Missing the Margins: A Systematic Literature Review on the Demographic Representativeness of LLMs

Indira Sen[1], Marlene Lutz[1], Elisa Rogers[1], David Garcia[2], and Markus Strohmaier[1,3]
[1]University of Mannheim, [2]University of Konstanz, [3]GESIS - Leibniz Institute for the Social Sciences
indira.sen@uni-mannheim.de, marlene.lutz@uni-mannheim.de, elisa.marie-rogers@students.uni-mannheim.de,
david.garcia@uni-konstanz.de, markus.strohmaier@uni-mannheim.de

Abstract

Many applications of Large Language Models (LLMs) require them to either simulate people or offer personalized functionality, making the demographic representativeness of LLMs crucial for equitable utility. At the same time, we know little about the extent to which these models actually reflect the demographic attributes and behaviors of certain groups or populations, with conflicting findings in empirical research. To shed light on this debate, we review 211 papers on the demographic representativeness of LLMs. We find that while 29% of the studies report positive conclusions on the representativeness of LLMs, 30% of these do not evaluate LLMs across multiple demographic categories or within demographic subcategories. Another 35% and 47% of the papers concluding positively fail to specify these subcategories altogether for gender and race, respectively. Of the articles that do report subcategories, fewer than half include marginalized groups in their study. Finally, more than a third of the papers do not define the target population to whom their findings apply; of those that do define it either implicitly or explicitly, a large majority study only the U.S. Taken together, our findings suggest an inflated perception of LLM representativeness in the broader community. We recommend more precise evaluation methods and comprehensive documentation of demographic attributes to ensure the responsible use of LLMs for social applications.

1 Introduction

In addition to their applications as general assistive technology, an emerging use case of LLMs in the (computational) social sciences is the simulation of human behavior, to replicate or augment existing social data like survey responses (Argyle et al., 2023), behavioral experiments (Hewitt et al., 2024) or social network traces (Chang et al., 2024). For LLMs to be an effective tool in both assisting diverse human populations and simulating their behavior, LLMs would need to be representative, i.e., *their behavior would need to validly reflect the underlying target population.* For example, if providing personalized healthcare or educational recommendations, the LLM should be equally assistive to multiple groups of people, and not display lack of background knowledge for certain groups. Similarly, if an LLM is used in social simulations, then it should also be equally effective at emulating the behavior of different groups of people.

In the emerging field of social applications of LLMs, current studies reach opposing conclusions on their demographic representativeness, even when analyzing the same populations using similar techniques and models. For example, while Argyle et al. find that LLMs represent American populations via prompt-induced personas, Santurkar et al. conclude that LLMs only reflect the opinions of certain groups. Additionally, other researchers find that LLMs reduce the variance of behavior within groups and flatten (Wang et al., 2024a) and caricature people (Cheng et al., 2023b). While literature surveys on algorithmic bias in LLMs exist (Gupta et al., 2024; Gallegos et al., 2024; Chu et al., 2024), as well as a scoping review on LLMs supplementing humans in human-subject studies (Agnew et al., 2024), none comprehensively review the fine-grained demographic dimensions probed in social applications of LLMs and their connection to representativeness.

Motivated by this lack of systematic literature analyses, we survey 211 articles across a variety of LLM applications, asking and answering the following research questions: **RQ1: Which demographic dimensions are probed, and in which contexts?**[1] We investigate the demographic categories studied and evaluated in these papers, as well as how these categories are operationalized,

[1]We use the term 'demographic' to include both demographic and sociodemographic groups.

Findings of the Association for Computational Linguistics: ACL 2025, pages 24263–24289

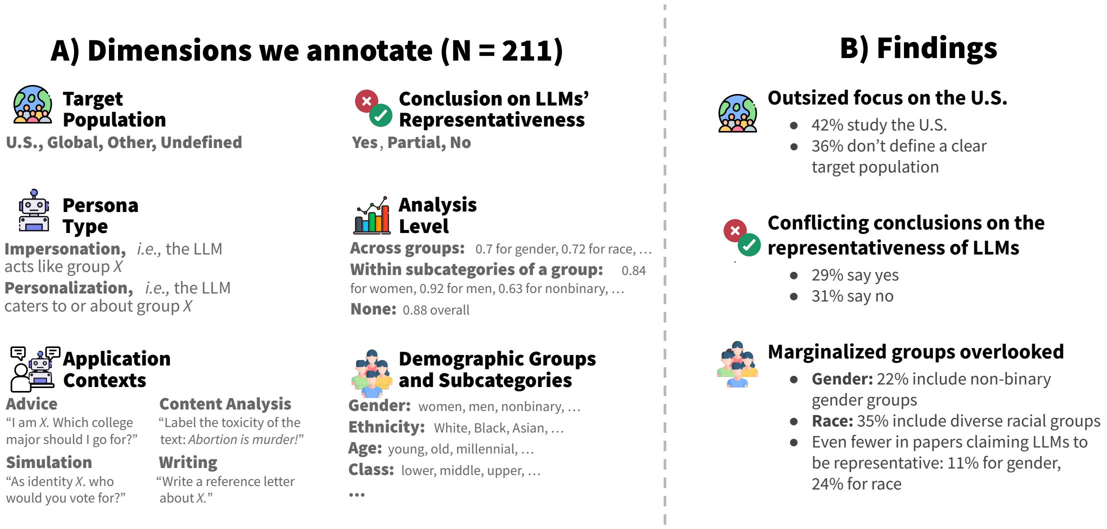


Figure 1: Description of the (A) dimensions we annotate for each paper in this review and (B) key findings related to the demographic representativeness of LLMs.

including which subcategories are considered (Figure 1A). We then assess **if there is consensus on the representativeness of LLMs? (RQ2)**.

Findings: While the majority of papers find no evidence for representativeness or that LLMs are *partially* representative for a certain group only, 29% claim that LLMs are representative. Assessing potential causes of this divergences, we find that among papers claiming LLMs to be representative, 30% do not evaluate representativeness across multiple demographic groups, nor across the subgroups of a particular demographic. Instead these papers report overall LLM representativeness. Another 35% of these papers make conclusions about gender representativeness without reporting the gender subcategories they study. The equivalent proportion for racial categories is even higher. This type of underreporting on demographic factors is comparatively lower in studies that claim partial or no representativeness. While studies with a negative outlook also include a higher portion of marginalized racial and gender categories in their study, only around a fifth of all studies (22%) include subcategories beyond the gender binary. Finally, most studies either explicitly or implicitly focus on people in the U.S. excluding other relevant (sub)populations.

Contributions. Our work contributes a systematic understanding of the state-of-the-art research on the demographic representativeness of LLMs, finding patterns of underreporting of crucial details required to establish representativeness. We provide a set of specific recommendations for future research on this topic for better documentation and evaluation. Our annotated list of papers and analysis code is publicly available.[2]

2 Background

Representativeness in NLP and Beyond. Chasalow and Levy describe representativeness, like bias, as a 'suitcase word'—a term used widely but with multiple definitions. In quantitative social sciences, representative samples allow studying large populations without surveying every member (Gobo, 2004). In Computational Linguistics, NLP, and LLMs, it refers to "the extent to which a sample includes the full range of variability in a population" (Biber, 1993). In sociolinguistics, representativeness is often linked to generalizing across languages and varieties (Grieve et al., 2025).

There are extensive studies of algorithmic bias in NLP, including sociodemographic bias (Gupta et al., 2024). Bias has many definitions, some of which often focus on misrepresentation or underrepresentation of certain (demographic) groups (Ferrara, 2023a). Recent research has also focused on 'AI Alignment', which is broadly defined as making "AI systems behave in line with human intentions and value" (Ji et al., 2023). There are several parallels between alignment and representativeness, especially for personalized LLM agents; however such parallels have not been widely explored, possibly due to the lack of a concrete vocabulary for operationalizing alignment (Kirk et al., 2023).

[2] https://github.com/Indiiigo/LLM_rep_review

Source	Longlisted	Deduplicated	Included
Agnew et al.	13	13	4
ArXiv	291	290	156
ACL	196	41	9
Semantic	86	4	1
ACM DL	117	108	5
OpenAlex	362	160	29
Other	24	12	7
Total	1076	615	211

Table 1: Summary statistics of papers considered in this review.

Repurposing Bias for Representativeness? Argyle et al. conceptualize 'algorithmic fidelity', positing that biases in LLMs conditioned on demographic attributes can mirror ideas and opinions of those demographics. They state: *"... the "algorithmic bias" within one such tool—the GPT-3 language model—is instead both fine-grained and demographically correlated, meaning that proper conditioning will cause it to accurately emulate response distributions from a wide variety of human subgroups."* However, algorithmic bias in NLP systems, including LLMs, often has competing definitions, with some emphasizing underrepresentation and misrepresentation as key factors (Ferrara, 2023b; Gallegos et al., 2024). Given its multifaceted nature, can bias enhance equitable representativeness across all groups, including marginalized ones?[3] In this review, we attempt to find the current consensus w.r.t to the demographic representativeness of LLMs and unearth potential reasons behind seemingly conflicting findings.

3 Literature Search and Annotation

In this literature review, we are interested in the intersection of LLMs and demographics; as such we only include papers that conduct a study which incorporates demographics somewhere in the pipeline, i.e., either use demographic dimensions in input to LLMs or include demographic variables in the evaluation of LLMs. Therefore, we search for papers containing the keywords "Large Language Models"/"LLM" and "demographic*" available online before December 1st, 2024. We first started with the 13 papers assessed in the scoping review by Agnew et al. on the potential of replacing human participants with LLMs in human-subject studies.[4] To expand this list, we utilize five sources — arXiv, the ACL Anthology, Semantic Scholar, ACM Digital Library, and OpenAlex. The latter is an open-source version of the Microsoft Academic Graph. Finally, we also include existing community resources, i.e., papers identified in Simmons and Hare and a paper list on public opinion simulation with LLMs.[5] After a semi-automatic deduplication step, three authors split the 615 papers between them to manually assess whether a paper should be included in the literature review.

3.1 Scope and Inclusion Criteria

We restrict our literature review to research papers (not necessarily peer-reviewed) with empirical findings. As such, we exclude other literature reviews, perspective and theoretical articles, pay-walled articles, and extended abstracts (but include short papers and workshop papers). The first content-related criterion for inclusion is that the study should touch on demographics. Only four of the 13 papers studied in Agnew et al. do so (Argyle et al., 2023; Park et al., 2022; Aher et al., 2023; Gerosa et al., 2024); the other nine discuss the potential of LLMs replacing humans, but do not state *which* humans. To balance coverage of relevant literature with the annotation workload, we only include generative LLM-based studies which are text-only, i.e., excluding vision, speech, or multimodal applications. Based on these criterion, we include 211 papers. The literature search and inclusion process is summarized in Table 1.

3.2 Codebook Categories

We have a three-part annotation scheme whose most important categories of the codebook are exemplified in Table 2, while the full codebook can be found in the Appendix (Section B).

Contexts and LLMs. Contexts refer to the scenario in which LLMs are used, either as proxies for humans or for providing services to or about humans. These categories are based on how people use LLMs (Mireshghallah et al.), restricted to those where demographics play an explicit role. In addition, in line with Tseng et al., we define and annotate two types of representativeness or *personas* in LLMs: their ability to *impersonate* group X and their ability to *personalize* to group X. We also

[3] Studies affirming LLMs' algorithmic fidelity, e.g., Argyle et al.; Kim and Lee, do not explicitly define bias.

[4] Out of the 16 artifacts studied in Agnew et al., there are 13 research papers, while the other 3 are product offerings.

[5] `https://github.com/CaroHaensch/public_opinion_llms`

Category	Subcategory	Definition and Examples
Contexts	Advice	Providing help with decision-making, or giving suggestions, recommendations, or advice, e.g., (Levy et al., 2024; Liu et al., 2024c; Lahoti et al., 2023)
	Simulation	Synthetic data generation to study human behavior directly, e.g., simulating survey respondents (Bisbee et al., 2024) or platform simulations (Park et al., 2022).
	Content Analysis	Qualitative content labeling, evaluation, and labeling, e.g., sentiment analysis, hate speech (Beck et al., 2024; Giorgi et al., 2024b; Sun et al., 2023)
	Writing	Fiction or non-fiction writing, could also include translation or rewriting content e.g., (Wan and Chang, 2024; Sourati et al., 2024)
	Generic	General investigation of LLMs, without any downstream context e.g., (Zhao et al., 2023; Jiang et al., 2022).
Persona Type	Impersonation	Persona induced in the LLM, e.g., "answer this question as a *Democrat*" using personas from survey data in Argyle et al. (2023); von der Heyde et al. (2024)
	Personalization	Persona that the LLM needs to act upon, e.g., text written by a group "would you hire this *man* based on his resume?" (Gaebler et al., 2024) or about a group, e.g., targets of hate speech "annotate: [content targeting *women*]" (Beck et al., 2024)
Conclusion on Representative -ness	Yes	The study is positive, e.g. Argyle et al., who say *"We suggest that language models with sufficient algorithmic fidelity thus constitute a novel and powerful tool to advance understanding of humans and society across a variety of disciplines."*
	Partial	The study has mixed results, finding LLMs to be successful at representating some groups but not others, e.g. Gabriel et al., who say *"We find that while GPT-4 can reflect and amplify harmful biases found in peer-to-peer support, these biases vary significantly based on prompt design and can be mitigated through..."*
	No	The study has a negative outlook on representativeness, e.g., von der Heyde et al. noting *"We have shown that in its current state, GPT-3.5 is not suitable for estimating public opinion across (sub)populations..."*

Table 2: **Key categories, definitions, and examples in our annotation codebook.** An expanded version of the codebook with all categories and the instructions given to annotators are included in the Appendix (Section B).

note the LLMs used and the approach to induce or improve their representativeness, e.g., prompting.

