Responsible NLP Checklist

Paper title: *Spontaneous Giving and Calculated Greed in Language Models*
Authors: *Yuxuan Li, Hirokazu Shirado*

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 8. Ethical Considerations

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

☑ B1. Did you cite the creators of artifacts you used?
References to model providers and related methods are found in: Section 2.2 (LLMs: Reasoning and Non-Reasoning Models); References (see citations for OpenAI, Google, Anthropic, DeepSeek, Together AI, and supporting papers); Details about model API documentation and access are also given in the References and Appendix A (Economic Games Settings).

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
Licensing/terms of use for model APIs are referenced in Appendix A (Economic Games Settings), where it is stated that models were accessed via their respective APIs. Model API documentation is cited in the References section (see OpenAI, 2025, Google, 2025, Anthropic, 2025, Together AI, 2025, "Alibaba Cloud, 2025").

☑ 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)?
Section 3 (Evaluation Framework: Economic Games on Social Dilemmas) and Appendix A describe how model APIs were used as intended (for research and evaluation of language models decision-making capabilities). The study aligns with the intended research use of these APIs and does not extend beyond the scope described by providers.

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?
No human subjects or personally identifiable information were used or collected; all data consists of simulated economic game decisions from language models, with no linkage to individuals. This is evident from the methodology described throughout Section 3 and Appendix A.

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.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Documentation of the games, prompts, and model configurations is detailed in Appendix A (Economic Games Settings), which includes descriptions and full text of all game prompts. Coverage of the language domain (English) is discussed in Section 7 (Limitations), acknowledging linguistic/cultural constraints.

☑ 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?
Experimental statistics are presented in Section 4 and Table 1 (e.g., 100 trials per model-game pair, cooperation and punishment rates, descriptive statistics, standard deviations, etc.). Iterated game details and numbers of runs are given in Section 4.3 and figures.

☑ **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

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

☑ 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 3, 4 and Appendix.

☑ 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?
Section 3, 4 and Appendix.

☒ **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.?
This study does not include human subjects.

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)?
This study does not include human subjects.

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)?
This study does not include human subjects.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
This study does not include human subjects.

N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
This study does not include human subjects.

☒ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

N/A E1. If you used AI assistants, did you include information about their use?
We did not use AI assistants.