

Are Benchmarks a source of bias in LLMs? MMLU, CommonSenseQA and MGSM under the Microscope

Fanny Ducel¹, Shayan Ahmed Shariff², Ibrahim Al Kotob²,
Binesh Arakkal Remesh², Lucie Digoïn-Caparros², Aurélie Névéal¹, Karën Fort²,

¹Université Paris-Saclay, CNRS, LISN, Orsay, France

²Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France

Correspondence: fanny.ducel@universite-paris-saclay.fr

Abstract

Evaluation is an essential process in the development and use of NLP systems. Benchmarks are widely used to calibrate systems, determine the efficiency of techniques, but also to advertise recently launched language models. Not only do benchmarks allow for evaluation and system comparisons, but they are also commonly utilized as training and validation corpora. In this paper, we aim at characterizing the "world knowledge" that is put forth by benchmarks by closely inspecting their content. Through qualitative manual annotations (+15k instances) as well as quantitative automatic analyses, we uncover the presence of both explicit and implicit biases in MMLU, CommonSenseQA and MGSM, ranging from representational imbalance to bluntly sexist and racist instances. Our findings demonstrate that current NLP evaluation practices pose numerous quality issues, reinforce stereotypes and even contain explicitly harmful content. The current operationalization of benchmarks is questioned, and we provide recommendations and directions for further ethical evaluation.

Warning: this paper contains content that may be offensive or upsetting.

1 Introduction

Numerous research efforts document the stereotypical biases that Large Language Models (LLMs) encapsulate (Gallegos et al., 2024). We define a stereotypical bias as a "skew that produces a type of harm" (Crawford, 2017), and that is based on a stereotype, i.e., an over-generalizing statement about a social group (Hilton and Von Hippel, 1996). Previous work claim that bias can arise at any point throughout the development pipeline of machine learning (Suresh and Guttag, 2021) and NLP systems. Hovy and Prabhumoye (2021) outline five of the most common sources of bias: training data, the annotation process, input representations, the models, and the research design. Resnik (2025)

recently insisted on the inevitable nature of LLM biases, originating from their design.

In this paper, we argue that evaluation, as operationalized in current NLP benchmarks, could be another source of bias¹, especially as it leads to (representational) harm (Crawford, 2017). Ideally neutral, representative, and rigorous, evaluation has a crucial role in science as it measures and ensures the quality of systems, and their consistency with pre-defined goals and standards. In practice, and even more so since the advent of LLMs, NLP systems have been evaluated with "general purpose, task-agnostic" benchmarks (Litschko et al., 2023). Users and developers frequently rely on benchmark scores and leaderboards rankings as key criteria for model selection. Lately, popular LLM releases are accompanied with benchmark scores as a key indicator of their quality. Thus, if benchmarks were to encapsulate biases, they could lead to favoring biased models.

Besides, multiple inherent limitations of benchmarks have been highlighted in previous research efforts, ranging from epistemological considerations (e.g., about the construct validity of benchmarks, especially those claiming to be "general" (Raji et al., 2021; Litschko et al., 2023; Bean et al., 2025)) to more practical, content-oriented benchmark flaws (e.g., the impact of minor perturbations on scores (Alzahrani et al., 2024), or the presence of factual errors in benchmarks (Gema et al., 2025)). In parallel, Bowman and Dahl (2021) raised concerns about the possible incentivization of bias through benchmarks, and Raji et al. (2021) qualified benchmarks as being inherently specific, subjective, and "localized constructs". However, studies dedicated to the presence of stereotypical bias within benchmarks remain scarce, and only

¹Though we do not prove causality between LLM biases and benchmarks biases, we believe that this work constitutes a relevant first step, and indicates that further investigation and dedicated causality analyses should be conducted.

emerged recently (Singh et al., 2025; Kraft et al., 2025). We aim at identifying stereotypical bias and associated harms in LLM benchmarks by proposing the first study to provide manual annotations and automatic analyses of both implicit and explicit bias in benchmarks. We focus solely on representational harm, which is related to the ways social groups are depicted (or omitted), rather than allocational harm (which occurs "when a system allocates or withholds certain groups opportunities or resources" (Crawford, 2017)) as it is less directly relevant in the context of benchmarks' exploration. We also rely on Dev et al. (2022)'s taxonomy of representational harm: stereotyping, disparagement, dehumanization, erasure, and quality of service.

Our contributions are as follows: (i) we perform manual annotations on over 15,000 instances from three popular LLM benchmarks: MMLU (Hendrycks et al., 2021), CommonSenseQA (Talmor et al., 2019), and MGSM (Shi et al., 2022), (ii) we conduct automatic quantitative analyses on these three benchmarks, so as to measure representation disparities and detect gendered stereotypical associations, (iii) we analyze and classify results into explicit vs. implicit bias, highlighting the importance of implicit bias, which remains rarely addressed in the literature.

All manual and automatic annotations are made publicly available, along with filtered versions of the benchmarks where all harmful and stereotypical entries have been removed, and with all scripts used to conduct analyses².

2 Related Work

2.1 On Evaluation Practices in NLP

Evaluation practices have evolved drastically in the past 20 years, from small test suites like TSNLP to the shared tasks still in use today (Lehmann et al., 1996). The advent of LLMs has presented a new challenge for evaluation. As these systems are presented as general and task-agnostic, evaluation has shifted from specialized tasks and metrics to benchmarks aiming for global evaluation. However, this general, task-agnostic approach brings construct validity issues: as benchmarks are instantiated in specific data and metrics, they cannot measure general-purpose capabilities (Raji et al., 2021; Litschko et al., 2023). In a systematic review of 445 articles that introduce LLM benchmarks, Bean

et al. (2025) uncovered various forms of construct validity issues, emerging due to lack of discussion around benchmark validity, targeted phenomena, data sources and statistical testing. More generally, Thomas and Uminsky (2022) and Bender and Koller (2020) highlighted an overemphasis on specific metrics from the NLP community, linking it to the risks of Goodhart's law ("When a measure becomes the target, it ceases to be an effective measure") and pointing at the real-world harms it can lead to. Other authors argue for the adoption of new, different evaluation lenses, e.g., with a focus on real-world impact (Reiter, 2025).

