Responsible NLP Checklist

Paper title: *SimpleDoc: MultiModal Document Understanding with DualCue Page Retrieval and Iterative Refinement*
Authors: *Chelsi Jain, Yiran Wu, Yifan Zeng, Jiale Liu, Shengyu Dai, Zhenwen Shao, Qingyun Wu, Huazheng Wang*

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 Section 6 (Limitations) and Appendix A.5 (Error Analysis), where risks such as retrieval failure, hallucinations, misalignment, and layout errors are described.

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

☑ B1. Did you cite the creators of artifacts you used?
We created the SimpleDoc codebase and used external datasets. Cited dataset and model creators in Section 4 (Experiments) and References (e.g MMLongBench, LongDocURL, PaperTab, FetaTab, ColPali, M3DocRAG, MDocAgent).

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We relied on publicly available academic datasets and models released for research. Their terms permit research use, so explicit license text was not included.

☑ 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 4.1 (Datasets). All datasets were used strictly for research/benchmarking as intended by their creators.

☒ 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 datasets (MMLongBench, LongDocURL, PaperTab, FetaTab) are curated academic/benchmark datasets without PII 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.?
Dataset coverage and domains are described in Section 4.1 (Datasets).

☑ 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?
Reported in Section 4.1 (Datasets) includes number of documents, questions, pages, and modality details.

☑ **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?
Model sizes (Qwen2.5-VL-32B, Qwen3-30B, Qwen2.5-VL-7B, Qwen3-8B) and token-level I/O statistics are reported in Section 4 (Experiments) and Appendix A.4 (Computational Statistics).

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Details in Section 4 (Implementation Details) and Appendix A.6 (Prompts Used). We specified models, retrieval top-k, summarization, and evaluation protocol.

☑ 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?
Tables in Section 4.1 (Table 1, Table 2, Table 3, etc.) report averages across runs. Appendix A.4 includes error analysis and breakdown statistics.

☑ 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?
Implementations referenced (ColBERT, ColPali, PyPDF2, pdfminer.six). Details of retrieval/evaluation methods are in Section 3 (Method) and Section 4 (Implementation Details).

☒ **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?**

- ☑ E1. If you used AI assistants, did you include information about their use?
  Documented in Appendix A.2 (Usage of AI assistant) AI assistants were used for debugging code, building utility functions, and refining writing.