Quantifying Cognitive Bias Induction in LLM-Generated Content

Anonymous ACL submission

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 to which extent 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.

1 Introduction

LLMs perform well across numerous tasks (Albrecht et al., 2022), such as content summarization (Laban et al., 2023), translation (Elshin et al., 2024), question-answering (Lin et al., 2025), sentiment analysis (Zhang et al., 2024). In many of these tasks, humans rely on LLMs for daily decision-making support (Rastogi et al., 2023; Li et al., 2022) in expert contexts such as writing policy documents (Choi et al., 2024) or summarizing medical documents (Spotnitz et al., 2024). However, models have been shown to exhibit several societal biases (Zhao et al., 2018; Nadeem et al., 2020; Liang et al., 2021; He et al., 2021), e.g., favoring specific genders (Zhao et al., 2018) and ethnicities (Caliskan et al., 2017), or fall into similar decision-making-patterns as cognitively biased humans (Echterhoff et al., 2024). Any bias within the model can have an effect on how textual context is processed for the human task, and introduce biased content to a user. Our work specifically focuses on qualifying and quantifying how much biased content LLMs induce to users, as visualized in Figure 1 with examples shown in Figure 2. We sumamrize our contributions as follows:

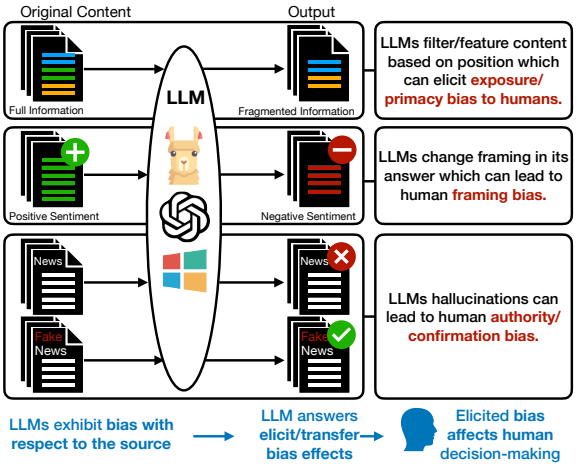

Figure 1: Our work evaluates how LLMs alter information a source text when performing a task for the user (e.g. when an LLM summary has a different sentiment compared to the original text). This model behavior introduces biased content to humans and can hence affect their decision-making. We evaluate how LLMs highlight source content leading to exposure bias for the user, how LLMs reframe text which leads to human experiencing framing bias and how LLMs hallucinate which leads to authority/confirmation bias.

1. We analyze the extent of content alteration that LLMs introduce in summarization and question-answering tasks. We find that LLMs alter the sentiment or framing of a text in 21.86% of summaries, disproportionately fo-

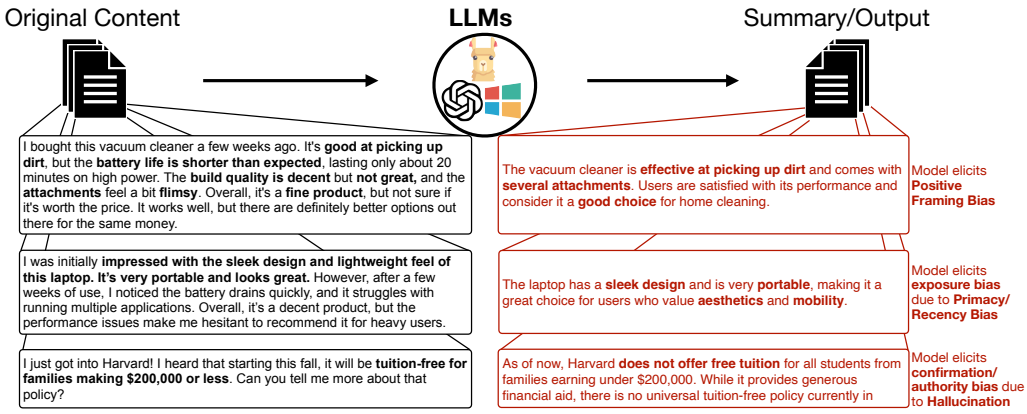

Figure 2: When LLMs process content for users to consume, they may change this content, e.g., by changing its sentiment or omitting some relevant parts. Exposure to the altered content may elicit cognitive biases such as positive framing bias or primacy bias, and subsequently have humans make different decisions than they would take if they saw the original text.

cus on the beginning of a source text in 5.94% of summaries, and hallucinate 60.33% on post-knowledge-cutoff questions.

2. We provide an incrementally self-updating dataset along with news published date to enable analysis of model hallucination after a model's cut-off date. At the time of writing this paper, it has 8554 instances. [1].

3. We evaluate 18 distinct mitigation methods and highlight which ones were effective for specific types of content alterations.

This work highlights how much content change LLMs introduce in various tasks, which subsequently has an effect on the user, how to quantify this content change, and sheds first light on the efficacy of several mitigation methods.

2 Related Work

2.1 Social and Representation Biases in LLMs

Previous work demonstrates the existence of various biases in LLMs. For example, prior work explores ways to evaluate human-like cognitive biases (Jones and Steinhardt, 2022; Echterhoff et al., 2024). Primacy and recency bias in LLMs have been shown in question-answering settings, where the models tend to be more accurate when the answer is presented in earlier or later documents (Liu et al., 2024; Peysakhovich and Lerer, 2023; Wang et al., 2023; Zhang et al., 2023; Eicher and Irgolič, 2024). When presented with options (e.g., A/B/C/D), LLMs have been shown to exhibit sensitivity to their order (Zheng et al., 2023a). Other studies have demonstrated the existence of framing bias in LLMs (Leto et al., 2024). In framing bias, alterations in the tone, e.g., from positive to negative, can change the model's decision-making (Jones and Steinhardt, 2022; Echterhoff et al., 2024).

In contrast to prior work evaluating biases inherent in model behavior, and affecting model decisions, we quantitatively assess the extent to which LLM responses alter the contextual content and hence the amount of bias introduced to humans through these modifications, hence affecting human decisions.

2.2 Cognitive Effects of LLM-Generated Content on Humans

LLMs generate content to assist users with various tasks, such as summarization (Laban et al., 2023), coding (Tong and Zhang, 2024), and writing (Spangher et al., 2024) for personal or professional (Draxler et al., 2023) purposes. When given a specific task, a model can alter, omit, or add information to produce the final output, which can affect users in several ways (Wester et al., 2024; Choi et al., 2024). Previous work demonstrates that LLMs can influence users by reinforcing their existing opinions (Sharma et al., 2024) increasing confirmation bias, or influence political decision-making (Fisher et al., 2024). The *style* in which LLMs present advice can significantly influence users' perceptions of its usefulness(Wester et al., 2024).

Prior work demonstrates that altered headlines

[1]We publish this data upon acceptance on Huggingface.

can introduce a framing effect on individuals in collective decision-making (Abels et al., 2024). Choi et al. (2024) found that even when experts use LLMs, they tend to adopt the suggestions with fewer revisions.

Prior work highlights how sensitive a user is to content displayed in different ways, but quantitative evaluation on the extent of LLM altering contextual content is still missing. Our study closes this gap by measuring the extent to which LLMs alter content through sentiment change, hallucination, and biased emphasis in summaries.

3 Background

Humans are prone to mental shortcuts that negatively influence rational decision-making, so called *cognitive biases*. These mental shortcuts are often amplified by simple changes of text, framing, or highlighting, that could be introduced by an LLM. We use three cognitive biases for our evaluation setup, and briefly introduce them here.

Framing bias refers to changes in human decisions based on *how* a problem is presented. Individuals make different choices based on the wording of questions, even when the underlying options remain unchanged (Tversky and Kahneman, 1981). Based on prior work, *we assume that a change in framing affects the user and should be avoided because it can affect their decisions*, e.g., when comparing the framing of a factual context with its summarization.

