

# Practical Issues: Non-Independent Work and Argumentation

Cenhao, Ruiwei, and Yang

# Outline

- **MicroTalk**: Using Argumentation to Improve Crowdsourcing Accuracy
- **Revolt**: Collaborative Crowdsourcing for Labeling Machine Learning Datasets
- **Atelier**: Repurposing Expert Crowdsourcing Tasks as Micro-internships
- **Cicero**: Multi-Turn, Contextual Argumentation for Accurate Crowdsourcing

# MicroTalk: Using Argumentation to Improve Crowdsourcing Accuracy

— Ryan Drapeau, Lydia B. Chilton, Jonathan Bragg, Daniel S. Weld

# Warm-Up

**Claim: Nicolas Sarkozy "lived in" France**

France is a Country.

**Sentence: Nicolas Sarkozy leads Bastille Day celebrations, his first after being elected as France's president.**

- ☐ True
- ☐ False

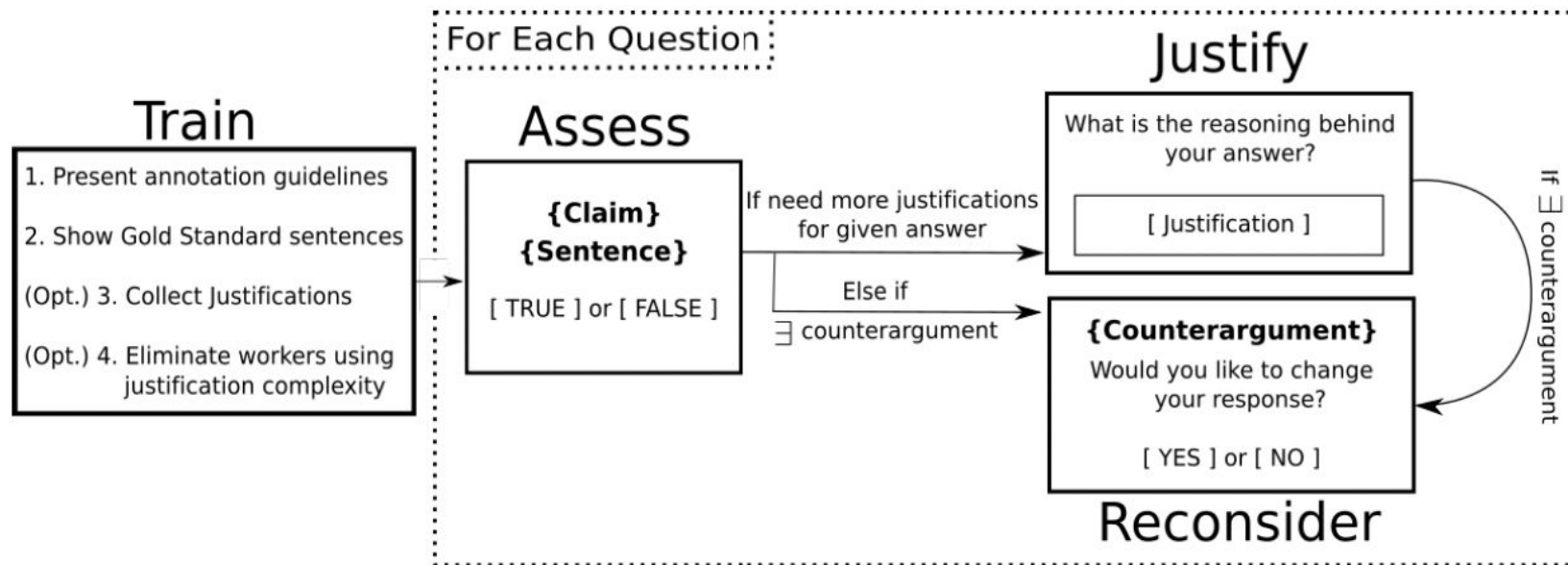
Submit your answer here: <https://pollev.com/ruiweixiao965>

In one minute

# Introduction

- Workers can be wrong on extremely difficult problems.
- Existing aggregation mechanisms (i.e. Majority Voting, EM) fail on such problems with local minima.
- Existing collaborative approaches do not help individual voters make well-informed decisions.
- This paper presents a new quality-control workflow ***MicroTalk*** to improve the accuracy.
- The experiment is tested on Relation Extraction Domain.

# Workflow (Prototype)



# Workflow (Worker Interface)

**Claim: Nicolas Sarkozy "lived in" France**

France is a Country.

**Sentence: Nicolas Sarkozy leads Bastille Day celebrations, his first after being elected as France's president.**

☐ True  
☐ False

Figure 3: The Assess microtask.

**Claim: Nicolas Sarkozy "lived in" France**

France is a Country.

**Sentence: Nicolas Sarkozy leads Bastille Day celebrations, his first after being elected as France's president.**

**What is the reasoning behind your answer?**

Justification \_\_\_\_\_

Figure 4: The Justify microtask.

**Claim: Nicolas Sarkozy "lived in" France**

France is a Country.

**Sentence: Nicolas Sarkozy leads Bastille Day celebrations, his first after being elected as France's president.**

**Some workers answered True with the following reasons:**

- Nicolas Sarkozy holds national office in France, therefore we can conclude that he also lives in the country he represents.

