Responsible NLP Checklist

Paper title: *QualBench: Benchmarking Chinese LLMs with Localized Professional Qualifications for Vertical Domain Evaluation*
Authors: *Mengze Hong, Wailing Ng, Chen Jason Zhang, Di Jiang*

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?
(left blank)

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

☑ B1. Did you cite the creators of artifacts you used?
This paper leverages publicly available qualification exam papers, with the specific exam names provided in Table 11.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We discuss the liscence and ethical use of presented benchmark data in Ethical Consideration section.

☑ 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)?
This paper leverages publicly available qualification exam papers, which is intended for knowledge evaluation and practice.

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?
The dataset consists only of publicly available information (e.g., facts, historical events) and contains no personal identities or offensive content.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
The dataset statistics is presented in Table 3, and the intended use is provided in open-sourced GitHub repository.

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 statistics of the dataset is provided in Section 3.3, and the demonstration of dataset QA is presented in Appendix A and 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 experimental details are provided in Section 4.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The experimental details are provided 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?
The experimental results are provided in Section 5.

☑ 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 code implementation is provided in open-sourced GitHub repository.

☑ **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 instruction to human annoatator is brieftly mentioned in Section 3.2.

☑ 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)?
The recruitment was conducted internally, with industry professionals invited to participate in data quality assurance.

☑ 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)?
The data annotation serves for the purpose of research use only, with a detailed ethical statement presented in Ethical Consideration section.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
This paper raises no ethical concerns.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
The information is provided in Section 3.2.

☑ **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?
The use of AI assistants is limited to minimal involvement, primarily for minor proofreading to enhance writing quality.