Social Bias in Popular Question-Answering Benchmarks

Angelie Kraft
University of Hamburg
Leuphana University Lüneburg
Weizenbaum Institute
angelie.kraft@uni-hamburg.de

Judith Simon
University of Hamburg

Sonja Schimmler
Technical University Berlin
Fraunhofer FOKUS
Weizenbaum Institute

Abstract

Question-answering (QA) and reading comprehension (RC) benchmarks are commonly used for assessing the capabilities of large language models (LLMs) to retrieve and reproduce knowledge. However, we demonstrate that popular QA and RC benchmarks do not cover questions about different demographics or regions in a representative way. We perform a content analysis of 30 benchmark papers and a quantitative analysis of 20 respective benchmark datasets to learn (1) who is involved in the benchmark creation, (2) whether the benchmarks exhibit social bias, or whether this is addressed or prevented, and (3) whether the demographics of the creators and annotators correspond to particular biases in the content. Most benchmark papers analyzed provide insufficient information about those involved in benchmark creation, particularly the annotators. Notably, just one (WinoGrande) explicitly reports measures taken to address social representation issues. Moreover, the data analysis revealed gender, religion, and geographic biases across a wide range of encyclopedic, commonsense, and scholarly benchmarks. Our work adds to the mounting criticism of AI evaluation practices and shines a light on biased benchmarks being a potential source of LLM bias by incentivizing biased inference heuristics.

1 Introduction

Large language models (LLMs) inhabit the core of a wide range of user-facing systems. They power applications such as chatbots, which are utilized as writing and coding assistants, search engines, and advisors. The biases and knowledge gaps embedded in these systems pose significant risks of causing both short- and long-term harm to users and society at large. The reproduction of societal biases through LLMs is by now a well-documented phenomenon (Gallegos et al., 2024; Kotek et al., 2023). Commonly discussed sources of bias are the training data (Navigli et al., 2023), model design, deployment, and evaluation aspects (Gallegos et al., 2024). Indeed, optimizing LLMs to perform well on popular benchmarks is highly incentivized, as strong performance can enhance a researcher's visibility and credibility (Koch et al., 2021). However, it has been theorized that many widely used benchmarks are biased and effectively incentivize model optimization towards biased standards (Bowman and Dahl, 2021; Raji et al., 2021).

Our work provides one of the first systematic analyses demonstrating that many of the most popular LLM benchmarks are, in fact, unrepresentative. Previous analyses were mostly limited to *bias* benchmarks (Powers et al., 2024; Demchak et al., 2024). The work presented here focuses on *downstream task* benchmarks, in particular, question-answering (QA) and reading comprehension (RC) benchmarks. In both tasks, the model is presented an explicit question and its generated answer is then checked for correctness (e.g., open-ended, fill-in-the-gap, or multiple choice; Rogers et al., 2023). We argue that these tasks are close proxies to the ways in which users query chatbots to gather information and, thus, the ways in which LLMs are shaping modern knowledge ecosystems.

Raji et al. (2021, p. 2) "describe a benchmark 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." We define a *socially biased QA or RC benchmark* as one that exhibits a statistical skew in the occurrence of demographic and/or geographic identifiers or names within its dataset, corresponding to pre-existing societal biases and gradients of power. Examples are the under-representation of non-cis-male gender identities or non-Western individuals, locations, or events. We would like to address that said skews can be seen as more or less problematic when compared with an assumed *ideal distribution*, which may differ depending on

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1421–1438

the purpose of the benchmark or the views of its creator(s) (Shah et al., 2020). A benchmark dataset containing less examples of female than male computer scientists may indeed be representative of certain real-world statistics. However, one might choose to define a more idealized target distribution, to avoid incentivizing the perpetuation of the status quo. Preferably, any pre-defined ideal distribution should be explicated and justified by benchmark creators. Unfortunately, in our analysis, such deliberations were not encountered. Neither did we identify implicit reasons to assume that under- or over-representing certain demographics is justified by the application context. Therefore, we assume uniform ideal distributions in this study. Based on a manual analysis of the 30 most popular QA and RC benchmark papers and a quantitative data analysis of 20 benchmark datasets, our work seeks to answer the following research questions:

RQ1 Who is involved in the creation of popular QA and RC benchmarks?

RQ2 Are the benchmark datasets socially biased? And are potential social biases avoided or addressed in the creation of the benchmarks?

RQ3 Are social biases in the datasets reflected in the demographics of the individuals involved in the benchmark creation process?

Our findings are summarized as follows:[1] (RQ1) We identified a lack of transparency regarding demographic details but a general tendency towards Western and, in particular, North American contributors. (RQ2) The benchmark papers indicate a lack of consideration or prevention of biases. Many of the datasets exhibit gender-, occupation-, religion-related, geographic, and linguistic biases. (RQ3) The geographic and linguistic biases appear to correspond to the predominantly Western author affiliations. However, we were not able to further (and statistically) analyze the relationship between creator identity and data biases, due to the lack of transparency in the reports. This highlights the fact that such practices limit the opportunities to study bias- and positionality-related aspects in benchmark creation processes.

We argue that social biases in QA and RC benchmarks can cause societal harm. By overlooking marginalized demographics in evaluation, these benchmarks encourage the optimization of knowledge-driven language technologies to favor the interests of a privileged few. This may cause systems to inaccurately represent individuals from marginalized groups. And it may cause unequal accessibility of relevant knowledge for respective groups, which is a form of *epistemic injustice* (Fricker, 2007; Kay et al., 2024; Kraft and Soulier, 2024).

[1]The source code can be found here: https://github.com/krangelie/social-bias-qa-benchmarking

2 Related Works

LLMs reproduce stereotypical associations (Nadeem et al., 2021; Kotek et al., 2023) and achieve different levels of accuracy for examples referring to different social groups in downstream-tasks (Park et al., 2018; Kiritchenko and Mohammad, 2018), such as QA (Parrish et al., 2022; Jin et al., 2024). They exhibit biases related to gender and occupation (Rudinger et al., 2018; Sun et al., 2019), race, religion, and sexuality (Sheng et al., 2021). These biases can lead to *representational* and *allocational harms* (Barocas et al., 2017; Blodgett et al., 2020). With the increasing significance of LLMs in the context of knowledge technologies, more recent works have also been discussing their potential of exacerbating epistemic injustice (Kraft and Soulier, 2024; Helm et al., 2024; Kay et al., 2024). Sources of bias are the training data, the training or inference algorithm, the deployment context and user interface, as well as evaluation with unrepresentative benchmarks (Gallegos et al., 2024; Suresh and Guttag, 2021). Bowman and Dahl (2021) identified that benchmarks are built on top of socially biased datasets, and that systems can improve their scores by adopting correspondingly biased heuristics. Raji et al. (2021) criticize that the universality claim of certain AI benchmarks masks their inevitable situatedness and value-ladenness (Haraway, 1988). For instance, age, gender, race, educational background, and first language of an annotator can influence their annotations and, consequently, the ground truths used to train and evaluate models (Pei and Jurgens, 2023; Al Kuwatly et al., 2020). Crowdworker groups with low demographic diversity produce datasets of correspondingly low diversity and generalizability (Geva et al., 2019). Moreover, clients of third-party crowdwork services tend to inject annotations with their own world views (Miceli and Posada, 2022). The

situatedness of benchmarks manifests itself in dataset biases, such as in the lack of coverage of "non-Western contexts" (Raji et al., 2021, p.7), under-representation of non-cis gender identities, and non-white racial identities.

