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

European Commission

Swiss National Science Foundation

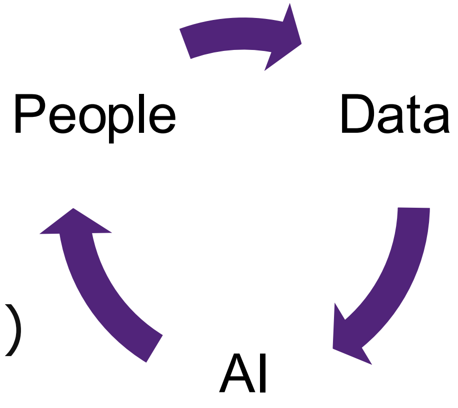
EPSRC
Engineering and Physical Sciences Research Council

Meta

WIKIMEDIA FOUNDATION

Google AI

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

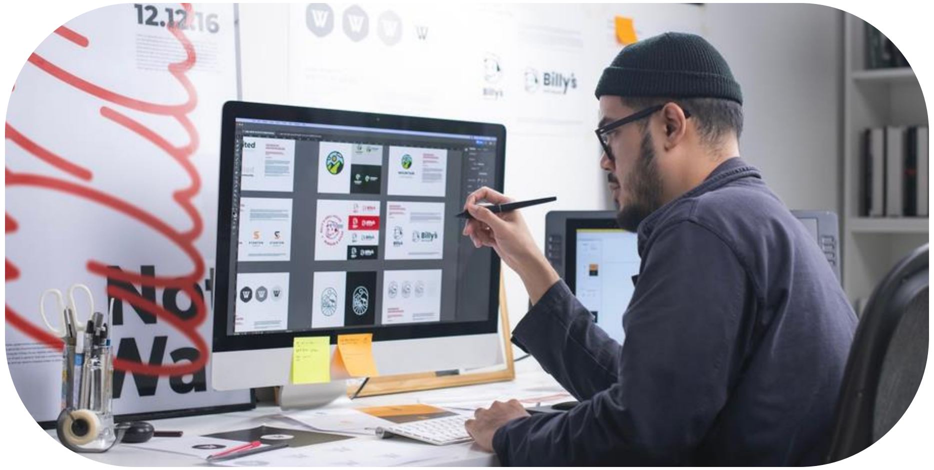
- Persona-based LLMs (ACM TheWebConf 2025)
- Bias Management (CACM Jan 2024)
- The BiasNavi tool (ACM TheWebConf 2025)

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.

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.