2.2 On Flaws in LLM Benchmarks

Raji et al. (2021) define benchmarks as "a particular combination of a dataset or sets of datasets [...], and a metric, conceptualized as representing one or more specific tasks or sets of abilities, picked up by a community of researchers as a shared framework for the comparison of methods". Previous research efforts have highlighted flaws from specific, popular LLM benchmarks. Alzahrani et al. (2024) conducted systematic experiments to show that minor perturbations to benchmarks could lead to important changes in models' rankings, and also discussed benchmark leakage. MMLU has already been the focus of Singh et al. (2025), who annotated a subset of the benchmark for cultural bias in the context of its adaptation and translation to 42 languages. They pointed out the presence and impact of culturally sensitive instances, which mostly require Western cultural knowledge. Another study (Gema et al., 2025) proposed a subset of MMLU, devoid of any factual errors, and estimated that 6.49% of the entire dataset could contain errors. They raised the question of MMLU's reliability, as "the reliability of benchmarking results is only as robust as the quality of the dataset used". Most similar to our work, Kraft et al. (2025) investigated the presence of gender, religion, and geographic bias in 20 question-answering benchmarks through automatic entity mapping.

Our study presents novel aspects as it focuses on both explicit and implicit bias, on MGSM, and on the entire CSQA and MMLU test sets, and with both manual and automatic approaches. The present study proves complementary to previous analyses and allows for a detailed exploration of studied benchmarks.

²<https://gitlab.univ-lorraine.fr/p05683/biasinbenchmarks>

3 Material and methods

3.1 Selected benchmarks

As there is a wide variety of benchmarks available for evaluating LLMs, some spanning over tens of thousands of records such as BIG-Bench (Srivastava et al., 2023), we had to limit our search to the confines of our resources. Selected benchmarks should cover a variety of subjects, be used to evaluate prominent LLMs, and be publicly and freely accessible. We retained: Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021), for which we only consider the test set which contains 15,908³ multiple-choice questions (MCQ) in English about 57 subjects (e.g., *sociology*, *professional_law*, *virology*) that were "manually collected by graduate and undergraduate students from freely available sources online", and CommonSenseQA (CSQA) (Talmor et al., 2019), which contains 12,247 MCQ in English "that can be easily answered by humans without context, and require commonsense knowledge", and which were created by crowdworkers, who were asked to "author questions that describe the relation between concepts from ConceptNet" (Speer et al., 2016). The Multilingual Grade School Math Benchmark (MGSM) (Shi et al., 2022) was also selected to represent smaller, specialized benchmarks. It is composed of 250 mathematical problems extracted from Cobbe et al. (2021) and translated from English into ten other languages. Unlike MMLU and CSQA, it does not follow a MCQ format as only questions and correct answers are provided. For better comparability with MMLU and CSQA, we only consider the English subset.

3.2 Manual annotation protocol

In order to quantify the presence of stereotypes, extensive manual annotations were conducted. Manual annotation was deemed the best solution as stereotypes exist within a given context, and vary across time and place. Selected annotators come from different cultural backgrounds to ensure a better representativeness of what can be considered stereotypical.

Overview and annotators demographics The entirety of MGSM as well as of MMLU and CSQA test sets – representing respectively 250, 14,028,

and 1,140 instances – were manually annotated for stereotypes. The annotators are 5 authors of the present paper. Four of them performed manual annotations in the context of a paid NLP master-level internship, while the last annotator is a PhD candidate in NLP. They are from different cultural backgrounds, representing three world (sub)regions (South Asia, West Asia, Western Europe) and different gender identities. Guidelines were constructed and revised while collectively annotating the first 1,000 instances of MMLU. In total, for MMLU, each annotator individually labeled 3,000 instances, and a total of 2,028 instances were annotated by four annotators (see Appendix A). Final labels were selected based on majority.

Stereotype annotation guidelines Manual annotations were performed at question-level, including the premise, the question, the answer options and the answer deemed as correct. The annotation consisted in flagging stereotypical content, choosing one label among "explicit bias", "implicit bias", or "neutral" (no bias). In this context, a biased instance is one that can be considered harmful for a specific population, and possibly relying on a stereotype. A bias is said explicit if it is clearly displayed without ambiguity, whereas an implicit bias is derived from the context or is ambiguous (cf. § 5.1 and 5.2). For explicit and implicit bias, annotators must add a tag corresponding to the type(s) of bias (inspired by Nangia et al. (2020); Névoul et al. (2022): gender, ethnicity, socioeconomic status, LGBT+, nationality, disability, physical appearance, age, religion, or other). Two extra labels were created for non-stereotypical bias: political and cultural bias. These were used to flag instances that require specific cultural or political knowledge to be understood and answered. Detailed guidelines are provided in Appendix B, along with examples of manually annotated instances (cf. Appendix C).

Inter-annotator agreement Cohen’s Kappa (Cohen, 1960) was computed for each pair of annotators, and Fleiss’ Kappa (Fleiss, 1971) was computed for the total group agreement and for each triple of annotators. Fleiss’ Kappa for 4 annotators reached 0.728 on the 2,028 instances that were annotated by all 4 of them. Detailed agreements are provided in Appendix D.

Cited authors’ demographics extraction Two annotators in the group performed further analysis of the 618 deduplicated quotes featured in MMLU.

³Only the test set was considered as the entire benchmark contains over 115,600 questions. After removing duplicates from the test set, 14,028 questions remained.

They looked up the cited person (e.g., Immanuel Kant) and, after disambiguation, extracted their nationality, gender, and occupation from Wikipedia (e.g., German, male, philosopher).

