LabelAld  Just-in-time AI Interventions
for Improving Human Labeling Quality and Domain Knowledge

Chu Li*, Zhihan Zhang*, Michael Saugstad, Esteban Safranchik, Minchu Kulkarni, Xiaoyu Huang,
Shwetak Patel, Vikram Iyer, Tim Althoff, Jon E. Froehlich
System feedback can influence labeling confidence, quality & knowledge.
System feedback can influence labeling confidence, quality & knowledge.

Are you sure this is an alpaca?
How can we leverage system feedback?

crowd workers → question mark → quality data + educated crowd
Quality control remains a major challenge in crowdsourcing.
Prior Work

Quality Control + Learning Experience

Example Projects
- Shepherd
  Dow et al., 2012
- Review vs. Doing
  Zhu et al., 2014
- Learning from the Crowd
  Mamykina et al., 2016

Learning Facilitation Methods
- Peer review
- Expert feedback
- Self-assessment

Limitation
- Requires additional review commitments
- Impacts scalability
LabelAID Pipeline

def f(x):
    return perm.heuristic(x)
Project Sidewalk

5 Major Label Types

- Curb Ramps
- Missing Curb Ramps
- Obstacles
- Surface Problems
- Missing Sidewalk
Project Sidewalk

21 Cities
13,660+ Users
950,276+ Labels
Project Sidewalk Quality Control

1. Find, label, and assess sidewalks

2. Validate & correct other labels
1. Find, label, and assess sidewalks
2. Validate & correct other labels
Project Sidewalk Quality Control
Driveway is not a curb ramp!
Project Sidewalk Quality Control
Project Sidewalk Quality Control

Data noise at neighborhood scale
Adapting LabelAI to Project Sidewalk
Programmatic Weak Supervision

Domain Knowledge + Heuristics

Unannotated Data → Labeling Functions → Label Model → Automatically Imperfectly Annotated Data

```python
def if_ (x):
    return per_heuristic(x)
```
Labeling Function Example

Domain Knowledge
Set Of Labeling Functions

Planning Guidelines
Distance to urban infrastructure

Crowdsource Nature
In proximity with other user’s labels

User Behaviors
Optional inputs & labeling zoom level

Label Characteristics
Severity rating
LabelAId Pipeline

Unannotated Data → Labeling Functions → Label Model → Automatically Imperfectly Annotated Data → Pre-trained Model → Expert-Validated Data → Fine-tuned Model + Downstream Tasks

Programmatic Weak Supervision
Pre-Training & Fine-Tuning

- Automatically Imperfectly Annotated Data
- Pre-trained Model
- Expert-Validated Data
- Fine-tuned Model + Downstream Tasks
FT-Transformer-Based Model Architecture
Technical Evaluation

![Accuracy Graph](image1.png)

![F1-Score Graph](image2.png)

- **Accuracy**: Models' performance increase with the number of expert-validated data points.
- **F1-Score**: Models show a steady improvement, with LabelAId consistently performing better than the other models in terms of F1-Score.

Models compared:
- LabelAId
- XGBoost
- Random Forest
- MLP
- Logistic Regression
Technical Evaluation

LabelAId improves accuracy by up to 37% with just 50 downstream samples.
User Flow

1. Inferred as incorrect
2. Are you sure?
3. Yes, I am sure → Label kept
4. No, delete the label → Label deleted
5. View Common Mistakes
6. Your Label vs MISTAKES
7. Correct
Are you sure this is a Curb Ramp? 😁

Tip: do **not** label *driveways* as curb ramps.

- View Common Mistakes
- Yes, I am sure
- No, remove the label
Are you sure this is a Curb Ramp?  

Tip: do not label driveways as curb ramps.

View Common Mistakes  Yes, I am sure  No, remove the label
Are you sure this is a Curb Ramp?  ✉️
Tip: do **not** label **driveways** as curb ramps.

