Responsible NLP Checklist

Paper title: *Over-Generation and Compaction: A Prompting Strategy for Procedural Text Adaptation with Large Language Models*
Authors: *HyeongSik Kim, XU Yanheng, Chaoqun Dong, Fei Du*

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 Appendix Section for a detailed discussion of potential risks, including prompt sensitivity, semantic drift during compaction, evaluator bias in our scoring metrics, and domain-specific limitations of applicability.

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

☑ B1. Did you cite the creators of artifacts you used?
We used the Xiachufang recipe corpus introduced by Xiao Liu, Yansong Feng, Jizhi Tang, Chengang Hu, and Dongyan Zhao. 2022. Counterfactual Recipe Generation: Exploring Compositional Generalization in a Realistic Scenario. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 73547370. In addition, we employed the MyFixit dataset introduced by Nima Nabizadeh, Dorothea Kolossa, and Martin Heckmann. 2020. MyFixit: An Annotated Dataset, Annotation Tool, and Baseline Methods for Information Extraction from Repair Manuals. In Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020), pages 21202128. Citation details are provided in Section 3 (Evaluation).

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
The Xiachufang corpus was obtained from a publicly available recipe-sharing website, as in the prior work of Liu, Feng, Tang, Hu, and Zhao (2022). However, the license terms for redistribution were not explicitly specified in the original source. Consequently, we used the data strictly for research purposes and do not redistribute it. Similarly, the MyFixit corpus is derived from the iFixit website, where repair manuals are collaboratively created and shared. While the dataset was originally collected and annotated by Nabizadeh, Kolossa, and Heckmann (2020), the licensing conditions of the underlying manuals remain ambiguous. To avoid potential conflicts, we likewise restrict our use of this corpus to research purposes only and refrain from redistributing the processed data. Therefore, for both datasets we only release preprocessing code and prompt templates rather than the datasets themselves.

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.

☑ 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 used the Xiachufang and the myfixit dataset strictly for research purposes, consistent with its original use. This is discussed in Section 4 (Evaluation), and our usage adheres to research norms regarding non-commercial application and data transformation for procedural NLP tasks.

☒ 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 Xiachufang corpus consists solely of cooking recipes and does not contain user-identifiable fields. As such, we did not perform additional anonymization. This corpus has been previously used in peer-reviewed work and is considered non-sensitive. Likewise, the MyFixit corpus is composed of repair manuals collected from the iFixit platform. These instructions focus exclusively on procedural content, such as device components, tools, and stepwise actions, and do not include personal or user-identifiable information. The dataset has also been published and evaluated in peer-reviewed research, and it is therefore regarded as non-sensitive.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We document dataset coverage and characteristics in Section 3 (Evaluation), including the number of recipes, categories, and dataset cleaning steps for the Xiachufang corpus. Likewise, for the MyFixit corpus we report the number of repair instructions, the distribution of device domains, and details of the additional component substitution scenarios. Full outputs and annotations for both datasets are provided with the code and supplementary material.

☒ 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?
Section 4 (Evaluation) includes detailed statistics: we used 2,492 recipe pairs after filtering, with coverage across 50 dish categories and approximately 50 variants per category. For the MyFixit corpus, we used around 500 curated componentinstructionsubstitution triples, spanning multiple domains such as consumer electronics, household appliances, automotive parts, and clothing repair.

☑ **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?
Model names and sizes are described in Section 3 (Evaluation) and Appendix. Exact GPU hours or infrastructure details were not provided due to API-based usage of proprietary LLMs, which do not disclose hardware specifications.

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We did not conduct gradient-based training or hyperparameter tuning, as our experiments involved prompting API-based LLMs. Therefore, traditional hyperparameter tuning is not applicable. Prompting strategies are discussed in Section 3 and Appendix

☑ 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?
Yes, Section 3 reports mean, median, standard deviation, and 95% confidence intervals for the metrics we introduced scores across models and strategies.

☑ 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 employ the g-eval framework to facilitate prompting and evaluation workflows. All prompting strategies are documented in detail (Section 3 and Appendix), and the full configuration details and parameter settingscovering prompt templates, API configurations, and evaluation metric specificationsare provided in our public code repository.

☑ **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 expert evaluation component of our study involved internal reviewers with culinary expertise. As the reviewer was an internal collaborator, formal participant instructions (e.g., disclaimers or standardized prompts) were not necessary and thus not documented. The evaluation process was informal and integrated into internal research practices, aligned with institutional ethical standards.

☑ 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)?
Details regarding the recruitment and compensation of the human reviewer are included in the Ethics Statement section. The expert reviewer was an internal member of the research organization and was fairly compensated in accordance with internal ethical guidelines and standard remuneration practices.

☑ 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)?
Our study did not involve the use of third-party data derived from identifiable individuals or external annotators. The human review process was conducted internally, with the consent of the participating expert implicit through institutional affiliation and participation under standard organizational protocols. As such, no additional data consent procedures were applicable.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The study did not involve vulnerable populations or sensitive personal data, and the human subject component was limited to qualitative review by an internal expert affiliated with the research organization. Given the low-risk nature and internal scope of participation, the work was not submitted for formal ethics review and would likely qualify as exempt under most institutional review board guidelines.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Only internal reviewers with culinary expertise contributed to the qualitative evaluation. Because the annotator was not part of a broader or externally recruited population, we did not report demographic or geographic characteristics. The reviewers are full-time researchers based at the institution conducting the study.

☑ **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 (e.g., ChatGPT) were used in a limited capacity to refine wording and improve clarity during the writing process. As our experiments centrally involve testing LLMs for procedural text editing, their use in generating and evaluating outputs is an inherent part of the research itself. All methodological design, analysis, and interpretation were conducted by the authors.