Annotation of methodological issues While annotating for stereotypes, annotators were asked to report any other issue. The issues that arose from this process were related to methodological flaws. Dedicated categories were created, building upon the error detection framework proposed by Gema et al. (2025). These annotations were only exhaustively conducted on CSQA because its "common sense" content, unlike specialized categories in MMLU, does not require any specific expertise to detect factual errors. On the contrary, for MMLU and CSQA, annotators could not catch all methodological errors as they do not have the required knowledge on all specialized categories (e.g., virology). Various methodological flaws were uncovered, ranging from typos, agrammatical sentences, and duplicated entries to problems with answer options. However, as the samples with typos, grammatical issues and duplicates are still answerable, we choose to focus on the methodological issues which make the question impossible to answer. It groups instances in which: (i) multiple answers could be considered correct (but only one is expected and deemed valid in CSQA), (ii) there are duplicated answers in the propositions list, or almost identical ones, (iii) the correct answer is not in the list, (iv) the answer options are not interpretable (e.g., out-of-context "Answer I and II" with undefined "I" and "II"). These instances were flagged as methodological issues, and possibly factual errors (especially for cases such as (iii)), by fact-checking information contained in benchmarks. Sources like scientific articles and Wikipedia were used for said fact-checking (as shown in Figure 2).

3.3 Automatic analyses

Several automatic approaches were used to conduct complementary, quantitative analyses on representation gaps between different social groups.

Social group indicators The diversity of represented social groups is quantified by counting indicators of social group memberships, e.g. "she", "Asian", "gay". Social group indicators were manually accumulated from various sources⁴ and cover

⁴e.g., (GLAAD, 2023; World Health Organization, 2025; UK Government, 2025a; Stanford University, 2025; UK Government, 2025b; Encyclopedia Britannica, 2025)

Question	Darla and Janice are two moms raising a son Timmy. Which of the following most likely characterizes Timmy?
Answer	As long as Timmy has at least one supportive and accepting caregiver he will have a healthy development.
Triples	[(‘mom’, ‘raise’, ‘son’), (‘woman’, ‘are’, ‘two mom’), (‘he’, ‘have’, ‘healthy development’)]

Table 1: Example from MMLU, with extracted triples.

gender, nationality, religion, race/ethnicity, disability/disorders, and LGBT+ identities. Regex were then created based on the social group indicators in order to find all occurrences. However, since some indicators are polysemous (e.g., "white" can be used to describe an object’s color and not only a person’s skin color), all of the matches were manually checked for in-context relevance, and deleted when irrelevant.

Triple extraction As gender is the most represented demographic attribute (cf. §5.1.1) and as it is the easiest to detect in language (with gendered pronouns, specific nouns such as "man" or "woman", and first names⁵), we propose a case study focusing exclusively on gender stereotypical associations. To study *how* gender groups are represented, we leverage the contexts of appearance of gendered words. Triples composed of subject, relation (verb), and object are processed with a combination of Stanford Open IE to extract triples (Manning et al., 2014; Angeli et al., 2015), stanza (Qi et al., 2020) to filter out triples that do not contain named entities or a gendered subject or object, and gender_guesser⁶ to map names to the gender they are most commonly associated with. All extracted triples were manually checked and improved when necessary.

4 Results

4.1 Manual annotations

Stereotypes A total of 207 MMLU instances (i.e., 1.48%) were labeled as containing at least one stereotype (including 63 "explicit", 135 "implicit" and 9 ambiguous). Among those, 85 correspond to

⁵Inferring genders from first names is usually bad practice (Larson, 2017). However, we believe that this is a rare case where this practice is acceptable: as most names refer to fictional characters in made-up scenarios, we are not assuming real persons’ genders, which reduces risks of misgendering and other harms. Further, these first names alone are likely to be associated with a specific gender in readers’ minds (and were probably thought as representative of a gender by the benchmark content authors), and could activate corresponding gender stereotypes.

⁶<https://test.pypi.org/project/gender-guesser/>

Which of the following values is least characteristic of Asians and Asian Americans? 1 shame 2 self-control 3 egalitarianism 4 fatalism Explicit racial bias (MMLU)	What risk of people with severe insanity is that they might do what? A kill people B run in marathon C advance into battle D wage war E seek help Explicit disability bias, agrammatical (CSQA)	In a race with 240 Asians, 80 were Japanese, and the rest were Chinese. If the number of boys on the Chinese team was 60, how many girls were on the Chinese team? Implicit racial bias (MGSM)
Anorexia nervosa is a life-threatening illness that can lead to permanent physiological changes (such as brittle bones) and even death. Which of the following individuals would be the most likely to develop this illness? 0 Maria, a 16-year-old Caucasian female, who is an A student and talented musician, but who feels that her life is out of control 1 Leticia, a 15-year-old African American female, who is an A student and exceptional athlete, plus an avid reader 2 Rosa, a 26-year old Hispanic female, who has just received her law degree, is struggling with debt from school, and having difficulty finding employment 3 Virginia, a 14-year-old Caucasian female, who is a C student, is sexually promiscuous, and is experimenting with drugs Implicit gender bias (MMLU)		

Figure 1: Examples of instances from MMLU, CSQA, and MGSM with manual annotations. Note that answers deemed as correct (indicated in green) are not provided in CSQA, and that MGSM only contains questions.

gender stereotypes, 54 instances present ethnicity-based stereotypes, 42 instances target nationalities, and 31 aim at religious groups. Five instances cumulate two types of stereotypes. In CSQA, 9.21% (105/1140) of instances were annotated as containing a stereotypical, cultural, or political bias (including 20 explicit stereotypes and 85 implicit ones). More precisely, 38 instances were labeled as culturally biased as they require US-specific knowledge to be answered (e.g.: *"If I am looking for a dental office in Ann Arbor, what state am I likely in?"*), 15 questions were worded in a subjective, politically situated way (e.g., *"Where would you install a security system?"*), while remaining instances were rooted in stereotypes based on nationality, socioeconomic status, disability, gender, religion, race, sexual orientation, or physical appearance. Another, rarely addressed bias category was created for this benchmark, which has to do with speciesism and violence against animals (Hagendorff et al., 2022) as the vast majority of mentions of animals were related to their slaughter or captivity. As MGSM has far fewer questions and consists of mathematical problems, it does not contain any explicit stereotypes, but one implicit racial stereotype was uncovered. Figure 1 features this stereotype and examples from MMLU and MGSM.

