



THE UNIVERSITY
OF QUEENSLAND
AUSTRALIA

CREATE CHANGE

Humans and AI: Persuasion and Trust

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)
 - LLM (COLING25, **CHI25**), Misinfo (ECIR20 Best SP, SIGIR20, CIKM20, IP&M, ICWSM24)
- **Better Crowdsourcing Platforms** (since 2013)
 - Platforms (WWW15, CSCWJ18, **CACM25**), Experiments (CSCW21), Pricing (HCOMP14)
 - Task Allocation (WWW13, WWW16, COR), Workers (CHI15, CSCW20 Hon. Mention)
 - Metadata (IP&M), Attacks (HCOMP18 Best P, JAIR), Time (HCOMP16)
 - Modus Operandi (UBICOMP17, HT19, WSDM20, TOIS24), Complexity (HCOMP16)
 - Abandonment (WSDM19, TKDE, ACM TSC)
- **Data Bias** (since 2018)
 - Gender (w/ Wiki; SIGIR18, ACIS24, **WWW25**), Management (CACM24, **WWW25**),
 - Impact on ML (CIKM22), SES (WebSci22, **ICWSM25**), Political (**WWW25**)
- **Better Data** (since 2019)
 - Noise (WWW19), Data Workers (SIGIR20, TOIS, TKDE, WWW23), Behaviors (CIKM20)
 - Know. Graphs (ISWC19), Unknown Unknowns (ECAI20, HCOMP21)
 - Fairness (CIKM22, SIGIR23, FAccT24, KDD24), Active Learning (AAAI24)

Thanks to:



Australian Government
Australian Research Council



Outline

Generating Persuasive Content at Scale

- AI-generated personalised social media ads (ACM TheWebConf 2024)

Detecting Harmful Content with LLMs as Data Preprocessors

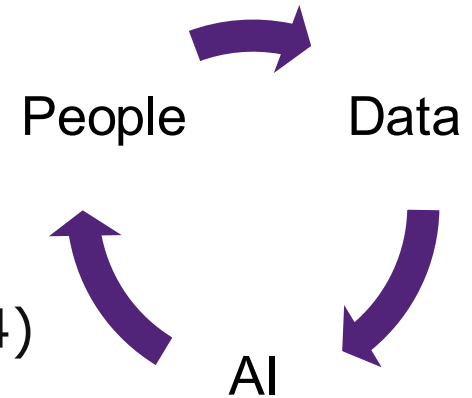
- LLM-based Data Pipelines (ICWSM 2025)

Controlling Bias in LLMs

- Persona-based LLMs (ACM TheWebConf 2025)
- Bias Management (CACM Jan 2024)
- The BiasNavi tool (ACM TheWebConf 2025)

Do we Trust LLM Agents?

- LLMs to complete tasks for us (ACM CHI 2025)
- Crowd-sourcing or AI-Sourcing? (CACM Apr 2025)



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.



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.

