far, along with any brief notes that community members may have included as rationale. This view is intended to help Wikipedians quickly understand the current level of disagreement associated with a given edit, and to examine how their own views align with or differ from others'. If Wikipedians think discussion on a given edit could be helpful, each entity page has a corresponding talk page for deliberation. This design resembles Wikipedia's article editing mechanisms, where each article has a corresponding talk page to discuss the article's content.

The mechanism by which Wikipedians choose the primary label for an edit through Wikibench is based upon the Wikipedia community's established norms for consensus-building (**D1**, **D2**). Similar to Wikipedia articles, the primary label is initially set to the value of the first submitted individual label. From that point on, it is open to modification by any Wikipedian. Wikibench does not automatically assign primary labels based the majority of individual labels, because Wikipedia follows the principle that "polling is not a substitute for discussion" when it comes to consensus building[10]. When disagreements arise, Wikibench's design explicitly encourages Wikipedians to employ their well-established consensus-building processes[11] (e.g., the bold, revert, and discuss cycle[12]), to boldly edit primary labels and engage in discussions when others disagree with the changes (see Figure 6). When the primary label is changed, Wikibench also notifies previous labelers, to facilitate discussion as needed.

[10]https://enwp.org/WP:POLL
[11]https://enwp.org/WP:CON
[12]https://enwp.org/WP:BOLD

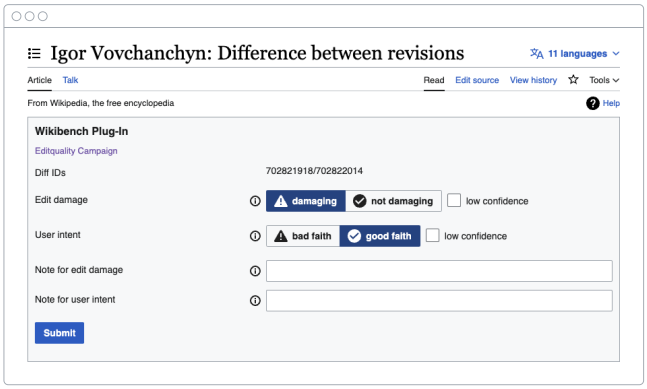

Figure 3: Wikibench's plug-in is embedded in Wikipedia's diff pages, where Wikipedians already assess edits during their regular patrolling activities. Through the plug-in, Wikipedians can label an edit's damage and user intent, specify their confidence level, and add notes if desired.

Figure 4: The message displayed after Wikipedians successfully submit their labels using the plug-in. The yellow message appears only when the submitted labels differ from the current primary label to facilitate discussion.

5.3 Campaign Page: Select Collected Data Points for Labeling and Discussion, or Discuss the Overall Curation Process

The campaign page publicly shows the entire dataset and surfaces edits that could benefit from additional attention, including edits with high disagreement and edits that could benefit from additional labelers (**D2**, **D4**). As shown in Figure 7, each row in the table is a link to an entity page for an edit and its label information. The four buttons above the table assist Wikipedians in sorting the table to easily find edits that may benefit from more labels or discussions. For example, the *provide more labels* button helps Wikipedians find and contribute to edits with fewer individual labels. The *build consensus* button surfaces edits that have high disagreement across community members' individual labels, to promote discussion among community members. The disagreement is measured as the standard deviation of encoded individual labels, with ±1 for damaging/not damaging, ±0.5 for damaging/not damaging submitted with low confidence; likewise for user intent.

The campaign page is also designed to enable Wikipedians to discuss and coordinate about the overall data curation process (**D1**, **D2**). In addition to the table, the campaign page serves as a living datasheet [36] that provides comprehensive information about the data curation campaign, such as label definitions and data statements, as outlined in Figure 7 on the left. Similar to Wikipedia articles, the campaign page can be edited by any Wikipedian and has an associated talk page for discussion.

5.4 Implementation

The current implementation of Wikibench is built upon Wikipedia's infrastructure to ensure its user interfaces and norms are familiar to Wikipedians. In the back-end, both entity and campaign pages, used for storing labels and campaign information, are standard Wikipedia article pages with built-in talk pages. Wikibench uses Wikipedia's user script feature[13] to re-render these article pages on the front-end. The plug-in is also a front-end element embedded in Wikipedia's existing diff page. To ensure that the front-end elements are familiar to Wikipedians and coherent with Wikipedia's existing interface, Wikibench uses Wikipedia's OOUI[14] and design system[15]. The creation and revision of Wikibench's labels are enabled through MediaWiki API[16]. Importantly, we adhere to our positionality statement by keeping Wikibench's campaign and entity pages within an author's user sandbox[17], a designated area for experimentation on Wikipedia, to minimize disruption to the site. This deep integration with Wikipedia also enables Wikipedians to easily use Wikibench by importing Wikibench into their Wikipedia account through a user script[18]. Wikibench is open-sourced on Wikipedia[19] and available to all Wikipedians.

6 EVALUATION STUDY

To understand how Wikipedians use Wikibench in practice, we conducted a two-part evaluation study on English Wikipedia, a highly active and extensively studied Wikipedia language community [11]. Prior research has demonstrated that community needs and norms for content moderation vary across different Wikipedia language communities [46, 51]. In the current study we focus on understanding how Wikibench can support community-driven data curation on one language community[20], before expanding the system to multiple communities.

In the remainder of the paper, we refer to English Wikipedia as "Wikipedia" for simplicity. We first conducted a one-week field study in which participants used Wikipedia to collectively curate a dataset. We then conducted a validation study with a separate set of participants, aimed at understanding whether labels generated collaboratively, through Wikibench, better reflect community consensus than those generated through Wikilabels.

[13] https://enwp.org/WP:JAVASCRIPT
[14] https://www.mediawiki.org/wiki/OOUI
[15] https://design.wikimedia.org/style-guide
[16] https://www.mediawiki.org/wiki/API:Main_page
[17] https://enwp.org/WP:SAND
[18] https://enwp.org/Wikipedia:User_scripts
[19] https://en.wikipedia.org/wiki/User:Tzusheng/Wikibench-Editquality.js
[20] Note that we focus at the level of a *language community* because this is the level at which AI-based content moderation tools are adopted on Wikipedia.

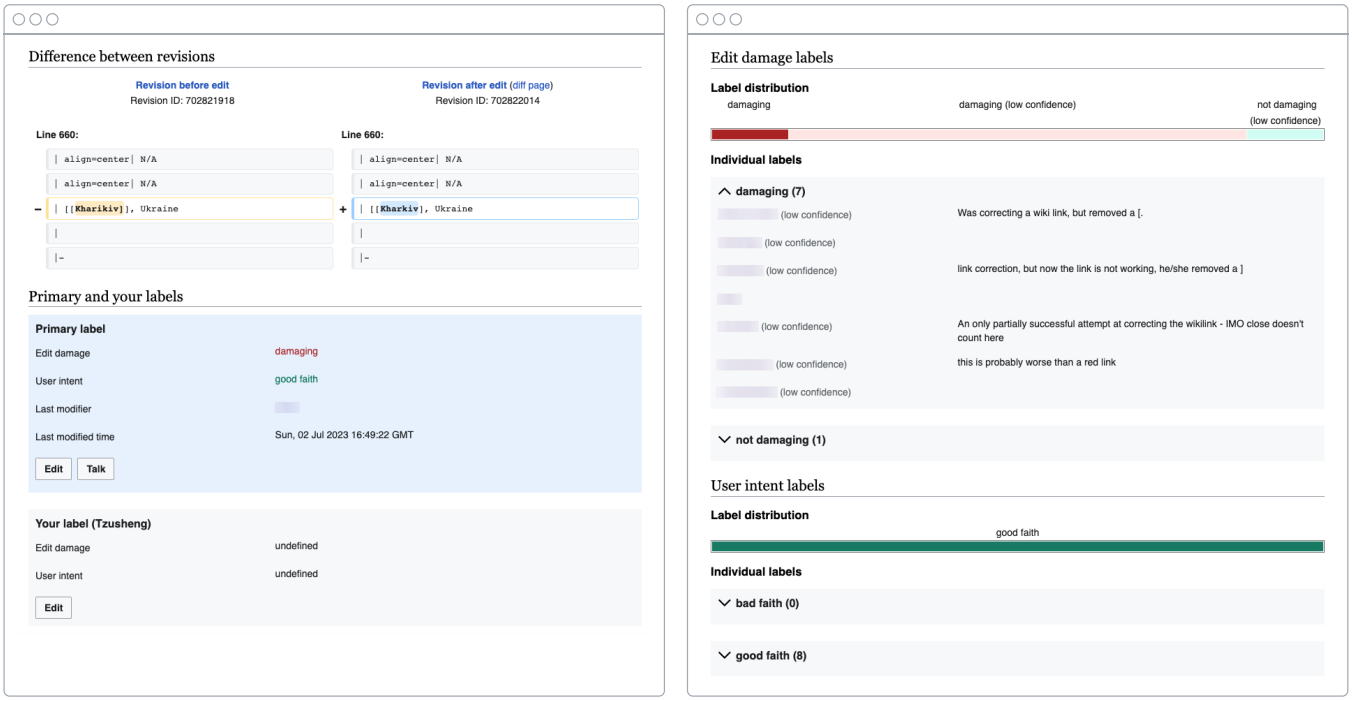

Figure 5: Wikibench's entity page for a given edit. The left side shows the top half of an entity page, featuring the edit, its primary label, and the user's individual label. The right side shows the bottom half of an entity page, containing the full set of individual labels and accompanying notes. Participants' usernames are blurred to avoid identification.

Figure 6: The message displayed when Wikipedians edit primary labels to encourage them to be bold yet respectful of others' views.

6.1 Field Study

We conducted a one-week field study to observe how Wikipedians use Wikibench in the course of their regular activities on Wikipedia.