Primacy and Recency Bias. Primacy bias occurs when humans prioritize information encountered first, while recency bias occurs when humans prioritize information encountered most recently (Murphy et al., 2006). This bias skews decision-making based solely on the order in which information is encountered (Glenberg et al., 1980). When LLMs exhibit primacy or recency bias (Echterhoff et al., 2024), they may filter or highlight information, leading to exposure bias. *We assume that systematically highlighting any part of the context, regardless of its position, can influence users and should be avoided.*

Confirmation and Authority Bias. Individuals favor information confirming their existing beliefs or expectations (Nickerson, 1998). *We assume that if a user prompts a model with belief-consistent input and the model subsequently hallucinates, this can reinforce the user's prior beliefs, leading to biased or misinformed decisions.* Authority bias causes individuals to trust information from perceived experts or authoritative sources (Milgram, 1963; Cialdini, 2007; Gültekin, 2024). Research on human-AI interaction suggests that similar dynamics apply when interacting with artificial systems perceived as experts, which can create challenges when verifying facts (Glickman and Sharot, 2025). For example, if a user asks an LLM to verify whether a particular legislation was passed, a user with strong prior beliefs may exhibit confirmation bias, while others may exhibit authority bias.

4 Quantifying Content Alteration in LLMs Responses

To capture unintended content alterations introduced by LLMs, we propose the following metrics, motivated by the aforementioned human cognitive biases. We focus on content alteration introduced through content framing changes, content filtering/highlighting, and content truthfulness.

4.0.1 Framing Effects in Model-Generated Content

We test whether a model changes the sentiment of its summary compared to the source context. For example, a model may shift the summary sentiment to positive, whereas the source context was neutral. To evaluate these changes, we classify the framing of the source context f_c (positive, negative, or neutral) before summarization, as well as the framing of the generated summary f_s.[2].

We consider a model *consistent* if f_c equals f_s. We define the Bias-Inducing (BI) score as the framing-change fraction, denoted by φ_{frame} as follows:

$$\varphi_{\text{frame}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\big(f_c^{(i)} \neq f_s^{(i)}\big), \qquad (1)$$

where $\mathbb{I}(\cdot)$ is the indicator function that returns one if the condition is true and zero otherwise.

4.0.2 Primacy Effects in Model-Generated Content

We assess whether the model emphasizes a specific part of the review. For example, a model might focus solely on the beginning of an Amazon user review that starts with praise but concludes with criticism. We divide the source context into three

[2]We use GPT-4o-mini to classify the framing of the text, after evaluating different models for this task. Details are provided in Appendix A

equal segments and calculate the cosine similarity, using multilingual-E5 embeddings (Wang et al., 2024), between each segment and the summary. We denote the similarities of the summary and the beginning as s_{b_i}, the middle as s_{m_i}, and the end as s_{e_i} for each instance i. We define biased examples as those where the similarity to the beginning exceeds the similarity to the middle by at least a threshold α. We define this Primacy Bias Score as:

$$\psi_{\text{pri}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(s_{b_i} > s_{m_i} + \alpha\right) \quad (2)$$

with $\alpha = 0.05$ for this study. A higher score indicates that a greater proportion of summaries focus more on the beginning of the source content than on the middle.

4.0.3 Hallucination Effects in Model-Generated Content

We quantify whether a model generates hallucinated content when asked to fact-check news content from after its knowledge cutoff. For example, a model may incorrectly state that an event never occurred, even if it did occur after the model's training was completed. Using pre- and post-knowledge-cutoff data, we prompt a model to verify the truth of specific news items. We test the accuracy of both true, real-world news (Actual News Accuracy) and its falsified (negated) counterparts (Falsified News Accuracy). We prompt the model to evaluate the factual accuracy of a given news description. We propose evaluating the strict accuracy in correctly identifying both true and falsified news.

For strict accuracy, we evaluate paired instances of true news and its falsified counterparts:

$$s_H^T(D_j) = \frac{1}{|P_j|} \sum_{(x_i^t, x_i^f) \in P_j} \mathbb{I}(\hat{y}_i^t \wedge \neg \hat{y}_i^f) \quad (3)$$

where P_j is the set of paired instances in dataset D_j, $\hat{y}_i^t$ and $\hat{y}_i^f$ are the model's predictions for the true and falsified versions of the same news item, respectively, and T represents the time horizon (pre- or post-model-knowledge-cutoff).

We analyze the disparity between pre-cutoff and post-cutoff performance:

$$\delta_H = |s_H^{pre} - s_H^{post}| \quad (4)$$

A larger δ_H indicates a higher likelihood of generating incorrect information when responding to queries about data beyond the model's knowledge cutoff.

4.1 Mitigation Strategies

To mitigate biased patterns in LLM responses, we build upon our quantification of content alterations, as described in Section 4. We introduce targeted interventions designed to ensure balanced content coverage, reduce misleading emphasis, and maintain accuracy in model outputs. We analyze a total of 18 mitigation methods, but describe only the most promising in this section. All other mitigation methods are described in Appendix C.

4.1.1 General Purpose Mitigations

Self-Awareness Prompt (SA). Prior work suggests that simple awareness prompts can mitigate certain biases (Mair et al., 2014; Echterhoff et al., 2024). For example, in the summarization setting, we use the following prompt:

> "You are an unbiased summarizer; be mindful not to omit coverage of any portion of the text or alter its sentiment. Do not incorporate framing bias."

Chain-of-Thought Prompting (CoT). Chain-of-Thought prompting encourages the model to decompose complex problems by generating a series of intermediate reasoning steps before producing a final answer (Wei et al., 2023). In the context of summarization, we guide the model to articulate steps for processing different parts of the input text. For the question-answering setting, we explicitly define when to return specific responses during fact-checking.

4.1.2 Primacy Effect Mitigation

We group interventions into four categories: prompt-level cues, chunk-based summarization, attention/order re-ranking, and decoding-time control. Each category targets the tendency of LLMs to overemphasize early context (primacy) or alter the source sentiment. A compact reference of all the methods we experimented with is provided in Appendix C.

Weighted Summaries is a chunk-based approach that ensures balanced representation of the source text. The algorithm first splits the document into three segments (beginning, middle, and end). It then generates a partial summary for each segment independently, constraining the output length for each part according to a pre-allocated token budget (33%, 34%, and 33% of the total desired summary length). These partial summaries are then combined to form the final output.

In the *Mirostat Decoding* approach, the model's sampling randomness (temperature) is dynamically adjusted during generation. This ensures the output text maintains a consistent level of unpredictability, which can prevent the model from focusing too narrowly on initial sentences and encourage broader coverage of the source text. We use a target entropy of $\mu^{\star} = 2.0$ in our runs, similar to the original paper (Basu et al., 2021).

4.1.3 Framing Effect Mitigation

In the *Weighted Token Decoding* approach, we directly influence the model's token selection during generation. At each step, we modify the output probabilities by adding a logarithmic weight to the logits of specific tokens before sampling (Liu et al., 2021; Dathathri et al., 2020). In particular, we down-weight tokens associated with negative sentiment to reduce framing shifts, as detailed in Appendix C.

4.1.4 Hallucination Mitigation

The mitigation strategies for hallucination are detailed in Appendix C.

We propose *Knowledge Boundary Awareness* which explicitly defines the model's knowledge cutoff date, preventing the fabrication of information about events beyond the training data.

Epistemic Tagging requires models to express confidence levels alongside factual assertions, forcing self-evaluation of fact-checking abilities (i.e., the model additionally responds with a high or low confidence level for each response) (Lin et al., 2022). Prompt details are provided in Appendix E.

4.2 Datasets

News Interviews News summaries should provide broad coverage of the content, regardless of content position, as all information is potentially relevant to the summary. As these interviews are supposedly an objective depiction of current states and affairs, a summary should not be more positive or negative than the original content. We use MediaSum, a dialogue summarization dataset comprising media interviews from NPR and CNN (Zhu et al., 2021), to measure framing and primacy effects. We randomly select 1,000 data points with no more than 4,000 tokens each, ensuring that the full content fits within a model's context window.

