Responsible NLP Checklist

Paper title: *From KMMLU-Redux to Pro: A Professional Korean Benchmark Suite for LLM Evaluation*
Authors: *Seokhee Hong, Sunkyoung Kim, Guijin Son, Soyeon Kim, Yeonjung Hong, Jinsik Lee*

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.

N/A A2. Did you discuss any potential risks of your work?
We do not have potential risk in our work.

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

☑ B1. Did you cite the creators of artifacts you used?
In terms of data, we note the original KMMLU source in section1 for KMMLU-Redux and note the original source of KBL, KorMedMCQ in Table 1 which are part of KMMLU Pro. In terms of evaluation code, we refer the source code in section 4. All the references of evaluated models are noted in Table 2 and 7.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We note the license in Ethical Statements.

☑ 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 note the existing artifacts and our license in Ethical Statements.

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 do not have any personally identifying info or offensive content.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We note the coverage of domain in Table 1 and Appendix C.

☑ 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?
We note in section 2.2.3 and 3.3.

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.

☑ **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 note in Section 4.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We note in Section 4.

☑ 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 have our results in section 5.

N/A 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 refer the source code and settings in section 4.

☑ **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.?
We attach screenshots of the instructions in 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 report the hired workers payment in Appendix D.2.

☑ 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)?
We describe in Appendix D.1.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We do not contain data that ethics review requires.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We report the hired workers demographic in Appendix D.2.

☑ **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?
We use AI assistants to fix grammatical errors in writing.