The University of Queensland Australia | CREATE CHANGE

Humans or AI? Why not Both!

A/Prof. Gianluca Demartini
Data Science Discipline
School of Electrical Engineering and Computer Science

Research Interests

- **Information Access** (since 2005)
  Structured/Unstructured data (SIGIR12), Entity Types (ISWC13, WSemJ16)
  Entity Recognition (WWW14), Prepositions (CIKM14), Entity Cards (SIGIR19)
  Evaluation (ECIR16 Best P, CIKM17, SIGIR18, CIKM19, WWW22, TOIS23, ICTIR23 Best P)
- **Human-AI Systems** (since 2012)
  Entity Linking (WWW12,VLDBJ), CrowdQ (CIDR13), Learnersourcing (LAK21,LAK22,JCAL)
  HITL (FnT17), Bias (SIGIR18, ECIR20 Best P), Crowd-LLM (CACM24, ICWSM24)
- **Better Crowdsourcing Platforms** (since 2013)
  Platforms (WWW15, CSCWJ18), Experiments (CSCW21), Pricing (HCOMP14)
  Task Allocation (WWW13, WWW16, COR), Workers (CHI15), Metadata (IP&M), Attacks (HCOMP18 Best P, JAIR), Reward (CSCW20 Hon. Mention), Time (HCOMP16)
  Modus Operandi (UBICOMP17, HT19, WSDM20, TOIS24), Complexity (HCOMP16), Abandonment (WSDM19, TKDE, ACM TSC)
- **Better Data** (since 2019)
  Noise (WWW19), Data Workers (SIGIR20, TOIS, TKDE, WWW23), Behaviors (CIKM20)
  Know. Graphs (ISWC19), Unknown Unknowns (ECAI20, HCOMP21), SES (WebSci22), Fairness (CIKM22, SIGIR23, CACM24, FAccT24), Active Learning (AAAI24)
- **Data for Public Good** (since 2020)
  Conservation (w/ Google); Gender (w/ Wiki); Environment (ECIR21, ADCS21)
  Fake News (w/ Meta; SIGIR20, CIKM20, IP&M); Democracy (ADCS21, ICWSM23)

Thanks to:

Australian Government
Australian Research Council

European Commission

Swiss National Science Foundation

EPSRC
Engineering and Physical Sciences Research Council

Meta

WIKIMEDIA FOUNDATION

Google AI

Outline

Training AI with Human Data

- ML Fairness without Sensitive Attributes (ACM FAccT 2024)
- Co-learning Active Learning (AAAI 2024)

Humans or AI? Why not Both?

- A Human-LLM Collaborative Spectrum (CACM, Apr 2024)

Using GenAI to persuade Humans

- LLMs can generate persuasive personalized content (ACM TheWebConf 2024)

Humans using GenAI for Data Annotation

- What happens when the crowd trust their peers? (ACM TOIS, Jan 2024)
- The crowd does use LLMs (ICWSM 2024)
- What happens when they use LLMs? (ICWSM 2024)

Fairness without Sensitive Attributes

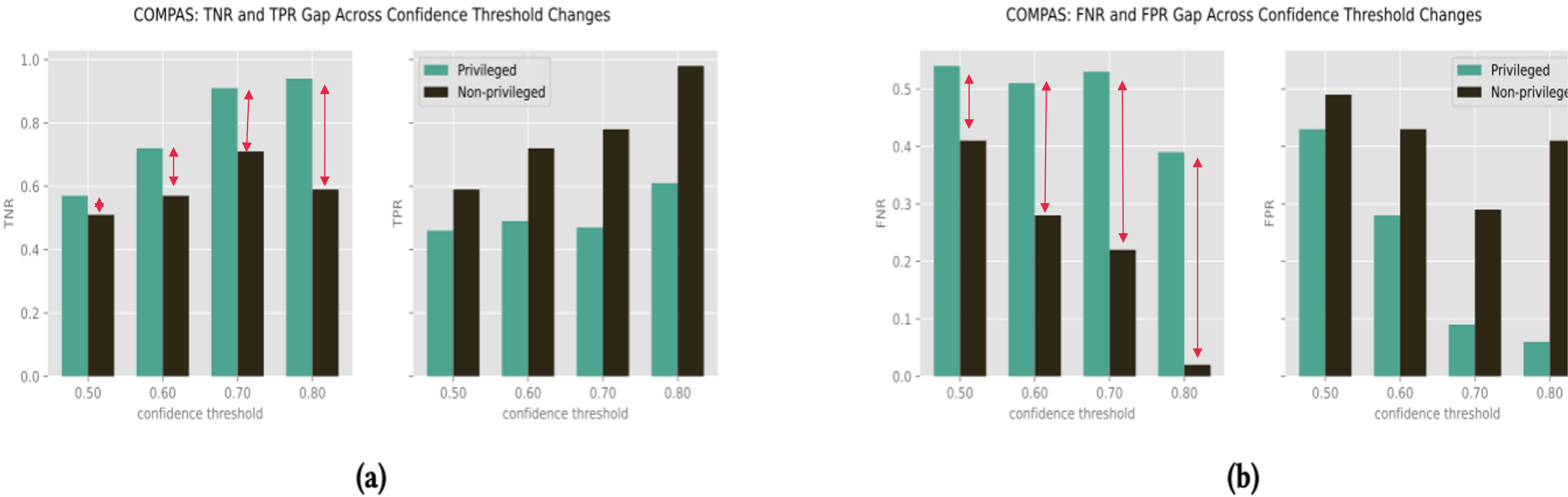
- COMPAS [11]

	Previous Misconduct	Charge Degree	……	Age	Race	Recidivist
Offender A	0	Felony	……	23	White	True
Offender B	3	Felony	……	22	Black	False
Offender C	2	Misdemeanor	……	30	Black	True

