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.
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 crowdsourcing 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 points1.

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 the 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 LabelAld significantly improves label precision by 19.2% without compromising efficiency.

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

<table>
<thead>
<tr>
<th></th>
<th>Seattle</th>
<th>Chicago</th>
<th>Oradell</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unannotated</td>
<td>Expert-Validated</td>
<td>Unannotated</td>
<td>Expert-Validated</td>
</tr>
<tr>
<td>Curb Ramp</td>
<td>70,690</td>
<td>5,333</td>
<td>5,710</td>
<td>2,386</td>
</tr>
<tr>
<td>Missing Curb Ramp</td>
<td>32,968</td>
<td>4,239</td>
<td>463</td>
<td>1,294</td>
</tr>
<tr>
<td>No Sidewalk</td>
<td>36,021</td>
<td>3,460</td>
<td>2,211</td>
<td>48</td>
</tr>
<tr>
<td>Surface Problem</td>
<td>26,912</td>
<td>2,909</td>
<td>2,136</td>
<td>1,651</td>
</tr>
<tr>
<td>Obstacle</td>
<td>10,103</td>
<td>407</td>
<td>1,254</td>
<td>320</td>
</tr>
<tr>
<td>All Label Types</td>
<td>176,694</td>
<td>16,348</td>
<td>11,774</td>
<td>5,699</td>
</tr>
</tbody>
</table>

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 $D_T$ 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 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.

### 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

$D ←$ 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 $\min_{i \in I} \text{distance}(l, i) > D$ then

$l ←$ wrong

else

$l ←$ correct

end if

end for

3.2.3 Multi-city Pre-training. To train a discriminative model, we start by pre-training on the AIA dataset $D_n$ 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 $D_n$ 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...
We evaluate the performance of our pipeline, which involves the PWS pipeline integrated with city-specific fine-tuning of a pre-trained model. 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 $D_g$, with a reduced learning rate of $3 \times 10^{-5}$ to prevent 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 $1 \times 10^{-4}$, and a weight decay of $1 \times 10^{-5}$ 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.

For testing purposes, we utilized the remaining $D_g$ 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 $D_g$ comprises 16,148, 5,499, and 3,652 expert-validated labels for Seattle, Chicago, and Oradell, respectively.

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 LabelAld’s performance through analysis of feature importance.

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

<table>
<thead>
<tr>
<th>City</th>
<th>Curb Ramp</th>
<th>Missing Curb Ramp</th>
<th>Obstacle</th>
<th>Surface Problem</th>
<th>Missing Sidewalk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seattle</td>
<td>0.971</td>
<td>0.966</td>
<td>0.766</td>
<td>0.861</td>
<td>0.942</td>
</tr>
<tr>
<td>Chicago</td>
<td>0.968</td>
<td>0.494</td>
<td>0.693</td>
<td>0.718</td>
<td>0.929</td>
</tr>
<tr>
<td>Oradell</td>
<td>0.972</td>
<td>0.768</td>
<td>0.793</td>
<td>0.944</td>
<td>0.988</td>
</tr>
</tbody>
</table>

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.

3.3.2 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.

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 $D_s$ 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.

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 = ±σ).
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.

4 LABEALAID: 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?

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.

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.
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 OpenStreetMap3. 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.

3https://www.openstreetmap.org/
For the actual study, two study moderators led six online sessions, three for each condition. Each session had five to seven 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. Then, 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. 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. Then, 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.

4.4 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 one. A sum score was calculated for all participants: each correct answer earned 1 point, and each incorrect answer was penalized with -1点.

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
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.

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

<table>
<thead>
<tr>
<th>Label Type</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curb Ramp</td>
<td>436</td>
<td>454</td>
<td>890</td>
</tr>
<tr>
<td>Missing Curb Ramp</td>
<td>265</td>
<td>245</td>
<td>510</td>
</tr>
<tr>
<td>Obstacle</td>
<td>309</td>
<td>298</td>
<td>607</td>
</tr>
<tr>
<td>Surface Problem</td>
<td>243</td>
<td>249</td>
<td>492</td>
</tr>
<tr>
<td>Missing Sidewalk</td>
<td>94</td>
<td>72</td>
<td>166</td>
</tr>
<tr>
<td>Overall</td>
<td>1347</td>
<td>1318</td>
<td>2665</td>
</tr>
</tbody>
</table>

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

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.

| 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.

| Table 6: Distribution of participants’ labels across all label types. C stands for Control group and I stands for Intervention group. |
We are also interested in self-efficacy and learning. Although the means were higher for the intervention group across the majority responded with "Somewhat Agree" or "Strongly Agree" the statements that the pop-up prompts were helpful and likable, able view of LabelAI. When asked to what extent they agreed with the overall user performance, confidence and learning gain, we now turn to the perceived usefulness and presence of AI in LabelAI.

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 LabelAI.

Perceived usefulness. Participants generally expressed a favorable view of LabelAI. 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 LabelAI 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 LabelAI 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 LabelAI.

5 DISCUSSION

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

5.1 Reflecting on LabelAI’s Performance

Below, we reflect on LabelAI’s 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 capture 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 over-cluttering 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 on 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 programatically 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 labeling 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
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.

Table 7: Labeling function analysis using label matrix. [0] = Wrong, [1] = Correct. Note: distance_i is the distance to 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.

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.

