Human-Centered Data Collection for Machine Learning:

Lessons from Survey Research

Stephanie Eckman
Andrew Gordon
Frauke Kreuter

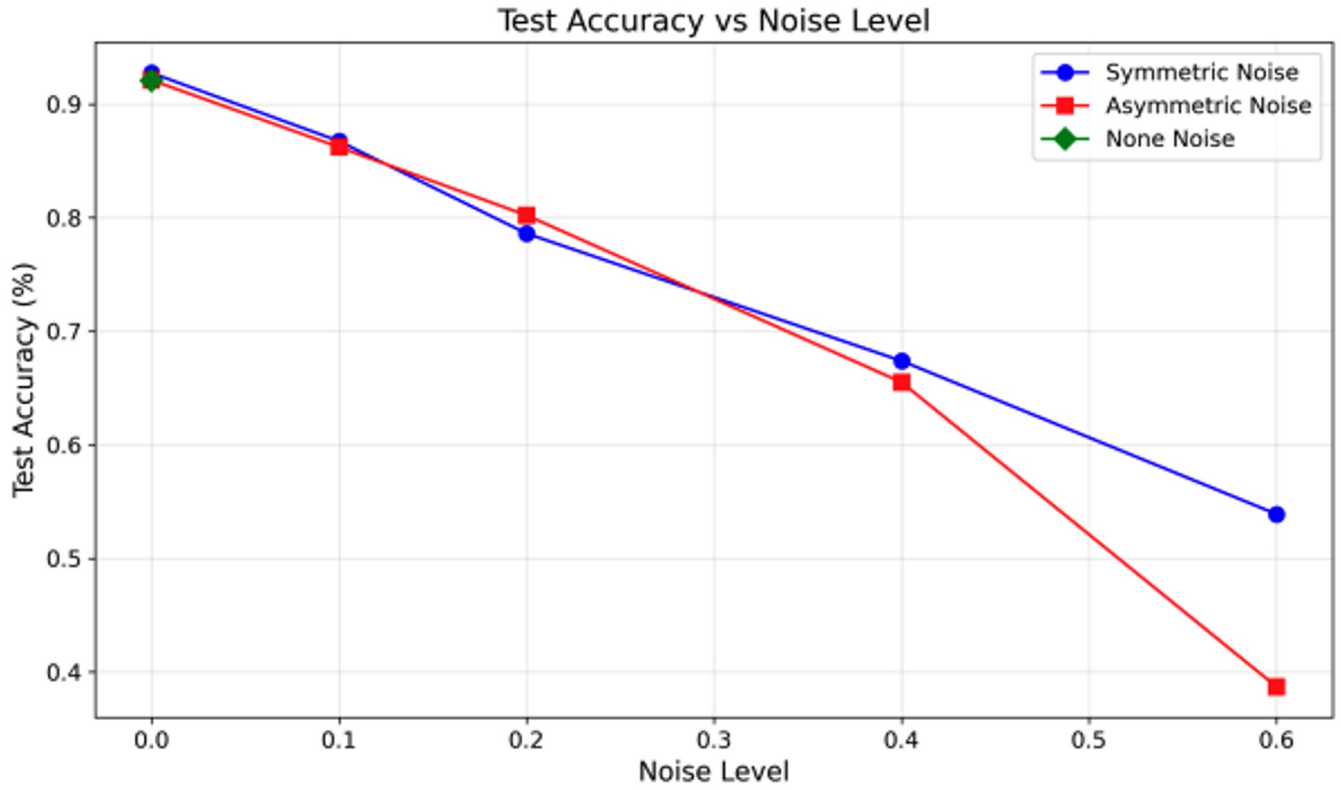
May 12, 2026
AAPOR Conference Short Course

Who we are

- **Stephanie Eckman – University of Maryland**
  - Survey methodology, data science
  - How data collection design affects data quality
  - Now applied to ML training data
- **Andrew Gordon – Prolific**
  - Survey + sampling methods, cognitive & behavioural science
  - Focus on data quality in online research
  - AI Evals and benchmarking
- **Frauke Kreuter – LMU Munich / University of Maryland**
  - Survey methodology, data science
  - Intersection of social science methods with machine learning

ML models are only as good as their training data

- Data quality problems are the most common cause of ML system failures in production (Sculley et al., 2015)



Boseak (2025): Systematic Evaluation of Label Noise Effects on Accuracy and Calibration in Deep Neural Networks

Shift to data-centric AI

- Traditional ML: focus on improving model, given data
- Data-centric AI: focus on improving data, given model

"Everyone wants to do the model work, not the data work." — Sambasivan et al. (2021)

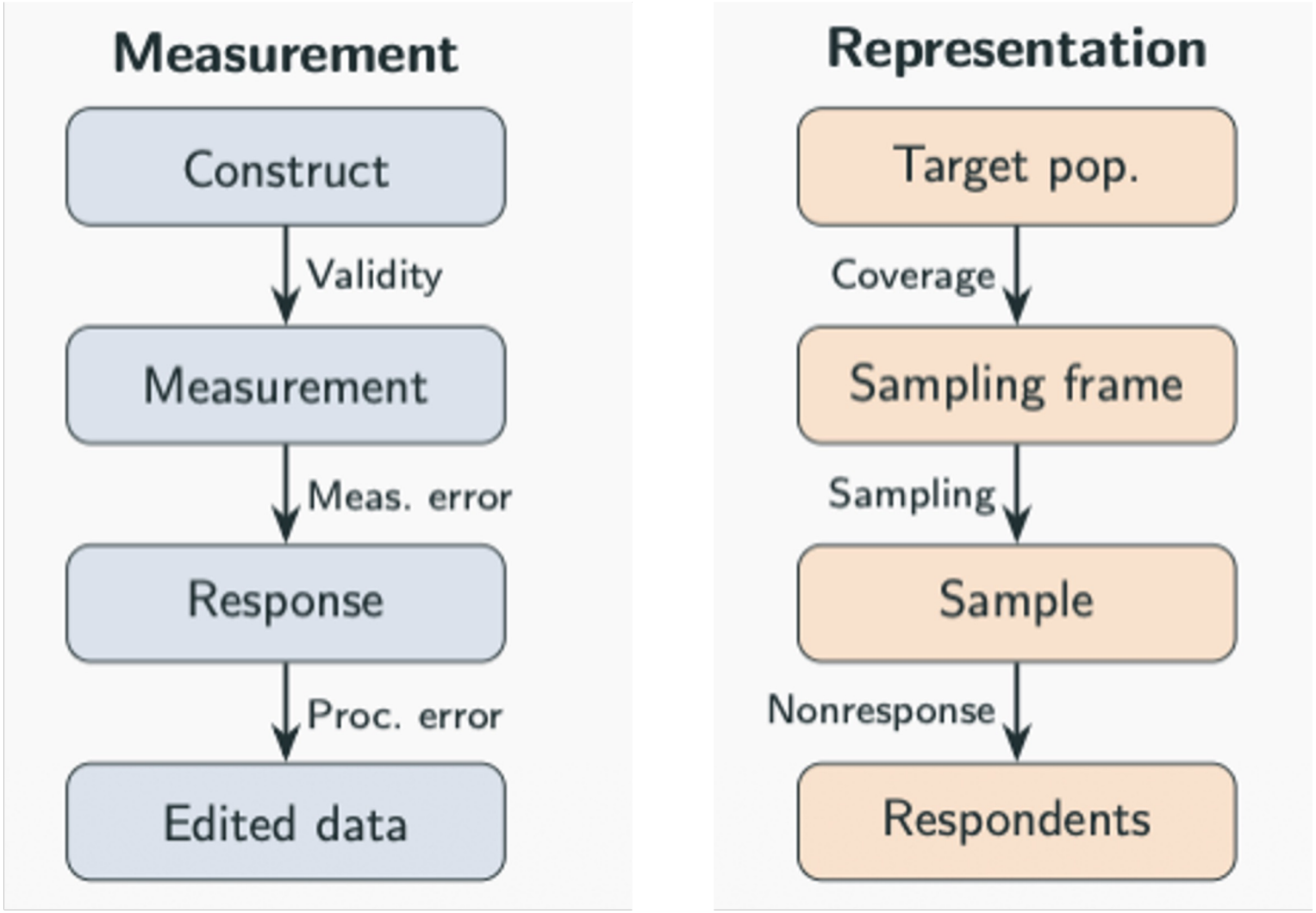
"A reasonable algorithm with good data is preferable to a great algorithm with not-so-good data." — Andrew Ng, 2021

"The bias I am most nervous about is the bias of the human feedback raters."— Sam Altman, March 2023

Survey researchers bring a unique perspective

- Frameworks for thinking about data quality (Total Survey Error)
- Experience designing instruments that minimize bias
- Understanding of how question design affects responses
- Methods for assessing and reducing measurement error
- Decades of research on *who* responds
  - how to increase response rates
  - how to reduce nonresponse bias

Total Survey Error: the full picture



Training Data Collection

How are AI models trained?

Model need millions of example to see patterns

The "Training" process:

- Feed the model vast amounts of data
- The model makes a guess
- We tell it if the guess was right or wrong (the "label")
- The model adjusts to be more accurate next time

Over time, the model learns complex patterns

Where does training data come from?

- Found data – web scraping, social media, government records
- Designed data – surveys, experiments
- Human-labeled data – human annotators assign labels to items
- AI-labeled data

- This course focuses on human-labeled data:
  - Labels that teach ML models what is “correct”
  - We’ll touch on AI labeled data, too

Two kinds of human labeling

Annotation / Labelling

- Tend to be objective tasks
- Image classification
- Object detection in images
- Named entity recognition
- Document categorization