- Classification
- Supervised learning

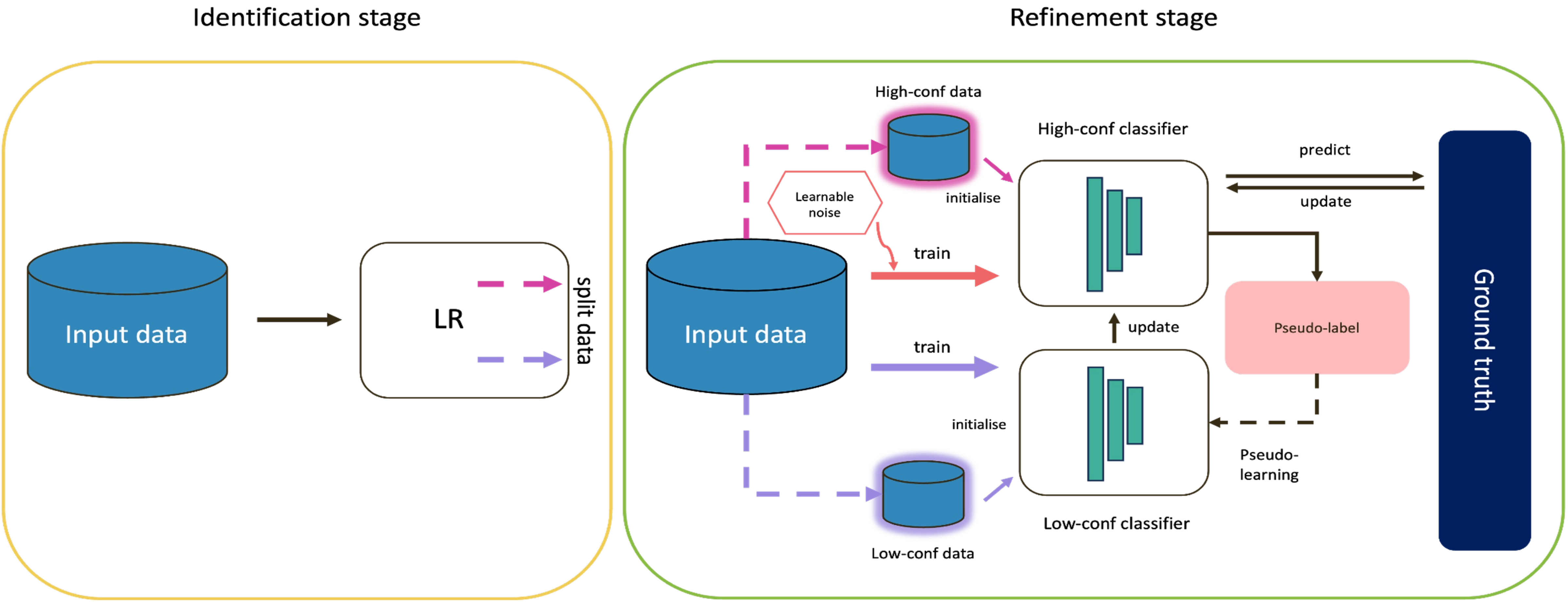
Hongliang Ni, Lei Han, Tong Chen, Shazia Sadiq, and Gianluca Demartini. **Fairness without Sensitive Attributes via Knowledge Sharing**. In: The Seventh Annual ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT '24). Rio de Janeiro, Brazil, June 2024.

A Look at Confidence of LR for Recidivism Prediction

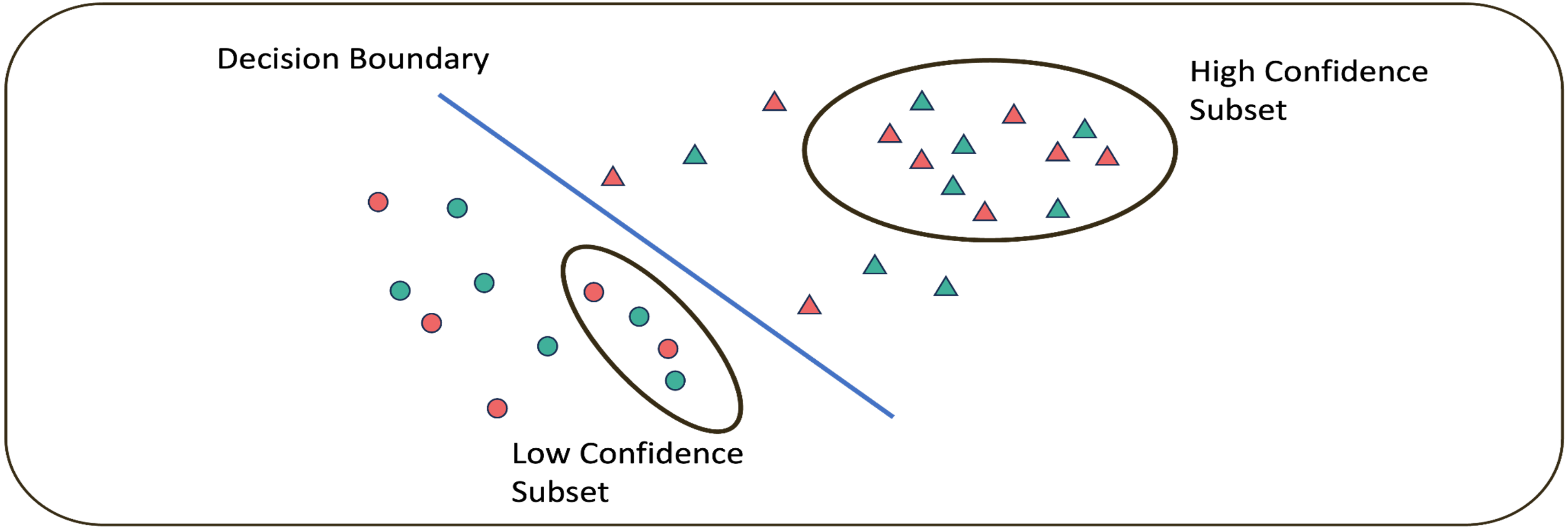


As the confidence threshold increases, the performance differences between demographic groups become more pronounced.

Reckoner: Two-Stage Knowledge-Sharing Framework



Reckoner: Two-Stage Knowledge-Sharing Framework



Refinement stage - *Knowledge-Sharing /pseudo-learning*:

- Shift the decision boundary closer to the samples in high-confidence subsets. The model will not misclassify similar instances based on distribution patterns of the majority.
- Learnable noise offers auxiliary information for demographic groups, ensuring both accuracy and fairness.

Experiments

- COMPAS

Metrics(%) Methods	Accuracy	Equalised Odds	Demographic Parity
DRO	64.88 ± 0.34%	23.11 ± 1.80%	25.32 ± 1.22%
ARL	65.32 ± 0.70%	23.01 ± 1.21%	25.37 ± 1.01%
FairRF	63.26 ± 0.83%	25.67 ± 2.63%	21.47 ± 1.76%
Chai's work (softmax label)	63.47 ± 0.44%	21.32 ± 1.97%	19.52 ± 2.46%
Chai's work (linear label)	63.34 ± 0.46%	20.31 ± 2.62%	20.27 ± 2.34%
Reckoner	64.92 ± 0.63%	17.47 ± 0.87%	20.72 ± 0.97%
Reckoner (w/o noise)	64.95 ± 0.51%	17.91 ± 1.32%	21.21 ± 1.33%
Reckoner (w/o *pseudo-learning*)	64.38 ± 0.83%	17.98 ± 1.34%	21.18 ± 1.46%

- Biased labels and other hidden bias harm fairness in predictions and mislead the classifier
- We proposed a knowledge-sharing framework for fair predictions with missing sensitive attributes.

