Responsible NLP Checklist

Paper title: *Personality Matters: User Traits Predict LLM Preferences in Multi-Turn Collaborative Tasks*
Authors: *Sarfaroz Yunusov, Kaige Chen, Kazi Nishat Anwar, Ali Emami*

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?
We discuss ethical considerations and limitations in the "Ethical Considerations" section and "Limitations" section, including data privacy risks and potential issues with demographic representation.

☑ **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 the creators of the LLMs (GPT-4 and Claude 3.5) in the Models paragraph of the Experimental Setup section with citations to OpenAI and Anthropic. We also cite the creators of the personality assessment tool (Keirsey Temperament Sorter) and the sentiment analysis model.

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We did not explicitly discuss the licenses for the commercial LLMs used. Both GPT-4 and Claude 3.5 are commercial products with terms of service that allow for research use.

☑ 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)?
In the Experimental Setup section , we describe our use of GPT-4 and Claude 3.5 for collaborative task completion, which is consistent with their intended use as conversational agents.

☑ 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?
In the Ethical Considerations section under "Data Privacy", we discuss how "All participant data were anonymized immediately after collection, with personally identifiable information replaced by anonymous identifiers."

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

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.

We document participant demographics in the "Participants and Recruitment" paragraph with detailed breakdowns in Appendix Figure 1 covering education level, gender distribution, and LLM familiarity.

☑ 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?
We report participant statistics in the Experimental Setup section, including the number of participants (32), their distribution across personality types (8 per type), and assignment to models (4 per personality type per model).

☑ **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 did not report the number of parameters or computational budget as we used the commercial API versions of GPT-4 and Claude 3.5, not locally trained models. The computing infrastructure used was the providers' own servers accessed via API.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We discussed experimental setup in our Experimental Setup section, and as for hyperparameters, our study focused on human-LLM interactions using commercial APIs with default settings rather than model training or hyperparameter tuning.

☑ 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 descriptive statistics in the Results section and in Appendix Table 1, including means and standard deviations of helpfulness ratings across personality types and tasks.

☑ 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?
We mention the use of BERT-Base-Multilingual-Uncased-Sentiment for feedback analysis in the "Feedback Analysis Confirms Personality-Based Model Preferences" paragraph with appropriate citation.

☑ **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.?
We provide task descriptions in Appendix Figures 2-5 and the survey question in Appendix Figure 6, showing the full text of instructions given to participants.

☑ 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)?
In the "Participants and Recruitment" paragraph, we mention that "Recruitment was conducted through university mailing lists and direct outreach." The "Ethical Considerations" section notes that "compensation was provided in accordance with institutional guidelines."

☑ 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)?
In the Ethical Considerations section under "Informed Consent", we state that "All participants provided explicit informed consent after receiving detailed information about the study purpose, tasks, data collection, and their rights."

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We provide this information in the Ethical Considerations section (we did indeed recdeive institutional ethics board approval)

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We report demographic characteristics of participants in the "Participants and Recruitment" paragraph and provide detailed breakdowns in Appendix Figure 1, including education level, gender distribution, and LLM usage patterns.

☒ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

N/A E1. If you used AI assistants, did you include information about their use?
(left blank)