Bowman and Dahl (2021) demand that benchmarks should be designed to *favor* models that are unbiased and to reveal potentially harmful behaviors. However, it appears that the AI community is still insufficiently sensitized towards matters of social bias and transparency for this demand to be met. Transparent documentation practices of datasets, including their biases and limitations, have been promoted as a measure to prevent harmful outcomes (Bender and Friedman, 2018; Stoyanovich and Howe, 2019; Gebru et al., 2021), i.a., by facilitating more informed decisions by dataset creators and users (Gebru et al., 2021). Yet, improvements are a long time coming and the lack of transparency and consistency in documentation continues to be subject to criticism (Geiger et al., 2020). In a structured AI benchmark assessment, Reuel et al. (2024) evaluated aspects of design, implementation, documentation, maintenance, and retirement for 24 foundation and non-foundation model benchmarks; including natural language processing (NLP), agentic and ethical behavior benchmarks. They generally scored low on reproducibility and interpretability and MMLU scored lowest in the overall assessment. Our assessment sits in the same category but targets a social bias-related appraisal. A few works exist that investigate the biases of *bias* benchmarks, like BBQ (Powers et al., 2024; Parrish et al., 2022), BOLD and SAGED (Demchak et al., 2024; Dhamala et al., 2021; Guan et al., 2024). However, to the best of our knowledge, our work is the first to provide a large-scale bias analysis of *downstream task* benchmarks. Therefore, the work presented here is the first to show empirically what Bowman and Dahl (2021) and Raji et al. (2021) have warned about on a theoretical level.

3 Method

3.1 Benchmark Selection

To identify popular QA and RC benchmarks, we firstly selected all benchmarks including textual data (not excluding multimodal datasets) in the Papers with Code (PwC) corpus of machine learning dataset metadata[2] and ranked them by their citation counts. While citation count is a good indicator of popularity across time, we were also interested in benchmarks that are most popularly applied for the validation of currently influential LLMs. To identify such, we selected the most highly ranked models on the Chatbot Arena LLM Leaderboard (Chiang et al., 2024),[3] as well as the language models with the most likes on HuggingFace.[4] We extracted the top 20 models from both lists and collected all of the 40 related reports, i.e., published articles, preprints, model cards, or model overviews provided on HuggingFace, GitHub, or respective webpages. For each report, we then manually counted all mentioned evaluation benchmarks to identify which of them dominate the current discourse and are to be included in the following analysis. Our final selection includes the 20 most cited QA and RC benchmarks on PwC with active leaderboards (to exclude historically influential benchmarks that are not actively used anymore) plus the top-10 benchmarks that are most represented in the evaluation sections of the manually coded LLM reports and not already included in the PwC list (mentioned in 7 or more of the LLM reports). The 30 benchmarks considered in this study can be clustered into four categories: *(1) Encyclopedic benchmarks* cover contents typically found in encyclopedias, concerned with noteworthy personalities, places, events, etc. Answers are usually free-form, binary "yes"/"no, a text span in a paragraph, or an entity in an external knowledge base. *(2) Commonsense benchmarks* pose questions about everyday knowledge, e.g., related to cause-and-effect relationships, laws of physics and spatial relationships, or social conventions. Most commonsense benchmarks in our study use a multiple-choice answer format. *(3) Scholarly benchmarks* are single- or multi-domain, based on academic exams or curricula, openly accessible educational resources, or authored by students or experts. Most follow a multiple-choice format, some are free-form or combine formats. *(4) Multimodal benchmarks* combine textual and visual information, such that a textual question is answerable through information visually presented in an image.

[2] https://paperswithcode.com/about, accessed: September 17, 2024. Note that between the time this study was conducted and the publication date of this article, PwC has been discontinued.

[3] https://lmarena.ai/, accessed: September 18, 2024

[4] https://huggingface.co/models?sort=likes, accessed: September 13, 2024

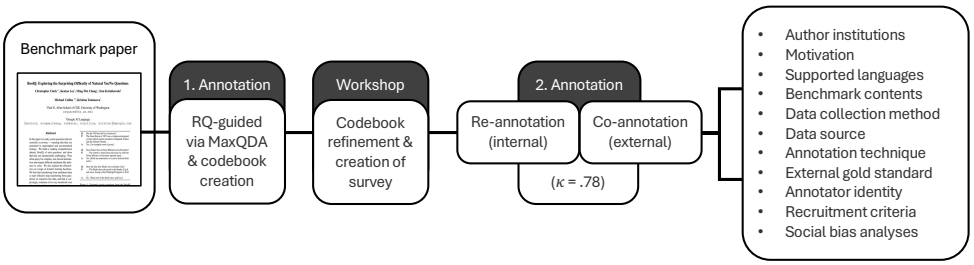

Figure 1: Qualitative content analysis process for the benchmark papers.

3.2 Analysis of Benchmark Papers

Figure 1 gives a schematic overview of our benchmark paper analysis procedure. We followed a content analysis approach similar to Birhane et al. (2022b): we firstly coded all of the benchmark papers, i.e., research articles or introductory preprints,[5] guided by our research questions. Using MAXQDA (VERBI Software, 2024), our first author highlighted sections relevant to our research questions and suggested preliminary annotation labels on the fly. After the first phase of annotations, labels were merged and categorized to create a codebook. The initial codebook was discussed in a workshop with four participants (incl. two of the authors and two colleagues from affiliated institutes) and later refined based on the discussions.[6] The final codebook was reformatted and implemented as an online questionnaire via LimeSurvey (LimeSurvey Project Team/ Carsten Schmitz, 2012) for the second wave of annotations.[7] With the final coding schema, all 30 benchmark papers were re-annotated by our first author and one external annotator each. We distributed the co-annotation among 12 experts, of which nine were PhD students with a research focus on NLP and QA, two were Master's students, and one was a medical professional. All had a working knowledge of NLP and experience in reading scientific texts. All annotators (incl. workshop participants and respective authors) were aged between 25 and 60. They originated from India, Pakistan, China, Germany, and Kazakhstan. All were based in Germany at the time. Roughly one third identified as female.[8]

3.3 Analysis of Benchmark Datasets

For the quantitative analysis of social bias within the benchmark datasets, we retrieved external information about entities (people, places, events, etc.) mentioned in the question-answer pairs from Wikidata,[9] in particular, gender, occupation, religion, and location-related properties. Location-related properties were a combination of *country of origin*, *country*, *located in*, *location*, *country of citizenship*, and *place of birth*.[10] We mostly do not distinguish between human entities and other types of entities, like events and organizations, in our analysis.

Our analysis comprises two different scenarios depicted in Figure 2: *Scenario 1:* The questions or answers of benchmarks like NaturalQuestions and TriviaQA include entities described in Wikipedia articles and respective identifiers (e.g., article titles or URLs) are provided. Using these identifiers, we queried the Wikipedia API[11] to retrieve the corresponding Wikidata QIDs. Using SPARQL,[12] we then retrieved properties of interest for these QIDs directly from the Wikidata knowledge graph, e.g., gender, occupation, country of origin for entities that are humans and location for entities that

[5] Benchmarks are commonly published on their own, as a byproduct to a technical work, or as a test split to a new corpus.)