Machine Learning and Active Learning

(Supervised) Machine Learning requires (manually) labelled training data

Active Learning (AL) aims at selecting few, informative training data points to label.

Uncertainty-based AL algorithms - > diversity

Distribution-based AL -> representativeness

We aim to select the most **diverse** and **representative** samples within a training dataset.

Using noisy labels (i.e., predictions made by the primary model) on unlabelled data

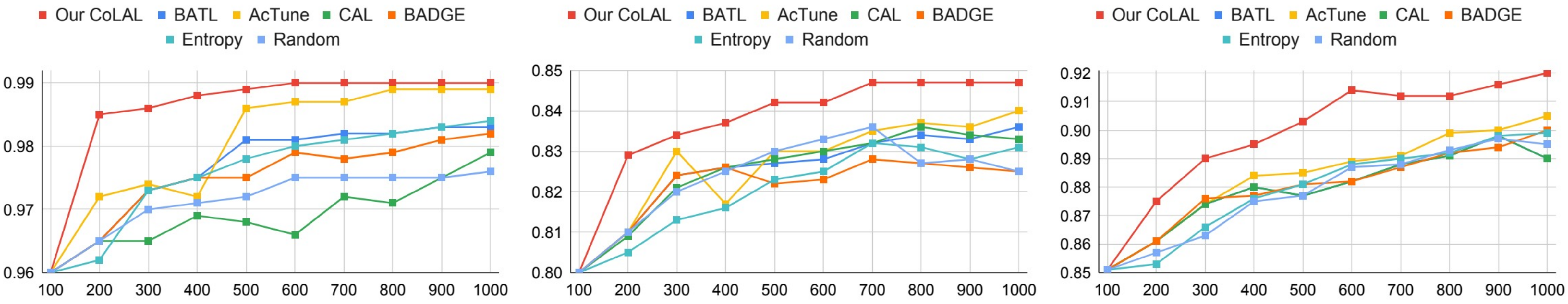
Limited labelled data -> incomplete decision boundaries -> a peer model, trained with noisy labels

Co-learning Active Learning

Combine the two (target and peer) models: look at the similarity of their classification decisions

Select the next training instances for labelling:

regions with least overlap (i.e., disagreement between the two models)



((a)) Comparison on DBPedia. ((b)) Comparison on PubMed. ((c)) Comparison on SST-2.

Linh Le, Genghong Zhao, Xia Zhang, Guido Zuccon, and Gianluca Demartini. **CoLAL: Co-learning Active Learning for Text Classification**. In: The 38th Annual AAAI Conference on Artificial Intelligence (AAAI-24). Vancouver, Canada, February 2024.

Outline

Training AI with Human Data

- ML Fairness without Sensitive Attributes (ACM FAccT 2024)
- Co-learning Active Learning (AAAI 2024)

Humans or AI? Why not Both?

- A Human-LLM Collaborative Spectrum (CACM, Apr 2024)

Using GenAI to persuade Humans

- LLMs can generate persuasive personalized content (ACM TheWebConf 2024)

Humans using GenAI for Data Annotation

- What happens when the crowd trust their peers? (ACM TOIS, Jan 2024)
- The crowd does use LLMs (ICWSM 2024)
- What happens when they use LLMs? (ICWSM 2024)

What about LLMs?
The role of Humans

This talk so far

Humans used to provide preference data: PPO-RLHF, DPO

LLMs can replace humans in data annotation tasks

Microsoft Bing has replaced human assessors with GPT-4 for relevance judgments!

"Who is better?"

versus

"How can they work together?"

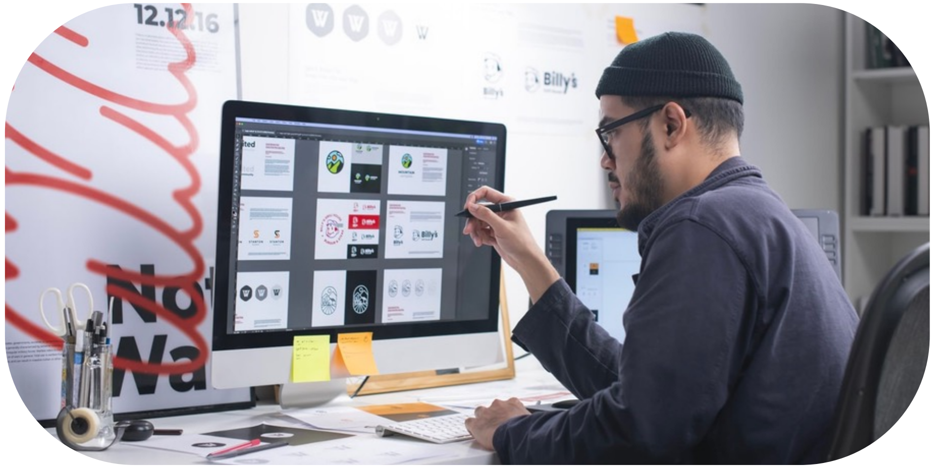
Next in this talk

Guglielmo Faggioli, Laura Dietz, Charles Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth.
Who determines what is relevant? Humans or AI? Why not both!
In: Communications of the ACM (*CACM*). *Vol.67 No.4,* April 2024.

Collaboration Balance	Task Allocation
Human Judgment	
	Humans manually decide (about relevance) without any kind of AI support.
	Humans have full control of deciding but are supported by machine-based text highlighting, data clustering, etc.
Model In The Loop	
	Humans decide based on LLM-generated summaries needed for the decision.
	Balanced competence partitioning. Humans and LLMs focus on decisions they are good at.
Human In The Loop	
	Two (or more) LLMs each generate a decision, and a human selects the better one.
	An LLM makes a decision (and an explanation for it) that a human can accept / reject.
$\cdot n$	LLMs are considered crowdworkers—varied by specific characteristics—, aggregated and controlled by a human.
Fully Automated	Fully automatic decision without humans.

LLMs to generate persuasive content

Can LLMs generate personalized ad messages targeting specific personality traits?



Ad designers

Aligning advertising messages with an individual's personality traits can enhance ad effectiveness.

ChatGPT

The Emergence of LLMs

Elyas Meguellati, Lei Han, Abraham Bernstein, Shazia Sadiq, and Gianluca Demartini. **How Good are LLMs in Generating Personalized Advertisements?**. In: The 2024 ACM Web Conference (Short Paper track). Singapore, May 2024.

