Responsible NLP Checklist

Paper title: *Decision-Making with Deliberation: Meta-reviewing as a Document-grounded Dialogue*
Authors: *Sukannya Purkayastha, Nils Dycke, Anne Lauscher, Iryna Gurevych*

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?
In the Limitations section of the paper

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

☑ B1. Did you cite the creators of artifacts you used?
In Sec 3.2, we attribute creators of the models, metrics employed, underlying dataset used and baselines. In Sec 3.3.1, we attribute creators of the human dialogue dataset. In Appendix A.1, we further detail the models used.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We use openly available datasets and models discussed in Sec 3.2 with further details in A.1.

☑ 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)?
We use all openly available datasets and models discussed in Sec 3.2 and A.1 intended for use in research 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?
The datasets used or released by us donont contain any personally identifiable information.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
The dataset analysis is in A.6 where we discuss the characteristics of the dataset such as domains covered, dialogue turns and lengths.

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?
The statistics of the underlying dataset used is reported in Sec 3.2 (Datasets), human dialogue dataset (Sec 3.3.1) and the generated dataset in Sec A.6

☑ **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?
The model size and computational budget are discussed in Sec 3.2 (Models) and A.1 for model details and computational budget.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The hyper-parameters used in dialogue generation are detailed in A.1 and the dialogue response generation in Sec A.7

☑ 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 standard deviation and average of the dialogue generation experiments are in Table 1 discussed in Sec 3.3.2 and full tables with the results in Table 7,8 and 9 respectively. The results of the dialogue response generation is in Sec 4.2 in Tables 4 and 5.

☑ 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?
We use OpenAI azure for all experiments with OpenAI models and huggingface transformers for other opensource models discussed in Sec A.1.

☑ **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.?
The instructions for evaluating dialogues in A.10 and response generation in A.11

☒ 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)?
This was a voluntary effort by the PhD students of the university

☒ 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 opensource datasets for generating dialogues and donot curate any data using annotators

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We didnot collect any data but rather used opensource datasets

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
The annotator details are provided in A.12

☒ **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 didnot use any AI assistants for writing the paper.