

Responsible NLP Checklist

Paper title: *The Impact of Off-Policy Training Data on Probe Generalisation*

Authors: *Adrians Skapars, Nathalie Maria Kirch, Samuel Dower, Ekdeep Singh Lubana, Dmitrii Krasheninnikov*

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, because we do not see much risk. Our work is more about calling out risky practices of existing work, rather than proposing a solution that itself could be risky. Additionally, the problems we point out cannot be leveraged by malicious actors.

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?

No, we did not need to process the dataset to check for personal information since the dataset creators already did not include personal information. There was some offensive content as part of our Jailbreak dataset, but this served a purpose in testing the safety profile of LLMs.

- 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, we detail this all in Section 4.3.

C. Did you run computational experiments?

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

Yes, we specify our hyperparameter search method and the best found values in Appendix F (and reference it in Section 4.3)

- 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?

Yes, see all of our bar chart. We use AUROC metric since it is more reliable than accuracy, for classifier performance. 95% confidence intervals are calculated using the analytical standard error formula from Hanley and McNeil (1982).

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.?

(left blank)

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)

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)

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?

(left blank)