**Would you like to change your answer to True?**

☐ Yes  
☐ No

Figure 5: The Reconsider microtask.

# Discussion

How would you select/generate the justification? (i.e. length of the justification)

How would you distribute (financial) incentive to this workflow?

**Claim: Nicolas Sarkozy "lived in" France**

France is a Country.

**Sentence: Nicolas Sarkozy leads Bastille Day celebrations, his first after being elected as France's president.**

**Some workers answered True with the following reasons:**

- Nicolas Sarkozy holds national office in France, therefore we can conclude that he also lives in the country he represents.

**Would you like to change your answer to True?**

☐ Yes

☐ No



# Workflow (Justification & Reconsider)

## Selecting Discerning Workers To Ensure Justification Quality

- LSAT Critical Thinking Task (✗)
- Length of Justification (✗)
- Flesch-Kincaid readability tests (✓)

## Proactive vs. Lazy Justification

$$S = \langle T, T, T, T, T \rangle \quad \langle A^1, J^1, A^2, A^3, A^4, A^5 \rangle$$

$$S' = \langle T, T, F, F, T \rangle \quad \langle A^1, J^1, A^2, A^3, J^3, R^3, A^4, \bar{R}^4, A^5, R^5 \rangle$$

# The Final Workflow



## Payment

- \$0.05 for finishing each microtask (Assess, Justify, Reconsider).
- \$0.05 bonus for every high-quality justification.
- \$0.05 bonus if they chose the correct response for Reconsider.

**Input** : A question  $q$ , budget  $B$ , and task costs  $C_a, C_j, C_r$

**Output**: An answer  $A(q)$

$j_T := j_F := ''$  ;

$b := 0$  ;

**for**  $i := 1$ , increase by 1, **while**  $b < B$  **do**

    Justifying = F ;

    Train and qualify worker  $w_i$ ;

$a_i := \text{Assess}(q, w_i)$  ;

$b := b + C_a$  ;

**if**  $j_{a_i} = ''$  **then**

$j_{a_1} := \text{Justify}(q, a_i, w_i)$  ;

$b := b + C_j$  ;

        Justifying := T

**end**

**if**  $j_{\overline{a_i}} \neq ''$  **then**

$a := \text{Reconsider}(q, j_{\overline{a_i}}, w_i)$  ;

$b := b + C_r$  ;

**if**  $a \neq a_i \wedge \text{Justifying}$  **then**

$j_{a_i} := ''$

**end**

$a_i := a$

**end**

**end**

**return**  $\text{Aggregate}(a_1, \dots, a_i)$ ;

**Algorithm 1:** The MicroTalk argumentation workflow.

# Experiment (Setup)

TAC KBP person-place relation questions

Training and Worker Selection

- Workers who had completed at least 1,000 tasks with a 97% acceptance rate
- TAC KBP annotation guidelines
- Answer five gold standard questions and had to get three (60%) or more correct

# Experiment (Questions)

- 1) Are workers able to formulate a convincing argument with their assessment?
- 2) Does argumentation have an effect on individual workers' accuracy?
- 3) Do workers perform better when higher quality justifications are shown?
- 4) How do we find high-quality workers and how much better at argumentation are they?
- 5) How does the MicroTalk workflow compare with other approaches and is it cost effective?

# Experiment

- 1) Are workers able to formulate a **convincing argument** with their assessment?
  - Workers with higher-performance (higher accuracy) tend to generate explicit reasoning and making reference to the annotation guidelines.
- 2) Does argumentation have an effect on individual workers' accuracy?
  - Workers given all three microtasks (71%) were significantly more accurate than those that only completed Assess tasks (59%).

# Experiment

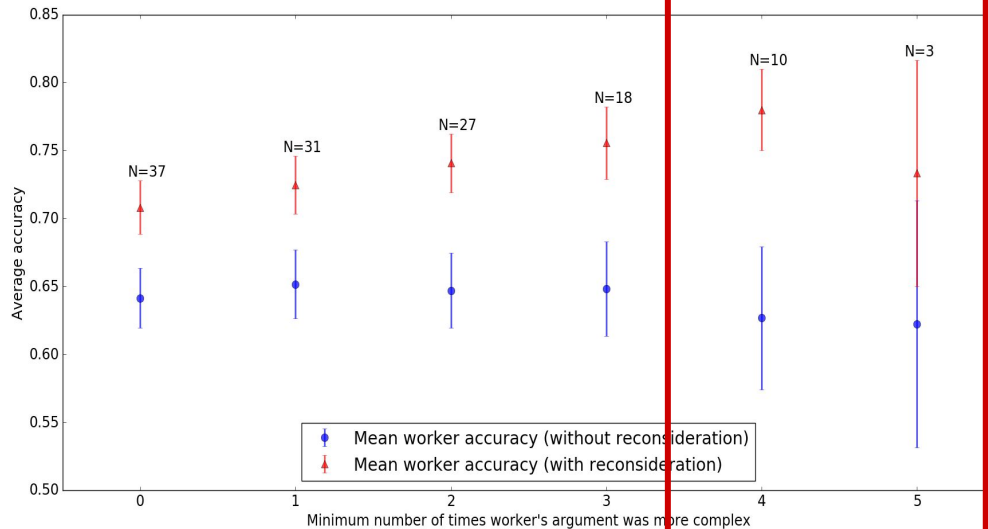
3) Do workers perform better when higher quality justifications are shown?