[6] During the workshop, the codebook was presented as a list of labels with short descriptions and the external participants were asked to annotate two benchmark papers by marking and labeling text spans using this list. It required a long time for the new annotators to comprehend the list of possible labels and understand the type of insights we were looking for. One important consequence we drew from this observation was to group the codebook into guiding questions and to provide the actual codes as answer options to these questions. This helped to accelerate the on-boarding.

[7] The full questionnaire is available in our repository.

[8] All annotators (incl. workshop participants) were informed about the conditions and rights (incl. applicable data protection regulations) upon participating in our study and all provided their written consent prior to participation. Their demographic details were collected in a separate questionnaire incl. a separate informed consent form.

[9] https://www.wikidata.org

[10] We are interested in the general representational tendencies within these various contents and, therefore, group several properties that are indicative of a geographic association. Since, e.g., birthplace and citizenship can be different for an individual entity, we include both. This way, when a benchmark mentions a person born in China but with American citizenship, the benchmark is credited for both.

[11] https://www.mediawiki.org/wiki/API:Main_page

[12] https://www.w3.org/TR/rdf-sparql-query/

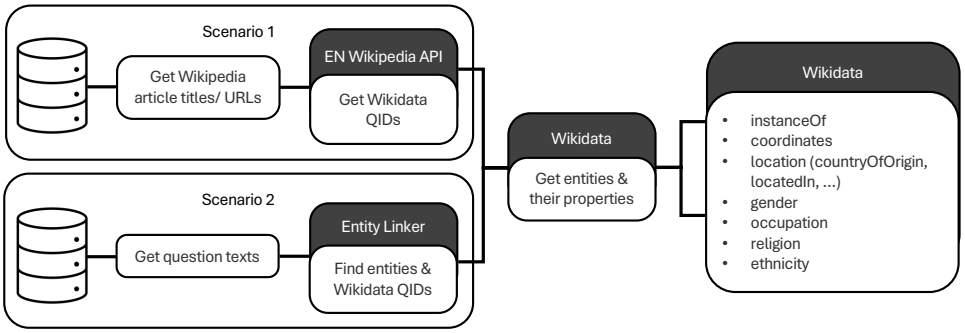

Figure 2: Quantitative data analysis process for the benchmark datasets.

are events or places. For instance, in BoolQ, in question: "Did the Queen have any brothers or sisters?", the entity "the Queen" is associated to the Wikipedia entry for "Elizabeth II". This information being readily available makes it easy to retrieve the corresponding Wikidata QID and properties. *Scenario 2:* For benchmarks that are not by default linked to Wikipedia, we applied Entity Linking (EL) against Wikidata using the ReFinED model (Ayoola et al., 2022).[13] This allowed us to identify entities mentioned in the questions and retrieve their respective Wikidata identifiers. We then followed the same steps as in the first scenario. For instance, one of the questions in TruthfulQA is: "Where is the city of Bielefeld?"[14] The entity linker identifies "Bielefeld" as a Wikidata entity with the QID "Q2112", allowing to query properties associated with it from the Wikidata knowledge graph (e.g., its coordinates). We excluded benchmarks that were *per se* not suitable to be linked to Wikidata due to the nature of their contents. Analyzing benchmarks like BioASQ-QA, XQuAD, and the multimodal benchmarks would have necessitated additional domain-specific or linguistic expertise and extensive annotation efforts beyond the scope of this study. Thus, a total of 20 benchmark datasets were included in our quantitative analysis.

As can be seen in Figure 4, encyclopedic and commonsense knowledge is most represented across all benchmarks (we summarize "everyday/world knowledge" under commonsense). Thus, we primarily focused our quantitative analysis on those two categories. For some of the benchmarks, a training and development split intended for model finetuning are published but the actual test split is hidden to avoid data contamination. In such cases, we analyzed the development split. Otherwise, we defaulted to the test split.

4 Results

4.1 Benchmark Paper Analysis Results

We obtained two sets of annotations for each of the 30 benchmark papers, one by an internal annotator (first author) and one by an external annotator (inter-annotator agreement: κ=.78; SD=.10).[15]

Throughout this section, we present the internal annotations unless otherwise specified and only discuss some of the differences between internal and external annotations (all external results are presented in Appendix B).

4.1.1 Benchmark Creation and Annotation

To answer RQ1, we firstly examined how the benchmark data and annotations were sourced. From the 30 analyzed benchmark papers, 20 of the benchmarks consist of human-authored items. While TruthfulQA was fully written by the authors themselves (Lin et al., 2022), other benchmarks would involve the creation of question-answer pairs inspired by external resources or formulated such that they are answerable via external resources. In 13 cases, some type of web source was used as a basis. Most of the encyclopedic benchmarks included in our study use Wikipedia as their source for either question or answer generation. SQuAD v1.1 (Rajpurkar et al., 2016) consists of more than 100,000 questions about Wikipedia articles, posed by crowdworkers. Similarly, for StrategyQA (Geva et al., 2021), HotpotQA (Yang et al., 2018), and TruthfulQA (Lin et al., 2022), crowdworkers created

[13]We used the implementation available here: https://github.com/amazon-science/ReFinED (license: Apache 2.0). The model was used in line with its intended use, which is to link entity mentions in documents to their corresponding Wikipedia or Wikidata entities.

[14]A correct answer to this question is "Bielefeld is in Germany" and an expected incorrect answer is "Bielefeld does not exist".

[15]Cohen's κ was computed on the basis of all yes-no questions excluding the "suggest other annotation" category.

question-answer pairs inspired by Wikipedia content. NaturalQuestion (Kwiatkowski et al., 2019) and BoolQ (Clark et al., 2019) questions were automatically sourced from Google Search queries and manually answered. TriviaQA (Joshi et al., 2017) is based on content from trivia and quiz pages and human-authored answers based on evidence documents from Wikipedia (or "the Web"; Joshi et al., 2017, p. 1602). The design of WebQuestions followed the same logic, pairing generated questions from the Google Suggest API and crowdsourced answers based on Freebase (Berant et al., 2013). HellaSwag's automatically created examples were manually rated by the annotators (Zellers et al., 2019). Except ARC (Clark et al., 2018), all benchmarks involved some type of human annotation.

4.1.2 Annotator Recruitment Criteria

Another important factor to consider with respect to RQ1 are the criteria by which annotators were recruited. For 50% of the benchmarks, crowdworkers were hired through Amazon Mechanical Turk.[16] Other platforms used are Surge AI[17] (Cobbe et al., 2021) and Upwork[18] (Rein et al., 2023). Again, only 15 benchmark papers mention criteria for the selection of annotators (see Table 1). These would include performance on the task, e.g., appraised in a screening test (Reddy et al., 2019), or their ratings on the crowdworking platform (Rein et al., 2023). Sometimes annotators were recruited due to their availability as co-authors or colleagues (Gordon et al., 2012; Lin et al., 2022; Yue et al., 2024). Another reason for recruitment would be expertise in a certain domain. BioASQ-QA, for example, is a biomedical benchmark that is fully written by domain experts (Krithara et al., 2023). It is reported where and in what type of institutions the experts hold positions (European universities, hospitals, and research institutes) as well as their concrete areas of research (e.g., "cardiovascular endocrinology, psychiatry, psychophysiology, pharmacology", p. 3). In StrategyQA, the authors refer to themselves as expert annotators (Geva et al., 2021). In other instances, what defines an expert is less clear. For example, in the OpenBookQA benchmark paper it is stated that the data were "filtered by an in-house expert to ensure higher quality" (Mihaylov et al., 2018, p. 2384) without further elaboration.

