Responsible NLP Checklist

Paper title: *Why Agents Compromise Safety Under Pressure*
Authors: *Hengle Jiang, Ke Tang*

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?
In Section 9, we explicitly address the risk that our "pressure injection" framework could be viewed as a method for generating adversarial behaviors. We clarify that the intent is defensive red-teaming to expose structural weaknesses. Additionally, Section 8 discusses the risks of deploying agents in high-stakes environments where real-world pressures may exceed simulated ones.

☑ **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?
The datasets used are either established public benchmarks or fully synthetic/simulated scenarios (as detailed in Appendix C) created specifically for this research, which by design do not contain Personally Identifiable Information (PII) of real individuals.

☑ 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?
Appendix C.1 (Dataset Overview and Composition) and Section 5 (Experiments and Analysis).

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

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The paper utilizes pre-trained LLMs (GPT-4o, Gemini 2.5, etc.) in an inference-only setting. The experimental setup, including prompts and baselines, is detailed in Section 4.1, Section 5, and Appendix A, but a traditional hyperparameter search (e.g., learning rate tuning) was not applicable as the models were not fine-tuned.

☑ 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?
Figure 5 presents a scatter plot with sigma ellipses visualizing the distribution of outcomes (Safety vs.

Utility) across episodes, showing the variance beyond simple means. Table 1 reports the aggregated success and adherence rates. Appendix A.1 details the micro-aggregation methods.

☒ **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 utilized AI assistants (e.g., ChatGPT) exclusively for grammatical error correction and to improve the clarity and polish of the manuscript's writing style. All scientific claims, experimental designs, data analysis, and the final vetting of the text remain the original work of the authors.