Responsible NLP Checklist

Paper title: *Scaling Unverifiable Rewards: A Case Study on Visual Insights*
Authors: *Shuyu Gan, James Mooney, Pan Hao, Renxiang Wang, Mingyi Hong, Qianwen Wang, Dongyeop Kang*

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?
Section 7

☑ **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?
All datasets used contain already publicly available and commonly used data and visualizations. In all datasets, no names or PII was present.No human subjects were involved in data collection for our work.

N/A 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 documented dataset domains and usage in Section 3, 4 and Appendix B, D, but did not include detailed numerical statistics (e.g., number of examples or splits) since our analysis focuses on test-time scaling behavior rather than dataset-level benchmarking.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3, Section 4, Appendix E

☑ 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 4, Appendix E

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.

☑ **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 D

☑ 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)?
Appendix D

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)?
No consent was required since all datasets used are publicly available for research use and contain no personal or identifiable information.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Ethics review board approval was not required, as the study used publicly available datasets and voluntary participation from graduate student annotators without collecting personal or sensitive data.

☑ **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?
Appendix A