Responsible NLP Checklist

Paper title: *Being Kind Isnt Always Being Safe: Diagnosing Affective Hallucination in LLMs*
Authors: *Sewon Kim, Jiwon Kim, SeungWoo Shin, Hyejin Chung, Daeun Moon, Yejin Kwon, Hyunsoo Yoon*

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?
Ethical Consideration Section, which discusses psychological and demographic limitations and emotional-safety risks.

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

☑ B1. Did you cite the creators of artifacts you used?
Section 3.2 (Reddit dataset), Sections 4.1-4.2 (LLaMA 3.1, Qwen 2.5, Mistral 7B), Appendix C.1 (License).

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
Appendix C.1 (License & Terms of 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)?
Ethical Consideration Section and Appendix C.1 (License & Terms of Use).

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?
All Reddit data were anonymized public text with no PII.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Appendix C (Data and Code Availability).

☑ 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 3.2 describes the number of posts (5,500 total, 500 in AHaBench and 5,000 in AHaPairs).

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.

☑ **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?
Section 4.1 (Experimental Setup), Appendix A (Environment Details)

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix A (Environment Details)

☑ 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 4.2 (Results, Table 2-6)

☑ 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?
Appendix A (Environment Details)

☑ **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.?
Section 3.3 (Response Generation), Appendix I (Human Evaluation Prompt)

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)?
Annotators were study authors and a clinical expert; no external recruitment or payment. (Ethical Consideration Section)

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)?
All Reddit data were public and anonymized. (Ethical Consideration Section)

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No direct human subjects research was conducted. Data were fully anonymized public Reddit posts. (Ethical Consideration Section)

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Ethical Consideration Section

☑ **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?
Sections 3.2, 3.4, and 3.5 describe the use of GPT-4o API for prompt augmentation, evaluation, and preference scoring. The model was used as a research tool, not for writing assistance or authorship.