Responsible NLP Checklist

Paper title: *ClimateViz: A Benchmark for Statistical Reasoning and Fact Verification on Scientific Charts*
Authors: *Ruiran Su, Jiasheng Si, Zhijiang Guo, Janet B. Pierrehumbert*

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?
The Ethics Statement section discusses potential risks. This section acknowledges several risks associated with our work, including: Automation Bias: The risk that over-reliance on a system trained on CLIMATEVIZ, even with high performance, could lead to the amplification of misinformation if used without human oversight. Adversarial Vulnerability: The possibility that malicious actors could develop new manipulation strategies not covered in the benchmark to bypass detection. Dual-Use: The concern that the models designed to verify claims could be repurposed to generate plausible but false claims, thereby accelerating disinformation campaigns.

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

☑ B1. Did you cite the creators of artifacts you used?
Yes, creators of all used artifacts are cited. The primary artifacts used (the baseline models) are described and cited in Section 4.2 Baselines. Other artifacts and evaluation metrics are cited throughout the methodology in Section 3 and Section 4. The primary contribution of our paper is the creation of the CLIMATEVIZ dataset and the associated code.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
Yes, the licenses for the created artifacts are specified in the Ethics Statement 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)?
Yes, the Ethics Statement specifies the intended use for the created artifacts.

☑ 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?
Our paper clarifies that the nature of the collected data precluded the presence of personally identifiable information. The Ethics Statement explicitly states: The charts were sourced from

"publicly accessible, reputable scientific institutions, and no proprietary or confidential data was used". Crucially, it affirms that " No personally identifiable or sensitive information was collected".

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Yes, the artifacts are documented across several sections, primarily in Section 3.2 (Dataset Analysis) and Appendices B and D.

☑ 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, dataset statistics are provided in Section 3.2.1, and the train/dev/test split is detailed in Section 4.2.

☑ **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 total computational budget is listed in Appendix F Experiments.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Our paper thoroughly describes the main experimental setup, including the different input and output settings, the models used, and the evaluation protocol . For the few-shot experiments, the prompts and exampleswhich act as the primary "hyperparameters" in this contextare fully documented in Appendix F.

☑ 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 results in Section 5 (Tables 5 and 6) report performance metrics from a single evaluation run on the held-out test set.

☑ 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 paper cites the original works for the evaluation metrics used (e.g., BLEU, ROUGE-L, BERTScore) in Section 4.1.

☑ **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, the full text of the instructions provided to the annotators for all three annotation tasks is reported in Appendix A, specifically in Section A.2.2.

☑ 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)?
Section 3.1 that participants were recruited through the citizen science platform Zooniverse.

☑ 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)?
Yes, this is discussed in the Ethics Statement section.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Ethics Statement confirms that the source data is publicly accessible and that annotators provided informed consent voluntarily. The Committee of Zooniverse committed an ethics review before full launch of the project.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
According to Zooniverse, the citizen science platform where annotators are from, we should not acquire information about the characteristis of the annotators.

☑ **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, the use of AI models as a core part of the research methodology is detailed in the paper. Specifically, the use of GPT-4o for generating refuted claims is described in Section 3.1.2 , and its use for knowledge graph construction is described in Section 3.1.3.