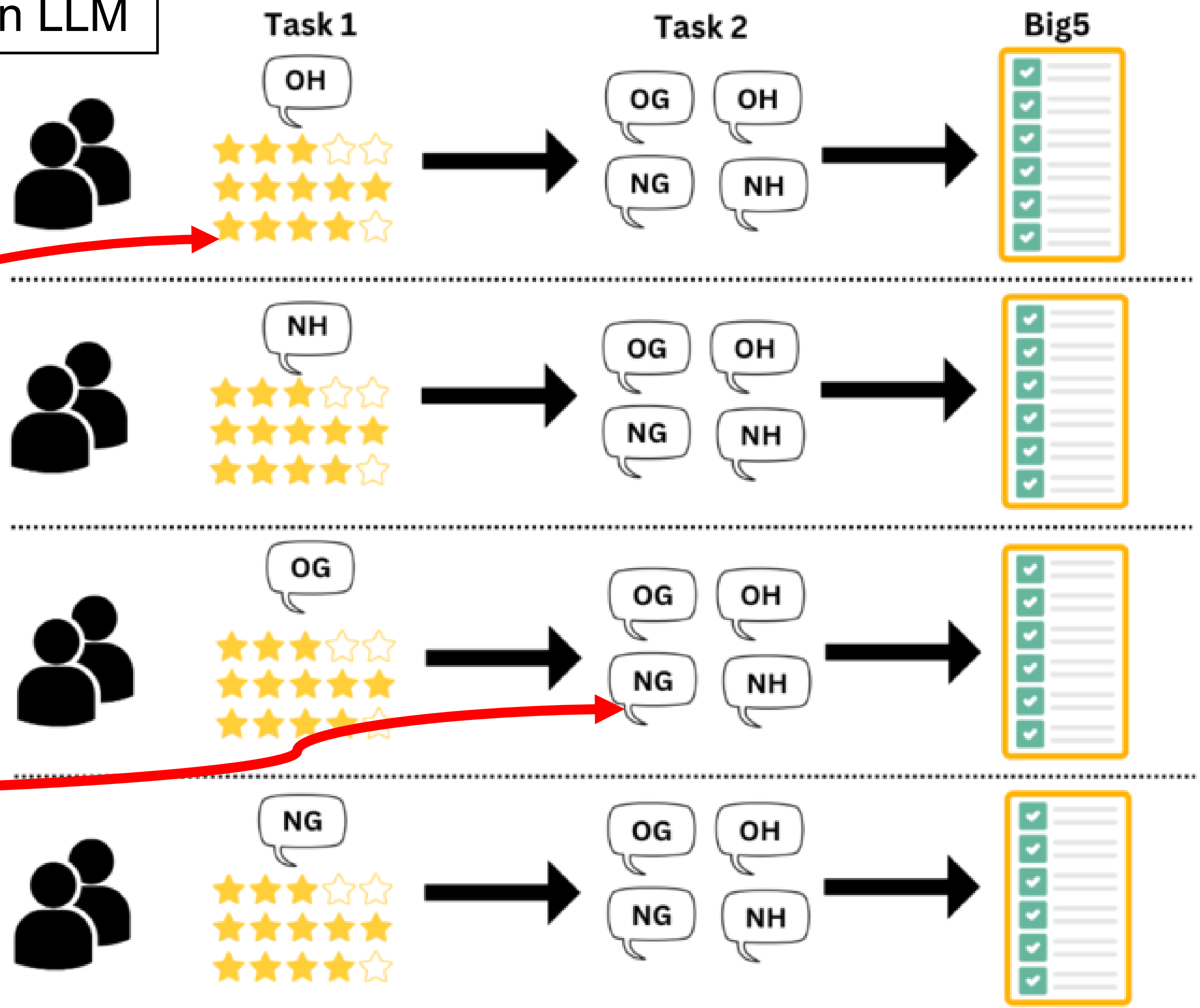
OH: **O**penness, written by a **H**uman
OG: **O**penness, **G**enerated by an LLM
NH: **N**euroticism, written by a **H**uman
NG: **N**euroticism, **G**enerated by an LLM

