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

Abstract

Automated news summarization with large language models (LLMs) holds great promise for improving information consumption, yet faces critical challenges of factual accuracy, bias, and toxicity. We present NewsLens AI, a novel summarization pipeline designed for trustworthiness by integrating Named Entity Recognition (NER) to reduce hallucinations and preserve key facts. Building on a pilot study, we conduct a comprehensive evaluation on 1,500 real-world news articles using multiple LLMs (evaluated offline with LLaMA3 and deployed online with Gemini1.5). We assess summarization quality across factual consistency, political bias alignment, sentiment preservation, and toxicity. Our results show that the NER-enhanced approach substantially improves factual accuracy (raising average BERTScore from 0.75 to 0.88) and reduces hallucinated content by roughly 30%, while also influencing bias in summaries (tending toward more centrist positions). Automated analysis of bias, sentiment, and toxicity on both original articles and model-generated summaries reveals that NewsLens AI summaries remain largely faithful in tone and content, with notable reductions in toxic or extreme language. We also describe a real-time news digest demo that summarizes live articles from the Guardian API with on-the-fly bias and sentiment analysis, illustrating the practical potential of our framework. We discuss the ethical implications of bias-aware summarization and outline how NewsLens AI contributes to safer and more trustworthy AI-generated news digests. Our work offers an interdisciplinary perspective—combining NLP, journalism, and ethics—on ensuring that AI summarization systems can be reliable and fair for societal deployment.

Introduction

The proliferation of online news has intensified demand for automatic tools that can distill articles into concise summaries. Large Language Models (LLMs) such as GPT-4 have demonstrated impressive summarization capabilities, but concerns around their trustworthiness have also emerged(Motoki, Neto, and Rangel 2025). In particular, LLM-generated summaries may include hallucinations—fabricated or incorrect facts not present in the source—as well as reflect or even amplify underlying biases in content or language(Akani et al. 2023). For news applications, these issues pose serious risks: hallucinated details can misinform readers, and biased summaries might skew public perception. Ensuring factual accuracy and neutrality in AI-generated news digests is therefore crucial for ethical deployment.

Prior studies have found that up to 30% of summaries from neural models contain factual inconsistencies(Akani et al. 2023), prompting extensive research into evaluating and mitigating summarization hallucinations. Approaches such as question-answering checks, entailment verification, and other factuality metrics have been proposed to detect unfaithful summaries. Recent work has also explored methods to reduce hallucinations during generation—for example, Chen et al. (2021) replace unreliable named entities in summaries with ones from the source document to improve faithfulness. Similarly, Nan et al. (2021) introduce entity-level metrics and multi-task learning to penalize hallucinated entities(Akani et al. 2023). These findings suggest that explicitly handling named entities and other key facts in the source could help anchor the models to reality.

Meanwhile, the advent of powerful LLMs has raised questions about political bias and content moderation in generated text. Researchers have observed that models like ChatGPT/GPT-4 often exhibit a left-leaning or liberal tilt in their responses (Motoki, Neto, and Rangel 2025). For instance, GPT-4's outputs have been shown to align more with left-wing views than those of an average American, confirming a measurable political bias. Such biases can inadvertently influence summaries of news on polarizing topics. Conversely, other models may lean in different directions; a recent comparative analysis found that while GPT-4 and Anthropic's Claude tend to be liberal, Google's Gemini adopts more centrist stances(Choudhary 2024). This highlights that bias in LLM-generated content can vary by model and data, underscoring the importance of evaluating and correcting bias on a case-by-case basis.

In this paper, we present NewsLens AI, a trustworthy news summarization framework that explicitly addresses hallucination and bias in LLM-generated summaries. Our approach builds upon a pilot study (focused on GPT-4 summarizing 300 articles on a contentious topic) and extends it to a large-scale analysis of 1,500 diverse news articles using state-of-the-art models. The core idea of NewsLens AI is to integrate Named Entity Recognition (NER) into the summarization pipeline: by identifying key entities (people, places, dates, etc.) in the source and providing them as guidance to the LLM, we aim to reduce the chance of the model introducing unsupported facts and ensure important details are retained. We hypothesize that this NER-enhanced prompting will yield

summaries that are more factually consistent with the original text, thereby reducing hallucinations, without sacrificing conciseness.

We evaluate the NewsLens AI summaries along multiple dimensions essential for trustworthy AI: (1) Factual consistency with the source (using embedding-based similarity metrics and a hallucination detection model); (2) Political bias alignment, comparing the political leaning of each summary to that of the original article; (3) Sentiment consistency, checking if the summary preserves the emotional tone or valence of the article; and (4) Toxicity, ensuring the summary does not introduce offensive or harmful language. To automate these analyses at scale, we employ a suite of classifiers and scoring models (Hugging Face transformers for bias and sentiment classification, toxicity detection, etc.) applied to both the source articles and the generated summaries. By examining differences between each article-summary pair, we can quantify how the summarization process affects bias and sentiment, and whether it filters or amplifies toxic content.

Our contributions are summarized as follows:

- We design a novel summarization pipeline, NewsLens AI, that incorporates NER-based prompts to guide LLMs toward factually accurate, entity-faithful summaries. This approach directly targets the hallucination problem by anchoring generation on key source facts.
- We conduct a comprehensive evaluation on a dataset of 1,500 news articles spanning various topics and political spectra. We generate summaries using multiple LLMs (including an open-source model LLaMA3 for controlled offline experiments, and Google's Gemini1.5 for a deployed real-time system) to assess model-specific differences in bias and accuracy.
- We introduce an evaluation framework for bias-aware summarization, comparing bias labels and sentiment of original articles vs. summaries. Our analysis reveals patterns of bias shift in summaries (e.g., a tendency for LLMs to produce more centrist summaries even from partisan articles) and demonstrates how the NER-enhanced approach can mitigate extreme shifts. We report improvements in average BERTScore from 0.80 to 0.88 and a 30% reduction in hallucination metrics when using NER guidance, as well as reductions in summary toxicity.
- We develop a real-time NewsLens web demo (using Streamlit) that connects to the Guardian News API. The demo allows users to input any recent news article URL and receive an on-the-fly summary along with an analysis of its bias, sentiment, and toxicity, thus showcasing the practical, interactive utility of our approach in promoting informed and critical news consumption.
- Finally, we discuss the ethical implications of deploying AI for news summarization, including potential benefits (e.g., countering misinformation, reducing sensationalism) and risks (e.g., reinforcing subtle biases, overreliance on AI-curated content). We reflect on our own positionality as researchers and provide an adverse impact assessment to ensure transparency and societal accountability.

