LabelAId: Just-in-time AI Interventions for Improving Human Labeling Quality and Domain Knowledge in Crowdsourcing Systems

Chu Li*
Zhihan Zhang*
chuchuli@cs.washington.edu
zzhihan@cs.washington.edu
University of Washington, USA

Michael Saugstad
University of Washington, USA
saugstad@cs.washington.edu

Esteban Safranchik
University of Washington, USA
estebans@cs.washington.edu

Minchu Kulkarni
University of Washington, USA
minchu@uw.edu

Xiaoyu Huang
University of California, Berkeley, USA
haytham.huang@berkeley.edu

Shwetak Patel
University of Washington, USA
shwetak@cs.washington.edu

Vikram Iyer
University of Washington, USA
vsiyer@cs.washington.edu

Tim Althoff
University of Washington, USA
althoff@cs.washington.edu

Jon E. Froehlich
University of Washington, USA
jonf@cs.washington.edu

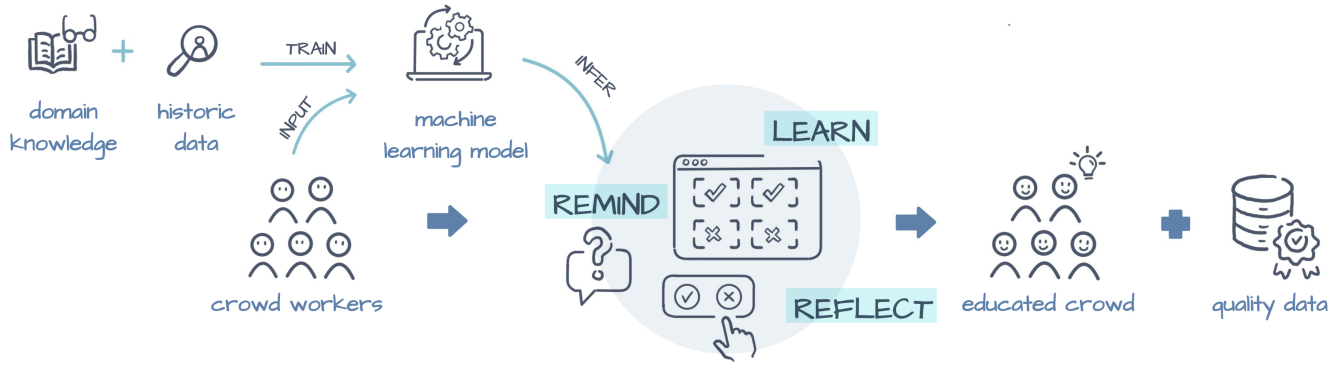


Figure 1: We introduce *LabelAId*, an ML-based inference system to provide just-in-time feedback during crowdsourced labeling to improve data quality and user expertise. LabelAId consists of: (1) a novel ML-based pipeline for detecting labeling mistakes, which is efficiently trained to infer label correctness based on user behavior and domain knowledge; (2) a real-time ML model and UI that tracks worker behavior and intervenes when an inferred mistake is occurring.

ABSTRACT

Crowdsourcing platforms have transformed distributed problem-solving, yet quality control remains a persistent challenge. Traditional quality control measures, such as prescreening workers and refining instructions, often focus solely on optimizing economic output. This paper explores just-in-time AI interventions to enhance both labeling quality and domain-specific knowledge among crowdworkers. We introduce LabelAId, an advanced inference model combining *Programmatic Weak Supervision* (PWS) with *FT-Transformers* to infer label correctness based on user behavior and domain knowledge. Our technical evaluation shows that our LabelAId pipeline consistently outperforms state-of-the-art ML baselines, improving mistake inference accuracy by 36.7% with 50 downstream samples. We then implemented LabelAId into Project Sidewalk, an open-source crowdsourcing platform for urban accessibility. A between-subjects study with 34 participants demonstrates that LabelAId significantly enhances label precision without compromising efficiency while also increasing labeler confidence. We discuss LabelAId's success factors, limitations, and its generalizability to other crowdsourced science domains.

*Both authors contributed equally to this research.

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

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642089

CCS CONCEPTS

• **Computing methodologies → Machine learning**; • **Information systems → Crowdsourcing**; • **Human-centered computing → Interactive systems and tools**.

KEYWORDS

crowdsourcing, community science, quality control, machine learning, programmatic weak supervision (pws), urban accessibility, human-ai collaboration

ACM Reference Format:

Chu Li, Zhihan Zhang, Michael Saugstad, Esteban Safranchik, Minchu Kulkarni, Xiaoyu Huang, Shwetak Patel, Vikram Iyer, Tim Althoff, and Jon E. Froehlich. 2024. LabelAId: Just-in-time AI Interventions for Improving Human Labeling Quality and Domain Knowledge in Crowdsourcing Systems. 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, 21 pages. https://doi.org/10.1145/3613904.3642089

1 INTRODUCTION

Crowdsourcing systems have transformed distributed human problem-solving, enabling large-scale collaborations that were previously infeasible [41]. Quality control, however, remains a persistent challenge leading to noisy or unusable data [16, 48]. Existing quality control measures such as prescreening crowdworkers [20, 44], refining instructions [24, 47, 81], manipulating incentives [24, 47, 81], and majority vote filtering are designed to optimize economic output: data quality and worker efficiency. Our research explores a subset of crowdsourcing that focuses on community science, or *crowdsourced science* [74]. Platforms like *Zooniverse* [84] and *FoldIt* [46] engage non-professionals in scientific tasks and serve as important means of public engagement and education [74, 90]. Since participants are primarily volunteers, crowdsourced science presents unique quality control challenges: users are primarily motivated by intrinsic interest, learning opportunities, and making a difference but may be unfamiliar with the domain [68]. Previous work in crowdsourcing has explored the dual objectives of enhancing work quality as well as learning experience in crowdsourcing systems by providing feedback to crowdworkers [21–23, 90, 99]. Yet, these approaches are less scalable because they require additional commitments from either crowdworker peers or external experts [22, 23, 90, 99].

Building on this prior work, we present *LabelAId*, a real-time inference model for providing just-in-time feedback during crowdsource labeling to improve data quality and worker expertise. LabelAId is composed of two parts: (1) a novel machine learning (ML) based pipeline for detecting labeling mistakes, which is efficiently trained on unannotated data that contain those very mistakes; (2) a real-time system that tracks worker behavior and intervenes when an inferred mistake occurs. Unlike previous approaches that improve crowdworkers' learning experience through peer or expert feedback [23, 99], LabelAId reduces the reliance on human input, leveraging human-AI collaboration to provide targeted feedback for enhancing crowdworker performance and domain knowledge.

To study LabelAId in a real crowdsourcing context, we instrumented the open-source crowdsourcing tool, *Project Sidewalk*, where online users virtually explore streetscape imagery to find, label, and assess sidewalk accessibility problems for people with mobility disabilities [80]. Since its launch in 2015, over 13,000 people across the world have used Project Sidewalk to audit 17,000 km of streets across 20 cities in eight countries including the US, Mexico, Ecuador, Switzerland, New Zealand, and Taiwan, contributing over 1.5 million data points[1].

Project Sidewalk provides a compelling use case for LabelAId because, unlike traditional image labeling tasks for object detection (*e.g.*, *ImageNet* [19], *COCO* [58], *Open Images Dataset* [30]), crowdworkers are asked to make careful judgments about a labeling target, which requires domain knowledge and training—similar to agricultural image recognition [29], medical imagery labeling [79, 98], and wildlife image categorization [4]. Such labeling tasks reflect a broader trend of crowdwork becoming increasingly complex, domain-specific, and potentially error prone [48]. Second, as a community science project, Project Sidewalk aligns with the growing emphasis on both educational impact and data quality in crowdsourcing [21–23, 90, 99], which LabelAId provides. Finally, Project Sidewalk currently employs a common but limited quality control mechanism: users validate labeled images by other users. Since both labelers and validators are drawn from the same user population, repeated errors can pervade the system.

To evaluate LabelAId, we conducted: (1) a technical performance evaluation of LabelAId's inference model; and (2) a between-subjects user study of 34 participants. For the former, we demonstrate that the LabelAId pipeline consistently outperforms state-of-the-art baselines and can improve mistake inference accuracy by up to 36.7%. With fine-tuning on as few as 50 expert-validated labels, LabelAId outperforms traditional ML models such as *XGBoost* [17] and *Multi-layer Perceptron* (MLP) [87] trained on 20 times the amount of expert-validated labels. Furthermore, we showcase the robust generalizability of our pipeline across different deployment cities in Project Sidewalk. Since its initial deployment in Washington D.C., Project Sidewalk has expanded to 20 cities, with ongoing plans for further growth. To support future city deployments, it is important to minimize labor and configuration overhead of the mistake inference model in new cities. Our study shows that LabelAId, even without fine-tuning, performs comparably in a new city to those in the pre-training set.

For the between-subjects user study, participants were randomly assigned to one of two conditions: using Project Sidewalk in its original form (control) or using Project Sidewalk with LabelAId (intervention). Our findings reveal that the intervention group achieved significantly higher label precision without sacrificing labeling speed. While using Project Sidewalk enhanced participants' understanding of urban accessibility and their confidence in identifying sidewalk problems in both groups, participants in the intervention group reported that LabelAId was helpful with decision-making, particularly in situations where they were initially uncertain.

To summarize, our contributions are as follows:

- A novel ML pipeline that allows for the integration of domain-specific knowledge and heuristics into the data annotation process, which facilitates the training of AI-based inference models for detecting crowdworker labeling mistakes across

[1] https://projectsidewalk.org/

various contexts, while minimizing the need for manual intervention in downstream tasks.
- A human-AI (HAI) collaborative system designed to create teachable moments in crowdsourcing workflows. This system not only improves the quality of crowdsourced data, but also enriches the learning experience for participants.
- A between-subjects user study involving 34 participants with no prior experience using Project Sidewalk, demonstrating that LabelAId significantly improves label precision by 19.2% without compromising efficiency.

While our empirical results focused on the performance of LabelAId within the context of Project Sidewalk, we believe our framework can be generalizable to other crowdsourcing platforms as well as the PWS-based ML pipeline and the two-step module design intervention are easily replicable and tailorable in different contexts.

2 RELATED WORK

Our work draws on, and contributes to research in improving the quality of crowdsourcing, enhancing crowdworkers' domain knowledge, and inferring the correctness of labels using ML methods.

2.1 Improving Quality of Crowdsourced Labels

Distributed crowdwork has transformed how loosely connected individuals collaborate together to solve large-scale problems such as protein folding [18] map building [35], and writing compendiums of knowledge [93]. Despite decades of research, however, large-scale crowdwork remains susceptible to quality control problems [14, 48]. For example, studies have shown that over 30% of MTurk submissions are likely to be poor quality [5, 47]. Current quality control methods can be broadly categorized into two groups: preventive techniques and post-hoc detections. Preventive measures include screening crowdworkers based on capabilities [20, 44], dividing work into fault-tolerant sub-tasks [5, 49, 65], improving instructions [24, 47, 81] and changing payment structures [13, 76, 81]. Post-hoc measures involve filtering based on majority vote [86] and employing additional crowdworkers to review others' work [11, 36]. Project Sidewalk currently uses both strategies: an interactive tutorial to train crowdworkers as their "first mission" and *post-hoc* validation where crowdworkers "vote" on the correctness of other users' labels.

Other quality control research examines how workers do their work rather than the end product itself, using ML algorithms to predict the quality of crowdworkers' output based on their behaviors [28, 77, 78]. This method captures behavioral traces from workers during task execution and uses them to predict quality, errors, and the possibility of cheating [28, 77, 78]. These behavioral traces are gathered by logging user interactions, which are then formulated into interaction patterns for monitoring real-time worker compliance [78]. This methodology, termed *"fingerprinting"* by Rzeszotarski and Kittur [78], has demonstrated its efficacy in predicting crowdworker output quality. Expanding Rzeszotarski and Kittur's work, Kaza and Zitoun [45] investigated using the behavior of trusted, trained judges to identify low-performing workers. Their study, which involved assessing the relevance of web pages to specific queries, showed that the classification accuracy nearly doubled in some tasks. However, the approach is challenging to scale due to the need for trained judges.

Building on this body of research, we introduce an ML pipeline that combines crowdworker behavioral data with expert domain knowledge (in our case, drawn from urban accessibility but the approach should generalize to other domains). This model aims to more effectively and automatically guide crowdworkers through their efforts in identifying street-level accessibility issues.

2.2 Teachable Moments in Crowdsourcing for Community Science

Crowdsourcing for community science are initiatives where professional scientists seek the assistance of crowds in contributing to scientific research [35, 74]. Platforms like *Zooniverse* [84], *FoldIt* [46], and *SciStarter* [40] are notable for having involved non-professionals in significant scientific discoveries. Beyond contributing to science, these platforms serve as tools for public engagement, outreach, and education [90]. Unlike crowdworkers driven by monetary incentives (*e.g.,* on MTurk and Prolific), participants of community science projects are primarily volunteers motivated by desires to learn and contribute to scientific research [69, 73].

Recent crowdsourcing research has been investigating ways to not only enhance the quality of work but also the learning experience of participants [21–23, 90, 99]. For instance, Dow *et al.* [23] demonstrated that timely, task-specific feedback can help crowdworkers learn, preserve, and produce better results. Projects like *Crowdclass* [54] and *CrowdSCIM* [90] introduced in-task learning modules for community science initiatives. Despite these advances, current methods facilitating learning through crowdsourcing, such as peer-review [23, 99], expert feedback [23], and self-assessment [23], all require additional commitments from either the crowdworkers or external experts, limiting their scalability.

Recent developments in HAI collaboration presents new ways to tackle these scalability issues. Matsubara *et al.* [61] suggest using machine predictions as reference answers for self-correction. Nakayama *et al.* [62] extended this concept, proposing workflows where AI learns alongside human workers without prior training. Inspired by this evolving landscape, we propose an add-on system for crowdsourcing systems like Project Sidewalk that enhances both task quality and educational outcomes. It leverages HAI collaboration to enable in-context learning without requiring additional commitments from participants.

2.3 Machine Learning to Infer Label Correctness

To create teachable moments in crowdsourcing workflows, it is essential to develop inference models for detecting crowdworker labeling mistakes. Recent research has harnessed the power of ML to infer crowdsourced label quality. For instance, computer vision-based neural networks are applied to validate crowdsourced labels of sidewalk accessibility problems in *Google Street View* (GSV) imagery [91], and also offer reference answers to image recognition labelers enabling self-correction [61, 62]. These deep learning methods present promising solutions for aiding crowdsourced labeling tasks, but they often require substantial training data. While general labeled image datasets such as [19, 30, 58] are available, domain-specific datasets are relatively rare and expensive to produce.