Study

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

Results - Task 2

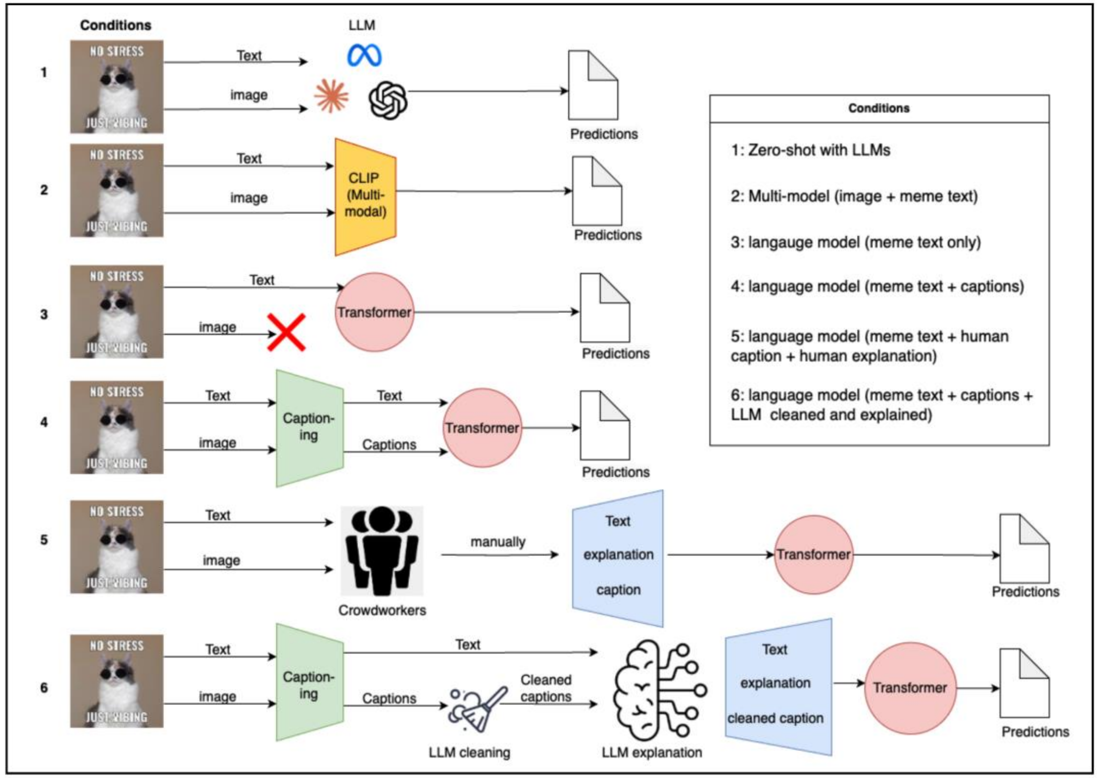
Ads crafted for openness works best
Human and AI generated ads perform equally good

