Responsible NLP Checklist

Paper title: *Powerful Training-Free Membership Inference Against Fine-Tuned Autoregressive Language Models*
Authors: *David Ili, David Stanojevi, Kostadin Cvejoski*

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?
Ethics Statement section discusses dual-use concerns and defensive motivations

☑ **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?
We use established public benchmark datasets (WikiText-103, AG News, XSum, Enron, PubMed, mC4, Swallow-Code) that have been widely used in prior NLP research. We did not collect or create new data involving individuals.

☑ 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 and Appendix A report dataset sizes (10k members, 10k non-members, 500 validation), sequence lengths (128 tokens), and train/test splits. We also provide full code.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4 and Appendix A (Table 3) report fine-tuning configurations including optimizer, learning rates, batch sizes, epochs, and LoRA parameters for all models

☑ 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?
Section 5 reports AUC and TPR metrics. Figure 2 includes 95% bootstrap confidence intervals (1,000 resamples). Section 4 notes computational resources (single NVIDIA H200 GPU).

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.?
No human annotators or subjects were used.

☒ 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)?
No human annotators or subjects were used.

☒ 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)?
No human annotators or subjects were used.

☒ N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No human annotators or 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?
AI assistants were used for proofreading, code debugging and change suggestion.