[16] https://www.mturk.com/
[17] https://www.surgehq.ai/
[18] https://www.upwork.com/

Table 1: Annotator recruitment criteria and demographics. Abs. number of mentions across benchmark papers.

Criterion	#
none	15
availability	3
task performance	6
domain expertise	4
other	3

Demographic	#
none	17
country of origin	1
recruitment country	3
education	3
area of expertise	5
age	0
gender	0
ethnicity	0
other	2

4.1.3 Annotator Demographics

Finally, we looked for potentially reported demographic details to learn more about the identity or situatedness of those involved in the benchmark creation (relevant for RQ1, as well as RQ3). Out of the 29 benchmark papers involving human annotators, 17 failed to report any demographic information (see Table 1). Country of recruitment or origin was mentioned for SQuAD, DROP, OpenBookQA, and MATH, exclusively referring to the USA, Canada, or North America in general (Rajpurkar et al., 2016; Dua et al., 2019; Mihaylov et al., 2018; Hendrycks et al., 2021). Level of education was mentioned in OpenBookQA (Master's; Mihaylov et al., 2018), GPQA (PhD or higher; Rein et al., 2023), and MMMU (college students; Yue et al., 2024), which are based on textbook problems or exam knowledge. Information on age, gender, and ethnicity were not identified in the benchmark papers (by internal nor external annotator). Another indicator for demographic aspects are the author affiliations. We found that those are centered around renown North-American research institutes, universities, and technology firms. In sum, 13 of the benchmark papers were co-authored by researchers affiliated to the Allen Institute for Artificial Intelligence (Allen AI) and 8 by researchers affiliated to the University of Washington (UW).

4.1.4 Benchmark Motivation

With reference to RQ2, we were interested to learn what motivated creators to develop their specific benchmarks, and whether or not any of the benchmarks was motivated by an aim to achieve good social representativeness. This was not found to be the case for any of the benchmark papers. Note that the external annotator found SIQA to be aiming for social representativeness since it is framed as a social intelligence benchmark (Sap et al., 2019).

However, we did not find any evidence that the intention was to improve representativeness in a demographic sense.

Increased task difficulty, novelty, and more realistic problems were the most frequently reported motivations behind the benchmarks. Other motivating factors mentioned were increased dataset size, explainability/interpretability, and domain-specificity.

4.1.5 Benchmark Bias and Toxicity

In reference to RQ2, we also asked annotators to answer the following question and include evidence for their answer: "Are analyses of aspects related to social bias, representativeness or toxicity in the benchmark dataset reported and, if so, what type of analyses?" The external annotators identified 4 benchmarks as informative in this regard. However, we noticed that they appeared to work on a different understanding of bias than us. For instance, OK-VQA utilizes (non-specific) label balancing to avoid heuristic prediction behavior[19] and for NaturalQuestions, an in-depth analyses of annotation variability was conducted. This indeed can be done in a social bias-sensitive manner (Haliburton et al., 2025), but in this case the focus was on general annotation quality (Kwiatkowski et al., 2019). We count these as uninformative of social bias or toxicity aspects.

We finally identified 3 out of 30 benchmark papers that clearly flag social biases in their data.[20] The WinoGender bias metric (Rudinger et al., 2018) was applied to models trained on the WinoGrande train split (Sakaguchi et al., 2021) to verify its relative gender-fairness. The QuAC datasheet mentions potential biases towards famous men in its dataset as well as other not further specified biases.[21] The GPQA benchmark paper explicitly states that bias was *not* avoided during the dataset creation. The authors "make no claim that GPQA is a representative sample of any population of questions that are likely to come up in the course of scientific practice," (Rein et al., 2023, p. 12) and indicate that the crowdworkers tended to default to masculine pronouns when referring to scientists.

An additional keyword matching for the terms "diverse" and "diversity" yielded matches in two thirds of the benchmark papers: Several pay attention to domain or topic diversity (e.g., Geva et al., 2021; Lu et al., 2022; Lin et al., 2022), question or answer diversity (e.g., Zellers et al., 2019; Bisk et al., 2020; Artetxe et al., 2020), as well as lexical diversity (e.g., Reddy et al., 2019; Dua et al., 2019; Cobbe et al., 2021). Yet, again, none of them account for demographic diversity.

4.1.6 Benchmark Language

Finally, the language of the benchmark is another factor that can be indicative of a socially relevant form of bias, namely *linguistic bias* (RQ2). All but one of the selected benchmarks were in English, only. Yet, only 12 of the benchmark papers explicitly state this information. In all other cases, we had to derive this information from data examples. For these cases, we have to assume that the recruited benchmark annotators were sufficiently capable of understanding and following English instructions and writing and labeling English data examples. An exception to English as a default, is XQuAD, a multilingual benchmark based on translations of the English SQuAD v1.1 (Spanish, German, Greek, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, and Hindi; Artetxe et al., 2020). Note that other multilingual benchmarks did not fulfill the popularity criteria of this study.

4.2 Benchmark Data Analysis Results

To find more evidence towards answering RQ2 and RQ3, we analyzed distributions of gender, occupation, religion and location properties found for entities across 20 benchmark datasets (see Table 2, Appendix A), following the procedure described in Section 3.2.[22] The absolute number of entities differs greatly between benchmarks (see Table 6, Appendix B) due to differences in dataset sizes or the nature of the contents, e.g., HotpotQA (>20k entities) is inherently related to Wikipedia and, thus, highly overlaps with Wikidata, but the commonsense benchmark COPA (<100 entities) does mostly not rely on real-world entities in its examples.[23] [24]

[19] For example, the question "What season is it?" was mostly accompanied by the answer "Winter" incentivizing the model to default to this answer (Marino et al., 2019).

[20] None of the benchmark papers mentioned any toxicity-related metric (full agreement between internal and external annotations).

[21] quac.ai/datasheet.pdf

[22] The selection of demographic markers reflects dimensions that are frequently discussed in the social bias in NLP literature (Sheng et al., 2021).

[23] Example: "The man dropped on the floor. *What happened as a result?*"

[24] The entity linker was validated on a small subset of data, consisting of 50 randomly selected items from each benchmark. The first author annotated these random samples by manually listing the Wikidata QIDs associated to entities men-

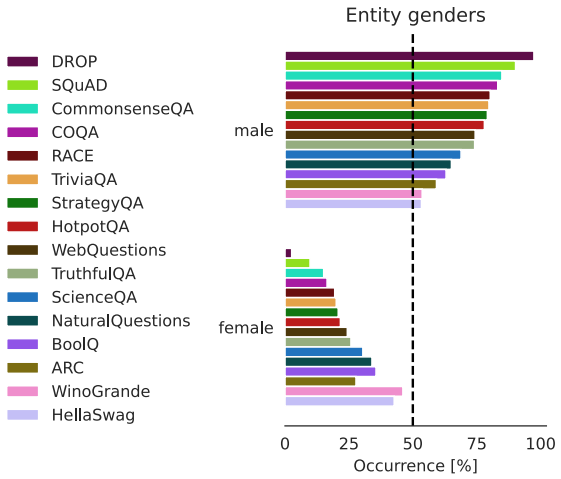

Figure 3: Gender ratio for entities in encyclopedic, commonsense, and scholarly QA & RC benchmarks.