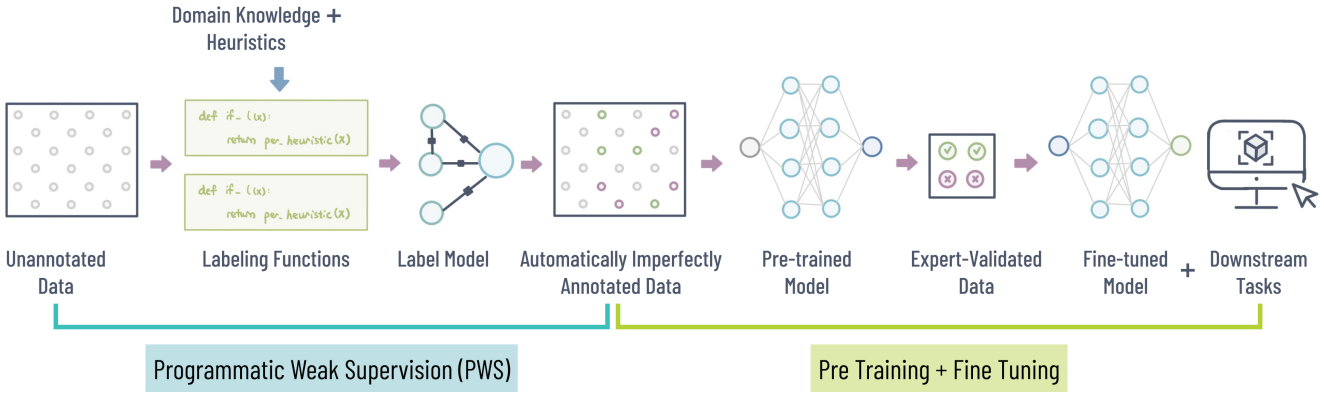


Figure 2: An overview of our LabelAId pipeline. Programmatic weak supervision, utilizing domain-specific knowledge and heuristics, is employed to annotate the raw data. Subsequently, the automatically imperfectly annotated data generated from PWS are used to pre-train the inference model. Lastly, the inference model is fine-tuned using expert-validated labels for the target downstream task. Diagram adapted from [70].

In this paper, we aim to minimize the need for manually-annotated data for training AI-based inference models. The recently proposed *Programmatic Weak Supervision* (PWS) framework [70–72] provides a promising approach: aggregating noisy votes from domain knowledge, heuristics, external patterns and rules to assign annotations to the raw data. Subsequently, this annotated data serves to train models for a range of domain-specific tasks, including video analysis [89], text classification [83], and sensor data analysis [27]. However, its potential for inferring and improving the quality of crowdsourced labels remains unexplored.

3 LABELAID: A LABEL CORRECTNESS INFERENCE FRAMEWORK

LabelAId is a novel ML pipeline designed to provide just-in-time intervention in crowdsourced labeling tasks, inferring and identifying labeling mistakes as they occur. At its core, LabelAId tackles a major hurdle in deploying such AI-based inference models: the need for large volumes of annotated training data, which is particularly scarce in crowdsourced environments. LabelAId introduces both (1) a programmatic pipeline to train an efficient ML inference model to detect crowdworker labeling mistakes, which is trained on unannotated data that contain those very mistakes and minimal expert-validated data, and (2) an example application of the LabelAId pipeline to a crowdsourcing system to recognize and intervene when a user is making a labeling mistake. Our overarching goal is to create a full end-to-end HAI pipeline with minimal expert involvement for model training, thereby facilitating rapid deployment of LabelAId mistake inference models in real-world crowdsourcing contexts. In this section, we describe LabelAId's pipeline design, how we implemented LabelAId in Project Sidewalk, and our technical evaluation. Our code is available on Github.[2]

3.1 LabelAId Pipeline

Developing an inference model capable of discerning specific user mistakes while also being generalizable to general crowdworkers' behaviors is challenging and necessitates the consideration of a variety of quality signals. Our LabelAId pipeline is composed of three core phases (Figure 2): (1) *Programmatic weak supervision* (PWS) uses domain-specific knowledge about crowdsourcing tasks and historical heuristics from crowdworker behavior; this phase generates a set of *Automatically, Imperfectly Annotated* (AIA) probabilistic training data. (2) We then pre-train an ML model using the AIA data generated from PWS. The inference model learns general features and representations of the target crowdsourcing task in this phase. (3) Finally, we fine-tune the inference model using a small number of expert-validated labels to further enhance its performance for the target task.

3.1.1 Programmatic Weak Supervision (PWS). Due to the inherently noisy nature of crowdsourced data, LabelAId adopts PWS as its foundational architecture. PWS takes unannotated data and produces probabilistic training labels, and has demonstrated effectiveness in tasks like document and numerical data classification [6, 97]. We use the popular Snorkel [70] platform as the backbone of our PWS pipeline. PWS allows for the integration of domain knowledge and heuristic guidelines into the data annotation process and provides a method for estimating their conflict and correlation in a programmatic manner. As a result, the probabilistic training labels can be reweighted and combined to create high-quality labels. This approach aligns well with our objectives for two key reasons: first, it allows us to train models on unannotated data, eliminating the need for manually labeling large datasets; and second, it enables the incorporation of domain-specific knowledge related to crowdsourcing tasks into our pipeline.

One of the core components of Snorkel is *Labeling Functions* (LFs), which are rules or heuristics humans code to annotate raw data programmatically. When incorporating a set of LFs, they introduce distinct characteristics and correlations that extend beyond disparities in accuracy and coverage. It is important to recognize that LFs are not equal in their contribution to the annotation. LFs may also overlap and conflict, which cannot be resolved by a simplistic hard-rule-based approach (see Appendix A.2 for additional

[2] https://github.com/makeabilitylab/LabelAId

	Seattle		Chicago		Oradell		
	Unannotated	Expert-Validated	Unannotated	Expert-Validated	Unannotated	Expert-Validated	**Total**
Curb Ramp	70,690	5,333	5,710	2,386	660	859	**85,638**
Missing Curb Ramp	32,968	4,239	463	1,294	325	396	**39,685**
No Sidewalk	36,021	3,460	2,211	48	3,949	1,217	**46,906**
Surface Problem	26,912	2,909	2,136	1,651	2,544	1,222	**37,374**
Obstacle	10,103	407	1,254	320	106	158	**12,348**
All Label Types	**176,694**	**16,348**	**11,774**	**5,699**	**7,584**	**3,852**	**221,951**

Table 1: Set sizes of unannotated and expert-validated labels, contain Project Sidewalk labels from Seattle, WA; Chicago, IL; and Oradell, NJ. The unannotated set is used to pre-train the model after our PWS annotation process. The expert-validated set is used to fine-tune and evaluate the inference model, which was created from labels manually-validated by the Project Sidewalk research team.

discussion). While previous approaches used majority-vote-based models to handle these intricacies, it can result in an overrepresentation of specific signals when two features are highly correlated [70]. To address this, we instead opt for probabilistic graphical models to integrate the outputs of LFs. We use the Snorkel label model to take the complete set of LFs as its input and generate a matrix of LFs, Λ. We aim to maximize the probability of the outputs of the LFs [96] in the context of our label correctness inference task, *i.e.*, a binary classification. Assuming θ is the set of model parameters and Y is the prediction class, this objective transforms into an optimization problem as the following equation:

$$\max_{\theta} P(\Lambda; \theta) = \sum_{Y} P(\Lambda, Y; \theta)$$

The label model generates a single set of noise-aware, probabilistic training labels, which can be used to train an ML model or neural network.

3.1.2 Transfer Learning from AIA to Expert-Validated Labels. The ultimate goal is to train a discriminative ML model—such as a neural network classifier—that can generalize beyond the label model. The choice of the specific ML model architecture should be tailored to the requirements of the downstream task.

While PWS produces a set of AIA probabilistic training data, an ML model trained on a large yet noisy dataset could potentially overfit to the noise, which should be avoided. Fine-tuning large pre-trained models is a popular methodology for leveraging knowledge from a related source domain to improve the learning performance in a target task with limited clean samples [100].

Since the source domain and target task in our LabelAId pipeline share the same feature space, we introduce a two-phase transfer learning pipeline from AIA data to expert-validated labels: (1) Pre-training the ML model on the AIA dataset $\mathcal{D}_n$ produced by our PWS pipeline; (2) Fine-tuning the pre-trained model on a small number of expert-validated downstream labels $\mathcal{D}_g$ for a specific task. Our pre-training and fine-tuning pipeline is illustrated in Figure 2. This approach aims to minimize the requirement of manual intervention in annotating downstream data for a target task.

3.2 Applying LabelAId to Project Sidewalk

We present a specific implementation of our end-to-end LabelAId pipeline using domain knowledge in urban accessibility for Project Sidewalk (Figure 3), a large-scale and widely used sidewalk accessibility platform leveraging crowdsourced labels. We describe the dataset, details of labeling functions and features, and model training process below.

3.2.1 Dataset Description. At the time of our analysis, Project Sidewalk had 757,730 crowdsourced image labels applied on top of *Google Street View* (GSV) panoramas across 15 cities in the US, Mexico, and Europe. Project Sidewalk has five main label types: *Curb Ramp*, *Missing Curb Ramp*, *Sidewalk Obstacle*, *Surface Problem*, *Missing Sidewalk*. Each label includes a severity assessment on a scale of 1 to 5, with 5 being the most severe, indicating a sidewalk issue that is impassable for a wheelchair user. Labels may also include an open-ended description and label-specific tags. Additionally, all labels are accompanied by implicitly captured metadata, such as the GSV image date, the label timestamp, and geographical location (latitude and longitude).

As urban composition and sidewalk accessibility guidelines differ across regions, our initial proof-of-concept of LabelAId focuses on three U.S. cities: Seattle, WA; Chicago, IL; and Oradell, NJ. These three cities offer distinct urban and geographical characteristics: Seattle represents a major city in the Pacific Northwest, the Chicago data provides a mix of dense downtown area and satellite towns in the Midwest, and Oradell is a suburban locale on the East Coast outside of New York City. We anticipate that this selection will help us develop ML models that could effectively account for diverse urban compositions. The ground truth dataset $\mathcal{D}_g$ is validated by the Project Sidewalk research team; two researchers worked collaboratively to verify each crowdsourced label, reaching unanimous agreement. The finalized dataset sizes alongside the distribution of each label type are shown in Table 1.

3.2.2 Input Features & Labeling Functions. The suitability of PWS for Project Sidewalk comes from two key factors: the integration of domain-specific knowledge (*e.g.*, urban planning guidelines) and user behavior insights into our labeling process.

Drawing on observations of user behavior in Project Sidewalk and research in urban planning guidelines, we propose the following hypotheses:

- **Severity Rating.** Project Sidewalk's label severity ranges between 1 to 5. Severity ratings closer to the extremes (1 & 5) are more likely to be correct.

Figure 3: (A) Project Sidewalk Labeling Interface. (B) Project Sidewalk Label Types. (C) Examples of Project Sidewalk severity ratings for surface problems. Severity 5 is the most severe, indicating a scenario impassable by wheelchair users.

- **Optional input.** Labels that include optional data are more likely to be accurate, because such information requires additional thought and effort. These optional fields include a free-form description (comment) and relevant tags, such as *fire hydrant* and *pole* for the *obstacle* label type.
- **GSV zoom/pitch/heading.** In most cases, changing the default parameters of GSV results in a more accurate label. For example, when a user zooms in to place a *Surface Problem* label, it is more likely to be correct.
- **Distance to other crowdworkers' labels.** A label is more likely to be correct if it is placed closer to existing labels of the same type. To determine this distance, we adopt the two-step spatial clustering approach employed in Project Sidewalk [80].
- **Distance to urban infrastructure.** The positioning of a label in relation to urban infrastructure can serve as an indicator of its accuracy. For example, US federal legislation [66] requires the installation of curb ramps at all intersections and at midblock locations where pedestrian crossings are present. Given that midblock crossings are relatively rare compared to those at intersections, and considering that the most common error in Project Sidewalk is mislabeling driveways as curb ramps [91], we hypothesize that a *Curb Ramp* label situated outside a specified radius from an intersection is likely to be incorrect.

We then derived eight LFs from our hypotheses for Project Sidewalk and integrated all these LFs into the PWS pipeline to ensure a diverse coverage and minimize overfitting (see Appendix A.1 for additional discussion). One example algorithm (Algorithm 1) is based on the observation that users often mislabel driveways as curb ramps in residential areas. The algorithm proposes that a *Curb Ramp* or *Missing Curb Ramp* in a residential area is likely to be wrong when it is far away from an intersection.

3.2.3 Multi-city Pre-training. To train a discriminative model, we start by pre-training on the AIA dataset $\mathcal{D}_n$ to initialize the weights for high-level patterns of sidewalk accessibility labels and user behaviors. Due to the mix of categorical and numerical features within Project Sidewalk's datasets, we chose the *Feature Tokenizer + Transformer* (FT-Transformer) [31]. FT-Transformer represents a novel adaptation of the *Transformer* architecture for tabular data domains. Prior research [31, 56] has shown that the FT-Transformer is a more universal architect for tabular data, and consistently outperforms other state-of-the-art deep tabular models across a variety of downstream sample regimes. A high-level view of our FT-Transformer-based model architecture is illustrated in Figure 4.

We employ three distinct datasets to train, validate, and test our backbone FT-Transformer-based model. We used the AIA datasets $\mathcal{D}_n$ from three cities generated from our PWS pipeline, which consists of 176,694, 11,774, and 7,584 labels for Seattle, Chicago, and Oradell, respectively. We balanced and randomly partitioned these labels into training, validation, and testing sets following a 70/20/10 split. In each subset, a minimum of 20 labels were guaranteed for every class of every label type.

