Responsible NLP Checklist

Paper title: *HumanLLM: Benchmarking and Improving LLM Anthropomorphism via Human Cognitive Patterns*
Authors: *Xintao Wang, Jian Yang, Weiyuan Li, Rui Xie, Jen-tse Huang, Jun Gao, Shuai Huang, Yueping Kang, Yuanli Guo, Hongwei Feng, Yanghua Xiao*

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 Statement section (pages 9-10), covering safety-fidelity tension, manipulation risks, parasocial attachment, and stereotype amplification.

☑ **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?
All training conversations are synthetically generated by LLMs (Gemini 2.5 Pro, Claude Sonnet 4.5). Character names are sampled from a public name-dataset library without linking to real individuals. No personally identifying information is included. Discussed in Section 3.3 and Appendix B.3.

☑ 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?
Dataset statistics reported in Table 1 (Section 3), Table 13 (Appendix B.5), and Figure 4 (Appendix C.1). Training/evaluation splits detailed in Appendix D.1.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Training hyperparameters detailed in Appendix C.2, Table 14. Evaluation protocol in Section 4.2 and Appendix D.

☑ 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?
Results reported as meanstd across 3 independent GPT-5-mini judge runs in Table 2 (Section 5.2). External benchmark results in Table 4.

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.?
Full annotation protocol reported in Appendix B.6: five evaluation dimensions with definitions, 4-point Likert scale with anchors, calibration procedure with 3 example patterns, and materials provided (LLM summaries + ~50 source papers per pattern). No risk disclaimers were required as the task involved only academic evaluation of publicly available literature with no personal data collection.

☑ 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)?
Three graduate students with psychology training at the authors' institution (Fudan University, China) participated voluntarily in the annotation task as part of academic collaboration. No monetary compensation was provided, as the task aligned with their own research interests and constituted a form of peer academic exchange rather than contracted labor; the time commitment was modest (evaluating 80 pattern entries). This arrangement is consistent with norms for voluntary research collaboration among graduate students at the institution. Protocol details in Appendix B.6.

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 study did not use or curate data from human subjects in the sense intended by this question. Annotators evaluated LLM-synthesized pattern summaries against publicly available academic literature; their ratings are part of the research methodology (validation metric), not curated human-subject data. Annotators were fully informed about the purpose and use of their ratings before participating and agreed to their inclusion in the paper.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The study was conducted under the norms of Fudan University's research practices, which do not mandate formal IRB review for low-risk internal academic evaluation tasks. The task involved: (1) voluntary participation by adult graduate students who are colleagues of the authors; (2) evaluation of publicly available academic literature and LLM-generated text, not human-subject data; (3) no sensitive personal information, no deception, no vulnerable populations, and no physical or psychological risk. Under common IRB exemption criteria, such tasks typically qualify as exempt 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?
Use of AI assistants is explicitly documented throughout the paper: Gemini Deep Search for literature retrieval (Section 3.2), Gemini 2.5 Pro and Claude Sonnet 4.5 for pattern data synthesis and scenario generation (Sections 3.2-3.3), and GPT-5-mini as evaluation judge (Section 4.2). Full prompts in Appendix F.