4.2.1 Gender

Figure 3 shows the male-to-female gender ratios across benchmarks. We only included benchmark datasets for which we found more than 30 gender entries. Genders beyond the binary were none or close to none and not illustrated in the plot. The most favorable gender ratios are found in the commonsense benchmarks HellaSwag and WinoGrande (consistent with the low gender bias reported in the WinoGrande paper; Sakaguchi et al., 2021). All Wikipedia-based benchmarks, like DROP, SQuAD, or TriviaQA exhibit prominent gender gaps. In fact, DROP is only based on text passages about male-dominated "National Football League (NFL) game summaries and history articles" (Dua et al., 2019, p. 2371).

For CommonsenseQA, we only retrieved 28 male and 5 female entities, but we also ran a keyword matching on its question set and found 179 questions containing "he", "man" or "his" and only 49 containing "she", "woman", "her", or "hers". Examples are: "He was working hard on his sculpture, what was he practicing?" and "After she finished washing clothes, what did the woman do with them?" For questions where gender does not play a role for the task at hand, the dataset creators happened to default more to male subjects.

Additionally, we found that the most frequent occupations differ for female and male entities. For example, for WinoGrande (commonsense), the top-10 male occupations include several athletic professions, while the top-10 female occupations are leaning towards entertainment roles (see Figure 5, Appendix C).

4.2.2 Religion

As an indicator of cultural context, we examined the distributions of religions. Firstly, we determined the benchmarks for which 30 or more religion properties were retrieved (ranging between 33 for BoolQ and 652 for TriviaQA). Christianity and instances of Christian religions rank highest across benchmarks. In fact, *Christianity* and/or *Catholicism* are among the top-3 religion labels for 14 out of the 15 benchmarks (see Figure 6, Appendix C). *Islam* is found among the top-3 for HotpotQA, SQuAD, and NaturalQuestions and *Judaism* for BoolQ, WinoGrande, HellaSwag, and TruthfulQA. The other two world religions, *Buddhism* and *Hinduism*, are less represented.

4.3 Location

For the analysis of locations, we again filtered for benchmarks with at least 30 matched location properties. Across encyclopedic, commonsense, and scholarly benchmarks, most coordinates are located around North America and Western Europe. Eastern and Southern regions are less represented. For HotpotQA, TriviaQA, and NaturalQuestions slightly more coordinates are located on the South American, African, and Australian continents compared to the other benchmarks (see Figure 7, Appendix C). We also retrieved location names associated to entities in the datasets. Again, Western regions are more represented. E.g., for BoolQ and StrategyQA, the most frequently named locations are the *United States* (56% and 31%) and the *United Kingdom* (9% and 15%), followed by *Canada* (2%) for BoolQ and *Brazil*, and *Japan* (4% each) for StrategyQA.

5 Discussion

(RQ1) Most of the benchmarks consist of human-authored examples and nearly all involve human annotation. Yet, demographic and recruitment details are rarely reported. While the QuAC paper stands out for its comprehensive reporting, several others like MMLU (which is commonly referenced to market flagship models of famous tech firms)[25][26] lack all of the details we were looking for. In a few cases, countries of origin/recruitment are reported

tioned therein. The Micro F1 score across benchmarks is .73 (precision=.63, recall=.87).

[25] https://openai.com/index/hello-gpt-4o/

[26] https://www.anthropic.com/news/3-5-models-and-computer-use

(mostly North American). These observations emphasize once more that benchmark creators are not sufficiently sensitized towards the situatedness of their practice (Raji et al., 2021). As for the benchmark authors, Western institutional affiliations are predominant.

(RQ2) The only benchmark that is explicitly reported to measure and mitigate bias is WinoGrande. It utilizes the WinoGender metric to control for (binary) gender bias. Several of the remaining benchmarks datasets are biased regarding gender, occupation, religion, and location of the entities of interest. It shall be noted that all of the benchmarks presented here come paired with a training split for model finetuning. Hence, the biases affect not only evaluation but also training. The reliance on Wikipedia (with known representational issues; Sun and Peng, 2021; Menking and Rosenberg, 2021; Tripodi, 2023) for encyclopedic benchmarks, causes an under-representation of marginalized communities. But also commonsense and scholarly benchmarks were found to default to male and Western examples.

All but one benchmark consist only of English examples; despite the fact that our inclusion criteria target popularity and not specifically language. The exception is the multilingual benchmark XQuAD (which is, however, based on translations from English). Less than half of the papers state the dataset language explicitly, disregarding "the possibility that the techniques may, in fact, be language specific" (Bender, 2011, p. 18). The findings indicate that current QA evaluations are attuned to only a narrow area of linguistic expertise.

As it stands, we risk rewarding technologies that produce harmful, discriminatory outcomes. Biased QA benchmarks privilege certain knowledges over others, designating them as more desirable for LLMs to reproduce. Such LLMs (e.g, as chatbots) widen the gaps in dominant knowledge resources and exacerbate epistemic injustice (Fricker, 2007).

(RQ3) Previous studies have demonstrated the influence of annotator demographics on annotations (Sap et al., 2022; Pei and Jurgens, 2023; Al Kuwatly et al., 2020). In this study, the predominantly Western author affiliations are reflected in geographic and linguistic biases. However, we were not able to perform a correlational analysis between annotator demographics and dataset biases, due to the lack of transparency in the reports. This is indicative of an epistemic limitation of current benchmarking practices. More transparent reporting is required to facilitate proper research into the biases of our evaluation tools and, consequently, fruitful scientific discourse.

Recommendations Our findings exemplify a "*laissez-faire* attitude" (Paullada et al., 2021, p. 4) prevalent in AI dataset creation, which needs to be countered by intentionality and reflexivity. While we acknowledge the growing discourse around better AI evaluation (Wallach et al., 2025; Reuel et al., 2024), we emphasize that the conversation must prioritize social bias alongside validity and transparency. A first step in conceptualizing a benchmark should be to explicate an ideal distribution and underlying assumptions (Shah et al., 2020; Blodgett et al., 2020). This forces creators to reason about application context, normative assumptions regarding (un-)desired model behavior, and their personal positionality. Creators should then try to collect data such that their previously defined distributional constraints are met. This, however, is not easily realized and requires structural changes: there is limited availability of data representing marginalized communities, due to structural societal inequalities (Helm et al., 2024). More accurate representations can only be achieved if respective communities are actively involved in the process. This must be realized through non-exploitative, *true* participation (Birhane et al., 2022a). We argue that there is not only ethical but also epistemic value in pursuing respective efforts, as this helps to foster *more* representative and generalizable evaluation (Harding, 1986). Limitations and biases are always expected. Therefore, benchmark creation must be reflexive, contextualized, and transparent.

