Responsible NLP Checklist

Paper title: *Causal Evidence Extraction and Triangulation in Crisis Reports using Large Language Models: A ReliefWeb-based Study*
Authors: *Yuanjun Zhang, Mourad Oussalah*

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?
See the Ethical Considerations section, which discusses potential risks related to reporting biases, incidental mentions of vulnerable populations in humanitarian reports, and the appropriate use of the system as decision support rather than causal or operational evidence.

☑ **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?
We use publicly available ReliefWeb reports, which are curated before publication. We do not release raw report text as a new dataset and mitigate privacy/offensive-content risks by reporting only aggregated statistics and limited supporting snippets, with caution noted in 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?
See the Data section and Experimental Setup, which report the number of documents, annotated instances, and evaluation subsets used in the experiments.

☑ **C. Did you run computational experiments?**

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
See LoRA fine-tuning details, which describes the training setup and hyperparameters.

☒ 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?
Results are reported for fixed model configurations for LoRA fine-tuning. Due to computational cost, we do not report multi-run summary statistics such as mean/std or error bars.

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.

☑ **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.?
Annotations were conducted by a small number of domain experts on publicly available humanitarian reports. The annotation task did not expose participants to offensive content or collect personal identifying information, and posed minimal risk. Therefore, the full instruction text was not included in the paper.

☒ 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 annotations were performed by a small number of domain experts (not via a crowdsourcing platform), and no participant compensation was provided. We therefore did not include recruitment/payment details in the paper.

☑ 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)?
Expert annotators participated as part of a research collaboration and were informed that their labels would be used for research evaluation and comparison with model outputs. No personally identifying information was collected.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The work involved a small number of expert collaborators, posed minimal risk, and we did not collect personally identifying information.

☑ **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?
AI assistants were used for language polishing. We did not include a dedicated statement in the paper; however, all experimental design, data annotation, analysis, and scientific conclusions were developed and verified by the authors.