NewsLensAI: NER-Guided Summarization for Mitigating Hallucination and Bias in LLM-Based News Summaries (Student Abstract)

Gaurank Maheshwari, Ambika Taploo, Ashiqur R. KhudaBukhsh

Rochester Institute of Technology
gm8189@rit.edu, al5797@rit.edu, axkvse@rit.edu

Abstract

Automated news summarization using large language models (LLMs) offers great potential to enhance information accessibility. However, critical challenges, such as hallucinations, bias, and toxicity, threaten their reliability and societal acceptance. In this paper, we present *NewsLensAI*, a novel summarization framework explicitly designed to address these trustworthiness concerns through Named Entity Recognition (NER)-guided prompting. By anchoring summaries in key factual entities extracted from source articles, our method significantly reduces factual inaccuracies without altering model weights or architectures. We evaluated *NewsLensAI* on a dataset of 1,500 real-world news articles using open-source (`LLaMA 3`) and proprietary (`Gemini 1.5`) LLMs. Our analysis encompasses factual consistency, political bias shifts, sentiment preservation, and moderation of toxicity. Our results indicate substantial improvements in factual alignment, demonstrated by an average increase in the BERTScore from 0.80 (baseline) to 0.88 (NER-enhanced), and an approximately 60% reduction in hallucinated entities. To capture contextual terms that are relevant beyond the core entities, we use TF-IDF salience scoring to supplement standard NER categories, particularly for legislative terms and event identifiers. Furthermore, we identify and characterize a notable "centrist drift," wherein summaries tend to moderate extreme biases present in source articles, along with a measurable reduction in toxic or emotionally charged language. Complementing our empirical findings, we introduce a real-time *NewsLensAI* demo that summarizes live news feeds from the Guardian API, providing dynamic bias and sentiment analysis. This practical implementation underscores the real-world applicability and potential societal benefit of our approach. Finally, we discuss critical ethical implications, including potential impacts on media literacy and information diversity. Our interdisciplinary approach, linking NLP, journalism, and ethical analysis, positions *NewsLensAI* as a meaningful step towards safer, fairer, and more trustworthy AI-generated news consumption.

Introduction

Abstractive news summarization with large language models (LLMs) improves accessibility but also poses risks of hallucination, bias, and toxic language. Approximately 30% of neural-generated summaries contain unsupported facts (Maynez et al. 2020; Kryściński et al. 2019), which can undermine trust and distort perceptions. Prior mitigation strategies include subsequent verification and entailment checks, or interventions during training, and these approaches are resource intensive (Chen et al. 2021; Nan et al. 2021; Gao et al. 2022). We propose a lightweight alternative: *NewsLensAI*, a model-agnostic framework that uses NER-guided prompting to improve factuality and preserve perspective in news summarization.

Methodology

Pipeline. We (1) extract entities (people, organizations, locations, dates) from each article using open source machine learning models; (2) construct a summarization prompt with an explicit "Important Entities" list; and (3) generate summaries with LLaMA 3 and Gemini 1.5 under zero-shot settings. This explicit entity anchoring follows prior observations that entity-aware constraints improve faithfulness (Nan et al. 2021; Chen et al. 2021; Akani et al. 2023).

Dataset and Models. The corpus comprises 1,500 real-world news articles that cover political and general interest topics. For controlled procedures, we use an open LLaMA 3 model; for the online demo, we use Gemini 1.5. No fine-tuning is applied; decoding uses a low temperature for determinism.

Prompt. The instruction requests a concise, factual, and unbiased summary; the entity list emphasizes verifiable actors and events. We generate (i) a baseline summary (no entity list) and (ii) an NER-guided summary per article and model, isolating the effect of entity anchoring.

