A Needle in a Haystack:

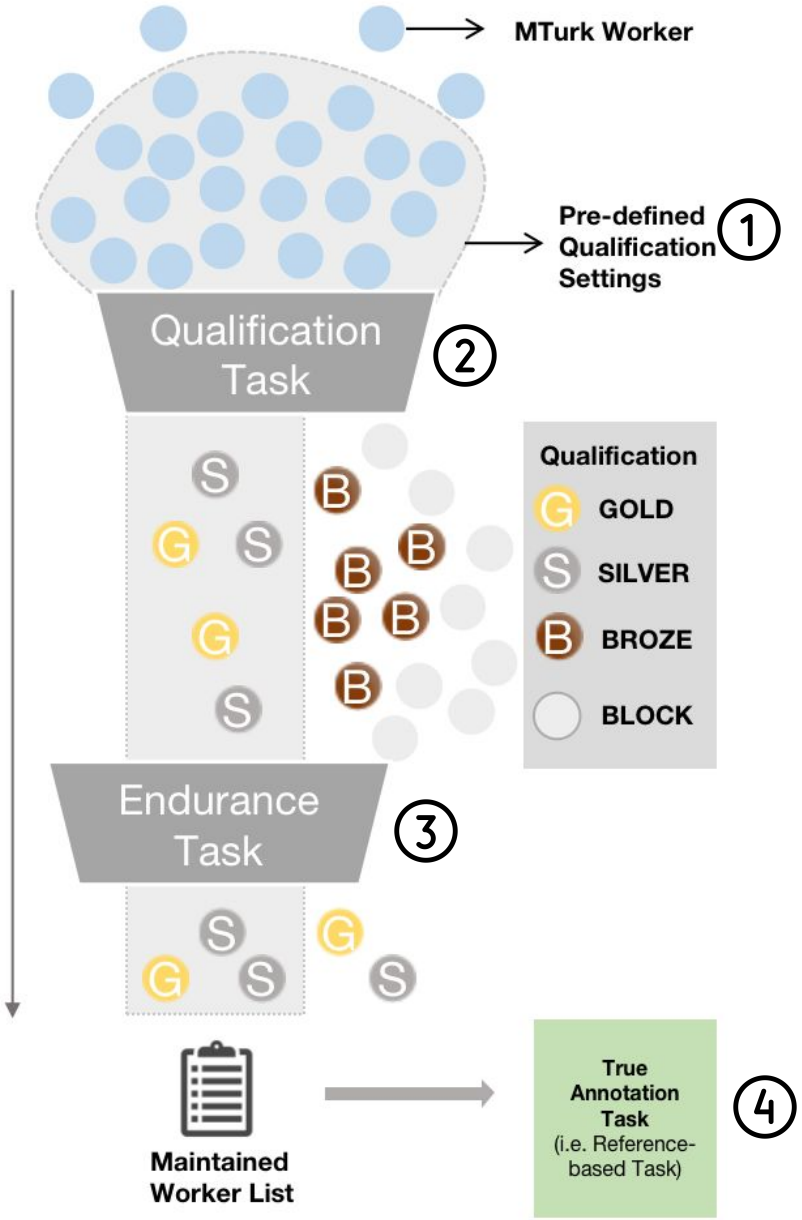
An Analysis of High-Agreement Workers on MTurk for Summarization

Lining Zhang,[1]* Simon Mille,[2] Yufang Hou,[3] Daniel Deutsch,[4] Elizabeth Clark,[5] Yixin Liu,[6]
Saad Mahamood,[7] Sebastian Gehrmann,[5] Miruna Clinciu,[8] Khyathi Chandu,[9] João Sedoc[1]
[1]New York University, [2]ADAPT Centre, DCU, [3]IBM Research, [4]Google, [5]Google Research, [6]Yale University, [7]trivago N.V.,
[8]University of Edinburgh, [9]Allen Institute for AI

NYU GEM

Motivation of Pipeline

- Automatic metrics: **problematic**
- **Best practices** for recruitment on *MTurk*[1]: **poorly understood**

