Responsible NLP Checklist

Paper title: *A Picture is Worth a Thousand Words? An Empirical Study of Aggregation Strategies for Visual Financial Document Retrieval*
Authors: *Ho Hung Lim, Yi Yang*

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?
This work is a diagnostic study analyzing publicly available financial datasets (FinQA and TAT-DQA). It does not introduce new models or systems that could be directly misused. The findings highlight risks of existing retrieval systems rather than introducing new ones.

☑ **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 use publicly available financial datasets (FinQA and TAT-DQA) that contain only publicly disclosed corporate financial reports. These documents do not contain personally identifying information 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?
Section 4.1 and Table 3 report dataset statistics, including the number of document pairs (N=200) for each experiment across FinQA and TAT-DQA.

☑ **C. Did you run computational experiments?**

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix B describes the experimental setup, including the GPU used (NVIDIA RTX 5880 Ada Generation, 48GB), model weights, and image resizing procedure. Our evaluation uses cosine similarity as the metric with no hyperparameter tuning required, as we use pre-trained models without fine-tuning.

☑ 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?

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.

Tables 4, 5, 6, and 7 report mean cosine similarity scores averaged over N=200 document pairs for each experiment condition. Results are reported across all models and datasets without additional variance metrics, as the diagnostic benchmark is designed to measure systematic behavior rather than statistical variance.

☒ **D. Did you use human annotators (e.g., crowdworkers) or research with human subjects?**

☒ 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 annotators or participants were used in this study.

☒ 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 annotators or participants were used in this study.

☒ 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)?
We use publicly available datasets (FinQA and TAT-DQA) that are openly released for research purposes.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
This study uses only publicly available datasets and does not involve human subjects. No ethics review board approval was required.

☑ **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?
AI language assistance was used for grammar refinement and stylistic editing. All technical content, experimental design, analysis, and conclusions were developed and verified by the authors.