## Responsible NLP Checklist

Paper title: *Leveraging High-Resource English Corpora for Cross-lingual Domain Adaptation in Low-Resource Japanese Medicine via Continued Pre-training*
Authors: *Kazuma Kobayashi, Zhen Wan, Fei Cheng, Yuma Tsuta, Xin Zhao, Junfeng Jiang, Jiahao Huang, Zhiyi Huang, Yusuke Oda, Rio Yokota, Yuki Arase, Daisuke Kawahara, Akiko Aizawa, Sadao Kurohashi*

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 "Limitations" section discusses the constraints and scope of the study, such as small sample size, computational costs, and data sharing restrictions due to licensing. It does not address potential societal risks or harms that could arise from the work itself.*

☑ **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, Appendix A*

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
*The "Limitations" section (Section 6) discusses the terms of use for the translation engine and the resulting restrictions on sharing the generated data.*

☑ 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 "Limitations" section (Section 6) discusses the inability to share derivatives (the translated corpus) due to the original access conditions of the translation engine, confirming that the use is consistent with the specified terms.*

☒ 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 paper uses publicly available and commercially allowed subsets of medical literature (e.g., PMC). However, it does not explicitly discuss specific steps taken to screen these public sources for any residual personally identifying information or offensive content.*

---

*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.?
*Section 3.1, Figure 2, and Appendix C provide detailed descriptions and compositional statistics of the multilingual corpora created for the study.*

☑ 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 provides details on the evaluation datasets, including the number of test samples. Appendix C provides detailed token counts and ratios for the corpora used in pre-training.*

☑ **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?
*Appendix D ("Training Details") reports the model size (13B parameters), computational budget (e.g., 256 H100 GPUs for about 24 hours for the largest corpus), and the computing infrastructure used (AWS SageMaker).*

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Appendix D ("Training Details") provides a detailed description of the experimental setup, including hyperparameters for both continued pre-training and supervised fine-tuning.*

☑ 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 paper reports results from a single, deterministic experimental run. It transparently reports the exact "score difference" as the primary metric for evaluation in Tables E.1 through E.6.*

☑ 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?
*Appendix B reports the specific models and libraries used for translation evaluation (SacreBLEU, MeCab, COMET). Appendix D reports the frameworks used for training (Megatron-LM v0.3.0, NeMo) and the use of FlashAttention.*

☒ **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?**

N/A E1. If you used AI assistants, did you include information about their use?
*(left blank)*