Responsible NLP Checklist

Paper title: *Role-Sensitive Neurons: A Neuron-Level Gain Control Mechanism for Confidence Steering*
Authors: *Peiwen Huang, Chih-Hao Hsu, Tzu-Hung Huang, Shou-De Lin*

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 the risk of "unwarranted certainty" in the Limitations section and Appendix A.13, where we explicitly warn that amplifying RSNs in knowledge-deficient models induces overconfident but incorrect outputs. This is framed as a safety concern relevant to AI calibration.

☑ **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?
(left blank)

☑ 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?
Dataset size is illustrated in Figure 1 (4,900 divergent answer pairs for RSN identification). All other benchmarks are standard public datasets with statistics reported in their original papers.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Experimental setup and hyperparameters (layer range L1119, steering coefficient , sparsity threshold =0.5%) are detailed in Section 4 and Appendix A.3.

☑ 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 report mean accuracy and abstention rates (E-ratio) aggregated across tasks and domains (e.g., Table 1, Table 2). The primary evaluations use greedy decoding on final-token logits, yielding deterministic results without run-to-run variance. For the self-evaluation experiment on MMLU-Pro, we additionally visualize score distributions via box plots (Appendix A.12), showing the spread across 90 tasks.

☒ **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?
See the Ethics Statement.