Responsible NLP Checklist

Paper title: *Towards Understanding the Robustness of Sparse Autoencoders*
Authors: *Ahson Saiyed, Sabrina Sadiekh, Chirag Agarwal*

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?
Section 9 (Ethics Statement)

☑ 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 use only established public safety benchmarks (HarmBench, Salad-Data, SafeEval, Prompt Injections Benchmark) designed for red-teaming research. These benchmarks are curated to test model safety and do not contain personally identifying information. The prompts are adversarial in nature by design.

☑ 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?
Section 4.1 reports the number of HarmBench batches evaluated (24 batches, approximately 218 prompts per model-condition). Appendix F reports black-box evaluation across 1,500 jailbreak prompts from three datasets. Appendix B (Tables 11, 12) provides full per-model ASR statistics with confidence intervals.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.1 (Attack protocol) details GCG (500 optimization steps, suffix length 20 tokens) and BEAST (beam-search parameters k1=k2=15, search depth L=20). Section 3.4 specifies bfloat16 precision on NVIDIA GPUs. Appendix A (Tables 7, 8) details SAE configurations including layer placement, model dimension, and dictionary width.

☑ 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,

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.

etc. or just a single run?
Table 1 reports median ASR with Mann-Whitney U test p-values. Table 2 reports median ASR with 95% bootstrap confidence intervals. Figure 1 shows error bars representing standard error. Table 18 reports descriptive statistics (mean, standard deviation) with Wilcoxon signed-rank test p-values across n=218 paired optimization runs.

☒ **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?
Section 9 (Ethics Statement): We used AI assistants for assisting with code completion during the preparation of this work.