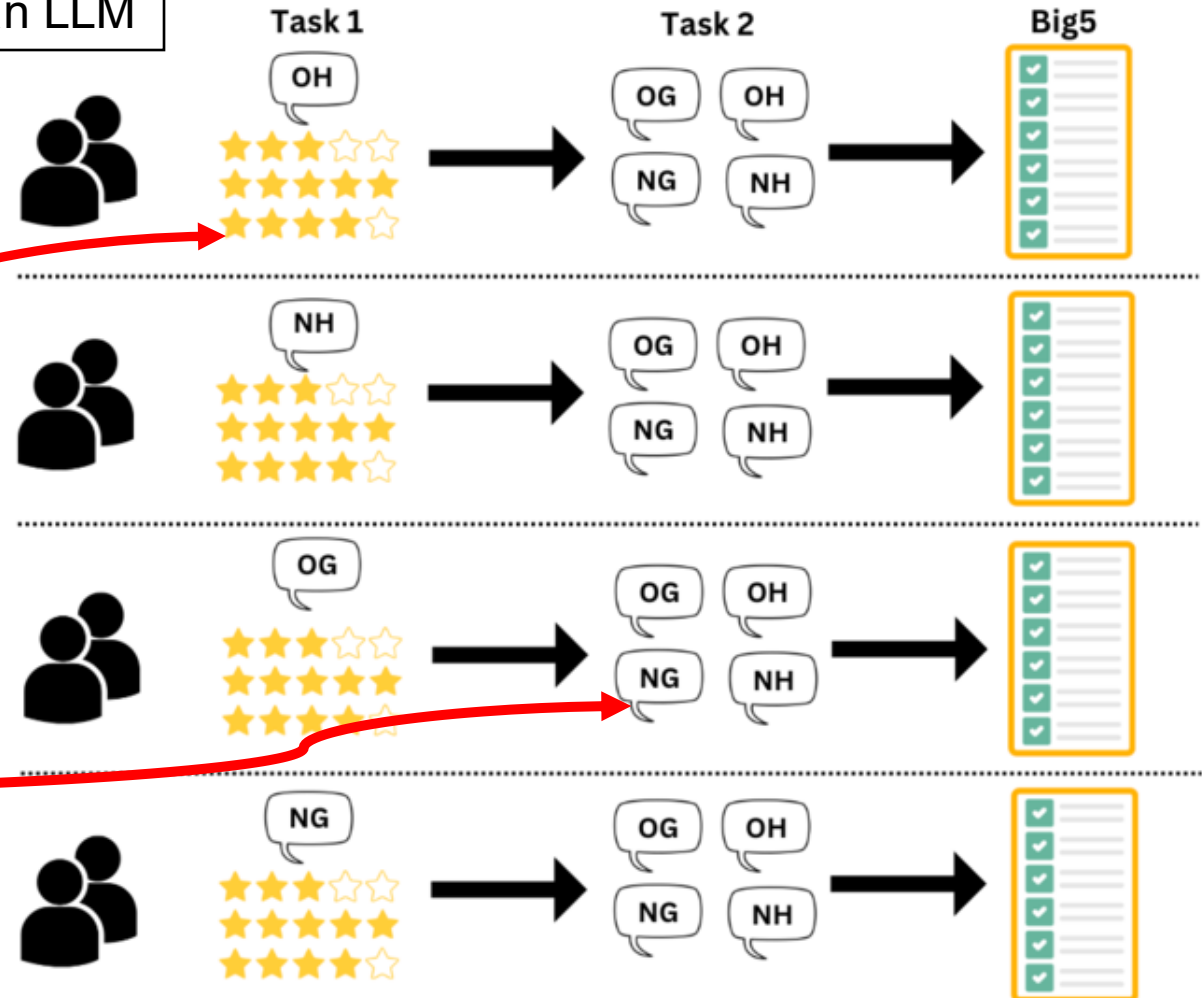
Study

OH: Openness, written by a **H**uman
OG: Openness, **G**enerated by an LLM
NH: Neuroticism, written by a **H**uman
NG: Neuroticism, **G**enerated by an LLM

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

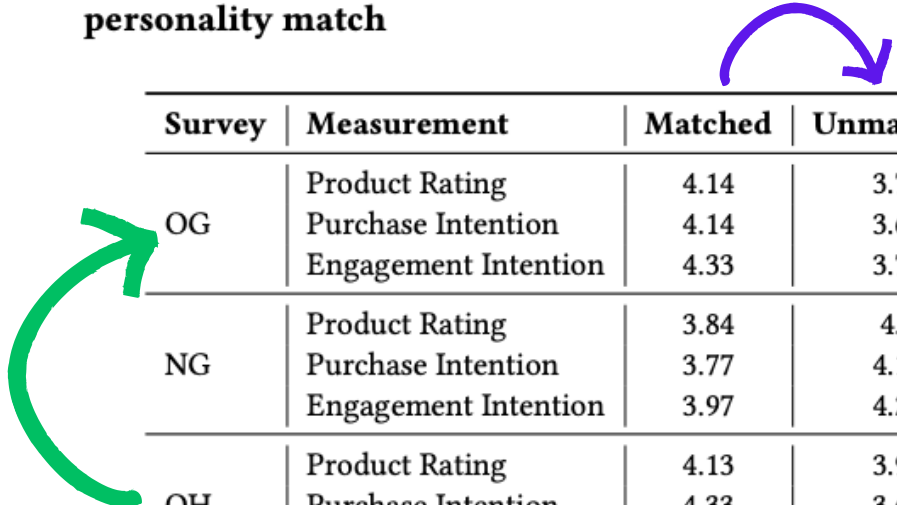
1. Product attitude
2. Purchase intention
3. Engagement intention

