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

# Quantifying Cognitive Bias Induction in LLM-Generated Content

PDF (/pdf?id=jnd8pFD4GL)

*Abeer Alessa (/profile?id=~Abeer_Alessa2), Akshaya Thenkarai Lakshminarasimhan (/profile?id=~Akshaya_Thenkarai_Lakshminarasimhan1), Param Somane (/profile?id=~Param_Somane1), Julian Skirzynski (/profile?id=~Julian_Skirzynski1), Julian McAuley (/profile?id=~Julian_McAuley1), Jessica Maria Echterhoff (/profile?id=~Jessica_Maria_Echterhoff1)*

28 Jul 2025 (modified: 19 Aug 2025) ACL ARR 2025 July Submission Everyone Revisions (/revisions?id=jnd8pFD4GL) BibTeX 

**Abstract:**
Large language models (LLMs) are increasingly integrated into applications ranging from shopping review summarization to medical diagnosis support, where they affect human decisions. Although LLMs often perform well on common evaluation metrics, they may inherit societal or cognitive biases. When humans get exposed to content that was processed by an LLM that shows bias, for example, a summary of a piece of text, any bias introduced through content processing by the LLM will inadvertently have an effect on the human. We investigate the extent to which LLMs expose users to biased content. We assess three LLM families in summarization and news fact-checking tasks, evaluating the consistency of LLMs with their context and their tendency to hallucinate. Our findings show that LLMs expose users to content that changes the sentiment of the context in 21.86% of cases, hallucinates on post-knowledge-cutoff data questions in 60.33% of cases, and highlights context from earlier parts of the prompt (primacy bias) in 5.94% of cases. To alleviate the issue, we evaluate 18 distinct mitigation methods across three LLM families and find that targeted interventions can be effective.

**Paper Type:** Long
**Research Area:** Ethics, Bias, and Fairness
**Research Area Keywords:** model bias/fairness evaluation, model bias/unfairness mitigation, ethical considerations in NLP applications, reflections and critiques
**Contribution Types:** Model analysis & interpretability, Data resources
**Languages Studied:** English
**Preprint:** yes
**Preprint Status:** There is a non-anonymous preprint (URL specified in the next question).
**Existing Preprints:** https://arxiv.org/pdf/2507.03194 (https://arxiv.org/pdf/2507.03194)
**Preferred Venue:** AACL
**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.
**Association For Computational Linguistics - Blind Submission License Agreement:** On behalf of all authors, I agree
**Previous URL:** /forum?id=KdhQcCwxkt (/forum?id=KdhQcCwxkt)
**Explanation Of Revisions PDF:** pdf (/attachment?id=jnd8pFD4GL&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:** Yes, I want a different set of reviewers
**Justification For Not Keeping Action Editor Or Reviewers:** The previous review did not offer any specific technical critiques. It also dismissed the value of our contribution without explaining how our approach, metrics, or findings are redundant or

lacking. The review also rated the reproducibility and dataset 5, but the overall assessment was one without elaboration.
**Data:** zip (/attachment?id=jnd8pFD4GL&name=data)
**A1 Limitations Section:** This paper has a limitations section.
**A2 Potential Risks:** Yes
**A2 Elaboration:** 7
**B Use Or Create Scientific Artifacts:** Yes
**B1 Cite Creators Of Artifacts:** Yes
**B1 Elaboration:** 4.2
**B2 Discuss The License For Artifacts:** Yes
**B2 Elaboration:** 7
**B3 Artifact Use Consistent With Intended Use:** Yes
**B3 Elaboration:** 4.2
**B4 Data Contains Personally Identifying Info Or Offensive Content:** N/A
**B5 Documentation Of Artifacts:** Yes
**B5 Elaboration:** 1,3,4.2, Appendix B
**B6 Statistics For Data:** Yes
**B6 Elaboration:** 4.2
**C Computational Experiments:** Yes
**C1 Model Size And Budget:** Yes
**C1 Elaboration:** 7
**C2 Experimental Setup And Hyperparameters:** Yes
**C2 Elaboration:** Appendix B
**C3 Descriptive Statistics:** Yes
**C3 Elaboration:** 5
**C4 Parameters For Packages:** Yes
**C4 Elaboration:** Appendix B
**D Human Subjects Including Annotators:** No
**D1 Instructions Given To Participants:** N/A
**D2 Recruitment And Payment:** N/A
**D3 Data Consent:** N/A
**D4 Ethics Review Board Approval:** N/A
**D5 Characteristics Of Annotators:** N/A
**E Ai Assistants In Research Or Writing:** Yes
**E1 Information About Use Of Ai Assistants:** No
**E1 Elaboration:** ChatGPT and Grammarly were used only to improve grammar, latex formatting and writing clarity. Their use did not influence the research content or findings, and therefore, we did not include this information in the main paper.
**Author Submission Checklist:** yes
**Submission Number:** 932