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

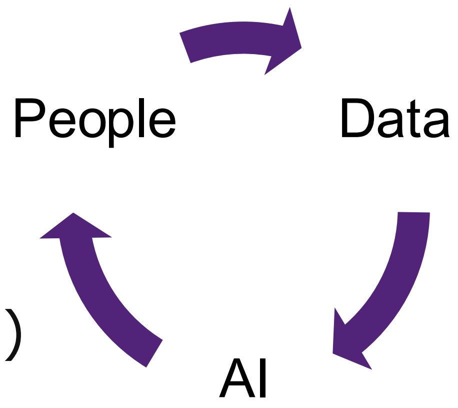
- LLM are not good harmful content detector if used as zero-shot classifiers
- LLM are good at segmenting, explaining, and providing more context for downstream harmful content classification
- LLM-based semantic augmentation is effective for context-dependent tasks
  - Reduced manual annotation costs

- **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)

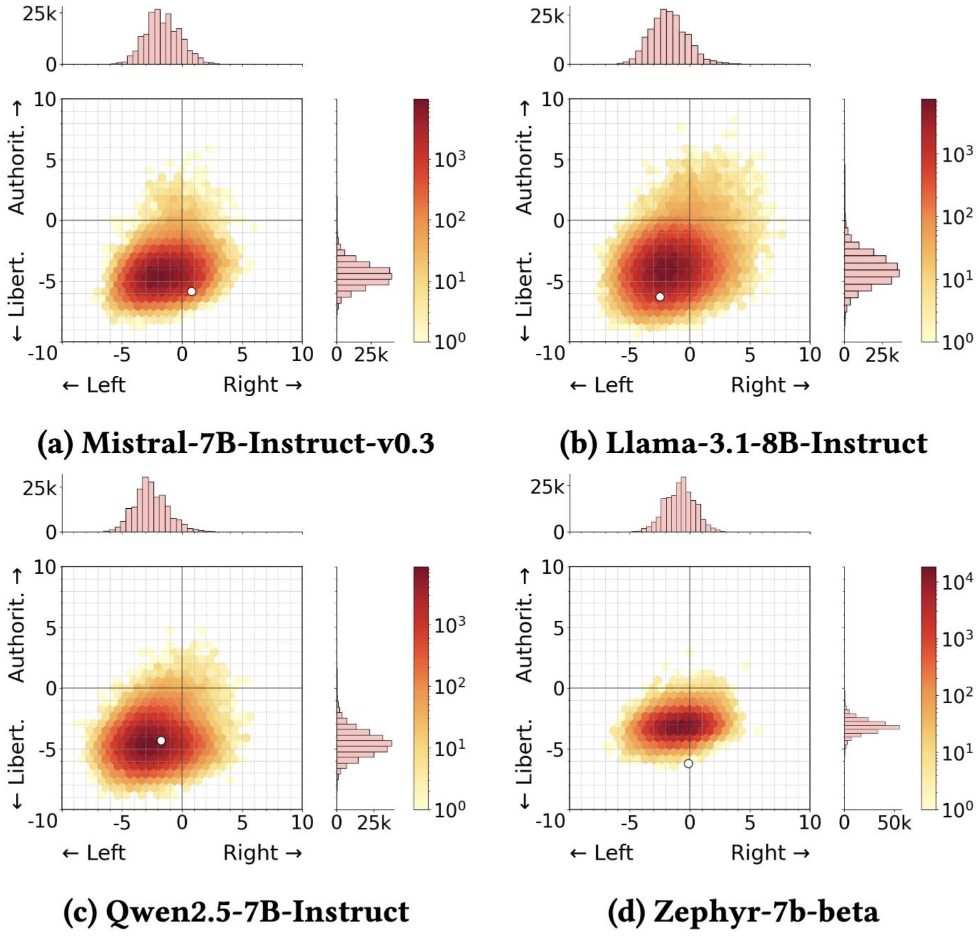
Controlling Bias in LLMs

- 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



(a) Mistral-7B-Instruct-v0.3
(b) Llama-3.1-8B-Instruct
(c) Qwen2.5-7B-Instruct
(d) Zephyr-7b-beta

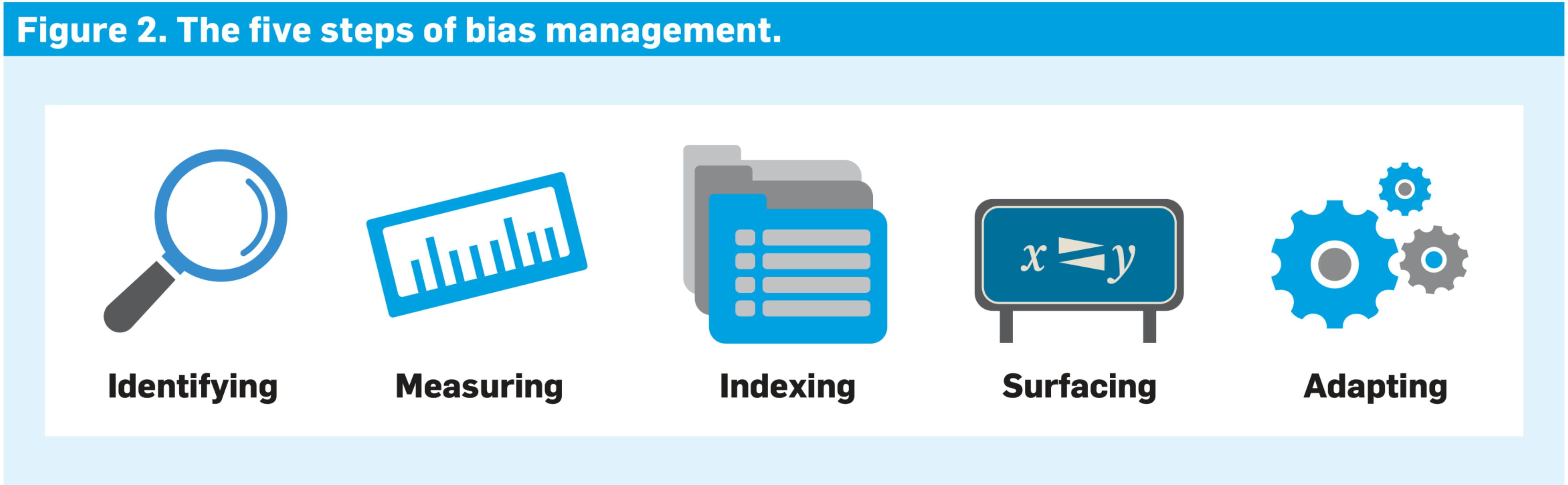
- We make LLMs answer the Political Compass Test
- We then make them impersonate 200,000 personas and answer the PCT again

