Responsible NLP Checklist

Paper title: *FaST: Feature-aware Sampling and Tuning for Personalized Preference Alignment with Limited Data*
Authors: *Thibaut Thonet, Germn Kruszewski, Jos Rozen, Pierre ERBACHER, Marc Dymetman*

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?
We include an "Ethical statement" section, just after the "Limitations" section at the end of the main body 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?
The datasets (Personalized Soups, ELI-5) which were exploited to build our own datasets are mentioned in Section 4.1 ("Datasets"). The existing LLMs used in our experiments are discussed in Section 4.2 ("Baselines").

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
The licence of the ELI-5 paper is specified in Appendix B.2; Personalized Soups was not released with a license.

N/A 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 intended use of the ELI-5 and Personalized Soups datasets are not available, to the best of our knowledge.

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?
Although our work is on personalization, we use LLM-generated synthetic data which prevents the use of sensitive personal information. Furthermore, our use of GPT-4o (explicitly aligned for safety and harmlessness) on non-toxic contexts ensures that no harmful language was used. The contexts used in the ELIP dataset have also been manually curated, and any potentially offensive questions were discarded (the manual curation is discussed in Appendix B.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.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
The construction of the datasets is briefly described in Section 4.1 ("Datasets") and further detailed in Appendix B.2.

☑ 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?
Statistics of the datasets are indicated both in Section 4.1 ("Datasets") and Appendix B.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?
The number of parameters are mentioned in the experimental setup paragraph of the Experiments sections (Sections 4.2 and 4.3) and further elaborated in Appendix C.2. The computational resources used are discussed in Appendix C.5.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The hyperparameters are discussed in Appendix C.3.

☑ 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?
Our preferred response prediction experiments (Section 4.2) were done on 5 train/val/test splits. Our personalized generation experiments (Section 4.3) were done on 3 train/val/test splits.

N/A 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?
(left blank)

☒ **D. Did you use human annotators (e.g., crowdworkers) or research with human subjects?**

N/A 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.?
(left blank)

N/A 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)?
(left blank)

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)

N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
(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?
The use of GPT-4o for the generation of the datasets is discussed in Appendix B, and its use for LLM-based evaluation is mentioned in Appendix C.4.