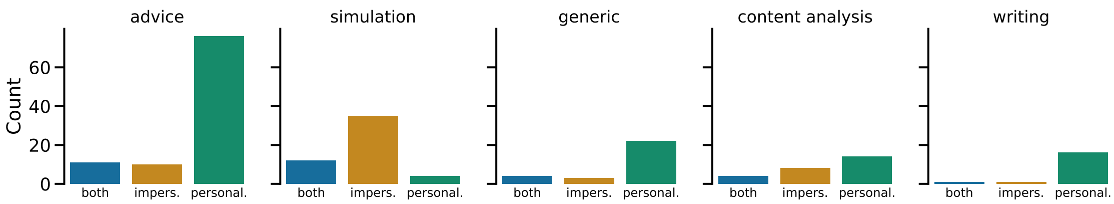
Evaluating and Improving Representativeness. We annotate the **response format** employed in each study, i.e., free-text responses from LLMs vs. closed-form responses, like multiple choice question-answering (QA). Except Lee et al. who study LLMs' ability to adapt to African American English, all papers in our sample focus on broad and/or multiple demographic categories. To establish an LLM's equitable representativeness for a population, it is important to understand how well´it represents different subgroups of that population. Therefore, we label whether studies conduct a **demographically disaggregated evaluation**, i.e., if they evaluate a model *across* multiple groups, *within* subgroups of demographic category, or both.

Demographics and Representativeness. We identify the demographic categories studied in a paper and how they are operationalized, i.e. the **subcategories** and **descriptors** used to represent these subgroups (examples in Table 3).[6] Subcate-

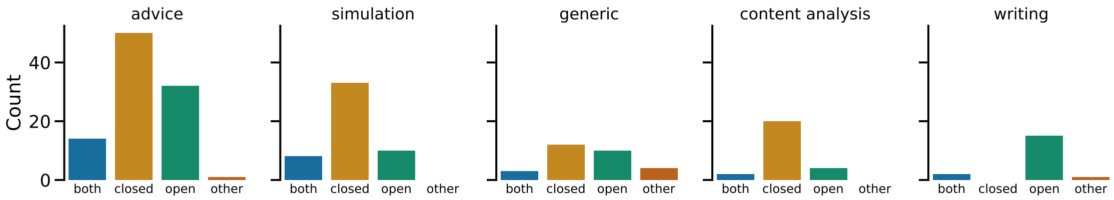
Paper	Dem. Category	Subcategories and Descriptors
Zheng	Gender	Male, Female, LGBTQ
	Race	American Indian or Alaskan Native, Asian, Black or African American, Filipino, Hispanic or Latino, Native Hawaiian or Pacific Islander, White
	Income	Disadvantaged, Non-disadvantaged
Alipour et al.	Gender	Man, Woman
	Race	Asian, Black, White, Hispanic
Steinmacher et al.	Gender	Not reported
	Location	
	Age	

Table 3: Examples of how demographic categories and subcategories are operationalized and reported in papers.

[6]The full list of demographic categories we include in this paper can be found in Table 5 in the Appendix.



(a) Context vs. Persona Type



(b) Context vs. Response Format

Figure 2: **Distribution of contexts over a) persona type and b) response format.** Most papers look into *personalization* of LLMs across different contexts, except *simulation* studies. Closed-form responses from LLMs are more widely studied except in *writing*.

gories refer to subgroups of a given demographic category, while descriptors refer to how these subgroups are described, e.g, Argyle et al. use the binary gender descriptors 'male' and 'female' vs. Cheng et al. who use 'man', 'woman', and 'nonbinary', while Deldjoo does not specify the gender subcategories or descriptors used. We also annotated the **target population** studied in a paper, e.g., the global population in Durmus et al.. While some papers do not explicitly specify a target population, for some of them we can infer whether the population is the U.S. based on racial and political leaning descriptors, i.e., using U.S. specific-categories like 'Native American' or 'Republican', e.g., Arzaghi et al..

Finally, we annotate the paper's **conclusion regarding the representativeness of LLMs** for the target population studied. Many of the papers included in our survey focus on mitigating demographic biases in LLMs through debiasing and alignment, e.g. Do et al. (2025a); He et al. (2025). We annotate these articles as concluding positively if they find their debiasing technique to be effective, as biases are considered a threat to representativeness (Ferrara, 2023a) and reducing them would lead to improved representativeness.

3.3 Annotation Process

Three annotators, who are also authors of this paper, independently annotated all 211 papers using the aforementioned codebook over three rounds. After each round, checks were done by randomly selecting three papers from each annotator's batch to be annotated by all three annotators to ensure reliability. We found little disagreement across three rounds (3-8% of diverging annotations across the rounds). Disagreements were resolved after discussion (c.f. Appendix Section A.2 for more details.)

4 Results

4.1 Descriptive Findings

A majority of studies fall under *advice* (43%), followed by *simulation* (23%), *generic* (13%) and *content analysis* (11%). *Advice* scenarios span many different topics including medical (Rawat et al., 2024), hiring (Gaebler et al., 2024), and education (Weissburg et al., 2024). Simulations often focus on replicating surveys (Gerosa et al., 2024) or social media behavior (Chang et al., 2024). *Content analysis* studies often focus on annotating subjective tasks with LLMs (Jiang et al., 2024a).[7] Figure 2a shows that across most contexts, *personalization* is more common, except *simulation* where *impersonation* or both are studied. Figure 2b shows the distribution of response formats across contexts; besides *content analysis* where mainly closed evaluation is conducted and the opposite in *writing*, we see both open and closed in other contexts, with closed format being more prevalent.

[7] Some have multiple contexts (c.f. Appendix A.2), while Mori et al. use LLMs for the sole purpose of generating training data. As this was the only paper for this application, it did not justify creating a new context. We therefore exclude it from the analysis on contexts.

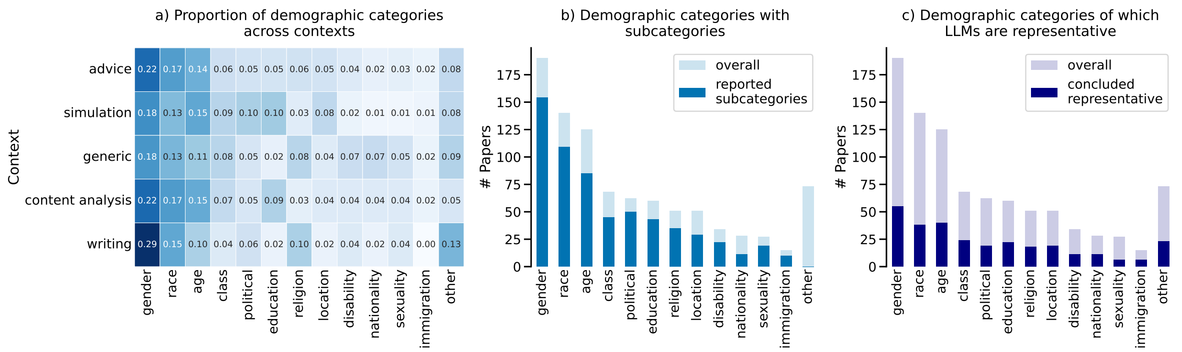


Figure 3: **Demographic Dimensions a) proportionally across contexts, b) with subcategories, and c) concluded to be represented by LLMs.** *Gender* and *Race* are widely studied, but *simulation* tends to have comparatively more balanced distribution of demographics.

LLMs and Methods for Steering them. While, most studies (62%) include more than one model, a majority of studies only use commercial LLMs; 80% of the papers include at least one OpenAI model. This is followed by open-weight models like LLaMa (39%), and Mistral or Mixtral (21%).[8] Finally, in terms of measuring and steering representativeness, prompt-based techniques are by far the most commonly used, appearing in 64% of the papers, followed by fine-tuning approaches (13%), e.g., Jiang et al. (2022); Wald and Pfahler (2023).

4.2 RQ1: Which demographics are most studied and in which contexts?

We first investigate the target populations mentioned in papers (Table 4). A large number of papers, i.e. 36%, do not specify a target population, instead aiming to study demographic characteristics in general with generic categories, e.g., white vs. non-white in Kamruzzaman et al. or do not mention either an explicit target population or demographic subcategories (Do et al., 2025a). 26% of the papers solely and explicitly focus on the U.S., while another 16% do so implicitly via the use of U.S.-specific racial or political subcategories.18% of studies explicitly mention target populations beyond the U.S., while a small proportion (4%) attempt to study the global population.

Figure 3a shows the proportion of demographic dimensions studied over different contexts, while Figure 3b and c show the count of different demographic dimensions. We confirm previous findings in NLP research showing gender and racial categories to be the most studied (Gupta et al., 2024), though the distribution is less skewed compared to research on biases (Gupta et al., 2024). We also note that for *simulation* studies, the distribution is more balanced compared to other contexts. Widely studied categories in 'other' include marital status, number of children, and occupation.

In terms of how specific demographic categories are operationalized, we find that many papers (38% on average across all demographics) do not explicitly report the demographic subcategories and descriptors they use in their study, i.e., they state that they study a particular demographic dimension but do not report the full list of subcategories and/or descriptors (Figure 3b).[9] We note that studies focusing on *nationality* and *class* tend to underreport the subcategories used more compared to other categories. The findings from our analysis of the target

Target Pop.	**#**	**Examples**	**% Repres -entative?**
Other	39	German Political Parties (Batzner et al.), Indians (Sahoo et al.)	23%
U.S. Explicit	54	Argyle et al., Santurkar et al.	22%
U.S. Implicit	34	Cheng et al., Giorgi et al.	14%
Global	9	Durmus et al., Jin et al.	11%
Undefined	75	Park et al., Lahoti et al.	45%

Table 4: **Target Populations and their proportion found to be represented by LLMs.** Studies on Global populations report the lowest rate (11%).

[8]To keep the annotation and analysis from blowing up, we do not report parameter size or versions of LLMs used. More descriptive results can be found in Appendix A.4.

[9]Note that reporting descriptors does not apply to papers that do not prompt LLMs with demographic personae, but reporting subcategories does.

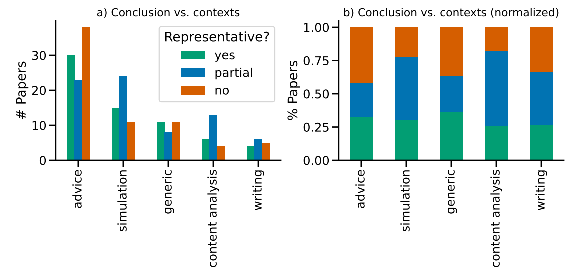


Figure 4: **Contexts vs. Representativeness.** While *advice* and *generic* have polarized responses, the rest report mainly partial representativeness.

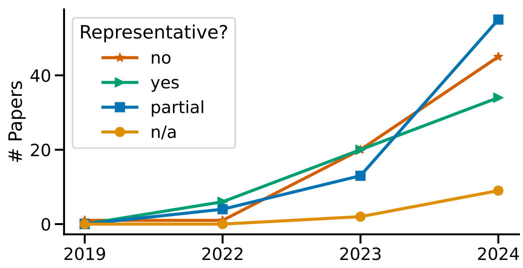


Figure 5: **Reported representativeness of LLMs over time.** We can observe (i) an increase of conflicting reports and (ii) an especially high increase in papers finding partial representativeness.

population and demographics studied reveal two patterns across most LLM application contexts — **i) an outsized focus on the U.S. population**, in line with previous research (Field et al., 2021) and persistent in studies with LLMs, and ii) **a tendency to under-specify the explicit target population and demographic subgroups being studied.**

4.3 RQ2: Is there consensus on the representativeness of LLMs?

Out of 211 papers, 29% conclude 'yes' on representativeness of LLMs, 34% conclude 'partial', and 32% conclude 'no'.[10] Figure 4a and Figure 4b shows the distribution and proportion of studies claiming representativeness of LLMs across different contexts, respectively. *Advice* and *generic* studies seem to have strongly diverging conclusions on representativeness while *simulations*, *writing*, and *content analysis* have a majority of partially representative findings. For the latter context's majority partial results, previous findings on the limits of demographic dimensions in subjective data annotation (Orlikowski et al., 2023), appears to extend to LLMs' annotations (Beck et al., 2024; Alipour et al., 2024). With respect to steering, 24% of the papers using prompting to induce personas conclude positively, while the number is much higher for fine-tuning (63%).

In terms of demographic factors, studies that implicitly target the US population or a global population have the lowest percentage of conclusions on the positive representativeness of LLM (Table 4, last column). However notably, studies which do not specify a target population had the most positive conclusions. Many of them are on debiasing (Lahoti et al., 2023) or alignment (Chen et al., 2024a). However, it is unclear to which exact populations these findings apply. In Figure 3c, we study the number of papers found to be representative across different demographic categories. This distribution generally mirrors the rank of demographic dimensions studied, however *political leaning*, *disability*, and *sexuality* have comparatively fewer studies claiming that LLMs are representative.

Finally, in Figure 5, we investigate trends related to LLM representativeness over time and note two points — a relatively slower growth of articles with a positive outlook and an increase in papers claiming partial representativeness, suggesting a move to more nuanced evaluations.

4.4 Disentangling Disagreements on Representativeness

To unpack disagreements regarding outlooks on representativeness of LLMs, we assess the relationship between these outlooks across a) evaluation approaches and b) demographic categories studied.

