

## Responsible NLP Checklist

Paper title: *TwinVoice: A Multi-dimensional Benchmark Towards Digital Twins via LLM Persona Simulation*

Authors: *Bangde Du, Minghao Guo, Songming He, Ziyi Ye, Xi Zhu, Weihang Su, Shuqi Zhu, Yujia Zhou, Yongfeng Zhang, Qingyao Ai, Yiqun LIU*

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?

*Section "Ethics Statement" (Page 10)*

### 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?

*Section "Ethics Statement" (Page 10)*

- 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?

*Table 1 (Page 5) and Section 3.1 "Data Pre-processing"*

### C. Did you run computational experiments?

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

*Appendix A (Evaluation Protocols and Full Prompts), Appendix B (mentions temperature settings, e.g., temperature 0), and Appendix F.2 (Models and Access)*

- 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 (Experiments), Table 3 (reports accuracy and mean scores), and Tables 4 & 6 (reports inter-annotator agreement statistics like Fleiss's kappa and Spearman correlation)*

*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 E "Human Annotation Guidelines" (Pages 15-17)*

- 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)?

*Appendix E.1 "Task Background and Objectives" (Page 15)*

- 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 "Ethics Statement" (Page 10)*

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?  
*The study utilized publicly available datasets and did not involve human subjects research requiring formal IRB approval. As stated in the "Ethics Statement", voluntary annotators provided informed consent and were fairly compensated.*

**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 F "Use of Large Language Models" (Page 18)*