Product Reviews Previous work suggests that changing the framing of the product influences consumer purchase decisions (Wei et al., 2024). To explore this effect in the context of LLM outputs, we use the Amazon Reviews dataset (Ni et al., 2019). This dataset contains customer reviews from Amazon across various product categories, providing information on product quality, usability, and customer satisfaction (Ni et al., 2019). We sample 1,000 random examples from the Electronics category, ensuring token lengths of no more than 4,000 to maintain full-context input for the model.

News Hallucination We provide a dataset to quantify how well models discern factual information across different time periods and between true and falsified versions of news items.

Pre-Knowledge Cutoff Data. For the pre-knowledge cutoff evaluation, we use the "All the news" data (all, 2020), which includes articles collected from 27 major American publications from 2016 to 2020. This dataset is available for non-commercial research purposes. We sampled 2,700 random data points in our experiment.

Post-Knowledge Cutoff Data. To evaluate model performance on recent events (i.e., events not included in a model's training data), we introduce NewsLensSync (Anonymous, 2025), a self-updating dataset specifically designed to measure ongoing hallucinations regarding content beyond models' knowledge cutoff dates. We evaluate 2,801 real news items and their negated versions.

For both post- and pre-cutoff examples, we include factual news items and carefully crafted counterfactual versions. We used a transformer-based model to generate the counterfactual versions (Anschütz et al., 2023).

4.3 Models

We use four open- and closed-source language models. Details on the models and our hyperparameter settings are provided in Appendix B.

5 Results

5.1 LLMs Alter Framing During Summarization

We observe that models alter the framing of summaries compared to the source content in both media interviews and the Amazon dataset (Table 1). Most models (Llama and GPT-3.5-turbo) exhibit a framing change fraction between 14.5% and 16%, whereas Phi-3-mini-4k-Instruct shows 34.5% framing changes. In Table 3, we show examples of a framing change between a summary and the ground-truth context. Table 2 provides a detailed

Method	Amazon Reviews					MediaSum News				
	$\varphi_{\text{frame}} \downarrow$	$\bar{s}_b \uparrow$	$\bar{s}_m \uparrow$	$\bar{s}_e \uparrow$	$\psi_{\text{pri}} \downarrow$	$\varphi_{\text{frame}} \downarrow$	$\bar{s}_b \uparrow$	$\bar{s}_m \uparrow$	$\bar{s}_e \uparrow$	$\psi_{\text{pri}} \downarrow$
GPT-3.5-turbo	16.0%	0.848	0.826	0.825	7.6%	24.8%	0.840	0.823	0.820	6.1%
Llama-3.2-3B-Instruct	14.9%	0.860	0.842	0.840	7.4%	22.1%	0.851	0.837	0.832	5.1%
Llama-3-8B-Instruct	14.5%	0.855	0.837	0.836	7.0%	21.9%	0.847	0.834	0.828	4.0%
Phi-3-mini-4k-Instruct	34.5%	0.836	0.822	0.823	5.4%	26.2%	0.828	0.827	0.823	4.9%

Table 1: Models introduce framing changes (φ_{frame}) and filter/feature content based on position (ψ_{pri}). φ_{frame} is the proportion of examples where the framing changed from the original to the summary (lower is better). $\bar{s}_b$, $\bar{s}_m$, and $\bar{s}_e$ represent the average cosine similarity between the summary and the beginning, middle, and end thirds of the source text, respectively. Higher values of $\bar{s}_b$, $\bar{s}_m$, and $\bar{s}_e$ indicate stronger content preservation from those respective sections. ψ_{pri} represents the percentage of examples with primacy bias as defined in Equation 2.

Dataset	Model	Neu→Pos	Neu→Neg	Pos→Neg	Neg→Pos	Pos→Neu	Neg→Neu
MediaSum	**GPT-3.5-turbo**	2.9%	8.6%	0.1%	0.1%	1.9%	11.2%
	Llama-3-8B-Instruct	2.6%	8.0%	0.0%	0.1%	2.3%	8.9%
	Llama-3.2-3B-Instruct	2.0%	7.7%	0.1%	0.1%	2.3%	9.9%
	Phi-3-mini-4k-Instruct	2.6%	7.3%	0.1%	0.7%	2.4%	13.1%
Amazon	**GPT-3.5-turbo**	2.6%	0.5%	0.2%	1.4%	7.4%	3.9%
	Llama-3-8B-Instruct	2.3%	0.6%	0.7%	1.4%	6.7%	2.8%
	Llama-3.2-3B-Instruct	1.8%	0.6%	0.5%	2.1%	6.2%	3.7%
	Phi-3-mini-4k-Instruct	1.6%	0.8%	1.8%	2.2%	21.2%	6.9%

Table 2: Framing category transitions in MediaSum and Amazon Reviews datasets. Values represent the percentage of summaries transitioning from framing category 'x' to 'y' (x→y).

evaluation of framing alterations (e.g., from neutral to positive, or positive to negative). In the MediaSum news interviews, most framing shifts are from neutral to negative between 7.3% and 8.6% and negative to neutral between 8.9% and 13.1%. In Amazon reviews, the most common shift is from positive to neutral between 6.2% and 7.4% for Llama and GPT-3.5-turbo models and 21.2% for Phi-3-mini-4k-Instruct. This suggests that the models tend to downplay positive sentiment in product reviews. This pattern is consistent across all evaluated models.

5.2 LLM Summaries have Imbalanced Context Coverage

We find that, on average, summaries align more closely with the beginning of the source text than with the middle and end sections (Table 1). For example, across all models and datasets, the average similarity to the beginning segment ranges from 0.828 to 0.860, consistently higher than the similarity to the middle (0.822-0.842) and end (0.820-0.840) segments.

We further quantify this coverage imbalance using our primacy bias score, which captures the proportion of examples where the beginning content is prioritized by at least 5% over other middle section. For the Amazon Reviews dataset, ψ_{pri} ranges from 5.4% to 7.6%. Similarly, MediaSum summaries exhibit ψ_{pri} scores between 4.0% and 6.1%, which shows primacy bias even in a more structured news content.

5.3 LLMs Hallucinate on Post-Knowledge-Cutoff Data

Baseline results in Table 4 show consistent decreases in strict accuracy (s_H^T) from pre-cutoff to post-cutoff data across all models. Llama-3-8B-Instruct declined from 26% to 21% ($\delta_H = 5\%$), Llama-3.2-3B-Instruct dropped from 19% to 13% ($\delta_H = 6\%$), and Phi-3-mini-4k-Instruct showed a $\delta_H = 4\%$ decline.

While the modest difference in strict accuracy between pre- and post-cutoff data suggests that a model's performance does not deteriorate dramatically when encountering newer events, the consistently low strict accuracy values highlight a critical limitation. Specifically, even when models achieve relatively high accuracies on true and falsified news separately, their low strict accuracy reveals a persistent inability to reliably differentiate fact from fabrication. Strong performance on isolated real or falsified news does not guarantee that models can reliably distinguish truth from falsehood when presented together, highlighting the need for improved prompting or training to enhance factual accuracy.

Original Review	GPT-3.5-turbo Summary
"This is a fun, useful tablet for the price. I love that you can make all different homepages. Love that you can pick wallpaper which you can't do with Kindle Fires, which are the only other tablets I've owned. Of course, if you want a tablet mainly for books, Kindle is the tablet you want. I also love all the apps you can get with the Samsung. What I don't like is the 8 GB storage. I was under the impression that if you got the 32 GB extra, it would integrate with the tablet rather than just be like a drive for MP3s and such. 8 gigs is barely enough for the apps you need. Also, the manual tells you nothing about how to add things to the extra drive. Had to go to YouTube. My tablet has connection issues from time to time. I still really enjoy it most of the time. It can be super fast and plays things well from the cloud. Very light and attractive. I am returning the tablet. 8 GB is just not enough. Extra storage is cumbersome to use and doesn't work half of the time. Not easy like the cloud. Speakers unimpressive. I am disappointed because I thought I would like the Samsung. Also a battery hog. Will try the Nexus 7."	"I enjoyed the Samsung tablet for its customization options and app variety, but the 8 GB storage was insufficient. The extra storage was cumbersome and unreliable. Connection issues and battery drain were also drawbacks. I will be returning it and trying the Nexus 7 instead."

