Responsible NLP Checklist

Paper title: *RISER: Orchestrating Latent Reasoning Skills for Adaptive Activation Steering*
Authors: *Wencheng Ye, Xiaoyang Yuan, Yi Bin, Hengyu Jin, Liang Peng, Pengpeng Zeng, Heng Tao Shen*

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?
We discuss potential risks, including the possibility of misuse for steering models toward unsafe behaviors, in Section Limitations. We emphasize that our experiments are restricted to standard public benchmarks and that deployment in safety-critical settings requires additional safeguards.

☑ 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?
We rely only on widely used public benchmarks that do not contain directly identifying personal information to the best of our knowledge. Some datasets (e.g., TruthfulQA) include misleading or harmful statements by design; we use them solely to evaluate factuality and safety, without amplifying or endorsing the underlying content. we discussed it in Implementation Details.

☑ 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 report the number of examples used for vector elicitation, supervised warm-up, and RL training, as well as the benchmarks used for evaluation in section 5.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
we discussed it in Implementation Details. We detail the full training setup, including learning rates, batch sizes, number of epochs, maximum context length, and GRPO configuration.

☑ 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?
we discussed it in Implementation Details. All results are based on random seeds; we report absolute accuracy and average improvements across tasks.

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.

☒ **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 used Claude-3.5 Sonnet as an automatic LLM judge to filter and score candidate reasoning pairs during the vector elicitation stage. The main manuscript text and analysis were written by the authors, with AI models used only as experimental components rather than for drafting or editing the paper. We discussed in section 4.