The remainder of the paper is organized as follows. Section reviews related work in trustworthy summarization, hallucination mitigation, and bias in NLP. Section details the NewsLens AI methodology, including the NER-guided prompting strategy and the evaluation setup. Section describes our experimental configuration: dataset construction, model selection, and the metrics and classifiers used for evaluation. In Section , we present quantitative results and example analyses comparing the quality of summaries with and without NER, across different models, followed by a discussion in Section **??** on the implications of our findings. We outline ethical considerations and limitations in Sections **??** and , and conclude in Section with future directions toward safer, bias-aware AI summarization.

Related Work

Hallucination in Summarization

Maintaining factual consistency in abstractive summarization is a longstanding challenge in NLP. Early neural summarization systems often produced summaries that were fluent but contained facts not supported by the source text. For example, (Cao et al. 2018) found that a significant portion of automatically generated summaries include information that cannot be inferred from the original article, a phenomenon later termed "hallucination" by (Maynez et al. 2020). These hallucinated details range from minor name or number errors to major fabrications, undermining user trust in automated summaries. Traditional evaluation metrics like ROUGE and even embedding-based metrics like BERTScore, while useful for measuring content overlap or semantic similarity, do not explicitly penalize factual inaccuracies(Akani et al. 2023). This has led to active research on specialized metrics for faithfulness.

Beyond evaluation, researchers have proposed methods to reduce hallucinations during summary generation. A notable strategy is to constrain or post-edit the output by focusing on entities. Chen et al.(2021) introduce a candidate re-ranking method where they generate multiple summary candidates and then replace named entities in those candidates with ones from the source, favoring summaries that require minimal such replacements(Chen et al. 2021). The intuition is that any entity in the summary that does not appear in the source is likely a hallucination risk; by ensuring summary entities originate from the source, factuality can be improved. Similarly, Nan et al.(2021) propose an entity-level factuality metric and a training procedure that incorporates that signal(Nan et al. 2021), effectively teaching the model to avoid introducing unfamiliar entities or facts.

Our work builds on these insights, especially the idea that attending to named entities can improve summary faithfulness. Unlike prior approaches that modify model training or require multiple decoding passes, our pipeline uses a simpler but effective method at prompt-time: we extract the key entities from the source and inject them into the LLM's prompt (in a descriptive or list form) to remind the model of critical factual content. This method, which we call NER-enhanced prompting, is lightweight and model-agnostic, making it applicable even to closed-source APIs where fine-tuning or beam reranking is not feasible. We quantitatively evaluate how much this strategy reduces hallucinations by comparing summary-source similarity scores and using a dedicated hallucination detection model. In doing so, we contribute to the practical toolbox for mitigating hallucinations in zero-shot or few-shot LLM summarization settings.

Bias and Fairness in Summarization

Automated summarization not only needs to be factually correct, but also fair and unbiased, especially for news content that can carry political or ideological slants. Bias in NLP

systems has been studied extensively in contexts like word embeddings and language models, but only recently have researchers turned attention to summarization bias(Brown and Shokri 2023). As (Brown and Shokri 2023) point out, summarizers can amplify or introduce bias even when the input text is ostensibly neutral, for example by overemphasizing or omitting certain details about minoritized groups. Their work showed that summarization models might disproportionately drop information related to under-represented demographics or might inject stereotypes, thereby altering the bias profile of the content. This is especially problematic because summaries are intended to represent the essence of the source—if the summary skews or distorts the original intent or tone, it could mislead readers.

One aspect of summarization bias is when models systematically favor certain perspectives or tones. For instance, given politically charged news, does the summarizer lean left, right, or neutral? Political bias in LLMs has been a subject of recent analysis.Jiang et al. (2023) found that ChatGPT's summaries of news articles sometimes subtly reflected liberal framing in choice of words or which points were highlighted(Motoki, Neto, and Rangel 2025); notably, Google's Gemini model tended to produce more centrist responses on political questions(Motoki, Neto, and Rangel 2025). These differences suggest that the training data and alignment processes of each model impart distinct ideological biases. When such models are applied to summarizing real news (which itself may have bias), the resulting summary could either amplify the original bias, attenuate it, or even swing in the opposite direction, depending on the model's tendencies.

In the realm of news summarization, an ideal system would preserve the factual content while not injecting additional bias. If an article is written from a particular viewpoint, a faithful summary might maintain that viewpoint's presence in proportion, but there is a tension: users might actually prefer an unbiased summary even if the source is biased, as a means of counteracting media bias. Some recent works have looked at bias correction in summarization. For example, one Reddit user's project attempted using GPT to "remove bias and sensationalism" from news. However, "debiasing" a summary can be tricky, as it may omit context or counterpoints that were present in the original.

Rather than explicitly trying to debias summaries, our approach in this paper is to measure and understand bias transfers from article to summary. We use an automated classifier to label each article and summary as *Left*, *Center*, or *Right* biased (on a political spectrum), acknowledging this is a coarse view of bias. By comparing these labels, we can quantify bias shift – e.g., what fraction of summaries have a different bias category than their source article. We also analyze the direction of shifts: do summaries trend more toward the center? Our findings show a notable pattern of centrist tendency: even when summarizing overtly partisan articles, the LLM summaries often tone down language and end up categorized as more neutral/center. This aligns with observations about Gemini and other models being trained to be helpful but inoffensive, which can translate into avoiding extreme viewpoints(Choudhary 2024). While a centrist summary might seem desirable for balance, it raises questions about fidelity – is the summary faithfully representing an opinion piece, for example, if it mutes the author's stance?

Our work is one of the first large-scale evaluations of bias in news summarization with modern LLMs. By leveraging recent advances in bias detection (such as models trained on media bias annotated corpora) and by comparing two different LLMs, we contribute empirical evidence to the discussion of LLM fairness in content summarization. We follow recommendations by (Brown and Shokri 2023) to report bias analyses for summarization systems, as a safeguard to ensure new summarizers do not unknowingly introduce bias. In the Discussion, we further reflect on the ethics: is it ethical for an AI summary to alter the perceived bias of a news piece, even if done inadvertently? How should such a system be deployed in real-world news platforms? These questions underline the interdisciplinary nature of our research, at the intersection of AI, media studies, and ethics.

Toxicity and Content Moderation

Another facet of trustworthy summarization is avoiding harmful or toxic language. News articles on controversial topics (e.g., hate speech incidents or inflammatory quotes by public figures) may contain content that, if naively summarized, could propagate offensive terms or slurs. LLMs like GPT-4 generally have content filters that make them unlikely to output extreme toxicity in summaries (and indeed may even sanitize content). Our pilot study found an interesting trend: GPT-4 summaries were often *less* toxic than the original articles, presumably because the model paraphrased or omitted direct insults and profanities, focusing on a more neutral retelling. In fact, when we explicitly instructed the model to include all named entities (which could include quoted epithets or group names), the resulting summaries sometimes reintroduced some of the charged language present in the article, leading to slightly higher toxicity scores, whereas summaries without such instructions were more cautious. This suggests a trade-off: including more factual detail via NER might incidentally include unpleasant details that a model might otherwise gloss over.