- Workers that saw arguments **written by other workers** changed their answer 20% of the time. However, when workers were shown arguments **written by the authors**, this rate increases to 46%
- No statistically significant difference in the number of responses that were **changed to the correct** answer in either condition

# Experiment

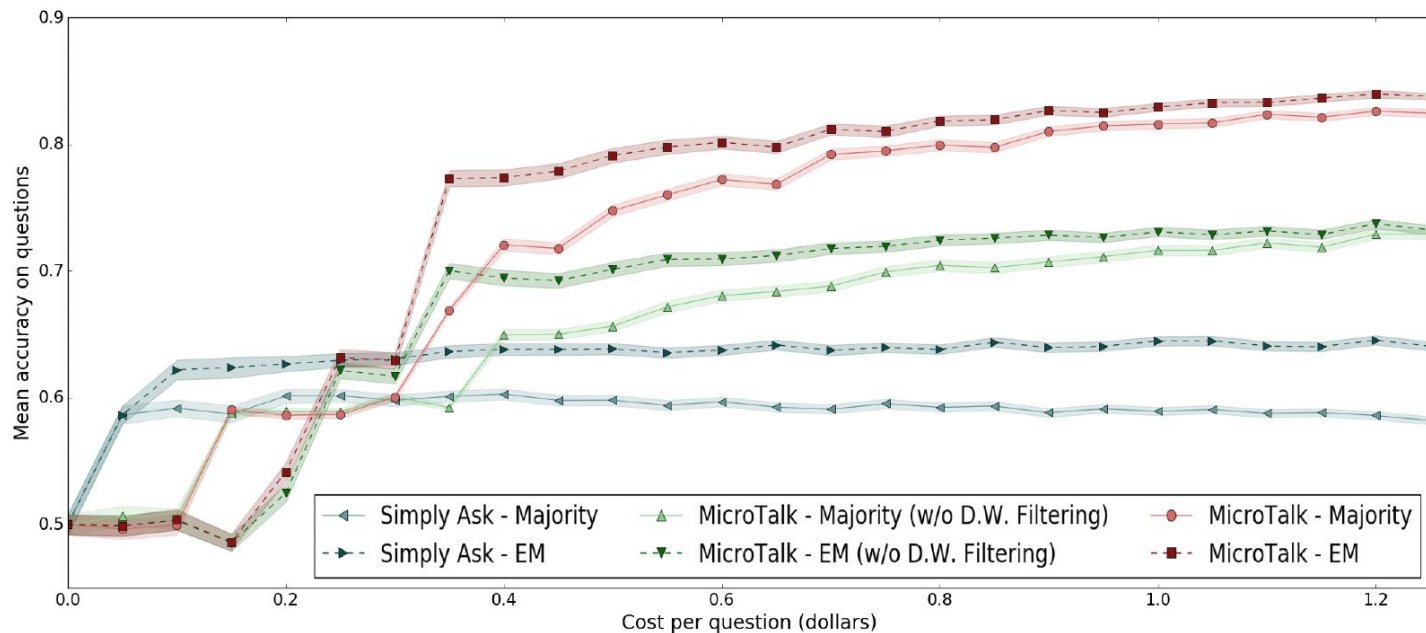
## 4) How do we find high-quality workers and how much better at argumentation are they?

- Select “Discerning workers”
- Incorporate reconsideration
- The average accuracy with reconsideration increased to 78% from 71% (general population).



# Experiment

5) How does the MicroTalk workflow compare with other approaches and is it cost effective?





# Discussion

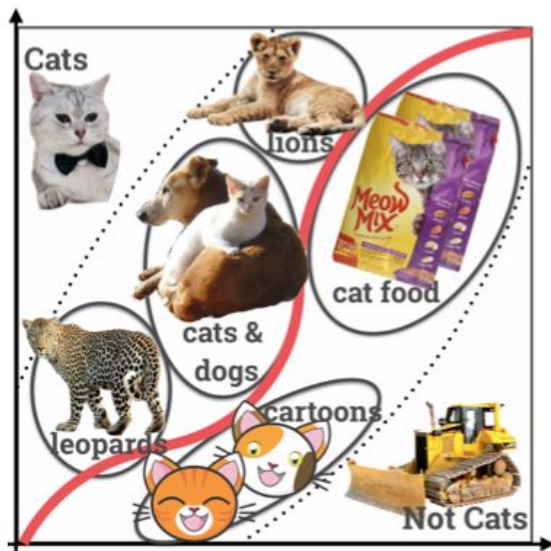
Can you think of one type of task that this workflow works well with and another one that applying this workflow is less effective?

# Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets

— Joseph Chee Chang, Saleema Amershi, Ece Kamar

# Ambiguous label guidelines

Cats or Not cats?



