Responsible NLP Checklist

Paper title: *TP-RAG: Benchmarking Retrieval-Augmented Large Language Model Agents for Spatiotemporal-Aware Travel Planning*
Authors: *Hang Ni, Fan Liu, Xinyu Ma, Lixin Su, Shuaiqiang Wang, Dawei Yin, Hui Xiong, Hao Liu*

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?
Limitations & Ethical Statement Sections

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

☑ B1. Did you cite the creators of artifacts you used?
Section 3.2, Section 4.1, Appendix B

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
Ethical Statement Section

☑ 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)?
Ethical Statement Section

☑ 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?
Ethical Statement Section

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We release a new dataset with the basic information introduced in Appendix A, but the raw non-aggregated data can not be publicly disclosed.

☑ 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?
Appendix A (we have not specified a train/val/test splitting method; users are free to apply their own approaches.)

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.

☑ **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 used GPU hours and GPT tokens are reported in Section 1, Appendix A, Appendix C.1. We benchmark existing LLMs and our simple method is also LLM prompting-based, so we do not report the number of parameters.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Some detailed setups are showed in Section 4.1, Appendix C.1, and Appendix D. However, our benchmark and simple method do not include complicated hyperparameters, and we do not do hyperparameter search.

☑ 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 emphasizes the reasons for using single run in Appendix C.1 and Appendix D. Since the considered methods are training-free, we do not split train/val/test data and just test the baselines and our method on the entire dataset.

☒ 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?
Data processing and evaluation are implemented by ourselves. The details of evaluation metrics are explained in Appendix B. We cite the paper or include the URL for all the used packages (e.g., TSP solver), and do not make any modifications of the parameters.

☑ **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 apply human annotators for dataset quality control (Section 3.2). We list some coarse-grained criteria (reported in Section 3.2) and highly rely on multiple rounds of meetings to determine the concrete criteria, instead of using structured and formal instructions. Also, we incorporate human evaluators (Appendix B). The instructions to human evaluators are similar to the prompts given to LLM evaluators (reported in Appendix E2), so we do not replicate that. We clarify this in Appendix B. We ensure the effectiveness of the assessment through repeated reviews of evaluation results and verbal discussions.

☒ 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)?
Our participants come from Baidu team, who are professional annotators. There are no specialized payments for our paper.

☑ 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)?
Ethical Statement Section. The document data was sourced internally from members of our research team, which ensures their consent. While other data are public and we use them under related licenses.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We just apply the human evaluators from our own team, without the needs to involve the ethics review board.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Our data annotation process is not sensitive to the demographic and geographic characteristics of annotators.

☑ **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?
Appendix F.