Responsible NLP Checklist

Paper title: *Collab-Overcooked: Benchmarking and Evaluating Large Language Models as Collaborative Agents*

Authors: *Haochen Sun, Shuwen Zhang, Lujie Niu, Lei Ren, Hao Xu, Hao Fu, Fangkun Zhao, Caixia Yuan, Xiaojie Wang*

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?
The "Ethics Statement" on page 9 discusses potential risks.

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

☑ B1. Did you cite the creators of artifacts you used?
Creators of used artifacts are cited throughout the paper. For example, the Overcooked-AI and ProAgent frameworks are cited in Section 4.1. The 13 LLMs used in the experiments are cited in Section 5.2.

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
The artifacts utilized are publicly available and were used in accordance with their standard open-source licenses or terms of service for API-based models. A detailed discussion of these widely understood terms was omitted for brevity, as all license information is accessible via the provided citations.

☑ 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)?
The paper specifies that the intended use of the created artifact, Collab-Overcooked, is to serve as a benchmark for evaluating the collaborative capabilities of LLM-based Multi-Agent Systems (Section 1 and 4). The use of the Overcooked-AI environment and various LLMs is consistent with their established applications in AI coordination and agent-based research.

☑ 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 "Ethics Statement" on page 9 confirms that "No personally identifiable information was collected during the experiments".

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.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Detailed documentation for the benchmark is provided in Appendix A. This includes descriptions of the environment (A.1), task construction (A.2), and the baseline agent architecture (A.3). A full list of the 30 tasks is provided in Table 5.

☑ 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?
Detailed statistics are provided in Section 5.1 and Appendix A.2.

☑ **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 5.2 reports the parameter sizes of the 13 models used.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5.2 details the experimental setup.

☑ 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 paper reports mean scores for its metrics over 10 repetitions, as stated in Section 5.2.

☑ 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?
Section 5.2 reports the specific parameters used.

☑ **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.?
Appendix C.2 and Figures 16 and 17 describe and show the human-computer interaction interface used in the experiments.

☑ 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)?
The "Ethics Statement" on page 9 discusses the recruitment and payment.

☑ 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)?
The "Ethics Statement" on page 9 discusses the data consent.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The study was assessed as minimal risk, as it involved a non-invasive behavioral task (playing a computer game) and no personally identifiable information was collected. Therefore, formal ethics review board approval was not deemed necessary for this type of user study.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Demographic data was not collected to ensure the privacy of the participants, particularly given the small sample size (10 volunteers). Furthermore, these characteristics were not considered relevant for establishing a general human performance baseline on the problem-solving task in our study.

☒ **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)