Responsible NLP Checklist

Paper title: *SCRIBE: Structured Chain Reasoning for Interactive Behaviour Explanations using Tool Calling*
Authors: *Fares Fawzi, Vinitra Swamy, Dominik Glandorf, Tanya Nazaretsky, Tanja Kser*

> 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?
In section 1, We provide the link to our repository. In section 3, we cite all other models we used.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
In section 1, we mention that our models and dataset are open-source

☑ 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 datasets are either synthetic or collected via a user study approved by the EPFL Human Research Ethics Committee (Section 3.1 and Appendix F). We used open-source Llama and ToolACE models for fine-tuning, as well as GPT-4o/GPT-4.1 for data generation and evaluation (Sections 3.1, 3.2 and 4).

☒ 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?
We used synthetic training data, and all human data were anonymized. The data does not contain any offensive content.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Artifacts are documented in sections: 3.1 (datasets, tools, synthetic data), 3.2 and 4 (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?
All the details are in Section 4 under "Experimental Protocol"

☑ **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?
All the details are in Section 4 under "Experimental Protocol"

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
All the details are in Section 4 under "Experimental Protocol".

☑ 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?
Sections 4.1, 4.2, 4.3, and 4.4. appendices: A, B, C, D, F, I

☑ 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?
All external models are mentioned in sections 3 and 4. Section 1 provides the link to the repository where all the implementation details with all external libraries can be found.

☑ **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 C and Appendix F

☑ 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 F

☑ 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)?
Participants in the user study gave informed consent; they could not proceed without reading and agreeing to the consent statement in Appendix F.2.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Our study was approved by the EPFL Human Research Ethics Committee (HREC) under application number HREC 065-2022/27.09.2022.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix F outlines all the demographics of the participants

☑ **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 7