Responsible NLP Checklist

Paper title: *TabReX: Tabular Referenceless eXplainable Evaluation*
Authors: *Tejas Anvekar, Junha Park, Aparna Garimella, Vivek Gupta*

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?
7. Ethics Statement

☑ **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?
Sources are public.

☑ 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?
2.2 TabReX-Bench; Table 2

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
3. Experiments and Appendix D (Prompt Templates): fixed prompts; provider default decoding (temperature, top-p, top-k), same across models; no hyperparameter search.

☒ 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?
Evaluation reports deterministic rank-correlation metrics over full datasets; no stochastic training or multiple seeds. Where LLM components are used, prompts/decoding are fixed and single-pass.

☑ **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 Human Evaluation Protocol (instructions + ranking rubric and tie-break rules)

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.

☑ 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)?
In-house expert annotators (co-authors). No external crowdworkers; no separate payment (effort covered by authors research time). Demographics are omitted to preserve anonymity.

☑ 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)?
2.2 TabRex-Bench Public or de-identified datasets only. Annotators (co-authors) consented to contribute rankings internally.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No new human-subjects data beyond co-author annotations; activity does not meet the definition of human-subjects research at our institutions.

☑ **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?
7. Ethics Statement: AI tools (e.g., Grammarly, ChatGPT) were used for language editing and clarity.