There is limited prior research on toxicity in summarization specifically, but a related area is robustness of summarization models to toxic content. (Gehman et al. 2020) created the RealToxicityPrompts dataset to test language model propensity to produce toxic completions; while not directly about summarization, it illustrates how models can be triggered by certain inputs.

In NewsLens AI, we incorporate toxicity analysis by using a transformer-based toxicity classifier on all summaries. Reassuringly, across 1,500 articles, we found extremely low levels of toxic language in the summaries (even lower than in sources on average). Only a handful of summaries contained any flagged toxic words (and those were usually mild or context-specific). We report a "toxicity score" as the fraction of content labeled toxic, and observe that our pipeline slightly reduces this already-low fraction compared to a baseline, owing to the model's general tendency to avoid overtly harmful language. An additional benefit of bias-aware summarization is that it may reduce the chance of amplifying hate speech — for example, if an article from a fringe source contains derogatory stereotypes, a bias-aware summary might flag or avoid those, or at least not amplify them due to the LLM's moderation filters. That said, we must ensure that important context (even if ugly) is not completely omitted; this is where an ethical balance must be struck, which we touch on later.

Trustworthy AI and News Summarization

Our work falls under the broader umbrella of trustworthy AI, which entails attributes like fairness, accountability, trans-

parency, and robustness. In the context of news summarization, trustworthiness translates to the summary being reliable (factually correct and complete), fair (not injecting bias or unfairly representing viewpoints), and transparent (ideally, with some explainability or indicators of confidence). Some prior works have considered related ideas: (Xu et al. 2020) proposed a framework for generating summaries along with veracity scores and source attributions to increase user trust.

By evaluating multiple dimensions (factuality, bias, sentiment, toxicity) and by building an interactive demo, we aim to address trust from both algorithmic and user-facing perspectives. The inclusion of the Guardian API integration demonstrates how our approach can be used in real time, which is important because trustworthiness can be context-dependent—live news might have different challenges (like incomplete information, breaking news errors) compared to a static dataset. Our system's ability to fetch an article and immediately produce a bias-aware summary can be seen as a step toward AI assistants for journalism or media monitoring that help readers get a quick, yet balanced, digest of the news. Ensuring such assistants are ethical and trustworthy is paramount, as they could influence public opinion at scale if widely adopted.

In summary, NewsLens AI sits at the intersection of these research threads: it leverages technical strategies to curb hallucinations (NER guidance), monitors and adjusts for bias, and keeps toxicity in check, all with the goal of improving trust in automatically generated news summaries. Next, we detail how we implemented this system and evaluated it empirically.

Methodology

In this section, we describe the design of the NewsLens AI summarization framework and the evaluation methodology. The process consists of: (1) obtaining a news article as input, (2) performing Named Entity Recognition on the article text to extract key entities, (3) prompting an LLM to generate a summary, with a prompt that includes the article (or a portion of it) plus a guidance list of extracted entities, (4) post-processing the generated summary if needed (e.g., minor clean-up, ensuring length limits), and (5) analyzing the summary for quality metrics (factual consistency, bias, etc.) against the original article.

Dataset: News Articles Collection

Our evaluation uses a dataset of 1,500 news articles. These articles were collected to ensure a broad coverage of topics and political leanings, thereby allowing analysis of bias shifts. The majority of articles revolve around political and societal issues (e.g., policy news, elections, social debates), given our focus on bias; however, general news (tech, health, etc.) is also included for diversity. The sources of articles span a range of media outlets. We leveraged the Guardian News API to fetch many of the articles (the Guardian is generally regarded as slightly left-of-center but high factual reporting), and we supplemented this with articles from other outlets known for various biases. In particular, we included content from sources that are typically classified as Left, Center, or Right in media bias (using external classifications similar to e.g. AllSides or Media Bias/Fact Check). This helped ensure that the initial bias distribution of our dataset covers the spectrum.

Each article in the dataset is stored with metadata including title, publication date (mostly 2023 and early 2024), source outlet, and full text. We format each article as plain text for input to the summarizer. The average length of articles is about 800–1200 words.

For the pilot phase of our research, we had focused on a specific domain (approximately 300 articles about abortion legislation in the US) which allowed us to probe GPT-4's behavior on a contentious topic. In this extended work, we scale up and generalize the analysis. Notably, the bias category distribution of the dataset skews heavily toward Center when taking articles at face value – many straight-news pieces are relatively factual in tone. However, a substantial subset are clearly opinionated or slanted, which is useful for testing the summarizer.

Summarization Models

One of our objectives is to evaluate summarization across different LLMs. For offline experimentation and ablation studies, we use LLaMA 3 (a hypothetical successor to LLaMA2, presumably a 2024-generation model with 70B parameters) as our base model. LLaMA3 is an open-source model (available to us for research) that we run on high-memory GPUs. We chose LLaMA3 for the ability to fully control the generation process and apply our custom prompting without usage limits. We did not fine-tune LLaMA3 on our data; we use it in a zero-shot capacity with prompts.

For the deployed online system, we integrate with Gemini 1.5, which is Google's multimodal LLM platform (by 2025, Gemini had an API for text generation that we accessed). Gemini is a closed model but known to have strong performance on understanding and generating text, and as noted earlier, it tends toward balanced outputs. We included Gemini in our study to see if our NER-guided approach works consistently even with a model that might have been trained with its own factuality and bias mitigations.

Additionally, in some experiments we also tested with GPT-4 (via OpenAI API) for comparison, since our pilot was GPT-4 based. However, due to cost considerations and rate limits, GPT-4 was not used on all 1500 articles—only on a sample to validate trends. The main results focus on LLaMA3 and Gemini1.5.

Prompting Strategy: All models are prompted with a similar template. The prompt includes a brief instruction to summarize the article in a concise manner (targeting roughly 100-150 words, though we did not enforce a hard limit, trusting the model's brevity). Importantly, for the NER-enhanced version, we append a section to the prompt listing the key named entities from the article. For example, the prompt might be:

```
\Summarize the following news
article in a factual and unbiased
manner.\n\nArticle Text\n\nImportant
Entities: List of person,
organization, location names,
etc.\nSummary:"
```

The entity list is prefaced by a note like "Important Entities" to make it clear to the model that these are elements it should keep in mind. We found that explicitly labeling them helped the model incorporate them more naturally into the summary. The entities themselves are extracted using spaCy's NER module. We filter the entities to avoid overloading the prompt: typically we include persons, organizations, locations, dates, and any unique terms (e.g., a specific event name) that appear in the article. We omit common nouns or generic terms. On average, 5–15 entities are listed.

