← Go to **ACL ARR 2025 October** homepage (/group?id=aclweb.org/ACL/ARR/2025/October)

4 Versions

When Benchmarks Age: Temporal Misalignment through Large Language Model Factuality Evaluation

(/pdf?id=6e5gG9usoy)

Xunyi Jiang (/profile?id=~Xunyi_Jiang2), Dingyi Chang (/profile?id=~Dingyi_Chang1), Julian McAuley (/profile?id=~Julian_McAuley1), Xin Xu (/profile?id=~Xin_Xu8)

07 Oct 2025 (modified: 02 Jan 2026) ACL ARR 2025 October Submission
October, Senior Area Chairs, Area Chairs, Reviewers, Authors
Revisions (/revisions?id=6e5gG9usoy)

TL;DR: This paper provides a testbed to explore the temporal misalignment of the static benchmarks with present LLMs, and real world facts, and its impacts on LLM factuality evaluation.

Abstract:
The rapid evolution of large language models (LLMs) and the real world has outpaced the static nature of widely used evaluation benchmarks, raising concerns about their reliability for evaluating LLM factuality. While substantial works continue to rely on the popular but old benchmarks, their temporal misalignment with real-world facts and modern LLMs, and their effects on LLM factuality evaluation remain underexplored. Therefore, in this work, we present a systematic investigation of this issue by examining five popular factuality benchmarks and eight LLMs released across different years. An up-to-date fact retrieval pipeline and three metrics are tailored to quantify benchmark aging and its impact on LLM factuality evaluation. Experimental results and analysis illustrate that a considerable portion of samples in the widely used factuality benchmarks are outdated, leading to unreliable assessments of LLM factuality. We hope our work can provide a testbed to assess the reliability of a benchmark for LLM factuality evaluation and inspire more research on the benchmark aging issue.

Paper Type: Short
Research Area: Resources and Evaluation
Research Area Keywords: evaluation methodologies, evaluation
Contribution Types: NLP engineering experiment, Data analysis
Languages Studied: English
Previous URL: /forum?id=ZcMCcuMJKC (/forum?id=ZcMCcuMJKC)
Explanation Of Revisions PDF: pdf (/attachment?id=6e5gG9usoy&name=explanation_of_revisions_PDF)
Reassignment Request Area Chair: No, I want the same area chair from our previous submission (subject to their availability).
Reassignment Request Reviewers: No, I want the same set of reviewers from our previous submission (subject to their availability)
Software: zip (/attachment?id=6e5gG9usoy&name=software)
Preprint: no
Preprint Status: We are considering releasing a non-anonymous preprint in the next two months (i.e., during the reviewing process).
Preferred Venue: EACL
Consent To Share Data: yes
Consent To Share Submission Details: On behalf of all authors, we agree to the terms above to share our submission details.

A1 Limitations Section: This paper has a limitations section.
A2 Potential Risks: N/A
B Use Or Create Scientific Artifacts: Yes
B1 Cite Creators Of Artifacts: Yes
B1 Elaboration: Section 2 and 3, Appendix B and C
B2 Discuss The License For Artifacts: N/A
B3 Artifact Use Consistent With Intended Use: Yes
B3 Elaboration: Section 2 and 3, Appendix B and C
B4 Data Contains Personally Identifying Info Or Offensive Content: N/A
B5 Documentation Of Artifacts: Yes
B5 Elaboration: Section 2 and 3, Appendix B and C
B6 Statistics For Data: Yes
B6 Elaboration: Section 2 and 3, Appendix B, C and D
C Computational Experiments: Yes
C1 Model Size And Budget: Yes
C1 Elaboration: Section 2 and 3, Appendix B and C
C2 Experimental Setup And Hyperparameters: Yes
C2 Elaboration: Section 2 and 3, Appendix B and C
C3 Descriptive Statistics: Yes
C3 Elaboration: Section 2 and 3, Appendix B and C
C4 Parameters For Packages: Yes
C4 Elaboration: Section 2 and 3, Appendix B and C
D Human Subjects Including Annotators: Yes
D1 Instructions Given To Participants: Yes
D1 Elaboration: Section 2, Appendix B and C
D2 Recruitment And Payment: Yes
D2 Elaboration: Section 2 and 3, Appendix B and C
D3 Data Consent: N/A
D4 Ethics Review Board Approval: N/A
D5 Characteristics Of Annotators: Yes
D5 Elaboration: Section 2 and 3, Appendix B and C
E Ai Assistants In Research Or Writing: Yes
E1 Information About Use Of Ai Assistants: Yes
E1 Elaboration: Section 2 and 3, Appendix B and C
Author Submission Checklist: yes
Association For Computational Linguistics - Blind Submission License Agreement: On behalf of all authors, I agree for this and all previous versions of this submission
Submission Number: 2602