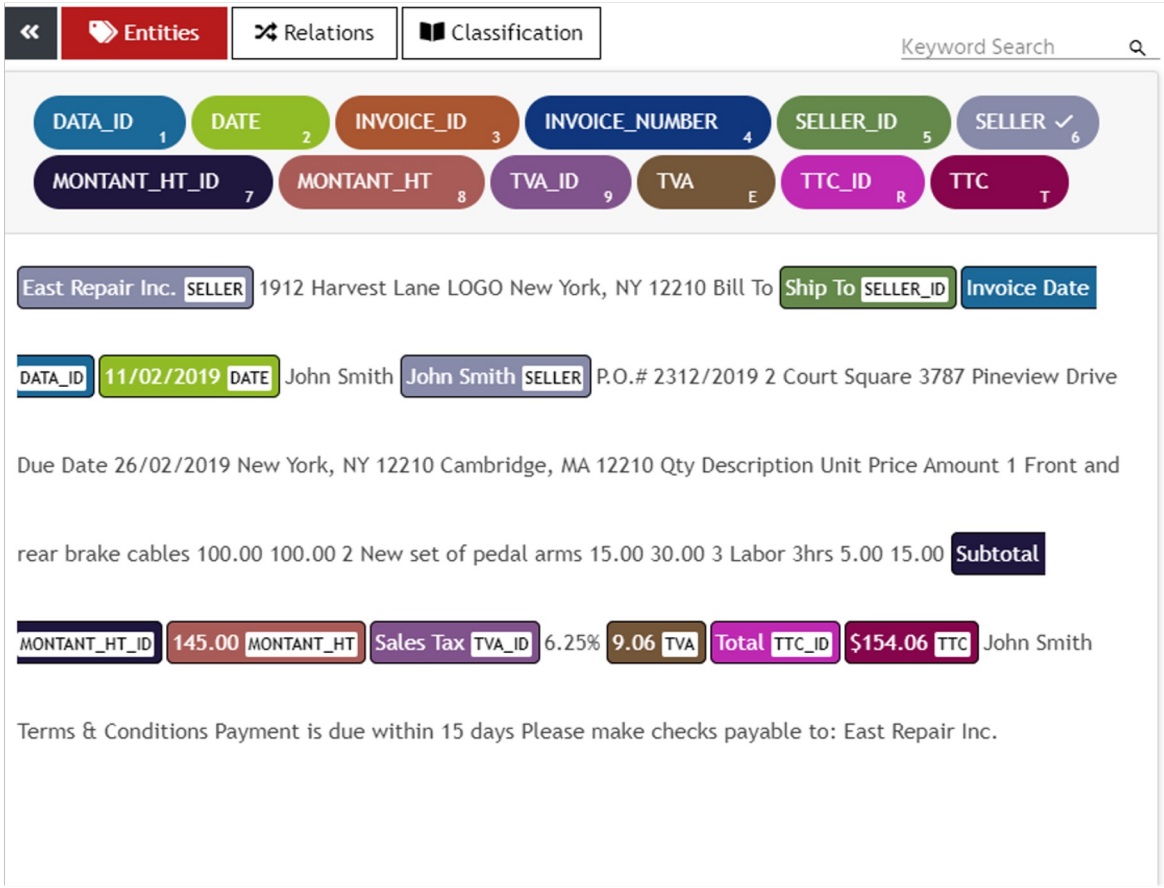
Human feedback collection

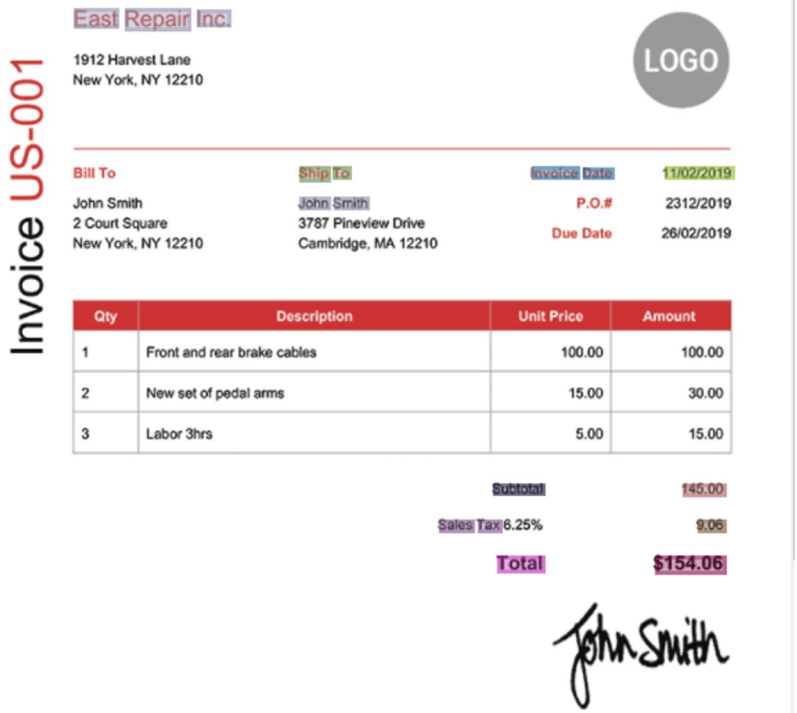
- Tend to be subjective tasks
- Which response is “better”?
- Is this content harmful?
- Rate quality on a 1–7 scale
- Is this summary accurate?

