Responsible NLP Checklist

Paper title: *DeepSieve: Information Sieving via LLM-as-a-Knowledge-Router*
Authors: *Minghao Guo, Qingcheng Zeng, Xujiang Zhao, Yanchi Liu, Wenchao Yu, Mengnan Du, Haifeng Chen, Wei Cheng*

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?
(left blank)

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

☑ B1. Did you cite the creators of artifacts you used?
See Section 3.2 and References. We cite the creators of all datasets used (e.g., MuSiQue~(?), 2Wiki~(?), HotpotQA~(?)) and retrieval models (e.g., ColBERTv2~(?)).

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We did not discuss licenses explicitly, as all datasets and retrievers used in our work are publicly available for research use under standard academic terms. See Section 3.2.

☑ 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)?
See Section 3.2. All public datasets (MuSiQue, 2WikiMultiHopQA, HotpotQA) are used according to their intended purpose (multi-hop QA). No restricted or private data is involved.

☒ 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?
See Section 3.2. All datasets are publicly available and contain no personally identifiable or offensive content. We do not collect or annotate any additional data.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
See Section 3.2 and Appendix A.1. We document dataset characteristics and retrieval corpus statistics, and specify the configuration of retrievers and sources in our experiments.

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.

☑ 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?
See Appendix A.1. We report number of samples, corpus sizes, and retrieval top-k values across three datasets. We also document subquestion decomposition and route counts.

☑ **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?
See Section 5.1. We report the LLM models used (e.g., DeepSeek-V3, ColBERTv2), their scale (e.g., 16B parameters), and total GPU hours used for inference (e.g., ~350 A100 GPU hours).

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
See Section 5.1. We describe the experimental configurations, including model backend (DeepSeek-v3-chat), top-k for retrieval (e.g., k=10), and decomposition/routing/reflect module settings.

☑ 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?
See Section 5 and Appendix A. We report average EM/F1 over 3 seeds for each method, with standard deviation across runs.

☑ 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?
See Section 5.1. We specify the retrieval packages used (e.g., ColBERTv2 for dense retrieval, FAISS for clustering), along with parameter settings such as top-k, batch size, and rerank thresholds.

☒ **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 were involved; all data was sourced from publicly available benchmarks. See Section 5.1.

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 were recruited or paid. All data used is from existing datasets (e.g., HotpotQA, MuSiQue). See Section 5.1.

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)?
All datasets used are publicly released under research-friendly licenses. No private or user-curated data was involved. See Section 5.1.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No human subjects involved. See Section 5.1.

N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No annotation or human annotators involved. All datasets used are publicly available. See Section 5.1.

☑ **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 used ChatGPT (GPT-4) to assist in editing and rephrasing some sentences in the paper.