We trained an FT-Transformer from scratch. To map all inputs into the same embedding space, we encode the numerical data through a single-layer perceptron with a dimension of four, and embed categorical data with one-hot embeddings of the same size. Then, we stack them to formulate the input embeddings for the Transformer module. The encoder was configured with a depth of two layers, each comprised of two attention heads. To prevent

Algorithm 1 Example Labeling Function encoding a heuristic about errors made in (Missing) Curb Ramp labels

Require: labels ∈ CurbRamp or NoCurbRamp
Ensure: labels ∈ Residential Area
  $D \leftarrow$ Intersection Distance Threshold
  for each label l **do**
    for each intersection i in nearby intersections I **do**
      Compute spatial distance between l and i
    end for
    if min distance$(l, I) > D$ **then**
      $l \leftarrow$ wrong
    else
      $l \leftarrow$ correct
    end if
  end for

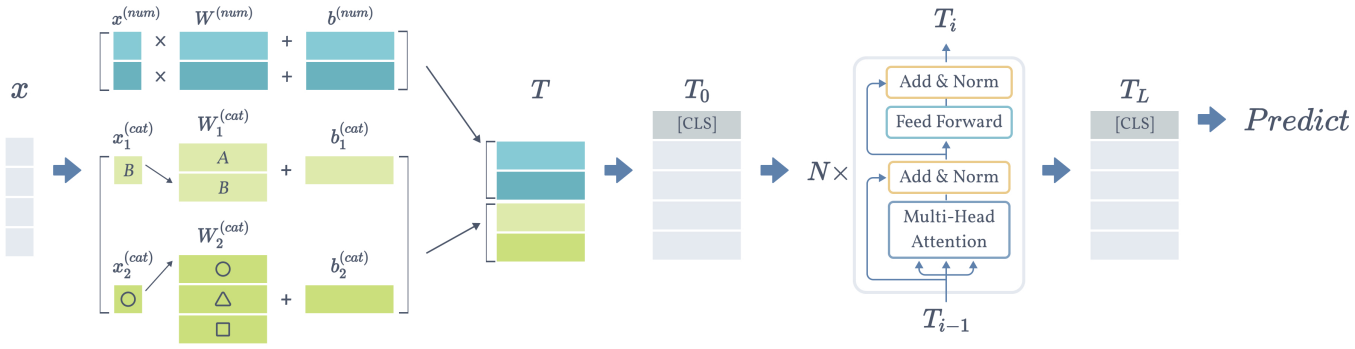


Figure 4: Conceptual diagram of our FT-Transformer-based model architecture. First, the model transforms the hybrid features (*e.g.*, two numerical and two categorical features) into unified embeddings. Subsequently, these embeddings are processed iteratively by the Transformer layer. The final output is based on the [CLS] token. Diagram adapted from [31].

overfitting and to provide regularization, we incorporated attention and feed-forward dropouts, both set at 0.2. Additional optimal hyperparameters were tuned via grid search, according to the full upstream dataset: the *AdamW* optimizer, a learning rate of 1×10^{-4}, and a weight decay of 1×10^{-5} were selected. Given our binary classification objective, we employed the *Binary Cross-Entropy* loss function coupled with a sigmoid activation function. We employed 200 epochs for training with early stops on validation loss. The average training duration was approximately 5 minutes on a single RTX 4070Ti GPU.

3.2.4 Fine-tuning on a Specific City. Upon delving deeper into our pipeline, the pre-trained model, which learned underlying patterns of sidewalk accessibility labels and a broad understanding of high-level user behaviors, can subsequently be fine-tuned on a smaller, city-specific $\mathcal{D}_g$ to elevate its performance for city-specific tasks. The rationale behind this is that each city has unique attributes that might not be entirely captured during this pre-training phase. Through fine-tuning, the model can adapt its previously acquired knowledge to the unique characteristics and nuances of the target city. This not only enables more precise inferences and understanding of the distinctive topologies of the target city, but also ensures that the model remains robust. To evaluate its performance enhancement, we conducted city-specific fine-tuning. The pre-trained FT-Transformer was fine-tuned end-to-end [56] with 200 downstream samples (40 per label type) sourced from Seattle, Chicago, and Oradell's $\mathcal{D}_g$, with a reduced learning rate of 3×10^{-5} to prevent overfitting to downstream samples.

3.3 Technical Evaluation

We evaluate the performance of our pipeline, which involves the PWS pipeline integrated with city-specific fine-tuning of a pre-trained inference model. We examine the following aspects: (1) How our pipeline performs in comparison to traditional ML methods in inferring label correctness; (2) The generalization of our model across cities after target-city-specific fine-tuning; (3) The model's generalizability to new cities that were not included in our pre-training dataset; (4) How different types of LFs complement each other in achieving LabelAId's performance through analysis of feature importance.

For testing purposes, we utilized the remaining $\mathcal{D}_g$ after removing the samples used for fine-tuning. This allowed us to evaluate the performance and generalizability of the final model on a previously unseen dataset. Specifically, the remaining $\mathcal{D}_g$ comprises 16,148, 5,499, and 3,652 expert-validated labels for Seattle, Chicago, and Oradell, respectively.

3.3.1 Pipeline Performance. We compared our performance against two widely adopted ML classifiers–random forest and logistic regression, and a GBDT tabular method XGBoost [17]. We also considered an MLP classifier which is known to be a consistent and competitive baseline [31]. All baselines are trained on the equivalent volumes of expert-validated downstream samples $|\mathcal{D}_g|$ and do not undergo pre-training using $\mathcal{D}_n$ produced by our PWS pipeline.

To ensure a robust evaluation, the optimal hyperparameters for baseline models were tuned via a grid search, executed on a single, randomly split downstream dataset. This procedure ensures that all baselines are tuned with an equivalent number of samples, given that the hyperparameters are profoundly influenced by sample size. Each baseline model underwent training using 50, 100, 200, 500, and 1000 expert-validated labels from Seattle $\mathcal{D}_g$. The samples were equally distributed across two classes and five label types. For our experiment, we fine-tuned the pre-trained FT-Transformer end-to-end [56] on the equivalent number of expert-validated downstream samples from Seattle $\mathcal{D}_g$ with the baselines.

The results demonstrate that our pipeline enhances the efficiency of neural network training. The integration of relevant domain knowledge through PWS bypasses the resource-intensive task of manually labeling and validating the unannotated Project Sidewalk repository. As shown in Figure 5, our proposed pipeline with as few as 50 expert-validated labels (10 per label type), achieved a test accuracy and precision of 73.3% and 83.6%, respectively. It outperformed all baselines of traditional ML methods, even when those were trained on substantially more expert-validated labels. Our pipeline improved accuracy by up to 36.7%, 28.0%, 15.3%, 16.3%, 16.5% for $|\mathcal{D}_g|$ = 50, 100, 250, 500, 1000, respectively. Our pipeline also demonstrated an average 0.0859 boost in F1 score compared to the second-best method.

	Chicago		Oradell	
	accuracy	F1	accuracy	F1
Pre-trained on multi-city' $\mathcal{D}_n$	0.679	0.787	0.808	0.893
Fine-tuned on target city's $\mathcal{D}_g$	**0.719**	**0.814**	**0.914**	**0.945**

Table 2: Improvements of inference model in accuracy and F1 score for Chicago and Oradell after city-specific fine-tuning (fold K = 5).

3.3.2 Generalizability Across Cities. As shown in Table 2, in Oradell, the pre-trained model, without fine-tuning, achieved an accuracy, precision, and recall of 80.8%, 91.9%, and 86.9%, respectively. After the fine-tuning process, these figures rose to 91.4%, 92.4%, and 96.8%. Similarly, for Chicago, the pre-trained model achieved accuracy, precision, and recall values of 67.9%, 80.4%, and 77.0%, respectively. Post fine-tuning, these metrics improved to 71.9%, 82.8%, and 80.1%. One plausible explanation for this lower performance could be the non-continuous geographic distribution of Project Sidewalk's data in Chicago, which includes a mix of dense urban downtown areas and pockets of suburbia. Variations in road width and intersection distances across these areas could complicate the model's ability to make accurate inferences.

3.3.3 Generalizability in New City. To fully evaluate the model's generalizability, we deployed it to a new city: Newberg, OR—a small town in the Portland metropolitan area with a population of 25k, similar in urban composition to Oradell. When applying the pre-trained model to Newberg, which was neither previously pre-trained nor fine-tuned, the model showcased accuracy, precision, and recall of 78.3%, 88.2%, and 86.0%, respectively. These scores represent on-par performance with the cities in the pre-training set, such as Oradell, NJ. This not only underscores the robust generalizable foundation of the multi-city pre-trained inference model, but also highlights that the pre-trained model can be deployed in a new city without any manual intervention and achieve respectable performance if the new city has a similar urban composition and crowdworker behavior to those in the pre-training set.

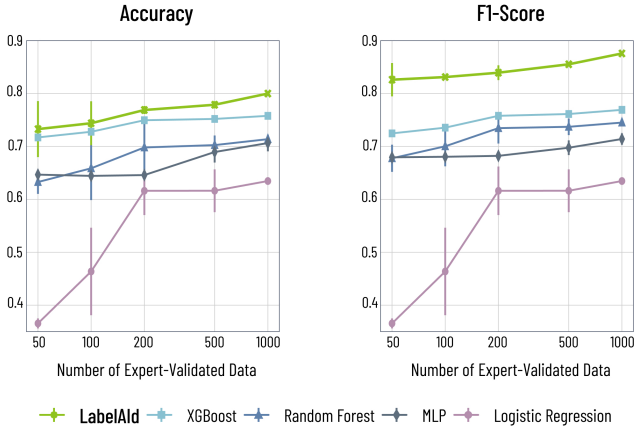


Figure 5: Overall performance of our LabelAId pipeline compared to the traditional ML methods as the number of expert-validated downstream labels increases. Note that the x-axis is on a log scale (N = 3, error bar = $\pm\sigma$).

City	Curb Ramp	Missing Curb Ramp	Obstacle	Surface Problem	Missing Sidewalk
Seattle	0.971	0.966	**0.766**	0.861	0.942
Chicago	0.968	**0.494**	**0.693**	0.718	0.929
Oradell	0.972	0.768	0.793	0.944	0.988

Table 3: Performance by label type of our inference model in F1 score for Seattle, Chicago, and Oradell. *Missing Curb Ramp* is a notable area of difficulty in Chicago. *Obstacle* is a low performer in Seattle and Chicago.

3.3.4 Performance by Label Type. We also analyzed performance as a function of label type for each city-specific fine-tuned model (Table 3). The model performs best for *Curb Ramp* and *Missing Sidewalk* across all cities, followed by *Surface Problem*. However, *Obstacle* is a low performer, especially in Seattle and Chicago. A close look at the tags associated with *Obstacle* labels revealed that the observed discrepancies might be explained by the complexity of sidewalk obstacles in these two cities. Specifically, in Oradell, obstacles were primarily associated with trees/vegetation (40%), whereas in Seattle, *Obstacle* were tagged with *poles*, *trash/recycling cans*, *vegetation*, and *parked cars* in similar frequencies of ~20%. Another low performer in Chicago is *Missing Curb Ramp*, with a low F1 score of 0.494. This is associated with user behavior exclusive to Chicago, where curb ramps lacking tactile strips are often mislabeled as *Missing Curb Ramp*. These findings highlight the high performance of our inference model for most scenarios, however further refinement is necessary to accommodate different urban environments and user behaviors.

3.3.5 Feature Importance. To explore feature importance, we used attention maps [31]. We expect that if certain features are used as input variables in LFs for a specific sidewalk label type, these features will be highly significant in the model's inference for that label type. For instance, the feature "distance to intersection" is an input feature for our model and also a variable in the LF (min distance(l, I) in Algorithm 1) for *Curb Ramp* and *Missing Curb Ramp*.

The feature importance results of the model fine-tuned on Seattle $\mathcal{D}_c$ are shown in Table 4. For label types *Curb Ramp* and *Missing Curb Ramp*, "distance to intersection" is the most important; while for *Obstacle* and *Surface Problem*, features like "zoom", "cluster" and "description" are more crucial. This difference suggests that the features influencing *Curb Ramps* are related to LFs based on urban planning knowledge, whereas those affecting *Obstacles* and *Surface Problems* are tied to user behavior. Specifically, mislabeled *Curb Ramps* exhibit a spatial pattern, making them identifiable using domain knowledge (*e.g.* Algorithm 1). In contrast, *Obstacle* and *Surface Problem* labeling mistakes are less about spatial distribution and more about user labeling diligence, characterized by zooming in and adding optional descriptions. The results of our feature importance ranking show how LFs based on urban planning and user behavior are complementary in achieving LabelAId's performance.

#Rank	Curb Ramp		Missing Curb Ramp		Obstacle		Surface Problem		Missing Sidewalk	
1	distance_i	0.173	distance_i	0.177	clustered	0.145	zoom	0.143	severity	0.129
2	way_type	0.103	way_type	0.106	zoom	0.135	clustered	0.135	tag	0.116
3	severity	0.103	tag	0.103	description	0.134	description	0.132	distance_r	0.101

Table 4: Top 3 features and their importance coefficient per label type in Seattle. Note: distance_i is the distance to intersection, distance_r is the distance to road, and way_type is the road hierarchy according to OpenStreetMap.

We believe such techniques generalize to other crowdsourcing platforms where user mistakes can be identified through a combination of domain guidelines as well as platform specific behaviors.

3.3.6 Inference Limitations. Finally, to understand the limitations of our model and identify opportunities for improvement, we conducted a qualitative assessment of our inference model by manually reviewing 100 randomly selected false positives and false negatives across each label type, and presented the results in Figure 6 & 7.

In analyzing false positives (where the model incorrectly infers the label as correct when, in fact, the label is wrong), we observed two key sources of error for *Curb Ramp* and *Missing Curb Ramp*: (1) The model fails to differentiate between a *Curb Ramp* and a *Missing Curb Ramp* (Figure 6a, c). (2) In edge cases, users labeled a drainage swale near an intersection as *Curb Ramp* (Figure 6b), and *Missing Curb Ramp* where there was no sidewalk present (Figure 6d). For *Obstacle*, *Surface Problem*, and *Missing Sidewalk*, misclassifications typically occurred when a user's label included attributes for a correct label, but there was in fact ample space for wheelchair users to avoid the problem (Figure 6e-j). For false negatives, common sources of errors for *Curb Ramp* and *Missing Curb Ramp* included: (1) When users labeled mid-block crossings, geospatial information for such footpaths/crossings are incomplete in OpenStreetMap, causing inaccuracies when computing the distance to the nearest intersection (Figure 7a-c). (2) The model struggled to correctly classify rare cases such as an exit for a public facility (Figure 7d). For *Obstacle*, and *Surface Problem*, misclassifications happened when the problem could be easily identified without zooming in, thus contradicting the hypothesis that labels placed without zooming in are likely to be incorrect. Similar mistakes were found when the labels lacked inputs of tags, severity, and description—all of which are signals for a diligent crowdworker who typically produces more accurate labels (Figure 7e-h). For *Missing Sidewalk*, misclassifications often occurred when a user's label had a low severity rating, since the absence of sidewalks is supposed to be a high-severity issue (Figure 7i, j).We refrain from further tuning of parameters in LFs post-analysis to prevent overfitting to specific scenarios in testing sets the model failed to learn, thereby preserving the model's generalizability.