Meta Review of Submission2602 by Area Chair broP

Meta Review by Area Chair broP 05 Dec 2025, 06:56 (modified: 15 Dec 2025, 11:12)

Senior Area Chairs, Area Chairs, Authors, Reviewers Submitted, Program Chairs, Commitment Readers

Revisions (/revisions?id=ZAsTcpp99P)

Metareview:
This paper investigates the temporal misalignment between static evaluation benchmarks and evolving real-world facts in the context of LLM factuality evaluation. The authors analyze five popular factuality benchmarks (TriviaQA, BoolQ, Natural Questions, TruthfulQA, SelfAware) and eight LLMs, introducing three metrics: (i) Dataset Drift Score (DDS), (ii) Evaluation Misleading Rate (EMR), and (iii) Temporal Alignment Gap (TAG), to quantify benchmark staleness. The work finds that substantial portions of these benchmarks contain outdated information, potentially penalizing models that provide correct, up-to-date answers.

However, several concerns persist: (1) the novelty is primarily in quantification rather than methodological innovation, (2) the authors did not propose any solutions, (3) the study focuses exclusively on factual QA, without demonstrating applicability beyond this domain, (4) time-sensitive sample extraction, fact retrieval, and answer alignment all rely on LLMs with limited human verifications.

Summary Of Reasons To Publish:

1. The work provides strong evidence that modern LLMs often align more closely with current, real-world facts than with the outdated answers in static benchmarks.
2. It also offers valuable insights, such as the finding that larger models tend to be more robust and better adapted to temporal changes in information
3. The study introduces new quantitative metrics, specifically the Drift Score (DS) and the Factual Staleness Score (FSS), providing the research community with concrete tools to measure the extent of this temporal drift (if they are able to extract the new gold labels).
4. The paper is well written.

Summary Of Suggested Revisions:
(1) The novelty is primarily in quantification rather than methodological innovation.

(2) The authors did not propose any solutions.

(3) The study focuses exclusively on factual QA, without demonstrating applicability beyond this domain.

(4) Time-sensitive sample extraction, fact retrieval, and answer alignment all rely on LLMs with limited human verifications.

Overall Assessment: 3 = Findings: I think this paper could be accepted to the Findings of the ACL.
Reported Issues: No
Publication Ethics Policy Compliance: I used a privacy-preserving tool exclusively for the use case(s) approved by PEC policy, such as language edits

Add: **Author-Editor Confidential Comment**

Official Review of Submission2602 by Reviewer AozG

Official Review by Reviewer AozG 10 Nov 2025, 12:30 (modified: 15 Dec 2025, 11:12)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer AozG, Commitment Readers

Revisions (/revisions?id=r0cqDRPkDN)

Paper Summary:
This paper investigates the "temporal misalignment" between static evaluation benchmarks, modern LLMs, and the real world. By extracting time-sensitive questions and comparing them against current web search results, the authors demonstrate that widely-used benchmarks often contain outdated facts.

The study finds that modern LLMs frequently provide more up-to-date answers than the benchmarks' gold standards, leading to unreliable and unfair performance evaluations.

Summary Of Strengths:

- The work provides strong evidence that modern LLMs often align more closely with current, real-world facts than with the outdated answers in static benchmarks. It also offers valuable insights, such as the finding that larger models tend to be more robust and better adapted to temporal changes in information
- The study introduces new quantitative metrics, specifically the Drift Score (DS) and the Factual Staleness Score (FSS), providing the research community with concrete tools to measure the extent of this temporal drift (if they are able to extract the new gold labels).
- The paper is well written and appendices contain some nice information.