- 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

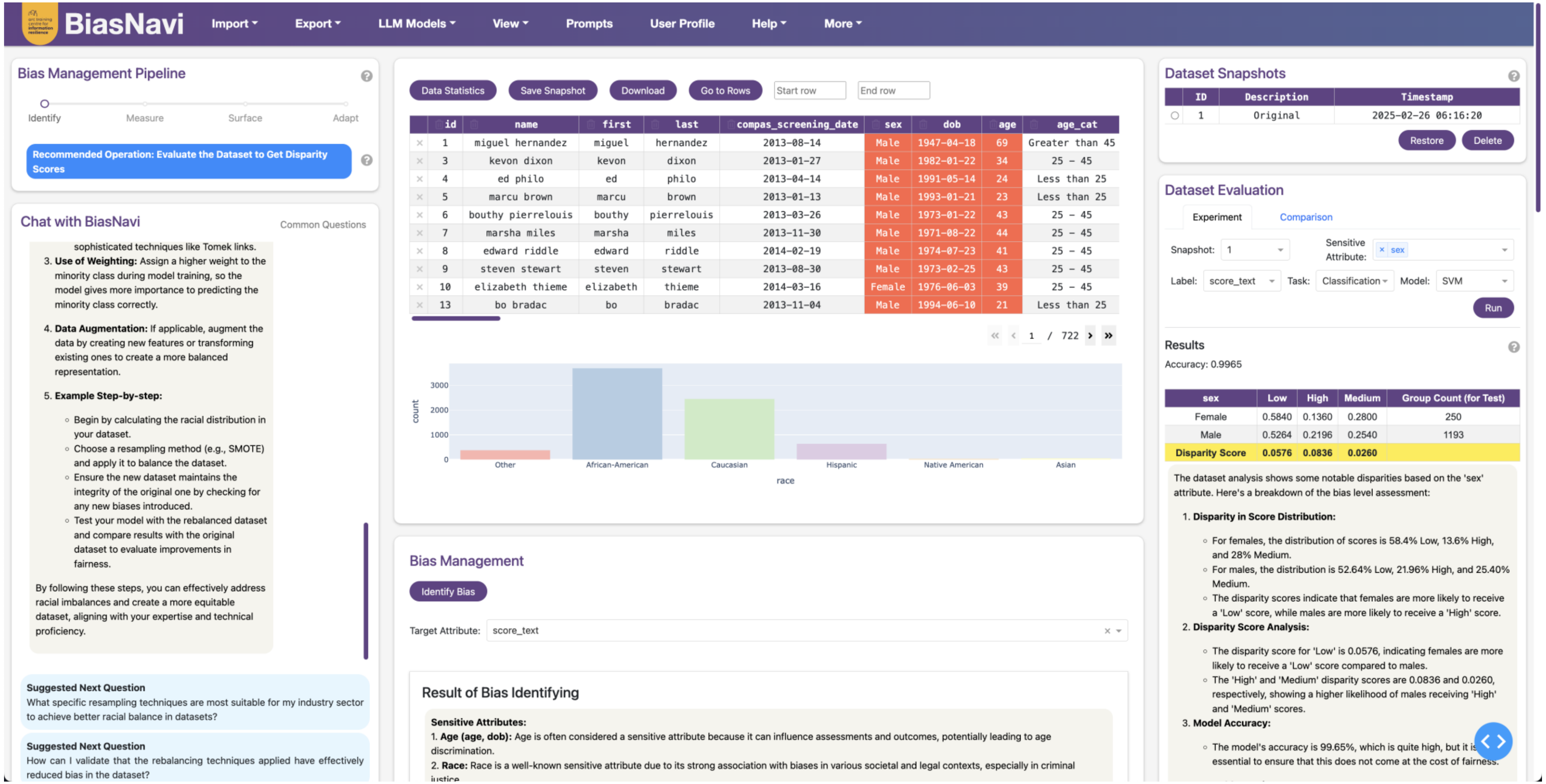
Employing an explicit and not transparent bias removal intervention might be potentially harmful to the user



