Responsible NLP Checklist

Paper title: *Multimodal Safety Evaluation in Generative Agent Social Simulations*
Authors: *Alhim Adonai Vera Gonzalez, Carlos Hinojosa, Karen Sanchez, Haidar Bin Hamid, Donghoon Kim, Bernard Ghanem*

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 Ethical Considerations (Page 9)

☑ **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?
This work does not involve the collection or use of real-world personal data. All data used in the study consists of fully synthetic, generated scenarios, plans, and agent interactions. As a result, there is no personally identifying information (PII) present in the dataset, and no anonymization or protection steps are required.

☑ 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?
Relevant dataset statistics are reported in Section 3.1 (Dataset Construction Pipeline), including the number of scenarios (1,000), the structure of the dataset, and details about categories and subcategories. Additional statistics and distributions are provided in the Appendix.

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

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
While the experimental setup is described in Section 4, this work does not involve hyperparameter search or 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?
Descriptive statistics are reported in Section 4 (Experiments), where results are averaged over multiple simulation runs (e.g., 100 runs per model). Figures such as Fig. 6 and Fig. 7 present mean trends and aggregated performance metrics across agents and scenarios.

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.?
While human annotators were involved in verifying the dataset, the full text of the instructions provided to them is not included in the paper. The annotation process is described at a high level in Section 3.1, but detailed guidelines or instruction materials are not reported.

☒ 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)?
While human annotators were involved in dataset verification, the paper does not provide detailed information about recruitment procedures or compensation, nor does it discuss the adequacy of payment relative to participants demographics.

☒ 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)?
This work does not involve real-world personal data. All data consists of synthetic, generated scenarios and interactions; therefore, no consent from human subjects was required.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
This study does not involve human subjects research with personal or sensitive data. The dataset consists of synthetic, generated content, and therefore ethics review board approval was not required.

☑ **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?
Information about the use of AI assistants is provided in the LLM Usage Disclosure section (page 9), where the use of ChatGPT and Grammarly for writing assistance is described.