Responsible NLP Checklist

Paper title: *Subtle Risks, Critical Failures: A Framework for Diagnosing Physical Safety of LLMs for Embodied Decision Making*
Authors: *Yejin Son, Minseo Kim, Sungwoong Kim, Seungju Han, Jian Kim, Dongju Jang, Youngjae Yu, Chan Young Park*

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.

N/A A2. Did you discuss any potential risks of your work?
(left blank)

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

☑ B1. Did you cite the creators of artifacts you used?
Yes. We cite the creators of the artifacts we used, including GPT-4o, iGibson, and BEHAVIOR, in Section 2 (Overview of EMBODYGUARD) and the References.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
Yes. We will release the EMBODYGUARD dataset under a CC-BY 4.0 license and the SAFEL code under an MIT license; both will be publicly available via our project repository.

☑ 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)?
Yes. We confirm that our use of existing datasets is consistent with their intended research use (see Appendix, Ethical Considerations), and we clearly specify the intended research use of our released dataset and code (see Appendix, Code and Dataset Availability).

☑ 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?
Yes. Our dataset does not contain any personally identifiable information or offensive content.

N/A B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
(left blank)

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.

☑ 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?
Yes. We report detailed statistics of the EMBODYGUARD dataset, including the total 942 scenarios and their distribution across risk types, in Section 3.2 (Benchmark Construction).

☑ **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?
Yes. We report the parameter scales of small (8B) and large (70B) models, along with the GPU infrastructure used (RTX 3090/4090, L40S, A6000) in Appendix J.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes. We describe inference settings such as engine (vLLM), precision (bfloat16), and decoding parameters (temperature, top-p, max tokens) in Appendix J.

☑ 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?
Yes. We report descriptive statistics such as model-level averages, error breakdowns, and scenario distributions in Section 5 (Results) and Appendix P.

☑ 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?
Yes. We describe the use of external packages such as iGibson and BEHAVIOR in Section 3.

☑ **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.?
Yes. We report the detailed annotation criteria and reject guidelines provided to human annotators in Appendix M.

☒ 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)

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No. Our work does not involve human subjects research requiring IRB review. The dataset consists of LLM-generated and simulated scenarios without personally identifiable information. Human annotators involved in expert review acted as collaborators, not research subjects, and therefore ethics board approval was not applicable.

☒ N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
(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?
We used AI assistants to refine the writing style and for preliminary coding assistance, as disclosed in the Acknowledgments section.