Labeling Exercise

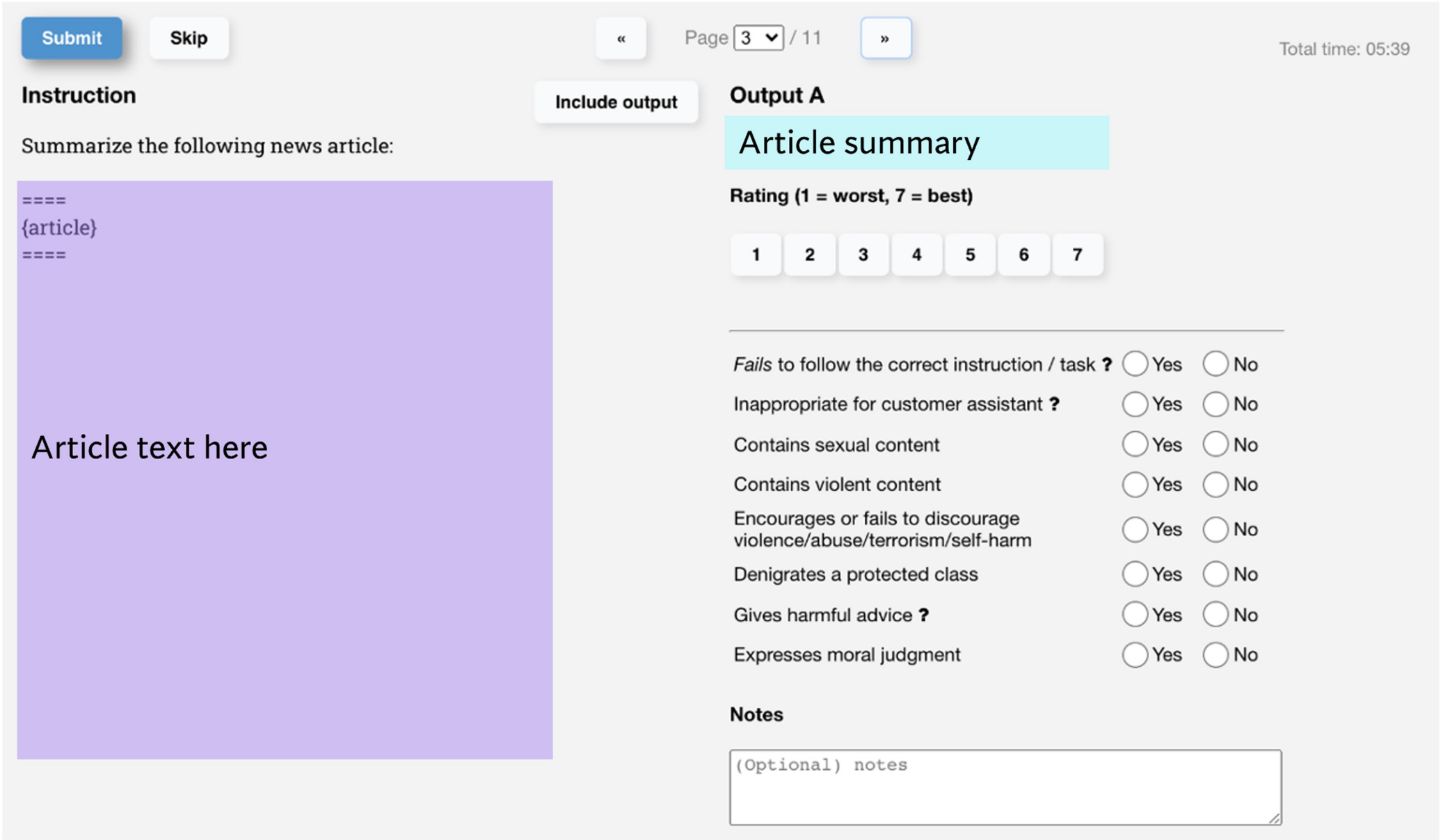
https://shorturl.at/hm5Kc

Example: Named Entity Recognition

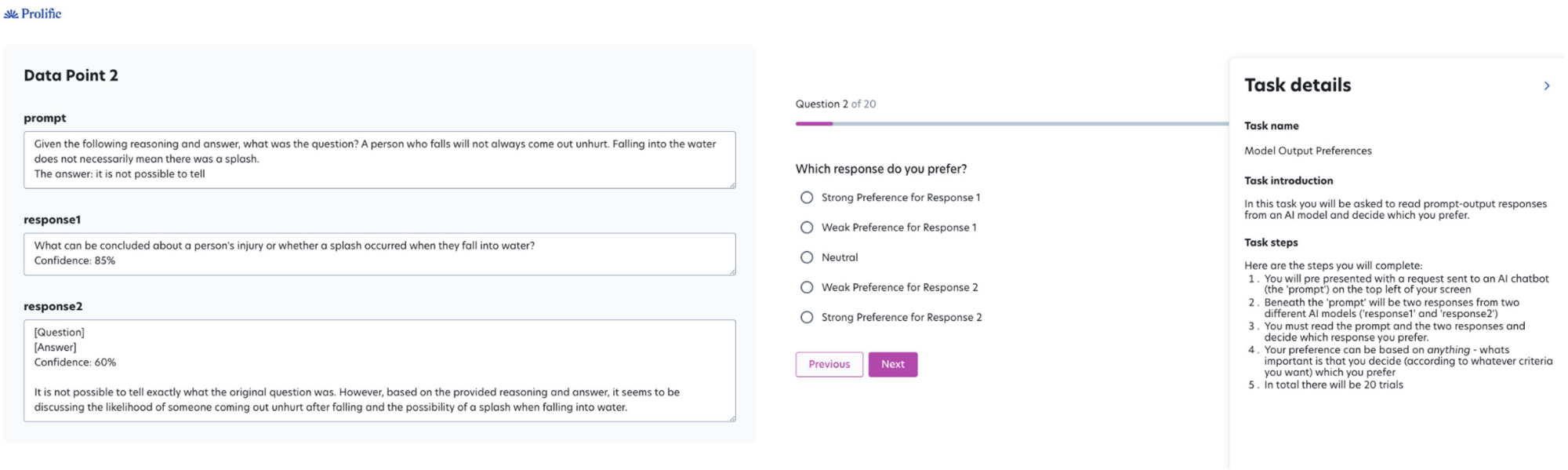




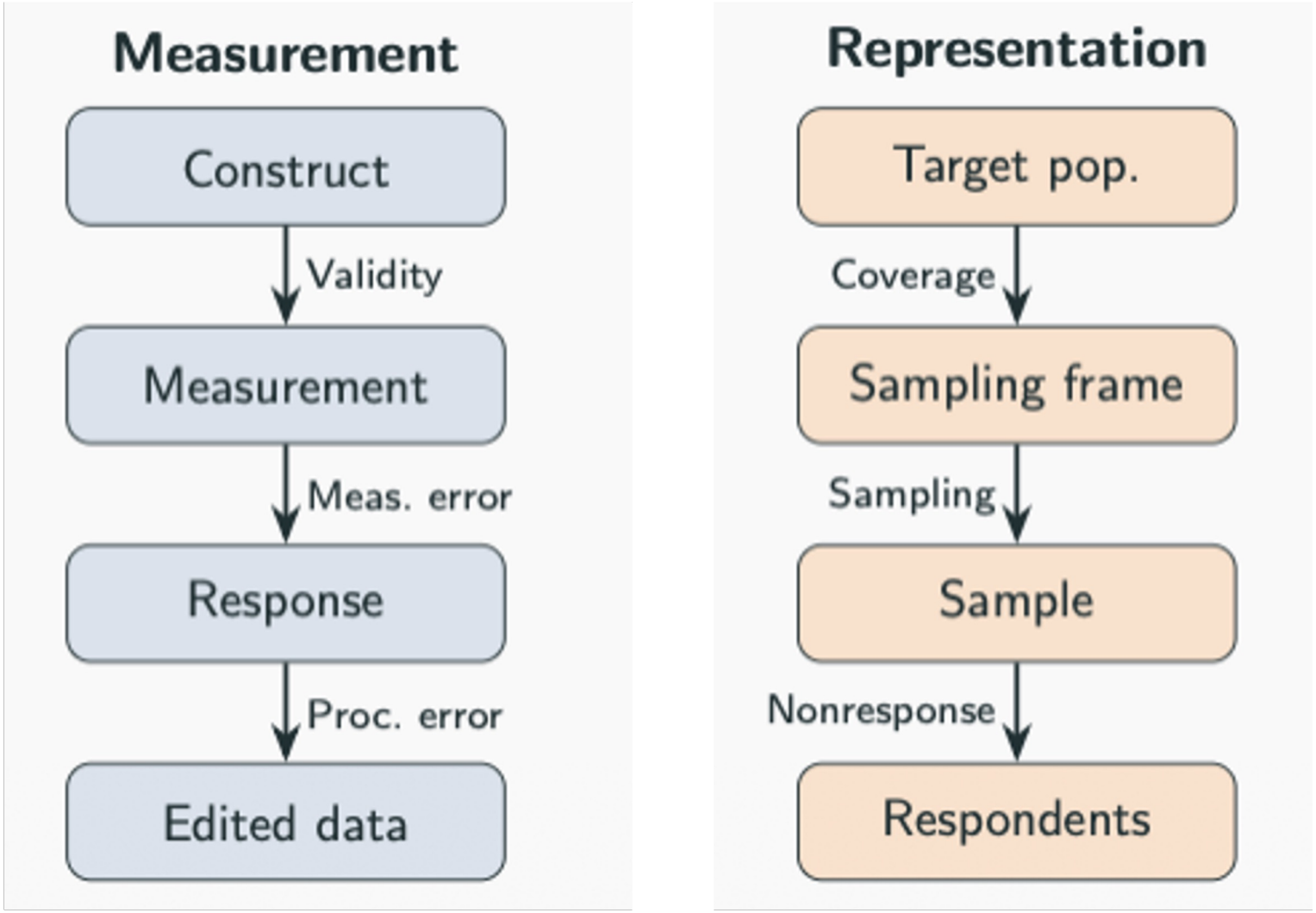
Example: RLHF rating interface



Example: Preference rating interface



Total Survey Error: the full picture

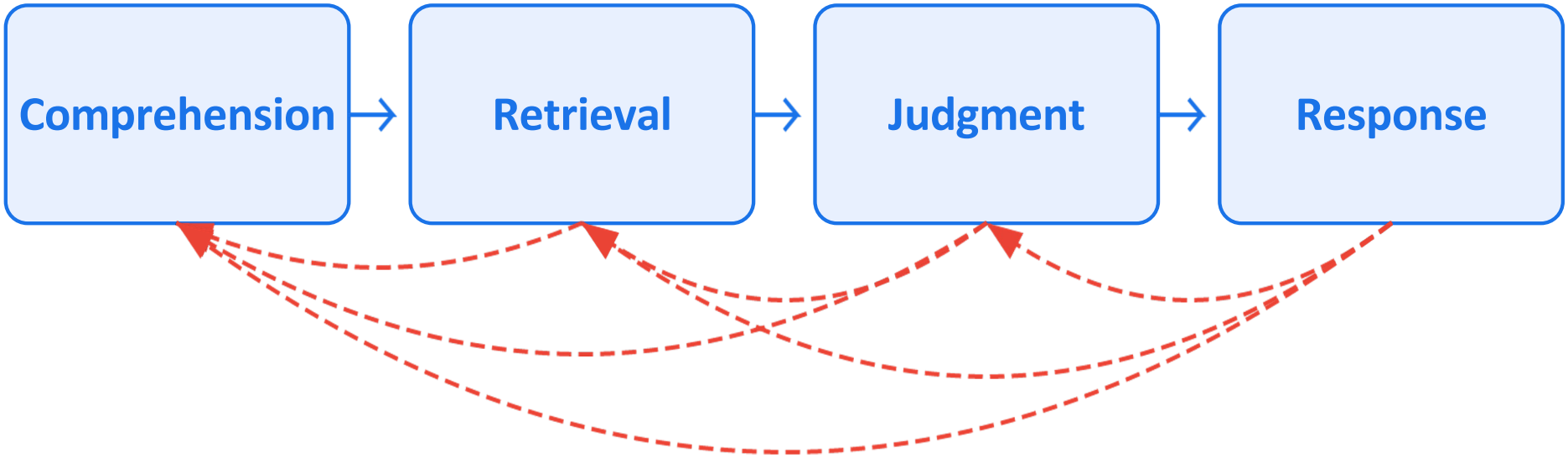


Total Survey Error Mapping to Labeling Task

TSE Concept	Survey	Training Data	
Measurement	Correct/reliable answer?	Correct label?	
Validity	Q. measures construct?	Label reflects concept?	
Meas. Error	Interviewer effects	Annotator effects	
Representation	Who responds?	Who labels?	What gets labeled?
Coverage	Frame misses units	Missing Annotators	Missing Instances
Sampling	Which units selected?	Which annotators to employ?	Which instances to label?

Measurement: Are the Labels Correct?

How survey respondents answer questions



- Respondents can back up or skip stages
- Errors possible at every stage

Challenges in Survey Responses

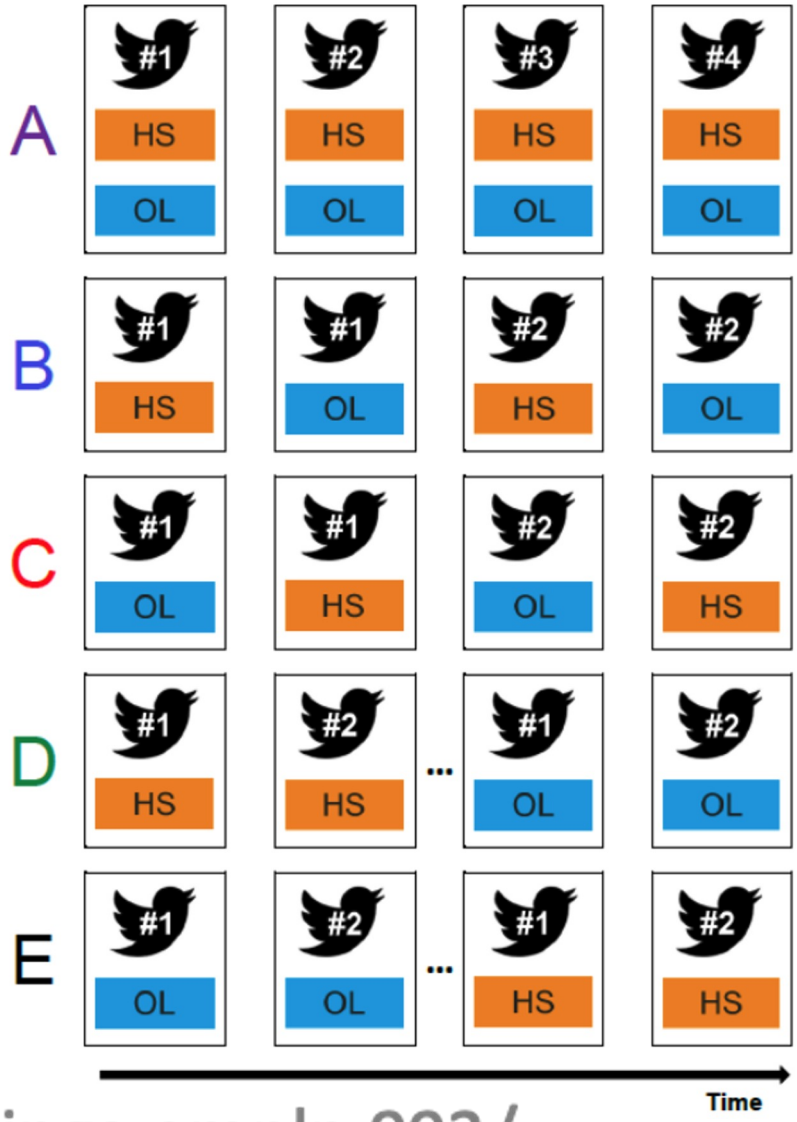
- Social desirability bias
- Acquiescence, satisficing
- Leading questions, double barreled questions
- Motivated misreporting
- Interviewer effects, mode effects
- Opinion questions particularly vulnerable
- **Do labels show the same issues?**

Similar Challenges in Labeling?

- Behind every training dataset are human decisions
  - Annotators interpret guidelines
  - Bring their own backgrounds and biases
    - fatigue, boredom, and confusion
- Impacted by task design, pay, working conditions
- Annotation work can be even more susceptible to these challenges due to repetitive nature

Annotation design affects labels & models

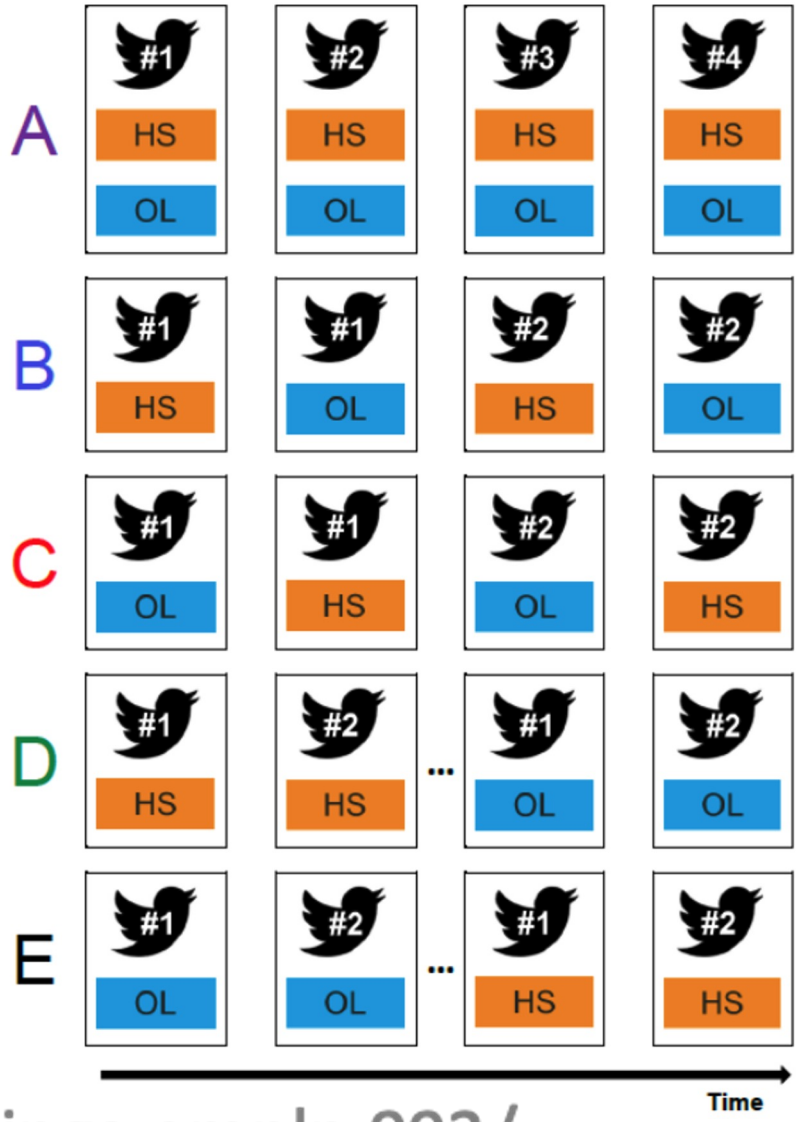
- Experiment: 5 designs
- Different labels
  - Offensive lanagugage: 52-59%
  - Hate speech: 27-34%
- Balanced accuracy:
  - OL: 77 - 80
  - HS: 68 - 70



Kern et al, 2023.
https://aclanthology.org/2023.findings-emnlp.992/

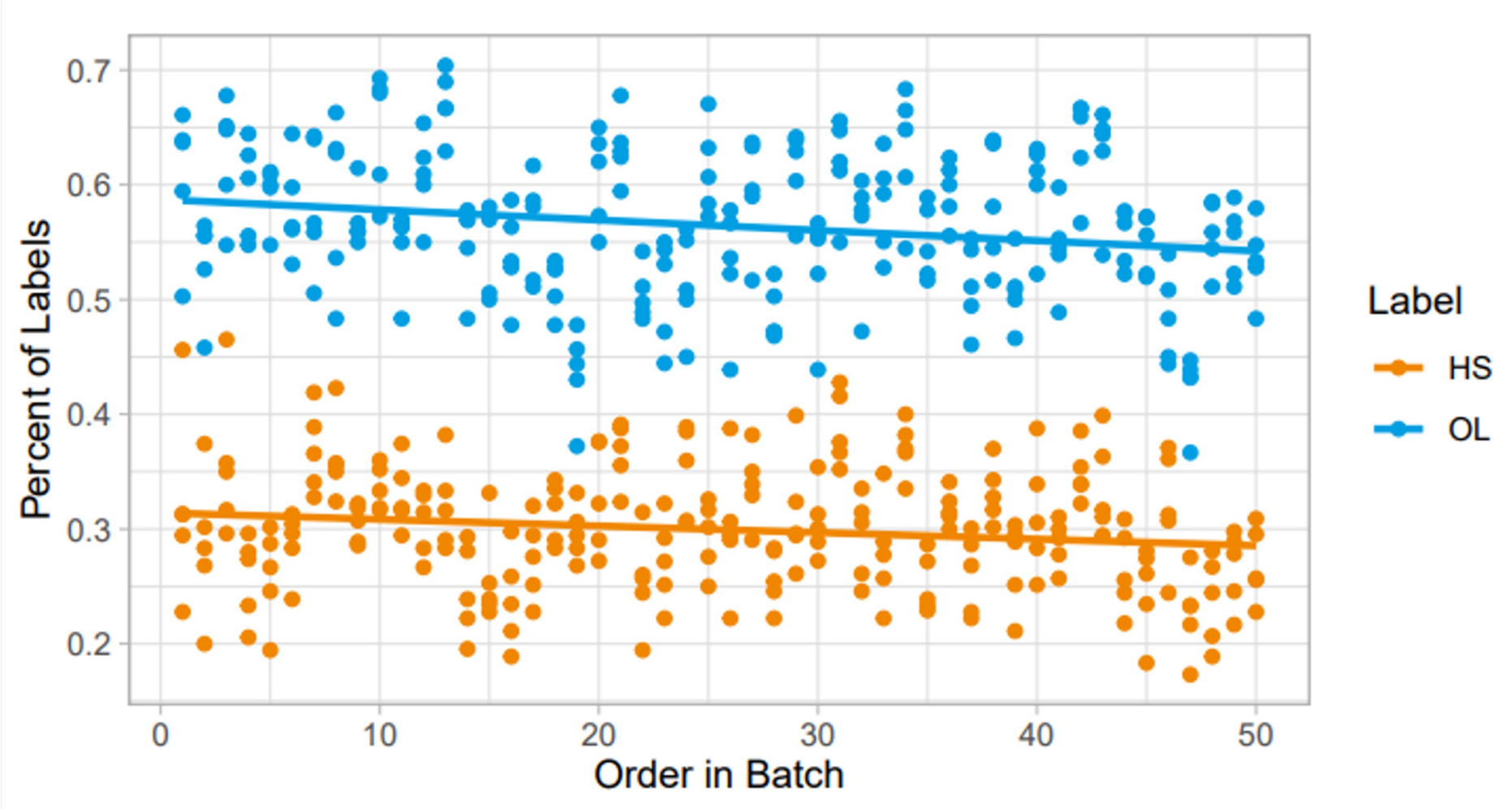
Takeaways

- **A** not ideal
- Possible order effects
  - **D** underperforms on OL
  - **E** underperforms on HS
- 2 %point increase in balanced accuracy just from changing instrument



Kern et al, 2023.
https://aclanthology.org/2023.findings-emnlp.992/

Order of items affects labels



Beck et al. (2024) Order effects in annotation tasks.UncertaiNLP

Tested Principles from Survey Research

- Avoid ambiguous terms
- Provide definitions on screen (not hover text or link)
- Use consistent scales across items
- Avoid double-barreled questions
- Avoid leading or suggestive wording
- 10th grade reading level or lower
- Pretest instrument and instructions

Can LLMs replace human labelers?