Table 3: Example of framing shift: from neutral (original text) to negative (summary).

Models	Prompt Strategy	Pre-cutoff data			Post-cutoff data		
		Actual News Acc.	Falsified News Acc.	Strict Acc. (s_H^T)	Actual News Acc.	Falsified News Acc.	Strict Acc. (s_H^T)
Llama-3-8B-Instruct	Baseline	74%	46%	26%	75%	42%	21%
	Prompt Calibration (CoT)	39%	85%	**27%**	36%	84%	**22%**
	Knowledge Boundary Aware	65%	54%	25%	50%	63%	18%
	Epistemic Tagging	80%	35%	18%	78%	37%	19%
Llama-3.2-3B-Instruct	Baseline	48%	58%	19%	35%	72%	15%
	Prompt Calibration (CoT)	48%	56%	17%	46%	63%	15%
	Knowledge Boundary Aware	25%	87%	15%	15%	94%	10%
	Epistemic Tagging	51%	68%	**26%**	53%	64%	**23%**
Phi-3-mini-4k-Instruct	Baseline	14%	97%	12%	9%	98%	8%
	Prompt Calibration (CoT)	3%	99%	3%	3%	99%	3%
	Knowledge Boundary Aware	3%	99%	3%	2%	99%	2%
	Epistemic Tagging	45%	76%	**31%**	53%	70%	**29%**

Table 4: Accuracies on real news items and their falsified counterparts, evaluated before and after each model's training cutoff date. Strict Accuracy $s_H^T(D_j)$ counts predictions where the model gets both members of each (True, False) pair correct.

5.4 Special Biases Require Special Mitigations

The 18 mitigations we study (all definitions in Appendix C) do not alleviate all biases simultaneously. Each method involves trade-offs among the different biases, and its effectiveness varies depending on the model and corpus. For example, techniques aimed at reducing position-based salience (e.g., primacy bias) often work by redistributing attention across the input. However, this can disrupt how the model maintains consistent sentiment in its responses.

5.4.1 Primacy Effect and Coverage

The ***Weighted Summaries*** method, which assigns a fixed token budget to each text segment, consistently increases overall content coverage. For instance, with Llama-3-8B-Instruct on Amazon Reviews, it boosts average coverage similarity from a baseline of 0.843 to 0.910. However, this comes at the cost of an increased primacy bias score (ψ_{pri}), which increases from 7.0% to 15.1%.

Method	$\bar{s}^{\dagger}$	$\psi_{\text{pri}}\%$ ‡	$\varphi_{\text{frame}}\%$ §
Baseline	0.843	**7.0%**	14.5%
Weighted Summaries	**0.910**	15.1%	15.1%
Mirostat Decoding	0.902	15.5%	13.8%
Pos.-Invariant Shuffle	0.891	11.5%	21.2%
Weighted Token Decoding	0.858	18.4%	**13.6%**
Self-Awareness Prompt	0.889	22.9%	15.2%

Table 5: Key mitigation results on Amazon Reviews for Llama-3-8B-Instruct. †Mean content-coverage similarity $(\bar{s}_b + \bar{s}_m + \bar{s}_e)/3$. ‡Primacy-bias score. §Framing-change rate. Lower is better for ψ_{pri} and φ_{frame}. Behavior on other corpora/models follows a similar qualitative pattern; see full results in Tab. 9.

Mirostat Decoding also improves coverage; it is the *only* method that reduces ψ_{pri} on the smaller Phi-3 model (-1.2 percentage points) but worsens it on larger models.

Position-Invariant Shuffle is useful diagnostically because it helps confirm that the model relies on word order. By randomly shuffling sentences, we

remove positional cues; the resulting change in output reveals the model's sensitivity to input structure. While this method increases overall coverage (e.g., from 0.843 to 0.891 for Llama-3-8B), it degrades other metrics, increasing the primacy score ψ_{pri} to 11.5% and the framing change rate φ_{frame} to 21.2%, and comes at the risk of loosing temporal information.

5.4.2 Framing Effect and Consistency

Weighted Token Decoding is the most effective method for mitigating framing bias. By downweighting negative lexemes during decoding, it achieves the lowest framing-change fraction (φ_{frame}) for Llama-3-8B, reducing it by 0.9 percentage points to 13.6%. However, this targeted intervention significantly worsens primacy bias, with ψ_{pri} increasing from 7.0% to 18.4%.

Prompt-only nudges. Lightweight prompts, such as **Self-Awareness** or **Chain-of-Thought**, alter φ_{frame} by only a few percentage points in most settings. However, the Self-Awareness variant yields the best framing result on MediaSum for Llama-3-8B-Instruct, reducing φ_{frame} to 21.0% from a 21.9% baseline at a negligible computational cost.

5.4.3 Prompt Calibration and Epistemic Tagging can improve Hallucination

Applying **Prompt Calibration with Chain-of-Thought** enhanced strict accuracy for Llama-3-8B Instruct but showed limited utility for others. It improved the falsified news detection by 40% for pre and post cut-off data.

Knowledge Boundary Awareness improved Llama-3-8B-Instruct's performance by 8% & 21% in falsified news detection in pre and post cut-off data, respectively, alongside maintaining its strict accuracy close to the baseline. For the other models, the performance declined, despite observing significant improvement in the falsified news data detection.

Epistemic Tagging performed best across all models. It associates a confidence level during the fact-checking process (i.e., high or low) with its response about whether an event occurred (Table 6) Phi-3-mini-4k-Instruct outperformed other models, achieving a substantial increase in strict accuracy of 21% on post cut-off data and by 19% on pre cut-off data. The Llama-3.2-3B-Instruct also performed well with strict accuracies improved by 8% on post cut-off data and by 7% on pre cut-off data along with a balanced independent accuracies for real and falsified news data above 50%.

Models	cutoff	Actual News		Falsified News	
		High	Low	High	Low
Llama-3-8B-Instruct	Pre	99.0%	1.0%	99.8%	0.2%
	Post	98.3%	1.7%	99.9%	0.1%
Llama-3.2-3B-Instruct	Pre	75.8%	24.2%	24.6%	75.4%
	Post	66.7%	33.3%	83.3%	16.7%
Phi-3-mini-4k-Instruct	Pre	93.8%	6.2%	86.2%	13.8%
	Post	98.1%	1.9%	95.4%	4.6%

Table 6: Epistemic Tagging prediction Confidence levels (high, low) for different models across cutoff conditions.

6 Conclusion

When LLMs modify the framing or emphasize certain aspects of the content they process, they can inadvertently shape human perception and decision-making when humans are exposed to the content. This paper quantifies such content changes in summarization and news fact-checking tasks. We find that models significantly alter content. On average, models introduce framing bias to the user in 21.86% of instances. The studied models introduce primacy bias in 5.94% of the cases, where summaries disproportionately reflect content from the beginning of the original text. We find that on average, users are exposed to hallucinations on post-knowledge-cutoff content in 60.33% of instances.

We evaluate 18 mitigation strategies, both specific to each bias category and generalized ones. We find that the effectiveness of each approach depends on the model and the targeted bias. *Weighted Summaries* and *Mirostat Decoding* showed promise in reducing framing changes and primacy biases, particularly in smaller and mid-sized models. *Epistemic Tagging* consistently improved factual reliability regarding post-knowledge-cutoff data for smaller models.

This paper represents a step toward careful analysis and mitigation of content alteration, reducing the risk of LLMs introducing systemic biases into decision-making processes across domains like media, education, and public policy.

