Responsible NLP Checklist

Paper title: *Bridging Internal Consistency and External Alignment: A Causal and Dynamic Interpretability Framework for LLM Generation*
Authors: *Shuyao Xiao, Shengling Wang, Ke Chao*

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?
This work focuses on the interpretability analysis of large language model generation. It does not introduce new model architectures, training objectives, or deployment mechanisms, and does not involve human subjects or sensitive data. As such, we do not identify any potential risks specific to this work.

☑ **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?
This work does not involve the collection or use of data containing personally identifiable information or offensive content. All experiments are conducted on publicly available datasets or generic model-generated examples, without reference to identifiable individuals or sensitive groups.

☑ 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.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.

☑ 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 4.

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.?
This work does not involve human subjects or human annotators. All analyses are conducted using model-generated outputs and automated metrics, without any human participation. Therefore, no participant-facing documentation is required.

☒ 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)?
This work does not involve human subjects or human annotators. All analyses are conducted without any human participation. Therefore, no participant-facing documentation is required.

☑ 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)?
Section 4.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
This work does not involve human subjects, human annotation, user studies, or new data collection. All experiments are conducted using publicly available data or model-generated outputs. As a result, ethics review board approval is 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?
Appendix B.