Responsible NLP Checklist

Paper title: *DiZiNER: Disagreement-guided Instruction Refinement via Simulating Pilot Annotation for Zero-shot Named Entity Recognition*
Authors: *Siun Kim, Hyung-Jin Yoon*

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?
As our work strictly focuses on a methodological framework for zero-shot NER using existing open-source LLMs and public benchmarks without deploying user-facing applications or handling sensitive data, we do not foresee any direct societal risks or dual-use concerns.

☑ 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?
This study exclusively utilizes well-established, publicly available benchmark datasets for Named Entity Recognition. As we did not collect any new data, we relied on the standard preprocessing and anonymization measures previously established by the original creators of these public corpora, and therefore did not include a separate discussion on identifying PII or offensive content.

☑ 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. Detailed statistics for all 18 datasets used in our experiments, including the number of examples across train/dev/test splits, number of entity types, and average tokens/entities, are comprehensively reported in Table 4 within Appendix B.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes. The detailed experimental setup, including deterministic decoding hyperparameters (e.g., temperature 0.0, top-p 1.0, and maximum output length) is described in Section 4.1 (Settings - Backbones and Implementation). Furthermore, the specific tuning parameters governing the instruction refinement process, along with the three distinct hyperparameter configurations explored, are detailed comprehensively in Appendix A and Table 3.

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.

☑ 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. We report average F1 scores along with their minimum and maximum values across multiple configurations in Tables 9 through 16. Furthermore, we report standard deviations from a sensitivity analysis across five random seeds in Section 4.2 (Main Results).

☒ **D. Did you use human annotators (e.g., crowdworkers) or research with human subjects?**

N/A 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.?
(left blank)

N/A 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)?
(left blank)

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)?
(left blank)

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
(left blank)

☑ **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?
We have explicitly disclosed the specific AI tools used for proofreading and code generation, along with our assumption of full responsibility, in the Acknowledgments section.