7 Limitations

This study examines the extent to which LLM outputs alter original content in ways that may influence human judgment. While our assumptions are grounded in prior cognitive science research, as discussed, we plan to validate these effects through a dedicated user study measuring how such alter-

ations affect human decision-making. Most of our mitigation methods were designed with one specific issue in mind (e.g., framing, hallucinations, or primacy effects). In our future work, we intend to explore more general mitigation techniques that address overlapping biases simultaneously.

Our current Primacy Bias score uses a fixed, generic threshold to identify summaries that disproportionately focus on the beginning of the source text. While this threshold provides a consistent way to detect bias, our future work aims to explore ways to refine it.

Our work aims to encourage the NLP community to continuously test LLMs for biases that can influence users' decision-making, which is why we introduce the self-updating NewsLensSync dataset (Anonymous, 2025). Currently, the dataset holds articles related to the "political" domain. In our future work, we aim to incorporate other areas to make it more generalizable and usable across a diverse set of researchers. However, our data could be misused for training models to reinforce existing ideological narratives or use falsified data. We urge researchers to be cautious in framing and contextualizing their use of the data. We provide our data under the CC-BY-4.0 license.

Ethical Considerations.

Our analysis shows that LLMs pay more attention to content from the beginning of the context, hallucinate facts, and alter the text framing in their outputs. These alterations can have an impact on the user and skew their decision-making. These findings underscore the risks of working with LLMs, even in everyday tasks.

Experiments. All experiments are run on NVIDIA RTX A6000 for open-source models and the official APIs for closed-source models with a fixed random seed. Our run time was approximately 360 GPU hours. All prompts details are included in Appendix E.

References

2020. All the news 2.0: 2.7 million news articles and essays from 27 american publications (2016–2020). https://components.one/datasets/all-the-news-2-news-articles-dataset. Accessed: 2025-05-16.

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Axel Abels, Elias Fernandez Domingos, Ann Nowé, and Tom Lenaerts. 2024. Mitigating biases in collective decision-making: Enhancing performance in the face of fake news. *arXiv preprint arXiv:2403.08829*.

Joshua Albrecht, Ellie Kitanidis, and Abraham J Fetterman. 2022. Despite" super-human" performance, current llms are unsuited for decisions about ethics and safety. *arXiv preprint arXiv:2212.06295*.

Anonymous. 2025. Newslenssync (revision 9c97ffe).

Miriam Anschütz, Diego Miguel Lozano, and Georg Groh. 2023. This is not correct! negation-aware evaluation of language generation systems. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 163–175. Association for Computational Linguistics.

Sourya Basu, Govardana Sachitanandam Ramachandran, Nitish Shirish Keskar, and Lav R. Varshney. 2021. Mirostat: A neural text decoding algorithm that directly controls perplexity. *Preprint*, arXiv:2007.14966.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Alexander S Choi, Syeda Sabrina Akter, JP Singh, and Antonios Anastasopoulos. 2024. The llm effect: Are humans truly using llms, or are they being influenced by them instead? *arXiv preprint arXiv:2410.04699*.

Robert B Cialdini. 2007. *Influence: The psychology of persuasion*. Collins New York.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. *Preprint*, arXiv:1912.02164.

Chenhe Dong, Yuexiang Xie, Yaliang Li, and Ying Shen. 2023. Counterfactual debiasing for generating factually consistent text summaries. *Preprint*, arXiv:2305.10736.

Fiona Draxler, Daniel Buschek, Mikke Tavast, Perttu Hämäläinen, Albrecht Schmidt, Juhi Kulshrestha, and Robin Welsch. 2023. Gender, age, and technology education influence the adoption and appropriation of llms. *arXiv preprint arXiv:2310.06556*.

Jessica Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in high-stakes decision-making with llms. *arXiv preprint arXiv:2403.00811*.

J. E. Eicher and R. F. Irgolič. 2024. Reducing selection bias in large language models. *Preprint*, arXiv:2402.01740.

Denis Elshin, Nikolay Karpachev, Boris Gruzdev, Ilya Golovanov, Georgy Ivanov, Alexander Antonov, Nickolay Skachkov, Ekaterina Latypova, Vladimir Layner, Ekaterina Enikeeva, and 1 others. 2024. From general llm to translation: How we dramatically improve translation quality using human evaluation data for llm finetuning. In *Proceedings of the Ninth Conference on Machine Translation*, pages 247–252.

Jillian Fisher, Shangbin Feng, Robert Aron, Thomas Richardson, Yejin Choi, Daniel W Fisher, Jennifer Pan, Yulia Tsvetkov, and Katharina Reinecke. 2024. Biased ai can influence political decision-making. *arXiv preprint arXiv:2410.06415*.

Arthur M Glenberg, Margaret M Bradley, Jennifer A Stevenson, Thomas A Kraus, Marilyn J Tkachuk, Ann L Gretz, Joel H Fish, and BettyAnn M Turpin. 1980. A two-process account of long-term serial position effects. *Journal of Experimental Psychology: Human Learning and Memory*, 6(4):355.

M. Glickman and T. Sharot. 2025. How human–ai feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, 9:345–359.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Duygu Güner Gültekin. 2024. Understanding and mitigating authority bias in business and beyond. In *Overcoming cognitive biases in strategic management and decision making*, pages 57–72. IGI Global Scientific Publishing.

Zexue He, Bodhisattwa Prasad Majumder, and Julian McAuley. 2021. Detect and perturb: Neutral rewriting of biased and sensitive text via gradient-based decoding. *arXiv preprint arXiv:2109.11708*.

Erik Jones and Jacob Steinhardt. 2022. Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems*, 35:11785–11799.

Philippe Laban, Wojciech Kryściński, Divyansh Agarwal, Alexander Richard Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. 2023. Summedits: Measuring llm ability at factual reasoning through the lens of summarization. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9662–9676.

Alexandria Leto, Elliot Pickens, Coen D Needell, David Rothschild, and Maria Leonor Pacheco. 2024. Framing in the presence of supporting data: A case study in us economic news. *arXiv preprint arXiv:2402.14224*.

Shuang Li, Xavier Puig, Chris Paxton, Yilun Du, Clinton Wang, Linxi Fan, Tao Chen, De-An Huang, Ekin Akyürek, Anima Anandkumar, and 1 others. 2022. Pre-trained language models for interactive decision-making. *Advances in Neural Information Processing Systems*, 35:31199–31212.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research (TMLR)*. https://openreview.net/forum?id=8s8K2UZGTZ.

Xin Lin, Zhenya Huang, Zhiqiang Zhang, Jun Zhou, and Enhong Chen. 2025. Explore what llm does not know in complex question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24585–24594.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. Dexperts: Decoding-time controlled text generation with experts and anti-experts. *Preprint*, arXiv:2105.03023.

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*, 12:157–173.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. *Preprint*, arXiv:1905.13164.

Carolyn Mair, Martin Shepperd, and 1 others. 2014. Debiasing through raising awareness reduces the anchoring bias.

Meta AI. 2024. Introducing llama 3: The next generation of open large language models.

Stanley Milgram. 1963. Behavioral study of obedience. *The Journal of abnormal and social psychology*.

Jamie Murphy, Charles Hofacker, and Richard Mizerski. 2006. Primacy and recency effects on clicking behavior. *Journal of computer-mediated communication*, 11(2):522–535.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 188–197.

Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*.

OpenAI. 2024. Gpt–3.5 turbo model card. https://platform.openai.com/docs/models/gpt-3.5-turbo. Model documentation, accessed 10 May 2025.

Litu Ou and Mirella Lapata. 2025. Context-aware hierarchical merging for long document summarization. *Preprint*, arXiv:2502.00977.

Alexander Peysakhovich and Adam Lerer. 2023. Attention sorting combats recency bias in long context language models. *arXiv preprint arXiv:2310.01427*.

Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, Harsha Nori, and Saleema Amershi. 2023. Supporting human-ai collaboration in auditing llms with llms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 913–926.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Preprint*, arXiv:2103.00453.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. *Preprint*, arXiv:1704.04368.