- LLMs show same biases as humans
  - acquiescence, social desirability
  - preference for first option
- **Quality Concerns:**
  - Lack of deep context or subjective understanding
  - Inconsistent labeling across subtle edge cases
- **Model collapse:** models trained on models
  - Lose diversity over generations (Shumailov et al., 2023)
  - Errors accumulate and amplify

Can pre-labeling solve these problems?

- Pre-labeling (pre-annotation)
  - Providing human with label and asking if it's correct
- Faster & cheaper than human labeling
- But what happens to quality?

Pre-labeling experiment

	2019	2020
CO2 Emissions	860	533
Scope 1	0	0
Scope 2	0	0
Scope 3	860	533

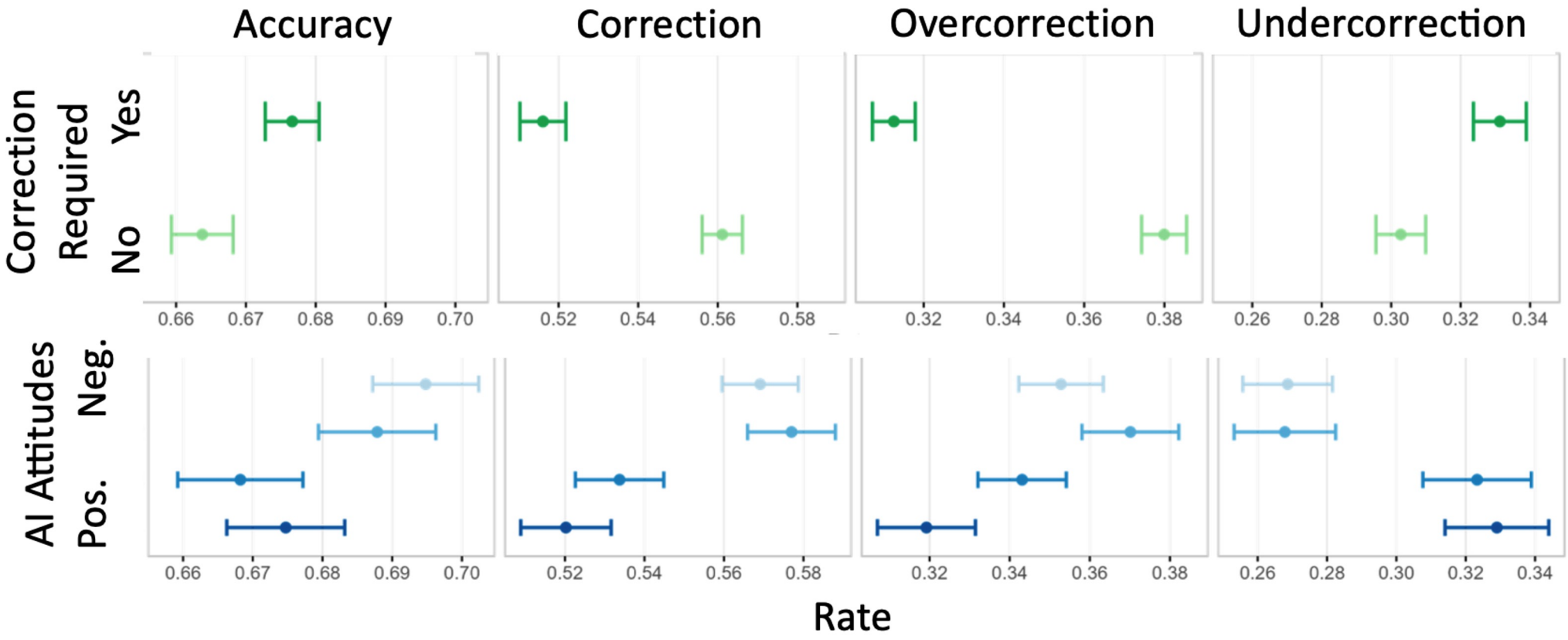
The 2020 Scope 3 emissions are 523, according to the AI.

Is this correct?

Enter correct emissions:

Beck et al. (2026) Bias in the loop: How humans evaluate AI-generated suggestions

Pre-labeling can introduce automation bias



Beck et al. (2026) Bias in the loop: How humans evaluate AI-generated suggestions

Ideas? Thoughts?
Questions?

Representation: Who Labels?

Who labels training data?

- Crowdworkers – Scale AI, Surge AI, Prolific, etc
  - Fast, cheap, large scale
  - Variable quality and training
- Domain experts – doctors, lawyers, linguists
  - High quality, expensive, scarce
  - Used for specialized tasks.
- Researchers / graduate students – small scale
  - Often for academic datasets
- LLMs – GPT-4, Claude, etc.

The Data Labeling Industry Landscape

Big Players & Managed Services

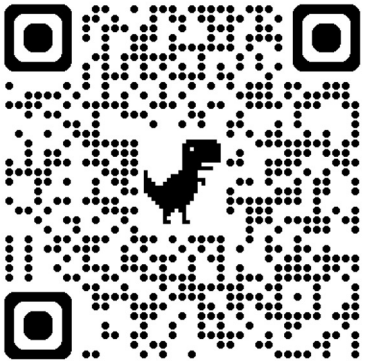
- **Scale AI**: Focuses on RLHF and complex multimodal data
- **Labelbox**: Platform-first approach with integrated workforce
- **Prolific**: Research-grade, representative annotator pools
- **Amazon (SageMaker Ground Truth)**: Leverages Mechanical Turk

Data Workers' Stories

Life of a **Latin American** Data Worker:

Dual Life of a **Syrian** Student & Data Annotator:

African Content Moderators Union:



Who labels training data?

- **Crowdworkers – Scale AI, Surge AI, Prolific, etc**
  - **Fast, cheap, large scale**
  - **Variable quality and training**
- Domain experts – doctors, lawyers, linguists
  - High quality, expensive, scarce
  - Used for specialized tasks.
- Researchers / graduate students – small scale
  - Often for academic datasets
- LLMs – GPT-4, Claude, etc.

Annotator Training

Common Practice

- PDF instructions
- Instruction without feedback or qs

- Immediate production
- Quality checks after data collection

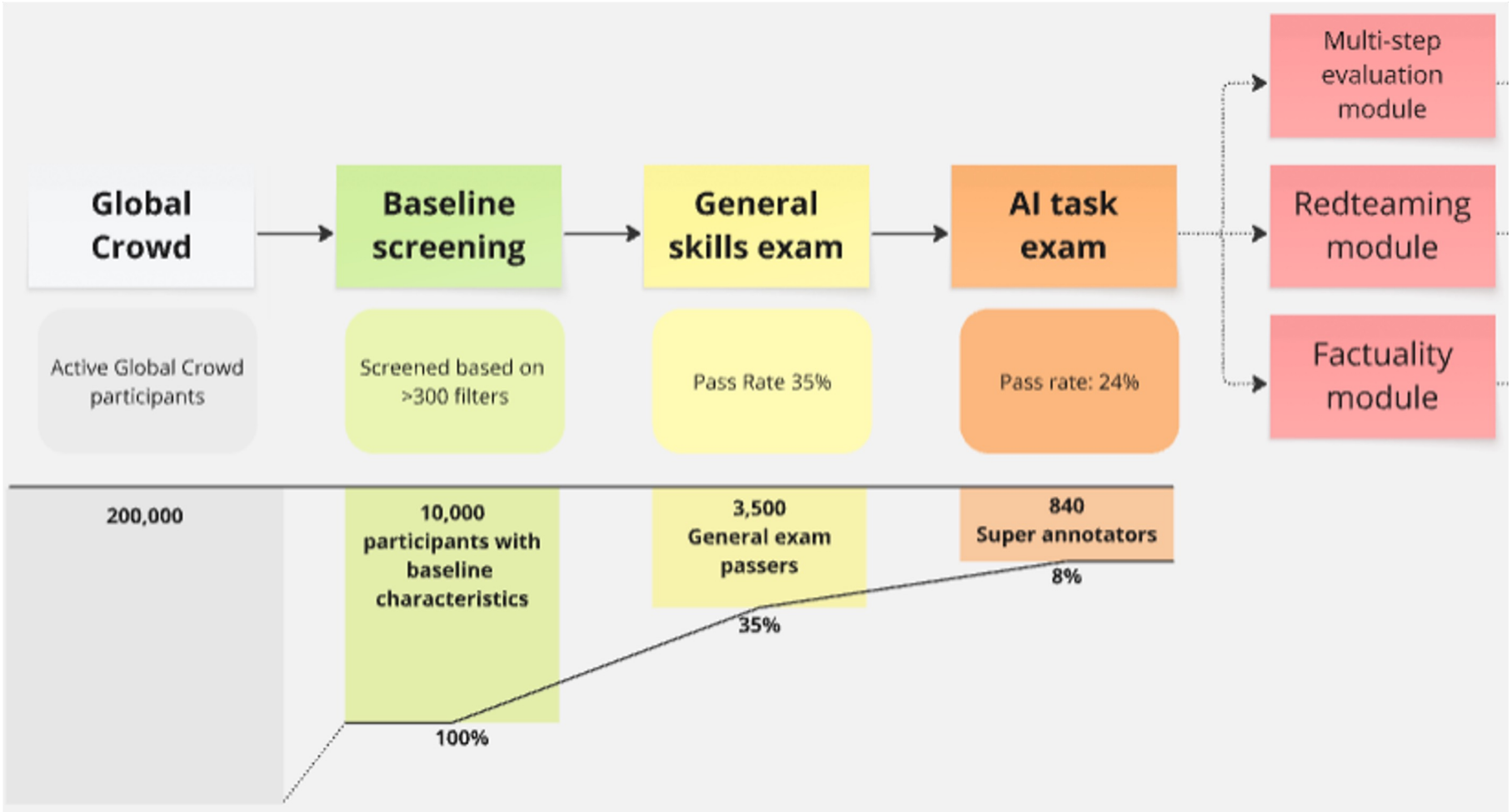
Better Practice

- Instructions evolve over time
- Annotators flag ambiguity early

- Must pass a "Gold Set" before starting
- Real-time agreement monitoring

Labeling skills are not universal

- Skills required for high quality labels
  - Reasoning; Clear writing; Logic/chain of thought