## Discussion (?id=jnd8pFD4GL#discussion)

Filter by reply type... | Filter by author... | Search keywords... | Sort: Newest First

Everyone | Submission932... | Submission932 Area... | Submission932 Authors | Submission932... | Program Chairs | Submission932... | Submission932... | Submission932... | Submission932...

*7 / 7 replies shown*

Add: **Author-Editor Confidential Comment** **Withdrawal**

## Meta Review of Submission932 by Area Chair dYek

Meta Review by Area Chair dYek 22 Sept 2025, 06:59 (modified: 14 Oct 2025, 08:12)

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

Revisions (/revisions?id=ymlCsawmLC)

**Metareview:**
The paper is concerned with how LLMs alter content during summarization and fact-checking tasks using three different dimensions/metrics: sentiment drift, prompt position bias (primacy / recency), and hallucinations. The authors evaluate 3 families of LLMs and propose a substantial number of mitigation strategies, even though most of the strategies only make it into the appendix.

The reviewers disagree quite markedly on this paper (overall scores between 2 and 4). According to my own reading, the paper is somewhere in the middle: I think that this is a worthwhile and solidly carried out study, but one that makes a couple of fundamentally limiting assumptions (see below for details). In its current state, I consider this a good Findings paper.

**Summary Of Reasons To Publish:**
Task:

- Innovative question, three types of biases in two types of tasks: summarization and QA (but see below)

Method:

- Comprehensive set of LLMs evaluated
- Comprehensive set of mitigation strategies
- Good experimental design

Resource:

- novel factuality dataset for the news domain

Presentation:

- Well structured and written

**Summary Of Suggested Revisions:**

- In my view, the most important remaining concern, raised by Reviewer rs8Y, is that the study has currently not a lot to say about any distinctions between the three types of biases, and I share the reviewer's concern that depending on the application, some of these might be much more relevant than others. I feel that the authors' response, "we pick text types where this is not the case", is only partially convincing: yes, for reviews, sentiment change is arguably always bad; but I do not agree that for interviews, "all parts are equally valid". If this were true, there would be no point in summarization. The paper would be much stronger if it had a convincing story to tell about this point; the "decision making" pilot study reported on in the author response can be a part of that. (I have discounted rs8Y's concerns on the assumption that you will do so in a final version of the paper.)
- Also, the paper investigates prompt position bias. One reviewer points out that it is a little sloppy regarding the distinction between primacy bias (things at the beginning of the prompt) and recency bias (things at the end of the prompt). The analyses in the paper only concern primacy bias. The authors argue that the analysis can be extended to recency bias. I agree, but the fact remains that the analysis is not carried out for recency bias. I would suggest to either carry out this analysis as well (ideal) or focus the paper on primacy bias (including a good conceptual motivation for doing so).
- The reviewers ask for additional details on the news hallucination data (given in author response -> integrate into paper)
- The reviewers ask for additional details on the use of GPT4o-mini for sentiment labeling (given in author response -> integrate into paper). (I do not share the fundamental concerns about the use of LLMs as judges as long as good practice and transparency are observed.)
- The reviewers ask for additional motivation on the threshold for the primacy metric (given in author response -> integrate into paper)

- The reviewers ask for additional details on input length statistics (given in author response -> integrate into paper)
- The reviewers ask for additional statistical testing on the results in Table 1 (given in author response -> integrate into paper)

**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 did not use any generative AI tools for this review

Add: **Author-Editor Confidential Comment**

## Official Review of Submission932 by Reviewer T6h9

Official Review by Reviewer T6h9 31 Aug 2025, 13:04 (modified: 14 Oct 2025, 08:12)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer T6h9, Commitment Readers
Revisions (/revisions?id=2L0QHElvbA)

**Paper Summary:**
The paper investigates the extent of exposure of biased content by LLMs to their users. The authors support their claim by presenting how different kinds of cognitive bias including hallucinations are exhibited by 3 families of LLMs. The authors also propose several mitigation techniques to reduce the encountered bias.