Nikhil Sharma, Q Vera Liao, and Ziang Xiao. 2024. Generative echo chamber? effect of llm-powered search systems on diverse information seeking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Alexander Spangher, Nanyun Peng, Sebastian Gehrmann, and Mark Dredze. 2024. Do LLMs plan like human writers? comparing journalist coverage of press releases with LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21814–21828, Miami, Florida, USA. Association for Computational Linguistics.

Matthew Spotnitz, Betina Idnay, Emily R Gordon, Rebecca Shyu, Gongbo Zhang, Cong Liu, James J Cimino, and Chunhua Weng. 2024. A survey of clinicians' views of the utility of large language models. *Applied Clinical Informatics*, 15(02):306–312.

Weixi Tong and Tianyi Zhang. 2024. CodeJudge: Evaluating code generation with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20032–20051, Miami, Florida, USA. Association for Computational Linguistics.

Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *science*, 211(4481):453–458.

David Wan, Jesse Vig, Mohit Bansal, and Shafiq Joty. 2024. On positional bias of faithfulness for long-form summarization. *Preprint*, arXiv:2410.23609.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Yiwei Wang, Yujun Cai, Muhao Chen, Yuxuan Liang, and Bryan Hooi. 2023. Primacy effect of chatgpt. *arXiv preprint arXiv:2310.13206*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Qiang Wei, An Bao, Dong Lv, Siyuan Liu, Si Chen, Yisi Chi, and Jiajia Zuo. 2024. The influence of message frame and product type on green consumer purchase decisions: an erps study. *Scientific Reports*, 14(1):23232.

Joel Wester, Sander De Jong, Henning Pohl, and Niels Van Berkel. 2024. Exploring people's perceptions of llm-generated advice. *Computers in Human Behavior: Artificial Humans*, page 100072.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

Zheyuan Zhang, Jifan Yu, Juanzi Li, and Lei Hou. 2023. Exploring the cognitive knowledge structure of large language models: An educational diagnostic assessment approach. *arXiv preprint arXiv:2310.08172*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023a. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. Mediasum: A large-scale media interview dataset for dialogue summarization. *arXiv preprint arXiv:2103.06410*.

A LLM as a framing judge

Previous work has shown that high-capability LLMs like GPT-4 can achieve over 80% agreement with human experts on question answering tasks, the level of agreement observed between humans themselves (81%)(Zheng et al., 2023b). GPT-4 reached 85% agreement with human judgments, suggesting that LLMs can serve as effective proxies for human preferences (Zheng et al., 2023b). Thus, we test three high-capability models to classify the framing (GPT-3.5-turbo, GPT-4-turbo, and GPT-4o-mini). We experimented on a sample of 500 randomly selected Amazon reviews, using the user-provided rating as ground truth to evaluate the accuracy of the models. We asked the models to rate the product based on the review, then we mapped ratings 1 and 2 as negative, 3 as neutral, and 4 and 5 as positive. The GPT-3.5-turbo, GPT-4-turbo, and GPT-4o-mini achieved accuracy of 0.89, 0.91, and 0.92, respectively. Based on this, we selected GPT-4o-mini to serve as the framing judge.

B Model setup and Hyperparameters

Our implementation utilizes the Hugging Face Transformers library using the official APIs. We employ Flash Attention 2 when available to enhance computational efficiency during inference. For consistency across experimental conditions, we maintain fixed generation parameters with `temperature=0.01`, `do_sample=False`, and `max_new_tokens= 500` set to a constant with fixed random seed. Table 7 shows the models we evaluated, their context window, and the rationale behind choosing them.

Model	Context Window	Reason
GPT-3.5-turbo	16k	A closed-source, highly capable model that is cost-effective (OpenAI, 2024).
Llama-3-8B-Instruct	8k	Outperforms competing models on both per-category win rate and average per-category score (Grattafiori et al., 2024).
Llama-3.2-3B-Instruct	8k	Lightweight yet powerful; offers multilingual text-generation abilities (Meta AI, 2024).
Phi-3-mini-4k-Instruct	4k	Matches high-performance closed-source models on several key benchmarks (Abdin et al., 2024).

Table 7: Models used in our evaluation, along with their context-window sizes and selection rationale.

C Detailed Descriptions of Mitigation Methods

Table 8 provides a concise, side-by-side reference for all 18 mitigation approaches evaluated in this work, including their intended bias target(s), operational definition, key hyper-parameters, and the original citation.

Family	Method	Operational definition	Target bias	Key hyper-parameters	Citation
Prompt	Self-Awareness Prompt	Prepend an explicit directive that the model remain neutral, preserve sentiment, and cover all parts of the source text.	General	–	(Mair et al., 2014; Echterhoff et al., 2024)
	Chain-of-Thought Prompting	Model generates intermediate reasoning steps for each segment before emitting a final summary, encouraging balanced coverage.	General	–	(Wei et al., 2023)
	Cloze-Style Prompt	Insert tags `BEGIN:__`, `MIDDLE:__`, `END:__`; model fills blanks, then emits `FINAL_SUMMARY`.	Primacy	Tags inside prompt; no extra threshold.	(Liu and Lapata, 2019)
	Cognitive Counterfactual Simulation	Draft → imagine how primacy, recency, or framing bias would distort it → rewrite to avoid those distortions.	Primacy, Framing	Two passes.	(Dong et al., 2023)
	Self-Help Debias	Draft, model self-critiques positional coverage, rewrites summary.	Primacy	Rewrite pass 300 tokens.	(Echterhoff et al., 2024)
	Task-Specific CoT Prompt Calibration	Establishes clear evaluation criteria by explicitly specifying when to return "True" or "False" during fact-checking tasks. For example: "Return true only if the described event has occurred, or if it is a direct consequence of a previously known event". Return false in all other cases	Hallucination	Chain-of-Thought sequence pertaining to your task	(Wei et al., 2023)
	Knowledge Boundary Awareness	Explicitly defines the model's knowledge cutoff date to create clear temporal boundaries for what the model can reasonably be expected to know. Prevents models from fabricating information about events beyond their training data by acknowledging these boundaries in prompts.	Hallucination	Knowledge cutoff date specification for the model being employed	–
	Epistemic Tagging	Requires models to express confidence levels alongside factual assertions, creating a more nuanced representation of the model's knowledge state. Forces the model to evaluate its own fact-checking abilities before responding and provides users with meta-cognitive signals about response reliability.	Hallucination	Confidence level scales - in this case we used High and Low. This can incude moderate, very high, very low depending on task.	(Lin et al., 2022)
Chunk	Partial Summaries Ensemble	Split the document into equal-length chunks and summarize each chunk independently (same generation hyperparameters). In a second pass, we merge the partial summaries into a final summary.	Primacy	Chunk size $\lfloor \lvert D \rvert / 3 \rfloor$; greedy merge.	(Ou and Lapata, 2025)
	Weighted Summaries	We pre-allocate the summary length budget so that 33% focuses on the beginning, 34% on the middle, and 33% on the end. The model is instructed not to exceed these chunk limits, which helps prevent overemphasis on earlier parts.	Primacy	Token ratio 0.33:0.34:0.33.	inspired by (Liu and Lapata, 2019)
Re-rank	Attention-Sort Re-ordering	Run a forward pass (without generating the final summary) to estimate cross-attention weights for each paragraph or segment, then reorder segments from lowest-attended to highest-attended. The model is then re-prompted with this new order, which pushes under-attended parts later in the context and encourages more balanced coverage.	Primacy	2 iterations; paragraph granularity.	–
	Position-Invariant Shuffle	Shuffle entire sentences (split by periods) to randomize their order before prompting the model, so it cannot rely on absolute position. This can disrupt semantic continuity slightly, but ensures that coverage does not depend on text location. Diagnostic ablation for positional sensitivity.	Primacy	Period-split; seed 42.	(Wan et al., 2024; Liu et al., 2024)
Decode	Forced Balanced Coverage	We measure coverage of the beginning, middle, and end sections via TF-IDF (Salton and Buckley, 1988), and add $\log \gamma$ ($\gamma = 1.5$) to logits of tokens from under-covered sections until $\lvert s_b - s_e \rvert \leq 0.05$.	Primacy	Threshold 0.05; boost $\gamma = 1.5$.	(See et al., 2017)
	Weighted Token Decoding	Multiply next-token probabilities by weights w_i (down-weight negative words, up-weight middle keywords).	Framing	$w_{\text{neg}} = 0.3$, $w_{\text{mid}} = 2.0$.	(Liu et al., 2021; Dathathri et al., 2020)
	Mirostat Decoding	After each token, compute its surprise $s_t = -\log p_t$; update the running state with $\mu_{t+1} = \mu_t - \eta\,(s_t - \mu^\star)$, set temperature $T_t = \exp(\mu_{t+1})$, and rescale logits as $\mathrm{softmax}(z/T_t)$. The feedback loop keeps the observed surprise near the target $\mu^\star$, smoothing positional coverage.	Primacy	$\mu^\star = 2.0,\ \eta = 0.1$	(Basu et al., 2021)
	Rejection Sampling	If top-1 token would increase chunk imbalance, set its logit to $-\infty$ and resample within top-k.	Primacy	$k = 5$.	–
	Self-Debias Decoding	Dual-pass decoding: at each step run a second forward pass on the current context preceded by a bias-inducing prefix that names the undesired attribute; obtain p_{bias}, compute $\Delta = p_{\text{main}} - p_{\text{bias}}$, and down-scale tokens with $\Delta < 0$ via $\alpha(\Delta) = e^{\lambda \Delta}$.	Framing	$\lambda = 10$; bias prefix < 30 tokens; refresh every 4 steps.	(Schick et al., 2021)
	Local-Explanation Guard	After each token the model explains its choice; if explanation says "ignoring middle" or "flipping sentiment", reject token and resample.	Primacy, Framing	Check every 5 steps; rule-based detector.	–