6.1.1 Study protocol. The study began with a one-hour, one-on-one onboarding session that introduced the study and the data curation campaign. As part of this onboarding, participants imported Wikibench into their Wikipedia user accounts and a researcher walked them through the system's features. At the end of the onboarding, participants received instructions for the week-long field study. We set minimal participation requirements to ensure that (1) participants would have ample opportunities for interaction during the field study period, while also (2) providing participants with flexibility to decide when and how much they want to contribute (cf. [97]). Each participant was asked to submit a minimum of 10 labels and to engage in at least 3 discussions per day using Wikibench, for 5 days out of the week. Finally, we conducted a 30-minute exit interview with each participant once they completed the field study, to learn about their experiences and gather feedback. As part of this exit interview, each participant was shown a randomly selected set of 10 edits from the dataset, and were invited to explore how Wikibench's labels compare with the predictions of ORES (described in Section 3.1) and a new AI tool[21] that is currently under development by the Wikimedia Foundation (with more details in Section 8.2). Our exit interview protocol is shown in Appendix A.2.1.

6.1.2 Recruitment. We adhered to our positionality statement and followed the norm of researching Wikipedia by including the recruitment message in our main project page on Meta-Wiki, where Wikipedians could review study details before signing up. We then shared this project page with Wikipedians through multiple channels, including English Wikipedia's Village Pump[22], r/wikipedia subreddit, and the Discord servers for the Wikimedia Community and Anti-Vandalism. We also reached out to several Wikipedians

[21] https://wikitech.wikimedia.org/wiki/Machine_Learning/LiftWing
[22] https://enwp.org/WP:VP

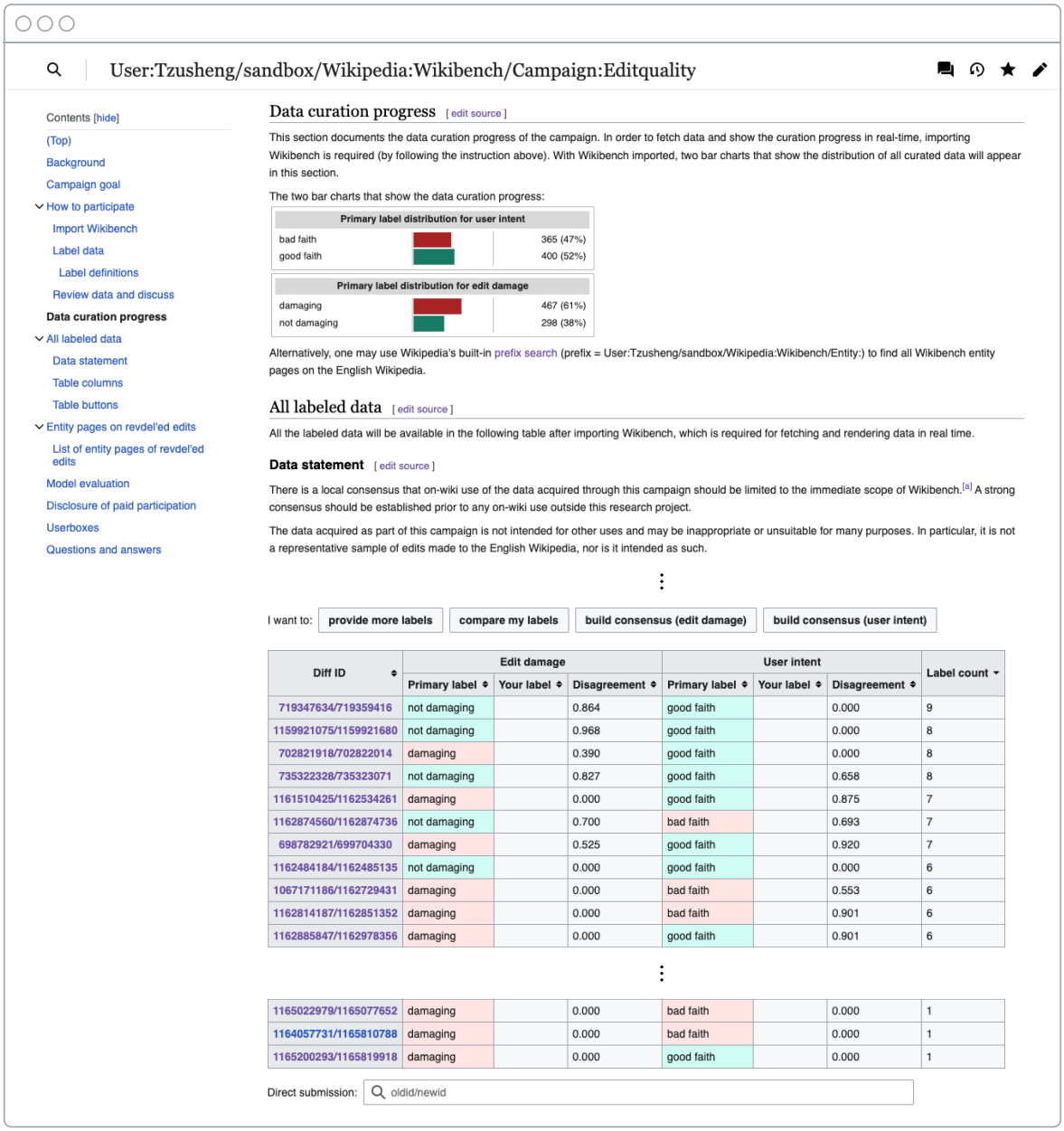

Figure 7: An excerpt of Wikibench's campaign page. The section at the top shows simple visualizations to help Wikipedians track the progress of a data curation campaign. The section at the bottom includes a table that helps Wikipedians navigate the entire dataset. The buttons above the table allow Wikipedians to sort the table and identify edits that may benefit from additional labels or discussion, including edits for which the current primary label differs from their own individual label.

who were actively patrolling edits by leaving messages on their user talk pages—an approach that aligns with existing norms for communication on Wikipedia.

In total, we recruited 12 Wikipedians with diverse experiences and backgrounds, as shown in Table 2. We conducted onboarding sessions and exit interviews via Zoom, with the exception of one participant who preferred to participate via text over Discord. We provided $150 USD as compensation, including $30 for onboarding, $100 for the field study ($20 per day for 5 days), and $20 for the exit interview. These compensation amounts align with prior studies on the English Wikipedia [94], as well as prior HCI research that has conducted similar week-long field studies [97]. As in our formative study, some participants declined compensation at the end of the exit interview, viewing the study as part of their voluntary work to improve Wikipedia.

Table 2: Field study participant demographics, including their self-identified experience and frequency of patrolling edits, registration year, edit count on English Wikipedia, and geographic location.

Participant ID	Patrol Experience	Patrol Frequency	Registered Since	Edit Count	Location
P1	Months	Daily	2023	9.1k	United States
P2	Years	Daily	2018	24k	Indonesia
P3	Months	Weekly	2023	1.6k	United States
P4	Years	Daily	2006	48k	Singapore
P5	Years	Daily	2021	2.6k	Germany
P6	Months	Daily	2018	7.4k	Italy
P7	Years	Daily	2013	0.9k	Hungary
P8	Months	Daily	2023	4.2k	United States
P9	Years	Yearly	2013	9.7k	Australia
P10	Years	Daily	2014	23k	United States
P11	Years	Daily	2019	7.1k	Ireland
P12	Years	Daily	2020	18k	United Kingdom

6.2 Validation Study

To understand whether Wikibench helped curate labels that more consistently reflect community consensus, compared with the previous approach (Wikilabels), we conducted a small-scale validation study following the conclusion of the field study. In particular, we recruited a separate group of Wikipedians to collaboratively label a subset of the edits using Wikipedia's default article and talk pages. These participants labeled edits anew, without knowledge of the labels each had previously received through either Wikibench or Wikilabels. While using the default interfaces for labeling and discussion without Wikibench's support was more involved and time-consuming for participants, this approach mirrors the standard process Wikipedians use to reach consensus on article pages. This validation study helped us understand whether Wikibench's primary labels, which are intended to be reflective of community consensus, are indeed aligned with the labels generated through Wikipedia's standard consensus-building process. Our validation study aimed to compare the consensus labels generated through this process, by an independent group of participants, with those generated through Wikibench and Wikilabels.

6.2.1 Edit selection. We first sampled 90 edits that had previously received labels through Wikilabels. We then had field study participants label them using Wikibench during onboarding sessions without being told about the validation study to prevent them from overly focusing on these edits more than they would naturally do. These edits then underwent the standard labeling and consensus-building process in Wikibench. Following the conclusion of the field study, we identified 33 edits where Wikilabels and Wikibench's primary labels differed. The resulting edits were selected for the validation study. Additional details of our sampling procedure of the 90 edits are available in Appendix A.3.1.

6.2.2 Study protocol. We created a standard article page in an author's user sandbox for the validation study. The page provided study purpose, instruction, and compensation information, along with a table where each row was an edit, and each column was available for one Wikipedian to provide their individual label. In addition, we asked Wikipedians to enter their label consensus in two extra columns, one for edit damage and another for user intent. They were also encouraged to use the talk page for discussion. The entire validation study lasted a week to ensure sufficient time for participants to label and discuss edits asynchronously.

6.2.3 Recruitment. We recruited five additional Wikipedians who had signed up or expressed interest in the field study but were unavailable during the study period. We shared the link to the article page we set up for the validation study and provided $90 USD as compensation at the end, with the estimated time required around three hours. As in the prior two studies, some participants declined compensation, viewing their participation as part of their voluntary work to improve Wikipedia. Participant demographics are shown in Table 4 in Appendix A.3.2.

6.3 Data Analysis

We adopted a mixed-method approach to analyze our data. We employed a top-down approach to quantitatively analyze the resulting community-curated dataset for indicators of quality, such as primary label composition, individual label variation, and label contributor diversity. We adopted a reflexive thematic analysis approach [9, 10] to qualitatively analyze both participants' interview data and their interactions with one another through Wikibench. In particular, two authors conducted open coding on 12 exit interviews and all discussions on Wikibench's campaign and entity talk pages. This process resulted in a total of 249 codes. Through iterative discussions, we synthesized higher-level themes using affinity diagramming. In total, we identified 64 first-level themes, 17 second-level themes, and 7 top-level themes through this process. Finally, we triangulated across findings from our quantitative and qualitative analyses. We report the results of this combined analysis in the next sections.