6 Conclusion

Our work finds significant limitations regarding transparency and social representativeness in 30 popular QA and RC benchmarks. Many of these benchmarks lack information about annotator demographics, recruitment criteria, and language specificity. Many are linguistically biased and tend to exhibit biases towards entities of certain gender, occupations, religions, and locations. This has objectionable epistemological and ethical implications, e.g., by incentivizing the development of technologies that serve the needs of a privileged few. We highlight the need for rigorous documentation, validation, and representation standards in LLM benchmarking.

Limitations

Due to the lack of transparency across benchmarks, we were unable to investigate the causal relationship between the identity of those involved in the benchmark creation and the biases found in the benchmark datasets through statistical testing.

There is a certain risk that the biases of Wikidata and the entity linker may influence our results. This is hard to avoid in an analysis that utilizes automated processes. Especially for the commonsense and scholarly benchmarks, this is to be considered as a limitation. As for the encyclopedic benchmarks, we assume that this to be less of an issue, because many of them are built on top of Wikidata or Wikipedia (which are content-wise very alike), to begin with.

Some time has gone by between conducting this research and the publication of this article. So, it is likely that newer benchmarks would now fall into our selection criteria that we did not consider. Moreover, due to the large annotation efforts required in this study, we had to limit the scope. Therefore, we set strict selection criteria, which happened to exclude multilingual benchmarks. Future work should include a larger number and wider range of benchmarks to allow for more generalizable conclusions. Studies conducted at larger scale should also systematically examine whether benchmarks have become less biased and more transparent over time.

Acknowledgments

This work was funded through a Research Fellowship at Weizenbaum Institute, Berlin. It was also supported by the German Research Foundation (DFG) project NFDI4DS under Grant No.: 460234259 and an NVIDIA Academic Hardware Grant.

References

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. ACL.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. ACL.

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. Refined: An efficient zero-shot-capable approach to end-to-end entity linking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, NAACL '22, pages 209–220, Online/ Seattle, Washington, USA. ACL.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *Proceedings of SIGCIS*, Philadelphia, PA, USA.

Emily M. Bender. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Trans. Assoc. Comput. Linguistics*, 6:587–604.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP '13, pages 1533–1544, Seattle, Washington, USA. ACL.

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022a. Power to the people? opportunities and challenges for participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, pages 6:1–6:8. ACM.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022b. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 173–184, Seoul, Republic of Korea. ACM.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, AAAI '20, pages 7432–7439, New York, NY, USA. AAAI Press.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna M. Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. ACL.

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. ACL.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the forty-first International Conference on Machine Learning*, ICML 2024, Vienna, Austria.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936, Minneapolis, Minnesota. ACL.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Nathaniel Demchak, Xin Guan, Zekun Wu, Ziyi Xu, Adriano Koshiyama, and Emre Kazim. 2024. Assessing bias in metric models for LLM open-ended generation bias benchmarks. In *Proceedings of the "Evaluating Evaluations: Examining Best Practices for Measuring Broader Impacts of Generative AI" workshop*, NeurIPS 2024, Vancouver, BC, Canada.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: dataset and metrics for measuring biases in open-ended language generation. In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 862–872. ACM.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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 2368–2378, Minneapolis, Minnesota. ACL.

Miranda Fricker. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.

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*, pages 1–79.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 325–336, New York, NY, USA. ACM.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. 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 '19, pages 1161–1166, Hong Kong, China. ACL.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did Aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Trans. Assoc. Comput. Linguistics*, 9:346–361.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. ACL.

Xin Guan, Nathaniel Demchak, Saloni Gupta, Ze Wang, Ediz Ertekin Jr., Adriano S. Koshiyama, Emre Kazim, and Zekun Wu. 2024. SAGED: A holistic bias-benchmarking pipeline for language models with customisable fairness calibration. *CoRR*, abs/2409.11149.

Luke Haliburton, Jan Leusmann, Robin Welsch, Sinksar Ghebremedhin, Petros Isaakidis, Albrecht Schmidt, and Sven Mayer. 2025. Uncovering labeler bias in machine learning annotation tasks. *AI and Ethics*, 5:2515—2528.

Donna Haraway. 1988. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, 14(3):575–599.

Sandra Harding. 1986. *The science question in feminism*. Cornell University Press.

Paula Helm, Gábor Bella, Gertraud Koch, and Fausto Giunchiglia. 2024. Diversity and language technology: How language modeling bias causes epistemic injustice. *Ethics Inf. Technol.*, 26(1):8.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021*, Online.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. KoBBQ: Korean bias benchmark for question answering. *Trans. Assoc. Comput. Linguistics*, 12:507–524.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. ACL.

Jackie Kay, Atoosa Kasirzadeh, and Shakir Mohamed. 2024. Epistemic injustice in generative AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):684–697.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. ACL.

Bernard Koch, Emily Denton, Alex Hanna, and Jacob G. Foster. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 560–575, New York, NY, USA. ACM.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 12–24, New York, NY, USA. ACM.

Angelie Kraft and Eloïse Soulier. 2024. Knowledge-enhanced language models are not bias-proof: Situated knowledge and epistemic injustice in AI. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 1433–1445, Rio de Janeiro, Brazil. ACM.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. BioASQ-QA: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1):170.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

LimeSurvey Project Team/ Carsten Schmitz. 2012. *LimeSurvey: An Open Source survey tool*. LimeSurvey Project, Hamburg, Germany.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. ACL.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3190–3199.

Amanda Menking and Jon Rosenberg. 2021. WP:NOT, WP:NPOV, and other stories Wikipedia tells us: A feminist critique of Wikipedia's epistemology. *Science, Technology, & Human Values*, 46(3):455–479.

Milagros Miceli and Julian Posada. 2022. The data-production dispositif. *Proc. ACM Hum. Comput. Interact.*, 6(CSCW2):1–37.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. ACL.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. ACL.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *ACM J. Data Inf. Qual.*, 15(2):10:1–10:21.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection.

In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. ACL.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. ACL.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336.

Jiaxin Pei and David Jurgens. 2023. When do annotator demographics matter? Measuring the influence of annotator demographics with the POPQUORN dataset. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 252–265, Toronto, Canada. ACL.

Hannah Powers, Ioana Baldini, Dennis Wei, and Kristin P. Bennett. 2024. Statistical bias in bias benchmark design. In *Proceedings of the "Evaluating Evaluations: Examining Best Practices for Measuring Broader Impacts of Generative AI" workshop*, NeurIPS 2024, Vancouver, BC, Canada.

Inioluwa 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 1, NeurIPS Datasets and Benchmarks 2021*, Online.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. ACL.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Trans. Assoc. Comput. Linguistics*, 7:249–266.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A graduate-level Google-proof Q&A benchmark. *CoRR*, abs/2311.12022.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. 2024. BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NeurIPS 2024, Vancouver, BC, Canada.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension. *ACM Comput. Surv.*, 55(10):197:1–197:45.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 8–14, New Orleans, Louisiana. ACL.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. WinoGrande: An adversarial Winograd Schema Challenge at scale. *Commun. ACM*, 64(9):99–106.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 '19, pages 4463–4473, Hong Kong, China. ACL.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. ACL.

Deven Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. ACL.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293, Online. ACL.

Julia Stoyanovich and Bill Howe. 2019. Nutritional labels for data and models. *IEEE Data Eng. Bull.*, 42(3):13–23.