https://doi.org/10.1145/3611641

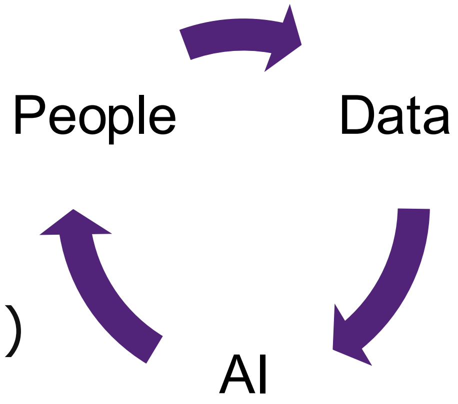
BiasNavi

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



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)

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

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

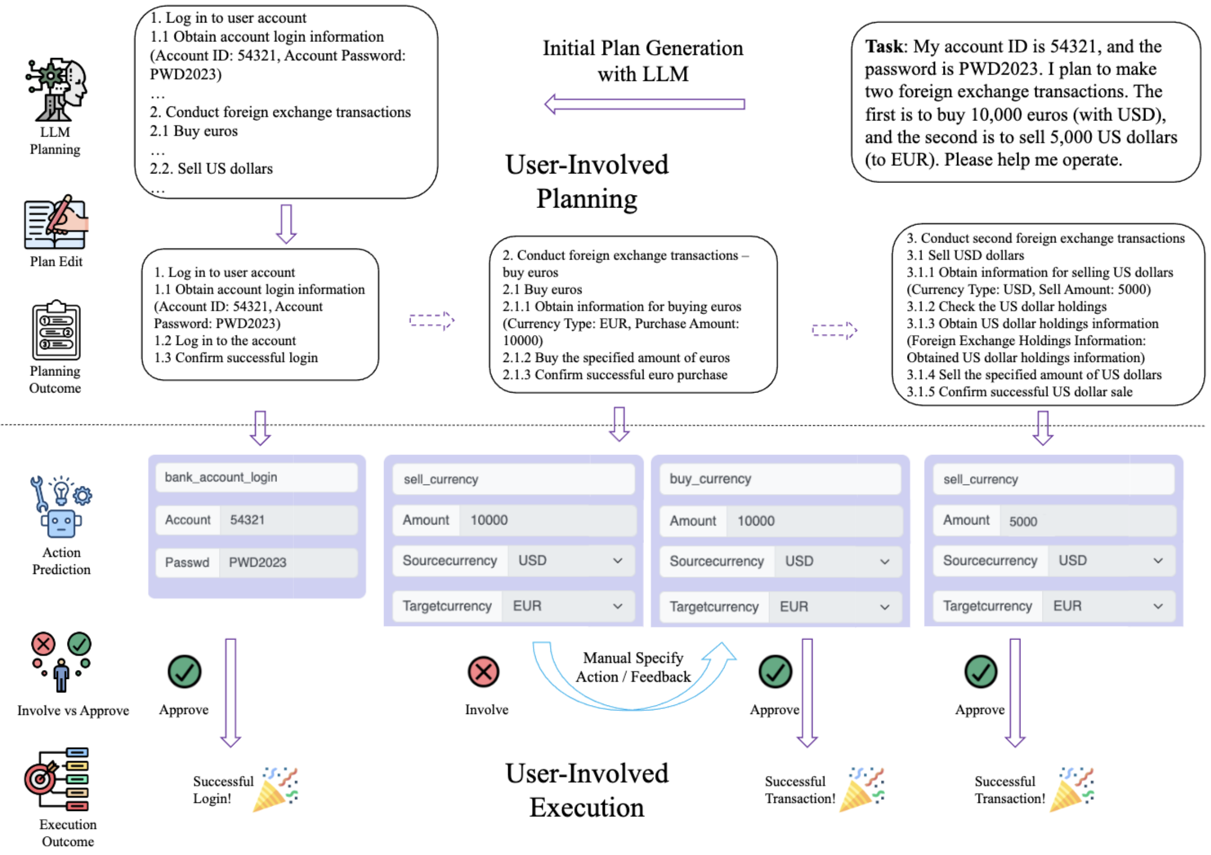


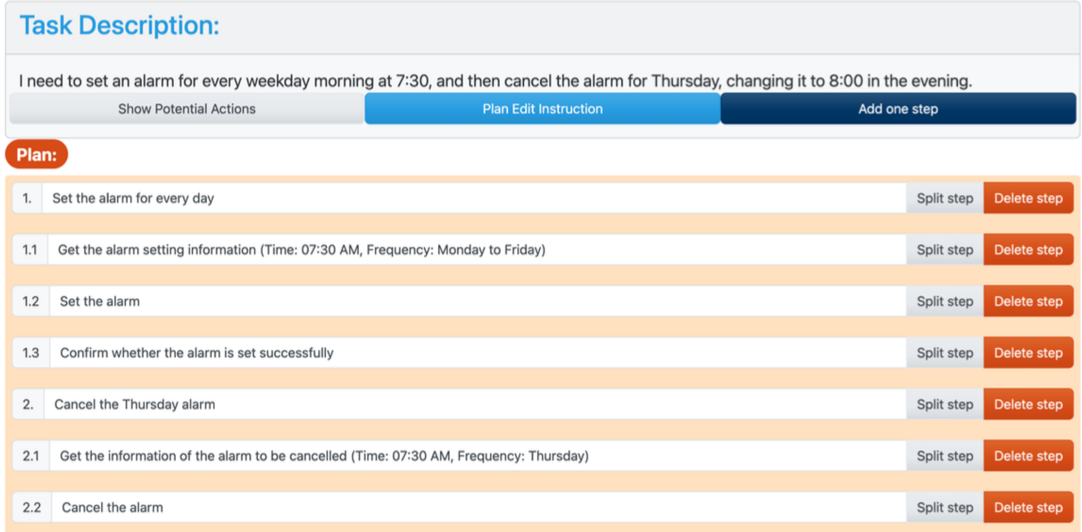
Figure 1: Illustration of the human-AI collaboration with plan-then-execute LLM agents.

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



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.

