Annotation Questionnaire for
"QA Benchmarks are Socially and Geographically Biased"
Authors: Angelie Kraft, Judith Simon, & Sonja Schimmler

I. Informed Consent

Privacy & data protection:
This data/annotation collection is conducted in compliance with [applicable data protection laws] and respects your rights, e.g. to withdrawal at any time throughout or after the annotation/data collection process. This study is conducted by [ANONYMIZED] - referred to as the investigator in this study. A table of annotator names and e-mail addresses linked to their survey access links is stored on the [ANONYMIZED] server and is only accessible to the investigator. This table is used to keep track of the recruitment and annotation progress as well as to potentially reach out to individual annotators if there are any things to be clarified about the answers provided. This deliberation is a necessary part of the annotation process and is meant to ensure the quality of the research results. However, no personal identifiable data is collected in this online form and your responses collected here will be stored separately from any type of data that would allow for reidentification. The results of this study will be published in aggregated and anonymized form.

Contact:
For data protection-related questions, you may get in touch with the data protection officer at [ANONYMIZED]

Consent:
By proceeding with this questionnaire, you agree that you have read all information regarding procedure and legal aspects of this data collection and that you consent to participation. In case you would not like to continue and/or wish for your information to be deleted from the server, please contact us via e-mail.

II. Annotation Instructions

General instructions:
You should have received a number of benchmark reports as PDF by the investigator prior to accessing this page. In the following you will see a list of questions regarding different aspects of a report. Please fill out one questionnaire per report.

This questionnaire is meant as an interactive way of collecting annotations for the reports. Some questions will ask for an open-ended text input and some will ask you to select from a number of response options. Each of these options is accompanied by a text field. Please use this text field to copy-paste words or sentences that contain your answer or hint to it (see screenshot below). This will help us analyze and trace your reasoning. Furthermore, we highly encourage you to suggest new answer categories

if you think they are missing from the provided list. Don't hesitate to add as many as you deem appropriate!

Options to pause and resume this questionnaire are given in the top right corner if this screen. Please also note that some benchmark papers might mention the creation of the benchmark dataset as a byproduct rather than its main contribution. In these cases, you may focus on the sections that address the benchmark creation itself. The benchmarks are the main point of interest in our research, not so much the algorithms or systems that are evaluated with them.

Annotation task instruction:

Please read the following questions carefully and try to think of the answer that is provided in the report. In some cases, you will be asked for an open-ended answer. In other cases, you will be given a list of answer options. In this case, please have a thorough look at the answer options and
A) use one or more of the available answer options if applicable
B) suggest a better answer option
C) combine pre-defined and own answer options
D) leave a comment in case you feel that neither A, B nor C are applicable or if you have difficulties understanding the question
Please copy-paste the sentence/word sequence that provides evidence for your answer, in case you went with option A, B or C.

III. Questions and answer options (codes)

(1) Please briefly describe what task this benchmark was designed for. [open-ended]
(2) What motivated the creation of the benchmark?
- ☐ increased difficulty compared to existing benchmarks
- ☐ increased difficulty compared to existing benchmarks
- ☐ more realistic questions compared to existing benchmarks
- ☐ better social representativeness compared to existing benchmarks
- ☐ to define a new task
- ☐ suggest other annotation

(3) Which institution/s conducted or funded this research? [open-ended]

(4) What language is or what languages are supported? [open-ended]

(5) How was the data sourced? I.e., what process was used to collect the dataset?

- ☐ synthetic data (e.g. generated)
- ☐ human authored (e.g. crowdsourced)
- ☐ exams or textbooks
- ☐ reuse of existing AI/ML dataset
- ☐ internal/private source (e.g. proprietary customer data)
- ☐ open access data/ web data

- ☐ suggest other annotation

(6) If applicable, where was the data precisely sourced from? [open-ended]

(7) How was the data annotated? E.g., was it manually annotated or automatically?

- ☐ human annotation
- ☐ automatic annotation
- ☐ suggest other annotation

(8) If applicable, where were the ground truth labels sourced from?

- ☐ the web (not further specified)
- ☐ a domain-specific website
- ☐ Wikipedia
- ☐ Freebase
- ☐ suggest other annotation

(9) If applicable, what type(s) of annotator(s) were involved?

- ☐ crowdworker
- ☐ expert
- ☐ student
- ☐ suggest other annotation

(10) If applicable, what were the recruitment criteria for the involved annotators and/or data authors?

- ☐ no recruitment criteria were mentioned
- ☐ expertise in a particular domain
- ☐ crowdworker ranking (e.g. rank on MTurk at the time of recruitment)
- ☐ availability (e.g. friends or colleagues)
- ☐ performance on a task (e.g. screening task)
- ☐ suggest other annotation

(11) What aspects of the annotator/s identity/identities are mentioned?

- ☐ no details about the annotator identity are mentioned
- ☐ country of recruitment
- ☐ country of origin
- ☐ level of education
- ☐ age
- ☐ sex or gender
- ☐ race or ethnicity
- ☐ domain/area of expertise
- ☐ suggest other annotation

(12) Are details about the contents of the dataset reported?

- ☐ no analyses of the dataset contents were reported
- ☐ lexical aspects were analyzed
- ☐ topics/domains were analyzed
- ☐ suggest other annotation

(13) What contents, like topics or knowledge domains are covered by the benchmark dataset?

- ☐ the presented benchmark is not about knowledge
- ☐ news & entertainment
- ☐ everyday knowledge
- ☐ education
- ☐ art, design & music
- ☐ maths
- ☐ language
- ☐ commonsense
- ☐ encyclopedic facts
- ☐ humanities
- ☐ social sciences
- ☐ science, technology & engineering
- ☐ medicine
- ☐ business, economics & finance
- ☐ suggest other annotation

(14) Are analyses of aspects related to social bias, representativeness or toxicity in the dataset reported and, if so, what type of

- ☐ analyses?
- ☐ no analyses regarding social bias, representativeness or toxicity are reported
- ☐ social bias analyses are reported
- ☐ toxicity analyses are reported
- ☐ suggest other annotation