Responsible NLP Checklist

Paper title: *PMPO: Probabilistic Metric Prompt Optimization for Small and Large Language Models*
Authors: *ChenZhuo Zhao, Ziqian Liu, Xinda Wang, Junting Lu, Chaoyi Ruan*

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.

N/A A2. Did you discuss any potential risks of your work?
This work focuses on prompt optimization using loss-based evaluation and does not involve sensitive data, human subjects, or downstream applications that directly raise ethical concerns. Therefore, potential risks were not discussed explicitly.

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

☑ B1. Did you cite the creators of artifacts you used?
Section 4(Experiment)

N/A B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
All artifacts used in this work (models, datasets, and tools) are publicly available under their respective open-source or permissive licenses; therefore, no additional licensing discussion is necessary.

N/A 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)?
All artifacts used in this work were employed in accordance with their publicly stated intended use. No modifications or derivative uses beyond standard research practices were introduced, and no new artifacts requiring usage specification were created.

N/A 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?
The work relies solely on publicly available benchmark datasets that are widely used in the community and do not contain any personally identifying information or offensive content. No additional data was collected.

N/A 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.

All artifacts used in this work are standard public benchmarks with existing documentation. No new datasets or artifacts were created that would require additional documentation.

☑ 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 Section 4.1 Experiment Settings, which provides details on dataset usage and train/test splits.

☑ **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?
See Section 4.1 Experiment Settings, which describes the model sizes (e.g., 0.5B32B parameters), computing infrastructure (single NVIDIA H800 GPU), and runtime per optimization run.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
See Section 4.1 Experiment Settings, which details the experimental setup, including hyperparameters such as the number of iterations, top- samples, number of prompt variants, and preference scaling factor .

☑ 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?
See Section 4.2 Experimental Results and Analysis, which reports summary statistics such as average accuracy across multiple tasks and models in Table 1, Table 2, and Table 3.

☑ 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?
See Section 4.1 Experiment Settings, which details the models and parameter settings used. No external evaluation packages were used; all results are based on internal model metrics.

☒ **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.?
This study does not involve human participants or annotators.

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)?
This study does not involve human participants or annotators.

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)?
This study does not involve human participants or annotators.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
This study does not involve human participants or annotators.

N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
This study does not involve human participants or annotators.

☒ **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?
(left blank)