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

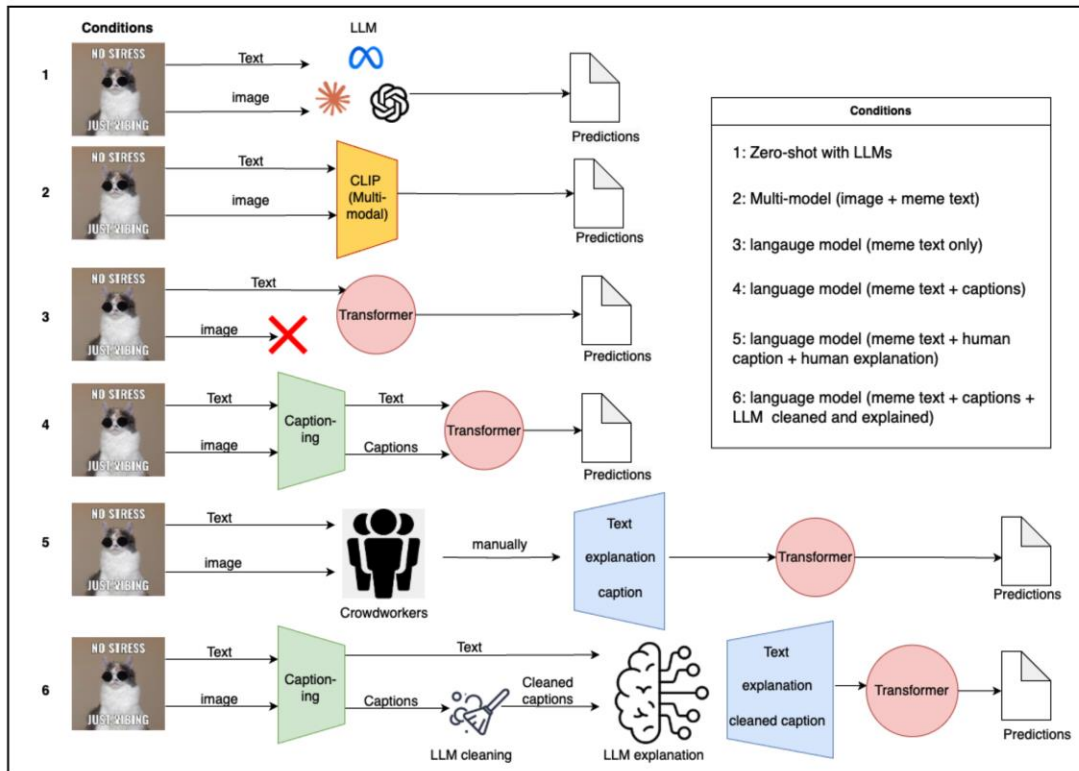
Results - Task 2

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

LLM-based Data Pipelines to Detect Harmful Content

Facebook Hateful Meme Challenge: classify content as hateful or non-hateful



1. Zero-Shot with Meme Image and Text
2. Image + Text (Multimodal Model; CLIP)
3. Meme Text Only (Language Model; DistilBERT, RoBERTa)
4. Text + Captions (Language Model)
5. Human Captions and Explanations (Language Model)
6. Meme Text + Cleaned Captions + LLM Explanations (Language Model):

Elyas Meguellati, Assaad Zeghina, Shazia Sadiq, and Gianluca Demartini. **LLM-based Semantic Augmentation for Harmful Content Detection**. In: 19th International AAAI Conference on Web and Social Media (ICWSM 2025). Copenhagen, Denmark, June 2025.

Findings

- **Does LLM-Based Caption Cleaning Work? (RQ1a)**
 - GPT-4o-cleaned captions showed significant improvements over the uncleaned captions for the classifier ($p = 0.0157$)
- **Does Adding Context Improve Performance? (RQ1b)**
 - Leveraging LLMs to augment each meme with a short, explanatory context yields performance gains
 - Including meme text, caption and LLM-generated explanation yields strongest performance
- **Generalizability Across Related Domains (RQ2)**
 - The approach generalizes well across social media tasks (Jigsaw Toxic Comments and Facebook Hateful Memes) with differing data modalities (text vs multimodal)

A fundamental distinction between LLM explanatory capabilities and predictive performance

Observations

- LLM are not good harmful content detector if used as zero-shot classifiers
- LLM are good at segmenting, explaining, and providing more context for downstream harmful content classification
- LLM-based semantic augmentation is effective for context-dependent tasks
- Reduced manual annotation costs
- **Safeguard mechanisms** embedded in LLMs limit performance on harmful content
 - Managing LLM safeguards by asking to preserve triggers
 - Important in domains where capturing explicit terms or themes is critical for model training

Explanation: “This meme implies that interacting with anything connected to Islam is dangerous or undesirable, feeding into a narrative that paints Muslims as inherently threatening or alien. By humorously suggesting that no one dares pull the doll’s string, the meme mocks and perpetuates fears of Islam.”