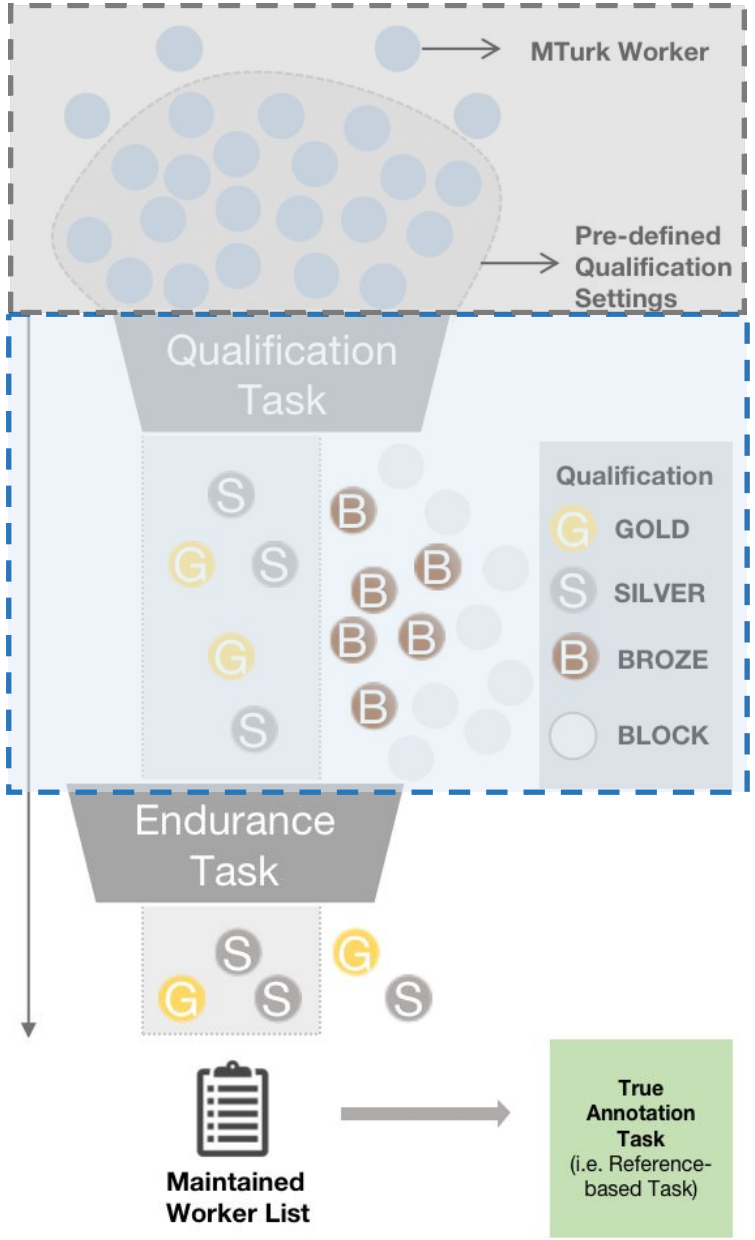


Two-step pipeline for finding high-agreement MTurk workers.

Outline

1. **Qualification Settings**
2. **Qualification Task**
3. **Endurance Task**
4. **Reference-based Task**
   - Baseline MTurk Workers (i.e. MACE)
   - CloudResearch MTurk Workers
   - Analysis of Correctness Across Annotation Sources
5. **Conclusion and Limitations**

[1]*MTurk: Amazon Mechanical Turk*



Two-step pipeline for finding high-agreement MTurk workers.

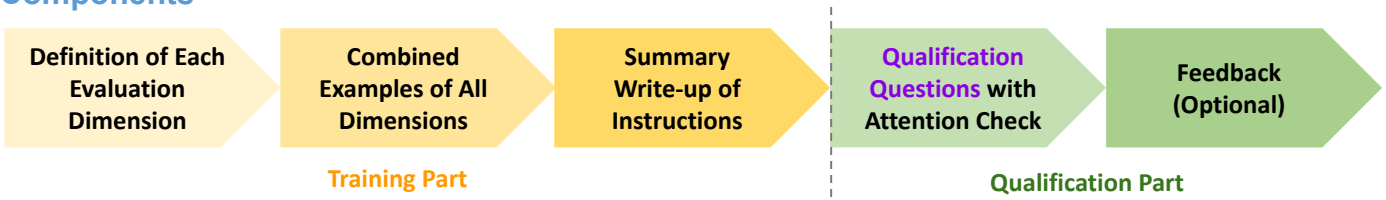
1. MTurk Qualification Settings

Pre-task **qualifications** on workers can be set:

- *Location **(US)***
- *Number of HITs[2] Approved **(>1,000)***
- *HIT Approval Rate (%) for all Requesters' HITs **(>=99)**, etc*

2. Qualification Task

❖ Components



Components of qualification task.

