Responsible NLP Checklist

Paper title: *All That Glisters Is Not Gold: A Benchmark for Reference-Free Counterfactual Financial Misinformation Detection*
Authors: *Yuechen Jiang, Zhiwei Liu, Yupeng Cao, Yueru He, Ziyang Xu, Chen Xu, Zhiyang Deng, Prayag Tiwari, Xi Chen, Alejandro Lopez-Lira, Jimin Huang, Junichi Tsujii, Sophia Ananiadou*

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?
For Limitation section and Ethical Consideration section

☑ 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?
No. The dataset is constructed from publicly available financial news articles (Yahoo Finance) and contains only corporate-level information. It does not include personally identifying information or content related to private individuals. All synthetic perturbations are restricted to financial narratives (e.g., numerical values, sentiment, or causal relations) and explicitly avoid defamatory or harmful content. Ethical considerations and potential biases are discussed in the Ethical Considerations section.

☑ 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?
Yes. We report detailed dataset statistics, including the number of articles, paragraph pairs, manipulation categories, and annotation agreement metrics. The final dataset contains 1,826 validated originalperturbed paragraph pairs across four manipulation types. Detailed statistics are provided in Section 2 and related appendices.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes. We describe the experimental setup including model selection, task definitions, prompting protocols, and evaluation metrics. All models are evaluated under a unified prompting framework, primarily in zero-shot settings, with few-shot variants used for ablation. Since the study evaluates

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.

pretrained LLMs via API-based inference, no model training or hyperparameter tuning is performed. Details are provided in Section 2.3 and Section 3.

☑ 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?
Yes. We report comprehensive descriptive statistics for experimental results, including accuracy, precision, recall, macro-F1, MCC, and AUROC across multiple models and tasks. Results are reported per task and per manipulation category, with additional human baseline comparisons. All evaluations are conducted on fixed datasets with deterministic protocols. Details are provided in Section 3.

☑ **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.?
Yes. Annotators followed standardized instructions and decision rules for both category correctness and rewrite validity. Detailed annotation guidelines and instructions are provided in the appendices.

☒ 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. All annotation and validation were conducted by volunteer contributors and co-authors of the paper. No external recruitment or monetary compensation was involved. All annotators were trained and followed standardized annotation guidelines.

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)

☑ **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?
Used GPT4.1 to generate the data and refine the task prompt