Responsible NLP Checklist

Paper title: *ADAPT: Benchmarking Commonsense Planning under Unspecified Affordance Constraints*
Authors: *Pei-An Chen, Yongching Liang, Jia-Fong Yeh, Hung-Ting Su, Yi-Ting Chen, Min Sun, Winston H. Hsu*

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?
Potential Risks

☑ **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 DynAfford benchmark is constructed entirely in simulated household environments and does not involve real human subjects, personal data, or user-generated content. All instructions and annotations describe object-centric household tasks and do not contain personally identifying information or offensive content. Therefore, no anonymization procedures are required.

☑ 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 3.3 reports dataset statistics, including the number of tasks, scenes, and train/test splits.

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

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We do not perform hyperparameter search in this work. All hyperparameters are fixed based on prior work or design choices, and our evaluation focuses on comparing methods under a consistent experimental setup rather than tuning for optimal performance.

☒ 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?
All reported results are obtained from a single deterministic evaluation run on the full test split of DynAfford, with metrics aggregated over all episodes in each split. While commercial APIs may still exhibit minor nondeterminism, our evaluation focuses on affordance classification rather than open-ended generation, and the observed performance gaps are consistent across object categories.

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.?
(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?
Appendix H