* View Common Mistakes  * Yes, I am sure  * No, remove the label
Your Curb Ramp Label

⚠️ Common Mistakes

**Driveways.** Driveways are not curb ramps. They are designed for vehicles and not pedestrians.

**Curb Missing.** When a curb ramp is missing, use the Missing Curb Ramp label instead.

**Not on pedestrian route.** Curb ramps are not needed at paths not intended for pedestrians.

See common mistakes 1/2  See correct examples
**Your Curb Ramp Label**

- **Correct Examples**
  - This is a good **curb ramp**. It's wide, has a yellow tactile warning, and is not too steep.
  - This is an OK **curb ramp**. It's missing a tactile warning strip and is angled into the street.
  - Label flat **curb ramps** with tactile warning strips.
  - Some corners have very wide **curb ramps** to support travel in both directions.

See common mistakes < 2/2 ➔ See correct examples
## Between-Subjects User Study

<table>
<thead>
<tr>
<th>Participants &amp; Task</th>
<th>Measures</th>
<th>Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 34 participants</td>
<td>• Task performance</td>
<td>• Data logging</td>
</tr>
<tr>
<td>• Randomly assigned to 2 conditions</td>
<td>• Labeling confidence</td>
<td>• Quiz</td>
</tr>
<tr>
<td>• 8 labeling routes</td>
<td>• Knowledge gain</td>
<td>• Questionnaire</td>
</tr>
<tr>
<td></td>
<td>• User preference</td>
<td>• Interview</td>
</tr>
</tbody>
</table>
Can LabelAld improve labeling performance?
Findings - Task Performance

- LabelAId improves precision by up to 19% without compromising labeling time

<table>
<thead>
<tr>
<th>Label Type</th>
<th>Control</th>
<th>Intervention</th>
<th>U</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>0.699 ± 0.199</td>
<td>0.891 ± 0.053</td>
<td>50.0</td>
<td>0.001 **</td>
</tr>
<tr>
<td>Curb Ramp</td>
<td>0.686 ± 0.346</td>
<td>0.956 ± 0.067</td>
<td>70.0</td>
<td>0.038 *</td>
</tr>
<tr>
<td>No Curb Ramp</td>
<td>0.802 ± 0.164</td>
<td>0.918 ± 0.091</td>
<td>80.5</td>
<td>0.025 *</td>
</tr>
<tr>
<td>Obstacle</td>
<td>0.7610 ± 0.126</td>
<td>0.812 ± 0.111</td>
<td>85.5</td>
<td>0.183</td>
</tr>
<tr>
<td>Surface Problem</td>
<td>0.812 ± 0.230</td>
<td>0.894 ± 0.116</td>
<td>100.0</td>
<td>0.423</td>
</tr>
<tr>
<td>No Sidewalk</td>
<td>0.842 ± 0.267</td>
<td>0.867 ± 0.208</td>
<td>66.5</td>
<td>0.480</td>
</tr>
</tbody>
</table>
How did the **perception of labeling ability** and **knowledge gained** differ between the two groups?
Findings - Self Efficacy & Learning Gains

- Higher self-efficacy in intervention group
- Similar objective learning gains
- Higher subjective learning gains in intervention group
How did participants perceive LabelAld?
Findings - Perceived Usefulness

- LabelAld was helpful (82.35%) & likable (64.7%)

“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.”

- Intervention Group Participant
Discussion

- Can AI-assistance replicate human feedback?
- How to design interactions with imperfect ML models?
- Cognitive forcing function reduces over-reliance on AI
Growing concern about AI-based assistance: over-reliance on AI, reduced human cognitive engagement
Discussion

Obstacles high false positive rate 36.2%

Users rejected 83 % incorrect suggestions

Users decide before seeing AI’s suggestion

Cognitive forcing functions

elicit analytical thinking at decision making time

effectively reduces reliance on AI
LabelAld  Just-in-time AI Interventions
LabelAld  Just-in-time AI Interventions for Improving Human Labeling Quality and Domain Knowledge

A novel pipeline that facilitates the training of AI-based inference models for detecting labeling mistakes

A human-AI collaborative system designed to create teachable moments in crowdsourcing workflows

Contacts
chuchuli@cs.uw.edu
zzhihan@cs.uw.edu