Responsible NLP Checklist

Paper title: *Explainable and Fine-Grained Safeguarding of LLM Multi-Agent Systems via Bi-Level Graph Anomaly Detection*
Authors: *Junjun Pan, Yixin Liu, Rui Miao, Kaize Ding, Yu Zheng, Quoc Viet Hung Nguyen, Alan Wee-Chung Liew, Shirui Pan*

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?
Although this work does not pose any direct risks, we also discuss potential ethical concerns, particularly regarding privacy, bias, and discrimination, in the Ethical Considerations section.

☑ **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?
Our study follows the experimental settings established in prior work on MAS safeguarding, which do not involve any personally identifying information or offensive content.

☑ 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?
We provide the details of the dataset in Section 4.1, Experimental Setups.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Details of the experimental setup are provided in Section 4.1. We carefully ensured that our settings match those of previous works, BlindGuard and GSafeguard. To avoid issues associated with unannounced LLM backend updates and to ensure comparability, we re-ran all baseline models using the latest LLM API. This concern is also discussed in the Limitations section.

☑ 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?
Details of the experimental setup are provided in Section 4.1. We carefully ensured that our settings match those of previous works, BlindGuard and GSafeguard. To avoid issues associated with

unannounced LLM backend updates and to ensure comparability, we re-ran all baseline models using the latest LLM API. This concern is also discussed in the Limitations section.

☒ **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)

☑ **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 only used AI to assist with language. According to the AI Writing/Coding Assistance Policy, this falls under case a: assistance purely with the language of the paper, which does not need to be disclosed.