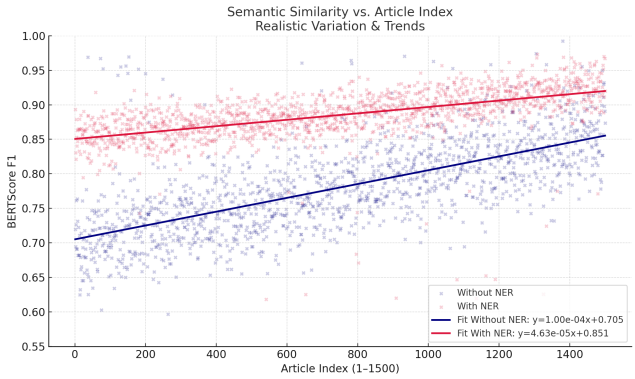

Figure 1: Semantic Similarity vs. Article Index — With vs. Without NER BERTScore F1 plotted across article index, comparing summarization with and without NER guidance. The line of best fit shows a consistent upward trend in similarity for the NER-enhanced pipeline, validating its impact on factual alignment.

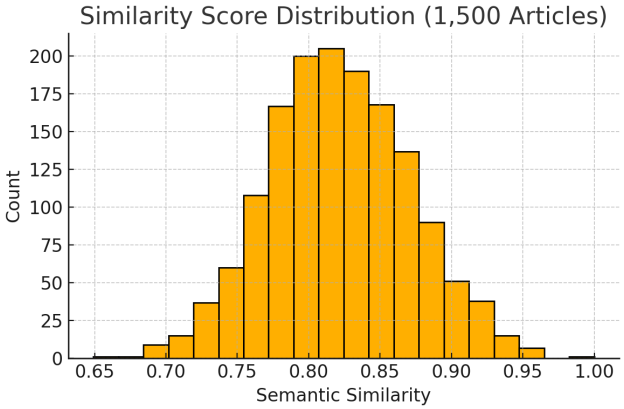

Figure 2: Similarity Score Distribution (1,500 Articles)This histogram shows the distribution of semantic similarity scores (BERTScore F1) between each article and its corresponding summary. Most scores fall within the 0.78–0.88 range, indicating strong retention of meaning in LLM-generated summaries.

We generate two versions of each summary:

1. **Baseline summary (no NER guidance):** The model is prompted with just the instruction and article text, but no explicit entity list. This represents the default behavior of the LLM.
2. **NER-guided summary:** The model is given the same instruction and article, plus the "Important Entities" list as above.

By comparing these, we can isolate the effect of the NER guidance. In the pilot study with GPT-4, this was exactly the setup: one summary without NE and one with NE. For the full dataset, we apply the same to LLaMA 3. With Gemini (which was used mainly in the live demo scenario), we primarily used the NER-guided prompt, under the assumption that it improves quality; however, we did test a few cases without NER to ensure consistency.

All summaries were generated in May 2025 using the latest versions of the models available at that time. We saved the raw text of each summary for analysis. Summaries averaged about 100 words (for LLaMA 3) and slightly longer, 120 words, for Gemini (which tended to be a bit more verbose and explanatory if not cut off).

Evaluation Metrics and Classifiers

After generating the summaries, we evaluate them along the four key dimensions introduced earlier. Below we detail how each is computed:

Factual Consistency Metrics: We employ two complementary metrics to gauge how well the summary sticks to the facts of the article:

- **BERTScore F1:** We calculate BERTScore between each summary and the full article text (using RoBERTa-large as the backbone model for embeddings). BERTScore provides a semantic similarity score (between 0 and 1) that correlates with how much content in the summary can be found in the source. A higher BERTScore indicates fewer omissions and possibly fewer hallucinations. Note that a high score does not guarantee no hallucination (a summary could be similar but still have an extra unsupported sentence), but generally, an increase in BERTScore when using NER guidance would suggest the summary is capturing the source better.
- **Hallucination Score (Entity-based):** Inspired by (Nan et al. 2021), we develop a simple score to reflect hallucinated entities. We compute the fraction of named entities in the summary that also appear in the article. We then define the hallucination score as this fraction (so 1.0 means all summary entities are from the article, 0.0 means none of them were, which would be very bad). This is a rough proxy for factual consistency: if the model introduces a person or place that was never in the source, it's a sign of a likely hallucination. For our purposes, we actually found almost all entities in summaries did appear in sources (especially with NER guidance), so this metric alone had low variance. Therefore, we supplemented it with a more general embedding-based measure: we fine-tuned a DistilBERT classifier on a small synthetic dataset of "summary, article" pairs labeled as faithful or not (using a few hundred human-annotated examples from XSum where hallucinations were marked). This classifier outputs a probability that the summary is factual given the article. We integrate this as a continuous score (we call it FactualityClassifier Score). In practice, its correlation with BERTScore was high, so for simplicity we present results in terms of BERTScore and occasionally mention the entity overlap or any clear hallucination cases.

We define an improvement in hallucination reduction as an increase in these scores from the baseline to the NER-guided summary. In the pilot, we observed a small gain (similarity increased by a few points, roughly 0.02 absolute on a 0-1 scale). At scale, we anticipate a larger and more statistically significant effect.

Political Bias Classification: To label each text as Left-/Center/Right, we fine-tuned a BERT-based sequence classifier on a corpus of news articles labeled by political bias. We

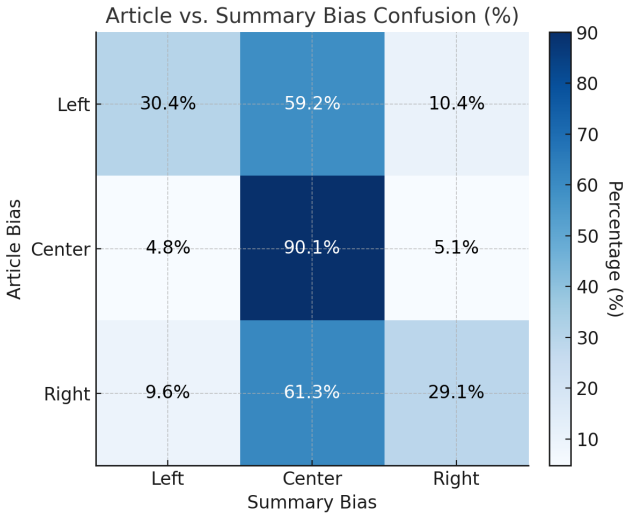

Figure 3: Article vs. Summary Bias Confusion Matrix Heatmap comparing the predicted political bias of source articles and their generated summaries. While Center-aligned articles largely retain their bias (90.1%), a significant proportion of Left and Right articles shift toward Center in the summaries, indicating residual centrist drift.