Triggers: Islamophobia, Stereotyping, Muslim doll, what the fuck, no one has the guts.

Outline

Generating Persuasive Content at Scale

- AI-generated personalised social media ads (ACM TheWebConf 2024)

Detecting Harmful Content with LLMs as Data Preprocessors

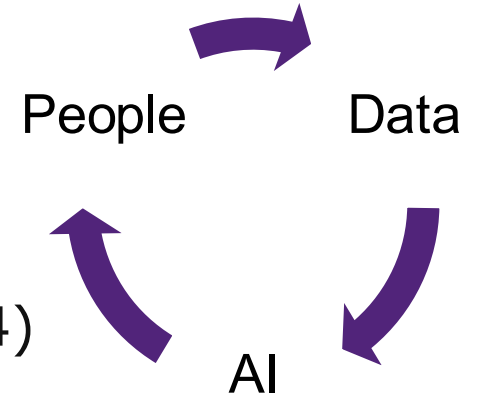
- LLM-based Data Pipelines (ICWSM 2025)

Controlling Bias in LLMs

- Persona-based LLMs (ACM TheWebConf 2025)
- Bias Management (CACM Jan 2024)
- The BiasNavi tool (ACM TheWebConf 2025)

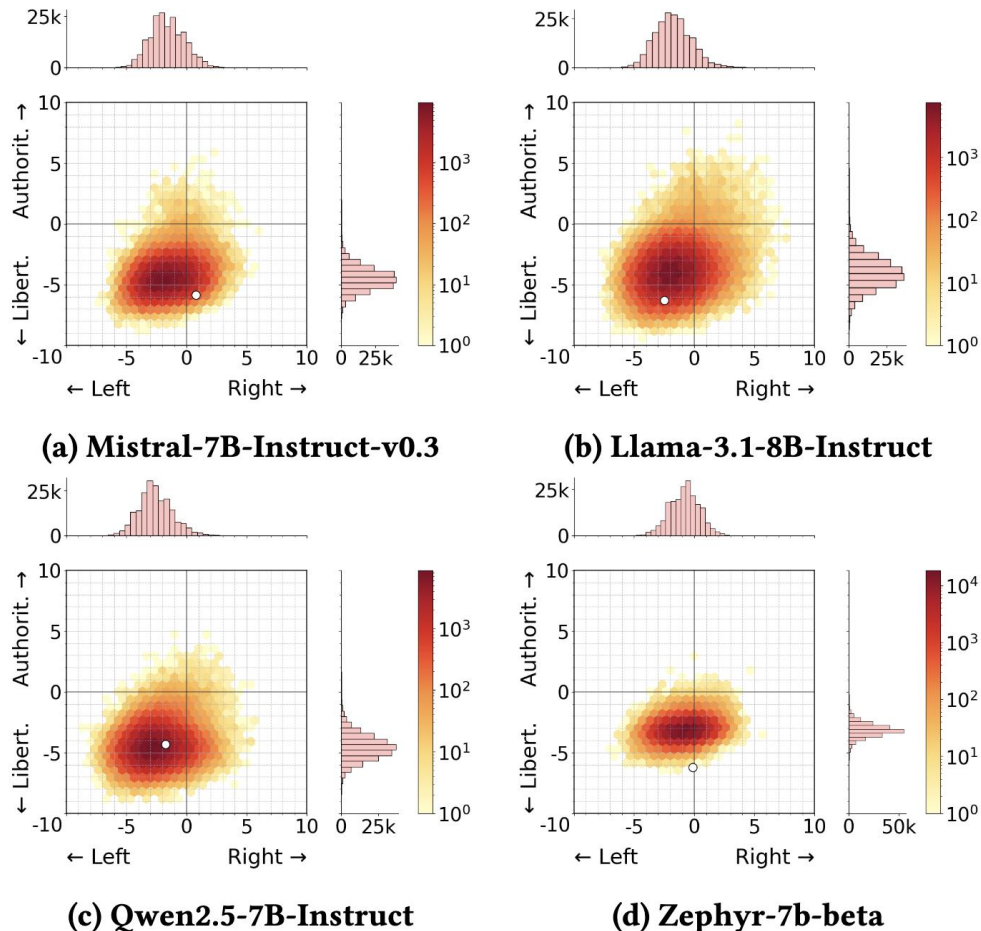
Do we Trust LLM Agents?

