Responsible NLP Checklist

Paper title: *TFDP: Token-Efficient Disparity Audits for Autoregressive LLMs via Single-Token Masked Evaluation*
Authors: *Inderjeet Singh, Ramya Srinivasan, Roman Vainshtein, Hisashi Kojima*

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?
Section Ethical Considerations

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

☑ B1. Did you cite the creators of artifacts you used?
Subsection 4.1 (Design Rationale and Origins) cites the original PDD base data (Bahrami et al., 2024) and explains our extension. We also document release details in Appendix Code and Data Availability and Footnote 1

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We publish code under BSD-3-Clause and data under CC-BY-4.0.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We extend a cited dataset and use external benchmarks only via published scores. Intended research use and access/derivative conditions are stated in Section 4 (dataset description), Section Ethical Considerations, and Appendix Code and Data Availability.

☑ 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?
Section Ethical Considerations. Also see Subsection 4.2 (human screening/QA filters) for curation checks.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4, esp. Subsections 4.14.3; Appendix Extended Dataset Construction Protocol; Appendix Code and Data Availability.

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.

☑ 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?
Subsection 4.3 (Dataset Statistics) and Table 1

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

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We use hosted APIs for commercial/open LLMs; parameter counts are undisclosed for several proprietary models and compute budget (GPU hours) is not applicable. We report models/endpoints, sampling settings, and token/call accounting in Section 5 (Experimental Setup) and Appendix Efficiency Analysis: Tokens and Payload Bytes / Reproducibility and Environment.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5 (Experimental Setup); Subsection 3.3 (Multi-Scale Semantic Alignment, incl. alpha, beta, r); Appendix Embedding and Similarity Hyper-parameters and Hyperparameter Sensitivity.

☑ 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?
Section 6 (Results and Analysis; mean +/- SD and per-model summaries); Appendix Statistical Procedures and Agreement between Single-token and Five-token Probes (CIs, tests).

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?
Section 5 (API sampling params: temperature, nucleus p); Appendix Embedding and Similarity Hyper-parameters (NV-Embed-v2, kernels); Appendix Reproducibility and Environment.

☑ **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.?
Screening was done by internal expert annotators; we report criteria and checks in Subsection 4.2 (Dataset Extension Protocol and Quality Assurance), but do not include full instruction text/screenshots (low-risk internal review).

N/A 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)?
Screening was performed by internal research staff (not crowdworkers); no separate recruitment or per-task payment beyond regular employment. See Subsection 4.2.

N/A 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)?
No personal data were collected/used; corpus consists of non-identifiable proverb-style sentences and LLM-generated paraphrases (Sections 9 and 4).

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No human-subjects research requiring IRB review; internal expert screening of non-identifiable text (Sections 9 and 4.2).

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
A small number of internal expert annotators performed QA; demographics are not reported to avoid identifiability (Subsection 4.2).

☑ **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?
  Subsection 4.2 details LLM-assisted candidate generation (GPT-4o/4.1) and subsequent human screening; Section 5 documents API settings; Appendix Reproducibility and Environment lists prompts/seeds/preprocessing used for experiments.