Responsible NLP Checklist

Paper title: *Dynamic PMI-Guided Contrastive Decoding Reduces Hallucination in Large Language Models: A Unified Framework of Fine-Grained Input Transformations*
Authors: *Dongsheng Chen, Yingqi Zhu, Xingyue Zhang, Wenqing Zhou, Lei Li*

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?
The paper focuses on an inference-time decoding algorithm to mitigate hallucinations in LLMs. It does not introduce new societal risks or harmful applications.

☑ **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?
The study uses standard public benchmarks (e.g., TruthfulQA, MMLU) and synthetically constructed sentence templates, which 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?
Section 4.1 and Appendix C report the sample sizes used for evaluation, such as the stratified random sample of 200 instances for generative benchmarks and 200 samples per category for the constructed diagnostic datasets.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 6.4 details the hyperparameter search and sensitivity analysis for the penalty intensity parameter (beta).

☒ 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?
The proposed contrastive decoding methods are deterministic inference strategies given a fixed context. Therefore, standard single-run performance metrics (Accuracy, MC1/2/3) are reported without error bars.

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 A reports the specific metrics (Truth, Info, Reject) and states strict adherence to the official, publicly available TruthfulQA evaluation standards. Since the evaluation was conducted internally by the research team who are already familiar with these established guidelines, rather than by external crowdworkers, providing redundant instructional text and UI screenshots was deemed unnecessary.

☒ 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)?
The human evaluation of the 100 sampled instances was conducted internally by the research team to validate automatic metrics. No external paid crowdsourcing platforms 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)?
The evaluation strictly involved rating machine-generated text for factual accuracy. No personal data was collected from the evaluators.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The task was limited to assessing the factuality of AI-generated responses and does not constitute human subjects research requiring Ethics Review Board approval.

☑ **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 4.3 explicitly states the use of GPT-4o for automatic evaluation of generative reasoning tasks, and Appendix C notes its use in constructing the diagnostic datasets.