Demographically Disaggregated Results. Overall, we find that 20% of the papers do not conduct any type of demographically disaggregated analysis, i.e., they report the LLMs' overall performance rather than performance across multiple demographic groups or within the subcategories of a group. This proportion is higher for papers claiming LLMs to be representative, i.e., 30% compared to 19% of the papers claiming LLMs to be partially representative and 10% claiming no representativeness (Figure 6a).[11]

Demographic Categories. Building on findings in past research on the tendency of LLMs to stereotype marginalized groups (Cheng et al., 2023b; Nguyen et al., 2024), we assess whether

[10] 11 papers or 5% do not provide a conclusion, hence we exclude them from the analysis in Sections 4.3 and 4.4.

[11] Papers which only focus on a single demographic are included under 'across + within' in Figure 6a if they report results within the subcategories of that demographic dimension.

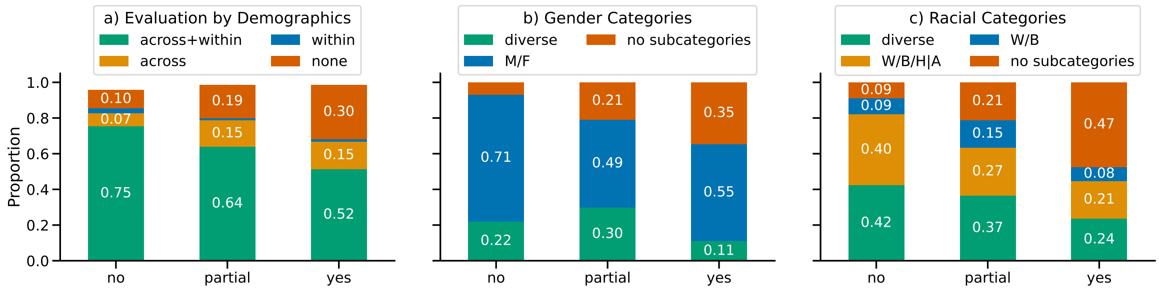


Figure 6: **Factors differentiating studies claiming LLMs are representative vs. those claiming otherwise.** We plot the proportion of papers conducting a) demographically disaggregated analysis, and the proportion of papers studying different types of b) gender and c) racial categories. We find that papers with a positive outlook show a higher tendency of not conducting any demographically disaggregated analysis. They also more often underreport demographic subcategories or do not include marginalized/diverse categories.

including marginalized or diverse categories is associated with findings on representativeness.

We investigate the two most studied demographic categories, *gender* and *race*, and investigate which, if any, descriptors have been used to operationalize these two demographic dimensions. For both *gender* and *race*, we devise a category for no reported subcategories and, as a consequence, no descriptors. For *gender*, we have one category encoding binary gender ('male/female') and another category including diverse categories ('diverse') if a study includes any gender subcategory besides 'male' or 'female'. Similarly for *race*, we define a 'Black/White' category, 'White/Black/Hispanic or Asian' if either Asian or Hispanic is included with White and Black,[12] and finally a 'diverse' racial category if the target population is explicitly beyond the U.S. (i.e., Global or Other) or if they include the identities 'Native American' or 'Pacific Islander'.

Figure 6b and 6c show the proportion of the aforementioned categories across studies claiming representativeness of LLMs. **For both *gender* and *race*, many studies claiming LLMs to be representative either do not report the demographic subcategories or have a lower proportion of diverse categories compared to studies claiming otherwise.** For example, from studies that claim that LLMs are representative of racial demographics, 47% do not report racial subcategories, compared to 9% in studies claiming no representativeness. Our findings across several papers strengthens the hypothesis that LLMs might be particularly unrepresentative of marginalized groups, e.g., Argyle et al.; Kim and Lee use binary gender and find LLMs to be capable of simulating the U.S. population, while Cheng et al.; Wang et al. investigate nonbinary personas as well, coming to the opposite conclusion. Similarly, for *advice*, only three of 30 papers concluding positively about the representativeness of LLMs, use diverse *gender* subcategories; they aim to debias (Lahoti et al., 2023; Tamkin et al., 2023) or align existing non-representative LLMs (Li et al., 2024b). However, we also note that while papers claiming LLMs to be representative tend to exclude marginalized groups more, the inclusion of these groups is generally low overall — 22% and 35% of all papers include diverse gender and racial subgroups, respectively. This indicates a greater need for studies on the representativeness of LLMs for marginalized groups.

Other Factors. Conducting a qualitative analysis of other factors driving disagreement, we find that many papers concluding positively rely more on closed-form response formats or often do not take into account the variance of LLM responses. More details are provided in Appendix A.6.

5 Discussion

LLMs, beyond their role as assistive chatbots, are increasingly used to supplement or replace humans in research (Gilardi et al., 2023). In all these applications, it is crucial to assess whether LLMs provide equitable assistance and adequately represent the populations they aim to supplement. Some studies (Argyle et al. (2023); Kim and Lee (2023) *inter alia*) argue that LLM biases enhance subgroup representation, while others highlight contradic-

[12]We do not find any studies that use 'Asian' or 'Hispanic' without also including 'White' and 'Black'.

tions between bias and representativeness (Ferrara, 2023b; Wang et al., 2024a). Empirical studies remain divided on LLM representativeness (Argyle et al., 2023; Bisbee et al., 2024). Our systematic review finds that studies incorporating demographics predominantly focus on the U.S. and often lack crucial details to assess representativeness.

5.1 Implications for using LLMs in Social Applications

LLMs' flexibility grants researchers broad methodological choices—e.g., persona induction (prompting vs. fine-tuning), response types, and model selection. These issues contribute to growing concern regarding reproducibility, even without factoring in the reproducibility issues associated with using closed-source models (Barrie et al., 2024). Our review shows that these degrees of freedom do not just affect the assessment of reproducibility, but also of representativeness. For instance, Argyle et al. and Bisbee et al. reach opposing results despite similar methodologies. Furthermore, marginalized groups remain underrepresented and underserved by LLMs (Wang et al., 2024a; Cheng et al., 2023b). Therefore, studies assessing LLMs for social applications should report their exact design choices, **while evaluating the representativeness of LLMs across multiple subgroups within the target population, rather than relying on an overall assessment.**

5.2 Recommendations for Reporting and Improving Representativeness

While many papers anticipate future LLM improvements in representativeness, the exact changes needed remain unclear. To gauge these improvements, context-specific evaluations are essential. Current LLM benchmarks like HELM (Liang et al., 2022) or BigBench (Srivastava et al., 2022) include some bias-related evaluations, however, in line with our findings, bias and representativeness are not necessarily equivalent. We advocate for tailored benchmarks explicitly defining target populations and demographic subcategories, along with demographically disaggregated analyses combining open and closed-form evaluations. To enhance transparency, we propose incorporating explicit population and demographic categories into reproducibility checklists and model/data documentation.

Additionally, future research should explore under-examined representativeness interventions, including model editing and RLHF. Algorithmically biased LLMs might only represent a particular group of people — non-marginalized Americans in line with previous findings (Durmus et al., 2023; Atari et al.), in narrow evaluation settings. To represent diverse populations, we need to move beyond repurposing algorithmic bias and think of intentionally designed representative LLMs, e.g., through approaches like more detailed personas (Moon et al., 2024) or pluralistic alignment (Sorensen et al., 2024). In data-driven approaches, robust sampling strategies from quantitative social sciences can also inform better demographic representation, especially for marginalized populations (Freimuth and Mettger, 1990). However, technical solutions might not overcome epistemological issues in the applications of LLMs to certain social applications, particularly subgroup simulations (Agnew et al., 2024; Kapania et al., 2024).

6 Conclusion

From our review spanning 211 papers with a focus on how demographic factors are operationalized in assessing LLMs' representativeness, we find that a significant number of papers underreport the target population being studied. Among those that do report it, most focus on the U.S. Additionally, demographic subcategories and descriptors are often omitted, while only a minority of studies include marginalized gender and racial groups when assessing the representativeness of LLMs. In terms of outlook on representativeness, roughly 29% of papers find positive results while a third do not, suggesting a degree of contention in the field. Articles with positive conclusions are more likely to underreport demographic subcategories and when they do include them, marginalized groups are often excluded comparatively more than papers that conclude negatively. Our findings suggest an inflated perception of LLM representativeness, particularly for marginalized groups and populations beyond the U.S. To improve the measurement of representativeness of LLMs, we need specific benchmarks explicitly assessing populations beyond the U.S. and encouraging demographic-based evaluation and documentation across and within subcategories.

7 Limitations

Our paper, like other systematic literature reviews, has a specific scope and to balance annotation effort and coverage, we had to exclude papers on multimodal uses of LLMs, e.g., Lin et al. (2024); Wu et al. (2024). A key reason for this is that applications of multimodal LLMs are broader than text-only LLMs, which would have also required thinking of new contexts such as graphic design.

Focus on Demographics. Another limitation of our work is that while social identity is complex (Stets and Burke, 2000; Cameron, 2004) and comprised of many different facets such as personality, interests, and affiliations, we focused solely on sociodemographics. However, demographic factors are of great interest to social scientists (Garza and Herringer, 1987; Smith, 2007) and often dictate real-world misrepresentation, e.g., sexism or racial discrimination. Furthermore, much of the literature on the intersection of human factors and LLMs focuses on demographic categories of (sub)populations, with a few exceptions studying individuals (e.g., Jiang et al.; Park et al.), personality traits and attitudes (Jiang et al., 2024a). Even papers focusing on attitudes often combine and correlate these with demographics (Jiang et al., 2024a) Within demographic dimensions, we do not focus on cultural identity, since it incorporates facets other than demographics such as cuisine or language. We point the interested reader to surveys on culture and LLMs (Adilazuarda et al., 2024; Liu et al., 2024b; Pawar et al., 2024). In principle, our categorization scheme is adaptable to other aspects of identity such as personality or interests.

Annotation Categories and Granularity. Our assessment of conclusion of representativeness of papers is based on their overall takeaway, i.e., we do not report the conclusion for specific demographics. This is too some extent impossible because many papers do not conduct any demographically disaggregated analysis (Figure 6a) and for those that do, the analysis of this conclusion across different categories would have made our, already complex, codebook even more complex. Our goal in this paper was summarizing the practices w.r.t. demographic representativeness of LLMs, and in the future, we hope to conduct a deeper meta-analysis of the reports on individual demographic factors in the future. Similarly, to keep the annotation and analysis from blowing up, we do not report parameter size or versions of LLMs used. Tracking model versions and parameters in a meaningful way is challenging due to several factors: (1) 62% of the papers test multiple models, complicating attribution, (2) many papers do not report precise model versions and parameter sizes e.g., 11% of papers say just "ChatGPT", and (3) the vast heterogeneity in metrics and evaluation methods makes direct comparisons difficult, e.g., e 1-Wasserstein distance in Santurkar et al., vs. tetrachoric correlation in Argyle et al.. A rigorous meta-analysis would be required to isolate the impact of model versions, but such an analysis extends beyond the scope of a systematic literature review. Crucially, our paper lays the groundwork for such a meta-analysis by mapping out and categorizing the existing literature—a necessary step before deeper quantitative synthesis.

Limited Timeframe. Finally, as the research on generative LLMs and social identity is still evolving, our temporal analysis is limited to mainly three years of research. The temporal granularity could be affected by discrepancies in reporting of year since some papers are still preprints while others have been published in peer-reviewed venues but would still be recorded under their preprint date.

Acknowledgments

We are grateful to Claire Jordan for her help with an initial version of the annotation codebook used in this paper. We thank Mattia Samory, Georg Ahnert, and Maximilian Kreutner for helpful comments and discussions.

8 Ethical Considerations

Our systematic literature review aims to shed light on the demographic representativeness of LLMs in social applications such as providing recommendations or simulating human behavior. Motivated by discordant findings on this topic, our review reveals an inflated sense of representativeness in papers that claim positively about LLMs' capability of mirroring human subpopulations. Many of these papers focus on people in the U.S. or do not include concrete evaluations required to establish representativeness. To that end, our work sets the stage for creating concrete reporting and evaluation protocols to better assess the representativeness of LLMs. Our findings apply to papers that study specific demographics, but even more to those papers that claim LLMs can replace or supplement humans but do not mention *which* people. For studies on personalization and simulation of people, we suggest explicitly reporting which target populations their findings apply to in reproducibility checklists for publications and data/model documentation sheets.

Finally, as a community, we need to incentivize, or at least not penalize, studying populations beyond the U.S., in the context of LLMs.

Our study of representativeness is limited to demographics, and even within that in operationalizing marginalized groups, we only focus on racial and gender categories. The main reason for this is because these are the two most widely studied categories. However, our annotations include how other categories were operationalized and one avenue of future research would be focusing on marginalized groups on other widely studied categories including age or political leaning, e.g., the elderly and political fringe groups. Last but not least, it is also vital to consider the arbitrariness of some of these demographic categories and subcategories, e.g., the variance in Table 5 in the Appendix. We should account for the process behind the construction of these categories and the impact of their definition on downstream applications (Bowker, 2000).

References

Ahmad M Abdelhady and Christopher R Davis. 2023. Plastic surgery and artificial intelligence: how chatgpt improved operation note accuracy, time, and education. *Mayo Clinic Proceedings: Digital Health*, 1(3):299–308.

ACL Anthology. 2025. Acl anthology search. Accessed: 2025-02-13.

ACM Digital Library. 2025. Acm digital library. Accessed: 2025-02-13.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling" culture" in llms: A survey. *arXiv preprint arXiv:2403.15412*.

William Agnew, A Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R McKee. 2024. The illusion of artificial inclusion. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Carlos Aguirre, Kuleen Sasse, Isabel Cachola, and Mark Dredze. 2024. Selecting shots for demographic fairness in few-shot learning with large language models. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 50–67.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.

