## Responsible NLP Checklist

Paper title: *Achieving Multi-Hop Calculation and Safe Abstention in Financial Numerical Reasoning by Metric Graph Constrained LLMs*
Authors: *Aoyuan Jiang, Liang Hong, Haoxuan Liu, Rui 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.*

N/A A2. Did you discuss any potential risks of your work?
*Our work uses standard and publicly available benchmark datasets. So, we do not foresee significant potential risks.*

☑ **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?
*Our work uses standard, publicly available financial benchmark datasets derived from public corporate filings. These datasets do not contain sensitive personally identifiable info of private individuals 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?
*See Appendix B.*

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*See Section 5.*

☑ 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*

*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.*

☑ **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.?
*See Appendix B.2*

☒ 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 were recruited internally from the authors' research laboratory, consisting of doctoral or master's students with expertise in finance. As the annotation and verification tasks were conducted as part of their academic research responsibilities, they were supported by their standard academic stipends or salaries, and no specific crowd-sourced payment or piece-rate compensation was involved.*

☒ 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 are publicly available benchmarks derived from public corporate filings, where individual consent is not applicable.*

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*This research relies entirely on publicly available financial datasets derived from public corporate filings. As the study involves no external human subjects, interventions, or sensitive personal data, and serves solely for technical evaluation, it was determined that formal ethics review board approval was not 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 assistants were used exclusively for linguistic polishing and grammatical correction. They made no intellectual contribution to the methodology or experimental results. Since this usage aligns with standard editorial assistance, we did not include a dedicated section describing their role.*