**Summary Of Strengths:**
1. The paper is very well written and the motivation is clearly explained.
2. The authors experimented with different families of LLMs to support their claim.
3. The authors experimented with tasks of summarization and question answering to validate their findings.
4. The authors also release a self-updating dataset to analyze the hallucinations generated by the LLMs on the news articles published after the model's cut-off date.
5. 18 different types of mitigation techniques are evaluated to figure out the most effective techniques along with the suggestions for content alterations.
6. The authors quantitatively measure the extent of LLMs altering the relevant contextual content through sentiment change, hallucination, and biased emphasis in summaries.
7. The authors define three cognitive biases to evaluate how these biases affect decision making capabilities of users.
8. The experiments are meticulously designed and the results are clearly detailed.

**Summary Of Weaknesses:**
1. The authors use Primacy bias and Recency bias interchangeably. But section 3 introduces both the biases as two separate types. If they refer to the same bias, it is more prudent to use a single name for better clarity.
2. It is not clear whether the end segment for a document has any impact on the primacy bias score. This part is completely ignored in the primacy bias computation.
3. The authors should explain the reason how they choose their LLMs. Is it based on the performance of certain benchmarks? Why are other LLMs such as Gemma and Qwen not selected for the study?

**Comments Suggestions And Typos:**
NA

**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:** 4 = Strong: This study provides sufficient support for all of its claims. Some extra experiments could be nice, but not essential.
**Excitement:** 4 = Exciting: I would mention this paper to others and/or make an effort to attend its presentation in a conference.
**Overall Assessment:** 4 = Conference: I think this paper could be accepted to an *ACL conference.
**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:** 4 = Useful: I would recommend the new datasets to other researchers or developers for their ongoing work.
**Software:** 4 = Useful: I would recommend the new software to other researchers or developers for their ongoing 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 did not use any generative AI tools for this review

Add: **Author-Editor Confidential Comment**

## Thank you for the constructive comments on clarity, thoroughness, and mitigation evaluation.

Official Comment

by Authors (Julian Skirzynski (/profile?id=~Julian_Skirzynski1), Abeer Alessa (/profile?id=~Abeer_Alessa2), Jessica Maria Echterhoff (/profile?id=~Jessica_Maria_Echterhoff1), Param Somane (/profile?id=~Param_Somane1), +2 more (/group/edit?id=aclweb.org/ACL/ARR/2025/July/Submission932/Authors))

15 Sept 2025, 20:41 (modified: 14 Oct 2025, 08:12)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer T6h9, Commitment Readers

Revisions (/revisions?id=om0gSShrq4)

**Comment:**

We thank reviewer T6h9 for recognizing that our paper is well written, with clearly articulated motivations, carefully designed experiments, and a thorough evaluation of the mitigation techniques. We thank the reviewer for highlighting our self-updating dataset as a useful resource for other researchers.

Primacy bias and Recency bias: We appreciate reviewer T6h9 for pointing this out. We revised our paper accordingly and added that in Sections 3 and 4, both biases describe similar issues in content alteration of LLMs, incorrectly focusing on the start (primary) or end (recency) of the prompt. As primacy bias has been reported to be very prevalent in LLMs [1], we focus our quantitative analysis on primacy bias. However, our evaluation can be adapted to measure recency bias as well. We specifically designed our metric ψ_pri to measure the incorrect deviation of the model's focus on different parts of the prompt (for our use cases of interview and review summarization, all parts of the prompt should be relevant). We adapted our wording in the paper to be more precise and ensured the terms are used distinctly throughout the paper.