Summary Of Weaknesses:

- The most interesting number that this paper from some reason does not present is the TS*DDS which is actually what driving the impact, looking at this number, we see that in fact the total % of examples in some of the benchmarks is large (BoolQ) and for some very small (TriviaQA), the authors should be truthful about this and say that in fact for 3/5 benchmarks tested, only <3.16% of the examples are impacted:

```
Dataset,TS % (from Table 2),DDS % (from Table 1),% of Total Dataset that is Outdated
TriviaQA,2.22%,37.05%,2.22% × 37.05% = 0.82%
BoolQ,13.76%,63.78%,13.76% × 63.78% = 8.77%
Natural Question,10.19%,24.19%,10.19% × 24.19% = 2.46%
TruthfulQA,19.58%,36.88%,19.58% × 36.88% = 7.22%
SelfAware,11.15%,28.26%,11.15% × 28.26% = 3.15%
```

Moreover, if one takes into account the real effect (with the EMR) it is even smaller. So the paper should consider this corruption in comparison to the bootstrap score and other effects described in the papers in the related work.

- The paper's central premise—that static benchmarks grow outdated—is well-known, with existing work like RealTimeQA and FreshQA already proposing solutions. While your quantification of the problem on older benchmarks is useful, it's not entirely clear what the novel, forward-looking contribution is beyond confirming an established issue. What is the single key takeaway that the community doesn't already intuit?
- The paper expertly diagnoses the problem but is light on proposing a scalable cure. You conclude by calling for "temporally aware strategies," but don't offer a concrete, novel proposal for how the community should adapt. Should all benchmarks now require a web-search component? Should they be retired after a certain date? A section discussing the trade-offs of potential solutions would significantly strengthen the paper's impact.
- The "Latest Fact Retrieval" stage seems incredibly complex and expensive to run, relying on multiple LLM calls and two different search APIs. Is this approach practical for maintainers of benchmarks to adopt? The paper should discuss the feasibility and cost implications of this method as a potential solution.

Comments Suggestions And Typos:

See Summary Of Weaknesses

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.
Soundness: 3 = Acceptable: This study provides sufficient support for its main claims. Some minor points may need extra support or details.
Excitement: 2 = Potentially Interesting: this paper does not resonate with me, but it might with others in the *ACL community.
Overall Assessment: 2.5 = Borderline Findings
Ethical Concerns:

There are no concerns with this submission

Needs Ethics Review: No
Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.
Datasets: 1 = No usable datasets submitted.
Software: 3 = Potentially useful: Someone might find the new software useful for their work.

Knowledge Of Or Educated Guess At Author Identity: No
Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources
Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.
Publication Ethics Policy Compliance: I used a privacy-preserving tool exclusively for the use case(s) approved by PEC policy, such as language edits

Add: **Author-Editor Confidential Comment**

Official Comment by Authors

Official Comment

by Authors (Xunyi Jiang (/profile?id=~Xunyi_Jiang2), Dingyi Chang (/profile?id=~Dingyi_Chang1), Julian McAuley (/profile?id=~Julian_McAuley1), Xin Xu (/profile?id=~Xin_Xu8))

22 Nov 2025, 18:59 (modified: 15 Dec 2025, 11:12)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer AozG, Commitment Readers

Revisions (/revisions?id=qL5zRakceN)

Comment:

(1) Clarity of Metrics

Thank you for the insightful comment. We appreciate the reviewer's computation of TS×DDS, and we fully agree that understanding the total fraction of impacted examples is important. However, we respectfully clarify two key misunderstandings:

