

Responsible NLP Checklist

Paper title: *K-MetBench: A Multi-Dimensional Benchmark for Fine-Grained Evaluation of Expert Reasoning, Locality, and Multimodality in Meteorology*

Authors: *Soyeon Kim, Cheongwoong Kang, Myeongjin Lee, Eun-Chul Chang, LEE JAEDEOK, Jaesik Choi*

How to read the checklist symbols:

- the authors responded 'yes'
- the authors responded 'no'
- 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?
Yes See the Ethical Considerations section, and Sections 6.16.3.

B. Did you use or create scientific artifacts? (e.g. code, datasets, 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?
Yes See the Ethical Considerations section.

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?
Yes See Section 3.1 and Tables 23.

C. Did you run computational experiments?

C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes See Section 4.1 and Appendix D.1.

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?
Yes See Table 5, Figure 4, and Appendix H.1H.2.

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.?
Yes See Appendix C.6 and C.9, as well as Tables 11 and 12.

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. ACL 2026 used a subset of ARR checklist form.

- 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)?

Yes See the Ethical Considerations section.

- 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 The dataset is derived from publicly released examination materials rather than personal data collected from individuals, and the paper states that the dataset contains no personally identifiable information. The consent question is therefore not directly applicable here.

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

N/A The paper does not describe human-subject experimentation or collection of personal/sensitive human data requiring IRB review; rather, it reports expert verification and scoring of benchmark materials. No IRB approval or exemption is discussed in the manuscript.

- 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?

Yes These uses are described in Sections 3.1, 3.3, 3.4, Appendix C.1C.9, and Appendix D.2.