Responsible NLP Checklist

Paper title: *We Politely Insist: Your LLM Must Learn the Persian Art of Taarof*
Authors: *Nikta Gohari Sadr, Sahar Heidariasl, Karine Megerdoomian, Laleh Seyyed-Kalantari, Ali Emami*

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 Ethical Considerations Section discusses potential risks of cultural evaluation and adaptation.

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

☑ B1. Did you cite the creators of artifacts you used?
All models (GPT-4o, Claude 3.5 Haiku, Dorna, Llama 3, DeepSeek) are cited in Section 3

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We do not explicitly discuss licenses for the models used or our dataset. Our dataset will be released with an open license.

☑ 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)?
In the Evaluation Methodology section (2.3), we describe how models were used within their intended conversational purposes.

N/A 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?
Our dataset consists entirely of fictional scenarios created by the authors and does not contain personal information or offensive content as described in Scenario Design (2.2)

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
In the Scenario Design (2.2) section, we document our dataset's coverage of topics, settings, and cultural relevance.

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?
We report statistics in the section 2.2, including number of scenarios (450), topic distribution, and setting types.

☑ **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 3 lists the models used in our experiments with their respective sizes, and Appendix 1.6 describes the computational resources used for fine-tuning.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In the Section 3, we report hyperparameters for fine-tuning experiments.

☑ 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?
Section 4 reports accuracy scores, p-values for significance tests, and performance differences across conditions.

☑ 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 describe the settings for Polite-Guard and GPT-4 judge (temperature=0) in Section 2.3 and 3.

☑ **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.?
Instructions given to participants are described in the Human Study paragraph (Section 3).

☑ 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 Study paragraph includes that compensation was provided in accordance with institutional guidelines.

☑ 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)?
In the Section 3, we mention that participants provided informed consent for data collection.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Section 3 notes that the human study protocol was approved by our institution's IRB.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We describe participant groups (11 native Persian, 11 heritage, 11 non-Iranian) with demographic details in Section 3 and Appendix 1.7.

☒ **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)