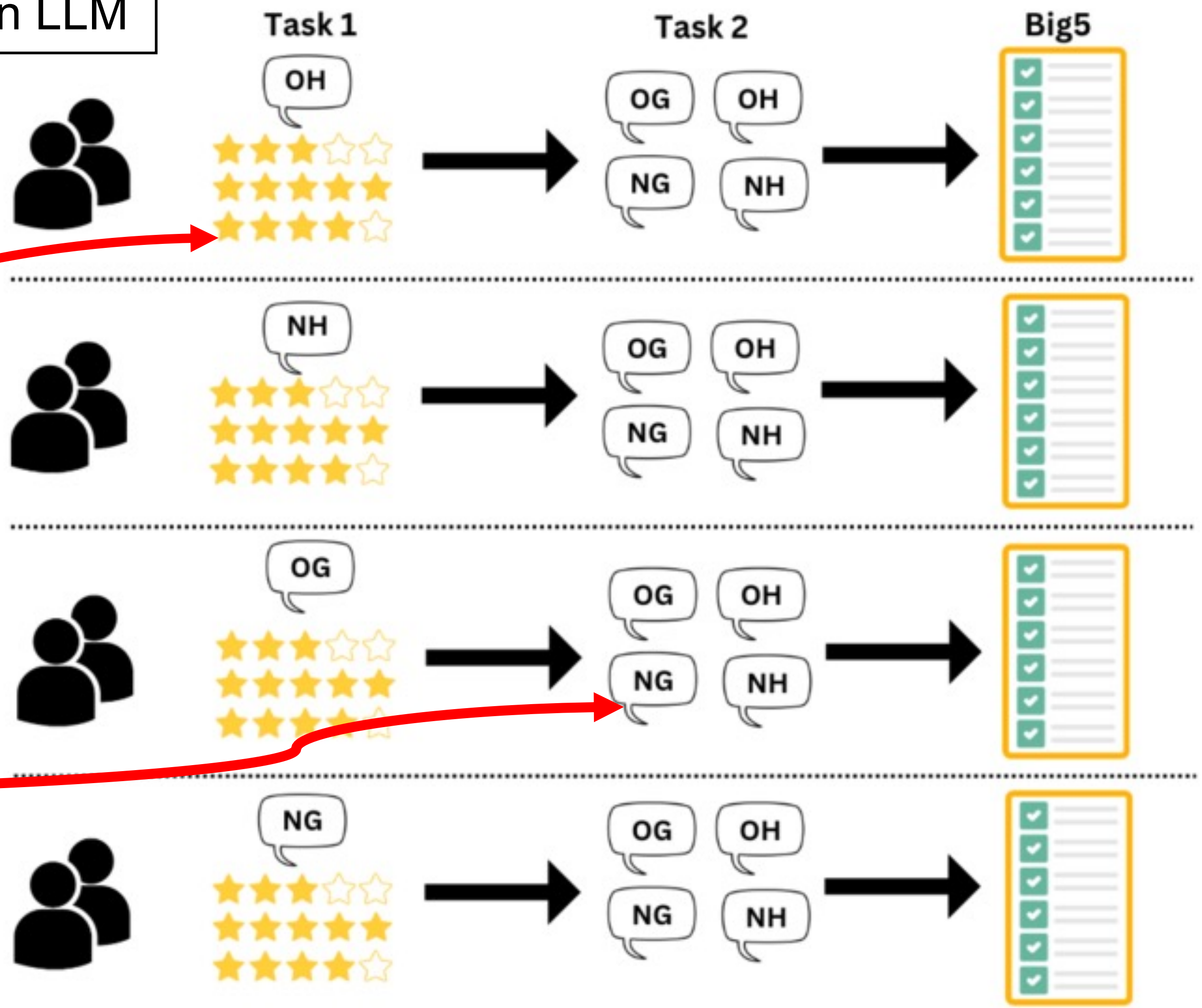
Study

OH: Openness, written by a Human
OG: Openness, Generated by an LLM
NH: Neuroticism, written by a Human
NG: Neuroticism, Generated by an LLM

Task 1: Assessed user reactions to ads in a social media feed

1. Product attitude
2. Purchase intention
3. Engagement intention

Task 2: Compared preferences for side-by-side presented ads in a shopping scenario



Results - Task 1

Table 1: Mean values of measurements for each survey and personality match

Survey	Measurement	Matched	Unmatched
OG	Product Rating	4.14	3.71
	Purchase Intention	4.14	3.69
	Engagement Intention	4.33	3.73
NG	Product Rating	3.84	4.0
	Purchase Intention	3.77	4.15
	Engagement Intention	3.97	4.29
OH	Product Rating	4.13	3.96
	Purchase Intention	4.33	3.68
	Engagement Intention	4.30	3.88
NH	Product Rating	3.61	3.76
	Purchase Intention	3.74	4.0
	Engagement Intention	3.71	4.15

Table 2: P-values of Ads between Match and Non-match Personalities after Benjamini-Hochberg Correction. A corrected P-value ≤ 0.05 is considered statistically significant.

Ad Type	Personality Trait	Product Rating	Purchase Intention	Engagement Intention
Generated	Openness	**0.02**	**0.02**	**0.01**
	Neuroticism	0.33	0.27	0.27
Human	Openness	0.50	**0.05**	0.15
	Neuroticism	0.54	0.54	0.47

Table 3: P-values of Human ads vs Generated ads for matched personalities after Benjamini-Hochberg Correction.

Ad's Personality	Product Rating	Purchase Intention	Engagement Intention
Openness	0.42	0.42	0.42
Neuroticism	0.46	0.42	0.90

Ads crafted for openness works best
Human and AI generated ads perform equally good

Results - Task 2

Table 4: Click Distribution and Percentages for Ads Displayed Side-by-Side for Task 2.

Ad Type	Clicks (%)
Human-written ad tailored to the openness trait	31.82
Generated ad tailored to the openness trait	26.21
Generated ad tailored to the neuroticism trait	24.93
Human-written ad tailored to the neuroticism trait	17.04

What about LLMs? The role of Humans

This talk so far

Humans used to provide preference data: PPO-RLHF, DPO
LLMs can replace humans in data annotation tasks
Microsoft Bing has replaced human assessors with GPT-4 for relevance judgments!

Next in this talk

"Who is better?"

versus

"How can they work together?"