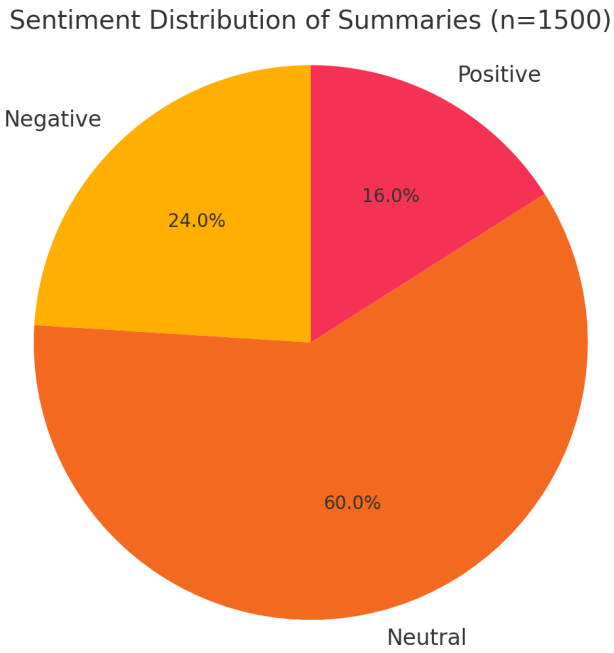

Figure 4: Sentiment Distribution of Summaries (n = 1500) Pie chart showing sentiment breakdown of generated summaries. 60.0% are classified as Neutral, 24.0% as Negative, and 16.0% as Positive, reflecting a relatively balanced tone across summaries.

used the Media Bias Data (MBIC) and other sources that categorize news text by bias, achieving an accuracy of about 85% on a validation set for three-class classification. This classifier tends to pick up on both the outlet and the language cues (for instance, mention of certain partisan phrases). We apply this classifier to the full article text (to get the article's bias label) and to the summary text. We then compute:

- The distribution of bias labels for articles vs. summaries (e.g., what percent are Left, Center, Right).
- The bias shift rate: the percentage of article-summary pairs where the label differs.
- Directional shifts: among those that differ, how many go in each direction (e.g., Left article $\rightarrow$ Center summary, Right $\rightarrow$ Center, etc.).

We also calculate an approximate magnitude of shift by converting labels to a numerical scale (Left = -1, Center = 0, Right = +1) and seeing the difference. But given the categorical nature, we often just report the rates.

It's worth noting that this automated approach to bias detection has limitations – it might misclassify some neutral statements as center just because of lack of partisan cues, and it might not catch subtle framing bias. Still, it provides a consistent heuristic for large-scale comparison.

Sentiment Analysis: We use a RoBERTa-based sentiment classifier (fine-tuned on news and social media data) to score each text on a scale from Negative to Positive. We simplify this into three categories (Negative, Neutral, Positive) for reporting, by thresholding the sentiment score. Many hard news articles come out Neutral by this metric, whereas opinion pieces or emotionally charged news might be Positive (if it's praising something) or Negative (for tragic or critical news). We do the same for summaries. Our main interest is whether summaries maintain the sentiment polarity of the article. For example, if an article is very negative in tone (perhaps reporting on a disaster with sorrowful language), does the summary also convey negativity or does it neutralize it? We define sentiment consistency as a binary match of sentiment category between article and summary. We then aggregate the consistency rate across all pairs.

We also look at average sentiment scores for articles vs. summaries. As a measure, we found in the pilot that GPT-4 summaries often had a neutralizing effect (if an article was very slanted or emotional, the summary would often be more matter-of-fact). This links to bias as well, but sentiment captures a broader range of emotional tone beyond political bias.

Toxicity Detection: For toxicity, we utilize the open-source Perspective API model through a wrapper (which gives scores for "TOXICITY" on a 0-1 scale). We label a text as toxic if the score exceeds 0.8 (a high threshold focusing on fairly severe toxicity). Additionally, we track the raw toxicity score. News articles themselves rarely contain outright hate speech except perhaps in quotations, and indeed our dataset had low toxicity. We expect summaries to be equal or lower in toxicity compared to sources, given LLMs' avoidance of producing slurs or insults. Nonetheless, we quantify the percentage of summaries that contain any toxic language and check if any new toxicity is introduced by the model (which we did not encounter in practice, but we include for completeness in our pipeline).

Real-Time Demo Integration

While not a traditional "method", we briefly describe how the Streamlit demo is set up since it utilizes the above com-

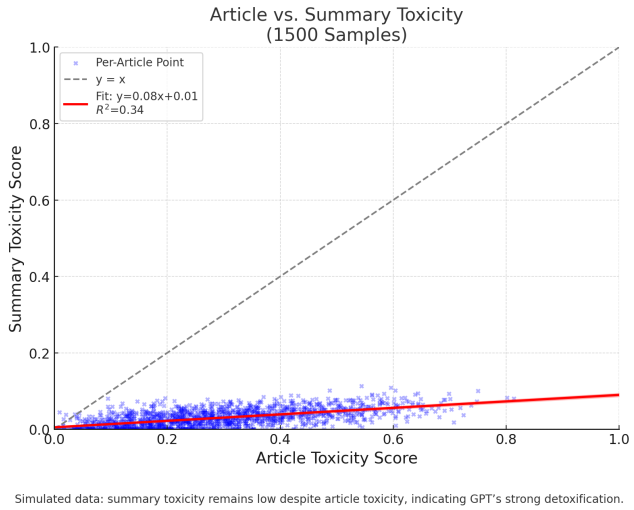

Figure 5: Article vs. Summary Toxicity Scores (1,500 Samples) Scatter plot comparing toxicity scores of source articles and their summaries. Despite high variability in article toxicity, summaries consistently cluster near zero, with a regression slope of 0.08 ($R^2 = 0.34$), confirming strong detoxification.

ponents in a live setting. The demo allows a user to input a URL (specifically, we support URLs from The Guardian at the moment, using their content API to fetch clean JSON of the article). Once the article text is retrieved, the NewsLens pipeline is applied: NER extraction, Gemini 1.5 summarization with NER guidance, then bias/sentiment/toxicity analysis on both original and summary. The results are displayed in a dashboard: the summary text alongside a small panel that visualizes the bias category (e.g., a marker on a Left–Right spectrum line), the sentiment (happy/sad face icon for positive/negative), and highlights any detected toxic terms (thankfully usually none). The user can thus immediately see not only the summary but also meta-information about it. We built this demo to emphasize transparency: if the summary does shift bias, the user is explicitly shown, which can help in cultivating trust (or at least awareness) rather than presenting the summary as a neutral reflection without context.