- LLMs to complete tasks for us (ACM CHI 2025)
- Crowd-sourcing or AI-Sourcing? (CACM Apr 2025)



Persona-based LLMs

- We make LLMs answer the Political Compass Test
- We then make them impersonate 200,000 personas and answer the PCT again
- This shows how we can measure and control the political bias of LLMs.
- It also highlights embedded stereotypes like “A business developer trying to bring new investments to the region, regardless of environmental cost” being authoritarian right



Pietro Bernardelle, Leon Fröhling, Stefano Civelli, Riccardo Lunardi, Kevin Roitero, and Gianluca Demartini. **Mapping and Influencing the Political Ideology of Large Language Models using Synthetic Personas.** In: The 2025 ACM Web Conference (TheWebConf 2025) - Short paper track. Sydney, Australia, April 2025.

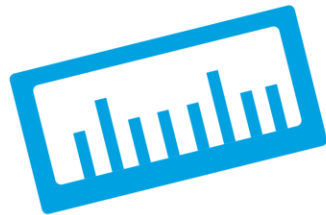
Bias Management, not bias removal

Employing an explicit and not transparent bias removal intervention might be potentially harmful to the user

Figure 2. The five steps of bias management.



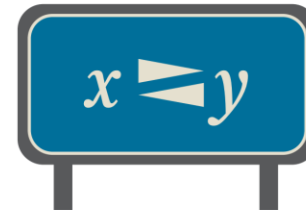
Identifying



Measuring



Indexing



Surfacing




Adapting

<https://doi.org/10.1145/3611641>

BiasNavi

<https://github.com/CIRES-Hub/BiasNavi/>


BiasNavi
Import ▾ Export ▾ LLM Models ▾ View ▾ Prompts User Profile Help ▾ More ▾

Bias Management Pipeline

Identify Measure Surface Adapt

Recommended Operation: Evaluate the Dataset to Get Disparity Scores

Chat with BiasNavi

Common Questions

sophisticated techniques like Tomek links.

3. **Use of Weighting:** Assign a higher weight to the minority class during model training, so the model gives more importance to predicting the minority class correctly.

4. **Data Augmentation:** If applicable, augment the data by creating new features or transforming existing ones to create a more balanced representation.

5. **Example Step-by-step:**

- Begin by calculating the racial distribution in your dataset.
- Choose a resampling method (e.g., SMOTE) and apply it to balance the dataset.
- Ensure the new dataset maintains the integrity of the original one by checking for any new biases introduced.
- Test your model with the rebalanced dataset and compare results with the original dataset to evaluate improvements in fairness.

By following these steps, you can effectively address racial imbalances and create a more equitable dataset, aligning with your expertise and technical proficiency.

Suggested Next Question

What specific resampling techniques are most suitable for my industry sector to achieve better racial balance in datasets?

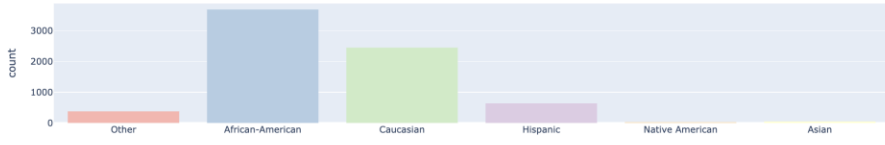
Suggested Next Question

How can I validate that the rebalancing techniques applied have effectively reduced bias in the dataset?

Data Statistics Save Snapshot Download Go to Rows Start row End row

id	name	first	last	compas_screening_date	sex	dob	age	age_cat
1	miguel hernandez	miguel	hernandez	2013-08-14	Male	1947-04-18	69	Greater than 45
3	kevon dixon	kevon	dixon	2013-01-27	Male	1982-01-22	34	25 - 45
4	ed philo	ed	philo	2013-04-14	Male	1991-05-14	24	Less than 25
5	marcu brown	marcu	brown	2013-01-13	Male	1993-01-21	23	Less than 25
6	bouthy pierrelouis	bouthy	pierrelouis	2013-03-26	Male	1973-01-22	43	25 - 45
7	marsha miles	marsha	miles	2013-11-30	Male	1971-08-22	44	25 - 45
8	edward riddle	edward	riddle	2014-02-19	Male	1974-07-23	41	25 - 45
9	steven stewart	steven	stewart	2013-08-30	Male	1973-02-25	43	25 - 45
10	elizabeth thieme	elizabeth	thieme	2014-03-16	Female	1976-06-03	39	25 - 45
13	bo bradac	bo	bradac	2013-11-04	Male	1994-06-10	21	Less than 25

« < 1 / 722 > »



count

Other African-American Caucasian Hispanic Native American Asian

race

Bias Management

Identify Bias

Target Attribute: score_text

Result of Bias Identifying

Sensitive Attributes:

- Age (age, dob):** Age is often considered a sensitive attribute because it can influence assessments and outcomes, potentially leading to age discrimination.
- Race:** Race is a well-known sensitive attribute due to its strong association with biases in various societal and legal contexts, especially in criminal justice.

