Responsible NLP Checklist

Paper title: *Seeing but Not Thinking: Routing Distraction in Multimodal Mixture-of-Experts*
Authors: *Haolei Xu, Haiwen Hong, Hongxing Li, Rui Zhou, Yang Zhang, Longtao Huang, Hui Xue, Yongliang Shen, Weiming Lu, Yueting Zhuang*

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?
No, the paper focuses on analyzing internal routing mechanisms and reasoning failures in multimodal MoE models and does not involve direct societal, safety, or environmental risk applications.

☑ **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?
N/A, the datasets used are standard mathematical and scientific reasoning benchmarks (such as MATH500, GPQA-Diamond, MathVerse, and GSM8K-V) that inherently do not contain PII 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?
Yes, Appendix E.1 details the number of test samples and specific subsets used for each benchmark (e.g., 788 samples for MathVerse, 304 for MATH-Vision, 1,319 for GSM8K-V, etc.).

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes, Section 5.1 outlines the implementation framework (vLLM, EasySteer), domain expert thresholds, and intervention layers. Appendix G provides further hyperparameter details, such as temperature (0), maximum generation length (8192 tokens), and repetition penalty configurations.

☑ 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?
Yes, Section 5.1 explicitly states that the results report the average accuracy across 16 trials with greedy decoding.

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?**

☒ 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.?
N/A, no human subjects were used.

☒ 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)?
N/A, no human subjects were used.

☒ 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)?
N/A, no human subjects were used.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
N/A, no human subjects were used.

☑ **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?
Yes, Appendix C details the use of an LLM (gpt-5.2-1211-global) along with its specific prompt template for conducting error analysis. Appendix E.2 details the use of the same model and provides the exact prompt used to generate textual descriptions of diagrams for MATH-Vision.