7 OVERVIEW OF FINDINGS

We present findings from our evaluation study in the following two sections. Section 8 examines the quality and properties of the resulting dataset curated using Wikibench (Section 8.1) and explores its use in evaluating different AI models' alignment with

community norms and values (Section 8.2). In total, Wikipedians curated 757 edits using Wikibench, with a relatively balanced primary label composition (61% of primary labels are "damaging" and 48% are "bad faith"), which can be useful in evaluating how well an AI model's decision boundary aligns with community-specified decision boundaries [35]. Overall, we find evidence that the dataset collected through Wikibench is broadly reflective of Wikipedians' perspectives, while also capturing ambiguity and disagreement among community members. We demonstrate how the resulting dataset can be used to investigate the relative strengths and limitations of two different AI models used on Wikipedia. The dataset is publicly available on Wikibench's campaign page and on GitHub[23]. Section 9 describes how participants used Wikibench throughout the study, and how they collectively steered the overall data curation process, in addition to labeling and discussing individual data points. Participants appreciated how Wikibench's design embedded seamlessly into their workflow (Section 9.1). They organically drove the overall data curation process, beyond just labeling data (Section 9.2) and believed that the collaborative approach supported by Wikibench was beneficial both for dataset quality and for community building (Section 9.3).

Taken together, our findings indicate potential for the approach embodied by Wikibench to support the curation of AI datasets that reflect community norms and values.

8 FINDINGS: QUALITY OF THE COMMUNITY-CURATED DATASET

A dataset curated by and for a community for the purpose of AI evaluation should meet the following criteria, based on our design requirements:

- **Label Quality**: Low-quality labels diminish overall dataset quality [53]. From a community standpoint, we consider high-quality primary labels as being reflective of community values, shared understanding, and consensus. This data criterion aligns with Design Requirements **D1** and **D2**.
- **Disagreement and Uncertainty**: Meanwhile, capturing the disagreement and uncertainty behind primary labels is equally important, to ensure that the dataset reflects both (1) substantive differences in perspective across individuals and (2) inherent ambiguity of a given data point [43]. This criterion corresponds to Design Requirement **D2**.
- **Collaborative Labeling**: Ideally, data points should be labeled and curated by multiple community members rather than being dominated by just a few voices. This criterion corresponds to Design Requirements **D2**, **D3**, and **D4**.

In the following subsections, we describe how the dataset curated through Wikibench aligns with the criteria above (Section 8.1) and showcase how the dataset can be used by comparing the community alignment of two AI models currently deployed on Wikipedia (Section 8.2).

8.1 Dataset Quality

8.1.1 ***Label Quality: Wikibench's labels reflect consensus among a broader set of community members.*** To understand whether Wikibench helped curate primary labels that better reflect community consensus, compared with Wikilabels, we examined results of the validation study (Section 6.2). In this study, a set of participants who had not participated in the field study were shown a sample of 33 edits where Wikilabels and Wikibench's labels differed in edit damage (27 edits) and/or user intent (18 edits). Participants in the validation study collectively labeled edits anew through Wikipedia's standard consensus-building process, without knowledge of the labels each edit had received previously through either Wikilabels or Wikibench. Figure 8 shows the counts of labels generated through Wikilabels versus Wikibench that align with the labels produced in the validation study. As shown, compared with labels generated through Wikilabels, Wikibench's primary labels tend to align with the consensus labels generated in the validation study. In exit interviews, field study participants also expressed a belief that Wikibench's collaborative approach would better reflect broader community perspectives: *"I believe that you're getting more of an overall viewpoint from the community itself, whereas that may not have always been the case for Wikilabels"* (P10).

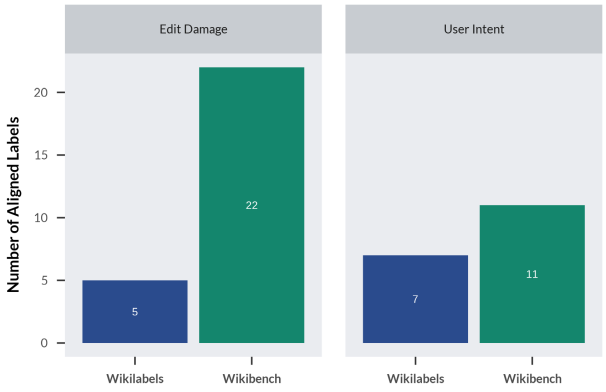

Figure 8: Counts of labels generated through Wikilabels versus Wikibench that align with validation study participants' consensus.

8.1.2 ***Disagreement and Uncertainty: Wikibench captures ambiguity and differences of perspective.*** In addition to the primary labels, Wikibench is designed to capture disagreement and uncertainty that may underlie these collectively-determined labels by allowing each person to provide unique individual labels representing their personal perspective, along with their self-reported confidence. In line with prior research, we use these signals to distinguish between *ambiguity* and *genuine differences in perspective* [16, 43]. Figure 9 shows edits that have multiple individual labels, plotted by labeler disagreement and confidence. For a given edit, overall confidence is measured by the proportion of individual labels that specify low confidence. Disagreement across individual labels is measured as the standard deviation of encoded individual labels, with -1 for damaging and 1 for non-damaging, and likewise for user intent[24].

[23] https://github.com/tskuo/Wikibench

[24] We experimented with various metrics [59] and chose standard deviation for simplicity. All resulting charts, measured by different metrics, closely resemble Figure 9.

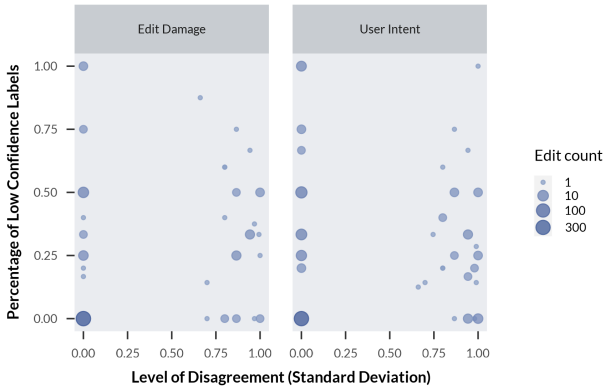

Figure 9: Edits with multiple individual labels, plotted by labeler disagreement (the standard deviation of individual labels) and confidence (the percentage of individual labels submitted with "low confidence").

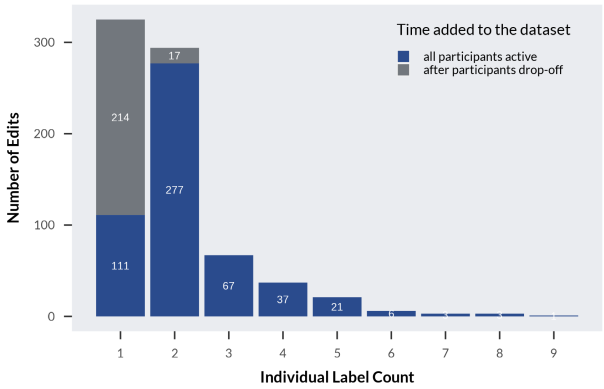

Figure 10: The count of edits with different numbers of individual labels received. Bar color denotes whether an edit was added to the dataset when all participants were active or after participant drop-off following exit interviews.

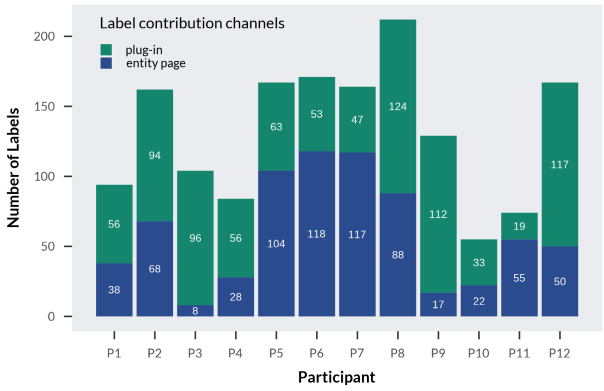

Figure 11: The number of labels contributed by each participant in the field study (bar heights), and the channels they used for contribution (colors). The upper and lower stacked bars denote plug-in and entity page, respectively.

As shown in Figure 9, the dataset curated in our field study captures a range of qualitatively distinct cases, represented by the four corners of the plot. The majority of edits fall in the lower-left corner, with *low disagreement and high confidence*. These can be interpreted as **clearer-cut cases**, where labelers tend to agree with high confidence. Edits toward the upper-right of this figure are ones with *high disagreement with low confidence*. These edits may be a consequence of **inherent ambiguity** regarding what label an edit should be assigned. By contrast, edits toward the lower-right of this figure are ones with *high disagreement with high confidence*. These edits are more likely to represent **genuine differences in perspective** among community members. For example, one of these edits[25] is a case that divides a wiki link. While some participants argued that the edit is damaging as it violates Wikipedia's style guidelines[26], others argued in the opposite direction, noting that the edit improved readability and that the style guideline is not strictly mandatory. In exit interviews, participants expressed that Wikibench was helpful in facilitating measured discussions even in cases where community members held strong opposing viewpoints: *"I was able to explain my rationale on the talk page. [...] I think that helped to make sure everyone's view was properly considered"* (P12). Finally, edits in the upper-left corner are ones with *low disagreement and low confidence*. These edits may represent **agreed-upon edge cases**: cases that are more ambiguous, but where community members nonetheless tend to agree. To better communicate about and capture cases like these through Wikibench, participants expressed desires for a means to explicitly indicate their *collective* confidence (or lack thereof) on a given edit: *"I think in some cases, we do want to tag it as low confidence because even after discussion we're not 100% sure"* (P9).

8.1.3 ***Collaborative Labeling: Most data points are labeled by multiple community members.*** To understand how effective Wikibench is in directing *multiple* labelers to edits, we examine the number of individual labels each edit received during our field study. As indicated by the blue bars in Figure 10, among the edits added to the dataset during the week when all participants were active, the majority (79%) received labels from at least two individuals. Even when considering the full set of edits, including those added to the dataset after participants began dropping off following exit interviews, over half (57%) received labels from at least two individuals.

