Responsible NLP Checklist

Paper title: *Parrot: A Training Pipeline Enhances Both Program CoT and Natural Language CoT for Reasoning*
Authors: *Senjie Jin, Lu Chen, Zhiheng Xi, Yuhui Wang, Sirui Song, Yuhao Zhou, Xinbo Zhang, peng sun, Hong Lu, Tao Gui, Qi Zhang, Xuanjing Huang*

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?
Yes, Section Limitations/Ethical Considerations.

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

☑ B1. Did you cite the creators of artifacts you used?
Yes, Section 4 and Appendix C.

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
No, the datasets from ReFT: Reasoning with REinforced Fine-Tuning are under Apache 2.0 License.

☒ 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)?
My use of these artifacts fully complies with the relevant licenses. I will release the artifacts I create and specify their intended use.

☑ 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?
Yes, Section Ethical Considerations.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Yes, in Section 4.1, we present the details of the datasets and models.

☑ 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?
Yes, Appendix C.

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.

☑ **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?
Yes, Section 4 and Appendix B.

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

☑ 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?
Yes, Section 4.

☑ 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?
Yes, Appendix B.

☑ **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.?
No, Section A.1, the authors identified the error types through manual inspection, and developed the corresponding instructions for model evaluation, so the manual inspection was solely for the purpose of setting instructions.

☒ 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)?
No, the participants in the manual evaluation are all the authors.

☒ 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, the participants in the manual evaluation are all the authors. We use the model to relabel the data after getting the corresponding instructions.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No, we use the dataset under the Apache 2.0 license and re-annotated them in paper.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No, the participants in the manual evaluation are all the authors.

☑ **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?
Yes, Appendix A.2 and Appendix C.