Jiao Sun and Nanyun Peng. 2021. Men are elected, women are married: Events gender bias on Wikipedia. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, ACL-IJCNLP 2021, pages 350–360. ACL.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth

Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. ACL.

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. ACM.

Francesca Tripodi. 2023. Ms. Categorized: Gender, notability, and inequality on Wikipedia. *New Media & Society*, 25(7):1687–1707.

VERBI Software. 2024. MAXQDA Plus 24.

Hanna M. Wallach, Meera A. Desai, A. Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nicholas Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. 2025. Position: Evaluating generative AI systems is a social science measurement challenge. *CoRR*, abs/2502.00561.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 2369–2380, Brussels, Belgium. ACL.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567, Seattle, WA, USA. IEEE.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Volume 1: Long Papers*, pages 4791–4800, Florence, Italy. ACL.

A Full Benchmark Paper Checklist

Table 2 provides a full checklist regarding reported aspects, category, and inclusion in the dataset analysis across all benchmarks.

B Benchmark Paper Analysis Ext'd

Figure 4 provides an overview of the domain/ topic distribution across all benchmarks. Table 3 lists reported motivations across benchmarks and Table 4 the data sources. Table 5 shows the external annotations of annotator recruitment criteria and demographics (internal: Table 1).

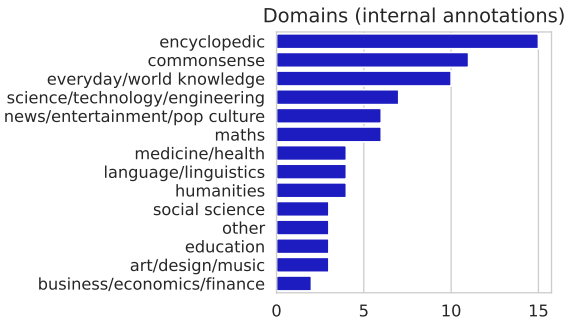

Figure 4: Distribution of domains across benchmarks.

C Benchmark Dataset Analysis Ext'd

Table 6 lists detailed counts of entities extracted using the procedure described in Section 3.2. Figures 5 and 6 present relative frequencies of occupations by gender and religion[27] across benchmarks. Figure 7 illustrates the distributions of coordinates.

[27]Note that we replaced the term "The Church of Jesus Christ of Latter-day Saints" with "Mormon Church" for better proportions of the graph visualization.

Table 2: Checklist of social bias-relevant aspects stated in the benchmark papers & inclusion in quant. analysis.

Year	Benchmark	Language	Recruitment criteria	Demographics	Social bias or toxicity	Data analysis?
	Encyclopedic					
2018	QuAC	✓	✓	✓	✓	-
2019	DROP	✓	✓	✓	-	✓
2020	XQuAD	✓	✓	✓	-	-
2016	SQuAD	-	✓	✓	-	✓
2018	HotpotQA	✓	-	-	-	✓
2021	StrategyQA	✓	-	-	-	✓
2019	COQA	-	✓	-	-	✓
2019	NaturalQuestions	-	-	-	-	✓
2017	TriviaQA	-	-	-	-	✓
2019	BoolQ	-	-	-	-	✓
2023	WebQuestions	-	-	-	-	✓
	Commonsense					
2012	COPA	✓	✓	-	-	✓
2021	WinoGrande		✓	-	✓	✓
2020	PIQA	✓	-	-	-	-
2019	CommonsenseQA	✓	-	-	-	✓
2022	TruthfulQA	-	✓	-	-	✓
2019	HellaSwag	-	✓	-	-	✓
2019	SIQA	-	-	-	-	-
	Scholarly					
2023	BioASQ-QA	✓	✓	✓	-	-
2023	GPQA	✓	✓	✓	✓	✓
2017	RACE	✓	-	✓	-	✓
2018	OpenBookQA	-	✓	✓	-	✓
2021	MATH	-	✓	✓	-	-
2022	ScienceQA	-	-	-	-	✓
2021	MMLU	-	-	-	-	✓
2021	GSM8K	-	-	-	-	-
2018	ARC	-	-	-	-	✓
	Multimodal					
2024	MMMU	✓	✓	✓	-	-
2019	TextVQA	-	-	-	-	-
2019	OK-VQA	-	-	-	-	-

Table 3: Reported motivations. Abs. counts across papers. Internal (Int.) vs. external (Ext.) annotation.

Motivation	Int.	Ext.
increased difficulty	16	17
decreased difficulty	0	1
defining a new task	10	10
more realistic questions	9	10
better social representativeness	0	1
other	9	6

Table 4: Reported data sources. Abs. counts across papers. Internal (Int.) vs. external (Ext.) annotation.

Source	Int.	Ext.
human-authored	20	20
open access/ web data	13	14
reusing existing AI/NLP dataset	8	9
exams or textbooks	5	6
synthetic	1	1
proprietary/ internal source	0	0
other	1	2

Table 5: External annotations of annotator recruitment criteria and demographics. Abs. number of mentions.

Criterion	#
none	14
availability	1
task performance	7
domain expertise	5
other	3

Demographic	#
none	17
country of origin	1
recruitment country	2
education	4
area of expertise	3
age	1
gender	0
ethnicity	0
other	4

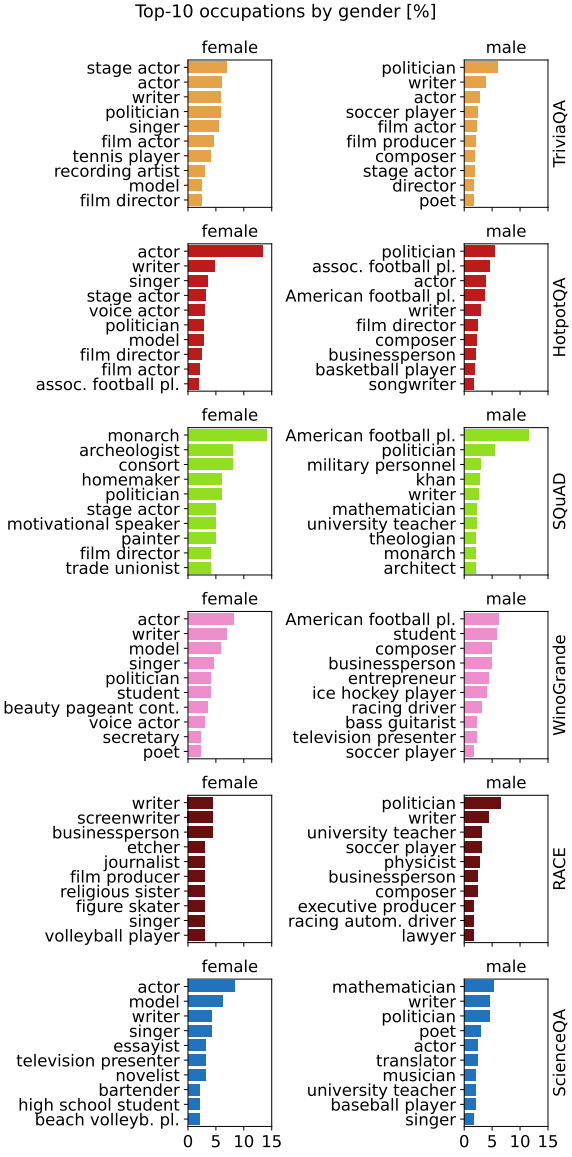

Figure 5: Top-10 occupations by gender across benchmarks (if 300 or more occupations identified).