All participants submitted labels that exceeded the minimum participation requirements, with no one strongly dominating label contributions. Participants were asked to submit at least 10 labels per day for up to 5 days. Figure 11 shows the number of labels each participant submitted. All participants surpassed the minimum requirement of 50 labels, with half of them contributing over triple this amount. Furthermore, as shown in Figure 11, participants

[25] https://enwp.org/Special:Diff/1163519177/1163595999
[26] https://enwp.org/WP:SEAOFBLUE

not only submitted labels using the plug-in during their regular patrolling activities but also contributed substantially via entity pages. Although we had asked participants to engage in at least three daily discussions on the campaign or entity talk pages, the observation that they also actively contributed many labels outside of their usual workflow suggests that they were highly engaged in the process, and may have been intrinsically motivated to contribute. This is corroborated by participants' comments in their exit interviews: *"I'm quite unsure about my [patrolling] decision sometimes, so it is good to have someone more experienced talk through why they concluded differently from me [...] It's very enlightening, this [study]"* (P11). In a longer-term deployment, these kinds of intrinsic motivations could play a pivotal role in promoting broader, sustained participation.

8.2 Potential for Use in AI Evaluation

While Wikibench's current interface primarily supports data curation, we are interested in understanding the potential of resulting datasets to support more informative AI evaluations downstream. In this section, we demonstrate how the dataset generated in our study can be used to compare two AI models for counter-vandalism deployed on Wikipedia (Section 8.2.1). Participants found these model comparisons informative and saw opportunities for the design of community-facing visualization and analysis tools to support such evaluations (Section 8.2.2).

8.2.1 ***Wikibench's dataset can help in understanding AI models' alignment with community perspectives.*** To showcase the potential of Wikibench datasets for use in AI evaluation, we used the dataset from the current study to evaluate the community alignment of two AI models deployed on Wikipedia for counter-vandalism. The first model is ORES, an AI system used extensively on Wikipedia for detecting damaging edits[27]. The second is the Revert-Risk model[28], which was recently developed by the Wikimedia Foundation to replace ORES. In contrast to ORES, which is trained on explicit labels of edit damage and user intent from Wikipedians, the new Revert-Risk model is trained solely on historical trace data. Specifically, the Revert-Risk model uses an edit's revert history on Wikipedia as ground truth for training and predicts the probability of an edit getting reverted. Although edits may be reverted on Wikipedia for a variety of reasons, beyond being damaging to articles [63], the Revert-Risk model's documentation[29] states *"the idea with Revert-Risk model is to use [edit] reverts as 'implicit annotations' [...] If we consider [ORES's] model as prediction for reverts, Revert-Risk is outperforming ORES in almost all scenarios."* Here, we present an alternative perspective made possible by a community-curated evaluation dataset.

We evaluated each model on Wikibench's dataset to examine their relative alignment with community perspectives. Given that the Revert-Risk model is meant to replace ORES in identifying damaging edits for counter-vandalism, we used the edit damage label in Wikbench's dataset to evaluate both models. Figure 12 shows the ROC curves of ORES and Revert-Risk models evaluated

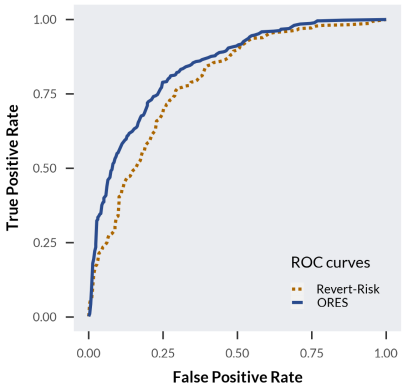

Figure 12: The ROC curves of ORES and Revert-Risk models evaluated on Wikibench's dataset. The AUC scores of ORES and Revert-Risk are 0.84 and 0.79, respectively.

on Wikibench's dataset. The AUC score of ORES and Revert-Risk is 0.84 and 0.79, respectively, showing that ORES performs better than Revert-Risk based on a dataset curated by community members. This result contrasts with the development team's evaluation on trace data, which had shown the opposite trend. This provides preliminary evidence that ORES's behavior aligns more closely with how our participants believe decisions *should* be made, compared with the Revert-Risk model.

Based on the evaluation results, we can further visualize the differences between the two AI models' predictions and Wikibench's primary labels, as shown in Figure 13. Each dot represents an edit in Wikibench's dataset encoded using a feature embedding[30], then projected onto a 2D space through t-SNE [91], and color-coded according to Wikibench's primary labels. The two lines, serving as decision boundaries for demonstrative purposes, are plotted using SVMs [15] based on the respective predictions of the two AI models. The prediction thresholds for ORES and Revert-Risk are 0.3810 and 0.6513, respectively, which were selected to maximize their prediction accuracy. Edits predicted by the model with a probability above the threshold are categorized as damaging. Figure 13 shows that the Revert-Risk model is more likely to incorrectly flag non-damaging edits. This result is likely because Revert-Risk is trained on trace data (an edit's revert history) rather than explicit labels, increasing the likelihood of incorrectly flagging types of edits that may have historically been reverted for reasons other than being damaging. For instance, in one case, both our participants and ORES considered an edit[31] non-damaging, whereas the Revert-Risk model predicted the opposite. Interestingly, the edit was eventually reverted, as predicted by Revert-Risk, but not for causing damage to the article. Instead, it resulted from an edit war, in which this edit was actually countering vandalism but was repeatedly reverted by the vandaliser. In this case, using Revert-Risk to identify vandalism might mistakenly flag edits that are actually combating vandalism, counter to its original purpose.

[27] https://www.mediawiki.org/wiki/ORES#Edit_quality
[28] https://meta.wikimedia.org/wiki/Machine_learning_models/Proposed/Language-agnostic_revert_risk
[29] https://wikitech.wikimedia.org/wiki/ORES
[30] https://www.mediawiki.org/wiki/ORES/Feature_injection
[31] https://enwp.org/Special:Diff/1163324435/1163324481

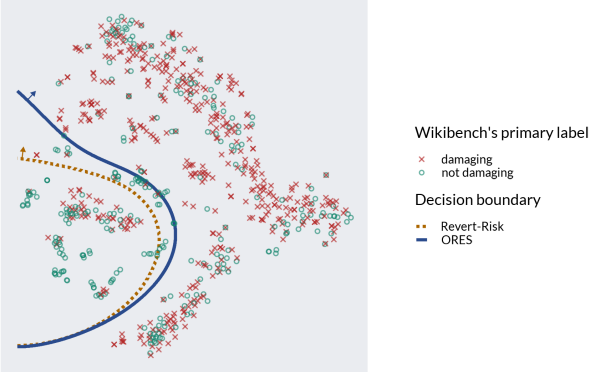

Figure 13: Edits from Wikibench's dataset projected onto a 2D space, along with decision boundaries visualized by training an SVM using the predictions from two AI models. The arrows indicate the direction in which SVMs predict damaging. As shown, the Revert-Risk model is more likely to incorrectly flag non-damaging edits.

Overall, this analysis showcases the potential of Wikibench's dataset for use in AI evaluation. It highlights the unique value of community-curated evaluation datasets, which may help identify misalignments between community perspectives and AI behaviors that would otherwise be overlooked in the development process.

8.2.2 ***Participants believe Wikibench's dataset can help the community understand gaps between their collective values and AI models' predictions.*** As mentioned in Section 6.1.1, during the exit interview we randomly selected edits from Wikibench's dataset and presented participants with both ORES's and Revert-Risk's predictions. Participants shared that *"it allows us to easily compare what Wikipedians believe vs what the AI believes. Using this data, we can find patterns of mistakes of AI models"* (P1). Through these comparisons, participants also gained a better understanding of which cases may be difficult for an AI model and why: *"Three people labeled it as damaging with low confidence, which might be why it wasn't picked up by the AI as well"* (P9). Participants recognized the dataset's potential to help their community evaluate among AI models: *"It would serve as a good benchmark for different tools, [... and] for whatever models that people come up with afterwards"* (P8). Finally, participants saw opportunities for community-facing visualization or analysis tools to help community members conduct more systematic comparisons: *"It will be really interesting if you compare these two samples, you can [visualize] nice charts [to show] how they look like"* (P7).

9 FINDINGS: HOW WIKIPEDIANS USE WIKIBENCH

In this section we present results from our analyses of interviews with participants to better understand their usage of Wikibench, their perceptions of how well our design requirements are fulfilled, and their visions for opportunities to improve Wikibench. We first discuss participant's perceptions of how Wikibench fits into Wikipedia's existing workflows and norms (Section 9.1). By cross-referencing participants' interview feedback and conversations on talk pages, we then highlight three cases where Wikibench's approach provided the community with the agency to drive the overall data curation process, beyond labeling individual edits (Section 9.2). We conclude this section by offering further insights into the ways participants use Wikibench for collaborative data curation and their suggestions for future improvements to the system (Section 9.3).

9.1 Data Curation within Wikipedians' Existing Workflows

> *"I would say it's almost like it was built-in [...] It was meant to be there"* (P10).

Compared to Wikilabels, participants find Wikibench much easier to use because it is embedded into their existing workflow: *"I think that Wikibench is currently the best way to achieve this because if you still use the old labeling options you lose a lot of time to even to login and try to work with the old interface just to label some edits"* (P7). Participants also shared that they felt comfortable editing labels and engaging in discussions because Wikibench followed Wikipedia's established norms: *"that creates a lot of comfort, because for me, I was navigating something that I knew very well. I know how discussions work on Wikipedia. I know what to expect. I know what's considered appropriate and what isn't. So I think it's really, really good"* (P6). For similar reasons, participants appreciated that in Wikibench's design, primary labels are established based on consensus rather than a majority vote: *"That is the core of decision making on Wikipedia. [...] I think it's smart that the primary label is not automatically [set as] the majority"* (P3).

9.2 Community Agency over the Data Curation Process

Wikibench was designed to empower communities to drive the data curation process, beyond labeling. This section showcases three examples where participants organically shaped the overall data curation process through Wikibench, by (1) refining the label definitions, (2) determining data inclusion criteria, and (3) authoring a data statement.

