Responsible NLP Checklist

Paper title: *AOT*: Efficient Synthesis Planning via LLM-Empowered AND-OR Tree Search*
Authors: *Xiaozhuang Song, Xuanhao Pan, Xinjian Zhao, Hangting Ye, Shufei Zhang, Jian Tang, Tianshu Yu*

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?
Our work is a beneficial application for drug discovery using public chemical data. No additional risks beyond those of the underlying LLM APIs.

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

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?
Our data consists solely of chemical molecules and reactions (SMILES representations). No personally identifying information or offensive content is possible in this data type.

☑ 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 3.1 reports dataset sizes (USPTO-Easy: 200, USPTO-190: 190, Pistachio Reachable: 150, Pistachio Hard: 100 molecules). Appendix B.1.1 (Table 5) provides detailed statistics.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3.1 (Implementation Details) and Appendix B.1.3 report hyperparameters. Section 3.5 and Appendix C.2 provide hyperparameter sensitivity analysis.

N/A 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?
API cost constraints limit multiple runs.

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.?
No human participants involved. This study involves only automated computational 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)?
No human participants were recruited or paid. This study involves only automated computational experiments.

☒ 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)?
Our data consists of chemical molecules and reactions from public/commercial databases, not human-generated content requiring consent.

☒ N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No human subjects involved. We use existing public chemical databases, which do not require 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 used ChatGPT/GPT-4 for language polishing and proofreading. This is disclosed in this checklist, and Appendix A (Section "Use of LLMs") discloses that LLMs were used for grammar checking, LaTeX formatting, improving clarity, and code refactoring. Core contributions are from the authors.