Responsible NLP Checklist

Paper title: *SycoBench-600: Measuring Sycophancy and Correction Selectivity in LLM Assistants*
Authors: *Debu Sinha*

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?
The benchmark uses publicly available factual MCQ items and does not involve sensitive or personal data. No foreseeable ethical risks identified.

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

N/A 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?
The dataset contains factual MCQ items drawn from public knowledge domains. No personally identifying information or offensive content is present.

☑ 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 (Dataset) and Table 2 report dataset statistics including number of instances, domain distribution, difficulty tiers, and train/test details.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4 (Experimental Setup) describes models evaluated, API parameters (temperature=0), and the perturbation protocol with 3 fixed prompt variants per condition.

☑ 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?
Section 5 (Results) and Tables 3-4 report cluster-bootstrap 95% confidence intervals for all sycophancy metrics.

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?**

☒ 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.?
N/A. No human annotators or subjects were used.

☒ 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)?
N/A. No human annotators or subjects were used.

☒ 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)?
N/A. No human subjects data was collected.

☒ N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
N/A. No human subjects research was conducted; no ethics review required.

☑ **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 coding assistants (Claude, ChatGPT) were used for code development. The benchmark design, experimental methodology, analysis, and scientific conclusions are entirely my own work.