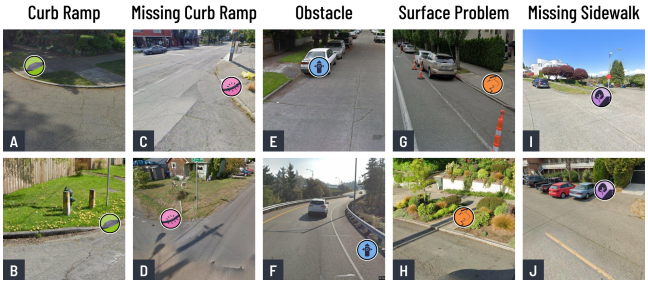


Figure 6: Selected typical inference false positives per label type (the actual label is wrong but was inferred as correct). a, c, failed to differentiate between a *Curb Ramp* and a *Missing Curb Ramp*. b, labeled a drainage swale near an intersection as *Curb Ramp*. d, labeled *Missing Curb Ramp* where there is no sidewalk. e-j, label has attributes for a correct label but there is ample space for a wheelchair user to pass.

4 LABELAID: IMPLEMENTATION & USER EVALUATION

Having demonstrated the technical efficacy of our LabelAId system in inferring label correctness, we implemented the LabelAId inference model in Project Sidewalk, and evaluated the user experience and performance of the end-to-end system with users in the loop. Our study aimed to answer the following questions:

- RQ1: Can LabelAId's feedback improve the ***performance*** of minimally-trained crowdworkers in labeling urban accessibility issues compared to a no feedback condition?
- RQ2: Can LabelAId's feedback enhance minimally-trained crowdworkers' ***self-efficacy*** and ***perceived learning*** when labeling urban accessibility issues compared to a no feedback condition?
- RQ3: How do participants perceive LabelAId's feedback in terms of ***usefulness, content, and frequency***?

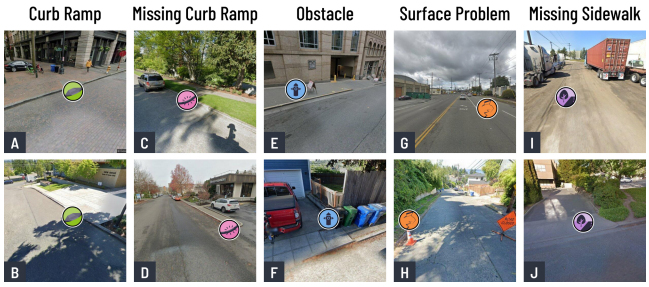


Figure 7: Selected typical inference false negatives per label type (the actual label is correct but was inferred as wrong). a-c, labeled mid-block crossings, geospatial information for such footpaths/crossings are incomplete in *OpenStreetMap*. d, an exit for a public facility. e-h, missing optional inputs. i, j, rated *Missing Sidewalk* with low severity.

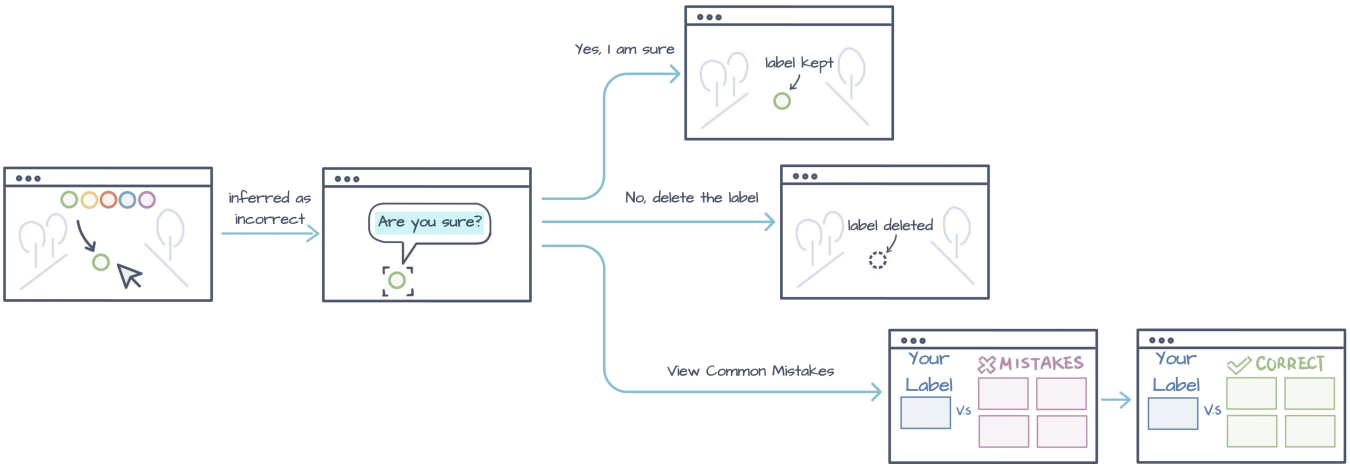


Figure 8: A user flow diagram of LabelAId implemented in Project Sidewalk. (1) A user places a label using the Project Sidewalk interface. (2) If LabelAId detects a mistake, the system displays a just-in-time intervention dialog. (3) The user can choose to keep the label, delete the label, or opt to view common mistakes associated with that label type. From the "View Common Mistakes" page, the user can navigate to the "View Correct Examples" page. See Figure 9 for actual screenshots.

To address these questions, we designed and conducted a between-subjects study of our LabelAId implementation, described below.

4.1 Implementing LabelAId in Project Sidewalk

To incorporate LabelAId into Project Sidewalk, we needed to integrate a real-time mistake inference model (as described in Section 3) and to design and develop a just-in-time UI intervention to help warn users of potential labeling mistakes (using the said inference model). We first highlight design considerations situated in the literature, before describing implementation details.

Design considerations. To design LabelAId's UI intervention, we first reviewed literature regarding the design space for crowd feedback [22, 23, 60, 90, 99] and guidelines for HAI design [1]. Studies have emphasized the importance of timeliness in feedback delivery [23], which led us to opt for real-time feedback, as it delivers feedback during a *teachable moment* when people are still thinking about the task. Additionally, the importance of contextual help for learning assistance has been well-documented in psychology literature [2] and demonstrated through HCI work (*e.g.*, [33, 95]). To further refine the user interface, we consulted best practices for dialog design [64], emphasizing specific response options that clearly outline the consequences of each choice, as well as employing progressive disclosure techniques [63] to help users understand the implications of their actions before committing to them [1]. Based on these insights, we iteratively designed LabelAId, starting with hand sketches and *Figma* mock-ups before implementing the tool in *JavaScript* (front-end) and *Scala* with *QGIS* (back-end).

System implementation. We integrated the city-specific, fine-tuned FT-Transformer into LabelAId using the *Open Neural Network Exchange* (ONNX) runtime standard. An important objective is to reduce latency and facilitate seamless HAI collaboration.The most time consuming step in the preparation stage is to assess whether the label belongs to a pre-existing cluster. To expedite calculation time, we simplified by calculating the spatial haversine distance of the input to a pre-computed cluster centroid, maintaining a threshold consistent with the clustering algorithm at 10 meters. We found in off-line experiments that this approach was 8-20 times faster (speed varies based on label type) and a mere 1.6% of labels (27 out of 1659) had a different clustering result.

We implemented the inference model on the front-end rather than server-side for the following reasons: (1) Latency: considering the small model size (~100 KB), inference can be performed locally in the user's browser, thereby avoiding communication with a remote server and network latency. (2) Privacy: we reduced potential user privacy concerns, as no data is transmitted to a remote server for processing. Notably, during the user study, we found an average preparation time of 1.5 *ms* and an average model inference time of 1.7 *ms* across various hardware and platforms.

User flow. Drawing on previous research on crowdworker feedback [23, 39], HAI [1], and UI design [63], we provide a two-stage intervention. After a user places a label, if LabelAId infers a mistake, we pop-up a just-in-time intervention dialog (Figure 9A) composed of three parts: a mistake title, a rotating set of labeling tips for that label type (*e.g.*, "Do not label driveways as curb ramps."; see Figure 9A), and three buttons: "Yes, I am sure," "No, remove the label" or "View Common Mistakes". Hovering over the "i" icon beside the mistake title will display an explanation that the reminder system is powered by AI and may make mistakes. If the user selects "View Common Mistake", they enter the second stage of customized information about common mistakes and correct examples for that label type. To minimize users' cognitive load [8], both the "View Common Mistakes" and "View Correct Examples" screens present a screen capture of the user's current label alongside three to four example labels, facilitating more straightforward comparison. These example images are curated based on an analysis of frequent mistakes and effective labeling practices on Project Sidewalk. Our user flow (Figure 8) prompts users to reflect on their labeling decisions

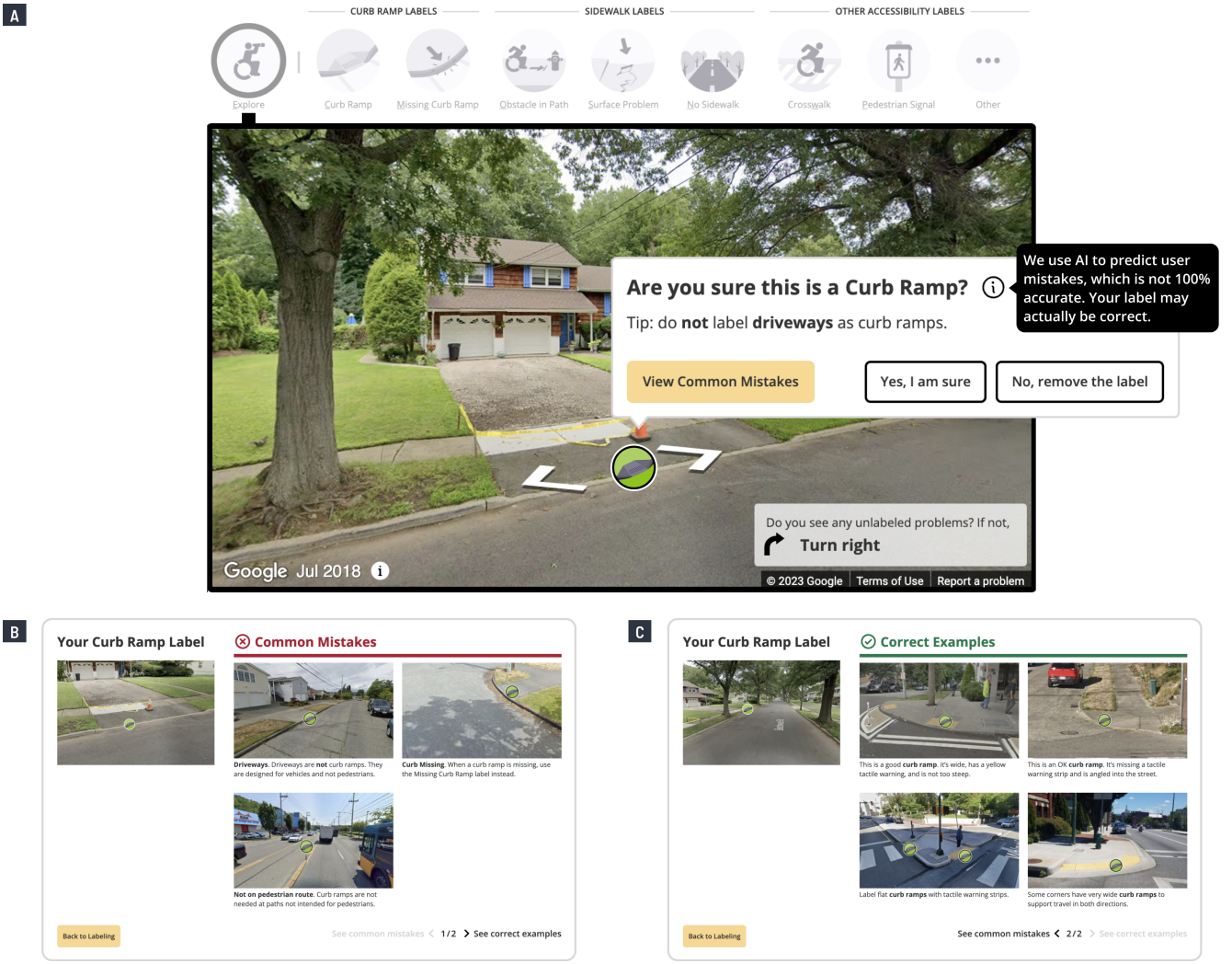


Figure 9: System screenshots of LabelAId implemented in Project Sidewalk. (A) When detecting a user label error: LabelAId pops up a just-in-time intervention dialog composed of three parts: a mistake title, a rotating set of labeling tips for that label type, and three buttons. "Yes, I am sure," "No, remove the label" or "View Common Mistakes". Hovering over the "i" icon will display a note explaining how the reminder is powered by AI and the system may make mistakes. (B) Common Mistakes Page. (C) Correct Examples Page. Both (B) and (C) present a screen capture of the user's current label alongside three to four example labels, facilitating more straightforward comparison.

and then educate them through examples, both of which have been proven to enhance crowdwork quality [23, 99].

4.2 Study Design

To examine our research questions, we conducted a between-subjects study with and without LabelAId. Inspired by previous Project Sidewalk mapathons, the study sessions were conducted in groups via Zoom based on condition. While this setup differs from traditional crowdsourcing studies conducted on platforms like MTurk or Prolific, mapathons and other synchronous social data collection events are key methods for participant involvement in crowdsourced mapping projects like Project Sidewalk and OpenStreetMap[3]. For example, in Project Sidewalk's 18-month deployment in Oradell, NJ, two single-day mapathons contributed over 2,056 labels, accounting for 22% of all labels [57].

Prior to the actual study sessions, we conducted pilot studies with one participant for each condition, during which two researchers observed the participants' labeling behaviors in-person and screen-recorded the process for post-analysis. Based on insights from these pilot studies, we refined the moderation workflow.