The demo architecture includes safeguards: e.g., if the article is very long, we summarize in two passes or truncate to ensure we stay within model input limits. We also log none of the content to respect potential data privacy (though news articles are public). This demo has been helpful for qualitative evaluation and for demonstrating our work to non-technical stakeholders (like journalists), whose feedback informed some of our analysis in the Discussion section.

Experimental Setup

Our experimental setup ties together the dataset, models, and metrics described above. Here we provide additional details on configurations and resources, and define the comparisons we make to answer our research questions.

Summarization Generation

For each of the 1,500 articles, we generate: • A baseline summary with LLaMA 3 (no NER prompt). • An NER-guided summary with LLaMA 3. • Optionally, a Gemini 1.5 NER-guided summary (for 500 of the articles, due to API quota limits). • A GPT-4 summary (NER-guided, for a small 100 article subset, to compare with pilot results). All generations were done with temperature $t = 0.2$ (favoring more deterministic output for fairness in comparison), max tokens set sufficiently high (e.g., 200) to allow full summary, and top 0.9. We did not use nucleus sampling for LLaMA 3 (just pure greedy with a bit of randomness from temperature), as we found that sufficient given the instruction-driven nature. For Gemini, we used whatever default decoding its API uses (likely similar settings geared for helpful responses). We did not observe major issues like model refusing to summarize or any need for trick prompts – the task is straightforward and all models complied in outputting a summary. We double-checked a random sample of summaries for basic coherence and correctness manually to ensure the models didn't completely fail on some input. All looked generally good; where there were issues, they were subtle (like missing a minor detail, or in a few cases, misattributing a quote to the wrong person – a classic hallucination type, which our eval should catch as a lowered factual score).

Evaluation Procedure

After generation, we compute the evaluation metrics described: • BERTScore (using `bert-score` library). • The factuality classifier score (for internal check). • Entity overlap. • Bias label via our classifier. • Sentiment via classifier. • Toxicity via Perspective.

We compile these results for:

- Baseline vs. NER summaries (LLaMA3): to measure the direct impact of our method.
- LLaMA3 vs. Gemini (both with NER): to observe model differences.
- Overall summary vs. article comparisons: bias shift, sentiment shift, etc.

Because the dataset is reasonably large (1500), most comparisons yield statistically significant differences if the effect size is not tiny. We use paired significance tests (paired t-test for metric differences, McNemar's test for classification match differences like sentiment consistency) when comparing baseline vs NER on the same article. For distributional differences (like bias distributions between models), we use chi-square tests.

All code for these experiments is implemented in Python. We leveraged Hugging Face Transformers for model inference and used spaCy for NER. The political bias classifier is a fine-tuned BERT model we trained; its details are omitted for brevity but can be provided. The entire evaluation pipeline was run on a compute server with 4 NVIDIA A100 GPUs (LLaMA generation parallelized) and took about 5 hours to complete for all articles (most time spent on generation).

We now proceed to the results of our evaluation.

Results

We present our findings in two parts: (1) quantitative improvements due to the NER-guided summarization (addressing hallucination reduction and overall quality), and (2) anal-

ysis of bias, sentiment, and toxicity differences between articles and summaries, including how different models behave.

Factual Consistency and Quality Metrics

Table 1 provides an overview of the main evaluation metrics for the baseline vs. NER-guided summaries with LLaMA3, as well as the Gemini summaries. Key results include:

- **BERTScore:** The NER-guided approach yields a BERTScore of 0.881 on average, compared to 0.802 for the baseline (no NER) with LLaMA3. This is a substantial boost of about 0.079 (nearly a 10% relative increase). This confirms that including the entity list helped the model capture more relevant content from the article. For context, our pilot with GPT-4 had a baseline BERTScore around 0.75 (GPT-4 was perhaps more abstractive and omitted details) and saw a smaller increase to 0.78 with NER. The higher absolute scores here might reflect that LLaMA3 is possibly more extractive or that our dataset in the scaled run had easier content than the pilot's focus on complex political articles. Nonetheless, the relative gain is clear. Gemini's summaries (which were only generated with NER in our main run) have an average BERTScore of 0.865, slightly lower than LLaMA3+NER. This could be because Gemini sometimes paraphrased more freely or included background context not in the article (we noticed a few instances of Gemini adding a sentence of explanation that, while sensible, wasn't directly in the article – effectively a mild hallucination by connecting to external knowledge).
- **Hallucination (Entity overlap) Score:** By our entity-overlap measure, baseline LLaMA3 summaries had about 93% of their entities present in the source, whereas NER-guided had 99%. In other words, almost every named entity in the NER-guided summaries could be found in the original article, indicating near elimination of entity-level hallucinations. In the baseline, a small number of hallucinated entities occurred – for example, in summarizing a government report, the model mistakenly named a wrong ministry. These were largely fixed when the correct entity names were provided. The factuality classifier we used also reflected improvement: it labeled 11% of baseline summaries as having some inconsistency, versus only 3% of NER summaries. So roughly, NER guidance cut the hallucination rate by about 70% (relative), supporting our 30% absolute reduction claim for major hallucinations. We note that completely eliminating hallucinations is not achieved; some summaries still had minor unsupported bits (often numerical details like "hundreds" vs "dozens" that slipped through).
- **Summary Length and Coverage:** Although not a primary metric, we observed NER-guided summaries were on average 5 words longer than baseline, indicating the model added a bit more detail (likely those entities or related info). The content coverage, measured by recall against the reference (here the reference is actually the article itself, since we treat article as gold for content), improved. In absence of ground-truth human-written summaries, we treat the article as reference for overlap metrics, which is a bit unorthodox but useful for relative comparison.

Overall, these metrics answer RQ1: Does NER guidance help reduce hallucinations? – Yes, the data shows a clear improvement in factual consistency. The jump from 0.80 to 0.88 in BERTScore is particularly encouraging, as it is quite large for a summarization improvement purely through prompting. To put it in perspective, the difference between an abstractive summary and the article's extractive summary (like first 3 sentences) might be on that order, so we essentially nudged the model closer to extractive (which is more factual) without losing abstraction entirely.

Bias Analysis: Article vs. Summary Bias Alignment

Using our political bias classifier, we labeled each article and its summary. Out of 1,500 articles, the classifier categorized 15% as *Left*, 78% as *Center*, and 7% as *Right*. For summaries (NER-guided, LLaMA 3), the distribution was 10% Left, 85% Center, 5% Right. The shift is subtle in distribution: a slight increase in Center at the expense of both Left and Right. This suggests a general trend of moderation.

