Responsible NLP Checklist

Paper title: *DRIV-EX: Counterfactual Explanations for Driving LLMs*
Authors: *Amaia Cardiel, Eloi Zablocki, Elias Ramzi, Eric Gaussier*

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 discuss possible risks in Section "Ethical Considerations

☑ **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?
In section "E. Appendix: License and terms of use", we mention that the data we use contains no personally identifying information as well as our verification process.

☑ 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 Appendix "B.2: Finetuning driving LLMs", we give statistics on the highD dataset that we use. We also provide statistics on the highD subsets that we use (cf Table 10 in "D. Appendix: Tuning candidate algorithms to the task" and Table 3 in "4.3.1 Retrieving injected biases".)

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Experimental setup is detailed in B. Appendix: Experimental details regarding finetuning and evaluation, including a subsection "B.4 Experimental compute budget", dedicated to describing the compute budget of our experiments. Hyperparameter search and best found values are displayed in Table 11 in section D.3 Exploration of hyperparameters.

☒ 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?
The compute heavy nature of our experiments (cf "B.4 Experimental compute budget") made it too costly for us to run each experiment more than once. Furthermore, our method, DRIV-EX, is not stochastic and its performance does not depend on random seeds.

☒ **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?
Open source LLMs were only punctually used to rephrase or shorten already written text. None were used to generate ideas nor to contribute code to the project.