

## Responsible NLP Checklist

Paper title: *Taming Actor-Observer Asymmetry in Agents via Dialectical Alignment*

Authors: *Bobo Li, Wu Rui, Zibo Ji, Meishan Zhang, Hao Fei, Min Zhang, Mong-Li Lee, Wynne Hsu*

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?  
*See 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?  
*See the "Ethical Considerations" section. We implemented safety filters and human-in-the-loop verification.*

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?  
*See Appendix A (Table 7) for dataset statistics and Section 3.1 for benchmark construction details.*

### C. Did you run computational experiments?

C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?  
*See Section 5.1 (Settings) and Appendix A (Supplementary Experiment Details).*

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?  
*See Section 5 and Appendix A (reward weight sensitivity analysis in Table 6).*

### 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.?  
*Human verification was performed internally by the authors, not by recruited annotators.*

---

*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.*

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)?

*No external annotators were recruited.*

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 data is synthetically generated.*

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

*No human subjects research was conducted.*

**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 GPT-5.1 for synthetic data generation (Section 3.1) and dialectical trajectory synthesis (Section 4.3).*