Responsible NLP Checklist

Paper title: *ChemReason-Bench: Benchmarking Large Language Models for Procedural Reasoning in Experimental Chemistry*
Authors: *Jinwei Zhang, Xucheng Liang, Yu Zhang, Ruijie Yu, Xiaokang Yang, Yaohui Jin, Yanyan Xu*

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?
We discuss potential risks in Section Limitations (Potential risks).

☑ **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?
We discuss the steps taken in Section E.1 (PII and offensive content screening).

☑ 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?
We report the relevant statistics in Section 1 and Section 4.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We discuss the experimental setup in Section H.2.

☑ 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?
We report descriptive statistics in Section F.3.5.

☑ **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.?
We did not include the full verbatim annotator instruction text in the paper. Instead, we report the annotation purpose, verification protocol, sampling procedure, acceptance thresholds, adjudication

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.

process, and quality-control steps used in manual review. The paper therefore summarizes the annotation procedure, but not the complete instruction text provided to annotators.

☒ 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)?
Human verification was conducted internally by the authors/collaborators; no external recruitment or participant payment was involved.

☒ 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)?
Our dataset is curated from publicly available sources (e.g., open repositories). We do not collect personal data from individuals; any incidental PII is screened and removed during canonicalization.

☒ N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The dataset is curated from publicly available sources (e.g., open repositories) and does not involve collecting personal data or conducting interventions with human subjects; internal verification was limited to checking and correcting task instances. Therefore, ethics board approval was not required.

☑ **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?
We include the use of AI in Section 4.