Table 8: Overview of the 18 bias-mitigation strategies evaluated in this work. Methods are grouped by the family that they belong to, annotated with the bias they aim to address, and accompanied by the main hyper-parameters and primary citation(s) used in our experiments.

D Full Results and Prompt Templates

Table 9 reports the complete results for all models and mitigation strategies.

Method	Amazon Reviews						MediaSum News					
	$\varphi_{\text{frame}} \downarrow$	$\bar{s}_b \uparrow$	$\bar{s}_m \uparrow$	$\bar{s}_e \uparrow$	$\psi_{\text{pri}} \downarrow$	$\rho_{\text{pri}} \downarrow$	$\varphi_{\text{frame}} \downarrow$	$\bar{s}_b \uparrow$	$\bar{s}_m \uparrow$	$\bar{s}_e \uparrow$	$\psi_{\text{pri}} \downarrow$	$\rho_{\text{pri}} \downarrow$
GPT-3.5-turbo												
Baseline	16.0%	0.848	0.826	0.825	7.6%	–	24.8%	0.840	0.823	0.820	6.1%	–
Llama-3.2-3B-Instruct												
Baseline	**14.9%**	0.860	0.842	0.840	**7.4%**	–	22.1%	0.851	0.837	0.832	**5.1%**	–
Self-Awareness Prompt	24.0%	0.908	0.883	0.881	19.7%	80.4%	**20.9%**	0.907	0.883	0.875	15.7%	85.0%
Chain-of-Thought	23.6%	0.909	0.883	0.881	19.9%	81.1%	22.6%	0.907	0.883	0.876	16.9%	85.9%
Cloze-Style Prompt	23.8%	0.908	0.883	0.881	19.8%	81.0%	22.6%	0.907	0.883	0.876	16.9%	85.9%
Cognitive Counterfactual Sim.	23.7%	0.908	0.883	0.881	20.2%	80.8%	22.5%	0.907	0.884	0.876	16.6%	84.8%
Self-Help Debias	23.0%	0.908	0.883	0.881	20.6%	81.3%	23.6%	**0.908**	**0.884**	**0.876**	15.9%	85.8%
Partial Summaries Ensemble	21.0%	0.886	0.870	0.869	14.8%	69.1%	27.6%	0.871	0.854	0.846	14.3%	78.2%
Weighted Summaries	20.7%	**0.921**	**0.903**	0.899	15.4%	75.6%	29.4%	0.876	0.862	0.852	9.8%	80.1%
Attention-Sort Re-ordering	19.8%	0.905	0.886	0.885	16.8%	71.4%	26.4%	0.859	0.841	0.830	14.6%	78.5%
Position-Invariant Shuffle	26.6%	0.903	0.890	0.889	10.8%	**67.1%**	30.8%	0.875	0.868	0.859	8.9%	**68.4%**
Forced Balanced Coverage	21.0%	0.905	0.884	0.883	17.5%	72.5%	26.2%	0.891	0.867	0.859	19.0%	83.0%
Weighted Token Decoding	18.9%	0.865	0.846	0.844	17.3%	71.0%	29.6%	0.855	0.841	0.834	14.4%	72.3%
Mirostat Decoding	20.0%	0.918	0.903	**0.901**	12.2%	70.0%	31.2%	0.901	0.882	0.875	11.7%	79.1%
Rejection Sampling	19.8%	0.904	0.885	0.883	16.7%	72.0%	23.8%	0.890	0.865	0.857	19.5%	82.7%
Self-Debias Decoding	22.9%	0.895	0.879	0.878	14.3%	68.9%	26.0%	0.879	0.854	0.847	17.6%	83.9%
Local-Explanation Guard	20.2%	0.865	0.844	0.842	17.9%	73.5%	26.5%	0.870	0.847	0.836	20.1%	81.8%
Llama-3-8B-Instruct												
Baseline	14.5%	0.855	0.837	0.836	**7.0%**	–	21.9%	0.847	0.834	0.828	**4.0%**	–
Self-Awareness Prompt	15.2%	0.909	0.881	0.878	22.9%	83.2%	**21.0%**	0.889	0.864	0.855	19.7%	85.5%
Chain-of-Thought	15.3%	0.909	0.881	0.878	22.6%	82.7%	22.5%	0.890	0.865	0.856	19.2%	85.6%
Cloze-Style Prompt	15.2%	0.909	0.881	0.878	21.9%	82.4%	21.9%	0.891	0.866	0.856	20.2%	85.3%
Cognitive Counterfactual Sim.	14.7%	0.909	0.881	0.878	22.5%	82.7%	21.8%	0.890	0.866	0.856	19.0%	84.8%
Self-Help Debias	15.2%	0.909	0.881	0.879	21.8%	83.0%	21.7%	0.891	0.865	0.856	19.2%	85.4%
Partial Summaries Ensemble	16.5%	0.890	0.871	0.870	15.4%	72.7%	28.5%	0.874	0.858	0.849	14.1%	79.3%
Weighted Summaries	15.1%	**0.925**	**0.904**	**0.902**	15.1%	77.0%	26.6%	0.880	0.866	0.855	9.0%	81.8%
Attention-Sort Re-ordering	16.9%	0.879	0.858	0.857	17.5%	74.7%	25.0%	0.869	0.845	0.831	20.7%	85.0%
Position-Invariant Shuffle	21.2%	0.899	0.888	0.887	11.5%	**64.4%**	27.0%	0.861	0.854	0.847	8.0%	**66.3%**
Forced Balanced Coverage	15.6%	0.900	0.878	0.877	19.5%	74.2%	23.8%	0.862	0.839	0.832	18.1%	81.3%
Weighted Token Decoding	**13.6%**	0.871	0.851	0.852	18.4%	70.0%	24.9%	0.852	0.831	0.823	18.1%	77.3%
Mirostat Decoding	13.8%	0.916	0.896	0.895	15.5%	75.8%	31.3%	**0.900**	**0.875**	**0.867**	18.3%	83.0%
Rejection Sampling	14.3%	0.911	0.890	0.889	18.9%	73.9%	26.8%	0.872	0.849	0.842	18.5%	80.0%
Self-Debias Decoding	16.0%	0.905	0.883	0.882	18.8%	75.1%	25.5%	0.879	0.855	0.848	18.2%	82.8%
Local-Explanation Guard	16.1%	0.859	0.837	0.836	18.9%	74.6%	25.9%	0.852	0.830	0.822	18.3%	79.1%
Phi-3-mini-4k-Instruct												
Baseline	34.5%	0.836	0.822	0.823	5.4%	–	26.2%	0.828	0.827	0.823	**4.9%**	–
Self-Awareness Prompt	25.5%	0.882	0.866	0.867	14.5%	69.5%	**24.9%**	0.850	0.832	0.826	16.6%	73.7%
Chain-of-Thought	23.9%	0.884	0.868	0.868	14.1%	69.7%	26.0%	0.851	0.832	0.827	15.7%	73.5%
Cloze-Style Prompt	24.7%	0.883	0.868	0.868	13.6%	67.7%	25.9%	0.850	0.832	0.826	16.4%	73.5%
Cognitive Counterfactual Sim.	22.8%	0.877	0.862	0.862	14.4%	67.7%	25.5%	0.850	0.832	0.826	16.2%	73.5%
Self-Help Debias	22.8%	0.879	0.863	0.864	14.3%	68.5%	25.1%	0.851	0.832	0.826	16.0%	75.5%
Partial Summaries Ensemble	28.0%	0.893	0.880	0.875	11.5%	71.3%	27.7%	0.858	0.845	0.836	11.4%	73.9%
Weighted Summaries	**22.6%**	**0.920**	**0.904**	**0.899**	11.2%	78.1%	27.8%	**0.880**	**0.871**	**0.863**	7.4%	72.1%
Attention-Sort Re-ordering	33.1%	0.887	0.877	0.876	9.9%	66.8%	28.1%	0.837	0.823	0.818	14.5%	67.9%
Position-Invariant Shuffle	28.2%	0.877	0.866	0.865	9.3%	67.6%	28.2%	0.837	0.834	0.825	7.7%	63.0%
Forced Balanced Coverage	31.5%	0.901	0.891	0.893	7.6%	61.5%	29.9%	0.844	0.831	0.830	11.8%	66.3%
Weighted Token Decoding	25.0%	0.860	0.849	0.854	9.8%	56.9%	28.9%	0.841	0.829	0.828	14.4%	63.0%
Mirostat Decoding	33.2%	0.891	0.883	0.884	**4.2%**	58.8%	34.3%	0.854	0.845	0.843	8.7%	61.3%
Rejection Sampling	22.7%	0.885	0.876	0.878	8.2%	60.5%	28.3%	0.853	0.836	0.835	14.6%	66.8%
Self-Debias Decoding	28.8%	0.894	0.885	0.888	7.1%	**56.4%**	31.3%	0.844	0.832	0.834	11.3%	**60.0%**
Local-Explanation Guard	**22.5%**	0.886	0.875	0.879	9.4%	59.7%	27.3%	0.841	0.828	0.826	12.4%	65.8%

