Responsible NLP Checklist

Paper title: *The Illusion of Randomness: How LLMs Fail to Emulate Stochastic Decision-Making in Rock-Paper-Scissors Games?*

Authors: *Zihao Guo, Hongtao Lv, Chaoli Zhang, Yibowen Zhao, Yixin Zhang, Lizhen Cui*

> 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?
The final section which is after limitations

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

☑ B1. Did you cite the creators of artifacts you used?
Section 4.1: Experiment Setup

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
Section 4.1: Experiment Setup

☑ 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 4.1: Experiment Setup

☒ 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?
Our research exclusively employs synthetic game-theoretic interaction data generated through controlled simulations of symmetric zero-sum games (specifically Rock-Paper-Scissors and extensions) with unique Nash equilibrium properties. All experimental data derives from model-generated outputs in response to standardized prompts, with no involvement of human-subject datasets, third-party user data, or external corpora.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Those artifact we used is the API from various LLMs service providers, such as OpenAI, Anthropic, Google, Meta, etc., whose documentation and lisence are publicly available and accessible.

☑ 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 4.1: Experiment Setup

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

N/A C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Our research use API from various LLMs service providers.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.1: Experiment setup

☑ 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: Experiment

N/A 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?
Our research use API from various LLMs service providers.

☒ **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.?
(left blank)

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)

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
(left blank)

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?
Throughout this study, AI assistants (specifically GPT-o3 and DeepSeek-R1) were solely utilized for minor text polishing and grammatical refinement of the manuscript. These tools were strictly limited to enhancing the clarity and readability of language in pre-existing content. No AI systems were employed in any research-related activities, including but not limited to: data analysis, experimental design, coding, theoretical reasoning, or decision-making processes. All intellectual contributions, interpretations, and conclusions remain exclusively human-generated. A detailed acknowledgment of AI usage scope will be provided in the final publication's acknowledgments section.