Primacy bias metric: This is an excellent point. Our primacy bias metric (ψ_pri) was intentionally designed to compare the beginning (s_b) to the middle (s_m), where initial content overshadows the body of the text. So the end segment does not have an impact on our primacy bias metric (ψ_pri). This was chosen to disentangle potential recency effects, which would mean a disproportionate focus on the end segment of the text.

We agree that the role of the end segment can give more information to complete the picture of positional bias. Our metric can be re-purposed to give more information on recency bias. We did calculate similarity with the end segment (s_e), and these results are reported in Table 1 for reference.

LLM's choice: We are happy to clarify reasoning on our initial model selection (detailed in Appendix B). We aim to cover a range of popular, high-performing open- and closed-source models of varying sizes to ensure our findings are not specific to a single architecture. Our work includes open-source models with small models (Phi-3-mini-4k-Instruct and Llama-3.2-3B-Instruct), medium-sized models (Llama-3-8B-Instruct), and closed-source (GPT-3.5-turbo).

We acknowledge that the LLM landscape is constantly evolving. To broaden the applicability of our findings, we have now run baseline evaluations on Qwen and Gemma models as requested and included them in Section 5 of the paper. The results are consistent with our original findings, showing similar patterns of framing changes and primacy bias.

For the MediaSum News dataset, the baseline metrics are:

- Qwen/Qwen3-4B-Instruct-2507: Framing Change (ϕ_frame): 45.9%; Primacy Bias (ψ_pri): 13.0%; Average Similarities: s̄_b = 0.898, s̄_m = 0.879, s̄_e = 0.869.
- google/gemma-3-27b-it: Framing Change (ϕ_frame): 41.9%; Primacy Bias (ψ_pri): 17.6%; Average Similarities: s̄_b = 0.902, s̄_m = 0.879, s̄_e = 0.868.

For the Amazon Reviews dataset, the baseline metrics are:

- Qwen/Qwen3-4B-Instruct-2507: Framing Change (ϕ_frame): 27.8%; Primacy Bias (ψ_pri): 19.6%; Average Similarities: s̄_b = 0.921, s̄_m = 0.896, s̄_e = 0.893.
- google/gemma-3-27b-it: Framing Change (ϕ_frame): 26.5%; Primacy Bias (ψ_pri): 23.7%; Average Similarities: s̄_b = 0.924, s̄_m = 0.896, s̄_e = 0.892. We thank the reviewer for this suggestion and believe this addition improves the camera-ready version.

References: [1] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. Transactions of the Association for Computational Linguistics

Add: **Author-Editor Confidential Comment**

## Official Review of Submission932 by Reviewer HufX

Official Review by Reviewer HufX 27 Aug 2025, 17:10 (modified: 14 Oct 2025, 08:12)

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

Revisions (/revisions?id=5yLFVL1PAJ)

**Paper Summary:**
The paper studies three types of bias in LLM responses: **framing** (sentiment drift), **primacy** (favouring the beginning of input), and **hallucination** (errors after training cutoff). It introduces evaluation metrics for each and reports results across several models and datasets. For hallucination, the authors present a new benchmark, "News Hallucination," built from pre- and post-cutoff news paired with negated counterfactuals. They also test 18 mitigation strategies, though only highlight a subset in the main text.

**Summary Of Strengths:**
The paper topic on biases like framing and hallucination affects trust in LLMs' responses, and defining measurable metrics for them is useful and timely.

The decomposition into three bias types (framing, primacy, hallucination) is clear and easy to follow.

The paper also contributes a new dataset (News Hallucination / NewsLensSync) and explores 18 mitigation ideas like Knowledge Boundary Awareness and Epistemic Tagging.

**Summary Of Weaknesses:**

The News Hallucination dataset is under-specified. The paper says it contains “2,801 real news items and their negated versions,” with counterfactuals generated by a transformer, but does not explain news sources, sampling, duplication handling, or whether humans validated any stage. It is also unclear how they separate predictable events (e.g., a president serving a known term) from unpredictable events (e.g., disasters, resignations). Since models have different cutoffs, the paper should also show that each item truly falls after the cutoff for each model. Without this, it is hard to know if results reflect real hallucinations or just predictable facts or cutoff mismatches.