- **Designed Motivation:** evaluate multiple dimensions **correctly**
- **3 documents** (1 w/ **attention check**), **1 summary** each, **6 dimensions**

❖ Worker Categorization

Categorize workers into **4 types**:

GOLD: *all correct + attention check passed* ✅

SILVER: *all but 1 correct + attention check passed* ✅

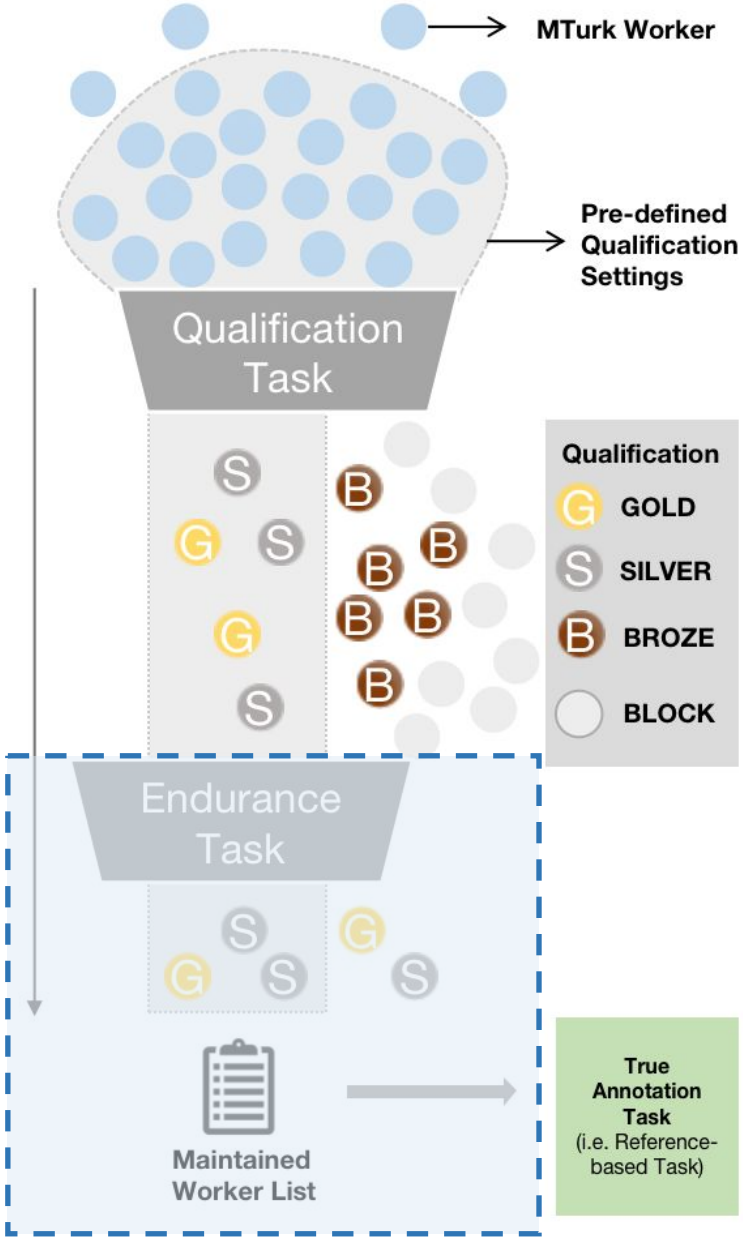
BRONZE: *attention check passed*

BLOCK: *attention check not passed*

❖ Results

- **26 (8 GOLD, 18 SILVER)** MTurk workers (**13%** of 200 participants) qualified

[2]*HIT: Human Intelligence Task*



Two-step pipeline for finding high-agreement MTurk workers.

