Responsible NLP Checklist

Paper title: *Beyond Task-Oriented and Chitchat Dialogues: Proactive and Transition-Aware Conversational Agents*
Authors: *Yejin Yoon, Yuri Son, Namyeong So, Minseo Kim, Minsoo Cho, Chanhee Park, Seungshin Lee, Taeuk Kim*

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?
See Ethics Statement section. The paper discusses dataset bias and privacy considerations.

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

☑ B1. Did you cite the creators of artifacts you used?
See Section 2 and 3. Prior datasets including MultiWOZ, SLURP, FusedChat, and InterfereChat are properly cited.

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
See Terms for Use of Artifacts section (Appendix). We specify MIT (MultiWOZ), CC BY 4.0 (SLURP)

☑ 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)?
See Terms for Use of Artifacts. All datasets were used within their intended academic and non-commercial purposes.

☒ 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?
As stated in the Ethics Statement, our dataset construction relies exclusively on public benchmarks (MultiWOZ, SLURP) that contain no personally identifiable information (PII). The augmented dialogues were synthetically generated, further minimizing any risk of privacy leakage or offensive content.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
See Section 3 (Dataset Construction) and Appendix A. We document TACT with domain coverage, flow types (TCT/CTC), intent schema, and characteristics of recoverable dialogues.

☑ 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?
See Table 1, Figure 4, and Appendix A.3, which report dataset statistics including dialogue counts, average turns, transition distributions, and train/dev/test splits.

☑ **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?
See Section 5.2 (Training Configuration) and Appendix D.1 (Environments), where we report model size (LLaMA-3.1-8B), computing infrastructure (2A100 80GB), and training setup.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
See Section 5.2 (Training Configuration) and Appendix D.1, which describe training epochs, batch size, learning rate, and optimization details.

☑ 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 6 (Experimental Results) and Appendix E.2E.3. We report comparative tables with multiple runs, human evaluation sample sizes, and summary statistics of model performance.

☑ 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?
See Appendix D.1 (Environments). We report the implementation frameworks (PyTorch, DeepSpeed ZeRO-3, vLLM) and model settings used for training and inference.

☑ **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.?
See Appendix D.2 (Evaluation Metrics), Figures 15 and 16, where we provide detailed instructions and screenshots of the evaluation interface.

☒ 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)?
Annotators (NLP practitioners and general users) participated voluntarily without monetary compensation; no crowdsourcing platform was used (Section 6.2).

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)?
All data used were from public benchmarks (MultiWOZ, SLURP) and synthetically generated augmentations; no personal data collection was involved.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
As no new personal data was collected and only public benchmarks were used, IRB approval was not required.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
ee Section 6.2 (Preference-Based Evaluation), where we describe annotators as including both NLP practitioners and general users.

☑ **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 only for grammar and spelling checks; not mentioned in the paper since no research contribution.