Need comprehensive label guidelines

You should **select** concepts like these:



Dog

and **NOT select** concepts like these:



Cartoon Dog



Wolf



Statue



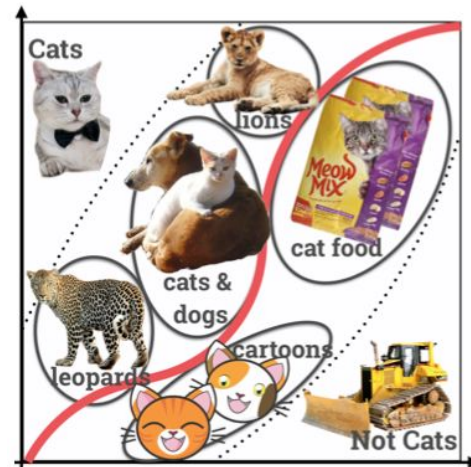
Robot Dog

# Ambiguous label guidelines

- Incomplete or ambiguous label guidelines can then result in **differing** interpretations of concepts and **inconsistent** labels.
- Lead to **rejection of honest work** and a missed opportunity to capture rich interpretations about data.
- Creating comprehensive label guidelines for crowdworkers is often prohibitive even for seemingly simple concepts.

# Revolt

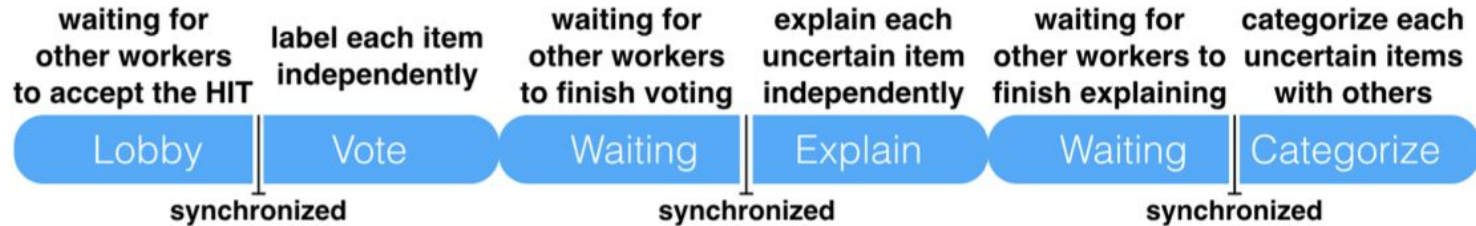
- Eliminates the burden of creating detailed label guideline
- Harnesses crowd disagreements to identify ambiguous concepts and create rich structures (groups of semantically related items) for post-hoc label decisions.
- Produce reusable structures that can accommodate a variety of label boundaries without requiring new data to be collected.



# Revolt

Three stages:




1. *Vote*: where crowdworkers label as in traditional labeling,
2. *Explain*: where crowdworkers provide justifications for their labels on conflicting items,
3. *Categorize* :where crowdworkers review explanations from others and then tag conflicting items with terms describing the newly discovered concepts.



# Vote stage

Revolt coordinates crowdworkers to create labels for *certain* items, and identify *uncertain* items for further explanation and processing. This method also avoids unfairly rejecting honest work.



We want to know if the main theme of the items below are "Cats". Label "Cat" if you think the main theme of the item is Cats, otherwise label "Not Cat". Label "Maybe/Not Sure" for items that you are uncertain about or if you think other workers might pick different labels.

	<input type="radio"/> Cat <input checked="" type="radio"/> Not Cat <input type="radio"/> Maybe/NotSure
	<input checked="" type="radio"/> Cat <input type="radio"/> Not Cat <input type="radio"/> Maybe/NotSure
	<input type="radio"/> Cat <input type="radio"/> Not Cat <input checked="" type="radio"/> Maybe/NotSure

# Explain Stage

Crowdworkers are asked to provide short explanations about their labels for items flagged as uncertain or the label disagreed by others in the previous stage.

The other workers have also finished labeling the same items you just labeled. The following items received different labels. Please provide an explanation for each of your labels below.



	<p>You labeled "Not Cat". Please focus on describing things about the item that could have made it difficult or ambiguous for others.</p> <p><input type="text" value="This is a tiger."/> <input type="button" value="Save"/></p>
	<p>You labeled "Maybe/NotSure". Please focus on describing things about the item that could have made it difficult or ambiguous for others.</p> <p><input type="text" value="This is a cartoon drawing of a cat."/> <input type="button" value="Save"/></p>



