## Responsible NLP Checklist

Paper title: *The Illusion of Insight in Reasoning Models*
Authors: *Liv G. d'Aliberti, Manoel Horta Ribeiro*

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 9 (Ethical Considerations). The paper discusses two main risks: misuse of methods that manipulate mid-trace behavior to steer models toward undesirable or deceptive outputs, and the risk of overstating model understanding through claims about insight or self-correction.*

☑ **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 9; Appendix B.4. Section 9 states that the datasets used are publicly available and contain no sensitive content or personally identifiable data. Appendix B.4 also states that no sensitive personal information or sensitive demographics were collected from annotators, and that released artifacts exclude operational contact data and use anonymized traces.*

☑ 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 4; Appendix A.1; Table 6. The paper reports dataset domains, train/eval sizes, and split details, including Xwords (50,000 train / 130 eval), Math (220,000 / 500), and RHour (180,000 / 500).*

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Section 5; Appendix A.3A.4; Table 9. The paper describes the training and evaluation setup, checkpoint cadence, decoding settings, model families, GRPO configuration, and per-domain hyperparameters such as learning rate, batch size, grad accumulation, epochs, num generations, and token budgets.*

☑ 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,

*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.*

etc. or just a single run?
*Sections 56; Tables 15; Appendix A.3, B.3, C, and D. The draft reports many descriptive statistics, including accuracies, prevalence rates, raw percentage-point differences, AMEs, p-values, prompt robustness mean/std/range, agreement statistics, bootstrap confidence intervals, and explicit notes on pooled/count-weighted averages.*

☑ **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 B.4. The appendix includes the annotator task description, labeling rubric, quick checklist, worked examples, and a sample annotation question.*

☑ 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)?
*Elaboration: Appendix B.4. The paper reports that annotators were 6 volunteer adult annotators, recruited from the authors academic networks, and that they were unpaid. Because no payment was provided, payment adequacy is not applicable.*

☑ 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)?
*Answer: Yes Elaboration: Appendix B.4. The paper states that participants gave informed consent on the task page and could withdraw at any time.*

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*Appendix B.4. The draft states that, under the authors institutional guidelines, the activity did not constitute human-subjects research, and therefore no IRB review was sought.*

☑ **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 9 (Ethical Considerations). The paper states that Claude, ChatGPT, and Elicit were used to help identify related work; ChatGPT was used to streamline and refine prose after drafting; and Claude/ChatGPT were also used for formatting tasks such as generating table templates and translating supplementary materials to LaTeX, with authors reviewing and correcting all resulting text.*