

## Responsible NLP Checklist

Paper title: *Stay Focused: Problem Drift in Multi-Agent Debate*

Authors: *Jonas Becker, Lars Benedikt Kaesberg, Andreas Stephan, Jan Philip Wahle, Terry Ruas, Bela Gipp*

How to read the checklist symbols:

- the authors responded 'yes'
- the authors responded 'no'
- 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?

*Not applicable because the work investigates the characteristics of multi-agent debate.*

### B. Did you use or create scientific artifacts? (e.g. code, datasets, models)

- B1. Did you cite the creators of artifacts you used?

*We cite all datasets in section 4.3. We provide a citation to the used multi-agent framework on page 11. We provide a link to the paper project code in the footnote of page 2.*

- B2. Did you discuss the license or terms for use and/or distribution of any artifacts?

*We do not list the license of each used dataset because of the number of datasets used (10). Our code is licensed under the Apache 2.0 License, as indicated by the GitHub repository. Our released data (DRIFTEval) is licensed under CC0.*

- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?

*Not applicable*

- 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?

*Not applicable*

- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?

*Table 6 of Appendix E*

- 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?

*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.*

*Appendix E gives information of dataset sampling and Table 6 of Appendix E details the number of samples used per dataset. The created dataset DRIFTEval is detailed in Appendix I.2.*

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

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?

*We detail the used Hardware and model parameters by the end of Section 4.1*

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

*Discussed in Section 4.1 and 4.2*

- 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 the standard deviations between three experiment runs for our results (table 1, 2, 4, 5, 7, and 8)*

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?

*Packages and parameters are detailed in Appendix B*

**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 I.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)?

*Yes, in Appendix I.2 "Human Assessment"*

- 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)?

*Not applicable*

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

*Not applicable*

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

*Appendix I.2 "Human Assessment" includes the demographic groups of our proposed dataset DRIFTEval.*

**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 K*