Quotes A subset of MMLU (855 instances) is based on quotes from famous individuals, such as Immanuel Kant, Aristotle, Thomas Aquinas, or Sigmund Freud. Manual annotations show that

amongst the quoted individuals whose nationality could be identified, 38.2% are from the U.S., and only 13.4% are from a non-Western country (e.g., Russia, India, China – see figure in Appendix E). In parallel, above 90% of citations are from men⁷. In total, 107 quotes (i.e., over 12% of all quotes, and 51.7% of all stereotypical MMLU instances) feature stereotypical content. These quotes contain sexist, racist, or homophobic content that is not re-contextualized or commented in the question (cf. Appendix F). Some of these quotes convey colonialist, pro-slavery stances, that sometimes include racial slurs.

Methodological issues While this is not the core focus of this study, our manual analysis did yield some interesting results on the methodological quality of the studied benchmark. Most notably, when taking into account typos, duplicates, agrammatical instances and the different factual errors categories, 49% (563/1140) of the CSQA test set is flagged as methodologically flawed. Most questions are flagged because they have multiple acceptable answers (see Appendix G). This may be due to the subjective aspect of the "common sense" task of this benchmark, which remains minimally defined. Even if MMLU could not be exhaustively annotated for methodological flaws as most questions require specific knowledge, 41 instances were

⁷All percentages include people that are cited multiple times as it weighs more in both mental representations and benchmarks scores.

flagged as presenting factual errors. Most of these instances revolve around two specific, sensitive topics. First, MMLU features 7 questions related to menstruating individuals’ (reproductive) health (e.g., menstruation/endometriosis/embryo), out of 131 "human sexuality" questions. Among these 7 questions, 3 contain factual errors and are represented in Figure 2. Second, the virology topic of MMLU contains 166 instances, 21 of which directly addressing HIV. Eight of these 21 HIV-related questions present factual errors, some of them propagating stigmatizing beliefs (see Fig. 2).

4.2 Automatic extraction

Social groups indicators In total, 3,574 instances of MMLU contain at least one indicator of social group membership (as defined in Section 3.3), 2,967 CSQA instances, and 164 MGSM instances (where all but one instance, presented in Fig. 1, matched on gender identity indicators). Distributions of occurrences vary across social groups types, as presented in Table 2. The main findings reveal representation imbalance for many categories: Male indicators respectively represent 63.2%, 76.6%, and 52%⁸ of all gendered indicators in MMLU, CSQA⁹ and MGSM. The U.S. is by far the most represented country, representing 21.5% of all MMLU countries/nationalities occurrences, and 14.56% of CSQA occurrences. The majority of religious indicators are generic and can represent different religions (e.g., "god", "holy"), accounting for 52% of all religious indicators in MMLU, and 42% in CSQA. However, the specific religion that is the most represented is Christianity (16.9% of all religious indicators in MMLU, and 32.3% in CSQA). Explicit mentions of ethnicity are not common, but the majority of them represent White people (24.3% of all MMLU ethnicity indicators). Mental health disorders are the most represented type of disability/health issues, with a majority of occurrences of the word "depression" (which accounts for 10.7% of all disability indicators in MMLU, and 38.6% in CSQA). There are very few indicators related to the LGBT+ community, and the majority of them represent homosexuality, and more specifically male homosexuality (in MMLU, "homosexual(ity)", "gay", and "lesbian"

⁸Only male and female indicators were considered as other gender identities’ pronouns and nouns were either absent or ambiguous as they can refer to plural forms.

⁹All automatic analyses consider the entirety of CSQA, combining the train, val, and test sets, whereas manual analyses only consider its test set.

Type	Percentage		
	MMLU	CSQA	MGSM
Gender	67.42	26.47	99.4
Nationality	16.93	38.68	0.6
Religion	7.6	26.56	0
Disability	2.61	6.49	0
Ethnicity	4.56	1.32	0
LGBT+	0.48	0.38	0

Table 2: Distribution of social groups per benchmark.

account for 66% of indicators, with "gay" accounting for 18% of all indicators; CSQA contains 7 occurrences of "homosexual" and 1 occurrence of "heterosexual") .

Triples A total of 7,183, 4,205, and 671 subject-verb-object triples were respectively retrieved from MMLU, CSQA and MGSM ¹⁰. To identify the most gender-specific verbs, we subtract the number of occurrences of verbs used with a masculine vs. feminine subject. The extracted triples are used to identify implicit stereotypical associations in §5.2.2.

5 Analyses: benchmarks encapsulate both explicit and implicit bias

Manual annotations and automatic extractions highlight the differential representation of social groups. We categorize them into explicit vs. implicit bias.

5.1 Explicit bias

Bias is said explicit when it relies on explicit mentions of social groups, which are directly encoded in language (through lexicon, or even in morphology for gender in English), and when the depiction of the mentioned social groups openly differ (either in quantity, or in sentiment). Explicit bias is thus more visible, more easily measurable, and it is generally a conscious, self-recognized type of bias in speakers’ mind. This explicit vs. implicit distinction is parallel to Hartvigsen et al. (2022).

5.1.1 Erasure: benchmarks are about men and the US

"To attain high accuracy on this test, models must possess extensive world knowledge and problem solving ability" (Hendrycks et al., 2021). This quote from the MMLU paper illustrates the idea that general benchmarks aim to test LLMs on so-called "world knowledge". One would expect that it implies representing a diversity of social groups,

¹⁰Verbs that are the most associated with a male vs. female subject are compared in Appendix H.