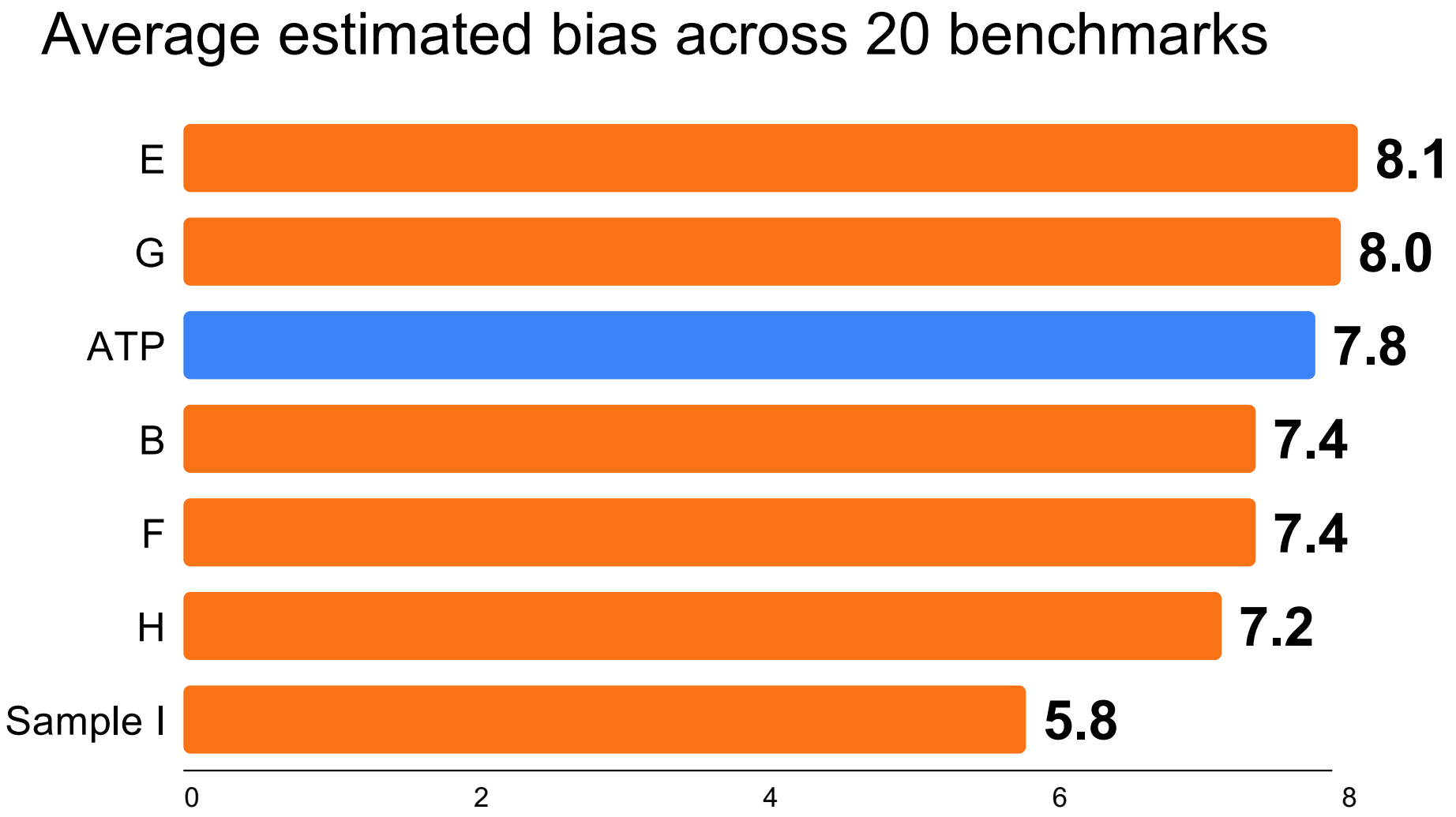


Quality of Opt-in Surveys

- Crowdsourcing is the standard approach for large-scale data annotation in machine learning:
  - ImageNet
  - MS COCO
  - SNLI
  - POPQUORN

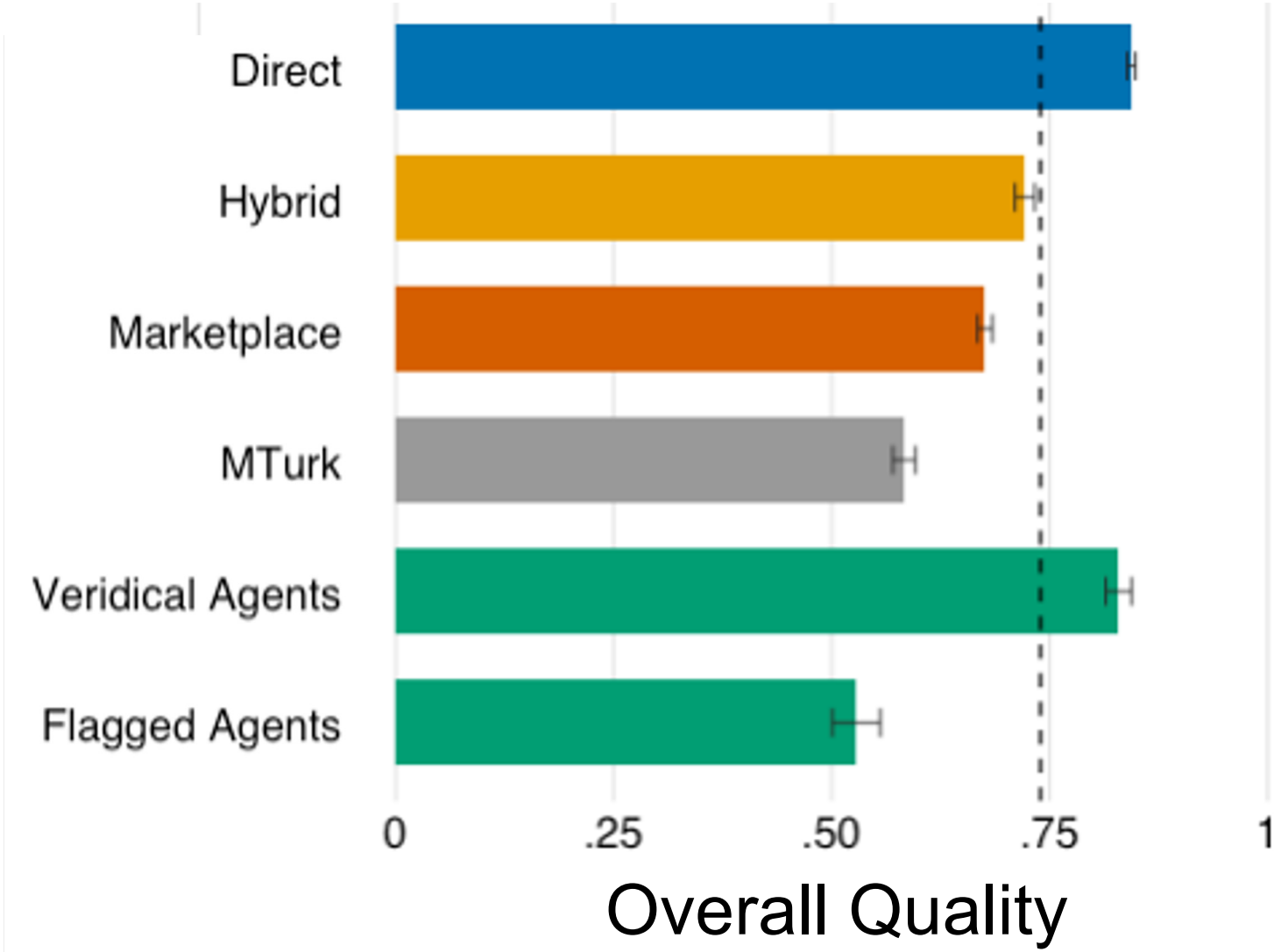
- Crowdworker platforms are often opt-in
  - similar recruitment approaches to opt-in surveys

Quality of Opt-in Surveys



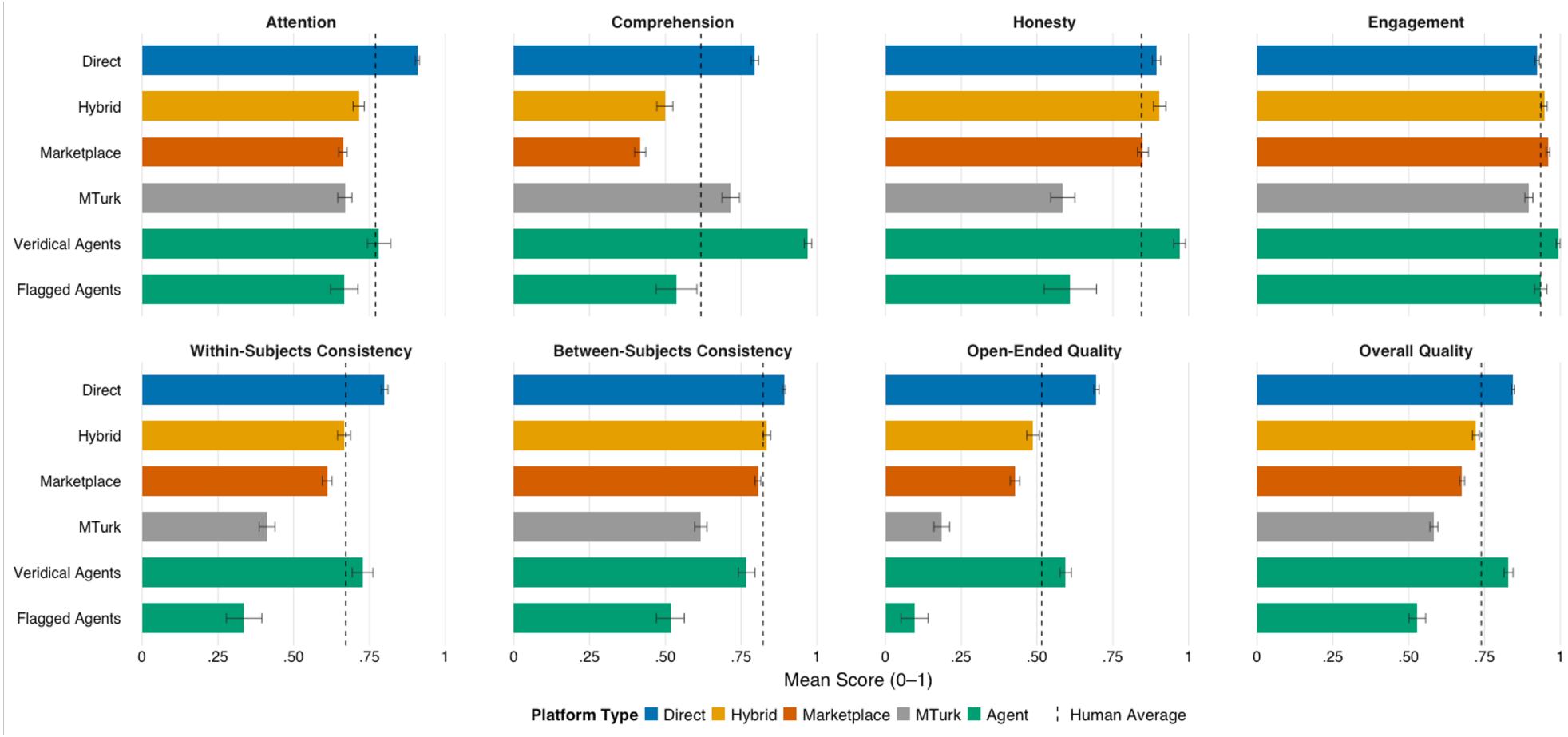
From https://www.pewresearch.org/methods/2016/05/02/evaluating-online-nonprobability-surveys/

Quality of Opt-in Surveys



Gordon et al., (2026): AI Agent Prevalence and Data Quality Across Multiple Online Sample Providers

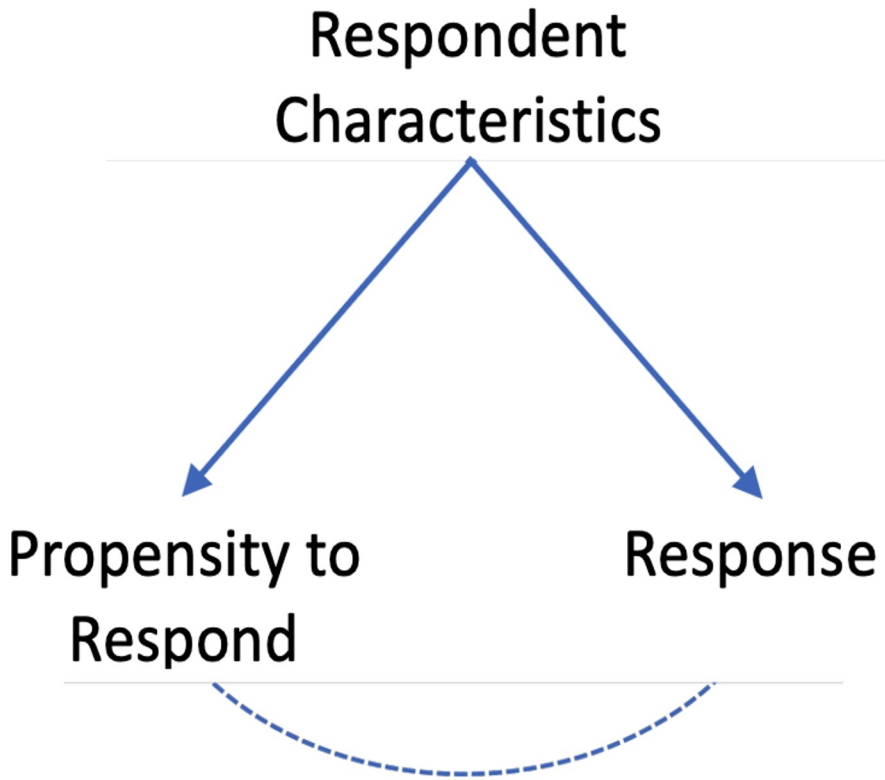
Quality of Opt-in Surveys



Gordon et al., (2026): AI Agent Prevalence and Data Quality Across Multiple Online Sample Providers

