Responsible NLP Checklist

Paper title: *Unmasking Deceptive Visuals: Benchmarking Multimodal Large Language Models on Misleading Chart Question Answering*
Authors: *Zixin CHEN, Sicheng Song, KaShun SHUM, Yanna Lin, Rui SHENG, Weiqi Wang, Huamin Qu*

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?
In section 1, we discussed how misleading charts can be wrongly used to manipulate perception and mislead audiences, leading to significant consequences.

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

☑ B1. Did you cite the creators of artifacts you used?
In section 2.1, 2.2 and section 5 and appendix A.1, we provide detailed citations to all the artifacts we used in our paper.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
In section 2.1, 2.2 and section 5 and appendix A.1, we include the discussion of the license and terms of the artifacts we used in our 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)?
In section 1, section 2.2, 2.3, section 5 and appendix A.1, we include the discussion of the artifacts we used in our paper consistent with their intended use.

☑ 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?
In section 2, we discuss that we didn't collect any data contains personally identifying Info.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
In the section 2 as well as the Appendix A.7 and A.8, we provide the documentation of artifacts.

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?
In the section 2.4, section 3 as well as the Appendix A.7 and A.8, we provide the statistics for data.

☑ **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?
In the section 3.1, 3.2, table1 and Appendix A.6, we provide the detailed computational experiments, model size and budget.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In the section Appendix A.6, we provide the detailed experiment setups.

☑ 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?
In section 3, table 1, table 2 and figure 5, we provide detailed descriptive statistics for our results.

☑ 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?
Privately revealed to ACL ARR 2025 May Program Chairs, ACL ARR 2025 May Submission422 Area Chairs, ACL ARR 2025 May Submission422 Authors, ACL ARR 2025 May Submission422 Reviewers, ACL ARR 2025 May Submission422 Senior Area Chairs In the Appendix A.6 and section 3.1, we provide parameters for packages we used.

☑ **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.?
In section 2.4 and Appendix A.4, A.5 we discussed the procedure that we instruct the human annotators to conduct the work.

☑ 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)?
In section 2.4 we discussed the recruitment and payment we provide to the participants.

☑ 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)?
In section 2.4 and Appendix A.4, A.5 we discuss the data consent when recruiting human annotators

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We don't need to collect any personal data that has any ethics issues for this study.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
In section 2.1, 2.4 and and Appendix A.4, A.5

☒ **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?
We did not use.