Responsible NLP Checklist

Paper title: *Too Nice to Tell the Truth: Quantifying Agreeableness-Driven Sycophancy in Role-Playing Language Models*
Authors: *Arya Shah, Deepali Mishra, Chaklam Silpasuwanchai*

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?
Section 5.5 (Broader Impact) - The paper discusses safety implications of persona-induced sycophancy, noting that "agreeable personas require additional safeguards in character AI and roleplay applications" and that "personality is not neutral in LLM deployment." Section 6 (Limitations) also discusses scope decisions and boundary conditions.

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

N/A 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?
The personas are synthetic/fictional characters created for research purposes (not real people), and the prompts are opinion-based on various topics. No PII is involved since all personas are fictional constructs.

☑ 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?
Section 3.1 and Appendix A.3 - The paper reports: 275 personas, 4,950 prompts across 33 categories, 13 models with parameter counts (0.6B to 20B), and complete computational statistics (17.9M total queries: 143,000 agreeableness queries, 64,350 baseline queries, 17,696,250 persona queries).

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix A.1 and A.2 - Complete hardware/software environment (NVIDIA RTX A6000 GPUs, PyTorch, Hugging Face Transformers), inference parameters (max tokens: 150, greedy decoding, bfloat16 precision, batch sizes 8-32), and generation settings for reproducibility.

☑ 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 and Appendix A.5 - The paper reports Pearson correlations (r), Spearman correlations (), p-values, Cohen's d effect sizes, Hedges' g, 95% confidence intervals, R values from regression, and complete statistical tables for all 13 models.

☒ **D. Did you use human annotators (e.g., crowdworkers) or research with human subjects?**

N/A 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.?
(left blank)

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)?
(left blank)

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)?
(left blank)

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

☑ **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?
AI writing assistants were used for drafting and editing paper text. The scientific methodology, experiments, and analysis were conducted independently.