Responsible NLP Checklist

Paper title: *Measuring the Effect of Disfluency in Multilingual Knowledge Probing Benchmarks*
Authors: *Kirill Semenov, Rico Sennrich*

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?
Within the field of knowledge analysis of LLMs, the most potential risks are seen in the knowledge editing approaches. Our research tackles evaluation of already existing models and uses the already retrieved neutral facts, thus we do not see significant risks of the current work.

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

☑ B1. Did you cite the creators of artifacts you used?
Since our paper is tackling comparison of the approach of the existing benchmark to its updated version, we cite the authors in all sections. The first elaborate explanation is provided in Sections 1 (Introduction) and 2 (Related Work).

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We are using the existing benchmark, which is available publicly, according to CC-BY-NC 4.0 license which was stated by the benchmark.

☒ 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)?
We are using the existing benchmark and our modification of it for the same purpose as the authors of that benchmark: for the factual evaluation of the pretrained language 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?
All data was retrieved from publicly available resource, Wikidata, and tackles neutral information such as capitals of countries, locations of headquarters of organizations, initial languages of pieces of art, etc.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
The most detailed description is provided in Appendix B.

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?
The most detailed description is provided in Appendix B.

☑ **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?
The information about the model size, GPU type and hours is provided in Section 3 (main experiment) and Appendix D (additional experiments).

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We used the default parameters of the LLaMA-2-7B model only.

☑ 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 information is provided in Sections 3, 4 and in Appendix C.

☒ 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?
The package parameters and model specifications will be available in the publicly available Github repository (the link is provided in the paper).

☑ **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.?
The human annotators were the colleagues from the NLP and linguistic fields, fluent in particular languages. They helped us in creation of the multi-shot prompts for ChatGPT and in interpretation of the results. Since different languages had differing grammatical and stylistic features, we organized our work through in-person talks, thus there was no fixed text of instructions. The resulting annotated data will be published with the code after the anonymity period.

☐ 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)?
We did not resort to help of crowdsourcing platform, students or paid participants.

☐ 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?
Since we did not work with any sensitive or personal data, nor our experiments were implying any new technological artifacts, we did not apply for an ethics review board

☐ 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?**

☑ E1. If you used AI assistants, did you include information about their use?
ChatGPT was used for creation of one of the alternative modes of the dataset; this is discussed in detail (and provided with prompt examples) in Section 3 and Appendix A. Additionally, we used the Jetbrains (PyCharm) AI assistant for autocompletion of the parts of the code, and ChatGPT for helping with prettifying the basic functions of the Huggingface interface (all examples can be found in Huggingface documentation). This was done due to the first experience of work with Huggingface for one of the co-authors. All generated pieces of code were then checked manually. We did not include information about that in the paper since the scope of use of the assistants was narrowed down to speeding up the process of code writing, while the main principles (and the main body of code) of the dataset creation, modification, experiment running and analysis of results were created by authors, and was not used for "creative" purposes such as suggesting new approaches. Finally, we used Writeful assistant for grammar checking of the text of the final paper.