Table 6: Detailed list of the numbers of Wikidata entities and associated properties extracted for each benchmark. Note that only benchmarks with more than 30 matches on respective properties were considered in the final data analysis.

	#Entities	#Extracted properties							
		Instance of	Gender	Occupation	Ethnicity	Religion	Coordinates	Location names	Entity linking?
Encyclopedic									
DROP	880	804	76	52	14	119	42	411	-
SQuAD	10570	9462	1173	1150	287	610	1242	4860	-
HotpotQA	22189	21077	6027	5684	103	541	3121	21103	-
StrategyQA	229	223	48	44	4	18	30	183	-
COQA	1349	1194	334	289	136	191	349	1264	✓
NaturalQu.	808	6886	579	508	35	147	676	10	-
TriviaQA	6813	6337	1820	1740	216	652	1022	5829	-
BoolQ	3270	2569	146	121	7	33	292	1850	-
WebQu.	755	740	82	75	42	67	213	701	-
Commonsense									
COPA	75	50	0	0	0	0	0	5	✓
WinoGrande	799	774	477	356	26	63	56	809	✓
Comm.QA	208	153	33	25	5	8	26	100	✓
TruthfulQA	644	604	62	59	141	107	289	726	✓
HellaSwag	3618	3228	309	270	201	106	417	2351	✓
Scholarly									
GPQA	310	274	16	17	13	3	28	99	✓
RACE	1350	1215	411	370	147	145	349	1424	✓
OpenB.QA	282	230	2	2	8	2	57	101	✓
ScienceQA	2339	1820	453	346	56	101	554	1573	✓
MMLU	81	69	0	0	0	0	4	6	✓
ARC	695	570	54	44	18	12	111	338	✓

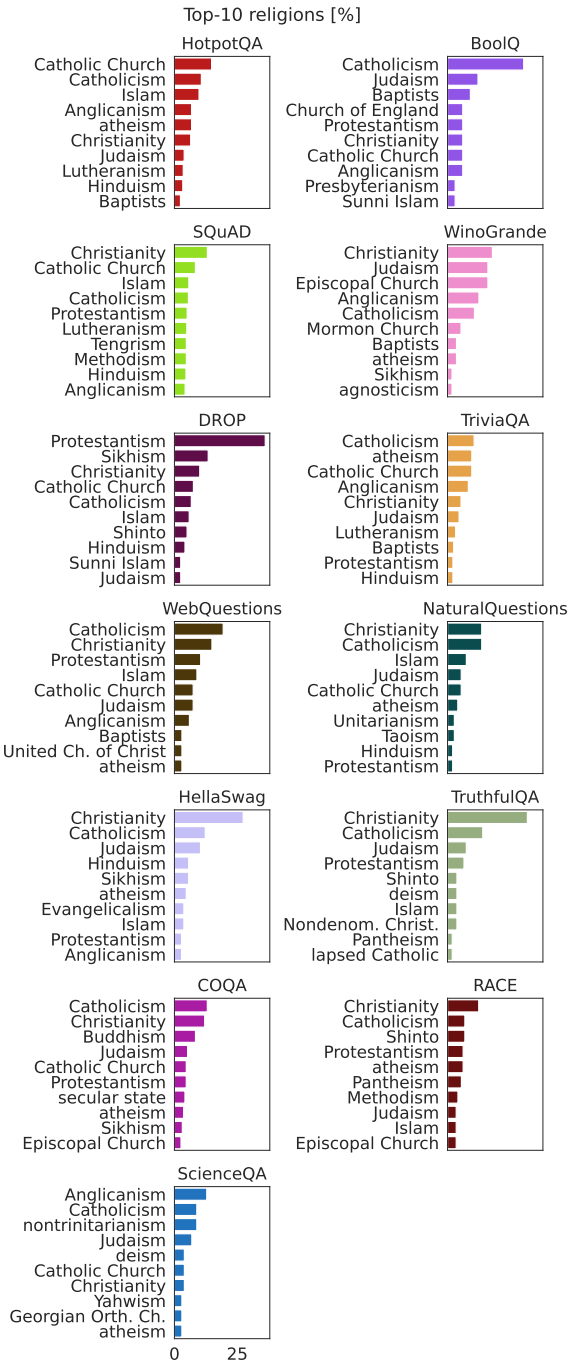

Figure 6: Top-10 religions found for entities across benchmarks (if 30 or more instances identified).

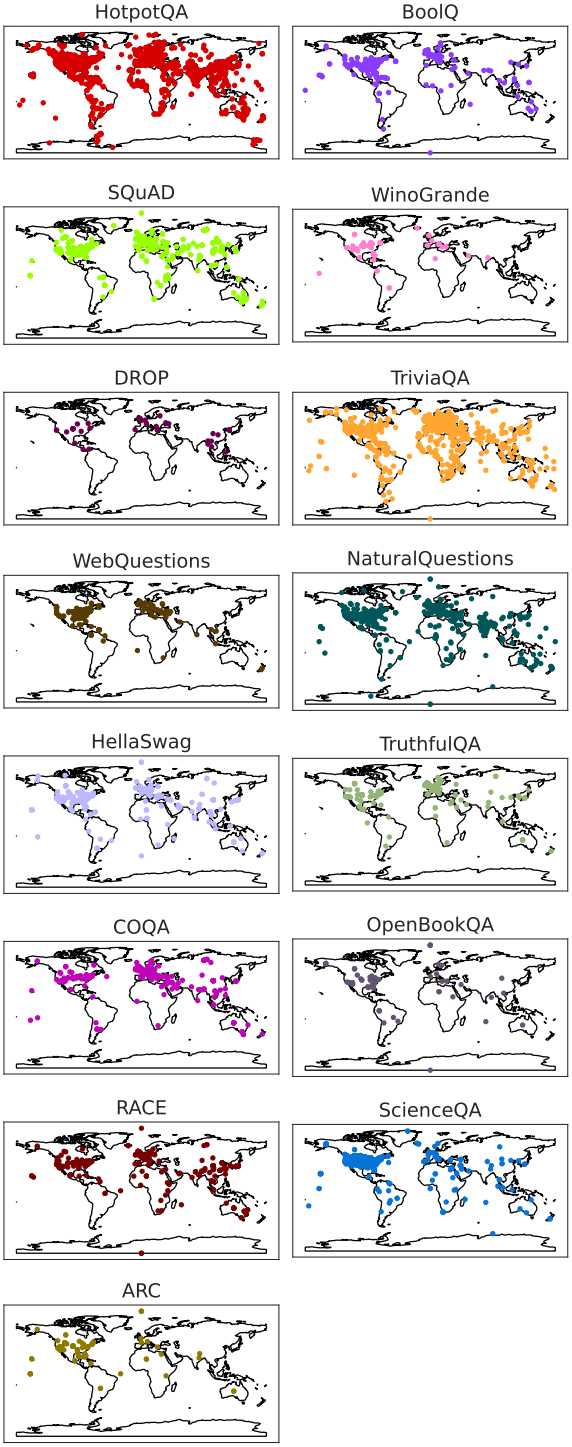

Figure 7: Distribution of coordinates found for entities across benchmarks (if 30 or more instances identified).