The framing metric depends on GPT-4o-mini to label sentiment. The paper does not show the exact prompt, so it is unclear what the model was instructed to look for. Many texts mix positive and negative points, but the method forces one label. Without details on how such cases were handled, $\phi_{frame}$ risks oversimplifying framing differences.

The primacy metric (Sec. 4.0.2) uses a fixed threshold $\alpha = 0.05$ without justification or sensitivity checks. In Eq. 2, the decision only compares beginning vs. middle, ignoring the end segment with no explanation. This makes the metric feel arbitrary and incomplete.

The paper mentions input caps (≤4000 tokens) but does not report average or median lengths of each dataset. Length likely affects both primacy and framing, so this omission matters.

primacy results in Sec. 5.2 (Table 1) shows beginning (0.828–0.860) slightly above middle (0.822–0.842) and end (0.820–0.840), but the ranges overlap heavily. Without statistical tests or effect sizes, it is not clear if this is a real effect.

**Comments Suggestions And Typos:**
Please provide full dataset details for Post-Knowledge Cuttoff Data: news sources, sampling rules, deduplication, human checks for counterfactuals, and other considerations.

Report input length stats (mean, median, and bins) for datasets. Plot $\psi_{pri}$ and $\phi_{frame}$ across short vs. long inputs.

**Confidence:** 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.
**Soundness:** 3 = Acceptable: This study provides sufficient support for its main claims. Some minor points may need extra support or details.
**Excitement:** 2.5
**Overall Assessment:** 3 = Findings: I think this paper could be accepted to the Findings of the ACL.
**Best Paper Justification:**
N/A

**Limitations And Societal Impact:**
The paper has a Limitations section.

**Ethical Concerns:**
There are no concerns with this submission

**Needs Ethics Review:** No
**Reproducibility:** 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.
**Datasets:** 2 = Documentary: The new datasets will be useful to study or replicate the reported research, although for other purposes they may have limited interest or limited usability. (Still a positive rating)
**Software:** 2 = Documentary: The new software will be useful to study or replicate the reported research, although for other purposes it may have limited interest or limited usability. (Still a positive rating)
**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**

− = ≡

## Thank You for Your Thorough Review: Addressing Points on Dataset Specification, Metric Clarity, and Statistical Validation

Official Comment

by Authors ( Julian Skirzynski (/profile?id=~Julian_Skirzynski1), Abeer Alessa (/profile?id=~Abeer_Alessa2), Jessica Maria Echterhoff (/profile?id=~Jessica_Maria_Echterhoff1), Param Somane (/profile?id=~Param_Somane1), +2 more (/group/edit?id=aclweb.org/ACL/ARR/2025/July/Submission932/Authors))

15 Sept 2025, 22:05 (modified: 14 Oct 2025, 08:12)

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

Revisions (/revisions?id=d09mbZhUS1)

**Comment:**

We thank the reviewer for their valuable feedback and for acknowledging the clarity of our bias decomposition, the thoroughness of our mitigation strategies, and the contribution of our new dataset and metrics.

**The News Hallucination Dataset:** To address the concerns about the dataset's specification, we have added the following details to Section 4.2. Our dataset is built using the NewsAPI [1] from ~150,000 publishers. We queried for "politics" articles in English from a set of trusted, high-engagement domains: bbc.co.uk, reuters.com, apnews.com, npr.org, pbs.org, theguardian.com, wsj.com, nytimes.com, and propublica.org.

Our system, NewsLensSync, collects key article attributes and generates negated descriptions for hallucination analysis. Crucially, **we manually validate these negated descriptions** to ensure semantic correctness. To handle duplicates, our system ingests daily articles but performs deduplication during the LLM evaluation phase. This ensures the analysis dataset is clean while maintaining efficient data ingestion. At the time of our experiments, the dataset contained 2,801 unique political articles with validated negated counterparts.

We agree that distinguishing between predictable and unpredictable events would enhance the evaluation and have added this as a limitation. However, our method of comparing model responses to both true and negated events provides a consistent measure of hallucination.