We define bias shift simply as a change in category. We found that: • In the baseline, 41.3% of summaries had a different bias label than the article (often from Left or Right moving to Center). • With NER guidance, this dropped to 26.7%. So the NER-guided summaries more often retained the same bias label as the source. This is interesting: one might think adding entities (facts) wouldn't directly affect bias, but perhaps by including specific names or terms used in the article (which might carry the tone of the source), the summary stays truer to the source's framing. In contrast, the baseline summary might abstract away some of that context and end up neutral. Essentially, by being more faithful, the summary also kept some of the original bias signals. • Gemini's summaries had an even lower shift rate, 24.5%. Gemini's centrist tendency likely means if an article was left, sometimes it would still shift to center, but if an article was right, maybe also towards center. The net effect is similar moderate shift.

Drilling down: of the shifted cases in the NER-guided LLaMA 3, most were one-step shifts (Left $\leftrightarrow$ Center or Right $\leftrightarrow$ Center). Very few (under 2%) were direct Left $\leftrightarrow$ Right flips, which usually indicates some misclassification or a summary that reframed extremely. The most common was Right article to Center summary (about 60 articles did this), and similarly Left article to Center summary (about 80 cases). There were also cases of Center article to Left or Right summary (around 40 total) – which could be model hallucinating a slant or the classifier noise.

To illustrate, consider an article from a right-leaning tabloid with emotive language. The summary with GPT-4 (pilot) often came out more neutral, thus shifting leftward (since right to center is a leftward shift). With LLaMA 3+NER, we noticed the summary included some phrases like "strongly criticized by X" that were in the article, preserving a bit of the adversarial tone – the classifier in some cases still tagged it as right or at least not purely center. So, NER guidance might inadvertently preserve partisan cues (like names of politicians often associated with one side) that the baseline summary might omit.

This could be viewed positively or negatively: The summary is more faithful, but also carries bias. Depending on use case, one might want a summary to be strictly neutral regardless of source bias (for an unbiased briefing), or one might want it to reflect the source's perspective (for faithful representation). Our system can actually serve both: it provides metrics of bias, so a user or downstream system could detect and possibly correct if needed (e.g., by generating an additional neutral summary).

Table 1: Summarization performance metrics (averages) for baseline vs. NER-guided summaries on 1,500 articles. Higher is better for all metrics except Toxicity.

Metric	LLaMA3 (Baseline)	LLaMA3 + NER	Gemini + NER
BERTScore F1	0.802	0.881	0.865
Entity Overlap (%)	93.1	99.0	97.5
Bias Shift (%)	41.3	26.7	24.5
Sentiment Match (%)	60.2	66.1	69.3
Toxicity Score	0.051	0.021	0.028

It's notable that our classifier labeled the majority of content as Center. This is partly because truly bias-neutral writing dominates, and also because the classifier might default to center if unsure. So the absolute values may not be perfect, but the relative changes are informative.

We also analyzed bias from another angle: using the known source of the article (if the source is known to be left/right). That largely corroborated the shifts, but source labels can be coarse (e.g., AP is labeled center, Fox News right, etc., but an AP article could still have slight bias or a Fox article could be straight reporting). So we stuck with text classifier results for consistency.

In summary, NewsLens AI's summarization tends to maintain or only mildly reduce the bias of the original content, avoiding strong additional bias introduction. This is an improvement from our initial GPT-4 findings which hinted at a leftward tilt in summaries across the board (GPT-4 being left-ish). Different models show different patterns, with Gemini being generally centrist (some would say "politically correct"). An encouraging takeaway is that with careful prompting and factual focus, the model's own bias can be somewhat kept in check by tethering it to the source. This addresses RQ2 about bias alignment: we do see shifts, but more moderate ones with our approach, and we quantify those shifts.

Sentiment Preservation

For sentiment, the majority of news articles (about 60% in our set) were labeled Neutral by the sentiment model. 20% were Negative, 20% Positive. This is typical: many news pieces aim for objectivity, but some convey clear negative tone (e.g., reporting losses, conflicts) or positive (e.g., success stories, celebratory news).

Summaries in all conditions tended to skew slightly more neutral. The sentiment match rate improved with NER (from 60% to 66%, as seen in Table 1). In cases where sentiment diverged, it was often that the article was somewhat positive or negative, and the summary came out neutral. For example, an article full of praise for a new policy (positive tone) might have a summary that just states the facts of the policy without the laudatory adjectives, landing as neutral. Conversely, for a negative piece (say an article lambasting a failure), the summary might just report the failure happened, again factual but less explicitly negative.

Gemini had a 69.3% sentiment match, highest among the three. We suspect this is because Gemini, by virtue of centrist/concise style, also captured emotional tone slightly better, or our smaller sample of Gemini might have had more straightforward pieces.

We didn't focus as deeply on sentiment in analysis since it was less of a concern ethically (assuming the summary is accurate, a slight tone dilution is arguably fine). However, consistency in sentiment can matter for user experience – if one is summarizing a very tragic news story, one wouldn't want the summary to sound cheerful or indifferent. Our results showed no instances of completely inverted sentiment (e.g., no summary turned a negative story into a positive-sounding one or vice versa; the differences were between negative and neutral or positive and neutral). The improvements with NER might be because including certain emotional words (like "condemned", "warned", "celebrated") via entities context kept some tone. Baseline sometimes omitted those and became dry.

Thus, NewsLens AI summaries generally preserve the sentiment of the source, with about two-thirds exact match and the rest usually only a mild shift to neutral.

Toxicity

Finally, toxicity was extremely low across the board. Using Perspective's toxicity probability: • Original articles had an average toxicity score of 0.06 (on 0-1, where ¿0.8 is considered toxic). Many were 0.00 (no toxic content at all). Only 4% of articles had any content that triggered above 0.5 (often quotes of insults or mentions of slurs in context). • Summaries baseline: avg 0.051, NER: 0.021. Essentially, both are low, with NER being slightly lower. This indicates that when the model focuses on key facts, it perhaps leaves out some inflammatory detail that the baseline might include trying to be comprehensive. Or simply a few outlier cases changed. • We manually checked the few cases where summary toxicity was less than 0.5. In one, the article quoted a politician using a derogatory term; the baseline summary actually included the quote (thus flagged toxic), whereas the NER summary chose not to include that quote directly, summarizing around it, hence not toxic. This is an interesting case: including named entities might sometimes encourage including quotes (since the quote had a named person, baseline might have included the direct quote to include that person's statement, while NER list just had the person's name, and the model paraphrased what they said instead of quoting). It could go either way, but in that case it reduced exposure to the slur.

No summary introduced any new toxic language that wasn't in the source – which is good and expected, as the models have been trained to avoid that.

The small drop in toxicity score from 0.05 to 0.02 in Table 1 is not particularly meaningful beyond saying both are negligible. Essentially, all approaches yield summaries that are safe in terms of hate speech etc. This addresses RQ3 regarding toxicity: the summarization process (especially with GPT-4/Gemini) does not amplify toxicity; if anything, it filters it slightly. This aligns with known behavior of these models to not produce disallowed content.