Table 9: **Coverage, framing-change, and primacy-bias metrics** for Amazon Reviews and MediaSum News. Metrics: φ_{frame} (framing-change fraction, $\downarrow$), $\bar{s}_b, \bar{s}_m, \bar{s}_e$ (cosine similarity with the first/middle/last third of the source, $\uparrow$), ψ_{pri} (share of summaries whose similarity to the beginning exceeds that to the middle by $> 5\%$, $\downarrow$), and ρ_{pri} (share of summaries exhibiting primacy bias, $\downarrow$).

E Prompt Templates Used in Our Experiments

Table 10 lists the exact prompt instruction templates supplied to each model under all the experimental settings.

Prompt Strategy	Prompt
Summarization and General Bias Mitigation Prompts	
Self-Awareness Prompting	"You are an unbiased summarizer. Be mindful not to introduce any framing bias or omit the middle. Preserve the original sentiment. Please summarize the following text: `[DOCUMENT_TEXT]` FINAL_SUMMARY:"
Chain-of-Thought (Summarization)	"Please read the text below carefully. Then break down the text into beginning, middle, and end, describing each portion in detail. After that, produce a final summary. Use the following format: BEGIN_ANALYSIS: [describe the beginning] MIDDLE_ANALYSIS: [describe the middle] END_ANALYSIS: [describe the end] FINAL_SUMMARY: [your final concise summary] Text: `[DOCUMENT_TEXT]`"
Cloze-Style Prompt	"Fill the blanks for each part of the text: BEGIN: ____ MIDDLE: ____ END: ____ Text: `[DOCUMENT_TEXT]` When you fill them in, finally produce: FINAL_SUMMARY:"
Cognitive Counterfactual Simulation (Final Revision)	"Original text: `[DOCUMENT_TEXT]` Initial Draft: `[DRAFT_SUMMARY]` Here are the potential biased deviations: `[LIST_OF_SIMULATED_BIAS_DEVIATIONS]` Please produce a final summary that avoids these biases. FINAL_SUMMARY:"
Self-Help Debias (Rewrite Stage)	"Rewrite the following summary to remove any bias, preserve the original sentiment, and ensure equal coverage of beginning, middle, and end: `[DRAFT_SUMMARY]` FINAL_SUMMARY:"

Partial Summaries Ensemble (Merge Prompt)	"Combine the following partial summaries into one coherent summary:
	`[PARTIAL_SUMMARY_1]`
	`[PARTIAL_SUMMARY_2]`
	...
	FINAL_SUMMARY:"

Weighted Summaries (Chunk Prompt)	"Summarize this portion in about `[PORTION_TOKEN_BUDGET]` tokens: `[CHUNK_TEXT]` FINAL_SUMMARY:"
Attention-Sort Re-ordering (Final Prompt)	"CONTEXT: Segment 1: `[SORTED_SEGMENT_1_TEXT]` Segment 2: `[SORTED_SEGMENT_2_TEXT]` ... TASK: Summarize: FINAL_SUMMARY:"
Position-Invariant Shuffle (Prompt)	"The text is out of order; please summarize it fully: `[SHUFFLED_DOCUMENT_TEXT]` FINAL_SUMMARY:"
Hallucination Mitigation Prompts	
Baseline	"You are a fact-checking AI. For each statement, predict if the event has occured or not and respond with ONLY 'True' if it's factual or 'False' if it's not. Do not provide any explanation. Respond with TRUE or FALSE only and nothing else."

Prompt Calibration (CoT)	"You are a fact-checking AI. Follow these strict guidelines:
	- Respond with 'True' ONLY if:
	  * You are 100% certain the statement describes an event that has occurred
	  * You have clear knowledge of the event from your training data
	  * You can identify specific details confirming the event happened
	- Respond with 'False' if:
	  * You have ANY doubt about the statement's accuracy
	  * The statement contains ANY misleading or incorrect information
	  * You lack reliable information about the statement
	Respond with EXACTLY one word: 'True' or 'False'. No explanations."

Knowledge Boundary Awareness	"You are a fact-checking AI with knowledge up to `{knowledge_cutoff}`. For each statement, predict if the event has occurred or not. We check your decision-making abilities on futuristic events based on current data. Respond with only 'True' or 'False' for each news event. If the event is scheduled to occur after `{knowledge_cutoff}`, make a good prediction based on your knowledge up to `{knowledge_cutoff}`. RESPOND WITH EXACTLY ONE WORD: 'True' or 'False'."
Epistemic Tagging	"You are a fact-checking AI. Analyze this statement and respond with: - 'True [High Confidence]' or 'True [Low Confidence]' - 'False [High Confidence]' or 'False [Low Confidence]' Use these epistemic tags to guide your reasoning: [Certain] = Information you know with high confidence based on your training data [Uncertain] = Information you're less confident about or may be outside your knowledge [Reasoning] = Step by step analysis to determine veracity [Conclusion] = Your final determination with confidence level Respond ONLY with one of the four exactly formatted options listed above and nothing else."

Table 10: Prompt strategies and the prompts used in our experiments.