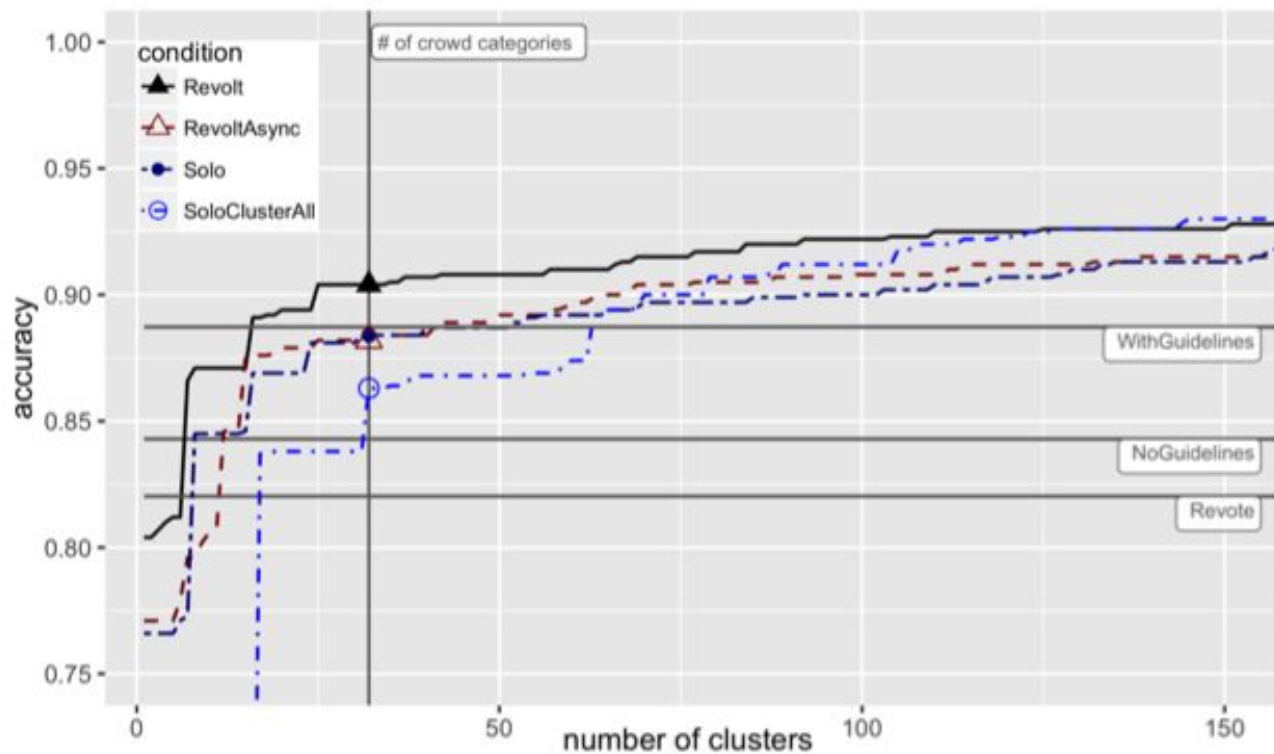
# Categorize Stage

Crowdworkers were tasked with grouping uncertain items into categories based on explanations from others in the group. Categories could either be selected from a list of existing categories presented next to each item or added manually via a text input field.

You labeled differently on the following items. Please review all the explanations provided by other workers and pick or come up with good category names so the requesters can make an informed decision afterwards.

	<input type="text"/> <input type="button" value="Create"/>	<b>worker1:</b> This is a tiger.
	<div><div>big cats</div><div>cartoon cats</div><div>cats with dogs</div></div>	<b>worker2:</b> This is a big cat. <b>worker3:</b> Do lions and other big cats
	<input type="text"/> <input type="button" value="Create"/>	<b>worker1:</b> This is a cartoon drawing of a cat.
	<div><div>big cats</div><div>cartoon cats</div><div>cats with dogs</div></div>	<b>worker2:</b> Cat drawing. <b>worker3:</b> Do cartoon cats count?

# Evaluation



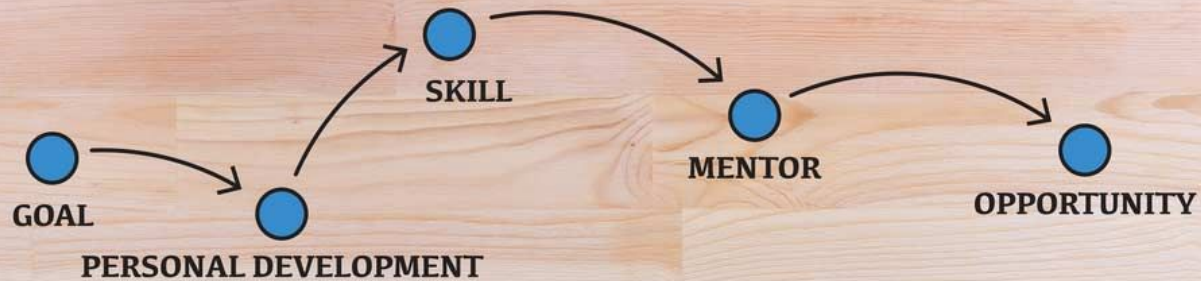
# Discussion

1. As a requester, which methods you prefer, MicroTalk or Revolt? Based on the monetary and time cost.
2. Each worker may has their own explanations for uncertain items and add a new category, how can reuse these new added category and reduce redundancy?

# Atelier: Repurposing Expert Crowdsourcing Tasks as Micro-internships

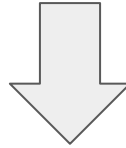
— Ryo Suzuki, Niloufar Salehi, Michelle S. Lam,  
Juan C. Marroquin, Michael S. Bernstein

# INTERNSHIP



# ABSTRACT

Many workers cannot afford to invest the time and sacrifice the earnings required to learn a new skill, and a lack of experience makes it difficult to get job offers even if they do.



Lower the threshold to skill development by repurposing existing tasks on the marketplace as mentored, paid, real-world work experiences.

# Micro-internships

# Micro-internships

For crowd work to stand as a viable *long-term* career option, online crowd experts must be able to grow and continually refresh their skills.