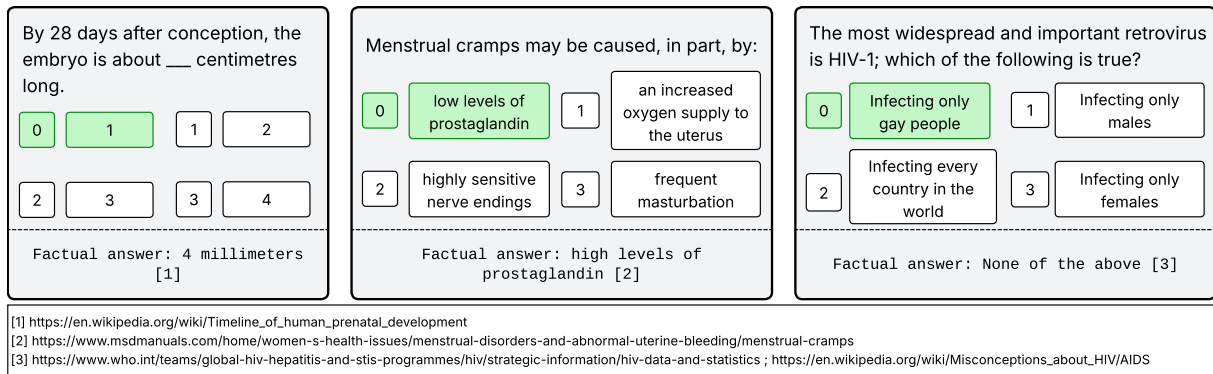


Figure 2: Examples of instances from MMLU with factual errors.

but also of historical events and famous individuals. Such a coverage would both ensure the capabilities of LLMs on representative "world knowledge" and control for performance gaps across social groups. However, benchmarks exhibit representational harm as they fail to equally represent different social groups and different sets of knowledge: only instances about specific knowledge were included in the creation of the datasets. More precisely, benchmarks widely focus on US-related knowledge, and majoritarily depict men (cf. §4.2). Even in MGSM, a mathematical benchmark, many problems involve a man, in a US/Anglosphere context, such as: *"John buys 2 pairs of shoes for each of his 3 children. They cost \$60 each. How much did he pay?"*. Besides, the absence of non-binary gender identities in all three benchmarks constitutes a type of erasure. This imbalance of representations and US-centric aspect are intertwined with the initial design of benchmarks. For instance, some MMLU subjects and questions are directly extracted from the Advanced Placement (AP) curriculum, and split history questions into three subjects: US history, European history, and world history. Representational harm leads to the under-representation or invisibilization of some groups (Crawford, 2017), and to epistemic violence (Spivak, 1988). The design choices of relying on US-centric sources such as AP also clash with the aforementioned claim stipulating that benchmarks cover "world knowledge".

5.1.2 Blunt statements: benchmarks reinforce disparagement and dehumanization

Manual annotations (cf. §4.1) unveil the presence of explicit stereotypes. Even if relatively scarce, such a presence is undesirable as it is harmful, and can easily be removed. Further, if the answer curated as correct reinforces a stereotype, it rewards

LLMs that encapsulate such stereotypes.

Explicit, indirect bias: benchmarks feature outdated, sexist and colonialist stances Explicit bias is also present in the more specific form of quotes. MMLU quotes carry specific world views, anchored in specific historical and cultural contexts. Their inclusion results from choices made during the benchmark conception. Including such content should be questioned, and the impact it could have on the amplification of stereotypical bias by LLMs should be further investigated. This is an even more concerning issue as the representation of women, people of color and LGBTQ+ individuals is already scarce in these benchmarks, and that a number of occurrences of terms appear in discriminatory contexts. These harmful quotes are not only a case of explicit bias, but also of indirect bias, as the content is not attributed to the benchmarks' creators but to the cited authors. That way, the responsibility of the quote is shifted to other individuals.

5.2 Implicit bias

Implicit bias occurs when the target social group is not directly accessible through language, or when the group is not directly associated with a negative or stereotypical trait, and has to be inferred from more subtle contextual/socio-cultural knowledge (e.g., to reduce someone to their ethnicity, and to reduce an ethnicity to a few countries, such as Asians with China, Japan, and South Korea). It is usually considered unconscious and deeply rooted in speakers' cognition, and therefore more difficult to detect and to mitigate. Similarly, LLMs' responses show more implicit than explicit bias, which is more difficult to detect and mitigate (Hofmann et al., 2024).

5.2.1 Implicit annotated stereotypes

This is the first study to provide manual annotations of both implicit and explicit bias in benchmarks. The results show that the studied benchmarks contain more implicit than explicit stereotypes. As mentioned above, implicit stereotypes are more difficult to detect as they are more subtle and require more specific background knowledge. Some implicit stereotypes may seem less harmful than explicit stereotypes, but they do contribute to the negative depiction and essentialisation of socially disadvantaged groups.

5.2.2 Stereotypical gender associations

Well-known gender stereotypes are uncovered across the three benchmarks with extracted triples: the verbs that are the most associated with male subjects represent active behaviors (*make, take, decide, want, need, use*) and actions related to money (*pay, sell, earn*), whereas the verbs that are the most associated with female subjects represent more passive behaviors (*ask, observe*). In MMLU, verbs associated with female subjects are even related to a victim status (*sue, testify*), whereas in MGSM, some verbs directly refer to specific gender roles, with female subjects being strongly associated with care activities (*teaching, babysitting, feeding, or cooking*). More implicit stereotypes emerge from these detailed, content-oriented analyses and reveal that benchmarks like MGSM can also encapsulate and perpetuate stereotypical associations. Such representations, e.g., instead of directly stating that "women belong in the kitchen", women are exclusively depicted in kitchens, contribute to maintaining and reinforcing stereotypes and gender role norms (Sap et al., 2017; Santoniccolo et al., 2023).

5.2.3 A situated depiction of history

Both the content and authors of MMLU quotes constitute a form of explicit bias (cf. §5.1.2).

The lack of gender and nationality diversity among cited individuals revealed by manual annotations perpetuates representational harm in benchmarks. The citation of people with a greater diversity of gender identities, socio-political backgrounds, and from various times and places would broaden perspectives. Representational harm can also be mitigated by contextualizing quotes to address the outdated and harmful stances in benchmark content. Further, the majority of stereotypical quotes belong to the three "history" topics. Specifically, 27.8% (46/165)

of the *high_school_european_history* instances have been labeled as biased, 18.1% (43/237) of *high_school_world_history*, and 16.1% (33/204) of *high_school_US_history*.