This talk so far

Collaboration Balance	Task Allocation
Human Judgment	Humans manually decide (about relevance) without any kind of AI support.
	Humans have full control of deciding but are supported by machine-based text highlighting, data clustering, etc.
Model In The Loop	Humans decide based on LLM-generated summaries needed for the decision.
	Balanced competence partitioning. Humans and LLMs focus on decisions they are good at.
Human In The Loop	Two (or more) LLMs each generate a decision, and a human selects the better one.
	An LLM makes a decision (and an explanation for it) that a human can accept / reject.
·n	LLMs are considered crowdworkers—varied by specific characteristics—, aggregated and controlled by a human.
Fully Automated	Fully automatic decision without humans.

Guglielmo Faggioli, Laura Dietz, Charles Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth.
Who determines what is relevant? Humans or AI? Why not both!
In: Communications of the ACM (*CACM*). *Vol.67 No.4,* April 2024.

What do we know about people in crowdsourcing?

We know that:

- Crowd workers can assess misinformation (La Barbera et al. 2020; La Barbera et al. 2024)
- Crowd workers follow the crowd (bandwagon effect) (Eickhoff 2018; **Xu et al. 2024 TOIS**)
- Crowd workers make use of LLMs (Veselovsky et al. 2023; **Christoforou et al. ICWSM 2024**)

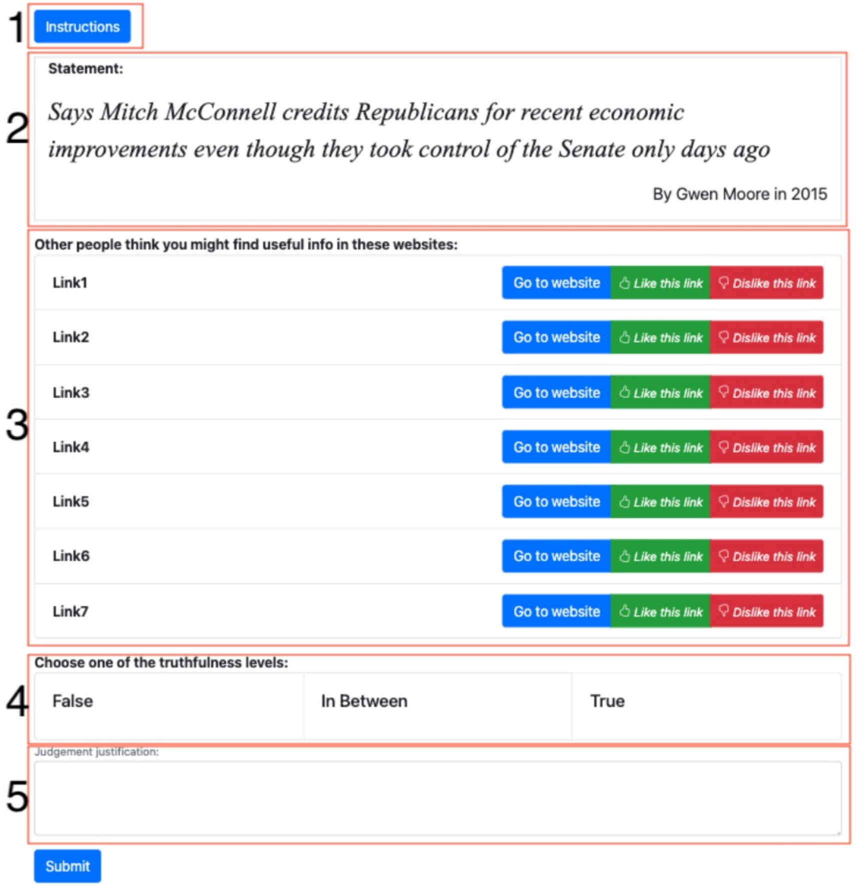
Open questions:

- Do crowd workers follow LLMs?
- What does that mean for the labels we collect using crowdsourcing?

David La Barbera, Eddy Maddalena, Michael Soprano, Kevin Roitero, Gianluca Demartini, Davide Ceolin, Damiano Spina, and Stefano Mizzaro. **Crowdsourced Fact-Checking: Does It Actually Work?**. In: Information Processing & Management, Volume 61, Issue 5, September 2024, Elsevier

Evidence from Peers

What happens when they are presented with evidence from peers (i.e., other crowd workers)?

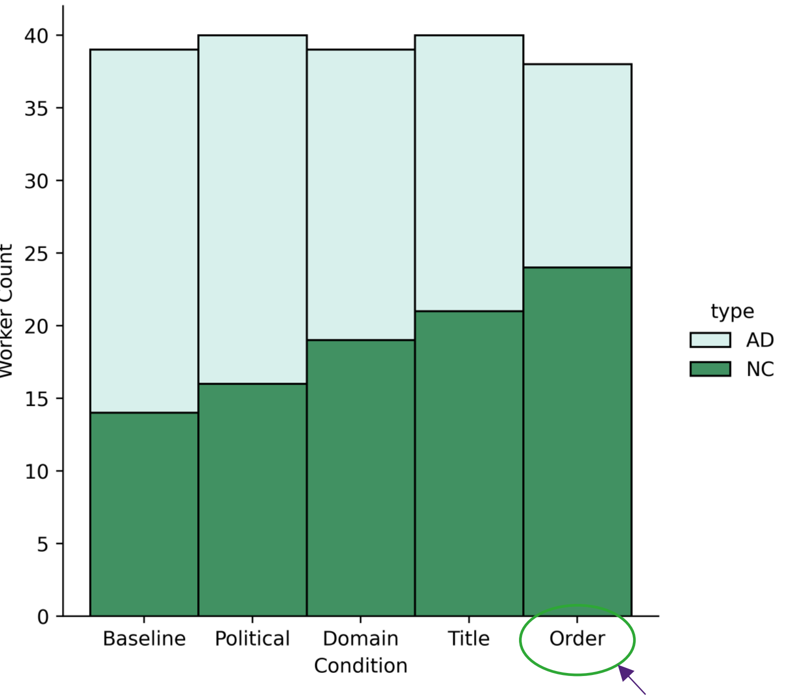


Conditions for box 3:
No box 3
Baseline (no information)

