Responsible NLP Checklist

Paper title: *GOBench: Stage-Wise Diagnostics and the Visual Paradox in Multimodal Graph Optimization*
Authors: *Yinghao Chen, Wantong Xie, Shuli Zeng, Sijia Zhang, Xiaotian Pan, Feng Wu, 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?
Yes. See the "Ethics Statement" 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 dataset consists entirely of procedurally generated synthetic instances (described in Section 3.1) for abstract graph optimization tasks. It does not involve the collection of real-world data or human subjects, rendering PII checks inapplicable.

☑ 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. See Appendix B.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes. See Section 4, Appendix C, D.

☑ 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. See Section 5, Appendix E, H.

☒ **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 evaluates AI models using procedurally generated synthetic data and solver-derived oracles. It does not involve any human participants, crowd-workers, or manual annotators.

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.

☒ 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)?
As stated in D1, this study does not involve any human participants or crowd-workers. The evaluation is fully automated using synthetic data and solver oracles.

☑ 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)?
The dataset is composed entirely of synthetic instances procedurally generated via algorithms (described in Section 3.1). No human data was collected or curated, so data consent is not applicable.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The dataset is composed entirely of synthetic instances procedurally generated via algorithms (described in Section 3.1). No human data was collected or curated, so data consent is not applicable.

☑ **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. See "Ethics Statement" Section.