Responsible NLP Checklist

Paper title: *ProactiveEval: A Unified Evaluation Framework for Proactive Dialogue Agents*
Authors: *Tianjian Liu, Fanqi Wan, Jiajian Guo, Xiaojun Quan*

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?
Our work measures models' proactive dialogue capabilities through experiments on LLM-generated synthetic data and simulated users. It is not deployed in the real world, and the content focuses on how to assist users without involving harmful content.

☑ **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 motivation of our work is to evaluate the model's proactive dialogue capabilities. We employ models with superior safety performance to generate data. During the data generation process, the topic tree it follows has undergone manual verification, and the ICL shots referenced are also manually edited for safety (as shown in Section 4). These measures ensure that the generated data remains as safe as possible.

☑ 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?
We report the domain and statistics in Section 5.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We elaborate on our considerations in parameter design and other setups in Section 5.1.

☑ 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?
We report the statistics of results in Sections 5.1 and 5.2. The experiments present results from single runs, but we report the standard deviation across multiple runs in Appendix D.2.

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?**

☑ 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 present this in Section 7.

☒ 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 human annotators were two authors with research experience in proactive dialogue and expertise in evaluating proactive dialogue systems.

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?
We employed AI assistants purely for language assistance LLMs polished our writing and corrected grammar without altering the authors' original intent. All ideas, experiments, and initial drafts were completed by human authors.