Responsible NLP Checklist

Paper title: *ReCoQA: A Benchmark for Tool-Augmented and Multi-Step Reasoning in Real Estate Question and Answering*
Authors: *Yindong Zhang, Wenmian Yang, Yiquan Zhang, Weijia Jia*

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?
All data used in this paper were obtained in a public environment and do not involve any sensitive information, personal privacy, or content that violates social ethics.

☑ **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?
Our data do not contain personally identifying information 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 3.5 (Data Annotation and Verification), and Appendix D (Statistics)

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5.1 (Implementation and Metrics), and Appendix F (Implements)

☒ 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?
Due to the high resource and time consumption of calling LLMs, the experimental results in this paper are based on single run results.

☑ **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.?
Section 3.1 (Data Collection and Preprocessing), Section 3.3 (QA Pair Generation), and Appendix B (QA Pairs Evaluation)

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 this experiment, we sought student volunteers through online recruitment. All participants were informed before the experiment that the project was a voluntary service, and they participated on a voluntary basis. For real estate experts and salespeople, we communicated with them through online consultation, and we did not pay any compensation to these participants

☒ 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)?
All the data used in this paper were obtained in a public environment and were utilized in accordance with basic ethical principles. The data does not involve any sensitive information, personal privacy, or the like, thus consent is not required

☐ 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?**

☑ E1. If you used AI assistants, did you include information about their use?
Section 3.3 (QA Pair Generation), and Appendix B (QA Pairs Evaluation)