- Traditional workplaces: utilize **on-the-job training** and **internships** to enable employees' skill development while providing financial support.
- Crowd workers: time spent learning is time spent not working, which **reduces income**, and it is **difficult to get hired** for new skills

Many workers' skills remain **static**, and workers today often view crowdsourcing marketplaces as places to seek **temporary** jobs for their preexisting skills rather than as venues for long-term career development.

# Micro-internships

Workers (interns): learn by applying their skills to achieve a real-world goal, get paid for their time, and gain a portfolio item and rating feedback to help break into the new area.

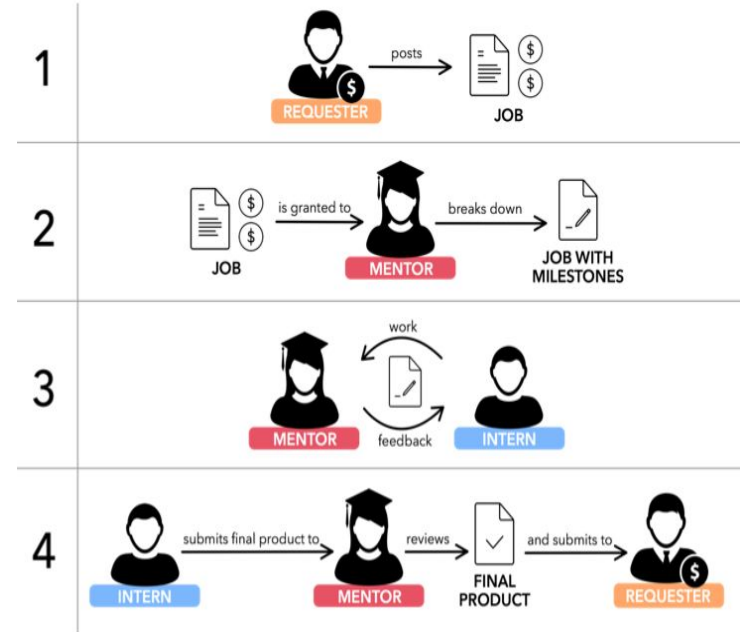


Experts (mentors): Guide, provide feedback, vouch for the final task quality, and maintain their usual payment rate by mentoring for less time. The mentors break down the task into *milestones*, provide *feedback*, *answer questions*, and interface with the task requester, in order to ensure the quality of the final product.



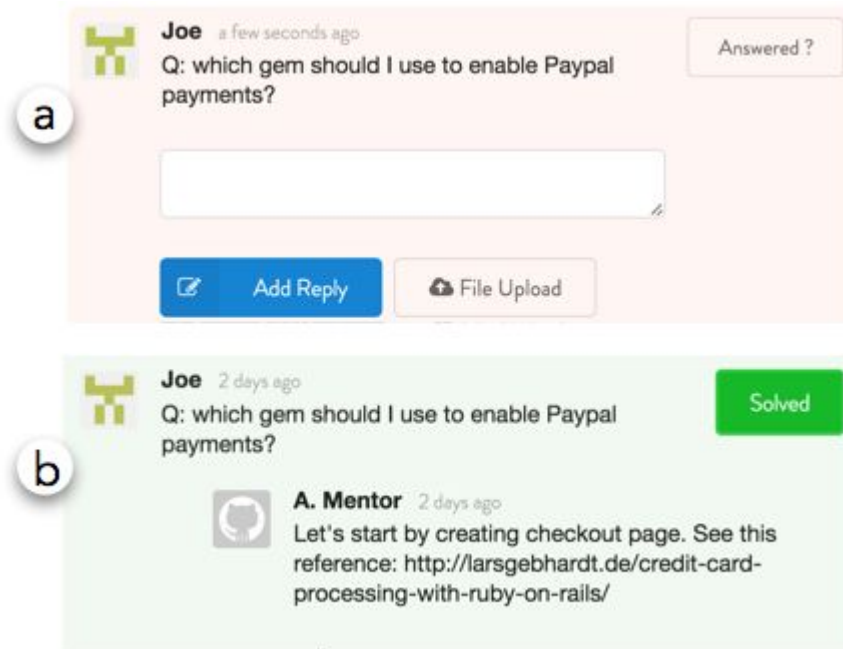
# Atelier

1. The requester chooses a mentor they trust.
2. The mentor uses the platform to author milestones, and chooses an intern from applicants.
3. The mentor and intern agree on office hours to review the work.
4. When the intern submits the final work product, the mentor reviews it and returns it to the original requester for both payment and feedback rating for the intern.



# Features

- **Milestones**
  - Designed by mentors
  - Builds scaffolding for learning
  - Facilitates communication
- **Office hours & question boards**
- **Feedback and Ratings**
  - Private feedback until work is accepted
  - Public ratings influence reputations
- **Publishing**
  - Optional publication of project proceedings



# Incentive Design

- Requestor
  - Reducing cost while assuring quality
  - Identifying promising future employees
  - Prosocial act
- Mentor
  - Extra income with small time commitment
  - Intrinsic motivation to educate
- Intern
  - Higher income compared to previous job
  - Skill development

# Discussion

1. What do you think about the incentive design? Do you think the current incentives are attractive enough, especially for the requesters and mentors?
2. A challenge identified by the authors is responsibility management: if the project fails, who and how does the platform hold responsible? How would you keep the mentors and interns accountable while preserving their motivation?