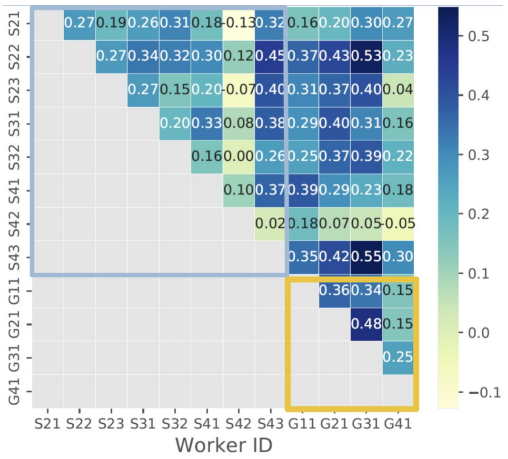
3. Endurance Task

❖ Components

- **Designed Motivation:** capacity for handling **heavy workload**
- **10 HITs**, **1 document** and **4 summaries** each, **saliency**

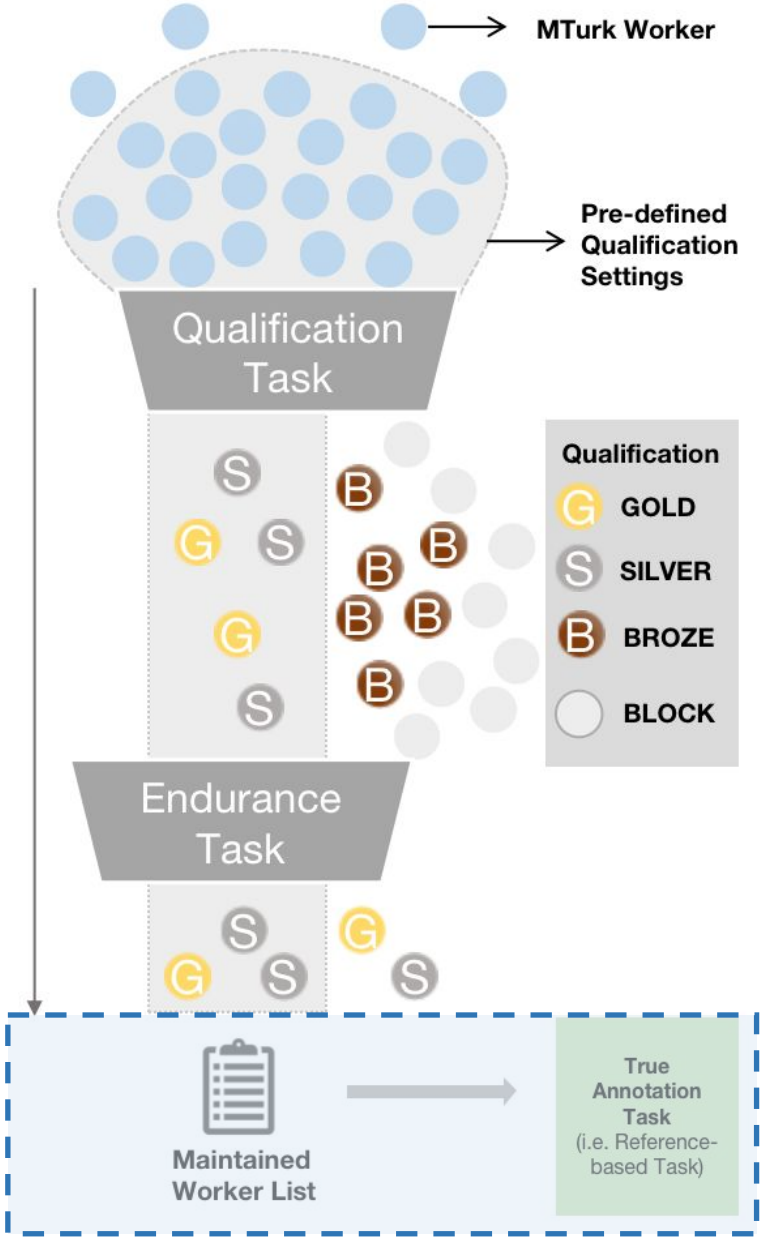
❖ SILVER & GOLD Workers Results

- **12** (**4 GOLD, 8 SILVER**) MTurk workers (**6%** of 200 participants) passed
- Achieved **high agreement (IAA[3])** than experts
  - *Best Cohen's Kappa:* **0.55** (Across Groups)
  - *Best Krippendorff's Alpha:* **0.443** (GOLD)



Cohen's Kappa for endurance task (grey: SILVER; yellow: GOLD)

[3] *IAA: Inter-annotator Agreement*



Two-step pipeline for finding high-agreement MTurk workers.

