Responsible NLP Checklist

Paper title: *InfoGain-RAG: Boosting Retrieval-Augmented Generation through Document Information Gain-based Reranking and Filtering*
Authors: *Zihan Wang, Zihan Liang, Zhou Shao, Yufei Ma, Huangyu Dai, Ben Chen, Lingtao Mao, Chenyi Lei, Yuqing DING, Han Li*

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.

N/A A2. Did you discuss any potential risks of your work?
Our paper first collected training data based on LLM inference, and then trained a 355M reranking model. The computational resources consumed are relatively minimal and will hardly pose harm to the environment. Moreover, our method helps LLMs answer human questions more accurately and reduces the interference caused by LLM hallucinations to users.

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

☑ B1. Did you cite the creators of artifacts you used?
Section 4

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
Section 4

☒ 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 only used datasets and models that are widely used in the academic community. The reranking model that we trained is solely for academic research purposes, aiming to design an efficient inference ranking module for RAG. It will not have any impact on marginalized or vulnerable groups.

☒ 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?
Our work primarily focuses on improving the retrieval-augmented generation (RAG) framework by introducing a novel document reranking method based on Document Information Gain (DIG). The datasets we used to construct training data is TriviaQA and wikipedia-2018, which are publicly available benchmark datasets commonly used in the research community for evaluating question

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.

answering. These datasets are designed for research purposes and do not contain personally identifying information or offensive content.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4 and AppendixD

☑ 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 D

☑ **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?
Section 4

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4

☒ 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?
Due to the substantial computational requirements of LLM inference for each evaluation, we conducted the evaluation only once. However, we have previously performed multiple evaluations under the same settings and found that the differences between the results were very minimal.

☒ 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?
The evaluation metric we used is exact match, which does not require the use of existing packages; it can be implemented on our own. We also don't need to perform complex processing on the data. The main package we use is huggingface.

☒ **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.?
Our research did not involve the use of human annotators or human participants. The experiments and evaluations were conducted using existing datasets and automated methods.

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)?
Our research did not involve recruiting or paying human participants, as it was based on automated experiments and evaluations.

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 in our experiments are publicly available and do not involve personal identifying information (PII). No consent was required for the use of these datasets.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Our research did not involve human subjects or data collection that required ethical review.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Our research did not involve human annotators or participants. The data used in our experiments are publicly available datasets (e.g., TriviaQA, NaturalQA, PopQA, FM2)

☒ **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?
We didn't use AI assistants for research, coding, or writing.