Shayan Alipour, Indira Sen, Mattia Samory, and Tanushree Mitra. 2024. Robustness and confounders in the demographic alignment of llms with human perceptions of offensiveness. *arXiv preprint arXiv:2411.08977*.

Khaled AlNuaimi, Gautier Marti, Mathieu Ravaut, Abdulla AlKetbi, Andreas Henschel, and Raed Jaradat. 2024. Enriching datasets with demographics through large language models: What's in a name? *arXiv preprint arXiv:2409.11491*.

AJ Alvero, Jinsook Lee, Alejandra Regla-Vargas, René F Kizilcec, Thorsten Joachims, and Anthony Lising Antonio. 2024. Large language models, social demography, and hegemony: comparing authorship in human and synthetic text. *Journal of Big Data*, 11(1):138.

A Amirova, T Fteropoulli, N Ahmed, MR Cowie, and JZ Leibo. 2024. Framework-based qualitative analysis of free responses of large language models: Algorithmic fidelity. *PLoS ONE*, 19(3):e0300024.

Toluwani Aremu, Oluwakemi Akinwehinmi, Chukwuemeka Nwagu, Syed Ishtiaque Ahmed, Rita Orji, Pedro Arnau Del Amo, and Abdulmotaleb El Saddik. 2025. On the reliability of large language models to misinformed and demographically informed prompts. *AI Magazine*, 46(1):e12208.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

arXiv. 2025. arxiv advanced search. Accessed: 2025-02-13.

Mina Arzaghi, Florian Carichon, and Golnoosh Farnadi. 2024. Understanding intrinsic socioeconomic biases in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 49–60.

Mercy Asiedu, Nenad Tomasev, Chintan Ghate, Tiya Tiyasirichokchai, Awa Dieng, Oluwatosin Akande, Geoffrey Siwo, Steve Adudans, Sylvanus Aitkins, Odianosen Ehiakhamen, et al. 2024. Contextual evaluation of large language models for classifying tropical and infectious diseases. *arXiv preprint arXiv:2409.09201*.

Mohammad Atari, Mona J Xue, Peter S Park, Damián Blasi, and Joseph Henrich. Which humans?

Yuqi Bai, Kun Sun, and Huishi Yin. 2024. Agentic society: Merging skeleton from real world and texture from large language model. *arXiv preprint arXiv:2409.10550*.

Pragyan Banerjee, Abhinav Java, Surgan Jandial, Simra Shahid, Shaz Furniturewala, Balaji Krishnamurthy, and Sumit Bhatia. 2023. All should be equal in the eyes of language models: Counterfactually aware fair text generation. *arXiv preprint arXiv:2311.05451*.

Ehsan Barkhordar and Sukru Atsizelti. 2024. Assessing the predictive power of social media data-fed large language models on vote preference. In *Companion Publication of the 16th ACM Web Science Conference*, Websci Companion '24, page 53–55, New York, NY, USA. Association for Computing Machinery.

Christopher Barrie, Alexis Palmer, and Arthur Spirling. 2024. Replication for language models problems, principles, and best practice for political science. *URL: https://arthurspirling.org/documents/BarriePalmerSpirling TrustMeBro.pdf*.

Harshavardhan Battula, Jiacheng Liu, and Jaideep Srivastava. 2024. Enhancing in-hospital mortality prediction using multi-representational learning with llm-generated expert summaries. *Preprint*, arXiv:2411.16818.

Jan Batzner, Volker Stocker, Stefan Schmid, and Gjergji Kasneci. 2024. Germanpartiesqa: Benchmarking commercial large language models for political bias and sycophancy. *arXiv preprint arXiv:2407.18008*.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615.

Cosmin A Bejan, Amy M Reed, Matthew Mikula, Siwei Zhang, Yaomin Xu, Daniel Fabbri, Peter J Embí, and Ryan S Hsi. 2024. Large language models improve the identification of emergency department visits for symptomatic kidney stones. *medRxiv: the preprint server for health sciences*.

Noam Benkler, Drisana Mosaphir, Scott Friedman, Andrew Smart, and Sonja Schmer-Galunder. 2023. Assessing llms for moral value pluralism. *arXiv preprint arXiv:2312.10075*.

Harel Berger, Hadar King, and Omer David. 2024. Dreaming with ChatGPT: Unraveling the challenges of LLMs dream generation. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 140–147, Miami, FL, USA. Association for Computational Linguistics.

Lorenzo Berlincioni, Luca Cultrera, Federico Becattini, Marco Bertini, and Alberto Del Bimbo. 2024. Prompt and prejudice. *arXiv preprint arXiv:2408.04671*.

Douglas Biber. 1993. Representativeness in corpus design. *Literary and linguistic computing*, 8(4):243–257.

Anindya Bijoy Das and Shahnewaz Karim Sakib. 2024. Unveiling and mitigating bias in large language model recommendations: A path to fairness. *arXiv e-prints*, pages arXiv–2409.

James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. 2024. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416.

Geoffrey C Bowker. 2000. *Sorting things out: Classification and its consequences*. MIT press.

James E Cameron. 2004. A three-factor model of social identity. *Self and identity*, 3(3):239–262.

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021, Bangkok, Thailand. Association for Computational Linguistics.

Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. 2025. Persona: A reproducible testbed for pluralistic alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11348–11368.

Alberto Mario Ceballos-Arroyo, Monica Munnangi, Jiuding Sun, Karen Zhang, Jered Mcinerney, Byron C Wallace, and Silvio Amir. 2024. Open (clinical) llms are sensitive to instruction phrasings. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 50–71.

Roberto Cerina and Raymond Duch. 2023. Artificially intelligent opinion polling. *arXiv preprint arXiv:2309.06029*.

Serina Chang, Alicja Chaszczewicz, Emma Wang, Maya Josifovska, Emma Pierson, and Jure Leskovec. 2024. Llms generate structurally realistic social networks but overestimate political homophily. *arXiv preprint arXiv:2408.16629*.

Kyla Chasalow and Karen Levy. 2021. Representativeness in statistics, politics, and machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 77–89, New York, NY, USA. Association for Computing Machinery.

Isha Chaudhary, Qian Hu, Manoj Kumar, Morteza Ziyadi, Rahul Gupta, and Gagandeep Singh. 2024. Quantitative certification of bias in large language models. *arXiv preprint arXiv:2405.18780*.

Khaoula Chehbouni, Jonathan Colaço Carr, Yash More, Jackie CK Cheung, and Golnoosh Farnadi. 2024a. Beyond the safety bundle: Auditing the helpful and harmless dataset. *arXiv preprint arXiv:2411.08243*.

Khaoula Chehbouni, Megha Roshan, Emmanuel Ma, Futian Andrew Wei, Afaf Taïk, Jackie Chi Kit Cheung, and Golnoosh Farnadi. 2024b. From representational harms to quality-of-service harms: A case study on llama 2 safety safeguards. In *ACL (Findings)*.

Quan Ze Chen, K. J. Kevin Feng, Chan Young Park, and Amy X. Zhang. 2024a. Spica: Retrieving scenarios for pluralistic in-context alignment. *Preprint*, arXiv:2411.10912.

Shan Chen, Jack Gallifant, Mingye Gao, Pedro Moreira, Nikolaj Munch, Ajay Muthukkumar, Arvind Rajan, Jaya Kolluri, Amelia Fiske, Janna Hastings, et al. 2024b. Cross-care: assessing the healthcare implications of pre-training data on language model bias. *Advances in neural information processing systems*, 37:23756–23795.

Xi Chen, Yang Xu, MingKe You, Li Wang, WeiZhi Liu, and Jian Li. 2024c. Evaluation of bias towards medical professionals in large language models. *arXiv preprint arXiv:2407.12031*.

Yiting Chen, Tracy Xiao Liu, You Shan, and Songfa Zhong. 2023. The emergence of economic rationality of gpt. *Proceedings of the National Academy of Sciences*, 120(51):e2316205120.

Yuen Chen, Vethavikashini Chithrra Raghuram, Justus Mattern, Rada Mihalcea, and Zhijing Jin. 2022. Causally testing gender bias in llms: A case study on occupational bias. *arXiv preprint arXiv:2212.10678*.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023a. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023b. Compost: Characterizing and evaluating caricature in llm simulations. *arXiv preprint arXiv:2310.11501*.

Zhibo Chu, Zichong Wang, and Wenbin Zhang. 2024. Fairness in large language models: A taxonomic survey. *Preprint*, arXiv:2404.01349.

Yun-Shiuan Chuang, Krirk Nirunwiroj, Zach Studdiford, Agam Goyal, Vincent Frigo, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. 2024. Beyond demographics: Aligning role-playing llm-based agents using human belief networks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14010–14026.

Amanda Cercas Curry, Giuseppe Attanasio, Zeerak Talat, and Dirk Hovy. 2024. Classist tools: Social class correlates with performance in nlp. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12643–12655.

Yashar Deldjoo. 2024. Understanding biases in chatgpt-based recommender systems: Provider fairness, temporal stability, and recency. *ACM Transactions on Recommender Systems*.

Yongxin Deng, Xihe Qiu, Xiaoyu Tan, Jing Pan, Chen Jue, Zhijun Fang, Yinghui Xu, Wei Chu, and Yuan Qi. 2024. Promoting equality in large language models: Identifying and mitigating the implicit bias based on bayesian theory. *Preprint*, arXiv:2408.10608.

Xuan Long Do, Kenji Kawaguchi, Min-Yen Kan, and Nancy Chen. 2025a. Aligning large language models with human opinions through persona selection and value–belief–norm reasoning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2526–2547, Abu Dhabi, UAE. Association for Computational Linguistics.

Xuan Long Do, Kenji Kawaguchi, Min-Yen Kan, and Nancy Chen. 2025b. Aligning large language models with human opinions through persona selection and value–belief–norm reasoning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2526–2547.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Questioning the survey responses of large language models. *Preprint*, arXiv:2306.07951.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Jane Dwivedi-Yu. 2024. Fairpair: A robust evaluation of biases in language models through paired perturbations. *LREC-COLING 2024*, page 28.

Tyna Eloundou, Alex Beutel, David G Robinson, Keren Gu-Lemberg, Anna-Luisa Brakman, Pamela Mishkin, Meghan Shah, Johannes Heidecke, Lilian Weng, and Adam Tauman Kalai. 2024. First-person fairness in chatbots. *arXiv preprint arXiv:2410.19803*.

David Esiobu, Xiaoqing Tan, Saghar Hosseini, Megan Ung, Yuchen Zhang, Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams, and Eric Smith. 2023. Robbie: Robust bias evaluation of large generative language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3764–3814.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024. Modular pluralism: Pluralistic alignment via multi-llm collaboration. *Preprint*, arXiv:2406.15951.

Emilio Ferrara. 2023a. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3.

Emilio Ferrara. 2023b. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in nlp. *arXiv preprint arXiv:2106.11410*.

Vicki S Freimuth and Wendy Mettger. 1990. Is there a hard-to-reach audience? *Public health reports*, 105(3):232.

Saadia Gabriel, Liang Lyu, James Siderius, Marzyeh Ghassemi, Jacob Andreas, and Asuman Ozdaglar. 2024a. Misinfoeval: Generative ai in the era of “alternative facts”. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8566–8578.

Saadia Gabriel, Isha Puri, Xuhai Xu, Matteo Malgaroli, and Marzyeh Ghassemi. 2024b. Can ai relate: Testing large language model response for mental health support. *arXiv preprint arXiv:2405.12021*.

Johann D Gaebler, Sharad Goel, Aziz Huq, and Prasanna Tambe. 2024. Auditing the use of language models to guide hiring decisions. *arXiv preprint arXiv:2404.03086*.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Preprint*, arXiv:2309.00770.

Raymond T Garza and Lawrence G Herringer. 1987. Social identity: A multidimensional approach. *The Journal of social psychology*, 127(3):299–308.

Marco Gerosa, Bianca Trinkenreich, Igor Steinmacher, and Anita Sarma. 2024. Can ai serve as a substitute for human subjects in software engineering research? *Automated Software Engineering*, 31(1):13.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Salvatore Giorgi, Tingting Liu, Ankit Aich, Kelsey Isman, Garrick Sherman, Zachary Fried, João Sedoc, Lyle Ungar, and Brenda Curtis. 2024a. Modeling human subjectivity in llms using explicit and implicit human factors in personas. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7174–7188.

Tommaso Giorgi, Lorenzo Cima, Tiziano Fagni, Marco Avvenuti, and Stefano Cresci. 2024b. Human and llm biases in hate speech annotations: A sociodemographic analysis of annotators and targets. *arXiv preprint arXiv:2410.07991*.

Michael Gira, Ruisu Zhang, and Kangwook Lee. 2022. Debiasing pre-trained language models via efficient fine-tuning. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 59–69, Dublin, Ireland. Association for Computational Linguistics.

Giampietro Gobo. 2004. Sampling, representativeness and generalizability. *Qualitative research practice*, 405:426.

Purva Prasad Gosavi, Vaishnavi Murlidhar Kulkarni, and Alan F Smeaton. 2024. Capturing bias diversity in llms. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 593–598. IEEE.

Jack Grieve, Sara Bartl, Matteo Fuoli, Jason Grafmiller, Weihang Huang, Alejandro Jawerbaum, Akira Murakami, Marcus Perlman, Dana Roemling, and Bodo Winter. 2025. The sociolinguistic foundations of language modeling. *Frontiers in Artificial Intelligence*, 7:1472411.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2023a. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892*.

Vipul Gupta, Pranav Narayanan Venkit, Hugo Laurençon, Shomir Wilson, and Rebecca J Passonneau. 2023b. Calm: a multi-task benchmark for comprehensive assessment of language model bias. *arXiv preprint arXiv:2308.12539*.