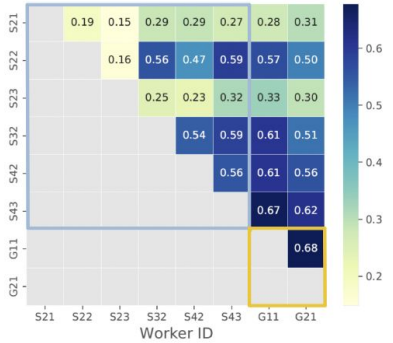
4. Reference-based Task

❖ Components

- **Designed Motivation:** test the **general performance** on true task
- **30 HITs**, **1 reference** and **4 candidate summaries** each, **information** coverage (**2 directions**)

❖ Qualified Pipeline Workers

- **8** (out of **12**) MTurk workers finished all HITs
  - *Best Cohen's Kappa:* **0.68** (GOLD)
  - *Krippendorff's Alpha:* **0.534** (all scores)



Cohen's Kappa for reference-based task (Pipeline).

❖ Baseline MTurk Workers

- IAA with Median
- Filter on Timing and Number of Finished HITs
- Statistical Filter **(MACE[4])**
  - *Krippendorff's Alpha* (threshold=0.5): **0.380**
  - **Incomplete** HIT coverage & **fewer** workers per HIT

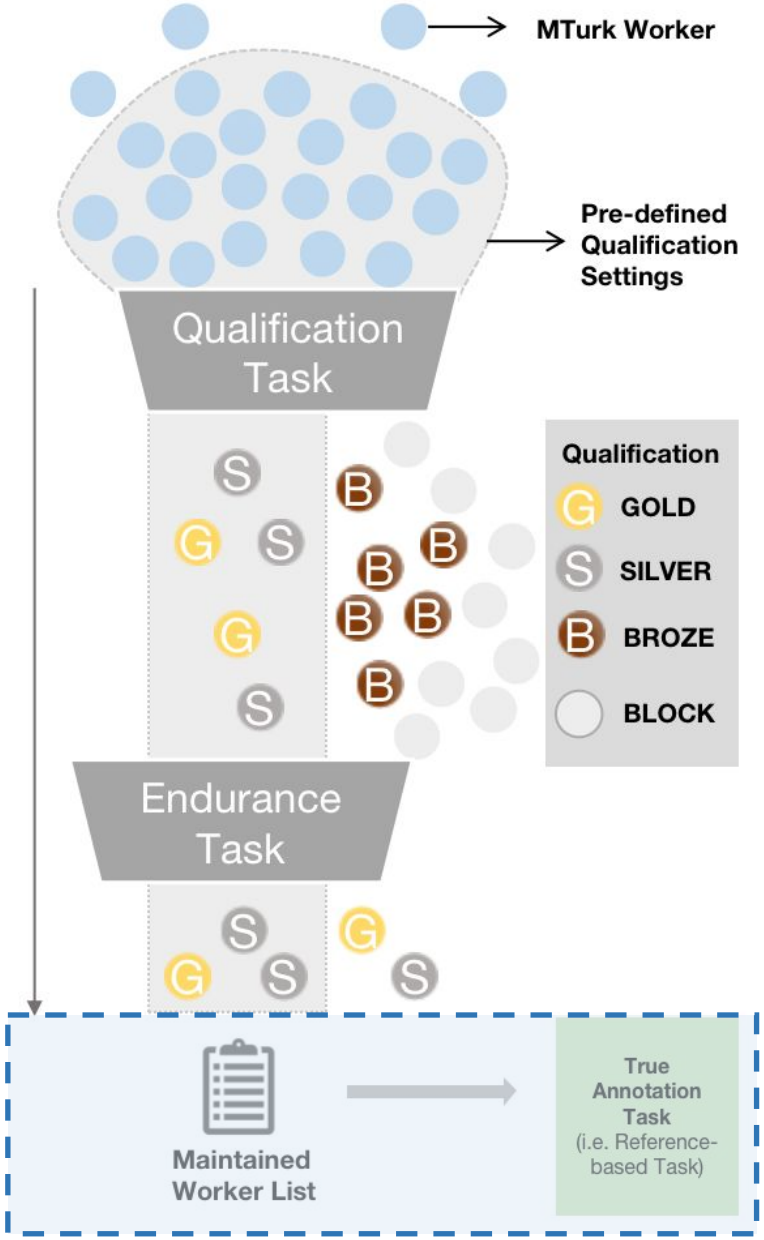
Threshold	0.5	0.6	0.7
% of workers kept	19.2%	15.9%	7.6%
HIT coverage	30/30	27/30	18/30
Avg. num. workers per HIT	2.4	1.9	1.2
Krippendorff's Alpha (all scores)	0.380	0.472	0.754
Spearman's coefficient (MACE workers)	0.351	0.414	0.770
Spearman's coefficient (pipeline workers)	0.558	0.565	0.577

