Responsible NLP Checklist

Paper title: *KRETA: A Benchmark for Korean Reading and Reasoning in Text-Rich VQA Attuned to Diverse Visual Contexts*
Authors: *Taebaek Hwang, Minseo Kim, Gisang Lee, Seonuk Kim, Hyunjun Eun*

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?
No; our benchmark uses copyright-free/public materials and our own field photos with non-identifiable content, and we discuss scope/coverage limits rather than safety risks in the Limitations section.

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

☑ B1. Did you cite the creators of artifacts you used?
Yes; we cite the creators of all third-party artifacts (e.g., VLMEvalKit, PaddleOCR, prior benchmarks/models) in References and describe our released code/data in the Abstract and Appendix.

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
No; we did not specify licenses in the paper text, but we will add explicit open licenses (e.g., CC-BY for data, MIT for code) for our dataset and code on GitHub and Hugging Face

☑ 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)?
Yes; we only used prior models, and tools (e.g., VLMEvalKit, PaddleOCR) in accordance with their intended research use, and our released benchmark is explicitly intended for research purposes 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?
No; our dataset consists of copyright-free/public images and our own field photos with non-identifiable content, and we manually filtered to exclude any personal or offensive material.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?

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.

Yes; Section 3 provides documentation of KRETA including domain coverage (15 domains), image types (26 categories), and reasoning levels (System 1 vs. System 2).

☑ 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; Section 3.1 reports dataset statistics including 2,577 total samples with 1,426 System 1 and 1,151 System 2 QA pairs, along with domain and image-type distributions.

☑ **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?
Yes; Tables 24 report the sizes of the evaluated models (ranging from 0.5B to 14B parameters), and we describe the evaluation setup using existing models without additional large-scale training cost.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes; Section 4 describes our evaluation setup, including model configurations, prompting strategies (baseline vs. Chain-of-Thought), and details of the seven-metric evaluation protocol.

☑ 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; Tables 24 report accuracy across multiple models, domains, and image types, showing aggregated results rather than single runs to provide transparent comparisons.

☑ 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?
Yes; we report the use of external packages such as PaddleOCR for text filtering and VLMEvalKit for evaluation, following their recommended settings without additional modifications.

☑ **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; authors followed a detailed annotation guideline (Human Annotation Refinement in Section 3.3), including checks for text utilization, alignment, inferential complexity, and review of grammar, factual correctness, and clarity.

☒ 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)?
No; no external annotators were recruited or paid, as all annotation work was performed by the authors.

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; our dataset is built from copyright-free/public materials and author-collected field photos, without involving personal data requiring consent.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No; an ethics review board approval was not required since the dataset was compiled from copyright-free/public sources and author self-annotation without human subject participation or personal data.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No; all annotations were conducted directly by the authors themselves, so no demographic or geographic information about external annotators was applicable.

☑ **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; AI assistants were used only for text polishing and presentation refinement, not for core research tasks such as dataset construction, experiments, or analysis.