

Responsible NLP Checklist

Paper title: *DatawiseAgent: A Notebook-Centric LLM Agent Framework for Adaptive and Robust Data Science Automation*

Authors: *Ziming You, Yumiao Zhang, Dexuan Xu, Yiwei Lou, Yandong Yan, Wei Wang, Huamin Zhang, Yu Huang*

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 7 Ethics Statement

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

B1. Did you cite the creators of artifacts you used?
Section 4 Experiments

B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
Yes we open-sourced our data and code at <https://github.com/Luffyzm3D2Y/DatawiseAgent> under the Apache 2.0 license (see the footnote in the Abstract).

B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Yes our data and code are released under the Apache 2.0 license at <https://github.com/Luffyzm3D2Y/DatawiseAgent>, intended for research purposes, which is consistent with the original access condition.

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?
All benchmarks used are data science tasks without personally identifying info or offensive content.

B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Yes we provide documentation in the repository README, including installation and usage instructions for code, as well as dataset details such as domains and languages (see the footnote in the Abstract).

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.

- 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 Experiments and Appendix

- C. Did you run computational experiments?**
 - C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4 Experiments
 - C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4 Experiments and Appendix
 - 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?
See Section 4 Experiment and Appendix. And we open-source the primary experimental results and agent trajectories in our repo for transparency and reproductibility.
 - C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?
Yes, all packages were used with default parameters unless otherwise specified.

- 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.?
Our work does not involve human participants or annotators.
 - 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)?
Our work does not involve human participants or annotators.
 - 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 use publicly available datasets released for research purposes under their original terms (see Section 4 Experiments).
 - D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No data collection is involved.
 - D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No data collection is involved.

- 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 AI assistants (ChatGPT) only for language refinement and grammar improvements. The scientific content, analyses, and results were entirely produced by the authors. This use was disclosed in the ARR submission form.