IAA for different thresholds of MACE.

❖ CloudResearch (cloudresearch.com) MTurk Workers

- Platform to recruit high-quality annotators
  - *Krippendorff's Alpha:* **0.513**
  - **lower** task acceptance rate

[4]*MACE: Multi-Annotator Competence Estimation*



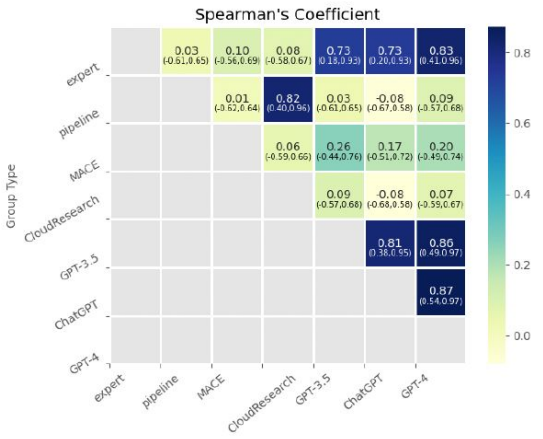
Two-step pipeline for finding high-agreement MTurk workers.

4. Reference-based Task (Continued)

❖ Analysis of Correctness Across Annotation Sources

50 random samples from the reference-based task:

- **Pipeline and CloudResearch** workers had a significant Spearman's **correlation**
- Pipeline may **not guarantee** the training of the **correctness**
- **GPT** models correlated well with **expert judgments**



Spearman's coefficient with 95% confidence interval on 50 samples.

❖ Discussion

Pre-task filtering of our pipeline:

- **avoid the waste** of time and resources (MACE)
- achieve **high agreement** at a **lower cost**
- **similar quality** (Spearman's correlation) to CloudResearch

	Pipeline	MACE (0.5)	CloudResearch
Num. of initial workers	200	276	45
% of workers kept	4%	19.2%	17.8%
HIT coverage	30/30	30/30	30/30
Avg. num. workers per HIT	8	2.4	8
Krippendorff's Alpha	**0.534**	0.380	0.513
Cost per worker (for Avg. num. workers per HIT)	**$27**	$175	$31

Comparison between approaches of crowd annotators.

5. Conclusion and Limitations

Conclusion

Pipeline result:
200 MTurk workers --> **4 GOLD, 8 SILVER (6%)**

Serves as the **best practice**:

- **high-agreement** annotations at **large scale** and **lower cost**
- **avoid resource waste** on discarded annotations

In the future:

- **high-quality** (high agreement; correctness)
- **multiple application** (tasks, languages, and platforms, etc)

Limitations

- **English summarization** on **MTurk** platform
- **Designed questions** not "panacea" solutions
- **No guarantee** for the training of **correctness**

Acknowledgement

Thank Google for the experiment fundings

NYU GEM


NYU
Thank You!