Responsible NLP Checklist

Paper title: *When Agents Look the Same: Quantifying Distillation-Induced Similarity in Tool-Use Behaviors*
Authors: *Chenghao Yang, Yuning Zhang, Zhoufutu Wen, Tao Gong, Jiaheng Liu, Qi Chu, Nenghai 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?
This work is an analytical framework for ecosystem transparency. It does not enable attacks, generate harmful content, or collect personal data. Behavioral similarity serves as an indicator rather than definitive proof of distillation.

☑ **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?
τ-Bench/ τ^2-Bench contains simulated customer service dialogues without real personal data.

☑ 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 (150 tasks, 50 per domain).

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.1 describes model families, benchmarks, and baseline metrics.

☑ 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?
Section 4.2 and Table 1 report similarity scores. Results are from single runs using deterministic API calls with default hyperparameters.

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 annotators were used. All evaluations were conducted using LLM judges.

☒ 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 involved.

☒ 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)?
No human data was collected. We used publicly available benchmarks.

☒ N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No data collection was performed. We used publicly available benchmarks (-Bench, -bench) and accessed models via official APIs.

☑ **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?
AI assistants were used for code debugging.