

Responsible NLP Checklist

Paper title: *Too Good to be Bad: On the Failure of LLMs to Role-Play Villains*

Authors: Zihao Yi, Qingxuan Jiang, Ruotian Ma, Xingyu Chen, Qu Yang, Mengru Wang, Fanghua Ye, Ying Shen, Zhaopeng Tu, Xiaolong Li, Liefeng Bo

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?

Our work primarily involves fundamental research, and we do not perceive any potential risks.

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

- ^{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?

The data we utilize does not involve any personal privacy information.

- 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?

Sec.2 The Moral RolePlay Benchmark

C. Did you run computational experiments?

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

Sec.2 The Moral RolePlay Benchmark

- 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?

Sec.3 Evaluating Moral RolePlay in LLMs

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.?

Human annotators were not involved throughout the entire process.

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.

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)?

Human annotators were not involved throughout the entire process.

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)?

Human annotators were not involved throughout the entire process.

D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

Human annotators were not involved throughout the entire process.

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?

The development and analysis for this project were conducted without the use of AI assistants.