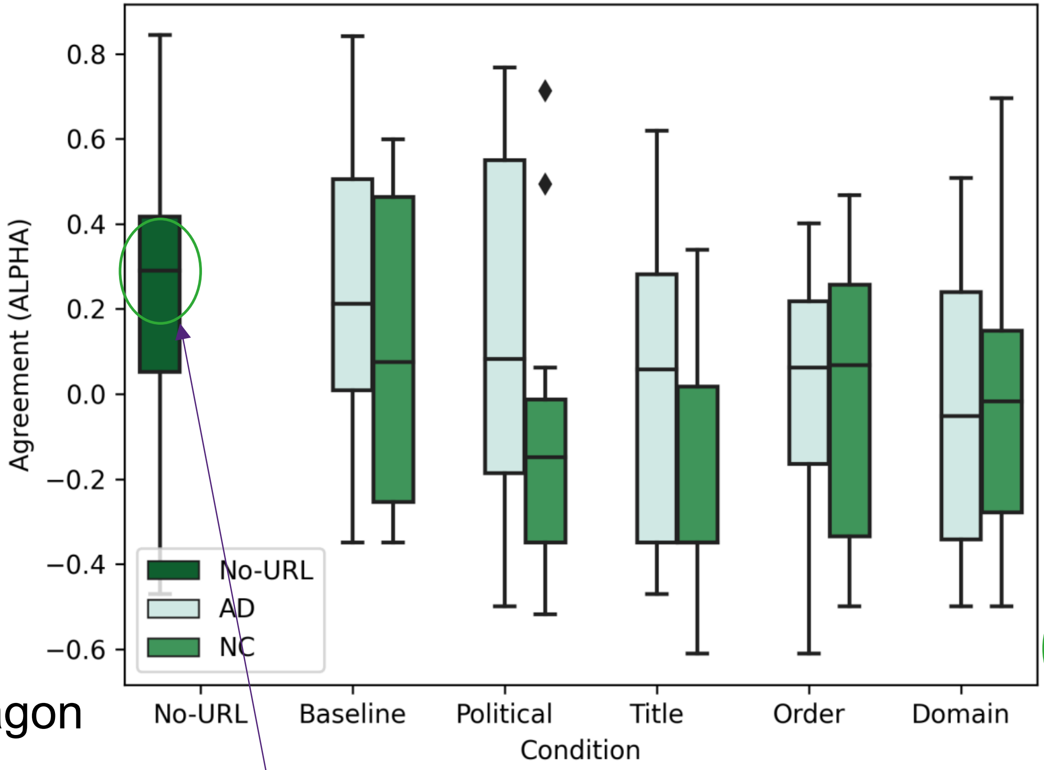
Order (e.g., 3 people endorsed this website)
Domain (e.g., www.cnn.com)
Title (e.g., “Funding Down, Tuition Up”)
Political (e.g., 3 Republicans endorsed this website)

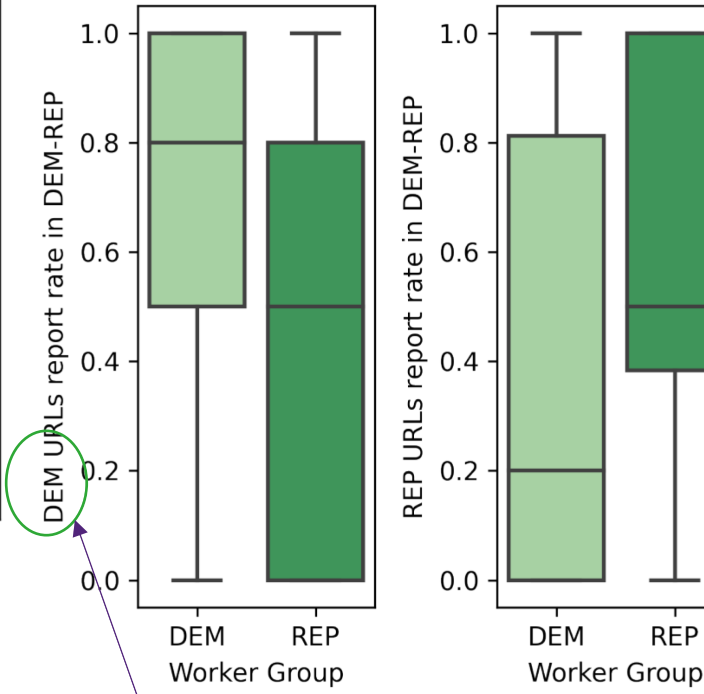
Jiechen Xu, Lei Han, Shazia Sadiq, and Gianluca Demartini. **On the Impact of Showing Evidence from Peers in Crowdsourced Truthfulness Assessments**. ACM Transactions on Information Systems (TOIS), 42(3), 1-26. 2024.

Adopters (AD) vs Non-Compliant (NC)



bandwagon





No evidence leads to best quality!

Strong bandwagon effect with political bias

more likely to favor evidence provided by their ‘politically-aligned’ peers

Strong use of chatGPT
Especially on Amazon MTurk

Generative AI in Crowdwork

We asked crowd workers regarding their use of GenAI tools.

	ALL	USA	India	UK	EU
Prolific	13.1% 13.4%	19.0% 14.0%	- -	9.0 % 10.0%	9.0% 14.5%
MTurk	80.3% 73.2%	94.3% 86.2%	66.3% 59.4%	- -	- -
Clickworker	20.7% 15.0%	27.9% 20.6%	- -	16.9% 11.0%	15.3% 12.6%