9.2.1 ***Participants revised label definitions to better capture emerging nuances and community norms.*** Unexpectedly, early in our field study, participants organically identified a need for better label definitions to guide subsequent data labeling and curation. The discussion started from an edit[32] that added more wiki links to an already overlinked article. P3 and P4 submitted opposite labels for edit damage, which attracted P2, P5, P6, and P8, from the campaign page. P6 noted: *"overlinking is damaging for legibility, especially in this section that's already 80% wikilinks."* Meanwhile, P5 noted: *"overlinking, yes, but that's not really damaging the article."* With these differing points of view, P2 initiated a discussion on the entity talk page where five participants responded. Even though participants, such as P3, replied on the talk page that they believed *"a healthy amount of differences of opinion (in the right places!) is the foundation of establishing positive consensus,"* they found it critical to have a clear label definition as a baseline to build upon. Given

[32] https://enwp.org/Special:Diff/719347634/719359416

Figure 14: An abridged screenshot of the campaign-level talk page where participants organically discussed and revised label definitions.

that general questions about label definitions went beyond the discussion about an individual edit, P5 initiated a discussion on the campaign talk page named *"What means 'damaging'/'not damaging' in general?"* (see Figure 14). Following a series of discussions, participants edited the campaign page to update the original label definition we provided:

> **Edit damage**: The edit damage label indicates whether this edit causes damage to the article or not.
> **User intent**: The user intent label indicates whether the edit was saved in good or bad faith.

into the new label definition that better captures the nuance and community norms:

> **Edit damage**: An edit is considered constructive when the post-edit revision is better than the pre-edit revision. For the purposes of evaluating edit damage, edits are not evaluated against what else could have been done to improve the article. "Not damaging" is a soft default; if an edit makes the article neither better nor worse at all, it is not damaging. Label data should still be provided in cases where edit damage depends on external factors. For example, an edit which introduces verifiably false information with appropriate style and formatting should be labelled as damaging.
> **User intent**: An edit is considered "good faith" when it is reasonably plausible that the editor's intention was to improve the article; per WP:AGF[33], good faith is the default until there is a concrete reason to suspect bad faith. Not all damaging edits are made in bad faith.

9.2.2 ***Participants defined inclusion criteria to ensure data was accessible by the full community.*** Midway through the study, participants found that some labeled edits were only visible to Wikipedia administrators. In response, they collectively decided to remove these edits from the dataset. It began when P4, P6, P9, and P10 each came across some entity pages where the edits were hidden from public view by Wikipedia administrators after being labeled. Even though these hidden edits, also known as revdel on Wikipedia[34], were still visible to some participants with administrator rights, P9 initiated a discussion on an entity talk page. P10 noticed this wasn't a one-time incident and raised the issue on the campaign talk page by creating a discussion topic *"How to handle diffs that have since been revdel'd?"* While some participants

[33] https://enwp.org/WP:AGF
[34] https://enwp.org/WP:REVDEL

wondered if *"labeling them as damaging or bad faith in Wikibench"* (P8) would make sense, others argued *"it might be better to have the dataset used for training and evaluating AI be entirely transparent"* (P6). Following discussions, participants collectively decided to exclude these edits from the dataset. P4, a participant with the administrator rights, volunteered to make an update so that Wikibench would no longer include these entity pages in the table on the campaign page given their updated prefix. Participants also added a new section on the campaign page named *"Entity pages on revdel'ed edits,"* where they compiled a list of archived entity pages and provided instructions for people to report here when encountering other such edits.

9.2.3 ***Participants authored a data statement to specify appropriate usage of the evaluation dataset***. During the field study, participants recognized that the resulting dataset likely would not align with the natural distribution of edits on Wikipedia. Given the observation, some participants raised the question on the campaign talk page: *"I wonder how useful this dataset will be for training and evaluating AI given that it does not accurately represent the totality of edits on Wikipedia. [...] For example, damaging and bad-faith edits are significantly overrepresented"* (P6). After we explained the potential use participants added the following data statement to clarify the usage of the evaluation dataset, preventing misuse by those unfamiliar with Wikipedia's context.

> **Data statement**: There is a local consensus that on-wiki use of the data acquired through this campaign should be limited to the immediate scope of Wikibench. A strong consensus should be established prior to any on-wiki use outside this research project. The data acquired as part of this campaign is not intended for other uses and may be inappropriate or unsuitable for many purposes. In particular, it is not a representative sample of edits made to the English Wikipedia, nor is it intended as such.

9.3 Collaborative Data Labeling

Participants appreciated Wikibench's collaborative approach to data labeling because it allowed contributors with complementary perspectives to build consensus and a stronger community. Participants found that Wikibench's campaign and entity pages facilitate collaborative data labeling by quickly pinpointing edits where more labels or discussions could be valuable. They also perceived Wikibench as effective in surfacing disagreements and facilitating consensus-building. In each area, participants also envisioned various opportunities to make Wikibench more effective.

9.3.1 ***Participants prefer Wikibench's collaborative approach to data labeling***. Participants found Wikibench enabled contributors with diverse and complementary expertise to discuss and reach a consensus on the final labels: *"It's very useful to have editors with different areas of expertise within Wikipedia working together. [...] In the end, this supports getting better data"* (P6). This is particularly relevant for P6, while reviewing an edit[35] about a coffee produced in Southeast Asia, initially found the primary label confusing: *"Adding this image seems fine to me. Damaging? Why?"* After reading P4's note, a participant from Singapore, P6 agreed with the label: *"Copyright? Oh, okay, this would be very difficult to investigate properly. It is a copyright violation because it has a company logo."* Without P4's local knowledge, P6 would have missed the copyright violation and labeled the edit as not damaging. Participants find these notes and discussions facilitated by Wikibench helpful because it allows them to learn new things from others: *"This allows me to discover things I might not have thought of"* (P1). They also find these discussions help the community collectively reflect on their patrolling standards: *"It allows the community to establish better consensus as to what typically should be reverted and what might require more care"* (P3). In turn, they believe these interactions could help build a stronger community: *"I haven't really talked to many of the other patrollers before this project. For people who are interested in a tighter community, Wikibench would be absolutely a great option"* (P11).

9.3.2 ***Participants find Wikibench helpful in quickly identifying edits where more labels or discussion could be valuable.*** For example, some participants found the campaign page helpful in identifying edits with fewer labels to ensure that edits are collaboratively labeled: *"I try to work on those which were alone because I previously complained about that the previous system allowed only one person to evaluate."* (P7). Some participants were particularly interested in checking edits with high disagreements and helping build consensus: *"I'm very interested in seeing where people differ. [...] I wanted to build consensus, or at least to try"* (P11). Given the desire to more effectively locate edits for contribution, participants suggested: *"Maybe you can add a label 'discussion: no/yes' to the big table so you can see where is a discussion and join there. The disagreement is a good way to find discussions, but isn't perfect"* (P5).

Once participants located an edit and opened its entity page, participants found the entity page provided an at-a-glance overview that helped quickly understand the current level of disagreement: *"You can see the colors and see one damaging, one not damaging low confidence, [...] The preview was really good to see"* (P5). Another participant echoed: *"Wikibench's feature of showing exactly where different editors align, and that one user cannot force the primary label, helped to facilitate discussions"* (P12). When deeper discussion was needed, participants found the talk page helpful: *"If I still have a question, I can open the talk page"* (P5). However, a participant also expresses concern that the transparent labeler information might affect people's judgment in undesirable ways, in some cases: *"If it wasn't transparent, you wouldn't know who did the labeling. [...] But on the other hand, if the quotation marks 'big guys' go on that way, then maybe the small guys will follow"* (P7).

9.3.3 ***Participants find Wikibench effective in surfacing disagreements and facilitating consensus-building.*** Participants perceived that discussions on Wikibench were typically sparked by disagreements, but concluded with a shared consensus: *"Most of the time, one of me or the other one says: Okay, I think you have the better argument, and I'm switching to your position, or it's okay for me"* (P5). Even when people hold strong opposing opinions, participants found Wikibench helpful in facilitating more productive discussions and avoiding emotionally charged confrontations: *"Wikibench provides an avenue for people to calmly discuss stuff because it's nothing personal"* (P4). Participants also shared that they adopted different strategies for consensus building, with some followed Wikibench's

[35] https://enwp.org/Special:Diff/1128051255/1163570328

nudge to boldly edit primary labels: *"I changed the primary label before I actually brought up the point, just mostly so I can actually get the person's attention"* (P11). Meanwhile, some participants prefer to initiate a discussion and wait for consensus to form first: *"'I might change it if I think the other editor has made a mistake, rather than made a decision that I disagree with. If it's a disagreement, we let consensus form first'* (P8). Overall, participants appreciated the design where the primary label was not the majority vote but could be edited by anyone: *"I like being able to edit the primary label to help better reflect community consensus. I find the warning helpful as it reminds users that even when being bold, your changes should reflect community consensus, not your personal opinions"* (P1).

9.3.4 ***Participants see value in the openness of Wikibench's datasets.*** Similar to Wikipedia articles, Wikibench's dataset is not owned by specific individuals but is open to the public. Participants appreciated the openness of Wikibench's dataset, and emphasized that data transparency is essential for the evaluation results to be trustworthy: *"These tools do so much. They're very highly trusted. [...] If we're evaluating that kind of tool, there needs to be an additional level of trust for the dataset we're using to do that"* (P6).

10 DISCUSSION

As AI tools are increasingly developed for community contexts, it is critical to ensure that they are aligned with community needs and values. In this paper, we introduce Wikibench, a system that enables community members to collectively curate evaluation datasets for AI tools that will be used in their communities. We conducted a field study on Wikipedia to understand how a real-world community might use Wikibench in practice. Overall, we found that Wikipedians' use of Wikibench yielded labels that are reflective of consensus among a broader range of community members, while also capturing ambiguities and differences in perspective among community members (Section 8.1). We find promising evidence of the utility of Wikibench's community-curated datasets in understanding areas of alignment and misalignment with community perspectives (Section 8.2). Finally, we present examples of how community members use Wikibench to shape the overall data curation process, and discuss their experiences using the system for data labeling and curation (Section 9). In this section, we highlight key takeaways and propose future directions for HCI systems that support community-driven data curation and AI evaluation.