<table>
<thead>
<tr>
<th>Label Type</th>
<th>With all LFs</th>
<th>Without distance_r</th>
<th>Without tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curb Ramp</td>
<td>0.971</td>
<td>0.909</td>
<td>0.910</td>
</tr>
<tr>
<td>Missing Curb</td>
<td>0.966</td>
<td>0.918</td>
<td>0.885</td>
</tr>
<tr>
<td>Obstacle Surface</td>
<td>0.766</td>
<td>0.695</td>
<td>0.656</td>
</tr>
<tr>
<td>Problem</td>
<td>0.861</td>
<td>0.803</td>
<td>0.677</td>
</tr>
<tr>
<td>Missing Sidewalk</td>
<td>0.942</td>
<td>0.836</td>
<td>0.722</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Label Type</th>
<th>LabelAId</th>
<th>Hard Rule-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curb Ramp</td>
<td>0.971</td>
<td>0.943</td>
</tr>
<tr>
<td>Missing Curb</td>
<td>0.966</td>
<td>0.752</td>
</tr>
<tr>
<td>Obstacle Surface</td>
<td>0.766</td>
<td>0.660</td>
</tr>
<tr>
<td>Problem</td>
<td>0.861</td>
<td>0.576</td>
</tr>
<tr>
<td>Missing Sidewalk</td>
<td>0.942</td>
<td>0.849</td>
</tr>
</tbody>
</table>

B USER EVALUATION TABLES

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>rho</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of times being intervened</td>
<td>-0.141</td>
<td>0.589</td>
</tr>
<tr>
<td>Total time spent interacting with UI</td>
<td>-0.230</td>
<td>0.374</td>
</tr>
<tr>
<td>Times viewed common mistakes</td>
<td>-0.066</td>
<td>0.801</td>
</tr>
<tr>
<td>Times viewed correct examples</td>
<td>-0.004</td>
<td>0.989</td>
</tr>
<tr>
<td>Total times viewed example screens</td>
<td>-0.074</td>
<td>0.779</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Question</th>
<th>Control</th>
<th>Intervention</th>
<th>U</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curb Ramp</td>
<td>4.65</td>
<td>4.71</td>
<td>142.0</td>
<td>0.914</td>
</tr>
<tr>
<td>Missing Curb</td>
<td>3.88</td>
<td>4.59</td>
<td>84.5</td>
<td>0.023*</td>
</tr>
<tr>
<td>Obstacles</td>
<td>4.35</td>
<td>4.76</td>
<td>98.0</td>
<td>0.061</td>
</tr>
<tr>
<td>Surface Problems</td>
<td>4.18</td>
<td>4.47</td>
<td>116.0</td>
<td>0.276</td>
</tr>
<tr>
<td>Missing Sidewalk</td>
<td>4.41</td>
<td>4.65</td>
<td>123.5</td>
<td>0.392</td>
</tr>
</tbody>
</table>

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 ≤ 0.05, **p ≤ 0.01, ***p ≤ 0.001).
<table>
<thead>
<tr>
<th>#</th>
<th>Question</th>
<th>Control</th>
<th>Intervention</th>
<th>U</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I feel that I have a better understanding of what a sidewalk curb ramp (or curb cut) is.</td>
<td>4.59</td>
<td>4.71</td>
<td>127.5</td>
<td>0.480</td>
</tr>
<tr>
<td>2</td>
<td>I feel that I understand better the accessibility challenges people with disabilities have to participate in society.</td>
<td>4.47</td>
<td>4.59</td>
<td>146.0</td>
<td>0.952</td>
</tr>
<tr>
<td>3</td>
<td>I feel that I have a better understanding of the sidewalk barriers that impact people who use wheelchairs or walkers.</td>
<td>4.71</td>
<td>4.71</td>
<td>150.5</td>
<td>0.788</td>
</tr>
<tr>
<td>4</td>
<td>I feel that I have a better understanding of the sidewalk barriers that impact people who are blind or low-vision.</td>
<td>4.00</td>
<td>4.29</td>
<td>120.5</td>
<td>0.373</td>
</tr>
<tr>
<td>5</td>
<td>I feel more confident about identifying problems on sidewalks faced by people with disabilities</td>
<td>4.47</td>
<td>4.53</td>
<td>153.5</td>
<td>0.721</td>
</tr>
<tr>
<td>6</td>
<td>Participating in the study gave me more ideas to make sidewalks accessible for people with disabilities.</td>
<td>4.35</td>
<td>4.82</td>
<td>87.5</td>
<td>0.017*</td>
</tr>
<tr>
<td>7</td>
<td>I enjoyed using Project Sidewalk.</td>
<td>4.24</td>
<td>4.47</td>
<td>119.5</td>
<td>0.336</td>
</tr>
<tr>
<td>8</td>
<td>It was easy for me to use Project Sidewalk.</td>
<td>4.47</td>
<td>4.29</td>
<td>153.5</td>
<td>0.728</td>
</tr>
<tr>
<td>9</td>
<td>I felt that an AI agent was watching my performance while I was labeling.</td>
<td>1.82</td>
<td>3.00</td>
<td>68.0</td>
<td>0.006**</td>
</tr>
<tr>
<td>10</td>
<td>I felt that an AI agent was helping me throughout the task.</td>
<td>2.29</td>
<td>3.24</td>
<td>83.5</td>
<td>0.028*</td>
</tr>
<tr>
<td>11</td>
<td>Overall, I desired more active help to complete the labeling tasks.</td>
<td>3.88</td>
<td>3.24</td>
<td>189.0</td>
<td>0.106</td>
</tr>
</tbody>
</table>

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.