Responsible NLP Checklist

Paper title: *TLUE: A Tibetan Language Understanding Evaluation Benchmark*
Authors: *Fan Gao, Cheng Huang, Yutong Liu, Nyima Tashi, Xiangxiang Wang, Thupten Tsering, BAN Ma-bao, RENZENG Duojie, Gadeng Luosang, Rinchen Dongrub, Dorje Tashi, XiaoFengCD, Yongbin Yu, Hao Wang*

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 potential risks in Section 7, noting that TLUE includes safety-critical content such as ethics and bias. However, all data was carefully curated and reviewed by native speakers and domain experts to ensure cultural appropriateness. The primary risk lies in possible misuse of model outputs in sensitive domains, not from the dataset itself.

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

☑ B1. Did you cite the creators of artifacts you used?
Sections 3 provide detailed descriptions of TLUE and its two sub-benchmarks. All data and evaluation scripts are publicly released.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
See Section 6 and Appendix A. We state that the TLUE benchmark will be released for academic research under a non-commercial license. We also acknowledge and respect the original licenses of the source datasets (e.g., CMMLU, SafetyBench), which are cited in the paper.

☑ 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)?
See Section 3.1 and References. We only use publicly available datasets (e.g., CMMLU, MMLU, SafetyBench) under terms consistent with their intended use. All adaptations are restricted to academic evaluation purposes. The newly created TLUE benchmark also specifies intended use for research only.

☒ 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 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.

See Section 8 (Ethics Statement). We confirm that TLUE does not contain any personally identifying or offensive content. All questions were manually reviewed by native Tibetan speakers and domain experts to ensure cultural appropriateness and neutrality.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
See Section 3.2 and Appendix A. We provide detailed documentation on the coverage of TLUE, including the number of questions, domains, subcategories, evaluation settings (zero-shot/few-shot), and category descriptions for both Ti-MMLU and Ti-SafetyBench.

☑ 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 report the total number of evaluation examples (22,963), along with a breakdown across sub-benchmarks (e.g., 11,528 examples in Ti-MMLU across 67 subjects) in Section 3.2. Further statistics, such as category coverage and subtask distributions, are detailed in Appendix A. TLUE is an evaluation-only benchmark without train/dev/test splits.

☑ **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?
Section 5

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix C

☑ 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?
Appendix C

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?
Our benchmark construction did not involve any evaluation or preprocessing tools such as NLTK, SpaCy, or ROUGE. All data processing steps were manually conducted or implemented through custom scripts.

☑ **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.?
We collaborated with two Tibetan language experts and fifteen domain annotators to ensure linguistic and cultural fidelity. Since they were part of a coordinated academic collaboration rather than crowdworkers or paid participants, no standardized instruction template was provided, and no formal risk disclosure was necessary.

☑ 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 describe our recruitment of two Tibetan language specialists and five annotators in Section 3.1. Annotators were compensated at an hourly rate of 28 USD, which is above the local professional rate, to ensure fairness and incentivize quality linguistic review.

☑ 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)?
All annotators were recruited through academic collaboration and were fully informed of the purpose, scope, and intended use of the data. As stated in Section 3.1, participants were compensated fairly and voluntarily contributed under explicit agreement, ensuring informed consent.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No personal data was collected, and all annotators were paid professionals who gave informed consent. The task involved no medical, psychological, or personally sensitive information. Thus, IRB approval was not applicable.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
See Section 3.1. We recruited 2 Tibetan language experts and a team of 5 annotators, all native Tibetan speakers from relevant regions of China. Their expertise ensured linguistic and cultural fidelity in the dataset curation.

☒ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

N/A E1. If you used AI assistants, did you include information about their use?
(left blank)