Responsible NLP Checklist

Paper title: *The Digital Dunning-Kruger Effect: Decoupling Hallucinations via Geometric Hidden-state Observation for Semantic Truthfulness*
Authors: *Yueheng Mao, Min Yu, Gengwang Li, Jianguo Jiang, Gang Li, Meng Zhang, Zhen Xu, Weiqing Huang, Ming Liu*

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 "Ethics Statement". We discuss risks including false negatives, false positives, and the potential misuse of GHOST as an automated censorship tool.

☑ 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 (FinanceBench, RAGTruth, HaluEval, PopQA) are publicly available benchmarks that do not contain personally identifying information.

☑ 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 Section A.1 (Table 5). We report total sample counts, train/test splits, and domain characteristics for all four datasets.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix Section A.3. We detail the Random Forest hyperparameter search space, selected values, and cross-validation protocol.

☑ 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?
All experiments are deterministic due to fixed random seeds and greedy decoding (temperature=0). Therefore, repeated runs yield identical outputs, and single-run deterministic results are reported. Variance statistics are not applicable in this setting.

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.?
Annotators followed standardized factuality assessment guidelines based on benchmark annotation protocols and internal expert review criteria. A summary of the annotation procedure is described in Section 4.1.

☒ 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)?
Annotators were domain experts from our research institution participating as part of institutional research activities; no separate monetary compensation was provided.

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)?
All datasets used are publicly available; no new data collection from human subjects was conducted.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Human evaluation involved expert review of model outputs by internal researchers and did not involve collection of sensitive personal data. According to our institutional guidelines, separate IRB approval was not 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 assistants were used exclusively for language editing and grammar refinement of the manuscript. All scientific content, experimental design, analyses, and conclusions are entirely the authors' own work.