Responsible NLP Checklist

Paper title: *Diagnosing and Mitigating Sycophancy and Skepticism in LLM Causal Judgment*
Authors: *Edward Y Chang*

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?
Ethics Statement section (page 10): discusses false assurance risk and data contamination risk.

☑ **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 data and code do not contain any private information nor offensive content.

☑ 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?
*4.1 and Appendix A.4: 454 cases across 10 domains; Table 7 gives per-domain breakdown (L1:50, L2:304, L3:100). No train/test split CausalT3 is an evaluation benchmark.C Computational Experiments*Yes

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4.1 (Experiment Setup): models, temperature T=0, and Table 10 (Appendix A.7) documents all protocol controls including label sets and prompting conditions.

☑ 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?
4.1 (Statistical analysis): 95% Clopper-Pearson CIs for all proportions, two-proportion z-tests with Bonferroni correction (=0.005) for 10 primary comparisons. CI half-widths reported.

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 C: full annotation guidelines including labeling rubric (C.1), trap identification protocol (C.2), and wise refusal guidelines (C.3).

☑ 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 C.4: 10 graduate students in Computer Science and Engineering, selected based on coursework familiarity with causal inference. Participation was voluntary as part of standard research training; no external crowdworkers were employed.

☑ 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)?
Appendix C.4: annotators were members of the research group who participated voluntarily as part of standard research training.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Annotators were voluntary members of the research group performing annotation as part of standard research training, not external human subjects. No external crowdworkers or vulnerable populations were involved.

☑ **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?
"Use of AI Assistants" section (page 10): LLMs (Claude and GPT) used for Python code generation in the evaluation pipeline, LaTeX formatting, and manuscript editing. All scientific claims, experimental designs, and annotations were generated and verified by human authors.