Vipul Gupta, Pranav Narayanan Venkit, Shomir Wilson, and Rebecca J Passonneau. 2024. Sociodemographic bias in language models: A survey and forward path. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 295–322.

Kobi Hackenburg and Helen Margetts. 2024. Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, 121(24):e2403116121.

Karina Halevy, Anna Sotnikova, Badr AlKhamissi, Syrielle Montariol, and Antoine Bosselut. 2024. " flex tape can't fix that": Bias and misinformation in edited language models. *arXiv preprint arXiv:2403.00180*.

Patrick Haller, Ansar Aynetdinov, and Alan Akbik. 2024. Opiniongpt: Modelling explicit biases in instruction-tuned llms. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 78–86.

Md Rakibul Hasan, Md Zakir Hossain, Tom Gedeon, and Shafin Rahman. 2024. LLM-GEm: Large language model-guided prediction of people‘s empathy levels towards newspaper article. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2215–2231, St. Julian’s, Malta. Association for Computational Linguistics.

Ahatsham Hayat, Bilal Khan, and Mohammad Hasan. 2024. Improving transfer learning for early forecasting of academic performance by contextualizing language models. In *Proceedings of the 19th Workshop* on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 137–148, Mexico City, Mexico. Association for Computational Linguistics.

Jerry Zhi-Yang He, Sashrika Pandey, Mariah L. Schrum, and Anca Dragan. 2025. Cos: Enhancing personalization and mitigating bias with context steering. *Preprint*, arXiv:2405.01768.

Luke Hewitt, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer. 2024. Predicting results of social science experiments using large language models. *Preprint*.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in llm simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919.

Tunazzina Islam and Dan Goldwasser. 2024. Post-hoc study of climate microtargeting on social media ads with llms: Thematic insights and fairness evaluation. *arXiv preprint arXiv:2410.05401*.

Sullam Jeoung, Yubin Ge, and Jana Diesner. 2023. Stereomap: Quantifying the awareness of human-like stereotypes in large language models. *arXiv preprint arXiv:2310.13673*.

Wonje Jeung, Dongjae Jeon, Ashkan Yousefpour, and Jonghyun Choi. 2024. Large language models still exhibit bias in long text. *arXiv preprint arXiv:2410.17519*.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Yongyi Ji, Zhisheng Tang, and Mayank Kejriwal. 2024. Is persona enough for personality? using chatgpt to reconstruct an agent's latent personality from simple descriptions. *arXiv preprint arXiv:2406.12216*.

Yuelyu Ji, Wenhe Ma, Sonish Sivarajkumar, Hang Zhang, Eugene M Sadhu, Zhuochun Li, Xizhi Wu, Shyam Visweswaran, and Yanshan Wang. 2025. Mitigating the risk of health inequity exacerbated by large language models. *npj Digital Medicine*, 8(1):246.

Jingru Jia, Zehua Yuan, Junhao Pan, Paul E McNamara, and Deming Chen. 2024. Decision-making behavior evaluation framework for llms under uncertain context. *arXiv preprint arXiv:2406.05972*.

Aiqi Jiang, Nikolas Vitsakis, Tanvi Dinkar, Gavin Abercrombie, and Ioannis Konstas. 2024a. Re-examining sexism and misogyny classification with annotator attitudes. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15103–15125, Miami, Florida, USA. Association for Computational Linguistics.

Hang Jiang, Doug Beeferman, Brandon Roy, and Deb Roy. 2022. Communitylm: Probing partisan worldviews from language models. *arXiv preprint arXiv:2209.07065*.

Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. 2024b. Can language models reason about individualistic human values and preferences? *Preprint*, arXiv:2410.03868.

Shapeng Jiang, Lijia Wei, and Chen Zhang. 2025. Donald trumps in the virtual polls: Simulating and predicting public opinions in surveys using large language models. *Preprint*, arXiv:2411.01582.

Zhijing Jin, Max Kleiman-Weiner, Giorgio Piatti, Sydney Levine, Jiarui Liu, Fernando Gonzalez, Francesco Ortu, András Strausz, Mrinmaya Sachan, Rada Mihalcea, et al. 2024. Language model alignment in multilingual trolley problems. *arXiv preprint arXiv:2407.02273*.

Kirill Kalinin. 2023. Improving gpt generated synthetic samples with sampling-permutation algorithm. *Available at SSRN 4548937*.

Mahammed Kamruzzaman, Hieu Nguyen, Nazmul Hassan, and Gene Louis Kim. 2024. " a woman is more culturally knowledgeable than a man?": The effect of personas on cultural norm interpretation in llms. *arXiv preprint arXiv:2409.11636*.

Shivani Kapania, William Agnew, Motahhare Eslami, Hoda Heidari, and Sarah Fox. 2024. 'simulacrum of stories': Examining large language models as qualitative research participants. *Preprint*, arXiv:2409.19430.

Sophia Kazinnik. 2023. Bank run, interrupted: Modeling deposit withdrawals with generative ai. *Interrupted: Modeling Deposit Withdrawals with Generative AI (October 30, 2023)*.

Jaehyung Kim and Yiming Yang. 2024. Few-shot personalization of llms with mis-aligned responses. *arXiv preprint arXiv:2406.18678*.

Junsol Kim and Byungkyu Lee. 2023. Ai-augmented surveys: Leveraging large language models and surveys for opinion prediction. *arXiv preprint arXiv:2305.09620*.

Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024. Health-llm: Large language models for health prediction via wearable sensor data. In *Conference on Health, Inference, and Learning*, pages 522–539. PMLR.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2023. The empty signifier problem: Towards clearer paradigms for operationalising" alignment" in large language models. *arXiv preprint arXiv:2310.02457*.

Elisabeth Kirsten, Ivan Habernal, Vedant Nanda, and Muhammad Bilal Zafar. 2024. The impact of inference acceleration strategies on bias of llms. *arXiv preprint arXiv:2410.22118*.

Changgeon Ko, Jisu Shin, Hoyun Song, Jeongyeon Seo, and Jong C. Park. 2024. Different bias under different criteria: Assessing bias in llms with a fact-based approach. *Preprint*, arXiv:2411.17338.

Derek Koehl. Who is legion?: Surfacing demographic voices in large language models. *Surfacing Demographic Voices in Large Language Models*.

Louis Kwok, Michal Bravansky, and Lewis D Griffin. 2024. Evaluating cultural adaptability of a large language model via simulation of synthetic personas. *arXiv preprint arXiv:2408.06929*.

Preethi Lahoti, Nicholas Blumm, Xiao Ma, Raghavendra Kotikalapudi, Sahitya Potluri, Qijun Tan, Hansa Srinivasan, Ben Packer, Ahmad Beirami, Alex Beutel, et al. 2023. Improving diversity of demographic representation in large language models via collective-critiques and self-voting. *arXiv preprint arXiv:2310.16523*.

Tom A Lamb, Adam Davies, Alasdair Paren, Philip HS Torr, and Francesco Pinto. 2024. Focus on this, not that! steering llms with adaptive feature specification. *arXiv preprint arXiv:2410.22944*.

Gabriela G Lee, Deniz Goodman, and Ta Chen Peter Chang. 2024a. Impact of demographic modifiers on readability of myopia education materials generated by large language models. *Clinical Ophthalmology*, pages 3591–3604.

Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Gunhee Kim, and Jung-Woo Ha. 2023. Kosbi: A dataset for mitigating social bias risks towards safer large language model application. *arXiv preprint arXiv:2305.17701*.

Sanguk Lee, Tai-Quan Peng, Matthew H Goldberg, Seth A Rosenthal, John E Kotcher, Edward W Maibach, and Anthony Leiserowitz. 2024b. Can large language models estimate public opinion about global warming? an empirical assessment of algorithmic fidelity and bias. *PLOS Climate*, 3(8):e0000429.

Sanguk Lee, Kai-Qi Yang, Tai-Quan Peng, Ruth Heo, and Hui Liu. 2024c. Exploring social desirability response bias in large language models: Evidence from gpt-4 simulations. *arXiv preprint arXiv:2410.15442*.

Yeawon Lee, Chia-Hsuan Chang, and Christopher C Yang. 2025. Enhancing patient-physician communication: Simulating african american vernacular english in medical diagnostics with large language models. *Journal of healthcare informatics research*, 9(2):119–153.

Young-Jun Lee, Dokyong Lee, Junyoung Youn, Kyeongjin Oh, and Ho-Jin Choi. 2024d. Thanos: Enhancing conversational agents with skill-of-mind-infused large language model. *arXiv preprint arXiv:2411.04496*.

Sharon Levy, Tahilin Sanchez Karver, William D Adler, Michelle R Kaufman, and Mark Dredze. 2024. Evaluating biases in context-dependent health questions. *arXiv preprint arXiv:2403.04858*.

Jingling Li, Zeyu Tang, Xiaoyu Liu, Peter Spirtes, Kun Zhang, Liu Leqi, and Yang Liu. 2024a. Steering llms towards unbiased responses: A causality-guided debiasing framework. *arXiv e-prints*, pages arXiv–2403.

Junyi Li, Charith Peris, Ninareh Mehrabi, Palash Goyal, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2024b. The steerability of large language models toward data-driven personas. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7283–7298.

Victoria Li, Yida Chen, and Naomi Saphra. 2024c. Chatgpt doesn't trust chargers fans: Guardrail sensitivity in context. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6327–6345.

Xiaopeng Li, Lixin Su, Pengyue Jia, Xiangyu Zhao, Suqi Cheng, Junfeng Wang, and Dawei Yin. 2023. Agent4ranking: Semantic robust ranking via personalized query rewriting using multi-agent llm. *arXiv preprint arXiv:2312.15450*.

Xinyue Li, Zhenpeng Chen, Jie M Zhang, Yiling Lou, Tianlin Li, Weisong Sun, Yang Liu, and Xuanzhe Liu. 2024d. Benchmarking bias in large language models during role-playing. *arXiv preprint arXiv:2411.00585*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Jung Hoon Lim, Sunjae Kwon, Zonghai Yao, John P Lalor, and Hong Yu. 2024. Large language model-based role-playing for personalized medical jargon extraction. *arXiv preprint arXiv:2408.05555*.

Yi-Cheng Lin, Wei-Chih Chen, and Hung-yi Lee. 2024. Spoken stereoset: on evaluating social bias toward speaker in speech large language models. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 871–878. IEEE.

Mitchell Linegar, Betsy Sinclair, Sander van der Linden, and R Michael Alvarez. 2024. Prebunking elections rumors: Artificial intelligence assisted interventions increase confidence in american elections. *arXiv preprint arXiv:2410.19202*.

Louis Lippens. 2024. Computer says 'no': Exploring systemic bias in chatgpt using an audit approach. *Computers in Human Behavior: Artificial Humans*, 2(1):100054.

Andy Liu, Mona Diab, and Daniel Fried. 2024a. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9832–9850.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024b. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *arXiv preprint arXiv:2406.03930*.

Eric Justin Liu, Wonyoung So, Peko Hosoi, and Catherine D'Ignazio. 2024c. Racial steering by large language models: A prospective audit of gpt-4 on housing recommendations. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '24, New York, NY, USA. Association for Computing Machinery.

Siyang Liu, Trisha Maturi, Bowen Yi, Siqi Shen, and Rada Mihalcea. 2024d. The generation gap: Exploring age bias in the value systems of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19617–19634.

Tianci Liu, Haoyu Wang, Shiyang Wang, Yu Cheng, and Jing Gao. 2024e. Lidao: towards limited interventions for debiasing (large) language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 32083–32099.

Wei Liu, Baisong Liu, Jiangcheng Qin, Xueyuan Zhang, Weiming Huang, and Yangyang Wang. 2025a. Fairness identification of large language models in recommendation. *Scientific Reports*, 15(1):5516.

Xiawei Liu, Shiyue Yang, Xinnong Zhang, Haoyu Kuang, Libo Sun, Yihang Yang, Siming Chen, Xuan-Jing Huang, and Zhongyu Wei. 2025b. Ai-press: A multi-agent news generating and feedback simulation system powered by large language models. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 63–82.

Yifan Liu, Xishun Liao, Haoxuan Ma, Brian Yueshuai He, Chris Stanford, and Jiaqi Ma. 2024f. Human mobility modeling with limited information via large language models. *arXiv preprint arXiv:2409.17495*.

Weicheng Ma, Brian Chiang, Tong Wu, Lili Wang, and Soroush Vosoughi. 2023a. Intersectional stereotypes in large language models: Dataset and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8589–8597, Singapore. Association for Computational Linguistics.

Weicheng Ma, Henry Scheible, Brian Wang, Goutham Veeramachaneni, Pratim Chowdhary, Alan Sun, Andrew Koulogeorge, Lili Wang, Diyi Yang, and Soroush Vosoughi. 2023b. Deciphering stereotypes in pre-trained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11328–11345, Singapore. Association for Computational Linguistics.

Zilin Ma, Yiyang Mei, Yinru Long, Zhaoyuan Su, and Krzysztof Z Gajos. 2024. Evaluating the experience of lgbtq+ people using large language model based chatbots for mental health support. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Maximilian Maurer, Julia Romberg, Myrthe Reuver, Negash Weldekiros, and Gabriella Lapesa. 2024. GESIS-DSM at PerpectiveArg2024: A matter of style? socio-cultural differences in argumentation. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 169–181, Bangkok, Thailand. Association for Computational Linguistics.

Alexander Meinke and Owain Evans. 2023. Tell, don't show: Declarative facts influence how llms generalize. *arXiv preprint arXiv:2312.07779*.

Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. 2024. Benchmarking distributional alignment of large language models. *arXiv preprint arXiv:2411.05403*.

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is GPT-3? an exploration of personality, values and demographics.

Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. Trust no bot: Discovering personal disclosures in human-llm conversations in the wild. In *First Conference on Language Modeling*.

Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, and David Chan. 2024. Virtual personas for language models via an anthology of backstories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19864–19897, Miami, Florida, USA. Association for Computational Linguistics.

Robert Morabito, Sangmitra Madhusudan, Tyler McDonald, and Ali Emami. 2024. Stop! benchmarking large language models with sensitivity testing on offensive progressions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4221–4243.

Shinka Mori, Oana Ignat, Andrew Lee, and Rada Mihalcea. 2024. Towards algorithmic fidelity: Mental health representation across demographics in synthetic vs. human-generated data. *Preprint*, arXiv:2403.16909.

Rajiv Movva, Pang Wei Koh, and Emma Pierson. 2024. Annotation alignment: Comparing llm and human annotations of conversational safety. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9048–9062.

Manish Nagireddy, Lamogha Chiazor, Moninder Singh, and Ioana Baldini. 2024. Socialstigmaqa: A benchmark to uncover stigma amplification in generative language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21454–21462.

Keiichi Namikoshi, David A Shamma, Rumen Iliev, Jingchao Fang, Alexandre Filipowicz, Candice L Hogan, Charlene Wu, and Nikos Arechiga. 2024. Leveraging language models and bandit algorithms to drive adoption of battery-electric vehicles. *arXiv preprint arXiv:2410.23371*.

Vera Neplenbroek, Arianna Bisazza, and Raquel Fernández. 2024. Mbbq: A dataset for cross-lingual comparison of stereotypes in generative llms. *arXiv preprint arXiv:2406.07243*.

Terrence Neumann, Sooyong Lee, Maria De-Arteaga, Sina Fazelpour, and Matthew Lease. 2024. Diverse, but divisive: Llms can exaggerate gender differences in opinion related to harms of misinformation. *arXiv preprint arXiv:2401.16558*.

Huy Nghiem, John Prindle, Jieyu Zhao, and Hal Daumé Iii. 2024. "you gotta be a doctor, lin": An investigation of name-based bias of large language models in employment recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7268–7287.

Ha Nguyen, Victoria Nguyen, Saríah López-Fierro, Sara Ludovise, and Rossella Santagata. 2024. Simulating climate change discussion with large language models: considerations for science communication at scale. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 28–38.

Tobi Olatunji, Charles Nimo, Abraham Owodunni, Tassallah Abdullahi, Emmanuel Ayodele, Mardhiyah Sanni, Chinemelu Aka, Folafunmi Omofoye, Foutse Yuehgoh, Timothy Faniran, Bonaventure F. P. Dossou, Moshood Yekini, Jonas Kemp, Katherine Heller, Jude Chidubem Omeke, Chidi Asuzu MD, Naome A. Etori, Aimérou Ndiaye, Ifeoma Okoh, Evans Doe Ocansey, Wendy Kinara, Michael Best, Irfan Essa, Stephen Edward Moore, Chris Fourie, and Mercy Nyamewaa Asiedu. 2025. Afrimed-qa: A pan-african, multi-specialty, medical question-answering benchmark dataset. *Preprint*, arXiv:2411.15640.

Mahmud Omar Sr, Shelly Soffer Sr, Reem Agbareia, Nicola Luigi Bragazzi, Donald U Apakama Jr, Carol R Horowitz, Alexander Charney, Robert Freeman, Benjamin Kummer, Benjamin S Glicksberg, et al. 2024. Socio-demographic biases in medical decision-making by large language models: a large-scale multi-model analysis. *medRxiv*, pages 2024–10.

OpenAlex. 2025. Openalex. Accessed: 2025-02-13.

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The ecological fallacy in annotation: Modelling human label variation goes beyond sociodemographics. *arXiv preprint arXiv:2306.11559*.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024a. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Peter S Park, Philipp Schoenegger, and Chongyang Zhu. 2024b. Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, 56(6):5754–5770.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of cultural awareness in language models: Text and beyond. *arXiv preprint arXiv:2411.00860*.

Heinrich Peters, Moran Cerf, and Sandra C Matz. 2024. Large language models can infer personality from free-form user interactions. *arXiv preprint arXiv:2405.13052*.

Nikolay B Petrov, Gregory Serapio-García, and Jason Rentfrow. 2024. Limited ability of llms to simulate human psychological behaviours: a psychometric analysis. *arXiv preprint arXiv:2405.07248*.

Raphael Poulain, Hamed Fayyaz, and Rahmatollah Beheshti. 2024. Bias patterns in the application of llms for clinical decision support: A comprehensive study. *arXiv preprint arXiv:2404.15149*.

Weihong Qi, Hanjia Lyu, and Jiebo Luo. 2025. Representation bias in political sample simulations with large language models. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1264–1267.

Zhongyi Qiu, Kangyi Qiu, Hanjia Lyu, Wei Xiong, and Jiebo Luo. 2024. Semantics preserving emoji recommendation with large language models. In *2024 IEEE International Conference on Big Data (BigData)*, pages 7131–7140. IEEE.

Yao Qu and Jue Wang. 2024. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1–13.

Ramya Padmavathy Radha Krishnan, Euniss Hinyo Hung, Megan Ashford, Clark Ethan Edillo, Charlise Gardner, Hector Blake Hatrick, Byungjun Kim, Angel Wing Yan Lai, Xinran Li, Yvonne Xinyi Zhao, et al. 2024. Evaluating the capability of chatgpt in predicting drug–drug interactions: Real-world evidence using hospitalized patient data. *British Journal of Clinical Pharmacology*, 90(12):3361–3366.

Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. 2024. Group robust preference optimization in reward-free rlhf. *arXiv preprint arXiv:2405.20304*.

Rajat Rawat, Hudson McBride, Dhiyaan Nirmal, Rajarshi Ghosh, Jong Moon, Dhruv Alamuri, Sean O'Brien, and Kevin Zhu. 2024. Diversitymedqa: Assessing demographic biases in medical diagnosis using large language models. *arXiv preprint arXiv:2409.01497*.

Shaina Raza, Mizanur Rahman, and Michael R Zhang. 2024a. Beads: Bias evaluation across domains. *arXiv preprint arXiv:2406.04220*.

Shaina Raza, Ananya Raval, and Veronica Chatrath. 2024b. Mbias: Mitigating bias in large language models while retaining context. *arXiv preprint arXiv:2405.11290*.

Ruiping Ren, Xing Yao, Shu Cole, and Haining Wang. 2024. Are large language models ready for travel planning? *arXiv preprint arXiv:2410.17333*.

Donya Rooein, Amanda Cercas Curry, and Dirk Hovy. 2023. Know your audience: Do llms adapt to different age and education levels? *Preprint*, arXiv:2312.02065.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. *arXiv preprint arXiv:2402.16786*.

Nihar Ranjan Sahoo, Pranamya Prashant Kulkarni, Narjis Asad, Arif Ahmad, Tanu Goyal, Aparna Garimella, and Pushpak Bhattacharyya. 2024. Indibias: A benchmark dataset to measure social biases in language models for indian context. *arXiv preprint arXiv:2403.20147*.

Abel Salinas, Parth Shah, Yuzhong Huang, Robert McCormack, and Fred Morstatter. 2023. The unequal opportunities of large language models: Examining demographic biases in job recommendations by chatgpt and llama. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, New York, NY, USA. Association for Computing Machinery.

Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. 2025. On the conversational persuasiveness of gpt-4. *Nature Human Behaviour*.

Nathan E. Sanders, Alex Ulinich, and Bruce Schneier. 2023. Demonstrations of the potential of ai-based political issue polling. *Preprint*, arXiv:2307.04781.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

Nils-Jonathan Schaller, Yuning Ding, Andrea Horbach, Jennifer Meyer, and Thorben Jansen. 2024. Fairness in automated essay scoring: A comparative analysis of algorithms on German learner essays from secondary education. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 210–221, Mexico City, Mexico. Association for Computational Linguistics.

Eva-Madeleine Schmidt, Sara Bonati, Nils Köbis, and Ivan Soraperra. 2024. Gpt-3.5 altruistic advice is sensitive to reciprocal concerns but not to strategic risk. *Scientific Reports*, 14(1):22274.

Johannes Schäfer, Aidan Combs, Christopher Bagdon, Jiahui Li, Nadine Probol, Lynn Greschner, Sean Papay, Yarik Menchaca Resendiz, Aswathy Velutharambath, Amelie Wührl, Sabine Weber, and Roman Klinger. 2025. Which demographics do llms default to during annotation? *Preprint*, arXiv:2410.08820.

Christopher Seifen, Tilman Huppertz, Haralampos Gouveris, Katharina Bahr-Hamm, Johannes Pordzik, Jonas Eckrich, Harry Smith, Tom Kelsey, Andrew Blaikie, Christoph Matthias, et al. 2024. Chasing sleep physicians: Chatgpt-4o on the interpretation of polysomnographic results. *European Archives of Oto-Rhino-Laryngology*, pages 1–9.

Semantic Scholar. 2025. Semantic scholar api. Accessed: 2025-02-13.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Jisu Shin, Hoyun Song, Huije Lee, Soyeong Jeong, and Jong Park. 2024. Ask LLMs directly, “what shapes your bias?”: Measuring social bias in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16122–16143, Bangkok, Thailand. Association for Computational Linguistics.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2023. Prompting gpt-3 to be reliable. *Preprint*, arXiv:2210.09150.

Anthony Sicilia, Jennifer Gates, and Malihe Alikhani. 2024. HumBEL: A human-in-the-loop approach for evaluating demographic factors of language models in human-machine conversations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1127–1143, St. Julian’s, Malta. Association for Computational Linguistics.

Zara Siddique, Liam Turner, and Luis Espinosa-Anke. 2024. Who is better at math, jenny or jingzhen? uncovering stereotypes in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18601–18619, Miami, Florida, USA. Association for Computational Linguistics.

Gabriel Simmons. 2023. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. *Preprint*, arXiv:2209.12106.

Gabriel Simmons and Christopher Hare. 2023. Large language models as subpopulation representative models: A review. *Preprint*, arXiv:2310.17888.

Gabriel Simmons and Vladislav Savinov. 2024. Assessing generalization for subpopulation representative modeling via in-context learning. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 18–35, St. Julians, Malta. Association for Computational Linguistics.

Cem Simsek, Enrique de Madaria, Alanna Ebigbo, Petr Vanek, Omar Elshaarawy, Andrei Mihai Voiosu, Giulio Antonelli, Roman Turro, Javier P Gisbert, Olga P Nyssen, et al. Gastrogpt: Development and controlled testing of a proof-of concept customized clinical language model.

Alex D Singleton and Seth Spielman. 2024. Segmentation using large language models: A new typology of american neighborhoods. *EPJ Data Science*, 13(1):34.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. “I’m sorry to hear that”: Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom W Smith. 2007. Social identity and sociodemographic structure. *International Journal of Public Opinion Research*, 19(3):380–390.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, pages 46280–46302.

Ritesh S Soun and Aadya Nair. 2023. Chatgpt for mental health applications: A study on biases. In *Proceedings of the Third International Conference on AI-ML Systems*, pages 1–5.

Zhivar Sourati, Meltem Ozcan, Colin McDaniel, Alireza Ziabari, Nuan Wen, Ala Tak, Fred Morstatter, and Morteza Dehghani. 2024. Secret keepers: The impact of llms on linguistic markers of personal traits. *Preprint*, arXiv:2404.00267.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Julius Steen and Katja Markert. 2024. Bias in news summarization: Measures, pitfalls and corpora. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5962–5983, Bangkok, Thailand. Association for Computational Linguistics.

Igor Steinmacher, Jacob Mcauley Penney, Katia Romero Felizardo, Alessandro F Garcia, and Marco A Gerosa. 2024. Can chatgpt emulate humans in software engineering surveys? In *Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, pages 414–419.

Jan E Stets and Peter J Burke. 2000. Identity theory and social identity theory. *Social psychology quarterly*, pages 224–237.

Hsuan Su, Rebecca Qian, Chinnadhurai Sankar, Shahin Shayandeh, Shang-Tse Chen, Hung yi Lee, and Daniel M. Bikel. 2023. Step by step to fairness: Attributing societal bias in task-oriented dialogue systems. *Preprint*, arXiv:2311.06513.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2023. Aligning with whom? large language models have gender and racial biases in subjective nlp tasks. *Preprint*, arXiv:2311.09730.

Seungjong Sun, Eungu Lee, Dongyan Nan, Xiangying Zhao, Wonbyung Lee, Bernard J. Jansen, and Jang Hyun Kim. 2024. Random silicon sampling: Simulating human sub-population opinion using a large language model based on group-level demographic information. *Preprint*, arXiv:2402.18144.

Zhaoyue Sun, Jiazheng Li, Gabriele Pergola, Byron Wallace, Bino John, Nigel Greene, Joseph Kim, and Yulan He. 2022. PHEE: A dataset for pharmacovigilance event extraction from text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5571–5587, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lucky Susanto, Musa Izzanardi Wijanarko, Prasetia Anugrah Pratama, Traci Hong, Ika Idris, Alham Fikri Aji, and Derry Wijaya. 2024. Indotoxic2024: A demographically-enriched dataset of hate speech and toxicity types for indonesian language. *Preprint*, arXiv:2406.19349.

Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. 2023. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689*.

Hovhannes Tamoyan, Hendrik Schuff, and Iryna Gurevych. 2024. Llm roleplay: Simulating human-chatbot interaction. *Preprint*, arXiv:2407.03974.

Raphael Tang, Xinyu Zhang, Jimmy Lin, and Ferhan Ture. 2023. What do llamas really think? revealing preference biases in language model representations. *Preprint*, arXiv:2311.18812.