- Our measured evaluation bias (EMR and TAG) mathematically depends only on the time-sensitive subset, not the whole dataset. The impact of outdated labels occurs only when: a) The question is time-sensitive (TS), and b) The benchmark gold answer is outdated (DDS), and c) The LLM gives the correct current answer. Therefore, evaluating corruption using TS×DDS relative to the entire dataset underestimates the actual bias. Our measures of DDS, EMR and TAG are metrics that quantify evaluation errors within the affected time-sensitive subset.
- EMR shows that bias is non-trivial, even if TS×DDS is numerically small. Across all models and benchmarks, EMR ranges: 8–24% for BoolQ, 10–23% for TriviaQA, 14–22% for SelfAware, and Up to 24% on Natural Questions. These are not small in the context of a modern LLM evaluation protocol. Even when TS×DDS is <3%, the LLMs are regularly penalized for being correct, which is exactly the phenomenon our work investigates. Thus, TS×DDS is a misleading proxy. EMR and TAG better capture the real evaluation corruption.
- TS% also underestimates time-sensitive content. As we mentioned in Appendix B.1 that TS% is strict, only counting questions whose answers demonstrably change over time. But many questions rely on context passages that themselves contain outdated information. For example, in BoolQ, outdated passages systematically anchor models to obsolete facts (Table 6). For example, question in the BoolQ: "Do you have to pay to use the ITV hub" is true, with passage indicating that the subscription is needed. ("ITV brought out a paid subscription service called ITV Hub+ which enables viewers to watch and download shows without adverts. While abroad in another EU member state, subscribers will continue to receive access but will be limited to catch-up."). However, the current up-to-date information is: "Viewers can either watch for free, supported by ads, or pay £5.99/month or £59.99/year to stream ad-free." The passage is outdated and leads LLMs to respond to wrong answers. This leads to much larger degradation than TS% alone suggests.

(2) Novelty and Contribution Beyond Prior Work

We appreciate the reviewer highlighting prior work such as RealTimeQA and FreshQA. We clarify that while these efforts propose solutions to temporal drift, no prior work has systematically quantified how severe benchmark aging already is across widely-used factuality benchmarks, nor its concrete impact on evaluation reliability. Our key novelty lies not in observing that benchmarks age, but in providing the first unified testbed to have a large-scale empirical measurement of how much, where, and to what extent this affects today's LLM evaluations.

(3) Provide a scalable cure

Thank you for raising this important point that discussing potential solutions would improve the paper. At the same time, we respectfully clarify that our contribution is intentionally diagnostic rather than prescriptive. As a short paper, our intent is to provide an empirical analysis and a testbed that highlights a growing problem, rather than to fully solve it.

(4) Feasibility and Cost of Latest Fact Retrieval

We agree that cost and practicality are crucial for benchmark maintainers, so we quantified the end-to-end expense of our pipeline on the most time-sensitive subset in our study. Using Natural Questions, we identify 798 time-sensitive questions. Running our full "Latest Fact Retrieval" stage on these 798 questions costs $49.53 in total:

- Web search cost: We rely on Google Custom Search and Brave Search for document retrieval. The combined API fees for these 798 questions are $48.50.
- LLM extraction cost. For the Google-search branch, we use an open-source Qwen-2.5-14B-Instruct model locally on a single RTX A6000 (no API charge). For the Brave-search branch, only 22.3% of the selected questions have Wikipedia as the retrieved source, and for these we use gpt-4o-mini to extract answers from the Wikipedia passages. Using the current OpenAI pricing (4 dollars per million input tokens, 16 dollars per million output tokens), we estimate the total answer-extraction cost to be $1.03.

Add: **Author-Editor Confidential Comment**

Official Review of Submission2602 by Reviewer DNVe

Official Review by Reviewer DNVe 02 Nov 2025, 21:42 (modified: 15 Dec 2025, 11:12)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer DNVe, Commitment Readers

Revisions (/revisions?id=krfdsNOQ9o)

Paper Summary:
This paper investigates the temporal misalignment problem in LLM factuality evaluation, i.e., the issue that static benchmarks become outdated as real-world facts evolve. The authors systematically analyze five popular factuality benchmarks (TriviaQA, BoolQ, NaturalQuestions, TruthfulQA, SelfAware) and eight modern LLMs, introducing three quantitative metrics—Dataset Drift Score (DDS), Evaluation Misleading Rate (EMR), and Temporal Alignment Gap (TAG)—to measure benchmark staleness and its effect on evaluation reliability. Results reveal that a substantial portion of factuality benchmarks (up to 63.8%) contain outdated information, and modern models such as GPT-4o-mini or Qwen2.5-14B may be unfairly penalized for providing correct but temporally updated answers.

Adequacy Of Revisions:
Yes

Summary Of Strengths:
1. Discusses a real and under-discussed problem in LLM evaluation.
2. The metrics (DDS, EMR, TAG) are intuitive and useful for quantifying benchmark staleness.
3. The analysis is systematic and transparent, with clearly documented experimental details.