Evaluation and Metrics

We quantify four dimensions, using automated metrics established in prior work:

- **Factual consistency:** BERTScore (F1) (Pagnoni, Balachandran, and Tsvetkov 2021; Goyal and Durrett 2021) computes semantic similarity between reference and generated summaries using contextual embeddings, combining precision and recall into a single F1 score. (Nan et al. 2021).
- **Political bias:** a classifier with three classes (left, center, right) that quantifies bias shifts from article to summary, following prior analyzes in news summarization

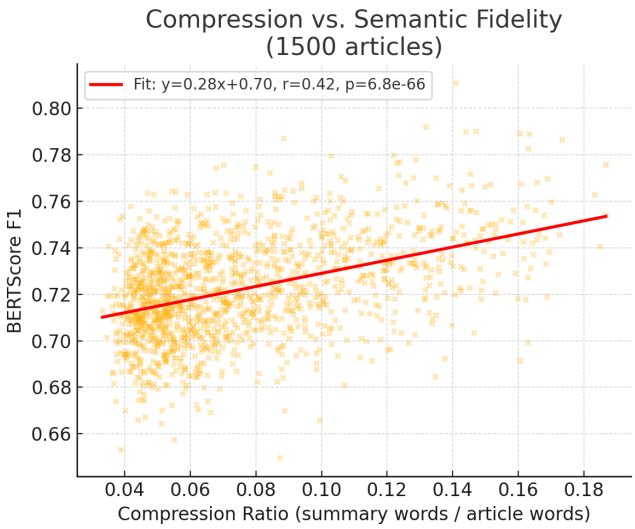

Figure 1: Relationship between compression ratio and semantic fidelity (BERTScore F1) across 1,500 articles. Higher compression correlates with improved factual alignment under NER-guided prompting.

and LLM output. (Steen and Markert 2023; Motoki, Neto, and Rangel 2025; Choudhary 2024).

- **Sentiment preservation:** polarity agreement (article vs. summary), capturing tone drift.
- **Toxicity:** toxicity scores to ensure that generation does not introduce harmful content (Gehman et al. 2020).

Where appropriate, we report paired comparisons between baseline and NER-guided summaries (the same article, the same model).

Results

Table 1 summarizes the performances. NER-guided prompting improves the BERTScore (0.80 $\rightarrow$ 0.88), increases the overlap of the entities (93.1% $\rightarrow$ 98.7%), reduces hallucinated entities by $\sim$ 60%, and reduces toxicity (Gehman et al. 2020). The bias shifts also decrease (41. 3% $\rightarrow$ 26. 7%), indicating a greater alignment with the original article framing; Gemini 1.5+NER shows comparable preservation (24.5% shift). These findings are consistent with previous observations that entity-based constraints improve faithfulness (Nan et al. 2021; Akani et al. 2023) and complement revision-based approaches such as RARR (Gao et al. 2022). We attribute the residual performance gap between Gemini 1.5 and LLaMA 3 outputs to their differing alignment objectives and safety-tuning strategies.

As shown in Figure 1, higher compression ratios correlate positively with BERTScore F1, illustrating that NER-guided prompting maintains strong factual alignment even in shorter summaries—supporting the observed reduction $\sim$60% in hallucinated entities.

Metric	LLaMA 3	LLaMA 3 + NER	Gemini 1.5 + NER
BERTScore F1	0.802	0.881	0.865
Entity Overlap (%)	93.1	98.7	97.5
Bias Shift (% articles)	41.3	26.7	24.5
Sentiment Match (%)	60.2	66.1	69.3
Toxicity Score	0.051	0.021	0.028

Table 1: Average summarization metrics for LLaMA 3 (baseline vs. NER-guided) and Gemini 1.5 (NER-guided) on 1,500 news articles. Higher is better for all metrics except Toxicity.

Related Work

Hallucination in abstractive summarization is well documented (Kryściński et al. 2019; Maynez et al. 2020; Pagnoni, Balachandran, and Tsvetkov 2021; Goyal and Durrett 2021). Existing mitigation strategies range from entity focused modeling (Nan et al. 2021; Chen et al. 2021; Akani et al. 2023) to revision methods applied after generation (Gao et al. 2022). Bias in summarization and LLM output have also been reported in multiple datasets and settings (Steen and Markert 2023; Motoki, Neto, and Rangel 2025; Choudhary 2024). Our findings indicate that NER guided prompting offers a complementary approach, substantially reducing unsupported content while more faithfully preserving the perspective of the source text.

