Responsible NLP Checklist

Paper title: *ReTRE: Benchmarking LLM Transfer Robustness with Structure-Preserving Variants*
Authors: *ZhongDong Li, Weijie Shi, Yue Cui, Haolun MA, Yuanjun Liu, Jiawei Li, An Liu, Jia Zhu, Jiajie 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.

N/A A2. Did you discuss any potential risks of your work?
This work introduces an evaluation benchmark and does not pose significant potential risks. The benchmark uses publicly available datasets (MATH500, GPQA) and generates synthetic task variants through controlled rewrites that do not introduce harmful or sensitive content.

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

N/A 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?
The datasets used in this work (MATH500 and GPQA) consist of mathematical reasoning problems and scientific multiple-choice questions, which do not contain personally identifying information or offensive content by design. Our generated transfer variants are structure-preserving rewrites of these mathematical and scientific problems, and similarly do not introduce any personal information or offensive material.

☑ 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?
Sections, Limitations and Appendix A.3.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Inference settings for all evaluated models (temperature=0, pass@1 metric) are described in Section 4.1.

☒ 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?
All experiments used temperature=0 for deterministic outputs, and results are from single runs

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.

with pass@1 metric (Section 4.1). Error bars were not reported as the deterministic setting yields reproducible results.

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

N/A 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.?
(left blank)

☑ 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)?
We report annotator recruitment and compensation in Appendix A.3.

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)?
(left blank)

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
(left blank)

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

N/A E1. If you used AI assistants, did you include information about their use?
We only use AI to refine the writing and the contents are checked by authors.