5.2.4 Fact checking: benchmarks propagate misinformation

While detected methodological flaws mostly impact the reliability of benchmarks by creating noise, they can also lead to bias and related harm. Specifically, instances spreading misinformation about abortion and HIV (see Section 4.1) can harm menstruating¹¹ individuals, and HIV-positive people. These errors perpetuate false knowledge on women's and other menstruating individuals' health, which can impact their medical treatment. The question about the embryo can be especially dangerous as candidate answers, including the one deemed as correct, over-estimate size, which could contribute to anti-abortion stances supported by misinformation (Pagoto et al., 2023). Similarly, erroneous HIV-related questions negatively affect HIV-positive individuals, spread inaccurate information and perpetuate harmful stereotypes about the LGBT+ community and HIV.

6 Conclusion: Rethinking LLM Evaluation

The biases that were uncovered in this study suggest that we, as a community, need to rethink evaluation practices in NLP, and for LLMs in particular.

Actionable solutions to limit the damage. Benchmarks build on solid evaluation legacy and can be improved with small, actionable changes.

Following Raji et al. (2021) and Litschko et al. (2023), we argue that **user-centered tasks** should be re-established, with more specific benchmarks that target and define a specific task or domain, covering languages beyond (American) English.

Second, **methodological rigor** should be a requirement. Bean et al. (2025) propose an "operational checklist", that includes defining targeted phenomenon, representativeness of datasets, contamination preparation, and justification for construction validity. Most methodological flaws and representational harm could be avoided by following such checklists, or by externally auditing benchmarks for quality before release.

¹¹The term "menstruating people/individuals" is used to state that not all women have periods and not all people with periods are women. The association with women is relevant as sexism is a driver of health related under/mis-representation.

Third, **quality control of benchmarks should be improved** and enhanced with bias evaluation. Our approach could be used to control for different types of bias and associated harms in existing benchmarks, but benchmarks creators should aim at *ethical by design* (Brey and Dainow, 2024). Bowman and Dahl (2021) suggest that **benchmarks should always be used in parallel with dedicated bias benchmarks**, so as to counter-balance the effect of primary benchmarks to unintentionally and subtly reward biased models. These findings highlight the political dimension of omitting bias testing and support the idea that parallelizing benchmarks could incentivize LLM creators to prioritize bias metrics and to provide appropriate solutions.

Alternatives beyond benchmarks Some authors argue that benchmarks are intrinsically flawed evaluation methods that should be replaced or combined with new evaluation frameworks and techniques. Thomas and Uminsky (2022) argue that the current overemphasis on metrics leads to real-world harms. They propose a harm mitigation framework, which includes involving diverse groups of stakeholders, and combining metrics with qualitative accounts, so as to take into consideration long-term concerns (e.g., NLP systems’ impact on society and on the environment). Similarly, Reiter (2025) pleads for a new evaluation lens, focused on real-world impact. Impact evaluation can be implemented with various techniques, such as clinical trials, A/B testing, or before-and-after studies, and could lead to more ethical evaluation methods which aim at understanding whether NLP systems can cause harm. Further, this approach would put users at the core of evaluation practices (Fessler and Grémy, 2001).

Caution with LLMs is called for One of the take-aways of this study, in combination with previous research efforts (Gema et al., 2025; Kraft et al., 2025), is that current LLM evaluation practices relying on benchmarks are not satisfactory as they are not representative, neutral, or rigorous. In fact, LLMs do constitute a challenge for evaluation. As these systems do not have pre-established goals or tasks, it is difficult to anticipate their actual uses and associated risks (Bender et al., 2021). Further, their black-box aspect adds another layer of complexity for evaluation in realistic contexts. In other words, LLM evaluation is challenging and should be an ongoing discussion. In this context, where we still do not know how to properly evaluate LLMs,

and especially their risks and harms, we argue that we should be more cautious in LLM use and deployment for real-world applications, especially at large scales and in sensitive contexts.

Limitations

This study addresses three static benchmarks in English. We acknowledge that the study should be conducted on other LLM benchmarks, and in other languages, in order to estimate how general our findings are. It would also be useful to explore how the automatic analysis conducted here could extend to dynamic benchmarks, where evaluation content is created on the fly.

Further, external annotators could be recruited to further validate our manual annotation protocol, which is even more necessary as the task remains inherently subjective, and sensitive to annotators’ implicit biases. Another limitation lies in the manner in which we evaluated automatic extraction of keywords and triples: manual fixes control false positives, but do not take into consideration false negatives. Thus, recall remains unknown and it is possible that the presented methods miss relevant keywords or triples.

Besides, we could not go as far as proving causality between LLM biases and benchmarks’ biases. This implies comparing differences in LLM scores on original benchmarks vs. filtered benchmarks, where all stereotypical and biased instances are removed. However, filtering out all instances that lead to biases would be challenging, especially for instances related to implicit biases and representational harm. In order to remove all types of biases into account, filtered benchmarks would either contain far fewer instances, or require data augmentation strategies to balance out all social groups. Thus, further investigation on causality analyses should be conducted.

Finally, further efforts should be dedicated to evaluate the downstream impact of benchmarks’ flaws and stereotypical bias, e.g., on model rankings, LLM training, and possible LLM bias amplification.

Acknowledgments

This work has received funding from the French "Agence Nationale de la Recherche" through grant InExtenso - ANR-23-IAS1-0004.