For the actual study, two study moderators led six online sessions, three for each condition. Each session had five to seven

[3]https://www.openstreetmap.org/

participants and lasted between 90 and 120 minutes. The sessions were composed of three parts, and the moderator adhered to a script to ensure consistency. First, we provided a brief orientation of urban accessibility and disability, guided the participants through platform account registration, and asked the participants to finish Project Sidewalk's standard ~5-minute interactive tutorial. Second, participants labeled eight curated routes on Project Sidewalk; the routes were carefully chosen by the research team to ensure they included frequent sidewalk accessibility features and problems. Both groups labeled identical routes. Participants were asked to mute themselves during the labeling tasks, and any questions were addressed privately via Zoom chat or in a breakout room. Although the intervention group had access to correct and incorrect examples through the LabelAId UI flow, both groups were shown illustrated tutorial screens in the beginning of each route, which is the standard Project Sidewalk UI (Figure 11). Furthermore, all participants could refer to these examples as well as the *How to Label* section on the platform during labeling (Figure 11), a practice we observed in both groups during the pilot studies. Third, after completing their routes, participants filled out a post-study questionnaire followed by a semi-structured group debriefing session. The debriefing sessions were video and audio recorded. Please see the supplementary materials for our orientation slide deck and pre- and post-study questionnaires.

4.3 Participant Recruitment

For our user study, we recruited participants via university mailing lists and snowball sampling. Our study size of 34 participants was determined through a power analysis using *G*Power* [25], aiming for an effect size of 1 and a statistical power of 0.8. Participants were randomly assigned to either the control or the intervention group depending on their availability. Based on self-reported demographics, we had 21 participants aged 18-24 (12 in the control group), 11 aged 25-34 (5 in the control group), and 2 aged 35-44 (none in the control group); 18 women (10 in the control group), 15 men (6 in the control group), and 1 non-binary individual (1 in the control group). As for computer experience, 2 participants reported having basic skills, 4 had intermediate skills, and 28 considered themselves experts; these numbers were evenly split between the two groups. Before the study session, all participants were required to sign a research consent form and complete a pre-study questionnaire. Each participant was compensated at a rate of $30 per hour for their participation.

4.4 Evaluation Measures

Our study had a dual focus of understanding the objective performance of LabelAId users compared to the baseline as well as to examine their subjective experiences. For our objective measures, we collected and examined:

- **Labeling precision.** The number of correct labels compared to the total number of labels, measuring the correctness of user input.
- **Labeling time.** Time for participants to complete the labeling tasks, recorded per each route.
- **Learning gain in urban accessibility.** We designed quiz questions that were included in both pre- and post-study questionnaires (see supplementary materials). Participants were shown four images for each of the five label types and were asked to select the correct ones. A sum score was calculated for all participants: each correct answer earned 1 point, and each incorrect answer was penalized with -1 point.

We also captured subjective measures through 5-point Likert scale questions:

- **Confidence in response.** *e.g.*, "How confident are you in labeling curb ramps?"
- **Self-efficacy gain.** *e.g.* "I feel more confident about identifying problems on sidewalks faced by people with disabilities."
- **Perceived learning gains in urban accessibility.** *e.g.* "Participating in the study gave me more ideas to make sidewalks accessible for people with disabilities."
- **Perceived usefulness.** *e.g.* "I liked the pop-up prompts."
- **Perceived AI intervention.** "I felt that an AI agent was watching my performance/helping me while I was labeling."

Full list of questions can be found in our supplementary materials.

4.5 Analysis Approach

To analyze our results, two researchers independently validated all participant labels (N=3,574). In cases of disagreements (N=74, IRR=0.98), a third researcher was consulted to reach a consensus. Validations were then used to calculate the precision of user input. For subjective measures captured through Likert scale questions, we mapped responses such as "Strongly disagree" to "Strongly agree" or "Not confident at all" to "Very confident" onto a numerical scale ranging from 1 to 5. We then use descriptive statistics to explore the dataset and to assess the participant performance across different conditions. Due to the between-subjects study and the distribution of the data, we use *Mann-Whitney U* tests to compare label precision, labeling time, and Likert scale responses between the two groups [75]. Additionally, both the debriefing sessions and the post-study questionnaire included open-ended questions to capture nuanced feedback about perceived learning experience, self-efficacy, and overall user experience. Our analysis for these responses focused on summarizing high-level themes. One researcher developed a set of themes through qualitative open coding [15] based on the video transcript and the questionnaire responses, then coded the responses according to the themes. Participant quotes have been slightly modified for concision, grammar, and anonymity.

4.6 Results

During the study, participants contributed a total of 3,574 labels, with 2,091 from the control group and 1,483 from the intervention group. A detailed breakdown of the labels' types and their correctness can be found in Table 6. Our open-encoding process highlighted several key themes, as outlined in Table 5. When asked what helped the participants to label, a majority of intervention participants mentioned the pop-up screens. Regarding labeling confidence, they reported that their confidence varied across different label types and generally increased as they progressed through the tasks. In terms of future improvements, many suggested implementing AI-assisted labeling followed by human verification. Below, we

Labeling confidence	Count	Helpful elements during the labeling process	Count	Future improvement ideas	Count
Confidence varies across different label types	10	Pop-ups	11*	Implementing AI labeling followed by human verification	10
Confident grows with the labeling process	6	Tutorial	6	Providing rationales/confidence levels for the pop-ups	3
High confidence in label type but uncertainty in severity ratings	4	Hover-over images	5	Introducing practice quiz to prefilter participants	2
Unsure of potential missed labels	1			Option to disable AI-generated pop-ups	1

Table 5: During the study's semi-structured group debriefing session, we asked participants (N=34) open-response questions about their confidence levels during the labeling task, what was most helpful during the labeling process, and ideas for future improvements. Participants were not required to answer all questions. We manually coded the participants' responses to identify themes. The count column indicates the number of participants who mentioned each theme. *Note that only the intervention group (N=17) was shown the pop-ups.

delve into an in-depth analysis that integrates both qualitative and quantitative evaluations to address each research question.

4.6.1 Task Performance (RQ1). We first seek to examine whether there are significant differences between groups in task performance and how intervention level correlates with labeling precision within the intervention group.

Labeling precision and task completion time. As summarized in Figure 10, the intervention group demonstrated higher precision overall and across all label types compared to the control group. The Mann-Whitney U results indicate a significant difference in precision between the two groups both overall ($p \leq 0.01$) and for *Curb Ramp* ($p \leq 0.05$) and *Missing Curb Ramp* ($p \leq 0.05$) label types. For route completion time, we found no significant difference between the two groups (p=0.693). The control group had a mean completion time of 2303.3 seconds (SD=1240.3), while the intervention group spent 2801.4 seconds (SD=2035.3). Similarly, no significant differences were observed when examining the time taken for each of the eight routes (p-values ranged from 0.143 to 0.971). These findings indicate that the use of LabelAId resulted in improved labeling precision without compromising labeling speed.

Label Type	Correct		Incorrect		Total	
	C	I	C	I	C	I
Curb Ramp	436	454	487	23	923	477
Missing Curb Ramp	265	245	61	29	326	274
Obstacle	309	298	124	77	433	375
Surface Problem	243	249	55	26	298	275
Missing Sidewalk	94	72	17	10	111	82
Overall	1347	1318	744	165	2091	1483

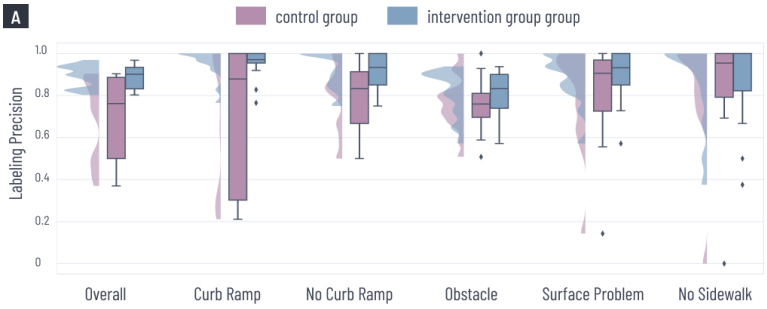
Table 6: Distribution of participants' labels across all label types. *C* stands for Control group and *I* stands for Intervention group.

Labeling precision and level of intervention. While the intervention group clearly performed better, two pertinent questions are: how ***often*** did a LabelAId participant receive a just-in-time AI-assisted prompt and how ***accurately*** did LabelAId perform, *i.e.*, what was the true positive and false positive rate for intervening?

Towards examining the first question: within the intervention group, there were a total of 172 instances where LabelAId intervened with a just-in-time prompt (10.9% of total labels; 10.1 per intervention group participant). When broken down by label type, LabelAId demonstrated high precision in predicting *Curb Ramp* (0.882), *Missing Curb Ramp* (0.750), and *Missing Sidewalk* (1.000) mistakes. However, the model's precision was notably lower for *Obstacle* (0.362) and *Surface Problem* (0.377). Upon closer examination, we found that these less accurate inferences often corresponded with user behaviors that are likely to result in incorrect labels, such as not zooming in or failing to provide severity ratings or tags.

Within the 17 participants in the intervention group, our analysis revealed no significant correlation between the frequency of interventions by LabelAId and participants' labeling precision, either overall or for specific label types. Similarly, the number of times participants viewed common mistakes or correct examples UI screens did not correlate with their labeling accuracy (Table 10). We will return to this point in Section 5.

Despite the relatively low view frequency of the "Common Mistakes" UI screens (24 views in total, 1.4 views per person) and correct examples (6 views in total, 0.4 per person), qualitative feedback indicated their usefulness for those who chose to engage with them. During the debriefing sessions, several participants cited these screens when asked about what helped them during the labeling tasks. For instance, one participant noted a shift in their labeling approach after viewing the AI-triggered common mistake screen, stating, *"Midway through, I saw the common mistakes, and it totally shifted my perspective. I had been labeling driveways from*



B

Label Type	Control	Intervention	U	p-value
Overall	0.699 (±0.199)	0.891 (±0.053)	50.0	0.001 **
Curb Ramp	0.686 (±0.346)	0.956 (±0.067)	70.0	0.038 *
No Curb Ramp	0.802 (±0.164)	0.918 (±0.091)	80.5	0.025 *
Obstacle	0.7610 (±0.126)	0.812 (±0.111)	85.5	0.183
Surface Problem	0.812 (±0.230)	0.894 (±0.116)	100.0	0.423
No Sidewalk	0.842 (±0.267)	0.867 (±0.208)	66.5	0.480

precision low — precision high

Figure 10: User labeling precision in the intervention group was higher across all categories, and the difference was statistically significant for the overall category, as well as for *Curb Ramp* and *Missing Curb Ramp* label types ($*p \leq 0.05$, $**p \leq 0.01$). (A) A raincloud plot (a half violin plot and a boxplot) shows user labeling precision between the control group and the intervention group, both overall and for the five specific label types. (B) A complementary table displays the precision mean, standard deviation, Mann-Whitney U value, and p-value for both the control and intervention groups.

houses, but the screen clarified that those should not be labeled as curb cuts."

4.6.2 Self-efficacy & Learning Gains (RQ2). While the above findings demonstrate users' improvements in terms of task performance, we are also interested in self-efficacy and learning.

Self-efficacy. In the post-study questionnaire, we asked all participants about their confidence in identifying sidewalk features or problems. On average, participants rated their self-confidence higher in the intervention group (Avg=4.47; SD=0.88) than the control group (Avg=4.53; SD=0.52) with a statistically significant difference for *Missing Curb Ramps* (Avg=4.6; SD=0.7 vs. Avg=3.8; SD=0.9, $p \leq 0.05$), as shown in Table 12. However, when participants were asked if they felt more confident about identifying problems on sidewalks faced by people with disabilities, the difference between groups was not statistically significant (p=0.721, see Q5 in Table 13).

Perceived learning gains. While task performance serves as one indicator of learning outcomes, we also used quizzes to assess objective learning gains and Likert scale questions to measure perceived learning gains. For objective learning gains, the mean improvement between the pre- and post-study quizzes was 1.35 (SD=1.73) for the control group and 1.31 (SD=1.54) for the intervention group, showing only a minor difference between the two. In terms of perceived learning gains, both groups demonstrated an enhanced understanding of curb ramps and accessibility challenges. Although the means were higher for the intervention group across all questions, no statistically significant difference was observed, except for the question, *"Participating in the study gave me more ideas to make sidewalks accessible for people with disabilities."*, where the mean score for the control group was 4.35 (SD=0.7), compared to 4.82 (SD=0.53) for the intervention group ($p \leq 0.05$).

4.6.3 Perceived Usefulness & Presence of AI (RQ3). Having explored the overall user performance, confidence and learning gain, we now turn to the perceived usefulness and presence of AI in LabelAId.

Perceived usefulness. Participants generally expressed a favorable view of LabelAId. When asked to what extent they agreed with the statements that the pop-up prompts were helpful and likable, the majority responded with "Somewhat Agree" or "Strongly Agree" (82.35% and 64.7%, respectively). In the post-study questionnaire and debriefing sessions, 11 out of 17 participants in the intervention group specifically cited the pop-up screens from LabelAId as a feature they appreciated or found helpful for labeling tasks. These timely reminders were particularly valued when participants were uncertain about their initial judgments. One participant mentioned: *"There were times when I was not sure if I should label it, and the system popped-up for me and said 'Are you sure about this?' I found that really helpful."* When asked about whether the prompts were distracting or appeared too frequently, the responses were more mixed—with a relatively even distribution across Likert responses.

Perceived presence of AI. We asked participants whether they felt an AI agent was observing their performance or assisting them during the labeling task and found a statistically significant differences between the two groups. This suggests that the presence of LabelAId had a noticeable impact on participants' perception of AI involvement. Interestingly, some participants in the control group explicitly expressed a desire for AI assistance. One control-group participant mentioned, *"There was a section [in the post-study questionnaire] asking how I felt about AI helping me to label. Honestly, I didn't notice any AI while I was labeling. It would be super convenient if there was one that could suggest labels and ask me to correct them or provide a confidence level."* This is exactly the intent of LabelAId.

5 DISCUSSION

Through our technical evaluation and user study, we showed how LabelAId improves both labeling data quality and crowdworkers' domain knowledge. We now situate our findings in related work, highlight key factors behind LabelAId's success, its limitations, and directions for future research. We also discuss how LabelAId can be generalized to other domains of crowdsourced science.

5.1 Reflecting on LabelAId's Performance

Below, we reflect on LabelAI'd performance and its relevance to future research, including comparing the differences between AI and human feedback, minimizing the overreliance on AI, and striking a balance between constructive feedback and perceived surveillance.

