Responsible NLP Checklist

Paper title: *Leveraging Knowledge Graph-Enhanced LLMs for Context-Aware Medical Consultation*
Authors: *Su-Hyeong Park, Ho-Beom Kim, Seong-Jin Park, Dinara Aliyeva, Kang-Min Kim*

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?
Discussed in Limitations and 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?
In Section 1, UMLS and the datasets HealthCareMagic and iCliniq are cited in footnotes, while Triple2Seq is cited through related work. In Section 3, the retriever and reranker used are described and the corresponding references are provided.

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
While explicit license terms were not discussed in the manuscript, all artifacts used in this study, including UMLS, are publicly available for non-commercial academic use, and we have complied with their respective licensing policies.

☑ 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)?
Section 3.1 discusses the use of the UMLS-based knowledge graph solely for research purposes, consistent with its intended use.

☑ 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?
Section 4.1: The datasets originally contained user-submitted medical consultation questions, but both have been thoroughly de-identified and are publicly available for academic research. No personally identifiable information is present in the data used in this study.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?

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.

Section 4.1.1 describes the UMLS knowledge graph used as an external resource, and Section 4.1.2 documents the HealthcareMagic and iCliniq medical consultation datasets, including their source and characteristics.

☑ 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 4.1.2; Table 1 provides statistics including the number of examples, token counts, and train/validation/test splits for both datasets.

☑ **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 A (Implementation Details); describes model sizes (e.g., Llama2 7B, Llama3.1 8B), GPU configuration (2RTX A6000 48GB), and fine-tuning setup.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix A (Implementation Details) specifies the fine-tuning hyperparameters such as the LoRA settings, learning rate, warmup steps, scheduler, number of epochs, and sequence length for the large language model. Additionally, Appendix C describes the training settings for the retriever and reranker, as well as the number of documents used for reranking.

☑ 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 4.4 (Result) and Table 2; reports metrics such as F1, METEOR, BLEU-4, ROUGE-2, Top-1 Hit Rate, and Avg NLI Score across in-distribution and out-of-distribution sets, with performance compared across multiple models and settings.

☑ 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?
Section 4.2 reports evaluation models and metrics, including BERTScore with RoBERTa Large, ROUGE-2, BLEU-4, METEOR, and MinosEval.

☒ **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.?
No human subjects or annotators were involved in this work.

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)?
No human participants or annotators were recruited or paid in this study.

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)?
The datasets used (HealthcareMagic and iCliniq) are publicly available and already de-identified for research purposes. No additional consent was required.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The datasets used in this study (HealthcareMagic and iCliniq) are publicly available and fully de-identified for research purposes. Therefore, ethics board approval was not required.

☒ N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No information on annotator demographics was included in the original publicly available datasets, and we did not conduct any new annotation work.

☑ **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 paper does not disclose that AI assistants such as ChatGPT were used during the development of this work to improve the fluency of the manuscript and to assist in debugging code (e.g., identifying syntax errors or resolving runtime issues). However, they were not used for generating novel research ideas, designing the model, or writing core experimental content.