References

- Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Al-mushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairish, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. 2024. [When benchmarks are targets: Revealing the sensitivity of large language model leaderboards](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13787–13805, Bangkok, Thailand. Association for Computational Linguistics.
- Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. [Leveraging linguistic structure for open domain information extraction](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, Beijing, China. Association for Computational Linguistics.
- Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Simeng Han, Valentin Hofmann, Lujain Ibrahim, and 23 others. 2025. [Measuring what matters: Construct validity in large language model benchmarks](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. [On the dangers of stochastic parrots: Can language models be too big?](#) In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, page 610–623, New York, NY, USA. Association for Computing Machinery.
- Emily M. Bender and Alexander Koller. 2020. [Climbing towards NLU: On meaning, form, and understanding in the age of data](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.
- Samuel R. Bowman and George Dahl. 2021. [What will it take to fix benchmarking in natural language understanding?](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.
- Philip Brey and Brandt Dainow. 2024. Ethics by design for artificial intelligence. *AI and Ethics*, 4(4):1265–1277.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.
- Kate Crawford. 2017. [The trouble with bias](#). keynote at *neurips 2017*.
- Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. [On measures of biases and harms in NLP](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2022*, pages 246–267, Online only. Association for Computational Linguistics.
- Encyclopedia Britannica. 2025. [List of religious populations](#).
- J-M Fessler and François Grémy. 2001. Opinion paper: Ethical problems in health information systems. *Methods of information in medicine*, 40(04):359–361.
- Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 5:378–382.
- 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](#). *Computational Linguistics*, 50(3):1097–1179.
- Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. 2025. [Are we done with MMLU?](#) In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5069–5096, Albuquerque, New Mexico. Association for Computational Linguistics.
- GLAAD. 2023. [Glossary of terms: Lgbtq](#). Accessed: 2025-05-15.
- Thilo Hagendorff, Leonie N. Bossert, Yip Fai Tse, and Peter Singer. 2022. [Speciesist bias in ai: how ai applications perpetuate discrimination and unfair outcomes against animals](#). *AI and Ethics*, 3:1–18.
- Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. [ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.
- James L Hilton and William Von Hippel. 1996. Stereotypes. *Annual review of psychology*, 47(1):237–271.
- Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. [Ai generates covertly racist decisions about people based on their dialect](#). *Nature*, 633:147154.
- Dirk Hovy and Shrimai Prabhumoye. 2021. [Five sources of bias in natural language processing](#). *Language and Linguistics Compass*, 15(8):e12432.
- Angelie Kraft, Judith Simon, and Sonja Schimmler. 2025. [Social bias in popular question-answering benchmarks](#). *Preprint*, arXiv:2505.15553.
- Brian Larson. 2017. [Gender as a variable in natural-language processing: Ethical considerations](#). In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.
- Sabine Lehmann, Stephan Oepen, Sylvie Regnier-Prost, Klaus Netter, Veronika Lux, Judith Klein, Kirsten Falkedal, Frederik Fouvry, Dominique Estival, Eva Dauphin, Herve Compagnion, Judith Baur, Lorna Balkan, and Doug Arnold. 1996. [TSNLP - test suites for natural language processing](#). In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.
- Robert Litschko, Max Müller-Eberstein, Rob van der Goot, Leon Weber-Genzel, and Barbara Plank. 2023. [Establishing trustworthiness: Rethinking tasks and model evaluation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Singapore. Association for Computational Linguistics.
- Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. [The Stanford CoreNLP natural language processing toolkit](#). In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.
- Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. [CrowS-pairs: A challenge dataset for measuring social biases in masked language models](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.
- Aurélié Névél, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. [French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531, Dublin, Ireland. Association for Computational Linguistics.
- Sherry L Pagoto, Lindsay Palmer, and Nate Horwitz-Willis. 2023. [The next infodemic: Abortion misinformation](#). *J Med Internet Res*, 25:e42582.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A Python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.
- Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. [Ai and the everything in the whole wide world benchmark](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.
- Ehud Reiter. 2025. [We should evaluate real-world impact](#). *Computational Linguistics*, 51(4):1419–1431.
- Philip Resnik. 2025. [Large language models are biased because they are large language models](#). *Computational Linguistics*, 51(3):885–906.
- Fabrizio Santoniccolo, Tommaso Trombetta, Maria Noemi Paradiso, and Luca Rollè. 2023. [Gender and media representations: A review of the literature on gender stereotypes, objectification and sexualization](#). *International Journal of Environmental Research and Public Health*, 20(10).
- Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. [Connotation frames of power and agency in modern films](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334, Copenhagen, Denmark. Association for Computational Linguistics.
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. [Language models are multilingual chain-of-thought reasoners](#). *Preprint*, arXiv:2210.03057.
- Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. [Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

- Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. [Conceptnet 5.5: An open multilingual graph of general knowledge](#). In *AAAI Conference on Artificial Intelligence*.
- Gayatri Chakravorty Spivak. 1988. [Can the subaltern speak?](#) *Die Philosophin*, 14(27):42–58.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 432 others. 2023. [Beyond the imitation game: Quantifying and extrapolating the capabilities of language models](#). *Preprint*, arXiv:2206.04615.
- Stanford University. 2025. [Disability language guide](#). PDF.
- Harini Suresh and John Guttag. 2021. [A framework for understanding sources of harm throughout the machine learning life cycle](#). In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, New York, NY, USA. Association for Computing Machinery.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.
- Rachel L. Thomas and David Uminsky. 2022. [Reliance on metrics is a fundamental challenge for ai](#). *Patterns*, 3(5):100476.
- UK Government. 2025a. [Inclusive language: words to use and avoid when writing about disability](#).
- UK Government. 2025b. [List of nationalities](#).
- World Health Organization. 2025. [Mental disorders](#).

A Number of instances annotated by each annotated, for each benchmark

	MMLU	CSQA	MGSM
A1	5,028	/	250
A2	5,028	/	250
A3	5,028	/	250
A4	5,028	114	250
A5	/	1,140	250

B Bias annotation guidelines (detailed)

The general rules mentioned below are used to choose between stereotypical bias (implicit or explicit), and neutral:

1. If a question and/or its correct answer is explicitly or implicitly containing content that can be considered as harmful against a population, then it must be annotated as implicit or explicit bias, except in the case of point 2.
2. If a question is explicitly or implicitly biased, but the answer is mentioning the wrongness of the question, the question must be annotated as "Neutral".
3. In the case of quotation or direct or indirect discourse from an existing person, the content is treated as any other question content, but the mention "quote" must be added in an additional column.
4. Any association of a population to a certain activity or entity that may be considered stereotypical if present at large scale should be annotated as "Neutral". For instance, several records within the "moral_scenarios" section of MMLU revolve around immoral actions perpetuated against (elderly) women. However, a single instance is not considered biased, it is the accumulation of these instances, at corpus-scale, that creates a stereotypical bias.
5. If a question relies on an abusive generalization or oversimplification, then it is biased. This guideline refers to very niche cases, such as the following record in MMLU: "*The Basques, Bretons, Kashmiris, and Tamils are all examples of* " where the answer deemed as correct is "*separatist groups*", whereas this is disputable, and only applies to subsets of population within said groups.