Summary Of Weaknesses:

1. The contribution is incremental: it diagnoses an obvious issue without proposing any mitigation or theoretical insight.
2. The novelty mainly lies in quantification rather than methodological innovation.
3. Limited to factual QA; no exploration of other domains.

Overall, this paper is well-executed and timely, but conceptually modest. It provides a clean empirical study and could serve as a useful reference point or diagnostic tool rather than a breakthrough contribution.
Comments Suggestions And Typos:
See weakness above.

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.
Soundness: 2.5
Excitement: 2.5
Overall Assessment: 3 = Findings: I think this paper could be accepted to the Findings of the ACL.
Ethical Concerns:
There are no concerns with this submission

Needs Ethics Review: No
Reproducibility: 5 = They could easily reproduce the results.
Datasets: 5 = Enabling: The newly released datasets should affect other people's choice of research or development projects to undertake.
Software: 5 = Enabling: The newly released software should affect other people's choice of research or development projects to undertake.
Knowledge Of Or Educated Guess At Author Identity: No
Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources
Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.
Publication Ethics Policy Compliance: I did not use any generative AI tools for this review

Add: **Author-Editor Confidential Comment**

Responses to Weaknesses

Official Comment

by Authors (Xunyi Jiang (/profile?id=~Xunyi_Jiang2), Dingyi Chang (/profile?id=~Dingyi_Chang1), Julian McAuley (/profile?id=~Julian_McAuley1), Xin Xu (/profile?id=~Xin_Xu8))

22 Nov 2025, 19:21 (modified: 15 Dec 2025, 11:12)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer DNVe, Commitment Readers

Revisions (/revisions?id=vMWUrMwbuR)

Comment:
We thank the reviewer for the positive assessment of our revisions and for recognizing the importance of addressing temporal misalignment, the usefulness of our metrics, and the clarity of our analysis. We would like to clarify the above concerns:

1.2.

We appreciate the reviewer’s perspective, but we respectfully clarify that our contribution goes beyond reporting obvious observations. Our work offers a focused and practical testbed for assessing the reliability of widely used factuality benchmarks under temporal misalignment, which despite being recognized, has not been systematically quantified or analyzed in prior literature. Specifically, we design a complete evaluation pipeline, including time-sensitive sample extraction, an up-to-date fact-retrieval workflow, and

three dedicated metrics for measuring temporal misalignment among benchmark labels, LLM outputs, and real-world facts (e.g., Table 1 and Figure 1 and 2 in the paper). This framework reveals substantial and previously unmeasured aging effects in benchmarks such as BoolQ and NaturalQuestions, directly impacting the reliability of LLM factuality evaluation. As a short paper, we follow the guideline “a small, focused contribution that can be made in a few pages,” to provide diagnostic tools and empirical insights that we believe will help the community make more reliable use of existing benchmarks.

3.

We appreciate the reviewer’s comment, but our study is intentionally scoped to LLM factuality evaluation, which is precisely the domain where temporal misalignment poses the most direct and measurable impact. Other domains, such as summarization, reasoning, or math, do not show temporal drift and thus cannot be evaluated for temporal misalignment. Our focus on factual QA benchmarks is thus deliberate for our research topic. We’re open to a kind discussion.

Add: **Author-Editor Confidential Comment**

Replying to Responses to Weaknesses

Response to Authors

Official Comment by Reviewer DNVe 23 Nov 2025, 04:32 (modified: 15 Dec 2025, 11:12)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer DNVe, Commitment Readers

Revisions (/revisions?id=nFpbBIWcxP)

Comment:
Thank you for your efforts, and I will maintain my score.

Add: **Author-Editor Confidential Comment**

Official Review of Submission2602 by Reviewer QW8H

Official Review by Reviewer QW8H 01 Nov 2025, 12:05 (modified: 15 Dec 2025, 11:12)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer QW8H, Commitment Readers

Revisions (/revisions?id=q1N7O6wiSe)

Paper Summary:
This paper addresses the temporal misalignment between static benchmarks, modern LLMs, and real-world facts. The authors develop a pipeline to (1) identify time-sensitive questions from popular QA benchmarks (TriviaQA, BoolQ, Natural Questions, TruthfulQA, SelfAware), (2) retrieve updated answers via Wikipedia/web search, and (3) compare benchmark gold labels, LLM responses, and current facts using three metrics (Drift Score, Cohen’s Kappa, Factual Staleness Score). Experiments on 10 LLMs show substantial divergence between benchmark labels and current reality. The authors added more clarifications in the new version compared to the last submission.