Dataset Snapshots

ID	Description	Timestamp
1	Original	2025-02-26 06:16:20

Restore Delete

Dataset Evaluation

Experiment Comparison

Snapshot: 1 Sensitive Attribute: sex

Label: score_text Task: Classification Model: SVM

Run

Results

Accuracy: 0.9965

sex	Low	High	Medium	Group Count (for Test)
Female	0.5840	0.1360	0.2800	250
Male	0.5264	0.2196	0.2540	1193

Disparity Score 0.0576 0.0836 0.0260

The dataset analysis shows some notable disparities based on the 'sex' attribute. Here's a breakdown of the bias level assessment:

- Disparity in Score Distribution:**
 - For females, the distribution of scores is 58.4% Low, 13.6% High, and 28% Medium.
 - For males, the distribution is 52.64% Low, 21.96% High, and 25.40% Medium.
 - The disparity scores indicate that females are more likely to receive a 'Low' score, while males are more likely to receive a 'High' score.
- Disparity Score Analysis:**
 - The disparity score for 'Low' is 0.0576, indicating females are more likely to receive a 'Low' score compared to males.
 - The 'High' and 'Medium' disparity scores are 0.0836 and 0.0260, respectively, showing a higher likelihood of males receiving 'High' and 'Medium' scores.
- Model Accuracy:**
 - The model's accuracy is 99.65%, which is quite high, but it is essential to ensure that this does not come at the cost of fairness.

Junliang Yu, Jay Thai Duong Huynh, Shaoyang Fan, Gianluca Demartini, Tong Chen, Hongzhi Yin, and Shazia Sadiq. **BiasNavi: LLM-Empowered Data Bias Management.** In: The 2025 ACM Web Conference (TheWebConf 2025) - Demo track. Sydney, Australia, April 2025

Outline

Generating Persuasive Content at Scale

- AI-generated personalised social media ads (ACM TheWebConf 2024)

Detecting Harmful Content with LLMs as Data Preprocessors

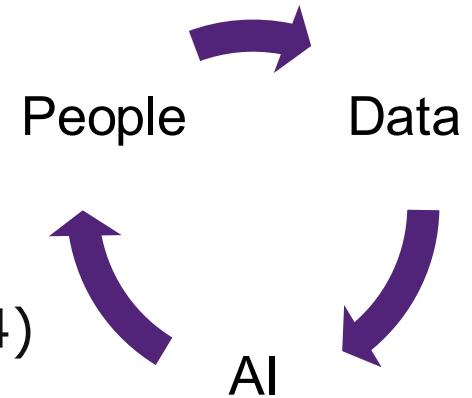
- LLM-based Data Pipelines (ICWSM 2025)

Controlling Bias in LLMs

- Persona-based LLMs (ACM TheWebConf 2025)
- Bias Management (CACM Jan 2024)
- The BiasNavi tool (ACM TheWebConf 2025)

Do we Trust LLM Agents?

- LLMs to complete tasks for us (ACM CHI 2025)
- Crowd-sourcing or AI-Sourcing? (CACM Apr 2025)



