Responsible NLP Checklist

Paper title: *Temporal Leakage in Search-Engine Date-Filtered Web Retrieval: A Retrospective Forecasting Case Study*
Authors: *Ali El Lahib, Ying-Jieh Xia, Zehan Li, Yuxuan Wang, Xinyu Pi*

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?
Our work focuses on auditing and improving the technical reliability of offline evaluation methodologies for forecasting agents. We do not introduce new models, datasets involving private individuals, or capabilities that we foresee posing immediate societal risks or safety concerns beyond standard research limitations.

☑ **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?
The dataset consists of public web search results retrieved via the Google Search and DuckDuckGo APIs. We did not explicitly check for personally identifying or offensive content, as the research objective is to audit standard retrieval behavior. The forecasting questions are curated from Metaculus tournaments.

☑ 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 3 (Table 1 reporting dataset-level leakage profile on 393 Google / 389 DuckDuckGo questions), Section 4.2 (Table 2 reporting the 93 binary questions subset used for the Brier score experiment), and Section 4.4 (Table 3 reporting per-cutoff-year URL statistics).

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3 describes the data collection and experimental setup. Section 3.2 details the LLM-as-a-judge model (gpt-oss-120b, temperature 0.5). Appendix A.1 details the document processing and the embedding model Qwen-0.6B. Appendix C contains the full prompts used for query generation, LLM judging, and the forecasting experiments.

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.

☑ 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?
Section 4.1 (Table 1) reports leakage prevalence. Section 4.2 (Table 2) reports Mean and Median Brier scores across retrieval conditions. Section 4.4 (Table 3) reports URL-level post-cutoff rates stratified by cutoff year for both search engines. Appendix A.2 (Figure 2) shows the human-LLM confusion matrix.

☑ **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.?
Section 3.3 describes the annotation protocol, in which annotators used the same scoring rubric provided to the LLM judge. The full rubric is included in Appendix C.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 the authors of the paper, so no recruitment or payment was involved.

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?
Discussed in the Acknowledgments section. We used LLMs only for sentence-level polishing (clarity, wording, and grammatical corrections) and limited implementation assistance for small refactors or boilerplate. All LLM-suggested changes were reviewed, edited as needed, and verified by the authors.