Responsible NLP Checklist

Paper title: *Escaping the Sisyphus Dilemma: Experience Replay for Robust Text-to-Optimization Modeling*
Authors: *Wantong Xie, Yinghao Chen, Yi-Xiang Hu, Feng Wu, Jieyang Xu, Sijia Zhang, Xiangyang Li*

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?
We discuss potential risks related to safety in critical infrastructure and data privacy in the "Ethical Considerations" section.

☑ 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?
The datasets used in this work (NL4Opt, MAMO-Easy, OptiBench, ComplexOR, ComplexLP, OptMATH, and IndustryOR) are established public benchmarks for mathematical optimization. They consist of synthetic or anonymized problem descriptions focused on logical reasoning and do not contain personally identifiable information 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?
Section 4.1 (Benchmarks) describes the seven datasets used for evaluation. Detailed statistics regarding the number of samples and difficulty levels are provided in Section 4.1 and the Appendix.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.1 (Experimental Setup) details the backbone model (DeepSeek-V3), random seed (42), and external solver (SCIP). Section 4.4 and Table 1 discuss the effect of the key hyperparameter.

☒ 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?
Due to the high computational cost of long-context reasoning and API interactions, we report results from a single run with a fixed random seed (42) to ensure reproducibility, as stated in Section 4.1. We provide average efficiency metrics across datasets in Table 2.

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?**

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 work relies entirely on automated evaluation using Large Language Models and external solvers on publicly available benchmarks. No human annotators or human subjects were involved in the experiments.

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)?
Not applicable. This study is purely computational and did not involve the recruitment or payment of any human participants.

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)?
Not applicable. We utilized previously published, open-source datasets (NL4Opt, OptMATH, etc.) which are standard benchmarks in the field. As stated in Appendix 4.3, these datasets are available under standard licenses (MIT or CC-BY-SA 4.0) and do not contain personal data requiring individual consent.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. This research focuses on computational methods using existing public benchmarks and did not involve any human subjects or new data collection protocols requiring ethics board approval.

☑ **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 employed Large Language Models (DeepSeek-V3, GPT-4o) as the primary subjects of our experiments, as detailed in Section 4. Additionally, we used general-purpose AI assistants (e.g., ChatGPT) solely for minor language polishing and checking LaTeX syntax. All scientific claims and ideas are our own.