Responsible NLP Checklist

Paper title: *NSF-CoT: Neuro-Symbolic Formal Verification of Chain-of-Thought Faithfulness in Contextual Question Answering*
Authors: *Vishal Pramanik, Maisha Maliha, Nathaniel D. Bastian, Alvaro Velasquez, Susmit Jha, Sumit Kumar Jha*

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?
Discussed in Sections 7 (Limitations) and 8 (Ethical Statement). Section 8 specifically addresses risks of over-trust in fluent rationales, inherited biases from underlying LMs and the LLM judge, potential for misuse to create an undeserved appearance of rigor, and recommendations for use as an auditing aid rather than definitive arbiter, especially in high-stakes domains.

☑ **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?
Section 8 (Ethical Statement) notes that the approach does not require access to personal user data and is evaluated on publicly released benchmark datasets (OpenBookQA, QASC, HotpotQA), which have been widely used and vetted by the research community. No new data was collected from human subjects.

☑ 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 3.1 describes the three QA benchmarks used (OpenBookQA, QASC, HotpotQA) and their compositional structure. Standard public splits from the original dataset papers (Mihaylov et al., 2018; Khot et al., 2020; Yang et al., 2018) are used without modification.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3.1 details the experimental setup and hyperparameters: M=128 ablation samples, LASSO =0.01, attribution threshold =0.5, SMT inference axioms (Appendix C), LLM judge at temperature 1.0, hybrid weight =0.5, faithfulness threshold _faith=0.5. Appendix F.1 provides a full sensitivity

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.

analysis of across {0.0, 0.25, 0.5, 0.75, 1.0}, and Appendix F.2 compares LLM judge choice (o1-preview vs. o3).

☑ 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?
The Experiment section contains all and the table captions also contains these informations

☒ **D. Did you use human annotators (e.g., crowdworkers) or research with human subjects?**

N/A 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)

N/A 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)

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)?
(left blank)

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?
AI assistants were used for light proofreading of prose and debugging of LaTeX and spelling checks. All research design, experiments, analysis, and final writing decisions were made by the authors.