10.1 Supporting Community-Driven Data Curation beyond Wikipedia

As the largest, most successful platform for collective knowledge-building and curation, we believe there is much to learn from Wikipedia for the design of effective community-driven data curation processes. Given that Wikibench was designed around Wikipedia's processes and norms, we anticipate that several aspects of Wikibench's design may be useful for community-driven data curation in other contexts. In particular, we expect that the four design requirements discussed in Section 4 are broadly relevant for the design of tools intended to support community-driven data curation. We also expect that the overall workflow embodied by Wikibench (Figure 2) will be generalizable to other community platforms, particularly those operating within the middle level of a multi-level governance structure [56], such as subreddits, Facebook Groups, Mastodon Nodes, other Wikipedia language editions, and more.

At the same time, we expect that the specific implementation of Wikibench's plug-in, entity page, and campaign page will require careful adaptation to align with the established norms of other communities. For example, imagine a subreddit community interested in adapting Wikibench to curate posts as data points and labels indicating whether a given post ought to be flagged for removal. In this context, integrating the data curation process into existing work practices (D3), would likely mean tailoring the specific implementation of a plug-in to fit the workflow of the "Knights of New"[36] on Reddit [41, 70]. Similarly, to encourage deliberation (D2), the entity and campaign page would likely need to be adapted to better align with existing collaboration and consensus-building processes on Reddit [14]. The system may also need to incorporate the community's established norms (D1) for safeguarding against bad actors. On Wikipedia, this may mean restricting access to Wikibench to registered users, whereas on Reddit this may mean considering a Reddit user's karma scores[37]. Future research should investigate how systems for community-driven data curation can be designed for use by other communities outside of Wikipedia, and what mechanisms in Wikibench's current design are most readily transferable across community contexts.

10.2 Balancing Costs and Benefits in Community-Driven Data Curation

Prior approaches to account for annotator disagreements tend to handle disagreements post-hoc, after individual labels have been gathered [4, 20, 26, 42, 93], instead of facilitating discussion and deliberation among annotators. By contrast, Wikibench provides community members the agency to navigate and resolve disagreements, leading to various benefits that would not be achievable by algorithmic methods alone. For example, discussions among community members can help resolve ambiguities in labeling, facilitate consensus building, and promote collective reflection on community standards, as discussed in Section 9.3. However, given that community members will generally have limited time and attention to contribute to the curation of AI datasets, further research is needed to find the right balance between *community agency* over curation processes, on the one hand, and time and labor *efficiency* on the other. The current version of Wikibench aims to make more effective use of community members' time by embedding the plug-in within their everyday workflows (D3) instead of asking them to use Wikilabels just for labeling. Wikibench also automatically surfaces edits that may benefit most from additional attention on the campaign page. Future research could explore ways to better optimize the use of community members' time, such as algorithmically prioritizing data points that are predicted as more likely to prompt disagreements for community discussions [6]. Relatedly, future research could investigate when the benefits of further community engagement in data curation becomes marginal, as the number of participants and contributions increase.

[36] A group of volunteers that review new posts instead of already popular posts.
[37] A score representing the positive social signals that a user's activity (e.g., posts and comments) has received.

10.3 Advancing Pluralistic Approaches to AI Evaluation

Our demonstration of Wikibench's use to compare different AI models' community alignment employed the primary labels collected through Wikibench in our field study. However, a major strength of Wikibench's community-curated datasets is their ability to capture additional signals such as ambiguity in labeling and differences in perspective among community members. Recent research has suggested the benefits of evaluating AI models using datasets that reflect diverse perspectives, with multiple labelers per data point (e.g., [16, 42–44, 60]). Where disagreements arise, these may represent ambiguity inherent to an edit or noise in labeling, or they may signal genuine differences in perspective among subgroups of a community whose voices deserve consideration. Wikibench datasets record signals to help AI evaluators tease apart these possibilities, which can support more nuanced and pluralistic analyses of AI models' community alignment. The development of evaluation methods and workflows to support more pluralistic approaches to AI evaluation is an emerging area of research [12, 89]. Future research could systematically compare Wikibench's process with alternative approaches to handling differences in perspective for pluralistic AI evaluation. Such comparisons could advance our understanding of trade-offs between different community-driven and algorithmic approaches to navigating disagreements in labeling. In turn, this may inform the development of new approaches that integrate complementary strengths of existing methods. It is our hope that systems for community-driven data curation can help to accelerate progress in this area through the development of relevant datasets [71, 78].

10.4 Designing Community-Facing Evaluation Interfaces

While Wikibench's current interface primarily supports community-driven *data curation*, future research should explore the design of community-facing interfaces that empower communities to effectively leverage the resulting datasets to inform decision-making about AI design and adoption. We envision that, in more complex evaluation scenarios that require caution in interpretation, community-driven AI evaluations may sometimes be facilitated through partnerships with technical experts [19]. Beyond supporting AI evaluation, our participants found value in Wikibench's collaborative data curation process during our field study because it helped them to reflect, both individually and collectively, on their edit patrolling standards. Thus, an additional promising direction for future research, in the Wikipedia context and beyond, is to explore how community-driven data curation processes can be more explicitly designed to support such reflection. This may include the design of interfaces that help community members leverage community-curated datasets to reflect upon *their own* decision-making—both individually and as a community—and identify potential areas for improvement (cf. [96]).

10.5 Supporting Communities in Steering Overall Dataset Composition

The current version of Wikibench was designed to deeply embed into community members' regular patrolling activities on Wikipedia. This "lowers the floor" of effort required to contribute to Wikibench datasets. However, a consequence of this design is that the data points included in Wikibench datasets may tend to reflect the distribution of edits that Wikipedians encounter while patrolling, which are not necessarily representative of *all* edits made to English Wikipedia. Indeed, as the participant-authored data statement from our field study acknowledges, the dataset curated through this study was not intended to be representative in this sense. This makes the dataset more useful for some evaluation purposes than others. Future research should explore how to design mechanisms for community-driven curation that can assist communities in steering datasets, depending on their specific goals, toward desired distributional properties (e.g., specific notions of representativeness, or oversampling along particular dimensions of interest). It is possible that to some extent, this may require community members to spend more time contributing outside of their regular workflows. However, it may be possible to design new mechanisms that minimize the disruption required. For example, in the context of a community member's regular patrolling activities on Wikipedia, a hypothetical future browser plug-in might be designed to occasionally present edits that they *would not have otherwise encountered* for labeling purposes, intended to help the community achieve their overall distributional goals for the dataset. We envision that, in many community contexts, decision-making regarding the distributional properties a campaign should aim for may benefit from accessibly-designed "explainers" that summarize relevant consideration (cf. [64]), and/or through partnerships with relevant technical experts [19].

10.6 Drawing Inspiration from Existing Content Curation Mechanisms

Several of Wikibench's *AI data* curation mechanisms are inspired by existing *content* curation mechanisms from Wikipedia and other online platforms. For example, Stack Overflow users can vote on and discuss individual posts [69] or have higher-level discussions about community norms on Meta Stack Overflow [29]. Similarly, Wikipedians can use Wikibench to label and discuss individual data points on entity pages or initiate higher-level conversations about the overall data curation process on the campaign page. We see opportunities for future versions of Wikibench, or other community-driven data curation platforms, to draw further inspiration from the design of existing content curation mechanisms. For example, a feature inspired by Reddit's post flairs (short tags attached to each post) [54], could be used to categorize data points and facilitate dataset navigation (e.g., via data slices [17]). In addition, future work in this space can take inspiration from HCI research related to content curation. For example, recent social media research has proposed curating content not solely based on engagement signals such as votes but on a community's shared visions [49] and values [57]. In the context of data curation, these ideas may inform new approaches to the community-driven prioritization of data points for AI evaluation.

11 CONCLUSION

In this work, we have demonstrated the potential for new approaches to community-driven curation of AI evaluation datasets, through the introduction of the Wikibench system and a field study investigating its use. Our findings demonstrate that community-driven curation on Wikibench can produce datasets that capture community consensus, disagreement, and uncertainty, while enabling community members to shape the overall data curation process. Building on this work, future research should explore the design of tools and processes that can support community-driven data curation across a broader range of contexts, and that can expand community agency in both the curation of datasets and their use in evaluation.

ACKNOWLEDGMENTS

The funding for this research was provided by UL Research Institutes through the Center for Advancing Safety of Machine Intelligence, CMU's Block Center for Technology and Society, and the National Science Foundation (NSF) under Award No. 1952085 and 2001851. We thank Wikipedians for their time and input that shaped this research, especially 1TWO3Writer, Actualcpscm, Alpha3031, Bencemac, Bluerasberry, Chtnnh, Ciell, Fehufanga, FenrisAureus, Illusion Flame, Loafiewa, Matthewrb, PriusGod, Robertsky, RonnieV, Schminnte, Skarmory, TenWhile6, Vermont, Zache, Zppix, and many others (including those who chose to remain anonymous). We also appreciate Jane Hsieh, Kimi Wenzel, Luke Guerdan, Ningjing Tang, Pranav Khadpe, Seyun Kim, Tiffany Chih, and Wesley Deng for their feedback on the paper draft. Finally, we thank Isadora Krsek for designing the Wikibench logo, as shown in Figure 1.

