Responsible NLP Checklist

Paper title: *Navigating Large-Scale Document Collections: MuDABench for Multi-Document Analytical QA*
Authors: *Zhanli Li, Yixuan Cao, Lvzhou Luo, Ping Luo*

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?
We discuss potential risks in the Ethical Considerations section, including privacy considerations of public financial disclosures and the risk that automated systems may produce incorrect financial analyses if deployed without expert verification.

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

☑ 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?
We discuss data privacy checks in the Ethical Considerations section. The source documents are public corporate disclosures. We do not annotate individual-level personal attributes, and the released benchmark focuses on company-level metadata and financial indicators. We also checked the released structured fields to avoid unnecessary personal identifiers or offensive content.

☑ 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?
Dataset statistics are reported in Section 3 Benchmark, including the number of documents, pages, questions, document types, metadata fields, and documents per question. Additional source and construction details are provided in Appendix Data Source Details and Benchmark Construction Detail.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The experimental setup is described in Section 5.1 Experimental Setup. We report the evaluated model versions, retrieval settings, chunk budgets, metadata variants, workflow settings, judge models, and temperature. We did not perform hyperparameter search; the retrieval chunk budgets were pre-specified experimental conditions.

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. ACL 2026 used a subset of ARR checklist form.

☑ 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?
Results are reported in Section 5.2 and Tables 35 as average accuracies over the evaluated questions. Each result is from a single deterministic run with temperature 0; we do not report error bars because of the high cost of running commercial API-based document workflows repeatedly.

☑ **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.?
The benchmark construction primarily relied on distant supervision and automated pipelines (Section 3.3). The minimal human involvement (e.g., converting metrics to descriptions and the "human performance" estimation) was conducted by internal researchers/authors familiar with the task. Therefore, formal crowdsourcing instructions or disclaimers were not applicable or utilized. D2 Recruitment And Payment: N/A

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)?
The annotators and volunteers mentioned in the paper were internal employees or researchers from the authors' institutions. They did not receive specific per-task payment as this work was part of their regular employment or research activities; thus, crowdsourcing payment rates are not applicable.

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 dataset is derived from publicly available corporate disclosures and does not involve collecting data from individuals; therefore consent is not applicable.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Our dataset is constructed from publicly available corporate disclosures (e.g., annual reports/ESG reports/announcements) and does not involve human participants, interventions, or collection of personal/sensitive data. Therefore, ethics board/IRB approval is not applicable.

☑ **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?
We used AI assistants for minor language polishing; they did not contribute to ideas, experiments, or results.