# Cicero: Multi-Turn, Contextual Argumentation for Accurate Crowdsourcing

# Drawbacks of MicroTalk

- the steering power of a one-round debate is limited
  - Only pre-collected justification for an opposing answer
  - No specific and personalized argument
  - No back-and-forth interaction that could illuminate subtle differences
  - Impractical for tasks with many answers

# Cicero

- Multi-turn
- Contextual discussions
- Real-time, synchronous argumentation

Russia 's relations with the West are a perennial topic at the press conference , which gives foreign journalists a rare chance to directly ask a question of Putin -- and gives Putin a chance to portray Russia , as he often does , as a country under attack from ill-wishers abroad .

Claim: Putin Lived In Russia.

You answered true while your partner answered false.

Partner

I think the answer should be false.  
I thought this was an example of the NatAff rule here, as he is a national public official in Russia. But then I realized that it doesn't say that anywhere in the sentence so under NoOutsideInfo rule, I have to choose False

2

Me

Don't you think that the sentence implies he is speaking for Russia, therefore he holds office in Russia?

Partner

I think based on the ruleset, we aren't supposed to make inferences like that

Partner

A claim is only true if it can be inferred solely by reading the sentence

Alright, you convinced me, that definitely makes sense!

Send Message

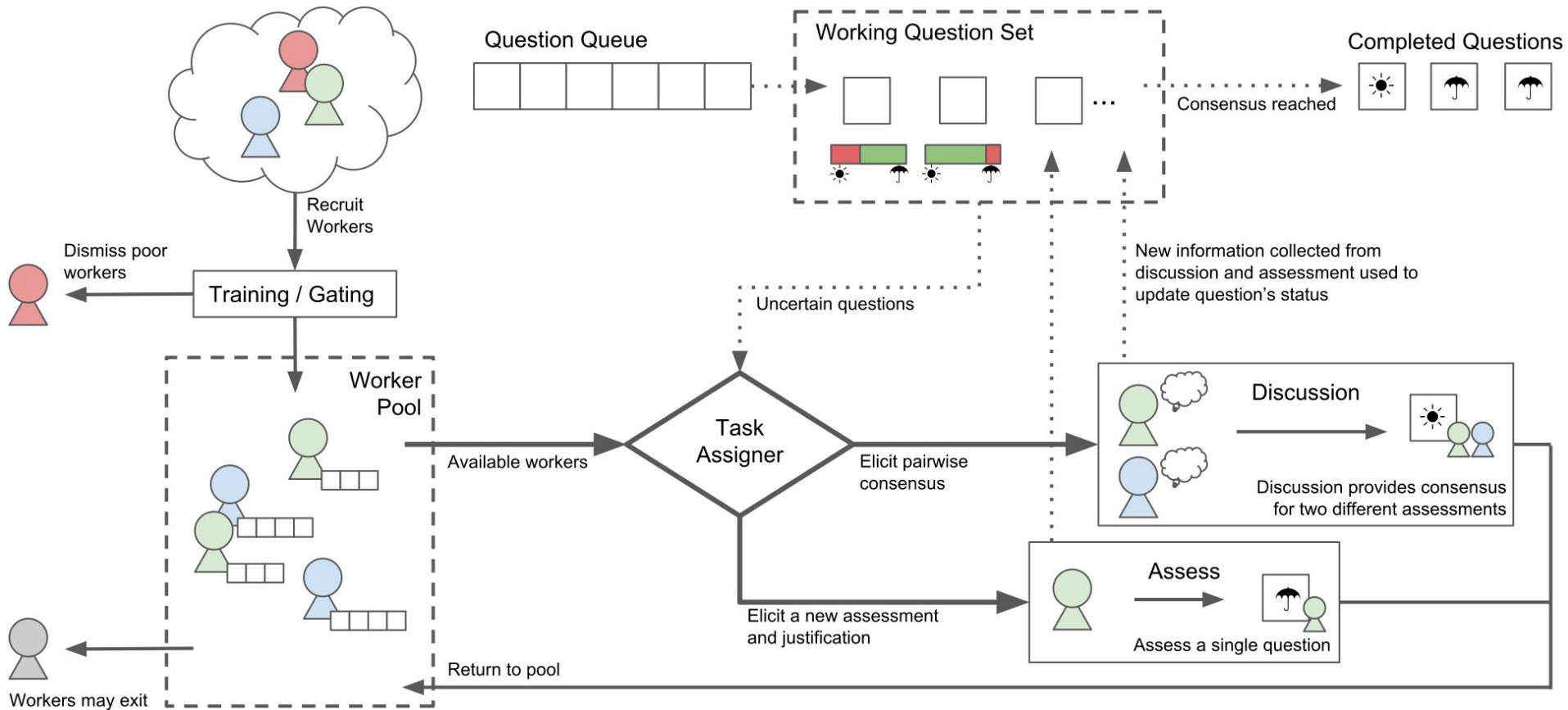
3

End the discussion and:

Agree with partner's judgment (false)