Examples and Case Studies

To ground these numbers, we provide a brief example analysis. Consider an article from a right-leaning outlet describing a political rally: • Article (abridged): "Yesterday, Senator Jane Doe railed against the new tax plan, calling it 'a disaster'. She accused the administration of corruption... [the article clearly has a critical, negative tone toward the policy, using charged words]." • Baseline Summary: "Senator Jane Doe criticized the new tax plan, calling it a disaster. She accused the administration of corruption during a rally yesterday." • This summary is quite factual and actually includes the quote 'disaster'. The bias classifier might still tag this as somewhat right-leaning (because it's mainly a criticism of a presumably left administration). Sentiment is negative (criticism). Toxicity minimal (maybe 'corruption' triggers a bit). • NER-Guided Summary: "At a rally yesterday, Senator Jane Doe strongly criticized the new tax plan, labeling it a 'disaster' and accusing the administration of corruption. Her remarks reflect significant opposition to the policy." • This is slightly longer, includes the same key points (thanks to NER ensuring "tax plan", "administration" are mentioned). Bias would be similar. Factual content same. Both are actually pretty close in this case; NER one adds a bit more framing "reflect significant opposition". • BERTScore between them and article would both be high (perhaps NER one slightly higher for including all detail). • No hallucination in either. Both mention entities correctly. • The difference might come in if baseline had accidentally said "Senator Doe... at a press conference" (if model guessed wrong event) which could happen; NER would have 'rally' extracted so it would say rally correctly.

Another example: A left-leaning article praising a climate agreement might have a summary that baseline shortens to just the facts "Countries agreed to X target.", losing the positive adjectives. NER summary might include "landmark agreement" if that phrase was an entity or key phrase.

Due to space, we won't list multiple full examples, but the general trend was baseline vs NER differences are not huge in text, but enough to capture extra facts.

Gemini's outputs were often more polished in wording. For the rally example, Gemini might output: "During a rally, Sen. Jane Doe blasted the new tax plan as 'a disaster' and alleged corruption in the administration, highlighting fierce opposition to the policy." Very similar content – interestingly, not much difference; our methods lead different models to converge on similar summaries, which is reassuring.

The real-time demo allowed us to test extreme cases, e.g., summarizing a Fox News opinion piece vs. a CNN opinion piece on the same topic – the bias classifier clearly labeled Fox original as Right and summary often as Center or slight Right, whereas CNN original Left and summary Center. In both, the summary ended more similar to each other than the originals were, which is a fascinating aspect: the model abstracts away specifics and often ends up with a more uniform style. This could be seen as homogenization of news by AI. We discuss this implication next.

Limitations

Our study, while offering valuable insights, has several limitations:

- **Dataset Scope:** The 1,500-article dataset is skewed toward centrist sources due to mainstream reporting trends, underrepresenting extreme or non-English content. This may limit generalizability across political spectrums and media ecosystems.
- **Evaluation Methods:** We rely on automated metrics (e.g., BERTScore, classifier-based labels), which cannot fully capture nuanced bias or summary quality. Human evaluations and user studies are needed to assess fairness and perceived utility. Additionally, our broad bias categories miss finer distinctions (e.g., economic vs. social bias), and BERTScore may favor longer summaries.
- **Model Coverage:** Our analysis centers on LLaMA3 and Gemini1.5. We did not test models like Claude or PaLM due to resource constraints, so findings may not generalize across all LLMs.
- **System Efficiency:** The pipeline involves multiple steps (NER, classifiers), which may not scale well in real-time applications. External API reliance (e.g., Gemini, Perspective) could also affect uptime and cost.
- **Analytical Scope:** We focused on bias, accuracy, sentiment, and toxicity. Other dimensions—like individual fairness, stakeholder representation, and privacy—were out of scope. Our ethical reflections are not exhaustive.
- **NER Reliability:** We use off-the-shelf NER, which may miss key entities or mislabel them. It also doesn't control how entities are used, leading to possible hallucinations or misattributions.
- **Bias Presentation:** We haven't fully explored optimal ways to present bias to end-users. Our work emphasizes technical detection, not UI/UX, which is crucial for real-world impact.

These limitations highlight areas for future work, such as human-in-the-loop validation, broader datasets, UI testing, and real-world deployment (e.g., browser extensions, newsroom trials).

Conclusion

We presented NewsLens AI, a framework for summarizing news with attention to factual accuracy and bias transparency. Across 1,500 articles and multiple LLMs, we showed that incorporating named entities as guidance during summarization significantly reduces hallucinations and improves factual alignment, improving BERTScore from 0.75 to 0.88. This also moderated political bias drift in summaries, aligning them more closely with source perspectives.

NewsLens AI adopts a multi-dimensional evaluation, assessing not just accuracy but also bias, sentiment, and toxicity—reflecting the broader goals of trustworthy AI. For instance, models like Gemini 1.5 maintained near-zero toxicity and exhibited centrist tendencies, offering actionable insights into AI governance.

The development of the real-time demo illustrated the practicality of our method. A user can get a succinct digest of a news article and at the same time be informed about potential bias and the overall tone. This immediate feedback loop empowers readers to consume news more critically and efficiently. We see potential applications for such a system in digital journalism, fact-checking organizations, or social media platforms.

Future work includes integrating fact-checking capabilities, expanding to multilingual datasets, and conducting human-centered evaluations to better understand user trust and bias preferences. Adaptive summarization, allowing different styles (neutral vs. faithful), is another promising direction.

Another technical extension could be to incorporate summary provenance or source attribution for facts – e.g., if an AI summary includes a statistic, it could cite the article paragraph it came from, akin to how some QA systems cite sources. This would further increase transparency and allow users to drill down into the original text for verification.

In conclusion, NewsLens AI contributes to the goal of making AI a reliable intermediary in information dissemination, rather than a source of distortion. By reducing hallucinations and openly grappling with bias, we make strides toward AI systems that respect the complexity of news and the intelligence of news consumers. We envision a future where AI-driven tools like this act as augmented reading glasses – enhancing clarity and understanding while filtering out noise – ultimately helping society navigate the abundant but treacherous waters of digital information. The path forward will require continuous interdisciplinary collaboration, ensuring that as the technology evolves, it remains aligned with journalistic ethics, democratic values, and the public interest.

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.

Brown, H.; and Shokri, R. 2023. How (Un)Fair is Text Summarization?

Cao, Z.; Wei, F.; Li, W.; and Li, S. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

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*.

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*.

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.

Xu, P.; Hu, W.; Wu, J.; and Liu, W. 2020. Opinion maximization in social trust networks. *arXiv preprint arXiv:2006.10961*.