Trust in AI Agent

Agents as daily assistants

Tasks with different levels of risk

LLM agents used in a plan-then-execute manner

A double-edged sword

(1) they can work well with a high-quality plan and necessary user involvement

(2) users can easily mistrust the LLM agents with plans that seem plausible

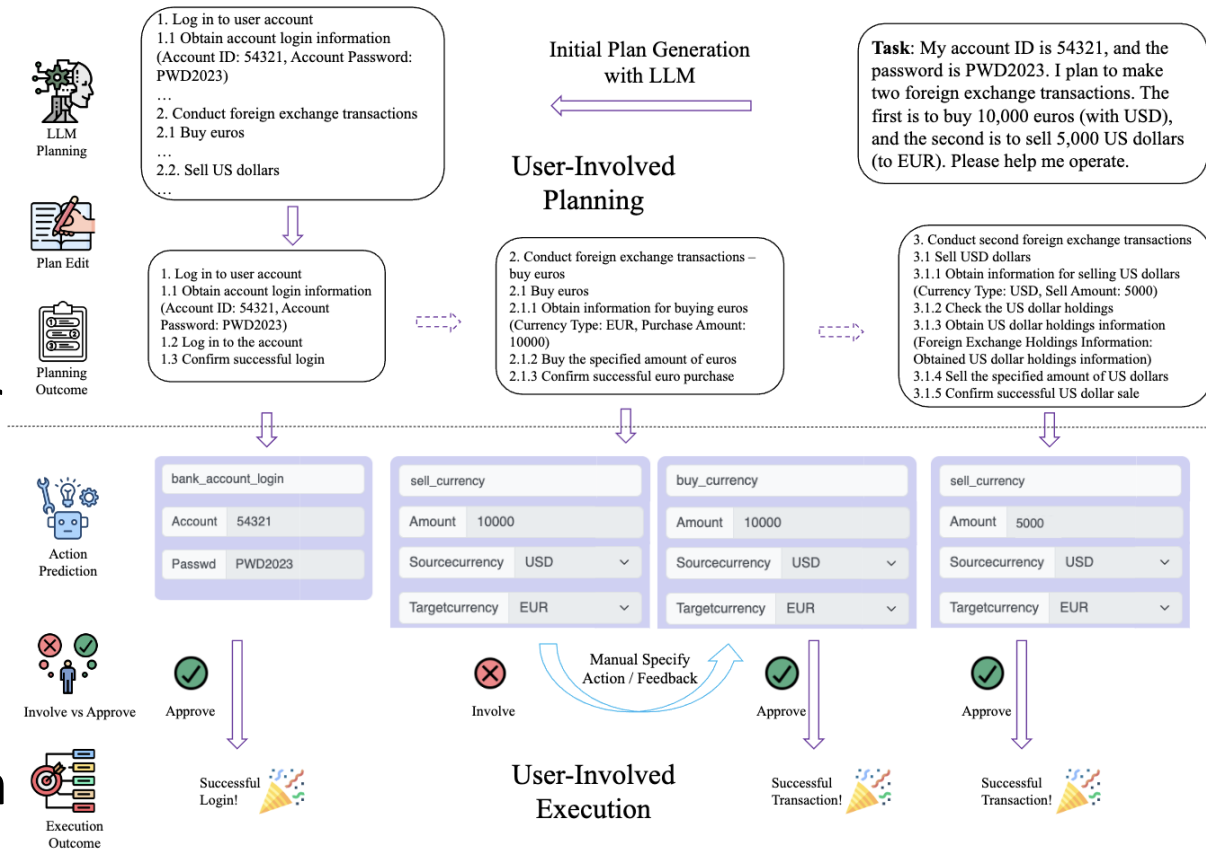


Figure 1: Illustration of the human-AI collaboration with plan-then-execute LLM agents.

Gaole He, Gianluca Demartini, and Ujwal Gadiraju. **Plan-Then-Execute: An Empirical Study of User Trust and Team Performance When Using LLM Agents As A Daily Assistant.** In: ACM CHI 2025 Conference on Human Factors in Computing Systems (CHI 2025). Yokohama, Japan, April 2025.

Key Findings

- User involvement does not significantly impact user trust and calibrated trust
 - User involvement in planning can harm plan quality in tasks with a high-quality plan
 - Plan quality has a significant positive correlation with calibrated trust
 - User involvement in planning can help address imperfect plans
-
- Recommended approaches:
 - Iterative LLM agent simulation to decide when users should be involved
 - Users may need to articulate or manually override the agent action, posing a high cognitive load

Task Description:

I need to set an alarm for every weekday morning at 7:30, and then cancel the alarm for Thursday, changing it to 8:00 in the evening.

Show Potential ActionsPlan Edit InstructionAdd one step

Plan:

1.	Set the alarm for every day	Split step	Delete step
1.1	Get the alarm setting information (Time: 07:30 AM, Frequency: Monday to Friday)	Split step	Delete step
1.2	Set the alarm	Split step	Delete step
1.3	Confirm whether the alarm is set successfully	Split step	Delete step
2.	Cancel the Thursday alarm	Split step	Delete step
2.1	Get the information of the alarm to be cancelled (Time: 07:30 AM, Frequency: Thursday)	Split step	Delete step
2.2	Cancel the alarm	Split step	Delete step

Strong use of chatGPT
Especially on Amazon MTurk

Generative AI in Crowdwork

	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%

We asked crowd workers regarding their use of GenAI tools. Table 4: Workers reporting self-initiated use of AI chatbots in tasks, by platform, country and T1/T2 [top/bottom].

