Responsible NLP Checklist

Paper title: *Precise In-Parameter Concept Erasure in Large Language Models*
Authors: *Yoav Gur-Arieh, Clara Haya Suslik, Yihuai Hong, Fazl Barez, Mor Geva*

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 have an ethical considerations section at the end

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

☑ B1. Did you cite the creators of artifacts you used?
Cited in Section G in the Appendix

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We will include licensing details in the public GitHub repository where our code and data will be released. In our work we used publicly available models and datasets with permissive licenses.

☒ 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)?
For each artifact we used, we ensured our work aligned with its intended use. Licensing information will be available in the public GitHub repository where we release our code and data. For the artifacts created, namely the code attached under "Software", we will include a license and clear README.md once we make this public on Github.

☒ 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?
The results and data in this paper and the datasets we used are general and do not include personal content. We do however have concepts we anlayze which are sensitive, but they are not personal and are at the conceptual level. Questions generation about said topics is detailed in Appendix section D.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
In Appendix section G.

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?
In section 5 and Appendix sections A and B we discuss the number of examples and details of validatation/test split of the data.

☑ **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?
In section 5 and Appendix section G.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In section 5 and Appendix sections A and B.

☑ 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 this in section 5, as well as in the main results in table 1, which includes confidence intervals which are further explained in its caption

☒ 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?
To the best of our knowledge, the results reported in the paper are not sensitive to specific implementation choices or parameter settings, aside from those explicitly detailed in Section 5.

☑ **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.?
Discussed in Appendix section C and Figure 8.

☒ 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)?
We relied on fellow graduate students, therefore no recruitment or compensation were needed.

☑ 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)?
Discussed in Appendix section C.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
There was no need to.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix section C.

☑ **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?
Appendix Section G.