REFERENCES

[1] Lora Mois Aroyo and Praveen Kumar Paritosh. 2021. Adversarial Test Set for Image Classification: Lessons Learned from CATS4ML Data Challenge. (2021).
[2] Joshua Attenberg, Panos Ipeirotis, and Foster Provost. 2015. Beat the machine: Challenging humans to find a predictive model's "unknown unknowns". *Journal of Data and Information Quality (JDIQ)* 6, 1 (2015), 1–17.
[3] Yuanchen Bai, Raoyi Huang, Vijay Viswanathan, Tzu-Sheng Kuo, and Tongshuang Wu. 2023. Measuring Adversarial Datasets. *arXiv preprint arXiv:2311.03566* (2023).
[4] Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems* 35 (2022), 38176–38189.
[5] Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics* 8 (2020), 662–678.
[6] Connor Baumler, Anna Sotnikova, and Hal Daumé III. 2023. Which examples should be multiply annotated? active learning when annotators may disagree. In *Findings of the Association for Computational Linguistics: ACL 2023*. 10352–10371.
[7] Ivan Beschastnikh, Travis Kriplean, and David McDonald. 2008. Wikipedian self-governance in action: Motivating the policy lens. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 2. 27–35.
[8] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*. 491–500.
[9] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.
[10] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.
[11] Susan L Bryant, Andrea Forte, and Amy Bruckman. 2005. Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. In *Proceedings of the 2005 ACM International Conference on Supporting Group Work*. 1–10.
[12] Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 6860–6868.
[13] Ángel Alexander Cabrera, Abraham J Druck, Jason I Hong, and Adam Perer. 2021. Discovering and validating ai errors with crowdsourced failure reports. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–22.
[14] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–25.
[15] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2, 3 (2011), 1–27.
[16] Quan Ze Chen and Amy X Zhang. 2023. Judgment Sieve: Reducing Uncertainty in Group Judgments through Interventions Targeting Ambiguity versus Disagreement. *arXiv preprint arXiv:2305.01615* (2023).
[17] Vincent Chen, Sen Wu, Alexander J Ratner, Jen Weng, and Christopher Ré. 2019. Slice-based learning: A programming model for residual learning in critical data slices. *Advances in neural information processing systems* 32 (2019).
[18] cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. 2017. Toxic Comment Classification Challenge. https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge
[19] Sasha Costanza-Chock. 2020. *Design justice: Community-led practices to build the worlds we need.* The MIT Press.
[20] Naihao Deng, Siyang Liu, Xinliang Frederick Zhang, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You Are What You Annotate: Towards Better Models through Annotator Representations. *arXiv preprint arXiv:2305.14663* (2023).
[21] Wesley Hanwen Deng, Boyuan Guo, Alicia Devrio, Hong Shen, Motahhare Eslami, and Kenneth Holstein. 2023. Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.
[22] Emily Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. 2021. Whose ground truth? accounting for individual and collective identities underlying dataset annotation. *arXiv preprint arXiv:2112.04554* (2021).
[23] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. 2020. Bringing the people back in: Contesting benchmark machine learning datasets. *arXiv preprint arXiv:2007.07399* (2020).
[24] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.
[25] Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. *arXiv preprint arXiv:1908.06083* (2019).
[26] Senjuti Dutta, Sid Mittal, Sherol Chen, Deepak Ramachandran, Ravi Rajakumar, Ian Kivlichan, Sunny Mak, Alena Butryna, and Praveen Paritosh. 2023. Modeling subjectivity (by Mimicking Annotator Annotation) in toxic comment identification across diverse communities. *arXiv preprint arXiv:2311.00203* (2023).
[27] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. "Be careful; things can be worse than they appear": Understanding Biased Algorithms and Users' Behavior around Them in Rating Platforms. In *Proceedings of the international AAAI conference on web and social media*, Vol. 11. 62–71.
[28] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User attitudes towards algorithmic opacity and transparency in online reviewing platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
[29] Jingchao Fang, Jia-Wei Liang, and Hao-Chuan Wang. 2023. How People Initiate and Respond to Discussions Around Online Community Norms: A Preliminary Analysis on Meta Stack Overflow Discussions. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. 221–225.
[30] Michael Feffer, Hoda Heidari, and Zachary C Lipton. 2023. Moral Machine or Tyranny of the Majority? *arXiv preprint arXiv:2305.17319* (2023).
[31] Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. 2018. Reddit rules! characterizing an ecosystem of governance. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.
[32] Andrew Flinn. 2007. Community histories, community archives: Some opportunities and challenges. *Journal of the Society of Archivists* 28, 2 (2007), 151–176.
[33] Denae Ford, Kristina Lustig, Jeremy Banks, and Chris Parnin. 2018. " We Don't Do That Here" How Collaborative Editing with Mentors Improves Engagement in Social Q&A Communities. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.
[34] Andrea Forte, Vanesa Larco, and Amy Bruckman. 2009. Decentralization in Wikipedia governance. *Journal of Management Information Systems* 26, 1 (2009),

49–72.

[35] Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. 2020. Evaluating models' local decision boundaries via contrast sets. *arXiv preprint arXiv:2004.02709* (2020).

[36] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[37] R Stuart Geiger and Aaron Halfaker. 2013. When the levee breaks: without bots, what happens to Wikipedia's quality control processes?. In *Proceedings of the 9th International Symposium on Open Collaboration*. 1–6.

[38] R Stuart Geiger and Aaron Halfaker. 2017. Operationalizing conflict and cooperation between automated software agents in wikipedia: A replication and expansion of'even good bots fight'. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–33.

[39] R Stuart Geiger and David Ribes. 2010. The work of sustaining order in Wikipedia: The banning of a vandal. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. 117–126.

[40] R Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 325–336.

[41] Eric Gilbert. 2013. Widespread underprovision on reddit. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 803–808.

[42] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[43] Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[44] Nitesh Goyal, Ian D Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is your toxicity my toxicity? Exploring the impact of rater identity on toxicity annotation. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–28.

[45] Luke Guerdan, Amanda Coston, Zhiwei Steven Wu, and Kenneth Holstein. 2023. Ground (less) Truth: A Causal Framework for Proxy Labels in Human-Algorithm Decision-Making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 688–704.

[46] Aaron Halfaker and R Stuart Geiger. 2020. Ores: Lowering barriers with participatory machine learning in wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–37.

[47] Aaron Halfaker, R Stuart Geiger, Jonathan T Morgan, and John Riedl. 2013. The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist* 57, 5 (2013), 664–688.

[48] Aaron Halfaker, Aniket Kittur, and John Riedl. 2011. Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work. In *Proceedings of the 7th international symposium on wikis and open collaboration*. 163–172.

[49] Wanrong He, Mitchell L Gordon, Lindsay Popowski, and Michael S Bernstein. 2023. Cura: Curation at Social Media Scale. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–33.

[50] Dorothy Howard and Lilly Irani. 2019. Ways of knowing when research subjects care. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–16.

[51] Sohyeon Hwang and Aaron Shaw. 2022. Rules and Rule-Making in the Five Largest Wikipedias. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 347–357.

[52] Lilly C Irani and M Six Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 611–620.

[53] Abhinav Jain, Hima Patel, Lokesh Nagalapatti, Nitin Gupta, Sameep Mehta, Shanmukha Guttula, Shashank Mujumdar, Shazia Afzal, Ruhi Sharma Mittal, and Vitobha Munigala. 2020. Overview and importance of data quality for machine learning tasks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 3561–3562.

[54] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. " Did you suspect the post would be removed?" Understanding user reactions to content removals on Reddit. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–33.

[55] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 5 (2019), 1–35.

[56] Shagun Jhaver, Seth Frey, and Amy X Zhang. 2023. Decentralizing Platform Power: A Design Space of Multi-level Governance in Online Social Platforms. *Social Media+ Society* 9, 4 (2023), 20563051231207857.

[57] Chenyan Jia, Michelle S Lam, Minh Chau Mai, Jeff Hancock, and Michael S Bernstein. 2023. Embedding democratic values into social media AIs via societal objective functions. *arXiv preprint arXiv:2307.13912* (2023).

[58] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 306–316.

[59] Gary D Kader and Mike Perry. 2007. Variability for categorical variables. *Journal of statistics education* 15, 2 (2007).

[60] Shivani Kapania, Alex S Taylor, and Ding Wang. 2023. A hunt for the Snark: Annotator Diversity in Data Practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[61] Anna Kawakami, Venkatesh Sivaraman, Logan Stapleton, Hao-Fei Cheng, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. "Why Do I Care What's Similar?" Probing Challenges in AI-Assisted Child Welfare Decision-Making through Worker-AI Interface Design Concepts. In *Designing Interactive Systems Conference*. 454–470.

[62] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking Benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 4110–4124. https://doi.org/10.18653/v1/2021.naacl-main.324

[63] Aniket Kittur, Bongwon Suh, Bryan A Pendleton, and Ed H Chi. 2007. He says, she says: conflict and coordination in Wikipedia. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 453–462.

[64] Tzu-Sheng Kuo, Hong Shen, Jisoo Geum, Nev Jones, Jason I Hong, Haiyi Zhu, and Kenneth Holstein. 2023. Understanding Frontline Workers' and Unhoused Individuals' Perspectives on AI Used in Homeless Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[65] Michelle S Lam, Mitchell L Gordon, Danaë Metaxa, Jeffrey T Hancock, James A Landay, and Michael S Bernstein. 2022. End-User Audits: A System Empowering Communities to Lead Large-Scale Investigations of Harmful Algorithmic Behavior. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–34.

[66] 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. *arXiv preprint arXiv:2308.15768* (2023).

[67] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

[68] Gerald S Leventhal. 1980. What should be done with equity theory? New approaches to the study of fairness in social relationships. In *Social exchange: Advances in theory and research*. Springer, 27–55.

[69] Lena Mamykina, Bella Manoim, Manas Mittal, George Hripcsak, and Björn Hartmann. 2011. Design lessons from the fastest q&a site in the west. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2857–2866.

[70] Adrienne Lynne Massanari. 2015. Participatory culture, community, and play. (2015).

[71] Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Alicia Parrish, Hannah Rose Kirk, et al. 2022. Dataperf: Benchmarks for data-centric ai development. *arXiv preprint arXiv:2207.10062* (2022).

[72] Danaë Metaxa, Joon Sung Park, Ronald E Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, Christian Sandvig, et al. 2021. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction* 14, 4 (2021), 272–344.

[73] 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. In *Conference on Human Factors in Computing Systems-Proceedings*. 86–94.

[74] Michael Muller, Christine T Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, et al. 2021. Designing ground truth and the social life of labels. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–16.