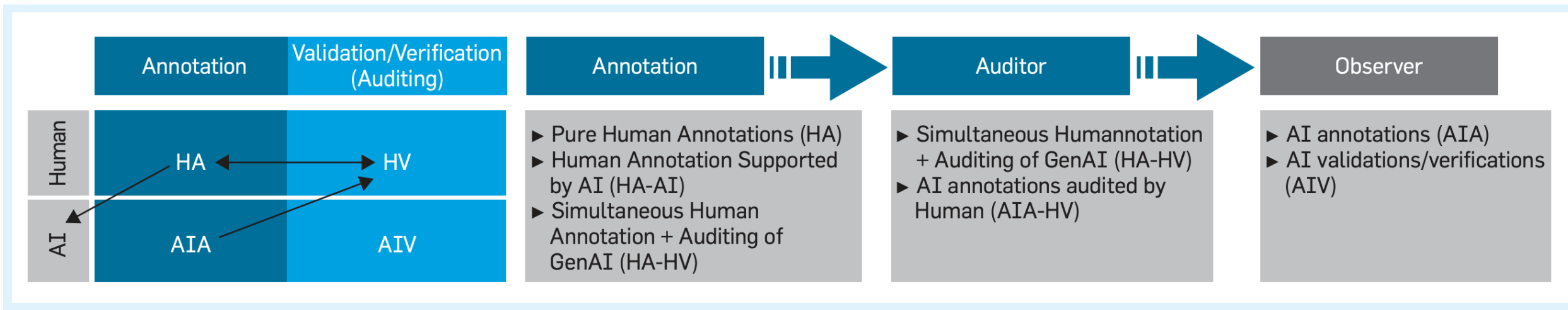
Prolific, Mturk, Clickworker; May 2023, and Dec 2023

- 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).

Crowd-Sourcing or AI-Sourcing?

There will always be a role for humans in AI pipelines, although GenAI is disrupting the crowdsourcing environment as we know it.



Evgenia Christoforou, Gianluca Demartini, and Jahna Otterbacher. **Crowd-Sourcing or AI-Sourcing? - The Impact of GenAI on Data Annotation Tasks.** In: Communications of the ACM (CACM), Vol. 68, No. 4 April 2025.

Lessons learned and what to do

- LLM can generate persuasive content and understand harmful content
- LLMs can replace humans in many tasks, but should they?
- Crowd workers over-rely on LLMs to label data. Is this the end of crowdsourcing?
- Track and profile data bias across the AI pipelines
- Select and diversify the sources of the labels (i.e., human annotators, LLMs)
- **Bias management** instead of bias removal



DOI:10.1145/3611641

Gianluca Demartini, Kevin Roitero, and Stefano Mizzaro

Opinion Data Bias Management

*Envisioning a unique approach
toward bias and fairness research.*

THE PRESENCE of bias in data has led to a lot of research include work looking at how to remove bias from learned word embeddings. increase fairness across groups when doing data augmentation,¹⁷ feature

Demartini et al. “**Data Bias Management**”, in *Communications of the ACM*, Vol. 67, No. 1, Jan 2024

To be continued ...

Visiting PhD Students Scheme

Visit us in Brisbane, Australia!

2 or 3 months visits for PhD students to work on a joint paper

Funding and application instructions: <https://cires.org.au/engagement/visitors/>

Application deadlines in 2025:

March 22; June 22; September 22



Since 2023, we hosted 8 PhD students based in 6 countries
(CH, NL, DE, NO, BE, CN)



Gaole He, Gianluca Demartini, and Ujwal Gadiraju. **Plan-Then-Execute: An Empirical Study of User Trust and Team Performance When Using LLM Agents As A Daily Assistant.** In: ACM CHI 2025 Conference on Human Factors in Computing Systems (CHI 2025). Yokohama, Japan, April 2025.

Mads Skipanes, Tollef Emil Jørgensen, Kyle Porter, Gianluca Demartini, and Sule Yildirim Yayilgan. **Enhancing Criminal Investigation Analysis with Summarization and Memory-based Retrieval-Augmented Generation: A Comprehensive Evaluation of Real Case Data.** In: The 31st International Conference on Computational Linguistics (COLING 2025).