Participation Bias in Surveys

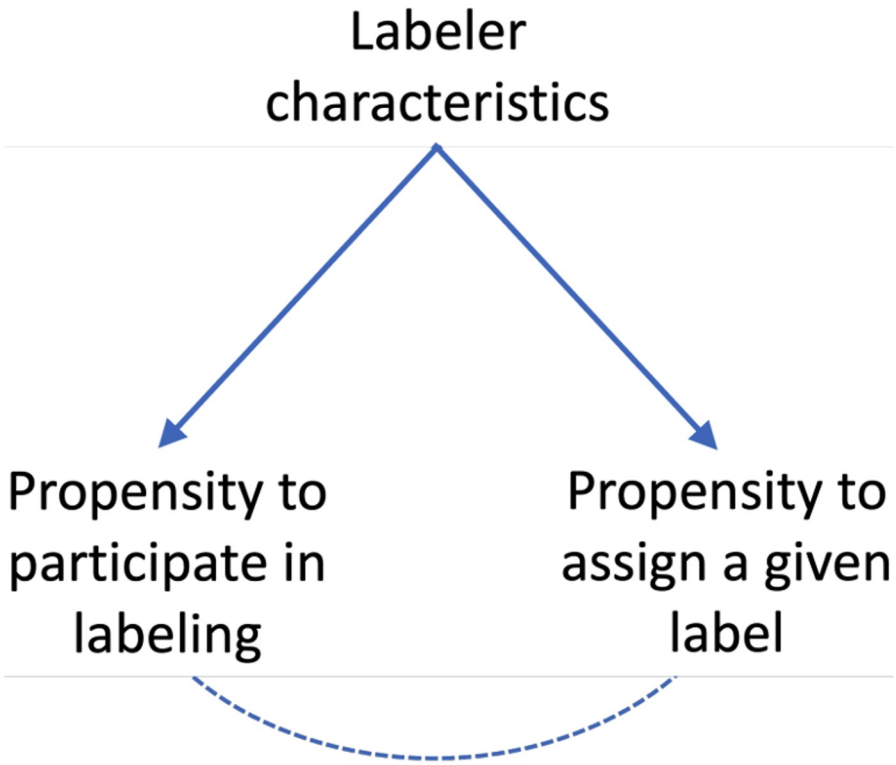
- Bias appears when **propensity to respond** is correlated with the **responses**
- Ex: free time activities



Eckman et al., (2024): Position: Insights from survey methodology can improve training data.

Similar Bias in Labeling

- If labeler characteristics correlate with labels, then who labels matters
- Get different **labels** if we use different labelers
- and different **models**



Eckman et al., (2024): Position: Insights from survey methodology can improve training data.

Crowdworkers Characteristics

- Demographic differences (Posch et al 2022)
  - younger
  - more educated
  - Lower income
- Time use patterns (Rinderknect et al 2025)
  - Spend more time at home, more time alone
- Global South labels for Global North requesters (Ouyang et al 2022)

Who Labels Shapes What Models Learn

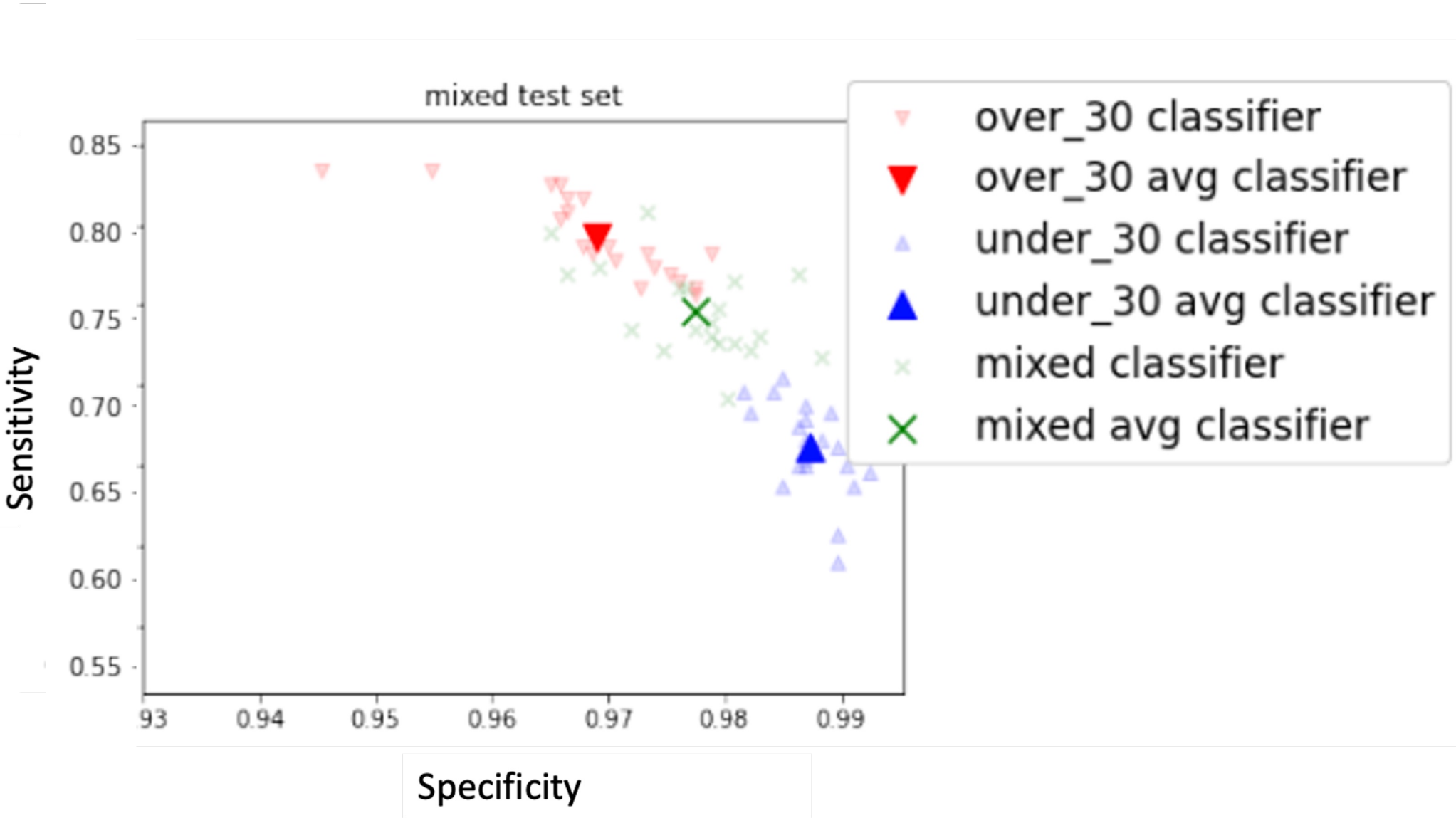
Influence of Annotator Characteristics on rating
DV: Offensiveness Rating (5-point Likert)

	Coef.	Std.Err.
Intercept	1.998	0.048
Gender: Non-binary	–0.235	0.060
Gender: Woman	–0.022	0.020
Race: Black or African Amer.	0.184	0.045
Race: Hispanic or Latino	–0.405	0.078
Race: White	–0.104	0.038
Educ.: College degree	–0.015	0.023
Educ.: Graduate degree	0.052	0.029

	Coef.	Std.Err.
Age: 25–29	–0.185	0.043
Age: 30–34	–0.165	0.041
Age: 35–39	–0.142	0.040
Age: 40–44	–0.037	0.043
Age: 45–49	–0.087	0.044
Age: 50–54	–0.141	0.046
Age: 54–59	0.001	0.039
Age: 60–64	0.309	0.050
Age: >65	0.117	0.042

Pei & Jurgens (2023): When Do Annotator Demographics Matter? Measuring the Influence of Annotator Demographics with the POPQUORN Dataset

Who Labels Shapes Model Performance



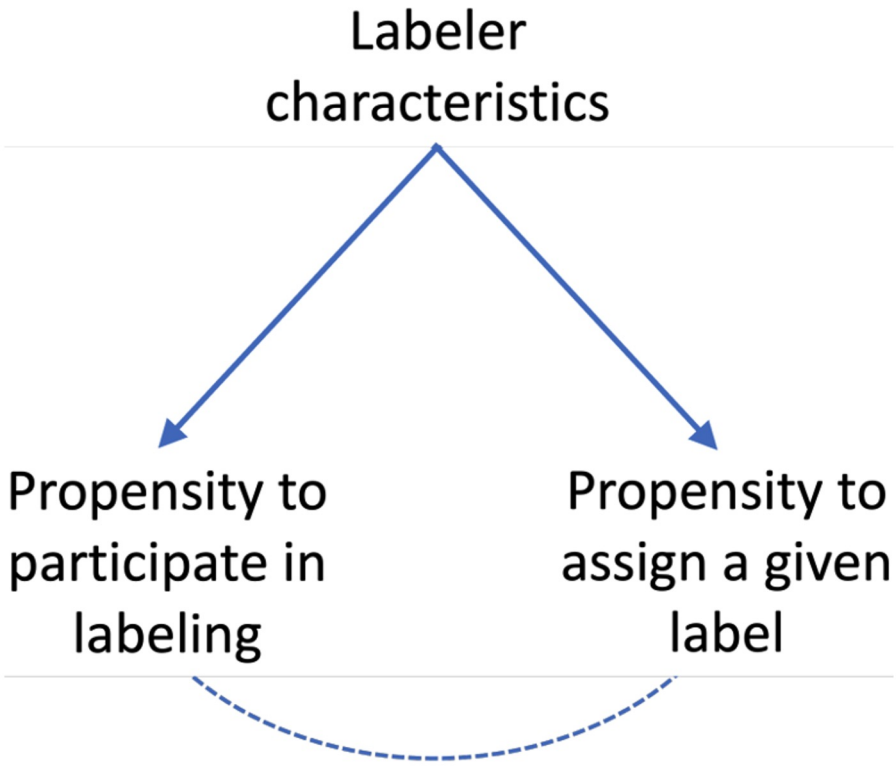
Al Kuwalty et al "Identifying and Measuring Annotator Bias Based on Annotators' Demographic Characteristics"

The 'Ground Truth' Problem: Majority ≠ Universal

- The "majority" perspective is not an objective truth:
  - If the majority does not find something offensive, that doesn’t mean it’s not offensive
  - Example: anti-gay tweets may be rated as "safe" by a straight majority but are harmful to the LGBTQ+ community
- Survey researchers understand this:
  - A sample must represent the target population
- But what is the “target population” for data labeling?

How can we reduce bias?

- Focus on right arrow
- Focus on left arrow
- Weighting



Eckman et al., (2024): Position: Insights from survey methodology can improve training data.