[75] Claudia Müller-Birn, Leonhard Dobusch, and James D Herbsleb. 2013. Work-to-rule: the emergence of algorithmic governance in Wikipedia. In *Proceedings of the 6th International Conference on Communities and Technologies*. 80–89.

[76] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial NLI: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599* (2019).

[77] Safiya Umoja Noble. 2018. Algorithms of oppression. In *Algorithms of oppression*. New York university press.

[78] Luis Oala, Manil Maskey, Lilith Bat-Leah, Alicia Parrish, Nezihe Merve Gürel, Tzu-Sheng Kuo, Yang Liu, Rotem Dror, Danilo Brajovic, Xiaozhe Yao, et al. 2023.

DMLR: Data-centric Machine Learning Research–Past, Present and Future. *arXiv preprint arXiv:2311.13028* (2023).
[79] Orestis Papakyriakopoulos, Severin Engelmann, and Amy Winecoff. 2023. Upvotes? Downvotes? No Votes? Understanding the relationship between reaction mechanisms and political discourse on Reddit. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–28.
[80] Kenny Peng, Arunesh Mathur, and Arvind Narayanan. 2021. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. *arXiv preprint arXiv:2108.02922* (2021).
[81] Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366* (2021).
[82] 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 *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.
[83] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 1668–1678.
[84] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997* (2021).
[85] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–29.
[86] Dilruba Showkat, Angela DR Smith, Wang Lingqing, and Alexandra To. 2023. "Who is the right homeless client?": Values in Algorithmic Homelessness Service Provision and Machine Learning Research. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.
[87] Divya Siddarth, Daron Acemoglu, Danielle Allen, Kate Crawford, James Evans, Michael Jordan, and E Weyl. 2021. How AI fails us. *arXiv preprint arXiv:2201.04200* (2021).
[88] C Estelle Smith, Bowen Yu, Anjali Srivastava, Aaron Halfaker, Loren Terveen, and Haiyi Zhu. 2020. Keeping community in the loop: Understanding wikipedia stakeholder values for machine learning-based systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
[89] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A Roadmap to Pluralistic Alignment. *arXiv preprint arXiv:2402.05070* (2024).
[90] John W Thibaut and Laurens Walker. 1975. Procedural justice: A psychological analysis. *(No Title)* (1975).
[91] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
[92] Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics* 7 (2019), 387–401.
[93] Shaun Wallace, Tianyuan Cai, Brendan Le, and Luis A Leiva. 2022. Debiased label aggregation for subjective crowdsourcing tasks. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–8.
[94] Zining Ye, Xinran Yuan, Shaurya Gaur, Aaron Halfaker, Jodi Forlizzi, and Haiyi Zhu. 2021. Wikipedia ORES explorer: Visualizing trade-offs for designing applications with machine learning API. In *Designing Interactive Systems Conference 2021*. 1554–1565.
[95] Kyra Yee, Uthaipon Tantipongpipat, and Shubhanshu Mishra. 2021. Image cropping on twitter: Fairness metrics, their limitations, and the importance of representation, design, and agency. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–24.
[96] Angie Zhang, Olympia Walker, Kaci Nguyen, Jiajun Dai, Anqing Chen, and Min Kyung Lee. 2023. Deliberating with AI: Improving Decision-Making for the Future through Participatory AI Design and Stakeholder Deliberation. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–32.
[97] Amy X Zhang and Justin Cranshaw. 2018. Making sense of group chat through collaborative tagging and summarization. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–27.

A STUDY DETAILS

A.1 Formative Study

As described in Section 4, we conducted semi-structured interviews with Wikipedians who self-identified themselves with one or more of the roles listed in Table 1. Here, we provide our interview questions for reference:

- Please describe your experience with AI tools, such as ORES, for counter-vandalism, new page review, or other tasks on Wikipedia.
- Please share your experience as a [participant's role] within your community on Wikipedia.
- How did your community make decisions about the design and use of AI tools?
- How were you involved in this decision-making process?
- How did your community evaluate whether these AI tools fit the community's needs and values before or after the deployment?
- How were you involved in the evaluation process?
- How did community members collaborate and resolve disagreements during the evaluation process?
- Was the evaluation process effective or not? Why?
- Are there forms of support that would be particularly helpful to have from your perspective as a [participant's role]?
- Are there forms of support that would be particularly helpful to have for the entire community?
- Did the evaluation change the community's perception and acceptance of AI tools? How?
- Have you ever participated in data labeling campaigns on Wikipedia? If so, would you please describe your experience?
- Can you envision ways to better support communities in evaluating AI tools before they are deployed to make more informed decisions about whether or not the community should adopt them?

A.2 Field Study

A.2.1 Exit interview questions. As described in Section 6.1.1, we conducted an exit interview with each participant once they completed the field study to learn about their experiences and gather feedback. Here, we provide our interview questions for reference:

- What is your overall experience using Wikibench?
- What is your best and worst experience?
- Does this process provide the community with agency over the curation of evaluation datasets?
- How well does this process align with Wikipedian's norms for editing and discussion?
- How well does Wikibench support Wikipedians in discussing and resolving disagreements?
- Do you feel the primary labels are the result of community consensus?
- How do you think Wikibench could be better designed to support consensus building?
- How well does Wikibench fit into Wikipedia's interface and workflow?
- How do you think Wikibench could fit better?
- Were you able to get a good overview of the data curation progress using Wikibench?
- Do you feel Wikibench shows all the data and edits with transparency? Is it good or bad?
- Is there anything you would like to be able to keep track of?
- How do you think we can improve Wikibench?

In addition to collecting feedback, we presented participants with a randomly selected set of edits from Wikibench's datasets,

including their labels and the predictions from the ORES and RevertRisk models (mentioned in Section 3.1 and 8.2). Specifically, we presented these edits in a table, with each row featuring an edit's ID, a link to its entity page, its primary label in Wikibench's dataset, and the models' predicted probabilities of the edit being damaging or reverted. The table displayed a random set of ten edits at a time and resampled each time as participants reloaded their browsers.

A.2.2 Exit interview thematic analysis. In Table 3, we provide a summary of the seven highest-level themes we identified through a reflexive thematic analysis of the exit interviews. We also list the sections in which each is discussed in the paper. Due to the limitation of word counts, we do not include the 17 second-level themes, 64 first-level themes, and 249 codes in the table.

A.3 Validation Study

A.3.1 Edit selection procedure. We selected the edits for the validation study based on a few considerations. First, these edits should have labels from Wikilabels and Wikibench for comparison. We achieved this by sampling existing, labeled edits from Wikilabel's dataset and having the field study participants provide initial labels for these edits using Wikibench during onboarding sessions, so that they would be added to Wikibench's dataset. To ensure participants did not focus on these edits more than they would otherwise, they were not told that these edits would be used in a validation study. Secondly, due to the limited onboarding time, we asked each participant to label only a small number of edits, which in turn limited the total number of edits we could sample from Wikilabels' dataset in the first place. Finally, considering the total limit and our goal of assessing Wikibench's label quality resulting from participants' navigation of consensus, disagreement, and uncertainty, we oversampled edits that were likely to spark discussion while also including straightforward ones that were more likely to receive unanimous labels, as described below. Given these three considerations, we sampled 90 edits from Wikilabels' datasets and had each of the 12 field study participants label a random subset of 15 edits. This design ensured that each edit was labeled by at least two participants, the minimum number needed to kickstart discussion.

In order to sample 90 edits from Wikilabels' dataset[38] that were likely to spark discussion, we first identified 4,407 edits with both edit damage and user intent labels available in the dataset when we conducted the study. We excluded 127 edits that were hidden from public view by Wikipedia administrators and 18 edits that were labeled by more than one person due to Wikilabels' system race conditions. Among the remaining 4,262 edits, we categorized edits into three categories: (1) potentially ambiguous edits, (2) contested edits, and (3) other edits (cf. [16, 43]). We considered *potentially ambiguous edits* as those with the "unsure" mark specified in Wikilabels' dataset. We identified *contested edits* as edits that had received higher-confidence labels (i.e., without the "unsure" mark), which were different from their actual reversion outcomes on Wikipedia (e.g., edits labeled as damaging in Wikilabels but didn't get reverted on Wikipedia). Given that these were cases where two Wikipedians had historically disagreed, we expected that these edits were ones for which Wikipedians are more likely differ in their perspectives. This categorization led to 365 potentially ambiguous edits, 660 contested edits, and 3,237 other edits. We sampled 30 edits from each category, resulting in 90 edits in total.

A.3.2 Participant demographics. The demographic information of the five additional Wikipedians we recruited for the validation study is shown in Table 4.

[38] https://labels.wmflabs.org/stats/enwiki/41

Table 3: The seven highest-level themes we identified through data analysis and sections where each is discussed in the paper.

Highest-Level Themes	Relevant Paper Sections
Participants perceive Wikibench as effective in surfacing disagreements and facilitating the development of a shared consensus.	8.1.1, 8.1.2, 9.3.3
Participants appreciate Wikibench's collaborative approach to data labeling because it allows contributors to build consensus and a stronger community.	8.1.3, 9.3.1
Participants find data produced by Wikibench helpful in understanding gaps between different AI models' predictions and community consensus.	8.2
Participants perceive that Wikibench's user interface and process fit naturally into Wikipedia's existing interface, workflow, and community norms.	9.1
Participants perceive that Wikibench provides them with the agency to collectively shape and reflect on the data curation process.	9.2
Participants find Wikibench's campaign and entity pages helpful for quickly pinpointing edits where more labels or discussions may be valuable.	9.3.2
Participants believe the transparency Wikibench provides into the data and the process by which it is curated is essential for evaluation results to be trustworthy to the community.	9.3.4

Table 4: Validation study participant demographics, including their self-identified experience and frequency of patrolling edits, registration year, edit count on English Wikipedia, and geographic location. A dash indicates that they chose not to provide that information.

Participant ID	Patrol Experience	Patrol Frequency	Registered Since	Edit Count	Location
V1	Years	Weekly	2021	6.6k	–
V2	Years	Monthly	2018	10k	United States
V3	Years	Daily	2008	55k	United States
V4	Years	Weekly	2010	12k	United States
V5	Months	Daily	2022	20k	United Kingdom