For knowledge cut-offs, our data was collected after April 2025, which is after the training cut-off for all models used. Each data point is timestamped to allow future researchers to filter as needed.

| Model | Cut-off date (data freshness) |
|---|---|
| Llama3-8b-instruct | Mar 2023 |
| Llama 3.2-3b-instruct | Oct 2024 |
| Phi-3-mini-4k-instruct | Oct 2023 |

**Table 1: Models knowledge cut-off dates**

**The Framing Bias Metric:** We chose GPT-4o-mini after it achieved the highest accuracy (92%) in an internal evaluation against human ratings on 500 Amazon reviews. We have now added the exact prompt used for sentiment classification to the appendix: *"Can you classify the framing of this text (Positive, Negative, Neutral) and answer ONLY with the class label: \n + {text}"*. While this method targets broad sentiment shifts over subtle nuances, we agree a more detailed evaluation is an area for future work.

**Fixed Threshold for Primacy Bias:** We chose α = 0.05 as a simple heuristic. We acknowledge this as a limitation in Section 7 and an avenue for future work.

**Dataset Statistics:** We cap inputs at ≤ 4,000 tokens to fit within all models' context windows (e.g., Phi-3-mini-4k-Instruct's is 4096). Since prior work notes that primacy bias is more pronounced in longer contexts, we selected the 1,000 longest Amazon reviews and 1,000 random MediaSum examples, both capped at 4,000 tokens. We now include descriptive statistics for these datasets.

| | Mean | Median | Std | Min | Max |
|---|---|---|---|---|---|
| **Amazon Reviews** | 2180.3 | 2185.5 | 354.5 | 790 | 3169 |
| **MediaSum News** | 1166.1 | 1054.0 | 621.8 | 89 | 2924 |

**Table 2: Dataset statistics.**

**Primacy Results Statistical Tests:** To address whether the observed primacy effect is real, we performed paired t-tests comparing the beginning-summary similarity (s_b) with the middle-summary similarity (s_m). The results confirm the differences are statistically significant across all models and datasets (p < 0.001), indicating the effect is not due to random chance. We have added a summary of these tests to Section 5.2.

| Model | Dataset | T-Statistic | P-Value | Sig. |
|---|---|---|---|---|
| Llama-3-8B-Instruct | Amazon Reviews | 28.2896 | $8.62 \times 10^{-130}$ | ** |
| Llama-3-8B-Instruct | MediaSum News | 26.8965 | $2.64 \times 10^{-120}$ | ** |
| Llama-3.2-3B-Instruct | Amazon Reviews | 28.3100 | $6.26 \times 10^{-130}$ | ** |
| Llama-3.2-3B-Instruct | MediaSum News | 25.1535 | $1.5 \times 10^{-108}$ | ** |
| Phi-3-mini-4k-Instruct | Amazon Reviews | 18.2547 | $1.81 \times 10^{-64}$ | ** |
| Phi-3-mini-4k-Instruct | MediaSum News | 19.9294 | $1.14 \times 10^{-74}$ | ** |
| Qwen3-4B-Instruct-2507 | Amazon Reviews | 24.9284 | $4.82 \times 10^{-107}$ | ** |
| Qwen3-4B-Instruct-2507 | MediaSum News | 21.6017 | $3.18 \times 10^{-85}$ | ** |
| gemma-3-27b-it | Amazon Reviews | 27.7788 | $2.64 \times 10^{-126}$ | ** |
| gemma-3-27b-it | MediaSum News | 24.1413 | $1.01 \times 10^{-101}$ | ** |

**Table 3: Paired t-test results (Beginning vs. Middle coverage similarity).**

*Note. Sig.: * p < .05, ** p < .01. All rows are significant at p < .01.*

**References:**


1. Zheng, L., et al. (2023). Judging LLM-as-a-judge with MT-bench and Chatbot Arena. *NIPS '23*.
2. Liu, N. F., et al. (2024). Lost in the Middle: How Language Models Use Long Contexts. *TACL*.
3. Peysakhovich, A., & Lerer, A. (2023). Attention sorting combats recency bias in long context language models. *arXiv:2310.01427*.