How Survey Researchers Handle Unrepresentative Samples

When a sample doesn't represent the target population, survey researchers correct for it:

- **Raking/Iterative proportional fitting**: Adjust weights so sample margins match known population margins
- **Propensity score weighting**: Model the probability of being in the sample; up-weight underrepresented
- **MRP (Multilevel Regression + Poststratification)**: Estimate subgroup opinions, then aggregate using population shares

How ML Researchers Handle Unrepresentative Samples

They typically don't

- Large scale annotation tasks utilise sparse annotation (N ~3) on thousands of data points, goal to assess IAA
  - Surveys typically have full annotation on fewer data points
- Achieving 'representativeness' for a single data point not possible without entire sample labelling
- For niche tasks expertise > representativeness anyway

PAIR: Population-Aligned Instance Replication

- Research Questions
  - Do non-representative annotators impact model performance?
  - Can weighting techniques increase performance?
- Data
  - 3,000 tweets
  - 15 offensive language (OL) annotations each
  - $p_{i,OL}$: % who think tweet i contains OL

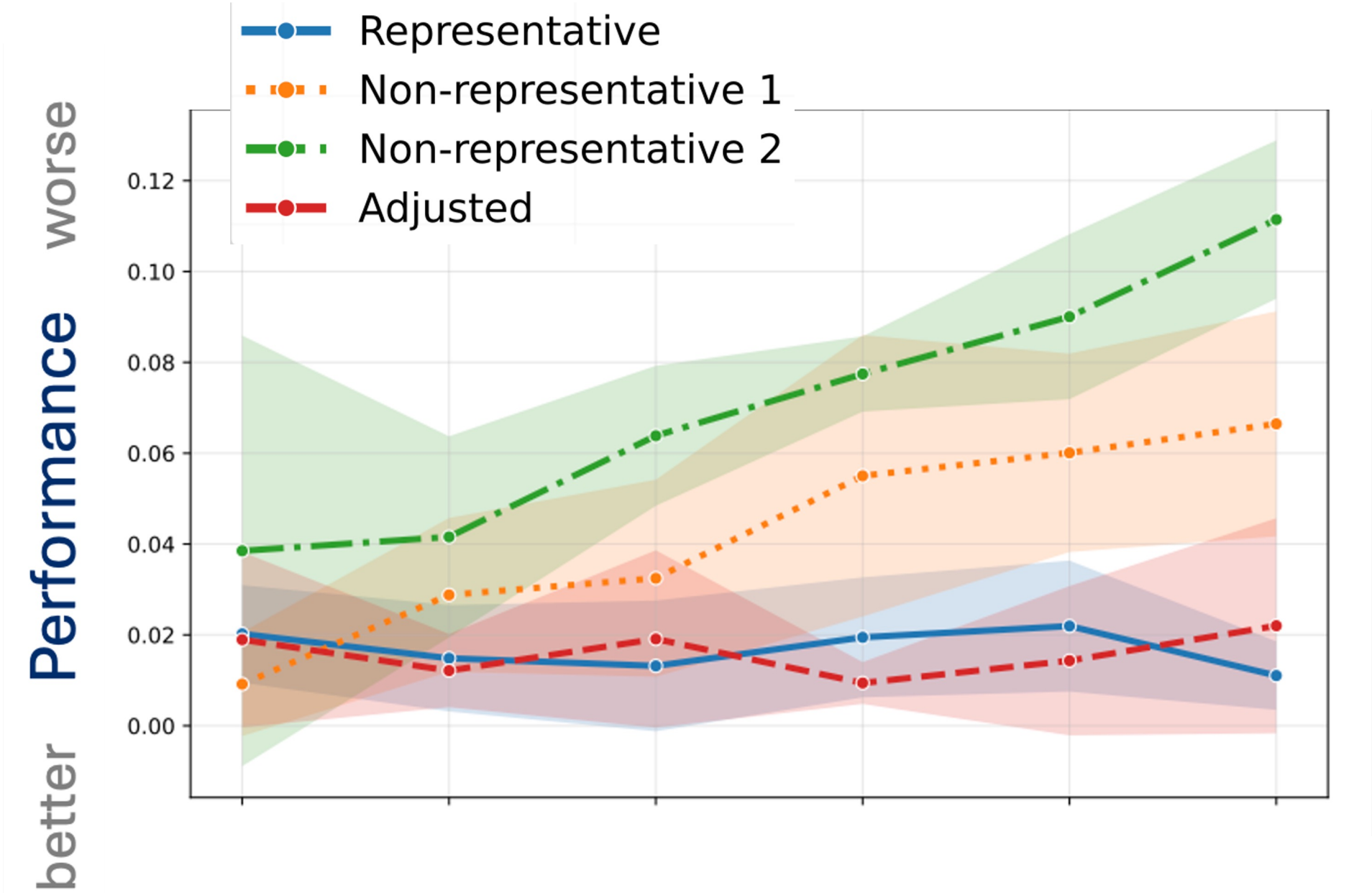
PAIR: Simulation Setup

- Simulate 2 types of annotators (A, B)
- Vary mix of types
- Convert probabilities to observations (0,1)
- Train models on each dataset
- Evaluate models

$$p^{A}_{i,OL} = \max(p_{i,OL} - \beta, 0)$$
$$p^{B}_{i,OL} = \min(p_{i,OL} + \beta, 1)$$

Dataset	A labels	B labels
Representative	6	6
Non-Rep 1	6	3
Non-Rep 2	9	3
Adjusted	6	3 + 3

PAIR: Results



Can LLMs replace human labelers?

- Match humans on narrow tasks (high agreement in structured settings; MSR 2025)
- Beat crowd workers in some cases (Bermejo et al., 2025)
- LLM labels ≈ human labels for training (MSR, 2025)
- Can fail on subjective / ambiguous tasks (ACL 2025)
- Potential to introduce systematic bias

What does survey research tell us?

Surface-level/directional agreement is relatively established (Argyle et al., 2023; Kim & Lee, 2024)

Statistical similarity ≠ validity

- Matching aggregates does not guarantee correct structure (Bisbee et al., 2024)
- Subgroup differences (Morris, 2025)
- Distribution tails / rare events not reproduced

Core limitation: no internal ground truth

- Synthetic data cannot be validated without external reference data
- Evaluation becomes circular if humans are removed

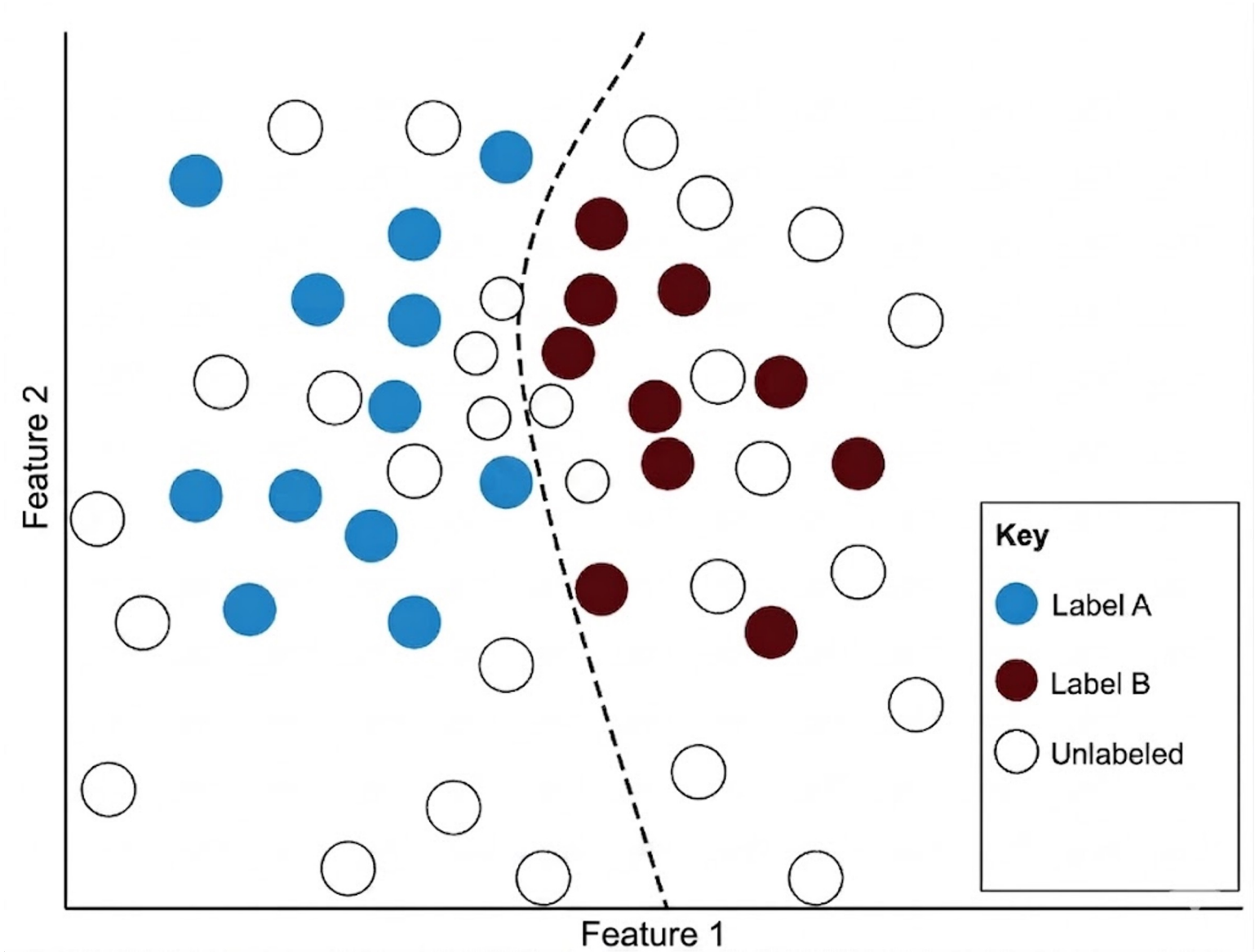
Ideas? Thoughts?
Questions?

Representation: Which Instances to Label?

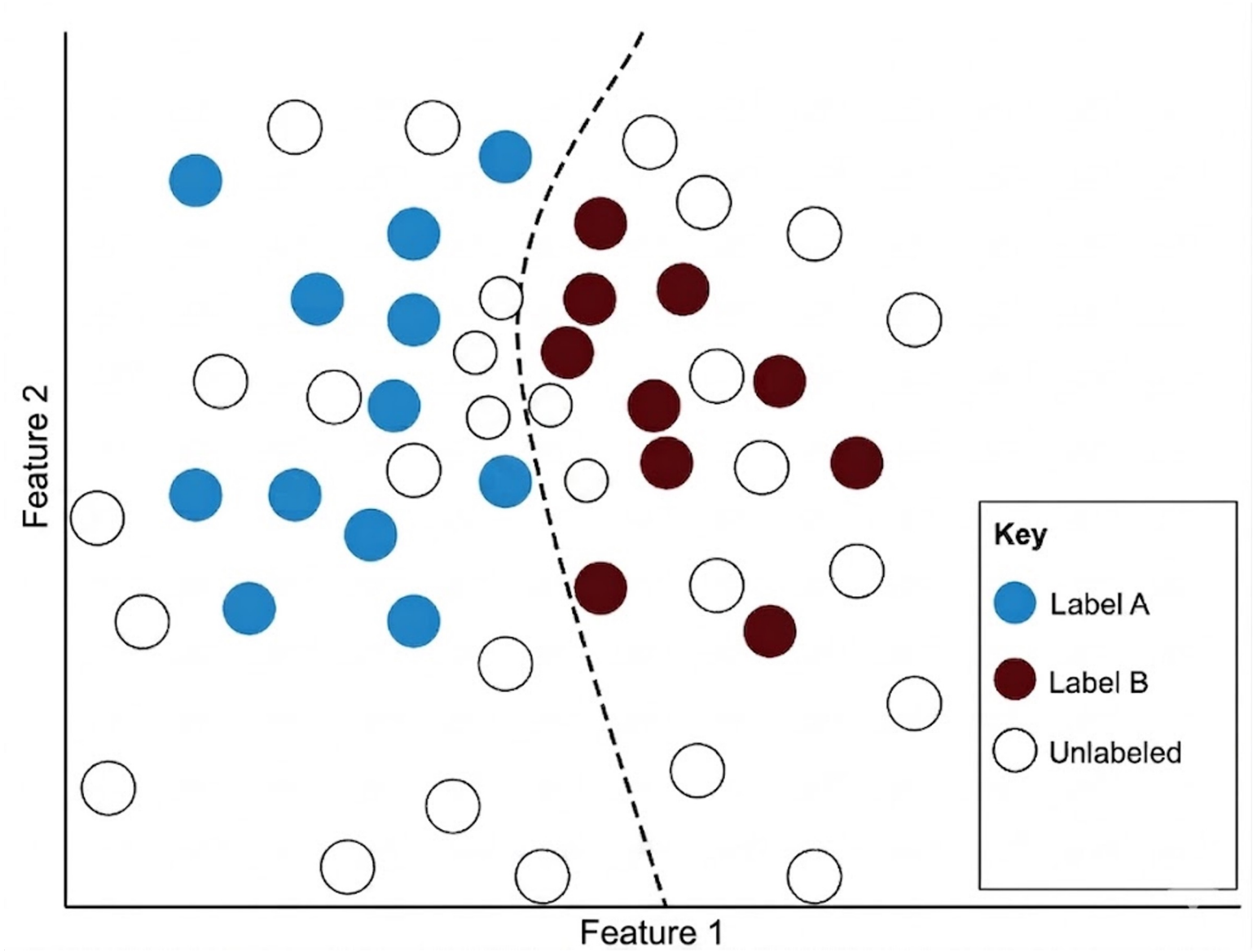
Sampling Approaches

- Simple random sampling from the pool
- Uncertainty sampling
  - where the model is confused
- Diversity sampling
  - cover the feature space
- Active learning
  - iterative, model-guided selection
  - (Monarch, 2021)

