Responsible NLP Checklist

Paper title: *Flip-Flop Consistency: Unsupervised Training for Robustness to Prompt Perturbations in LLMs*

Authors: *Parsa Hejabi, Elnaz Rahmati, Alireza Salkhordeh Ziabari, Morteza Dehghani*

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?
No substantial potential risks were identified. The work uses publicly available benchmark datasets and open-weight models, involves no new human data collection, and focuses on classification robustness/consistency. Relevant data, licensing, and privacy considerations are discussed in the Ethical Considerations section.

☑ **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?
Ethical Considerations. The work uses publicly available benchmark datasets and no new human data collection. Some datasets may include named entities (e.g., public figures in news excerpts), but to the best of our knowledge they do not contain contact details or other sensitive personal identifiers. The paper discusses these data considerations in the Ethical Considerations section.

☑ 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 A.2. We report the datasets used, task categories, and details of the train/validation setup, including how validation splits were derived from the original training data and the hold-out sizes.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.2 and Appendix A.3. We describe the experimental setup, including the base models, LoRA fine-tuning setup, model selection criterion, compute/infrastructure details, and the FC hyperparameters used for the reported results. We also state that the FC hyperparameters were chosen separately for each dataset using a small set of manual configurations rather than an exhaustive per-dataset search.

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.

☑ 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 3.4 and Section 5. We report descriptive statistics for results, including mean F1, observed agreement, and the standard deviation of F1 across prompt variations. We also report aggregate transfer statistics with positive/negative counts in Table 3, and Figure 2 includes error bands based on std of F1 over held-out formats.

☒ **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.?
(left blank)

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)?
(left blank)

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)?
(left blank)

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
(left blank)

☑ **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?
We used ChatGPT for fixing grammatical errors.