Add: **Author-Editor Confidential Comment**

## Official Review of Submission932 by Reviewer rs8Y

Official Review by Reviewer rs8Y 27 Aug 2025, 11:41 (modified: 14 Oct 2025, 08:12)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer rs8Y, Commitment Readers

Revisions (/revisions?id=e8HPtd2kOI)

**Paper Summary:**
This paper explored the new area by providing a systematic quantification of how LLMs alter content during summarization and fact-checking tasks. The experimental design is comprehensive, evaluating four different models across three distinct datasets with mathematically rigorous metrics for measuring framing bias, primacy bias, and hallucination rates. The authors demonstrate thoroughness by testing 18 different mitigation strategies across multiple bias categories, providing valuable empirical data on intervention effectiveness.

**Summary Of Strengths:**
This research investigates how LLMs perform summarization and news fact-checking tasks, potentially influencing human decision-making through exposure to altered information. The authors evaluate GPT-3.5-turbo and three Llama/Phi models on MediaSum interviews and Amazon Reviews datasets, finding that models change sentiment framing in 21.86% of summaries, exhibit primacy bias focusing on early content in 5.94% of cases, and hallucinate on 60.33% of post-knowledge-cutoff fact-checking questions. They test 18 mitigation methods ranging from prompt modifications to decoding-time interventions, discovering that targeted approaches can reduce specific biases but often involve trade-offs between different bias types.

**Summary Of Weaknesses:**

1. The study lumps together sentiment changes, content emphasis, and factual errors as if they're all the same type of problem. But these are quite different issues. When a model highlights certain parts of a text or shifts the tone slightly, that might actually represent reasonable editorial judgment rather than harmful bias. After all, summarization naturally involves making choices about what to emphasize and how to frame information. Not every deviation from the original source should be labeled as problematic bias.
2. The paper doesn't compare LLM performance to human summarizers or other established methods for detecting bias. Without these benchmarks, we have no idea whether the "biases" they found are actually worse than what humans would produce, or whether they exceed normal levels of variation in summarization tasks.
3. The researchers never establish how much content alteration actually matters for human decision-making. Finding that LLMs change framing 21% of the time sounds concerning, but what's the threshold where this becomes problematic? Would a 10% rate be acceptable? Without understanding the dose-response relationship, these percentages lack meaningful context.
4. The methodology relies on GPT-4o-mini to detect sentiment bias in other language models, which creates an obvious problem. If all these models share similar training approaches or underlying biases, GPT-4o-mini might systematically miss the same types of issues that affect the models being evaluated. It's like asking someone to proofread their own writing—they're likely to miss their own systematic errors.

**Comments Suggestions And Typos:**
The biggest problem is that the authors never actually tested whether these content changes affect human decisions. They assume that when LLMs alter text, it biases people, but they never proved this with real human experiments.

The evaluation method has issues. They use one LLM (GPT-4o-mini) to judge whether other LLMs are biased, which is like having a biased judge evaluate bias. They need human evaluators instead.

The paper treats all content changes as equally bad, but that's not realistic. Sometimes emphasizing certain parts of text or slightly changing the tone might be reasonable editorial choices, not harmful bias. They need to show what level of change actually matters for human decision-making.

**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 = Potentially Interesting: this paper does not resonate with me, but it might with others in the *ACL community.
**Overall Assessment:** 2 = Resubmit next cycle: I think this paper needs substantial revisions that can be completed by the next ARR cycle.
**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**

## Thank you for the constructive feedback on content alteration, human impact, and LLM-as-judge.

Official Comment

by Authors (Julian Skirzynski (/profile?id=~Julian_Skirzynski1), Abeer Alessa (/profile?id=~Abeer_Alessa2), Jessica Maria Echterhoff (/profile?id=~Jessica_Maria_Echterhoff1), Param Somane (/profile?id=~Param_Somane1), +2 more (/group/edit?id=aclweb.org/ACL/ARR/2025/July/Submission932/Authors))

15 Sept 2025, 23:08 (modified: 14 Oct 2025, 08:12)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer rs8Y, Commitment Readers

Revisions (/revisions?id=jGgCJldAAq)

**Comment:**
We thank reviewer rs8Y for noting our systematic quantification of LLM content alteration during summarization and fact-checking, our comprehensive experimental design, rigorous metrics, and mitigation strategies with valuable data on intervention effectiveness.