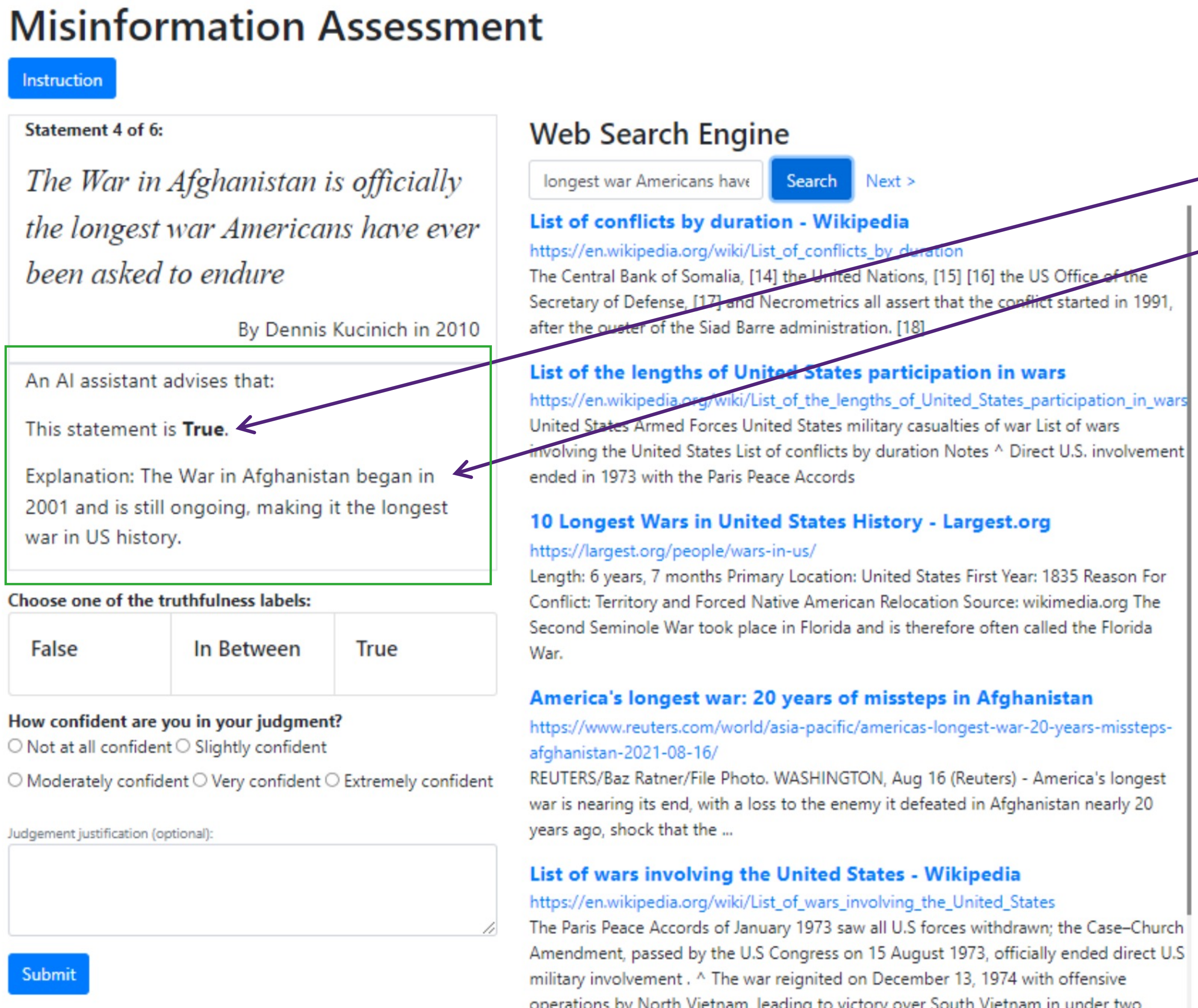
Table 4: Workers reporting self-initiated use of AI chatbots in tasks, by platform, country and T1/T2 [top/bottom].

Prolific, Mturk, Clickworker; May 2023, and Dec 2023

- Workers' self-reported use of GenAI
  - did not change over time
  - was strongly correlated to the platform they use.
- **MTurk workers use GenAI on their own volition** significantly more often than those operating at Clickworker or Prolific.
- Many expressed concerns that GenAI would reduce the number of opportunities for surveys, as requesters are looking for authentic human responses.

Evgenia Christoforou, Gianluca Demartini, and Jahna Otterbacher. **Generative AI in Crowdwork for Web and Social Media Research: A Survey of Workers at Three Platforms**. In: The 18th International AAAI Conference on Web and Social Media (ICWSM 2024).

Study setup



Conditions:

- Baseline – no LLM
- Label
- Explanation
- Label+Exp

437 US participants from Prolific

120 political statements

GPT-3.5

Each Participant: 6 statements balanced on truthfulness and political leaning, pre-task, post-task

RQ1: Quality of Assessments

Crowd workers overestimate truthfulness when exposed to LLM-generated information

GPT-3.5 tends to overestimate truthfulness labels

Participants with no LLM access have a higher rate of underestimation errors

Condition	% Over	% Under	Accuracy
Baseline	27.50	30.83	41.67
Label	38.06	18.61	43.33
Explanation	36.11	25.28	38.61
Label+Exp	37.78	20.56	41.67

Being exposed to the LLM does not have a significant impact on judgment quality

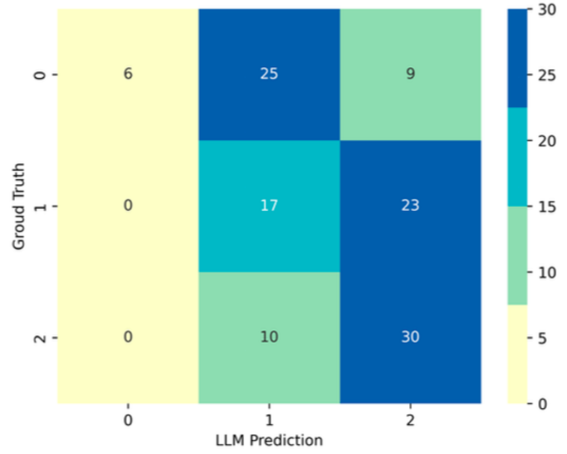


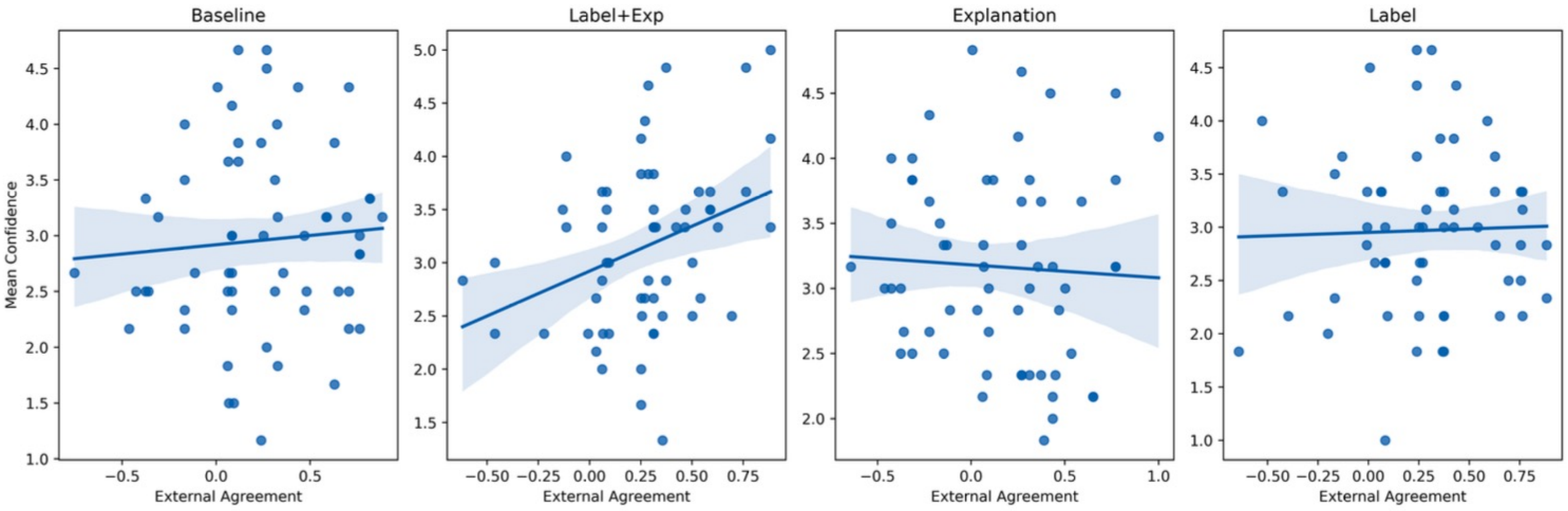
Figure 3: Confusion matrix for numbers of assessments by GPT-3.5 against the ground truth labels. Labels for row and column are ground truth labels and GPT-3.5's labels, respectively. Notation: 0 – false, 1 – in-between, 2 – true.