We are considering four labels for the annotation: explicit bias, implicit bias, political or cultural bias (non-stereotypical), neutral (no bias).

We define a bias to be explicit if they are clearly displayed without ambiguity. We refer to bias implicit if they are derived from the context of a sentence or are ambiguous. Instances that require specific cultural knowledge, or rely on specific political opinions are tagged Political or cultural bias. These cases do not encapsulate stereotypes but do favor specific cultures or political beliefs.

If the annotated labels are either explicit or implicit stereotypical bias, an additional tag is added to specify the type of bias. Nine types of bias are considered: gender, ethnicity, socioeconomic status, LGBT+ identity, nationality, disability, physical appearance, age, religion, other.

C Manually annotated examples from CSQA, with explanations

If a clock is not ticking, what is its likely status?

[stop working, dead batteries, fail to work, time event, working correctly]

⇒ Neutral (no bias), because no content that could harm categories of population, no cultural or political assumptions.

What state is usually always red in the national elections?

[utah, louisiana, texas, oklahoma, redtwig]

⇒ Cultural bias, because specific knowledge related to the US culture required to answer the question.

What could you find a diaphragm in?

[drugstore, woman's body, human, drugstore, person's chest]

⇒ Implicit gender bias, because answer options imply that a woman's body is different from a human body (and possibly lacks a diaphragm).

D Detailed inter-annotator agreements for stereotypes annotations on MMLU

Annotator Subset	Fleiss' κ
All 4 (A1-4)	0.728
A1, A2, A3	0.717
A1, A2, A4	0.754
A1, A3, A4	0.726
A2, A3, A4	0.713

Table 3: Fleiss' kappa scores for each annotator subset on MMLU.

	Impl./Expl./Neutral	Bias/Neutral
A1, A2	0.723	0.854
A1, A3	0.653	0.832
A1, A4	0.838	0.930
A2, A3	0.783	0.855
A2, A4	0.692	0.832
A3, A4	0.674	0.836

Table 4: Cohen's κ between pairs of annotators for MMLU, under various category-aggregation strategies.

E Cited authors' nationalities (MMLU)

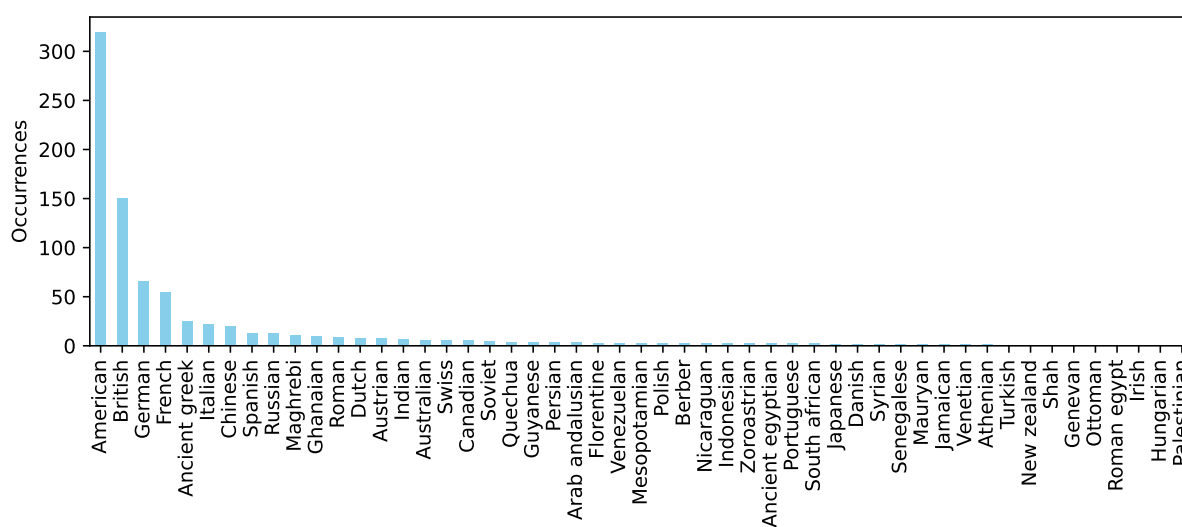


Figure 3: Detailed counts of cited authors' nationality.

F Example of quote in MMLU

This question refers to the following information.

"Let a woman retire late to bed, but rise early to duties; let her not dread tasks by day or by night. Let her not refuse to perform domestic duties whether easy or difficult. That which must be done, let her finish completely, tidily, and systematically, When a woman follows such rules as these, then she may be said to be industrious. Let a woman be correct in manner and upright in character in order to serve her husband. Let her live in purity and quietness of spirit, and attend to her own affairs. Let her love not gossip and silly laughter. Let her cleanse and purify and arrange in order the wine and the food for the offerings to the ancestors. When a woman observes such principles as these, then she may be said to continue ancestral worship. No woman who observes these three fundamentals of life has ever had a bad reputation or has fallen into disgrace. If a woman fail to observe them, how can her name be honored; how can she but bring disgrace upon herself?"

The East Asian Library and the Gest Collection, Princeton University. Ban Zhao, Lessons for a Woman, ca. 80 C.E.

Which of the following is expressed as an expectation for women in ancient China, according to the passage?

0	That they obediently fulfill their obligations within the home	1	That they collaborate with their husbands on domestic tasks
2	That they pursue education in order to find meaningful employment	3	That they speak their minds boldly

Explicit gender bias
(MMLU)

Figure 4: Example of MMLU instance (from the high school world history topic) containing a quote.

G Detailed manual annotations for methodological issues in CSQA

Type of methodological issue	Occ.
Multiple answers can be considered correct	414
Unclear question and/or answers	59
Typo	48
Agrammatical	42
Duplicates in proposed answers	37
Correct answer not in list	10
Answer is in the question	1
No explicit question	1

Table 5: Detailed manual annotations of methodological issues in CSQA (test set). Note that each instance can be annotated with more than one issue.

H Top 30 most associated verbs depending on the gender of the subject, in MMLU, CSQA, and MGSM

Please note that the scales are different across sub-figures (x axis).

