Responsible NLP Checklist

Paper title: *From Imitation to Discrimination: Progressive Curriculum Learning for Robust Web Navigation*

Authors: *PengChuang, Wei Zhang, Renshuai Tao, Xinhao Zhang, Jian Yang*

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?
Section 6 (Limitations) and Ethics Statement. We briefly discuss that our web navigation agent, while designed for legitimate automation tasks, could potentially be misused for unauthorized automated interactions with websites (e.g., automated data scraping violating Terms of Service). We emphasize that our research aims to improve the robustness and reliability of web agents for legitimate use cases, and we advocate for responsible deployment with proper human oversight and adherence to website policies.

☑ **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?
Our research utilizes existing, publicly released datasets (Mind2Web and WebSight). We did not perform or discuss additional PII or offensive content filtering, as these datasets were already anonymized and vetted by their original creators.

☑ 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 2.3 (Table 1) and Section 4.1 (Table 2)

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.1 (Implementation Details), Section 4.5 (Sensitivity of Discriminative Hyperparameters), and Appendix B.3 (Table 5)

☒ 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?

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.

Due to the significant computational constraints of progressively training and evaluating a large 32B-parameter model across three stages (SFT, ORPO, GRPO), we did not perform multiple independent training runs with different random seeds. Therefore, we do not report error bars, and the results reflect a single training run evaluated with greedy decoding.

☑ **D. Did you use human annotators (e.g., crowdworkers) or research with human subjects?**

☒ 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.?
We selected "No" because the human evaluation was a small-scale, internal pilot study (200 samples, mentioned in Section 2.2) conducted simply to sanity-check the automated data synthesis pipeline. Because the task was a straightforward binary verification of instruction-element pairs rather than a large-scale crowdsourcing effort, formal annotation instructions and interface screenshots were not included in the manuscript

☒ 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)?
We selected "No" because the human evaluation was a very small-scale internal pilot study (verifying only 200 samples) conducted by the researchers to sanity-check the automated synthesis pipeline. It did not involve recruiting external participants via crowdsourcing platforms, therefore formal recruitment and payment protocols were not reported.

☒ 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)?
We selected "No" because our work relies entirely on existing, publicly available datasets (Mind2Web and WebSight) and synthetic data generated by LLMs. We did not independently scrape, curate, or use new personal data from human subjects. Any data consent procedures were already handled by the original creators of the foundational datasets prior to their public release.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We selected "No" because our research relies on existing public datasets and synthetic data generation. The human involvement was strictly limited to an internal, small-scale pilot study (200 samples) to verify the accuracy of the automated data synthesis, which does not constitute human subjects research requiring formal Institutional Review Board (IRB) or ethics committee approval.

☒ **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)