Responsible NLP Checklist

Paper title: *EMNLP: Educator-role Moral and Normative Large Language Models Profiling*
Authors: *Yilin Jiang, Mingzi Zhang, Sheng Jin, Zengyi Yu, Xiangjie Kong, Binghao Tu*

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 8

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

☑ B1. Did you cite the creators of artifacts you used?
Appendix B

N/A B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
N/A

☑ 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)?
The Introduction (Section 1) and the EMNLP Framework Construction (Section 3) clearly define the intended use of the created artifacts (the EMNLP benchmark) for research and evaluation purposes. The Ethics Statement (Section 8) clarifies that the models were used in a controlled research setting, consistent with the intended use for research purposes.

☑ 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?
Section 8 (Ethics Statement) states that all human data collected for the benchmark were anonymized and aggregated.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Appendix A.2 provides a detailed breakdown of the moral dilemma inventory. The methodology in Section 4.1 specifies the languages used for different experiments.

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 3 and Section 4.1 report the number of dilemmas (88) and prompts created, along with the total number of experimental runs. Table 5 in Appendix A provides a categorical breakdown of the dilemmas.

☑ **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?
Appendix B (Table 9) reports model parameters where available. Appendix A.3 details the computing infrastructure. The total computational budget (e.g., GPU hours) is not reported.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.1 details the experimental setup for each research question. Appendix A.4 details the decoding parameters and hyperparameters used.

☑ 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?
Section 4.2 details the evaluation metrics and scoring protocols (Mode, Mean, Proportions). The results in Section 5 and the appendices present summary statistics. Appendix E provides detailed statistical significance testing (p-values, effect sizes).

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?
N/A

☑ **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.?
Appendix C.1 and C.2 provide the full text of the instructions given to the expert annotators for the moral stage and harmfulness rating tasks.

☑ 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)?
Appendix C reports the payment amount for expert annotators ($2.5). Appendix D describes the professional background of the annotators and the demographic diversity of the teacher sample. The specific recruitment method and a discussion of payment adequacy are not included.

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)?
N/A

N/A D4. Was the data collection protocol approved (or determined exempt) by 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?
Appendix D.1 reports the professional and academic background of the expert annotators. Appendix D.2 describes the professional characteristics (teaching level, experience) and geographic diversity of the teacher sample used for benchmarking.

☑ **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?
Section G (Use Of AI Assistants) states that GPT-4o was used to polish the language in the appendix, and clarifies that it was not used for generating scientific content or experimental results.