Can AI-assistance replicate human-based feedback? Prior work has shown that providing manual feedback to crowdworkers can improve task performance and enhance self-efficacy [22, 23, 60, 90, 99]. Our study further reveals that AI-feedback can improve labeling performance, increase participants' confidence, and enhance their domain knowledge—even with an imperfect ML inference model. While the nuances between human and AI-feedback in crowdsourcing have yet to be comprehensively studied, researchers in education have assessed the usage of automatic feedback as a learning tool [34, 38, 55, 92]. Findings suggest that automatic feedback can reduce bias and increase consistency in grading [38], liberate the instructor from grading to focus on other tasks [92], and allow more students to receive education simultaneously [85]. We believe that these benefits can well be extended to AI-generated feedback in crowdsourcing systems.

Yet, automated feedback in education contexts has limitations. It excels in grading tasks with clear-cut solutions (*e.g.* programming questions), but may be challenging to implement in more subjective disciplines [34]. Moreover, automatic graders fail to recognize when students are very close to meeting the criteria, whereas human graders would identify and assign partial grades accordingly [55]. Future research in crowdsourcing should incorporate these insights from education science when designing AI-based feedback systems, and borrow approaches such as AI-feedback combined with human feedback on request [55].

Cognitive forcing function reduces overreliance on AI. An overarching concern with AI-based assistance—including systems like LabelAId—is how the presence and behavior of AI may actually *reduce* active cognitive functioning in humans as they defer to AI's recommendations, which can then negatively impact overall task performance [42, 52]. For example, [9, 42] showed how users tend to overly depend on AI, following its suggestions even when their own judgment might be superior. Such a tendency is particularly problematic when the AI is inconsistent (*e.g.,* across class categories), as in our case. Recent work has explored *cognitive forcing functions* [10]—functions that elicit thinking at decision-making time. Because there is an *anchoring bias* [32] that occurs when presenting users with AI's recommendations, one effective strategy is to ask the user to make a decision prior to seeing the AI's recommendation [10]. Indeed, this is how LabelAId works: presenting suggestions only *after* the user makes an initial decision and places a label—which may mitigate such bias.

Specifically, in our user study, LabelAId performed particularly poorly for two label types *Obstacles* and *Surface Problems* with false positive feedback rates of 36.2% and 37.7% respectively. However, users rejected these suggestions 83% and 73% of the time, indicating that they preferred their own judgments to the AI. Although this design choice was dictated by LabelAI's model requirements, it encouraged analytical thinking that boosted participants' confidence in their own decisions. Our study contributes to the broader discourse of HAI, highlighting how system design can elicit analytical reasoning and reduce cognitive biases in decision-making.

Striking a balance between constructive feedback and perceived surveillance. We found a significant difference between the two groups regarding the perceived presence of AI (Section 4.6.3). Out of the 17 participants in the intervention group, eight felt observed and nine felt assisted by an AI agent, while in the control group, none felt observed and only three sensed AI assistance. We speculate that this difference in perceived surveillance also contributed to better intervention group performance, since they felt their work was being scrutinized. This observation raises questions regarding AI agents as a form of surveillance in crowdsourcing environments. When scholars apply a Foucauldian lens [26] to monitoring technology, some see AI monitoring as social control from existing power hierarchies [12], while others argue it can both restrict and empower individuals [50]. This dichotomy implies that, if well-implemented, AI can encourage self-regulation among crowdworkers. A recent study confirms that digital feedback improves crowdwork outcomes when learning is the primary objective [94], which is often the case in crowdsourcing in community science. Therefore, we advocate for crowdsourcing platforms where the AI system strikes a balance between constructive feedback and perceived surveillance.

5.2 LabelAId Limitations and Future Research

We now reflect on LabelAId's limitations and future work, focusing on designing interactions with imperfect ML models, promoting user agency in mixed-initiative interfaces, improving interaction efficiency in providing learning aids, and expanding participant diversity in future research.

Designing interactions with imperfect ML models. With LabelAId, we were able to determine when a user likely made a mistake, but not the exact source of the error, which limited the types of prompting we could provide. As one participant mentioned: *"It'll be great to provide some rationale or explanation on why there's a pop-up. Like maybe the location I placed my label is too far away from the obstacle."* Current approaches of offering AI explainability falls into two categories: communicating information about the model inferences on a local level (*e.g.* confidence score and local feature importance) and communicating information about the model itself on a global level (*e.g.* model accuracy and global explanations) [51]. However, LabelAId's current implementation does not incorporate explainability features.

On a global level, we recognize that our implementation could better communicate the model's varying accuracy levels across different label types. Despite a detailed technical analysis of LabelAId's performance in Section 3, we did not surface accuracy scores or global feature importance to participants. Future iterations should address this shortcoming. On a local level, we intentionally excluded confidence scores. This choice was informed by research indicating that confidence scores have limited impact on improving HAI collaboration [3, 10], coupled with our concern about overcluttering the already busy UI. Future work may incorporate recent approaches to model the user's level of confidence and provide adaptive recommendations, *i.e.,* only display AI's recommendations when the AI's confidence level is higher than the human's [59].

In summary, while our current design decisions were informed by a balance of user cognitive load considerations and technical constraints, future work should explore other methods to provide users with tailored explanations and rationale, enhancing their understanding and interaction with the ML model.

Promoting user agency in mixed-initiative interfaces. Participants had mixed opinions about the frequency of AI interventions,

with some finding them distracting. One participant noted, "*Sometimes the pop-ups were too frequent, so it might be helpful to give the user the option to disable them.*" In addition, we also noticed the diminishing returns of increased intervention. During the study, there is no significant correlation between the frequency of intervention and task performance (Table 10). One potential explanation is that users understood their mistakes after the first few interventions, thereby making fewer mistakes in subsequent tasks. These findings, consistent with learning science research demonstrating that additional exposure or intervention does not necessarily improve performance (known as the *saturation effect* [37]), are also supported by ongoing HAI research exploring ways to enhance human agency in mixed-initiative interfaces [1, 51, 82]. In future iterations, we would like to explore offering users overall control to enable or disable AI, to provide adaptive suggestion frequency based on labeling rate, and to allow users to request AI assistance only when needed [10].

Designing efficient UI for learning aids. In addition to a lack of correlation between how often participants viewed example screens and their performance levels (Table 10), we observed that common mistakes and correct examples were only viewed a total of 30 times–six of the 17 intervention participants never viewed either of the screens. This could be due to the *interaction cost* [7]: the common mistakes screen requires two clicks and the correct examples screen three. While click count alone is not a meaningful metric [53], it is important to minimize interaction costs [7] by making key information easily accessible. Future work should explore developing effective methods for presenting examples to crowdworkers while they are balancing high cognitive load tasks.

Expanding participant diversity in future research. While our study size of 34 aligns with typical HCI between-subjects studies (*e.g.*, [43, 67]), it is on the lower end for crowdsourcing research [48]. However, our study design choice facilitated in-depth interviews and focused analysis, allowing us to gather qualitative insights not typical in crowdsourcing studies. Participants were recruited through snowball sampling from the research team's contacts and university mailing lists, which may not represent the comprehensive user base of Project Sidewalk including disability advocates. In future studies, we aim to enhance the applicability of our findings by expanding our participant base.

5.3 Generalizability to Other Domains

Our study demonstrates the effectiveness of LabelAId in a crowdsourcing tool for urban accessibility, yet, its generalizability remains an open question. We believe there are two primary generalizable components:

- **LabelAId's PWS based ML pipeline.** PWS does not require annotated data, it works on a set of LFs generalized from domain knowledge and user behavior. This is particularly useful for crowdsourced community science because it allows organizers to transform their expertise and heuristic into LFs, which can then programmatically label large quantities of data. It is also more cost-effective compared to traditional ML models, as LabelAId improves inference accuracy by 36.7% with only 50 downstream data points.
- **LabelAId's mistake intervention design.** LabelAId's *in-situ* intervention design is rooted in literature on crowd feedback and contextual assistance, and aligns with recent HAI research on using cognitive theories to reduce over reliance on AI. Its simple two-step formula can be easily replicated in other platforms.

We believe our technique is most applicable to areas that require domain expertise and contextual understanding, such as medical image labeling [79, 98], galaxy classification [84], and wildlife categorization [4]. For example, the crowdsourcing application *iNaturalist* uses identification technology and taxonomic experts to assist people in identifying natural species, and it achieves the best results when combined with traditional field guides [88]. We envision these guides and knowledge from experts being translated into LFs in our pipeline, and with similar mistake intervention design, LabelAId can help iNaturalist users contribute data more effectively while learning more about biodiversity.

6 CONCLUSION

In conclusion, LabelAId offers a practical approach to improving both crowdsourced data quality and domain knowledge of crowdworkers. By using machine learning to provide real-time feedback, LabelAId reduces the need for extensive manual review while also helping workers learn throughout the crowdsourcing process. Our user study demonstrates that LabelAId can improve user label quality without sacrificing speed, thereby offering a scalable solution to enhance worker knowledge and label quality in crowdsourcing tasks. While our empirical results focused on the performance of LabelAId within the context of urban accessibility, our framework can be extended to other crowdsourcing platforms, such as agricultural image recognition, medical imagery labeling, and wildlife biology image categorization.

ACKNOWLEDGMENTS

We thank the reviewers for their insightful comments, our study participants, academic writing advisor Sandy Kaplan, and the Allen School Computer Science Laboratory Group. This work was supported by NSF SCC-IRG #212508, PacTrans, an Amazon Research Award, and a Google Research Scholar. Zhihan Zhang is supported by the University of Washington CEI Fellowship.

REFERENCES

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–13. https://doi.org/10.1145/3290605.3300233

[2] John Robert Anderson, C. Franklin Boyle, Robert Farrell, and Brian J. Reiser. 1984. *Cognitive principles in the design of computer tutors.* Department of Psychology, Carnegie-Mellon University.

[3] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3411764.3445717

[4] Tanya Y. Berger-Wolf, Daniel I. Rubenstein, Charles V. Stewart, Jason A. Holmberg, Jason Parham, Sreejith Menon, Jonathan Crall, Jon Van Oast, Emre Kiciman, and Lucas Joppa. 2017. Wildbook: Crowdsourcing, computer vision, and data science for conservation. *arXiv preprint arXiv:1710.08880* (2017).

[5] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, New York New York USA, 313–322. https://doi.org/10.1145/1866029.1866078

[6] Eran Bringer, Abraham Israeli, Yoav Shoham, Alex Ratner, and Christopher Ré. 2019. Osprey: Weak Supervision of Imbalanced Extraction Problems without Code. In *Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning*. ACM, Amsterdam Netherlands, 1–11. https://doi.org/10.1145/3329486.3329492

[7] Raluca Budiu. 2013. Interaction cost. https://www.nngroup.com/articles/interaction-cost-definition/

[8] Raluca Budiu. 2018. Working Memory and External Memory. https://www.nngroup.com/articles/working-memory-external-memory/

[9] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. ACM, Cagliari Italy, 454–464. https://doi.org/10.1145/3377325.3377498

[10] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 1–21. https://doi.org/10.1145/3449287

[11] Chris Callison-Burch and Mark Dredze. 2010. Creating Speech and Language Data With Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, Los Angeles, 1–12. https://aclanthology.org/W10-0701

[12] Marta B. Calás and Linda Smircich. 1999. Past Postmodernism? Reflections and Tentative Directions. *The Academy of Management Review* 24, 4 (1999), 649–671. https://doi.org/10.2307/259347 Publisher: Academy of Management.

[13] Dana Chandler and Adam Kapelner. 2013. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization* 90 (June 2013), 123–133. https://doi.org/10.1016/j.jebo.2013.03.003

[14] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2334–2346. https://doi.org/10.1145/3025453.3026044

[15] Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. SAGE, Los Angeles London.

[16] Quanze Chen, Jonathan Bragg, Lydia B. Chilton, and Dan S. Weld. 2019. Cicero: Multi-Turn, Contextual Argumentation for Accurate Crowdsourcing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–14. https://doi.org/10.1145/3290605.3300761

[17] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794. https://doi.org/10.1145/2939672.2939785 arXiv:1603.02754 [cs].

[18] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, and Foldit Players. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307 (Aug. 2010), 756–760. https://doi.org/10.1038/nature09304 Number: 7307 Publisher: Nature Publishing Group.

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848 ISSN: 1063-6919.

[20] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. 2013. Pick-a-crowd: tell me what you like, and i'll tell you what to do. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, Rio de Janeiro Brazil, 367–374. https://doi.org/10.1145/2488388.2488421

[21] Mira Dontcheva, Robert R. Morris, Joel R. Brandt, and Elizabeth M. Gerber. 2014. Combining crowdsourcing and learning to improve engagement and performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3379–3388. https://doi.org/10.1145/2556288.2557217

[22] Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. 2016. Toward a Learning Science for Complex Crowdsourcing Tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2623–2634. https://doi.org/10.1145/2858036.2858268

[23] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, Seattle Washington USA, 1013–1022. https://doi.org/10.1145/2145204.2145355

[24] Julie S. Downs, Mandy B. Holbrook, Steve Sheng, and Lorrie Faith Cranor. 2010. Are your participants gaming the system?: screening mechanical turk workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Atlanta Georgia USA, 2399–2402. https://doi.org/10.1145/1753326.1753688

[25] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191. ISBN: 1554-351X Publisher: Springer.

[26] Michel Foucault. 1920. *Discipline & Punish*. Random House, New York.

[27] Jonathan Fürst. 2020. Applying Weak Supervision to Mobile Sensor Data: Experiences with TransportMode Detection. In *Jonathan Fürst*. http://www.jofu.org/publication/furst-2020-transport/

[28] Ujwal Gadiraju, Gianluca Demartini, Ricardo Kawase, and Stefan Dietze. 2019. Crowd Anatomy Beyond the Good and Bad: Behavioral Traces for Crowd Worker Modeling and Pre-selection. *Computer Supported Cooperative Work (CSCW)* 28, 5 (Sept. 2019), 815–841. https://doi.org/10.1007/s10606-018-9336-y

[29] Sambuddha Ghosal, Bangyou Zheng, Scott C. Chapman, Andries B. Potgieter, David R. Jordan, Xuemin Wang, Asheesh K. Singh, Arti Singh, Masayuki Hirafuji, Seishi Ninomiya, Baskar Ganapathysubramanian, Soumik Sarkar, and Wei Guo. 2019. A Weakly Supervised Deep Learning Framework for Sorghum Head Detection and Counting. *Plant Phenomics* 2019 (June 2019). https://doi.org/10.34133/2019/1525874 Publisher: American Association for the Advancement of Science.