Discussion and Bridge Theme

NewsLensAI connects NLP with journalism and ethics, aligning with AAAI-26's collaborative bridge theme. By anchoring summaries in verifiable entities, it supports transparency and more trustworthy media consumption. Our real-time demo illustrates applicability beyond the lab, enabling readers to inspect bias and sentiment alongside generated summaries.

The differences between the LLaMA 3 and Gemini 1.5 outputs suggest that alignment and safety-tuning strategies play a critical role in the factual and stylistic results, emphasizing the need for transparent evaluation across model families. From an ethical standpoint, *NewsLensAI* highlights how interpretable prompting and verifiable entity anchoring can mitigate misinformation risks while promoting accountability in AI-generated journalism.

Limitations and Future Work

Future work includes extending to multilingual settings, incorporating citation grounding for factual claims, and conducting human evaluations to assess user trust and media literacy outcomes. In addition, as hallucination is substantially reduced by NER-guided prompting, a priority is to deepen *bias and political polarization analysis*, for example, through fine-grained framing and stance measures, as well as outlet- or issue-specific polarization dynamics, to characterize and mitigate residual perspective shifts that may remain even when factual errors are minimized. Manual human evaluation for bias perception and trust calibration is also planned to complement automated metrics.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Award No. DGE-2125362. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. KhudaBukhsh was partly supported by a gift from Lenovo.

References

Akani, E.; Favre, B.; Bechet, F.; and Gemignani, R. 2023. Reducing named entity hallucination risk to ensure faithful summary generation. In Keet, C. M.; Lee, H.-Y.; and Zarrieß, S., eds., *Proceedings of the 16th International Natural Language Generation Conference*, 437–442. Prague, Czechia: Association for Computational Linguistics.

Chen, S.; Zhang, F.; Sone, K.; and Roth, D. 2021. Improving Faithfulness in Abstractive Summarization with Contrast Candidate Generation and Selection. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tur, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5935–5941. Online: Association for Computational Linguistics.

Choudhary, T. 2024. Political Bias in Large Language Models: A Comparative Analysis of ChatGPT-4, Perplexity, Google Gemini, and Claude. *IEEE Access*.

Gao, L.; Dai, Z.; Pasupat, P.; Chen, A.; Chaganty, A. T.; Fan, Y.; Zhao, V. Y.; Lao, N.; Lee, H.; Juan, D.-C.; et al. 2022. Rarr: Researching and revising what language models say, using language models. *arXiv preprint arXiv:2210.08726*.

Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Goyal, T.; and Durrett, G. 2021. Annotating and modeling fine-grained factuality in summarization. *arXiv preprint arXiv:2104.04302*.

Kryściński, W.; McCann, B.; Xiong, C.; and Socher, R. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Maynez, J.; Narayan, S.; Bohnet, B.; and McDonald, R. 2020. On Faithfulness and Factuality in Abstractive Summarization. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1906–1919. Online: Association for Computational Linguistics.

Motoki, F. Y.; Neto, V. P.; and Rangel, V. 2025. Assessing political bias and value misalignment in generative artificial intelligence. *Journal of Economic Behavior & Organization*, 106904.

Nan, F.; Nallapati, R.; Wang, Z.; Nogueira dos Santos, C.; Zhu, H.; Zhang, D.; McKeown, K.; and Xiang, B. 2021. Entity-level Factual Consistency of Abstractive Text Summarization. In Merlo, P.; Tiedemann, J.; and Tsarfaty, R., eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2727–2733. Online: Association for Computational Linguistics.

Pagnoni, A.; Balachandran, V.; and Tsvetkov, Y. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. *arXiv preprint arXiv:2104.13346*.

Steen, J.; and Markert, K. 2023. Bias in news summarization: Measures, pitfalls and corpora. *arXiv preprint arXiv:2309.08047*.