Responsible NLP Checklist

Paper title: *Large Language Models Require Curated Context for Reliable Political Fact-CheckingEven with Reasoning and Web Search*
Authors: *Matthew R. DeVerna, Kai-Cheng Yang, Harry Yaojun Yan, Filippo Menczer*

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?
Section 4 (Discussion) and Section 5 (Limitations). Risks include: deployment of unreliable LLM-based fact-checkers by everyday users or professional organizations; ideological bias in web-search citation patterns (skew toward left-leaning sources); downstream harms from inaccurate fact-checks on public belief and trust; and structural risks to the fact-checking ecosystem, as wide adoption of LLM-based tools could reduce traffic and revenue to the organizations whose work these pipelines depend on.

☑ **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?
(left blank)

☑ 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?
Sections 2 (Data and Methods), 3 (Results), and Appendix (Performance Statistics, including full breakdowns by model, retrieval setting, and veracity label in Tables A1A3).

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Sections 2 (Data and Methods) and 3 (Results) and the Appendix. The experimental setup is described in Section 2, including the zero-shot and Curated RAG conditions, retrieval settings (k=3, 6, 9), temperature (set to 0), and prompt design. Full prompt details and a robustness analysis using a simpler prompt are provided in the Appendix.

☑ 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,

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.

etc. or just a single run?
Sections 2 (Data and Methods) and 3 (Results) and the Appendix. Results report macro F1 scores across all models and retrieval settings, with mean improvements and standard deviations across k values noted where relevant. Figure captions clarify what is being reported (e.g., averages, distributions, error bars). Full performance tables with precision, recall, F1, and support are provided in the Appendix.

☒ **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)

☑ **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 for coding assistance and for copy editing of the manuscript.