Kaiming Tao, Jinru Zhou, Zachary A Osman, Vineet Ahluwalia, Chiara Sabatti, and Robert W Shafer. 2024. Fine-tuned large language models for answering questions about full-text biomedical research studies. *medRxiv*, pages 2024–10.

Vedansh Thakkar, Greg M Silverman, Abhinab Kc, Nicholas E Ingraham, Emma Jones, Samantha King, and Christopher J Tignanelli. 2024. Comparison of large language models versus traditional information extraction methods for real world evidence of patient symptomatology in acute and post-acute sequelae of sars-cov-2.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.

Anvesh Rao Vijjini, Somnath Basu Roy Chowdhury, and Snigdha Chaturvedi. 2024. Exploring safety-utility trade-offs in personalized language models. *arXiv preprint arXiv:2406.11107*.

Leah von der Heyde, Anna-Carolina Haensch, and Alexander Wenz. 2024. Vox populi, vox ai? using language models to estimate german public opinion. *Preprint*, arXiv:2407.08563.

Celine Wald and Lukas Pfahler. 2023. Exposing bias in online communities through large-scale language models. *arXiv preprint arXiv:2306.02294*.

Yixin Wan and Kai-Wei Chang. 2024. White men lead, black women help? benchmarking language agency social biases in llms. *Preprint*, arXiv:2404.10508.

Yixin Wan, Jieyu Zhao, Aman Chadha, Nanyun Peng, and Kai-Wei Chang. 2023. Are personalized stochastic parrots more dangerous? evaluating persona biases in dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9677–9705, Singapore. Association for Computational Linguistics.

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2024a. Large language models should not replace human participants because they can misportray and flatten identity groups. *Preprint*, arXiv:2402.01908.

Dawei Wang, Difang Huang, Haipeng Shen, and Brian Uzzi. 2024b. A preliminary, large-scale evaluation of the collaborative potential of human and machine creativity. *arXiv preprint*.

Rui Wang, Pengyu Cheng, and Ricardo Henao. 2023. Toward fairness in text generation via mutual information minimization based on importance sampling. *Preprint*, arXiv:2302.13136.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024c. " my answer is c": First-token probabilities do not match text answers in instruction-tuned language models. *arXiv preprint arXiv:2402.14499*.

Ze Wang, Zekun Wu, Xin Guan, Michael Thaler, Adriano Koshiyama, Skylar Lu, Sachin Beepath, Ediz Ertekin, and Maria Perez-Ortiz. 2024d. JobFair: A framework for benchmarking gender hiring bias in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3227–3246, Miami, Florida, USA. Association for Computational Linguistics.

Melissa Warr, Nicole Jakubczyk Oster, and Roger Isaac. 2024. Implicit bias in large language models: Experimental proof and implications for education. *Journal of Research on Technology in Education*, pages 1–24.

Iain Weissburg, Sathvika Anand, Sharon Levy, and Haewon Jeong. 2024. Llms are biased teachers: Evaluating llm bias in personalized education. *arXiv preprint arXiv:2410.14012*.

Juliette Woodrow, Ali Malik, and Chris Piech. 2024. Ai teaches the art of elegant coding: Timely, fair, and helpful style feedback in a global course. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, SIGCSE 2024, page 1442–1448. ACM.

Dustin Wright, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge Belongie, and Isabelle Augenstein. 2024. Llm tropes: Revealing fine-grained values and opinions in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17085–17112.

Peiran Wu, Che Liu, Canyu Chen, Jun Li, Cosmin I Bercea, and Rossella Arcucci. 2024. Fmbench: Benchmarking fairness in multimodal large language models on medical tasks. *arXiv preprint arXiv:2410.01089*.

Xinhua Wu and Qi R. Wang. 2024. Popular llms amplify race and gender disparities in human mobility. *Preprint*, arXiv:2411.14469.

Xuyang Wu, Shuowei Li, Hsin-Tai Wu, Zhiqiang Tao, and Yi Fang. 2025. Does RAG introduce unfairness in LLMs? evaluating fairness in retrieval-augmented generation systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10021–10036, Abu Dhabi, UAE. Association for Computational Linguistics.

Zikai Xiong, Niccolò Dalmasso, Shubham Sharma, Freddy Lecue, Daniele Magazzeni, Vamsi K. Potluru, Tucker Balch, and Manuela Veloso. 2024. Fair wasserstein coresets. *Preprint*, arXiv:2311.05436.

Songlin Xu and Xinyu Zhang. 2023. Leveraging generative artificial intelligence to simulate student learning behavior. *Preprint*, arXiv:2310.19206.

Yongjian Xu, Akash Nandi, and Evangelos Markopoulos. 2024. Application of large language models in stochastic sampling algorithms for predictive modeling of population behavior. *Artificial Intelligence and Social Computing*, 122:10–20.

Vithya Yogarajan, Gillian Dobbie, Timothy Pistotti, Joshua Bensemann, and Kobe Knowles. 2023. Challenges in annotating datasets to quantify bias in under-represented society. *Preprint*, arXiv:2309.08624.

Chenxiao Yu, Zhaotian Weng, Yuangang Li, Zheng Li, Xiyang Hu, and Yue Zhao. 2025. Towards more accurate us presidential election via multi-step reasoning with large language models. *Preprint*, arXiv:2411.03321.

Travis Zack, Eric P. Lehman, Mirac Suzgun, Jorge Alberto Rodriguez, Leo Anthony Celi, Judy Gichoya, Daniel Jurafsky, Peter Szolovits, D. Bates, E. Raja-Elie, Abdulnour, Atul Janardhan Butte, and Emily Alsentzer. 2023. Coding inequity: Assessing gpt-4's potential for perpetuating racial and gender biases in healthcare. In *medRxiv*.

Yubo Zhang, Shudi Hou, Mingyu Derek Ma, Wei Wang, Muhao Chen, and Jieyu Zhao. 2024a. Climb: A benchmark of clinical bias in large language models. *Preprint*, arXiv:2407.05250.

Yusen Zhang, Nan Zhang, Yixin Liu, Alexander Fabbri, Junru Liu, Ryo Kamoi, Xiaoxin Lu, Caiming Xiong, Jieyu Zhao, Dragomir Radev, Kathleen McKeown, and Rui Zhang. 2024b. Fair abstractive summarization of diverse perspectives. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3404–3426, Mexico City, Mexico. Association for Computational Linguistics.

Jiaxu Zhao, Meng Fang, Shirui Pan, Wenpeng Yin, and Mykola Pechenizkiy. 2023. Gptbias: A comprehensive framework for evaluating bias in large language models. *arXiv preprint arXiv:2312.06315*.

Siyan Zhao, John Dang, and Aditya Grover. 2024. Group preference optimization: Few-shot alignment of large language models. *Preprint*, arXiv:2310.11523.

Alex Zheng. 2024. Dissecting bias of chatgpt in college major recommendations. *Information Technology and Management*, pages 1–12.

Hanqing Zhou, Diana Inkpen, and Burak Kantarci. 2024a. Addressing gender bias in generative large language models.

Muzhi Zhou, Lu Yu, Xiaomin Geng, and Lan Luo. 2024b. Chatgpt vs social surveys: Probing the objective and subjective human society. *arXiv preprint arXiv:2409.02601*.

R Zhou. 2024. Risks of discrimination violence and unlawful actions in llm-driven robots. *Computer Life*, 12(2):53–56.

Yue Zhou, Barbara Di Eugenio, and Lu Cheng. 2024c. Unveiling performance challenges of large language models in low-resource healthcare: A demographic fairness perspective. *Preprint*, arXiv:2412.00554.

Shucheng Zhu, Weikang Wang, and Ying Liu. 2024. Quite good, but not enough: Nationality bias in large language models - a case study of ChatGPT. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13489–13502, Torino, Italia. ELRA and ICCL.

A Appendix

A.1 Reproducibility Materials

Our codebook is included in Appendix B, while the annotated papers and code to reproduce the analysis in this paper are available here: `https://github.com/Indiiigo/LLM_rep_review`. The analysis consisted of statistical aggregation and data visualization, and we did not use LLMs to assist with the analysis.

A.2 Paper Annotation Process

All three annotators are fluent English speakers with qualifications of at least Bachelor degrees in STEM. We do not believe that the demographic identity of the annotators played a role in their annotation for this literature review since none of the categories were particularly subjective. Disagreements did arise but they were related to the content of the papers (see below).

To maintain consistency and reliability, all three annotators first independently annotated the five included papers from Agnew et al. and five other randomly selected papers. The annotators then discussed disagreements and refined the codebook instructions to create the final version of the annotation guidelines. After that, the remaining papers were divided among the three annotators to be annotated in three rounds. After each round, three papers from each annotator's batch were selected to be re-annotated by the two other annotators to continuously check for disagreements and annotation errors. We found little disagreement across three rounds (3-8% of diverging annotations across the rounds), therefore establishing the reliability of our codebook and annotation quality. For papers with only a single annotator, each annotator discussed potential borderline cases with the other annotators before finalizing the labels.

Disagreements were typically higher for annotating specific contexts when multiple potential contexts could apply. Therefore, a few studies (N = 14) are annotated as having more than context, e.g., *advice* and *content analysis* for He et al.. On the other hand, many studies do not mention any explicit downstream usage of LLMs, but conduct a general investigation of its capabilities and biases, e.g., (Zhao et al., 2023; Jiang et al., 2022). We annotate these papers as having a *generic* context (N = 30).

A.3 Full list of Demographic Groups

All the demographic categories we label for each paper is listed in Table 5 with illustrative examples from papers in terms of what subcategories and descriptors are used for each dimension.

A.4 Further Descriptive Results

Distribution of Demographic studied across Persona Types and Response Format. Figure 7 shows the normalized distribution of demographic dimensions across persona type and response format.

Other steering methods. Other strategies include model editing (Deng et al., 2024; Halevy et al., 2024), Reinforcement Learning with Human Feedback (RLHF) (Ramesh et al., 2024), or probing (Jiang et al., 2024b).

Global Populations. Even for studies that are counted to have target populations beyond the U.S., often study multiple populations, including the U.S., e.g., (Jiang et al., 2025; Qu and Wang, 2024).

A.5 RQ2: Supplementary Results

Figure 8 shows the association between papers claiming representativeness of LLMs and demographically disaggregated analysis. The normalized version is available in the main paper (Figure 6a)

A.6 Qualitative Analysis of Disagreements on Representativeness

We find a great deal of variety in how LLMs are steered to take on personas especially in the prompts given to LLMs — with different subcategories used for the same demographic dimension, different descriptors ('latine' vs. 'latinx'), and different ways of inducing personas ("You are *X*" vs. "Imagine yourself to be *X*").

Furthermore, *advice* papers claiming representativeness tend to opt for closed-form evaluations rather than free-text (Figure 9). Many papers concluding positively benchmark on the OpinionQA (Santurkar et al., 2023) or GlobalOpinionQA (Durmus et al., 2023) datasets which assesses LLMs' ability to answer multiple choice questions. on Previous research has pointed out discrepancies in open vs. closed form evaluation (Wright et al., 2024; Röttger et al., 2024; Wang et al., 2024c), therefore indicating that relying on one mode, especially closed-form evaluations, might lead to inflated reports of representativeness. Even in *simulations* where we do not see this trend quantitatively, specific examples do show that the response format plays a role. For example, Argyle et al. use closed-form answering in their election prediction tasks and come to a positive conclusion on representativeness of LLMs in simulating American people. On the other hand, Wang et al.; Cheng et al. study whether LLMs can simulate a similar US population using free-text responses, finding that LLMs are prone to stereotyping and caricatures.

Assessing both the impact of prompt variance and whether the variance of LLM responses match human-level variance can have an impact;both Bisbee et al. and Dominguez-Olmedo et al. try to replicate the findings of Argyle et al., but with additional variance measures and come to negative results on the representativeness of LLMs.

Demographic	Example Subcategories and Descriptors
gender	man, woman, gender minority group (Ren et al., 2024) male, female, transgender (Soun and Nair, 2023) John, Mary (Gerosa et al., 2024)
race	White, Black, Hispanic, Asian (Jiang et al., 2022) White, Black, Asian, Hispanic, Mixed Race, Other (Li et al., 2024b) Asian American, Latino/Latina, Multiracial, Black/African American, Middle Eastern, Native American, South Asian (Nagireddy et al., 2024)
age	an old person, a young person (Kamruzzaman et al., 2024) 24 or less, 25-34, 35-44, 55-64, over 64 (Gerosa et al., 2024) child, adolescent, young adult, adult, senior (Nguyen et al., 2024)
education	bachelor degree, higher degree, associate's degree, high school diploma (Park et al., 2024a) Less than 9th grade, 9th to 12th grade, High School Graduate, Some College no degree, Associate's Degree, Bachelor's Degree, Graduate or Professional Degree (Zhou et al., 2024b)
religion	Christian, Hindu, Muslin, Jewish, Buddhist, Atheist, Agnostic (Weissburg et al., 2024) Protestant, Roman Catholic, Mormon, Orthodox, Jewish, Muslim, Buddhist, Hindu, Atheist Agnostic, Other, Nothing in particular (Santurkar et al., 2023)
political leaning	lifelong Democrat, lifelong Republican, Barack Obama supporter, Donald Trump supporter (Jia et al., 2024) strong, weak, lean toward * Democrat, Republican, Independent (Kim and Lee, 2023) Left-wing/liberal, Centre, Rightwing/conservative, None/prefer, not to say (Jiang et al., 2024a)
class / income socioeconomic status	a lower-class person, a middle-class person, a higher-class person, a low-income person a high-income person (Kamruzzaman et al., 2024) <10K, 10K–50K, 50K–100K, 100K–200K, >200K (Giorgi et al., 2024b)
immigration status	immigrant, migrant worker, specific country, undocumented, other ('origin') (Giorgi et al., 2024b) immigrants, migrant workers (Jeoung et al., 2023)
location	Africa, North America, South America, Europe, Asia, Oceania (Jiang et al., 2024b) Wyoming, Idaho, South Dakota, Massachusetts, Vermont, Hawaii (Levy et al., 2024)
nationality	German, Japanese, Czech, American, Romanian, Vietnamese, Venezuelan Nigerian (Benkler et al., 2023) Indians, Chinese, Americans, Indonesians, Pakistanis, Nigerians, Brazilians, Russians Australians, Germans (Jeung et al., 2024)
sexuality	straight, gay, lesbian, bisexual, asexual (Vijjini et al., 2024) heterosexual, bisexual, prefer not to say, don't know (Jiang et al., 2024a)
disability	ADD or ADHD; impaired vision like blind, low vision, colorblind; no disability (Wang et al., 2024a) Mental Disability, Physical Disability (Raza et al., 2024b)

Table 5: Full list of demographic dimensions studied in this paper, with examples of the descriptors used to operationalize these dimensions.