[30] Google. 2022. Open Images V7. https://storage.googleapis.com/openimages/web/index.html

[31] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. Revisiting Deep Learning Models for Tabular Data. In *Advances in Neural Information Processing Systems*, Vol. 34. Curran Associates, Inc., 18932–18943. https://proceedings.neurips.cc/paper_files/paper/2021/hash/9d86d83f925f2149e9edb0ac3b49229c-Abstract.html

[32] Ben Green and Yiling Chen. 2019. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 50:1–50:24. https://doi.org/10.1145/3359152

[33] Tovi Grossman and George Fitzmaurice. 2010. ToolClips: an investigation of contextual video assistance for functionality understanding. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Atlanta Georgia USA, 1515–1524. https://doi.org/10.1145/1753326.1753552

[34] Marcelo Guerra Hahn, Silvia Margarita Baldiris Navarro, Luis De La Fuente Valentin, and Daniel Burgos. 2021. A Systematic Review of the Effects of Automatic Scoring and Automatic Feedback in Educational Settings. *IEEE Access* 9 (2021), 108190–108198. https://doi.org/10.1109/ACCESS.2021.3100890

[35] Muki Haklay. 2013. Citizen Science and Volunteered Geographic Information: Overview and Typology of Participation. In *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*, Daniel Sui, Sarah Elwood, and Michael Goodchild (Eds.). Springer Netherlands, Dordrecht, 105–122. https://doi.org/10.1007/978-94-007-4587-2_7

[36] Derek L. Hansen, Patrick J. Schone, Douglas Corey, Matthew Reid, and Jake Gehring. 2013. Quality control mechanisms for crowdsourcing: peer review, arbitration, & expertise at familysearch indexing. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, San Antonio Texas USA, 649–660. https://doi.org/10.1145/2441776.2441848

[37] B. Hauptmann and A. Karni. 2002. From primed to learn: the saturation of repetition priming and the induction of long-term memory. *Cognitive Brain Research* 13, 3 (May 2002), 313–322. https://doi.org/10.1016/S0926-6410(01)00124-0

[38] Emlyn Hegarty-Kelly and Dr Aidan Mooney. 2021. Analysis of an automatic grading system within first year Computer Science programming modules. In *Computing Education Practice 2021*. ACM, Durham United Kingdom, 17–20. https://doi.org/10.1145/3437914.3437973

[39] Danula Hettiachchi, Mike Schaekermann, Tristan J. McKinney, and Matthew Lease. 2021. The Challenge of Variable Effort Crowdsourcing and How Visible Gold Can Help. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–26. https://doi.org/10.1145/3476073

[40] Catherine Hoffman, Caren B. Cooper, Eric B. Kennedy, Mahmud Farooque, Darlene Cavalier, Catherine Hoffman, Caren B. Cooper, Eric B. Kennedy, Mahmud Farooque, and Darlene Cavalier. 2001. SciStarter 2.0: A Digital Platform to Foster and Study Sustained Engagement in Citizen Science. https://www.igi-global.com/gateway/chapter/www.igi-global.com/gateway/chapter/170184 Archive Location: scistarter-20 ISBN: 9781522509622 Publisher: IGI Global.

[41] Panagiotis G. Ipeirotis. 2010. Analyzing the Amazon Mechanical Turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students* 17, 2 (Dec. 2010), 16–21. https://doi.org/10.1145/1869086.1869094

[42] Maia Jacobs, Melanie F. Pradier, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational Psychiatry* 11, 1 (Feb. 2021), 1–9. https://doi.org/10.1038/s41398-021-01224-x Number: 1 Publisher: Nature Publishing Group.

[43] Sungchul Jung, Nawam Karki, Max Slutter, and Robert W. Lindeman. 2021. On the Use of Multi-sensory Cues in Symmetric and Asymmetric Shared Collaborative Virtual Spaces. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 72:1–72:25. https://doi.org/10.1145/3449146

[44] Ece Kamar, Severin Hacker, and Eric Horvitz. 2012. Combining human and machine intelligence in large-scale crowdsourcing.. In *AAMAS*, Vol. 12. 467–474.

[45] Gabriella Kazai and Imed Zitouni. 2016. Quality Management in Crowdsourcing using Gold Judges Behavior. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, San Francisco California USA, 267–276. https://doi.org/10.1145/2835776.2835835

[46] Ashley Rose Kelly and Kate Maddalena. 2015. Harnessing Agency for Efficacy: "Foldit" and Citizen Science. *Poroi* 11, 1 (May 2015). https://doi.org/10.13008/2151-2957.1184 Number: 1 Publisher: The Project on Rhetoric of Inquiry (POROI).

[47] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Florence Italy, 453–456. https://doi.org/10.1145/1357054.1357127

[48] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, San Antonio Texas USA, 1301–1318. https://doi.org/10.1145/2441776.2441923

[49] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E. Kraut. 2011. CrowdForge: crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, Santa Barbara California USA, 43–52. https://doi.org/10.1145/2047196.2047202

[50] Dany Lacombe. 1996. Reforming Foucault: A Critique of the Social Control Thesis. *The British Journal of Sociology* 47, 2 (1996), 332–352. https://doi.org/10.2307/591730 Publisher: [Wiley, London School of Economics and Political Science, London School of Economics].

[51] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 2023. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1369–1385. https://doi.org/10.1145/3593013.3594087

[52] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, Atlanta GA USA, 29–38. https://doi.org/10.1145/3287560.3287590

[53] Page Laubheimer. 2019. The 3-Click Rule for Navigation Is False. https://www.nngroup.com/articles/3-click-rule/

[54] Doris Jung-Lin Lee, Joanne Lo, Moonhyok Kim, and Eric Paulos. 2016. Crowdclass: Designing Classification-Based Citizen Science Learning Modules. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 4 (Sept. 2016), 109–118. https://doi.org/10.1609/hcomp.v4i1.13273

[55] Abe Leite and Saúl A. Blanco. 2020. Effects of Human vs. Automatic Feedback on Students' Understanding of AI Concepts and Programming Style. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*. ACM, Portland OR USA, 44–50. https://doi.org/10.1145/3328778.3366921

[56] Roman Levin, Valeriia Cherepanova, Avi Schwarzschild, Arpit Bansal, C. Bayan Bruss, Tom Goldstein, Andrew Gordon Wilson, and Micah Goldblum. 2023. Transfer Learning with Deep Tabular Models. https://doi.org/10.48550/arXiv.2206.15306 arXiv:2206.15306 [cs, stat].

[57] Chu Li, Katrina Ma, Michael Saugstad, Kie Fujii, Molly Delaney, Yochai Eisenberg, Delphine Labbé, Judy Shanley, Devon Snyder, Florian P Thomas, and Jon E Froehlich. 2024. "I never realized sidewalks were a big deal": A Case Study of a Community-Driven Sidewalk Accessibility Assessment using Project Sidewalk. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI'24)*. ACM, New York, NY, USA. https://doi.org/10.1145/3613904.3642003

[58] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common Objects in Context. http://arxiv.org/abs/1405.0312 arXiv:1405.0312 [cs].

[59] Yongfeng Ma, Shaojie Mo, Shuyan Chen, Guanyang Xing, Kun Tang, Jiguang Zhao, and Zhaoyan Guo. 2023. Evaluating the effectiveness of accessibility features for roadway users with visual impairment: A case study for Nanjing, China. *Transportation Research Part F: Traffic Psychology and Behaviour* 97 (Aug. 2023), 301–313. https://doi.org/10.1016/j.trf.2023.07.021

[60] Lena Mamykina, Thomas N. Smyth, Jill P. Dimond, and Krzysztof Z. Gajos. 2016. Learning From the Crowd: Observational Learning in Crowdsourcing Communities. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 2635–2644. https://doi.org/10.1145/2858036.2858560

[61] Masaki Matsubara, Masaki Kobayashi, and Atsuyuki Morishima. 2018. A Learning Effect by Presenting Machine Prediction as a Reference Answer in Self-correction. In *2018 IEEE International Conference on Big Data (Big Data)*. 3522–3528. https://doi.org/10.1109/BigData.2018.8622435

[62] Takumi Nakayama, Masaki Matsubara, and Atsuyuki Morishima. 2021. Crowd-Worker Skill Improvement with AI Co-Learners. In *Proceedings of the 9th International Conference on Human-Agent Interaction (HAI '21)*. Association for Computing Machinery, New York, NY, USA, 316–322. https://doi.org/10.1145/3472307.3484684

[63] Jakob Nielsen. 2006. Progressive Disclosure. https://www.nngroup.com/articles/progressive-disclosure/

[64] Jakob Nielsen. 2018. Confirmation Dialogs Can Prevent User Errors (If Not Overused). https://www.nngroup.com/articles/confirmation-dialog/

[65] Jon Noronha, Eric Hysen, Haoqi Zhang, and Krzysztof Z. Gajos. 2011. Platemate: crowdsourcing nutritional analysis from food photographs. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, Santa Barbara California USA, 1–12. https://doi.org/10.1145/2047196.2047198

[66] U.S. Department of Transportation (DOT). 2023. Curb Ramps. https://safety.fhwa.dot.gov/saferjourney1/library/countermeasures/03.htm

[67] Srishti Palani, Yingyi Zhou, Sheldon Zhu, and Steven P. Dow. 2022. InterWeave: Presenting Search Suggestions in Context Scaffolds Information Search and Synthesis. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3526113.3545696

[68] Tina Phillips, Norman Porticella, Mark Constas, and Rick Bonney. 2018. A framework for articulating and measuring individual learning outcomes from participation in citizen science. *Citizen Science: Theory and Practice* 3, 2 (2018). ISBN: 2057-4991 Publisher: Ubiquity Press.

[69] M. Jordan Raddick, Georgia Bracey, Pamela L. Gay, Chris J. Lintott, Carie Cardamone, Phil Murray, Kevin Schawinski, Alexander S. Szalay, and Jan Vandenberg. 2013. Galaxy Zoo: Motivations of Citizen Scientists. http://arxiv.org/abs/1303.6886 arXiv:1303.6886 [astro-ph, physics:physics].

[70] Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment* 11, 3 (Nov. 2017), 269–282. https://doi.org/10.14778/3157794.3157797

[71] Alexander Ratner, Braden Hancock, Jared Dunnmon, Frederic Sala, Shreyash Pandey, and Christopher Ré. 2019. Training Complex Models with Multi-Task Weak Supervision. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (July 2019), 4763–4771. https://doi.org/10.1609/aaai.v33i01.33014763 Number: 01.

[72] Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data Programming: Creating Large Training Sets, Quickly. In *Advances in Neural Information Processing Systems*, Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/hash/6709e8d64a5f47269ed5cea9f625f7ab-Abstract.html

[73] Jason Reed, M. Jordan Raddick, Andrea Lardner, and Karen Carney. 2013. An Exploratory Factor Analysis of Motivations for Participating in Zooniverse, a Collection of Virtual Citizen Science Projects. In *2013 46th Hawaii International Conference on System Sciences*. 610–619. https://doi.org/10.1109/HICSS.2013.85 ISSN: 1530-1605.

[74] Neal Reeves, Ramine Tinati, Sergej Zerr, Max G. Van Kleek, and Elena Simperl. 2017. From Crowd to Community: A Survey of Online Community Features in Citizen Science Projects. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Portland Oregon USA, 2137–2152. https://doi.org/10.1145/2998181.2998302

[75] Judy Robertson and Maurits Kaptein (Eds.). 2016. *Modern Statistical Methods for HCI*. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-26633-6

[76] Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. 2011. An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets. *Proceedings of the International AAAI Conference on Web and Social Media* 5, 1 (2011), 321–328. https://doi.org/10.1609/icwsm.v5i1.14105 Number: 1.

[77] Jeffrey Rzeszotarski and Aniket Kittur. 2012. CrowdScape: interactively visualizing user behavior and output. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. ACM, Cambridge Massachusetts USA, 55–62. https://doi.org/10.1145/2380116.2380125

[78] Jeffrey M. Rzeszotarski and Aniket Kittur. 2011. Instrumenting the crowd: using implicit behavioral measures to predict task performance. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, Santa Barbara California USA, 13–22. https://doi.org/10.1145/2047196.2047199

[79] Tim Rädsch, Annika Reinke, Vivienn Weru, Minu D. Tizabi, Nicholas Schreck, A. Emre Kavur, Bünyamin Pekdemir, Tobias Roß, Annette Kopp-Schneider, and Lena Maier-Hein. 2023. Labelling instructions matter in biomedical image analysis. *Nature Machine Intelligence* 5, 3 (March 2023), 273–283. https://doi.org/10.1038/s42256-023-00625-5 Number: 3 Publisher: Nature Publishing Group.

[80] Manaswi Saha, Michael Saugstad, Hanuma Teja Maddali, Aileen Zeng, Ryan Holland, Steven Bower, Aditya Dash, Sage Chen, Anthony Li, Kotaro Hara, and Jon Froehlich. 2019. Project Sidewalk: A Web-based Crowdsourcing Tool for Collecting Sidewalk Accessibility Data At Scale. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300292

[81] Aaron D. Shaw, John J. Horton, and Daniel L. Chen. 2011. Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. ACM, Hangzhou China, 275–284. https://doi.org/10.1145/1958824.1958865

[82] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (April 2020), 495–504. https://doi.org/10.1080/10447318.2020.1741118

[83] Kai Shu, Subhabrata Mukherjee, Guoqing Zheng, Ahmed Hassan Awadallah, Milad Shokouhi, and Susan Dumais. 2020. Learning with Weak Supervision for Email Intent Detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1051–1060. https://doi.org/10.1145/3397271.3401121 arXiv:2005.13084 [cs].