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].

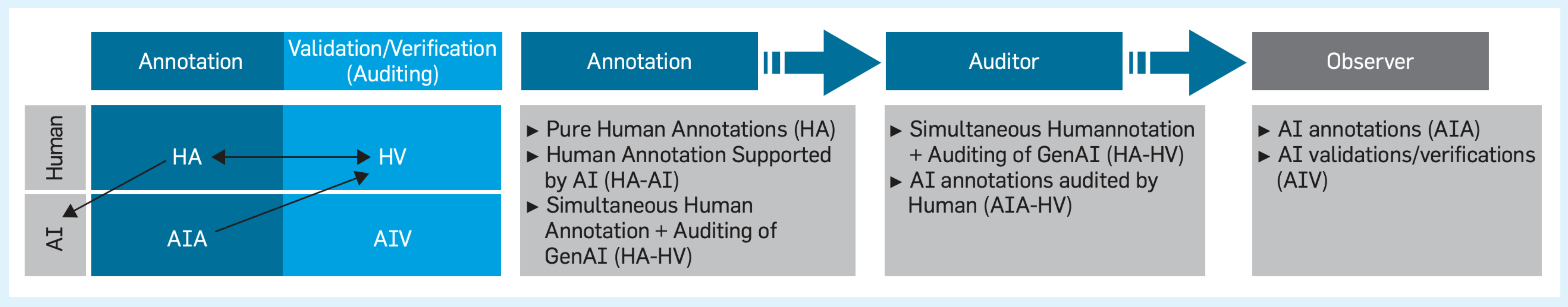
Prolific, Mturk, Clickworker; May 2023, and Dec 2023

- 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.

gianlucademartini.net
demartini@acm.org
@eglu81

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

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


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

To be continued …

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

arc training centre for **information resilience**

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).