Responsible NLP Checklist

Paper title: *Retracing the Past: LLMs Emit Training Data When They Get Lost*
Authors: *Myeongseob Ko, Nikhil Reddy Billa, Adam Nguyen, Charles Fleming, Ming Jin, Ruoxi Jia*

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 1 (Introduction) and Section 6 (Limitations) discuss privacy risks from training-data extraction/regurgitation.

☑ **B. Did you use or create scientific artifacts? (e.g. code, datasets, models)**

☑ B1. Did you cite the creators of artifacts you used?
Sections 2 (Related Work) and 4.1 (Experiment setup) cite prior work and creators of datasets/tools/models used (e.g., Llama family, OLMo, TruthfulQA, WikiQA, InfiniGram).

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We used publicly available research artifacts without redistribution; licenses/terms are not explicitly discussed in the paper.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Sections 3 (Proposed Methods) and Appendix A.2 (Fine-tuning Configuration) describe research-only use of public datasets and open-weight models in a controlled setting; no deployment beyond research contexts.

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)

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4.1 (models evaluated, prompts, metrics) and Appendix A.2 (mismatched SFT dataset construction) document artifacts and their composition.

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.

☑ 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 sample sizes and success-rate metrics; Appendix A.2 reports counts for the mismatched SFT dataset.

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

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We report model sizes by family and hardware (NVIDIA H100) in Appendix A, but we do not report total compute budget (e.g., GPU hours).

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4 (setup), Appendix A.1 (hyperparameters for CIA optimization) and A.2 (fine-tuning configuration: LoRA rank, LR, scheduler, batch size, iters) provide details.

☒ 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 point estimates (percent attack success) over fixed prompt sets but do not provide confidence intervals/error bars or multi-seed variance.

☒ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?
We cite tools (e.g., InfiniGram) and methods, but do not enumerate package versions or all preprocessing/evaluation parameter settings in the paper.

☒ **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)

N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
(left blank)

☒ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

N/A E1. If you used AI assistants, did you include information about their use?
(left blank)