[84] Robert Simpson, Kevin R. Page, and David De Roure. 2014. Zooniverse: observing the world's largest citizen science platform. In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, Seoul Korea, 1049–1054. https://doi.org/10.1145/2567948.2579215

[85] Lauren Singelmann, Ellen Swartz, Mary Pearson, Ryan Striker, and Enrique Alvarez Vazquez. 2019. Design and Development of a Machine Learning Tool for an Innovation-Based Learning MOOC. In *2019 IEEE Learning With MOOCS (LWMOOCS)*. 105–109. https://doi.org/10.1109/LWMOOCS47620.2019.8939621

[86] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Honolulu, Hawaii, 254–263. https://aclanthology.org/D08-1027

[87] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. 2021. MLP-Mixer: An all-MLP Architecture for Vision. In *Advances in Neural Information Processing Systems*, Vol. 34. Curran Associates, Inc., 24261–24272. https://proceedings.neurips.cc/paper/2021/hash/cba0a4ee5ccd02fda0fe3f9a3e7b89fe-Abstract.html

[88] Shem Unger, Mark Rollins, Allison Tietz, and Hailey Dumais. 2021. iNaturalist as an engaging tool for identifying organisms in outdoor activities. *Journal of Biological Education* 55, 5 (Oct. 2021), 537–547. https://doi.org/10.1080/00219266.2020.1739114 Publisher: Routledge _eprint: https://doi.org/10.1080/00219266.2020.1739114.

[89] Paroma Varma, Frederic Sala, Shiori Sagawa, Jason Fries, Daniel Fu, Saelig Khattar, Ashwini Ramamoorthy, Ke Xiao, Kayvon Fatahalian, James Priest, and Christopher Ré. 2019. Multi-Resolution Weak Supervision for Sequential Data. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/hash/93db85ed909c13838ff95ccfa94cebd9-Abstract.html

[90] Nai-Ching Wang, David Hicks, and Kurt Luther. 2018. Exploring Trade-Offs Between Learning and Productivity in Crowdsourced History. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–24. https://doi.org/10.1145/3274447

[91] Galen Weld, Esther Jang, Anthony Li, Aileen Zeng, Kurtis Heimerl, and Jon E. Froehlich. 2019. Deep learning for automatically detecting sidewalk accessibility problems using streetscape imagery. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 196–209. https://doi.org/10.1145/3308561.3353798

[92] Wim Westera, Mihai Dascalu, Hub Kurvers, Stefan Ruseti, and Stefan Trausan-Matu. 2018. Automated essay scoring in applied games: Reducing the teacher bandwidth problem in online training. *Computers & Education* 123 (Aug. 2018), 212–224. https://doi.org/10.1016/j.compedu.2018.05.010

[93] Dennis M. Wilkinson and Bernardo A. Huberman. 2007. Cooperation and quality in wikipedia. In *Proceedings of the 2007 international symposium on Wikis (WikiSym '07)*. Association for Computing Machinery, New York, NY, USA, 157–164. https://doi.org/10.1145/1296951.1296968

[94] Sut I Wong, Aldijana Bunjak, Matej Černe, and Christian Fieseler. 2021. Fostering Creative Performance of Platform Crowdworkers: The Digital Feedback Dilemma. *International Journal of Electronic Commerce* 25, 3 (July 2021), 263–286. https://doi.org/10.1080/10864415.2021.1942674

[95] Tom Yeh, Tsung-Hsiang Chang, Bo Xie, Greg Walsh, Ivan Watkins, Krist Wongsuphasawat, Man Huang, Larry S. Davis, and Benjamin B. Bederson. 2011. Creating contextual help for GUIs using screenshots. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. 145–154.

[96] Jieyu Zhang, Cheng-Yu Hsieh, Yue Yu, Chao Zhang, and Alexander Ratner. 2022. A Survey on Programmatic Weak Supervision. https://doi.org/10.48550/arXiv.2202.05433 arXiv:2202.05433 [cs, stat].

[97] Jieyu Zhang, Haonan Wang, Cheng-Yu Hsieh, and Alexander J. Ratner. 2022. Understanding programmatic weak supervision via source-aware influence function. *Advances in Neural Information Processing Systems* 35 (2022), 2862–2875.

[98] Yizhe Zhang, Tao Zhou, Shuo Wang, Ye Wu, Pengfei Gu, and Danny Z. Chen. 2023. SamDSK: Combining Segment Anything Model with Domain-Specific Knowledge for Semi-Supervised Learning in Medical Image Segmentation. https://doi.org/10.48550/arXiv.2308.13759 arXiv:2308.13759 [cs].

[99] Haiyi Zhu, Steven P. Dow, Robert E. Kraut, and Aniket Kittur. 2014. Reviewing versus doing: learning and performance in crowd assessment. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 1445–1455. https://doi.org/10.1145/2531602.2531718

[100] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2020. A Comprehensive Survey on Transfer Learning. https://doi.org/10.48550/arXiv.1911.02685 arXiv:1911.02685 [cs, stat].

A ADDITIONAL DISCUSSIONS ON LABELAID PIPELINE

A.1 Labeling Functions

LFs serve as flexible interfaces within the framework of PWS. We assess three aspects of each LF (Table 7): coverage (the proportion of examples each LF annotates), overlap (the proportion of examples each LF annotates that another LF also labels), and conflict (the proportion of examples each LF annotates and annotated differently by another LF). We note that it is necessary to apply as many LFs as possible for the best model performance due to the following reasons: (1) Improve coverage: each LF could capture different features of the data. More LFs can cover a higher proportion of raw data instances, leading to a larger AIA dataset generated from the PWS pipeline. (2) Reduce bias and overfitting: More LFs representing various heuristics or data insights, can mitigate systemic errors by averaging out individual LF bias. Incorporating multiple expert opinions and knowledge sources helps avoid overfitting to specific patterns or anomalies in the data, therefore making the model more generalizable. We assess the model performance with all LFs used during the PWS pipeline, and when removing one LF, the results (Table 8) show that even removing one LF during the PWS pipeline tends to hurt the end model's performance.

	Polarity	Coverage	Overlaps	Conflicts
distance_i	[0]	0.032	0.017	0.017
clustered	[1]	0.383	0.252	0.033
severity	[0, 1]	0.066	0.062	0.051
zoom	[0, 1]	0.479	0.288	0.071
tag	[1]	0.324	0.210	0.037
description	[1]	0.010	0.010	0.004
distance_r	[0]	0.027	0.027	0.019
way_type	[0, 1]	0.030	0.023	0.016

Table 7: Labeling function analysis using label matrix. [0] = Wrong, [1] = Correct.Note: distance_i is the distance to intersection, distance_r is the distance to road, and way_type is the road hierarchy according to OpenStreetMap.

A.2 Programmatic Weak Supervision vs. Hard Rule-based Approach

A key aspect of PWS is its ability to handle noise and conflicts in LFs [70–72]. Hard rule-based approaches would struggle in scenarios where LFs conflict or where the data presents ambiguities. For instance, if a user places a *Missing Curb Ramp* label within the distance threshold to the intersection but fails to provide a tag, then LFs of *distance_i* and *tag* provide contradictory annotations. PWS integrates these imperfect LFs into a probabilistic graphical model, so it can evaluate these conflicts based on the learned weight of each LF, whereas a hard rule-based approach would lack the mechanism to resolve such conflicts.

Our analysis indicates that LabelAId outperforms a hard rule-based method across all five label types (Table 9). To mitigate the complexity of resolving conflicts, we selected the most important rule from our feature importance analysis for each label type (Table 4). However, it is worth noting that hard rule-based approaches may still be valuable in low-resource scenarios. In situations where the raw dataset is small or when there is limited computational capacity to run an AI inference model, crafting a few expert-defined rules might be more feasible and efficient than establishing a complex PWS setup.

	Curb Ramp	Missing Curb	Obstacle	Surface Problem	Missing Sidewalk
With all LFs	**0.971**	**0.966**	**0.766**	**0.861**	**0.942**
Without distance_r	0.909	0.918	0.695	0.803	0.836
Without tag	0.910	0.885	0.556	0.677	0.722

Table 8: Performance decreases by label type of our LabelAId pipeline in F1 score for Seattle after one LF being removed. Note: distance_r is the distance to the road.

	Curb Ramp	Missing Curb	Obstacle	Surface Problem	Missing Sidewalk
LabelAId	**0.971**	**0.966**	**0.766**	**0.861**	**0.942**
Hard Rule-based	0.943	0.752	0.660	0.576	0.849

Table 9: Performance by label type of our LabelAId pipeline compared to the hard rule-based approach in F1 score for Seattle.

B USER EVALUATION TABLES

	rho	p-value
Number of times being intervened	-0.141	0.589
Total time spent interacting with UI	-0.230	0.374
Times viewed common mistakes	-0.066	0.801
Times viewed correct examples	-0.004	0.989
Total times viewed example screens	-0.074	0.779

Table 10: Spearman's rho correlation results for the level of intervention and precision.

Quiz	Control	Intervention	U	p-value
Pre-study	5.53 (2.07)	5.06 (1.57)	163.50	0.51
Post-study	6.88 (1.45)	6.38 (2.09)	174.00	0.30
Delta	1.35 (1.73)	1.31 (1.54)	152.50	0.78

Table 11: Quiz scores. In both pre- and post-study questionnaires, participants were shown four images for each of the five label types and were asked to select the correct ones. A sum score was calculated for all participants: each correct answer earned 1 point, and each incorrect answer was penalized with -1 point. There was no statistical difference between the two groups.

Question	Control	Intervention	U	p-value
Curb Ramp	4.65	4.71	142.0	0.914
Missing Curb	**3.88**	**4.59**	**84.5**	**0.023***
Obstacles	4.35	4.76	98.0	0.061
Surface Problems	4.18	4.47	116.0	0.276
Missing Sidewalk	4.41	4.65	123.5	0.392

Table 12: Responses to the question: "How confident are you that you can correctly recognize the following?" We mapped responses from "Not confident at all" to "Very confident" to 1-5. ($^{*}p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$).

#	Question	Control	Intervention	U	p-value
1	I feel that I have a better understanding of what a sidewalk curb ramp (or curb cut) is.	4.59	4.71	127.5	0.480
2	I feel that I understand better the accessibility challenges people with disabilities have to participate in society.	4.47	4.59	146.0	0.952
3	I feel that I have a better understanding of the sidewalk barriers that impact people who use wheelchairs or walkers.	4.71	4.71	150.5	0.788
4	I feel that I have a better understanding of the sidewalk barriers that impact people who are blind or low-vision.	4.00	4.29	120.5	0.373
5	I feel more confident about identifying problems on sidewalks faced by people with disabilities	4.47	4.53	153.5	0.721
6	**Participating in the study gave me more ideas to make sidewalks accessible for people with disabilities.**	**4.35**	**4.82**	**87.5**	**0.017***
7	I enjoyed using Project Sidewalk.	4.24	4.47	119.5	0.336
8	It was easy for me to use Project Sidewalk.	4.47	4.29	153.5	0.728
9	**I felt that an AI agent was watching my performance while I was labeling.**	**1.82**	**3.00**	**68.0**	**0.006****
10	**I felt that an AI agent was helping me throughout the task.**	**2.29**	**3.24**	**83.5**	**0.028***
11	Overall, I desired more active help to complete the labeling tasks.	3.88	3.24	189.0	0.106

Table 13: Responses to the question "To what extent do you agree with the following statements?". We mapped responses such as "Strongly disagree" to "Strongly agree" to 1-5. (*p ≤ 0.05, **p ≤ 0.01, ***p ≤ 0 .001).

C LABELING ASSISTANCE INTERFACE

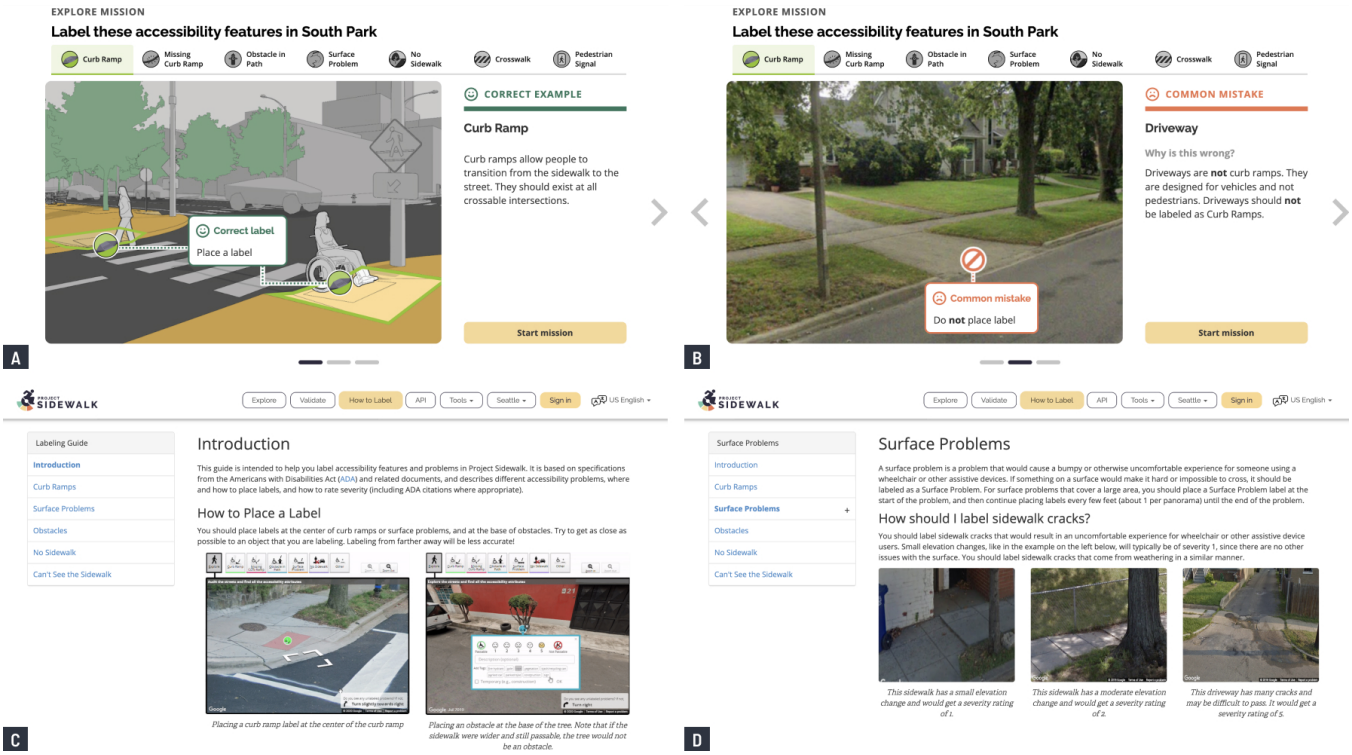


Figure 11: Project Sidewalk built-in labeling assistance. (A) & (B) Illustrated example screens shown in the beginning of each route. The label type is rotated every time. (C) & (D) The *How to Label* Section. Participants may access this section at any time during the labeling process.