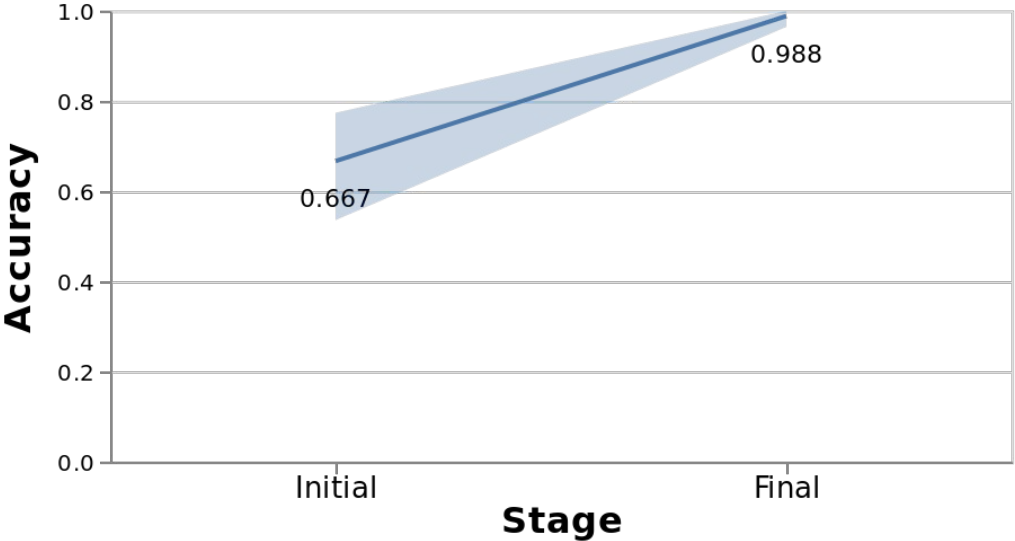
Keep my original judgment (true)

# Cicero

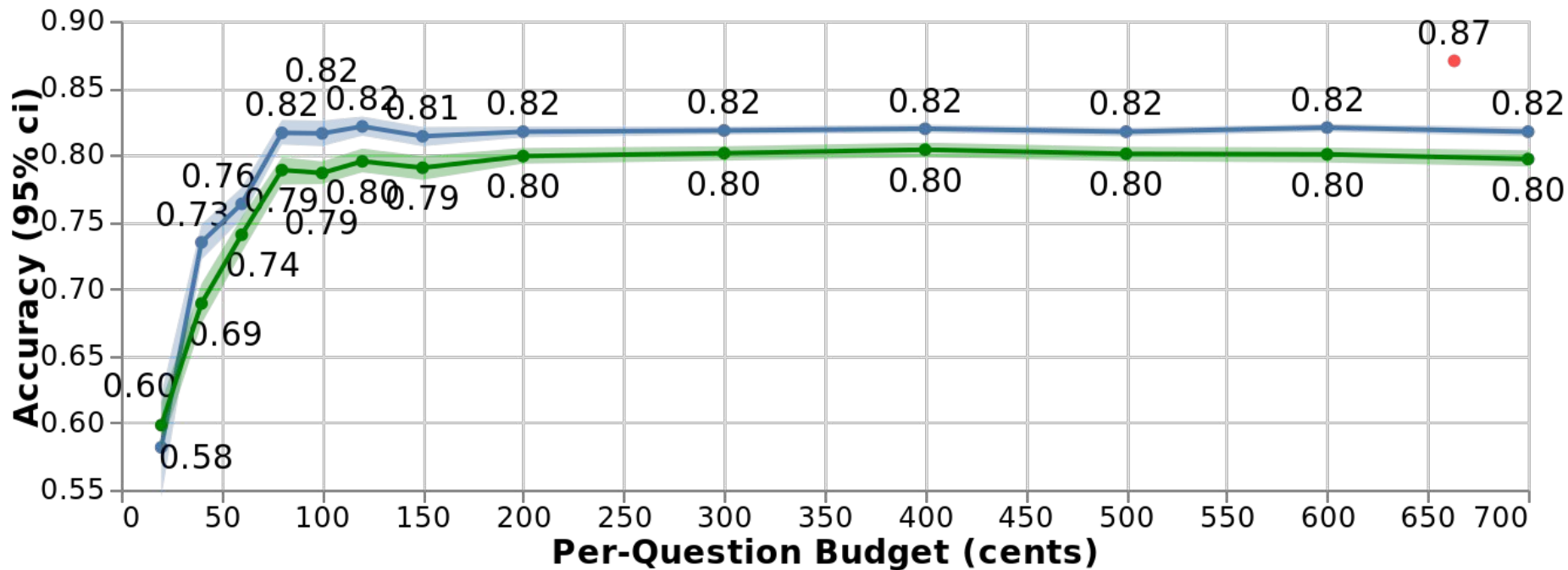




# Results



# Cost-to-benefit Analysis



**Green:** one-shot Microtalk (majority vote)

**Blue:** one-shot Microtalk (EM)

**Red:** multi-turn Cicero (EM)

# Discussion

1. Given the cost-to-benefit graph of multi-turn argumentation compared to single-shot argumentation (MicroTalk), in what applications do you foresee multi-turn argumentation being the most useful?
2. The authors suggest that a significant amount of time and therefore pay can be reduced by implementing Cicero in an asynchronous or semi-asynchronous manner. How do you think that would work? What are some of the challenges and benefits?