Responsible NLP Checklist

Paper title: *PersonaForge: Psychology-Grounded Dual-Process Architecture for Personality-Consistent Role-Playing Agents*
Authors: *Jizhou Tong, Sirui Zou*

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?
Discussed in the "Ethical Considerations" section under the "Risk Mitigation" framework, detailing content boundaries and anti-dependency measures.

☑ **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?
Addressed in "Ethical Considerations" (Copyright and Data Availability) and Appendix D.3 (Safety Architecture Details). The data involves literary characters rather than real individuals, and a multi-tier safety layer mitigates the generation of harmful or manipulative responses toward users.

☑ 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?
Data statistics, including character counts, dialogue sample sizes, and train/val/test splits for trigger learning, are detailed in Section 4.1, Appendix A.1, and Appendix B.6.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Experimental setups, hardware details, baseline configurations, and fine-tuning hyperparameters are documented in Section 4.3, Section 5.5, Appendix A.1, Appendix B.3, and Appendix D.23.

☑ 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?
Tables 1 and 11 report 95% confidence intervals and standard deviations to clearly reflect variance in evaluation metrics and latency. Additional bounds like standard error of the mean are discussed in Appendix D.11.

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.?
Evaluation rubrics, guidelines for identifying defense mechanisms, and LLM-as-Judge instructions corresponding to human tasks are provided in Appendix A.4 and A.5.

☑ 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)?
Recruitment details are specified in Appendix D.11, mentioning the use of domain experts and crowd workers via the Prolific platform filtered for high approval rates.

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)?
The dialogue generation and personality extraction rely on public domain literary text and character wikis, not on data derived from human subjects.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Human evaluation consisted of standard annotation tasks rating generated texts representing fictional literary characters, which typically falls under exempt categories for human subject research.

☑ **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?
Disclosed in the "Acknowledgements" section, specifying the use of the gemini-3-pro-preview model for text polishing and language refinement.