Responsible NLP Checklist

Paper title: *seqBench: A Tunable Benchmark to Quantify Sequential Reasoning Limits of LLMs*
Authors: *Mohammad Ramezanali, Mo Vazifeh, Paolo Santi*

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?
(left blank)

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

☑ B1. Did you cite the creators of artifacts you used?
Section 1.1 (Dataset Generation) cites Kruskal's algorithm used in the creation of the seqBench benchmark. Section 3 (Related Work) and Section 2.1 (Evaluated Models) cite existing benchmarks, models, and platforms used for comparison.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
Introduction, specifically paragraph 4 and footnote 1

☑ 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)?
The Abstract, Introduction, and Conclusion specify the intended use of the created seqBench benchmark which is for benchmarking sequential reasoning in LLMs

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?
(left blank)

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 1.1 (Dataset Generation), Appendix A (Dataset Generation Details), and Appendix B (Prompt Design and Model Configuration Details) provide extensive documentation on the generation of the seqBench benchmark. The domain is synthetic spatial pathfinding using templated English factual statements.

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?
Section 2.1 (Evaluated Models), the caption for Figure 1 and Figure 2, and Appendix A.3 (Dataset Parameters and Scope), specify the number of instances and configurations. While explicit train/test/dev splits are not applicable as the benchmark is used for evaluating pre-trained LLMs in this work, the sampling methodology for constructing evaluation sets is described.

☑ **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?
(left blank)

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 1.2 (Prompt Construction and Model Configuration) and Appendix B (Prompt Design and Model Configuration Details) detail the prompt structure. Section 1.2 specifies the inference hyperparameters used for all models (Temperature=1.0, nucleus sampling top-p=0.95, and maximum allowed output tokens per model).

☑ 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 2.1 Figure 2, Error bars are shown in figures reporting mean progress ratio, precision, and recall.

☒ 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?
(left blank)

☒ **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.?
(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)?
(left blank)

☒ 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)

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

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
(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?
For assistance primarily with grammar and minor edits