**Content alteration:** Our work facilitates discussion on content alteration: how much exists, how to measure it, and when it may be harmful. We carefully selected tasks like interview and Amazon review summarization where “editorial judgement” should not broadly apply. For review summarization, the summary should closely match the ground truth, keeping the original writer's judgment. Hence, the model should capture the writer's sentiment, not its own, and apply no sentiment change. For interview summarization, the model should not focus on certain parts of the prompt. In an interview, all parts are equally valid, preventing statements from being taken “out of context.” Systematic over-reliance on specific prompt parts should be avoided. We agree more work is needed for a detailed view of when content alteration occurs and to determine the threshold at which minor differences become problematic, but our work offers a first insight on tasks where this alteration is undesired. We added clarification in Section 4 of the paper, "Quantifying Content Alteration in LLMs Responses."

**LLM vs Human summary:** Thank you for this suggestion. We agree our study doesn't establish if LLM biases are better or worse than human summarizer biases. Our goal, however, is to quantify LLM-introduced biases to inform deployment decisions. In practice, LLMs are deployed at a massive scale where human summarizers are infeasible, making the relevant question if LLM biases are acceptable for automated systems.

**Content alteration affecting human decision-making:** We thank the reviewer for this valid question. Whether content alterations impact human behavior is important. We address this in the revised paper in a new section 6, "Impact on Human Decision-Making," by presenting an additional experiment. Participants made purchasing decisions based on original product reviews or LLM-generated summaries where the model changed the summarization from neutral or negative to positive framing. This study was designed to evaluate if framing alterations can causally impact decisions. Results showed participants chose products ~84% of the time when reading the model-generated, positive summaries (which incorrectly changed the framing) versus ~52% when reading original neutral/negative reviews (statistically significant - Chi-squared test; $\chi^2 = 78$, $df = 1$, $p < 0.001$). In an evaluation of willingness to pay, we saw a 4.5% increase in the price they would pay for the same product when seeing an incorrectly framed review by the model (also significant according to the linear mixed-effect model, $t = 4.406$, $p < 0.001$). Our work thus addresses two fundamental questions: (1) How often do alterations occur? (21% for framing bias), and (2) Can they meaningfully impact decisions when they do occur? (yes, significantly). We agree that establishing dose-response relationships is important, as it remains unknown how often the significant change in behavior can occur based on our results. That said, our contribution establishes the foundation for such analyses, demonstrating that content alterations are measurable and can have substantial behavioral consequences.

**GPT-4o-mini for framing classification:** LLM-as-a-judge approaches are increasingly common for evaluating tasks like summarization, QA, and dialogue. We apply these approaches following best practices, such as choosing a different and more powerful model as a judge, as otherwise, unexplained side-effects like self-preference bias may occur.

We selected OpenAI's GPT-4o-mini model after an initial validation study (Appendix A) where we selected 500 random Amazon reviews and asked each model to rate the product 1–5 given the review. We then mapped the star rating to 1–2 as negative, 3 as neutral, and 4–5 as positive, and computed accuracy against the user star-ratings. Compared to other models we tested (GPT-4-turbo and GPT-3.5-turbo), this model had the best performance on the Amazon Reviews ground-truth data, with accuracies of 92% (GPT-4o-mini), 91% (GPT-4-turbo), and 89% (GPT-3.5-turbo), suggesting it is a reasonable proxy for this task.

If satisfied, we ask Reviewer 3 to consider raising the score.

**References:**

Mitchell & Carson. 2013. Using surveys to value public goods: the contingent valuation method. RFF Press.

Zheng et al. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. NeurIPS.

Liu et al. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. EMNLP.

Chen et al. 2024. Humans or LLMs as the Judge? A Study on Judgement Bias. EMNLP.

Dietz et al. 2025. Principles and Guidelines for the Use of LLM Judges. ICTIR.

Huang et al. 2021. Efficient Attentions for Long Document Summarization. NAACL.