Summary Of Strengths:

- This is a timely and important problem, especially as LLMs are increasingly expected to remain factually up to date.
- The notion of temporal misalignment is well defined
- They provide insightful findings with new metrics introduced

Summary Of Weaknesses:
The authors made clarifications regarding related works, like RealTimeQA and FreshQA. However, the following weaknesses I mentioned in the last submission remain the same.

- Time-sensitive sample extraction, fact retrieval, and answer alignment all rely on LLMs with limited human verifications.

- The paper identifies the problem of temporal misalignment but does not offer concrete proposals for dynamic benchmark design.

Comments Suggestions And Typos:
See the weaknesses section.

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.
Soundness: 3 = Acceptable: This study provides sufficient support for its main claims. Some minor points may need extra support or details.
Excitement: 3.5
Overall Assessment: 3 = Findings: I think this paper could be accepted to the Findings of the ACL.
Ethical Concerns:
There are no concerns with this submission

Needs Ethics Review: No
Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.
Datasets: 3 = Potentially useful: Someone might find the new datasets useful for their work.
Software: 3 = Potentially useful: Someone might find the new software useful for their work.
Knowledge Of Or Educated Guess At Author Identity: No
Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources
Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.
Publication Ethics Policy Compliance: I used a privacy-preserving tool exclusively for the use case(s) approved by PEC policy, such as language edits

Add: **Author-Editor Confidential Comment**

Responses to Weaknesses

Official Comment

by Authors (Xunyi Jiang (/profile?id=~Xunyi_Jiang2), Dingyi Chang (/profile?id=~Dingyi_Chang1), Julian McAuley (/profile?id=~Julian_McAuley1), Xin Xu (/profile?id=~Xin_Xu8))

22 Nov 2025, 19:22 (modified: 15 Dec 2025, 11:12)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer QW8H, Commitment Readers

Revisions (/revisions?id=Whq5fqbuVy)

Comment:

- We thank the reviewer for raising this concern. While our pipeline uses LLMs as part of the workflow (e.g., time-sensitive sample detection, fact extraction, and answer alignment), these components are used with valid human verification. As detailed in Appendix C.2–C.4 of the paper , each stage is complemented by systematic human evaluation to ensure reliability: Time-sensitive sample extraction (Appendix C.2): We conduct human annotation on 150 samples using clear guidelines, achieving 100% recall, 90% accuracy, and 0.83 Cohen's Kappa, showing strong agreement between LLM predictions and human labels. Fact retrieval quality (Appendix C.3): Three domain experts manually reviewed 105 retrieved answers and achieved 89.52% factual accuracy. LLM-as-judge answer alignment (Appendix C.4): We again conduct human evaluation, yielding 97% accuracy and 0.72 Cohen's Kappa, confirming that LLM agreement judgments are highly consistent with human judgments. These human-verified evaluations demonstrate that our pipeline does not rely blindly on LLMs; rather, LLM components are

validated, calibrated, and shown to be reliable for this analysis. Given the scale of the benchmarks we study, combining LLM assistance with human verification is a practical and commonly adopted methodology in recent LLM evaluation research[1][2][3].

[1] LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. ACL 2025. [2] VIBECHECK: Discover & Quantify Qualitative Differences in LLM Behavior. ICLR 2025. [3] Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS 2023.

- We thank the reviewer for the thoughtful comment. As a short paper, our work focuses on constructing a comprehensive testbed for assessing the reliability of widely used factuality benchmarks under temporal misalignment, which, despite being recognized, has not been systematically quantified or analyzed in prior literature. Our contribution is to provide the community with a testbed, evaluation pipeline, and concrete metrics (DDS, EMR, TAG), which allow researchers to audit and measure temporal misalignment in current widely used benchmarks. As a short paper, we follow the guideline "a small, focused contribution that can be made in a few pages," to provide diagnostic tools and empirical insights that we believe will help the community make more reliable use of existing benchmarks.
Dynamic benchmark design is an important research topic. We will explore this in future work.

Add: **Author-Editor Confidential Comment**