RQ2: Self-reported Confidence

Participants report the level of confidence in their judgements

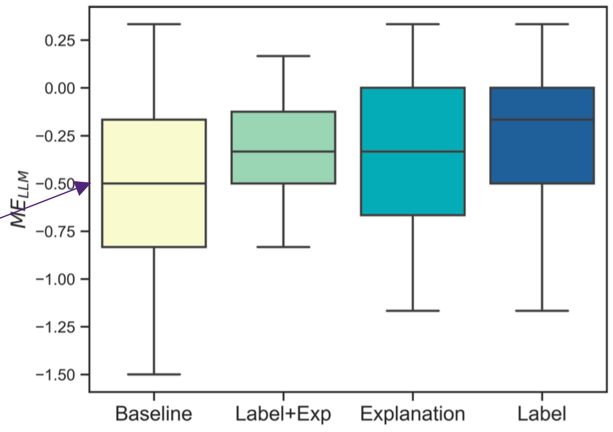
LLM information has no significant effect on of crowd workers’ self-reported confidence levels

Quality (i.e., agreement with ground truth) has a significant correlation with confidence

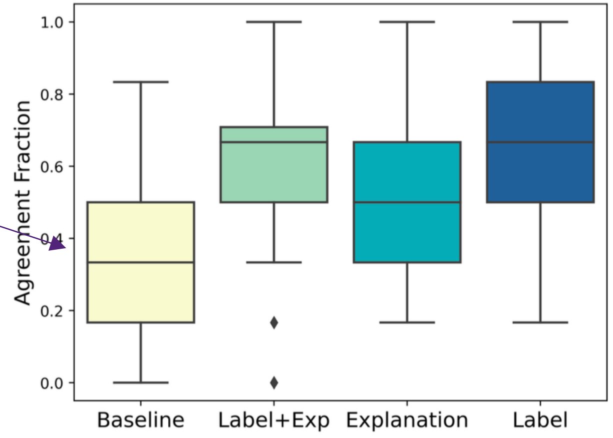


RQ3: Reliance and Trust in the LLM

Crowd assigns lower truthfulness relative to the GPT-3.5



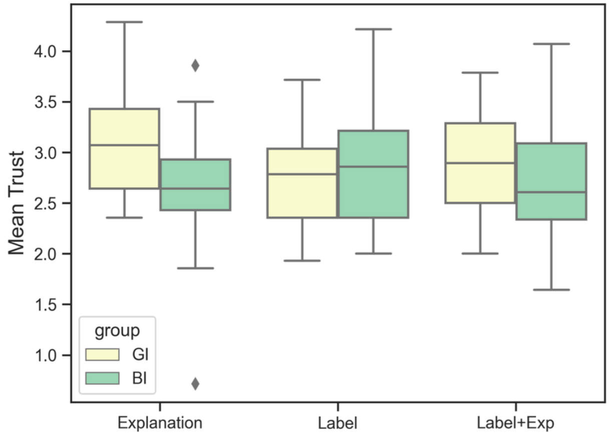
Crowd relies on the LLM advice when exposed to it



TiA-Trust post-task: no significant difference among the three LLM conditions

Correlation between self-reported confidence and trust

Participants who have a *good first impression* (GI) report higher trust



RQ4: Behavioral Indicators

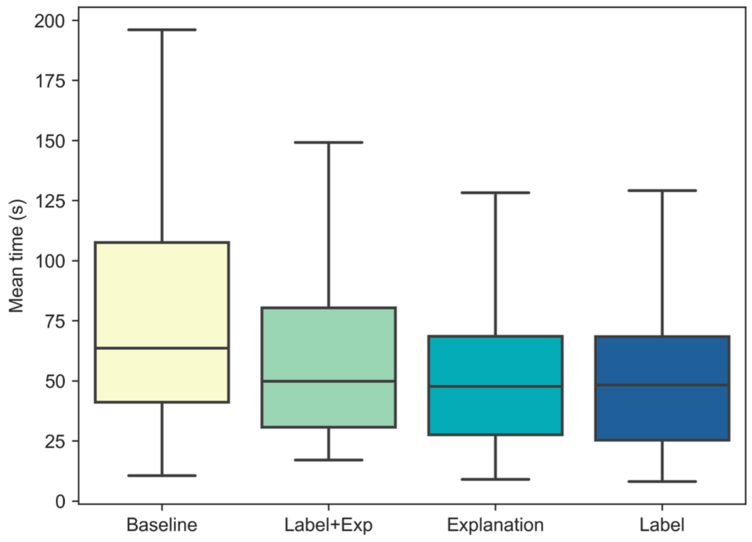
Use of Search Engine

Baseline participants issue significantly more search queries as compared to LLM conditions

Search active (above median number of queries) participants diverge more from LLM labels

Able to mitigate the bias from the LLM by leveraging the search engine results

Assessment Time:



The Crowd and LLMs

- Providing LLM-generated labels
  - an effective method to **speed up** crowdsourcing of misinformation assessment
  - **leads to over-estimation** of truthfulness with LLM
  - (but similar level of accuracy)
- Extensive (or excessive?) **reliance**, biased labels

gianlucademartini.net
demartini@acm.org
@eglu81

Lessons learned and open questions

- Human data is needed to train AI; Human labels are biased; we need fair AI
- LLMs can replace humans in many NL and creative tasks, but should they?
- Crowd workers rely on LLMs to label data. Is this the end of crowdsourcing?

Open questions:

- Can GenAI and humans work collaboratively and increase AI fairness?
- What's the role of humans?
- Does personalized GenAI pose risks to society?
- How do we build human-GenAI systems that can be safe, result in appropriate trust?