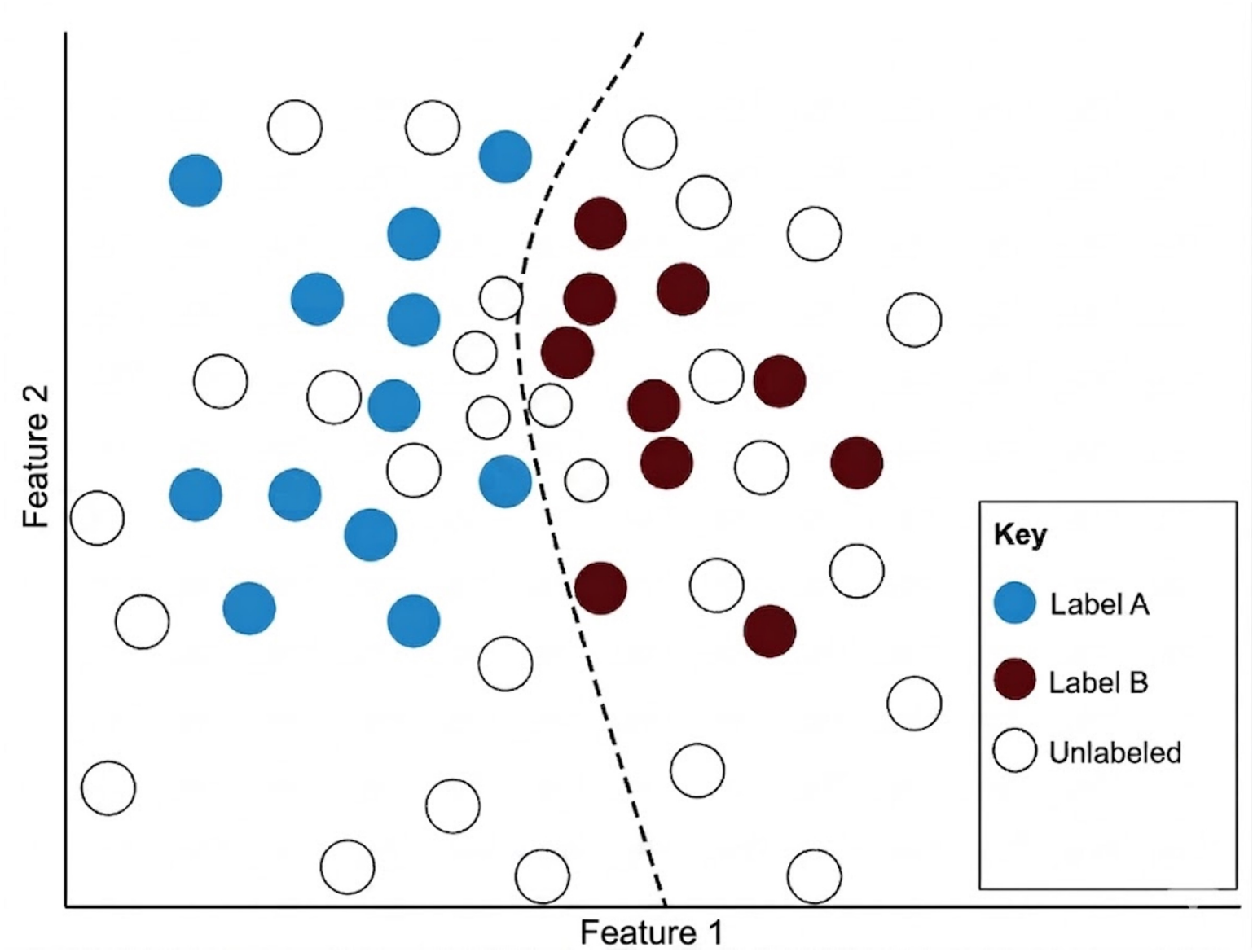
Model learns boundary between A and B



Uncertainty Sampling



Diversity Sampling



Active Learning

- Sample in batches
- Model looks at what it doesn't know yet and asks for specific labels to maximize information gain

- **Goal:** Get the best model performance with the *fewest* expensive human labels

“Clustering” means something different in ML

- In surveys:
  - Select groups of operational units (schools, blocks, households) as primary sampling units for cost efficiency
  - Then survey units within selected clusters
- In active learning:
  - Run k-means or similar to group similar instances (clusters)
  - Sample from each cluster

“Representative” means something different

	Surveys	**Machine Learning / AI**
Goal	Estimate population parameter	Learn a function
Rare Groups	Sampled proportionally	Oversampled
Success Metric	Unbiasedness	Generalization
Ideal Data Shape	True frequency in pop	Uniform Distribution: equal representation of A & B

Based on https://mindfulmodeler.substack.com/p/the-two-cultures-of-representativeness

How many labels do you need?

- In surveys: sample size is determined before data collection
  - power analysis, margin of error, design effect
- In active learning: sample size is determined during collection
  - stop when accuracy plateaus or budget runs out
  - depends on task complexity, class balance, and model architecture

Ideas? Thoughts?
Questions?

Ethics of Data Work

Ethics & Crowdworkers

- Wages: Many crowdworkers earn < minimum wage
- Benefits: No health insurance, no job security
- Harmful content
  - Labelers exposed to violence, hate speech, abuse
  - Psychological harm is well-documented
  - Workers often lack mental health support
  - High turnover signals the toll this takes
- Global dynamics: Wealthy countries commission work; lower-income countries perform it

Are labelers human subjects?

- Labelers are performing a task, not being studied – usually
- But research on labelers (behavior, accuracy, bias) clearly involves human subjects
- Current regulations lag behind practice

Data Workers' Inquiry

- Documents AI workers' experiences
- Influenced EU's Platform Workers' Directive
- Funded mental health support
- Supports the **Data Labelers Association**
    a. Workers produce zines, documentaries, and podcasts

Transparency in Labeling

- Survey industry benefits from transparency
- Call for ML researchers to release:
  - Labeling instructions and UI screenshots
  - Labeler selection protocols
    - and demographics (?)

Ideas? Thoughts?
Questions?

Practical Advice & Discussion

Concrete Recommendations

- Treat labeling as measurement, apply best practices
- **Pretest** – pilot with a few annotators first
- **Document** – task instructions, response options, labeler selection
- Randomize item order across annotators
- Use multiple annotators – disagreement is signal
- **Collect paradata** – time per item, revision patterns, confidence
- Report who labeled – demographics, expertise, selection method

What the ML pipeline is missing from surveys

- Standardized quality frameworks (like TSE)
- Pretesting protocols for labeling instruments
- Paradata collection (timing, process data)
- Transparent reporting of data collection methods
- Labeler selection documentation (like survey sampling reports)

What else?

Takeaway

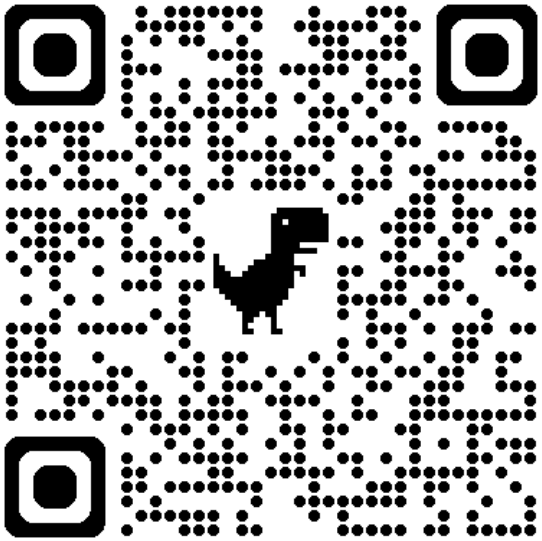
- Training data collection is data collection
- The principles that make surveys reliable also apply to labels
  - careful design
  - transparent methods
  - rigorous quality frameworks
- Survey researchers are uniquely qualified to improve this critical piece of the ML pipeline

Questions & Discussion

- What aspects of survey methodology are most transferable to training data collection?
- What is different enough that new frameworks are needed?
- How should the survey research community engage with the growing demand for labeled data?
- Should the survey industry pivot to training data collection?

Thank you

- Stephanie Eckman – steph@umd.edu
- Andrew Gordon - andrew.gordon@prolific.com
- Frauke Kreuter – fkreuter@umd.edu



Link to pdf version of slides

References 1

- Al Kuwatly, Hala, Maximilian Wich, and Georg Groh. Identifying and measuring annotator bias based on demography. In Proceedings of the Fourth Workshop on Online Abuse and Harms, pages 184–190, 2020.
- Beck, Jacob, Stephanie Eckman, Christoph Kern, and Frauke Kreuter. Order effects in annotation tasks. In Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP), pages 81–89, 2024. URL https://aclanthology.org/2024.uncertainlp-1.8.pdf.
- Beck, Jacob, Stephanie Eckman, Christoph Kern, and Frauke Kreuter. Bias in the loop: How humans evaluate AI-generated suggestions. *Harvard Data Science Review*, 2026. URL https://hdsr.mitpress.mit.edu/pub/nrcn4h7d/release/1.
- Chew, Robert, Stephanie Eckman, Christoph Kern, and Frauke Kreuter. From ground truth to measurement: A statistical framework for human labeling. arXiv preprint arXiv:2604.07591, 2026. URL https://arxiv.org/abs/2604.07591.

References 2

- Eckman, Stephanie, Barbara Plank, and Frauke Kreuter. Position: Insights from survey methodology can improve training data. In Proceedings of the 41st International Conference on Machine Learning, volume 235, pages 12268–12283. PMLR, 2024. URL https://proceedings.mlr.press/v235/eckman24a.html.
- Kern, Christoph, Stephanie Eckman, Jacob Beck, Rob Chew, Bo Ma, and Frauke Kreuter. Annotation sensitivity: Training data collection methods affect model performance. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 14874–14886, 2023. URL https://aclanthology.org/2023.findings-emnlp.992/.
- Krosnick, Jon A and Stanley Presser. Question and Questionnaire Design. Emerald Group Publishing, 2010. In: Handbook of Survey Research, 2nd edition.

References 3

- Monarch, Robert. Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI. Manning Publications, 2021.
- Ng, Andrew. A chat with Andrew on MLOps: From model-centric to data-centric AI. DeepLearning.AI, 2021. URL https://www.deeplearning.ai/the-batch/andrew-ng-data-centric-ai/.
- NLPerspectives Workshop. Proceedings of the 3rd workshop on perspectivist approaches to NLP. In Proceedings of ACL 2024 Workshops, 2024. URL https://aclanthology.org/2024.nlperspectives-1.1/.
- Otterbacher, Jahna, Jo Bates, and Paul Clough. Investigating user perception of gender bias in image search: The role of sexism. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pages 1–13, 2018.

References 4

- Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35:27730–27744, 2022. URL https://arxiv.org/abs/2203.02155.
- Posch, Lisa, Arnim Bleier, Fabian Flöck, Clemens M. Lechner, Katharina Kinder-Kurlanda, Denis Helic, and Markus Strohmaier. Characterizing the Global Crowd Workforce: A Cross-Country Comparison of Crowdworker Demographics. arXiv preprint arXiv:1812.05948, 2022
- Rinderknecht, R. Gordon, Long Doan, and Liana C. Sayer. The daily lives of crowdsourced U.S. respondents: A time use comparison of MTurk, Prolific, and ATUS. Sociological Methodology, 55(2):183–217, 2025.

References 5

- Sambasivan, Nithya, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pages 1–15, 2021. doi: 10.1145/3411764.3445518.
- Sculley, D, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. In Advances in Neural Information Processing Systems, volume 28, 2015.
- Shumailov, Ilia, Zakhar Shumilo, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. arXiv preprint arXiv:2305.17493, 2023. URL https://arxiv.org/abs/2305.17493.
- Sudman, Seymour and Norman M Bradburn. Asking Questions: A Practical Guide to Questionnaire Design. Jossey-Bass, 1982.