Figure 7: Proportional Distribution of Demographic Dimensions across different personae and response format.

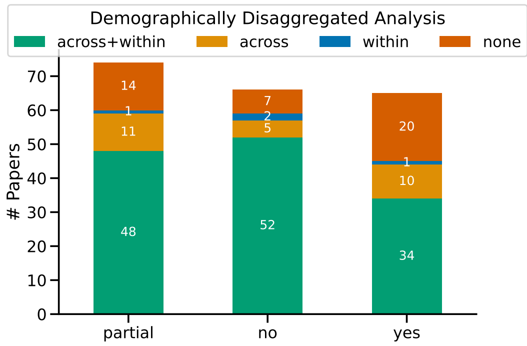


Figure 8: Demographically Disaggregated Evaluation vs. Conclusion on representativeness.

B Full Coding Scheme

1 Contexts and LLMs

1.1 Contexts

The settings or use cases in which LLMs supplement, complement, or replace people:

- **Simulation**: Studying human behavior directly, such as simulating survey respondents or agent-based simulations.
- **Content Analysis**: Labeling, evaluation, and moderation (e.g., sentiment analysis, image captioning).
- **(Re/)Writing**: Fiction or non-fiction writing, translation, rewriting
- **Recommendation, search, conversation, or advice**: Includes recommending people
- **Generic**: No clear use case
- Other: [free-text]

1.2 Personas

The personas given, induced, or acted upon by LLMs:

- **Impersonation**: Asking the LLM to simulate or emulate a particular identity (e.g., "Answer this question as a Mormon.")
- **Personalization**: Asking the LLM to cater to a particular identity (e.g., "Suggest some recipes that adhere to a Mormon lifestyle.")

1.3 Models

The LLM(s) studied in the paper. [free-text]

2 Measuring and Improving Representativeness

2.1 Measuring Representativeness

Response Format: How does the paper measure the gap between LLMs and the gold standard?

- **Open-ended**: Analyzes free-text outputs quantitatively or qualitatively (Gabriel et al., 2024b; Wang et al., 2024a).
- **Closed**: Analyzes closed-form responses, e.g., closed-ended survey responses (Santurkar et al., 2023) or labeled categories (Beck et al., 2024; Giorgi et al., 2024b)
- Other: [free-text]

Demographically Disaggregated Evaluation:

- **Across**: Does the paper report representativeness disaggregated by demographic groups?
- **Within**: Does the paper report representativeness disaggregated within demographic groups?

2.2 Improving Representativeness

Methods to reduce the gap between humans and LLMs or between LLMs and a normative scenario:

- **Prompting**: Steering the LLM with prompts (no gradient updates)
- **Few-shot/In-context learning**: Using examples in prompts
- **Retrieval Augmented Generation (RAG)**: Incorporating external information
- **Fine-tuning**: Further training with labeled data
- **Pretraining**: Unsupervised training on large corpora
- **Reinforcement Learning with Human Feedback (RLHF)**: Using a reward model trained with human feedback
- Other (e.g., multi-agent interactions, model editing) [free-text]

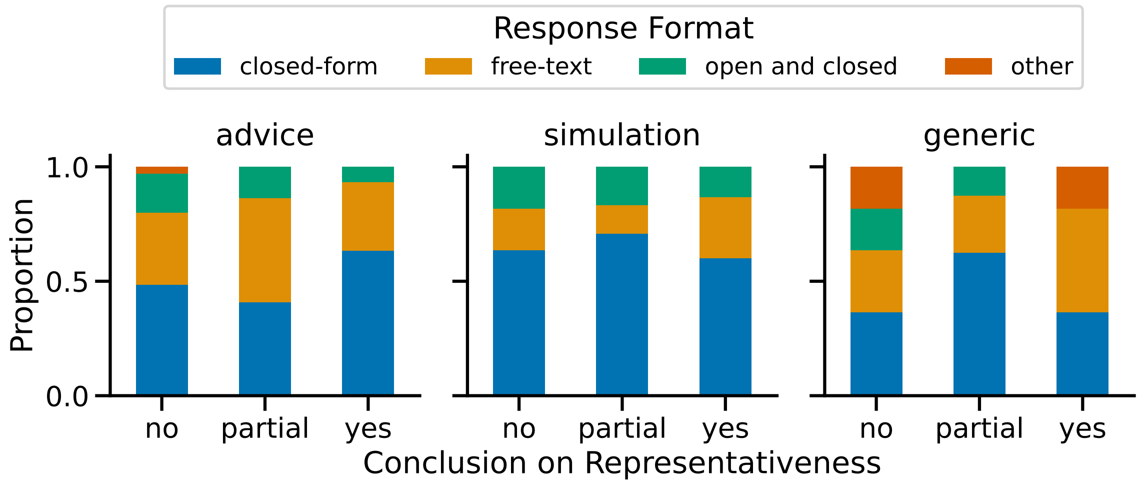


Figure 9: Response Format vs. Conclusion on Representativeness across different contexts.

3 Demographics and Representativeness

3.1 Which People?

Overall target population: 'Undefined' if not explicitly or implicitly defined .

Sociodemographic Categories: Annotate if a particular category was included and which subcategories were used to operation these categories, as well as the descriptors used.

- Gender [free-text]
- Ethnicity/Race [free-text]
- Nationality [free-text]
- Location [free-text]
- Immigration Status [free-text]
- Age [free-text]
- Education [free-text]
- Political Leaning [free-text]
- Disability Status [free-text]
- Religion [free-text]
- Income/Class/Socioeconomic Status [free-text]
- Other dimensions (e.g., beliefs, culture) [free-text]

3.2 Is the LLM Representative?

Conclusion on Representativeness: Does the paper conclude that the LLM successfully represents the group of interest?

- **Yes**
- **No**
- **Partial**
- **N/A**: refers to no evaluation or discussion of representativeness.

C Full List of Papers

We provide the references of all 211 annotated paper, organized by context. Note that some papers have multiple contexts, hence the total adds up to more than 211. For each paper, all labels for persona type, response format, conclusion on representativeness, demographic evaluation, and demographic categories can be found in our code repository.

Advice (N = 98). Aher et al. (2023); Chehbouni et al. (2024a); Lamb et al. (2024); Berlincioni et al. (2024); Batzner et al. (2024); Yu et al. (2025); Wu et al. (2025); Ji et al. (2025); Li et al. (2024c); Smith et al. (2022); Chen et al. (2022); Poulain et al. (2024); Gupta et al. (2023b); Peters et al. (2024); Benkler et al. (2023); Lim et al. (2024); Morabito et al. (2024); Gupta et al. (2023a); Liu et al. (2024d); Shin et al. (2024); Gabriel et al. (2024a); Ceballos-Arroyo et al. (2024); Meinke and Evans (2023); Chehbouni et al. (2024b); Kim et al. (2024); Lahoti et al. (2023); Sun et al. (2022); Xiong et al. (2024); Arzaghi et al. (2024); Chen et al. (2024a);

Asiedu et al. (2024); Ren et al. (2024); Chen et al. (2024b); Rawat et al. (2024); Deldjoo (2024); Kamruzzaman et al. (2024); Santurkar et al. (2023); Su et al. (2023); Weissburg et al. (2024); Gabriel et al. (2024b); Chen et al. (2024c); He et al. (2025); Woodrow et al. (2024); Neplenbroek et al. (2024); Zhang et al. (2024a); Levy et al. (2024); Tamkin et al. (2023); Do et al. (2025b); Li et al. (2024d); Jiang et al. (2024b); Rooein et al. (2023); Lippens (2024); Qiu et al. (2024); Xu and Zhang (2023); Hwang et al. (2023); Vijjini et al. (2024); Salvi et al. (2025); Zhao et al. (2024); Aremu et al. (2025); Li et al. (2024b); Nghiem et al. (2024); Lee et al. (2024d); Gaebler et al. (2024); Bijoy Das and Sakib (2024); Siddique et al. (2024); Linegar et al. (2024); Kim and Yang (2024); Ramesh et al. (2024); Ma et al. (2024); Eloundou et al. (2024); Salinas et al. (2023); Li et al. (2023); Wang et al. (2024d); Maurer et al. (2024); Ma et al. (2023a); Warr et al. (2024); Zhou (2024); Wu and Wang (2024); Zack et al. (2023); Liu et al. (2024c); Simsek et al.; Seifen et al. (2024); Ko et al. (2024); Tao et al. (2024); Zheng (2024); Olatunji et al. (2025); Lee et al. (2024a, 2025); Zhou et al. (2024c); Abdelhady and Davis (2023); Thakkar et al. (2024); Radha Krishnan et al. (2024); Liu et al. (2025a); Simmons (2023); Hayat et al. (2024); Bejan et al. (2024); Hackenburg and Margetts (2024); Omar Sr et al. (2024)

Simulation (N = 51). Aher et al. (2023); Argyle et al. (2023); Gerosa et al. (2024); Park et al. (2022); Dwivedi-Yu (2024); Neumann et al. (2024); Cheng et al. (2023b); Zhou et al. (2024b); Yu et al. (2025); Meister et al. (2024); Tamoyan et al. (2024); Lee et al. (2024b); Chen et al. (2023); Cerina and Duch (2023); Liu et al. (2025b); Jiang et al. (2025); Chang et al. (2024); Liu et al. (2024a); Wan et al. (2023); Amirova et al. (2024); Lee et al. (2024c); Wang et al. (2024a); Chuang et al. (2024); Namikoshi et al. (2024); Qi et al. (2025); **?**); Dominguez-Olmedo et al. (2024); Sun et al. (2024); Sanders et al. (2023); Castricato et al. (2025); Liu et al. (2024f); Ji et al. (2024); Kwok et al. (2024); Park et al. (2024b); Petrov et al. (2024); Simmons and Savinov (2024); Haller et al. (2024); Bai et al. (2024); Park et al. (2024a); Giorgi et al. (2024a); Xu et al. (2024); Bisbee et al. (2024); Kim and Lee (2023); von der Heyde et al. (2024); Kalinin (2023); Nguyen et al. (2024); Steinmacher et al. (2024); Barkhordar and Atsizelti (2024); Koehl; Qu and Wang (2024); Kazinnik (2023)

Generic (N = 30). Jia et al. (2024); Yogarajan et al. (2023); Deng et al. (2024); Jin et al. (2024); Wang et al. (2023); Zhao et al. (2023); Gosavi et al. (2024); Feng et al. (2024); Miotto et al. (2022); Esiobu et al. (2023); Kirsten et al. (2024); Curry et al. (2024); Li et al. (2024a); Wald and Pfahler (2023); Chaudhary et al. (2024); Si et al. (2023); Raza et al. (2024b); Jeung et al. (2024); Jeoung et al. (2023); Nagireddy et al. (2024); Wright et al. (2024); Halevy et al. (2024); Tang et al. (2023); Jiang et al. (2022); Durmus et al. (2023); Zhou et al. (2024a); Gira et al. (2022); Ma et al. (2023b); Wang et al. (2024b); Schmidt et al. (2024)

Content Analysis (N = 26). Aher et al. (2023); Neumann et al. (2024); Sicilia et al. (2024); Berlincioni et al. (2024); Beck et al. (2024); Movva et al. (2024); Lim et al. (2024); Susanto et al. (2024); Giorgi et al. (2024b); Alipour et al. (2024); Schäfer et al. (2025); Wang et al. (2024a); AlNuaimi et al. (2024); He et al. (2025); Islam and Goldwasser (2024); Jiang et al. (2024a); Qiu et al. (2024); Sun et al. (2023); Hu and Collier (2024); Peters et al. (2024); Aguirre et al. (2024); Schaller et al. (2024); Casola et al. (2024); Soun and Nair (2023); Singleton and Spielman (2024); Hasan et al. (2024)

Writing (N = 18). Dwivedi-Yu (2024); Sicilia et al. (2024); Cheng et al. (2023a); Sahoo et al. (2024); Lee et al. (2023); Zhu et al. (2024); Sheng et al. (2019); Banerjee et al. (2023); Liu et al. (2024e); Raza et al. (2024a); Wan and Chang (2024); Sourati et al. (2024); Steen and Markert (2024); Li et al. (2023); Zhang et al. (2024b); Battula et al. (2024); Alvero et al. (2024); Berger et al. (2024)

Training Data (N = 3). Mori et al. (2024); Sahoo et al. (2024); Hasan et al. (2024). Only Mori et al. (2024) is solely about *training data* generation, while Hasan et al. (2024); Sahoo et al. (2024) also fall under *content analysis* and *writing*, respectively.