Responsible NLP Checklist

Paper title: *Location Not Found: Exposing Implicit Local and Global Biases in Multilingual LLMs*
Authors: *Guy Mor-Lan, Omer Goldman, Matan Eyal, Adi Mayrav Gilady, Sivan Eiger, Idan Szpektor, Avinatan Hassidim, Yossi Matias, Reut Tsarfaty*

How to read the checklist symbols:

☑ the authors responded 'yes'

☒ the authors responded 'no'

N/A the authors indicated that the question does not apply to their work

☐ the authors did not respond to the checkbox question

For background on the checklist and guidance provided to the authors, see the Responsible NLP Checklist page at ACL Rolling Review.

☑ **A. Questions mandatory for all submissions.**

☑ A1. Did you describe the limitations of your work?
This paper has a Limitations section.

☒ A2. Did you discuss any potential risks of your work?
This work presents a diagnostic dataset and evaluation framework based on public, factual information (e.g., statutory laws, holidays, infrastructure). The dataset does not contain personally identifiable information (PII), offensive content, or subjective data that could harm specific groups. As an evaluation tool designed to expose bias rather than generate it, it does not pose direct safety risks or dual-use concerns.

☑ **B. Did you use or create scientific artifacts? (e.g. code, datasets, models)**

☒ B4. Did you discuss the steps taken to check whether the data that was collected/used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect/anonymize it?
Data consists of general factual questions and answers without risk of PII or offensive content

☑ B6. Did you report relevant statistics like the number of examples, details of train/test/dev splits, etc. for the data that you used/created?
Section B and C provide the question templates, the list of languages and locales. As this is a diagnostic benchmark, the entire dataset functions as a test set; no train/dev splits were created.

☑ **C. Did you run computational experiments?**

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5 ("Setup" paragraph) and Appendix E. We describe the zero-shot and 3-shot inference settings used for all 32 models. As this is an inference-only evaluation of pre-trained models, no hyperparameter search or training was performed.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Sections 5.1 and 5.2. We report aggregated metrics (Global Bias and Regional Bias) calculated over the full dataset of 2,156 examples.

The Responsible NLP Checklist used at ACL Rolling Review is adopted from NAACL 2022, with the addition of ACL 2023 question on AI writing assistance and further refinements based on ARR practice. ACL 2026 used a subset of ARR checklist form.

☑ **D. Did you use human annotators (e.g., crowdworkers) or research with human subjects?**

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix A. We include the guidelines and requirements provided to the professional annotators for the translation and answering tasks.

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Data collection was performed by a professional language services provider (vendor). While the vendor was paid their standard commercial rates for the project, we do not have visibility into the specific compensation received by individual annotators or their specific demographic details beyond their professional qualification as bilingual speakers.

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating (e.g., did your instructions explain how the data would be used)?
Section 3 explains that the dataset consists of objective, factual questions (e.g., laws, dates, infrastructure) generated specifically for this project. The task did not involve collecting personal data, private information, or subjective opinions from the annotators, so personal consent forms were not applicable beyond the standard vendor service agreement.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The study involved standard linguistic annotation tasks (translation and factual question answering) performed by professional vendors under commercial agreements. As the research did not involve psychological experimentation, intervention, or the collection of sensitive personal data, specific institutional ethics review was not sought.

☑ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

☑ E1. If you used AI assistants, did you include information about their use?
AI assistants were used for text editing and coding assistance.