Responsible NLP Checklist

Paper title: *Masked by Consensus: Disentangling Privileged Knowledge in LLM Correctness*
Authors: *Tomer Ashuach, Shai Gretz, Yoav Katz, Yonatan Belinkov, Liat Ein-Dor*

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)**

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?
All datasets used are publicly available benchmarks (Mintaka, TriviaQA, HotpotQA, MATH, GSM1K) that do not contain personally identifying information or 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?
Appendix B (Table 2) reports dataset sizes and disagreement subset sizes for all model pairs.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3.4 describes the experimental setup. Appendix D provides full implementation details including probe architectures, hyperparameters, and cross-validation procedures.

☑ 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?
We report 95% confidence intervals via bootstrap resampling, and statistical significance via paired t-tests with Bonferroni-Holm correction ($p < 0.05$) throughout the paper.

☒ **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.?
No human subjects or annotators were used in this study.

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.

☒ 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)?
No human subjects or annotators were used in this study.

☒ 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)?
No human subjects or annotators were used in this study.

☒ N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No human subjects or annotators were used in this study.

☑ **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?
Appendix J describes AI assistant usage for text refinement and coding assistance. All scientific claims and final text were written by the authors.