When AI Rewrites the News: How Sentiment, Framing, and LLM Disclosure Shape Perceptions

Prerana Khatiwada
Computer and Information Sciences
University of Delaware
Newark, Delaware, USA
preranak@udel.edu

Varun Pappu
Computer and Information Sciences
University of Delaware
Newark, Delaware, USA
varunp@udel.edu

Benjamin E. Bagozzi
Political Science & International Relations
University of Delaware
Newark, Delaware, USA
bagozzib@udel.edu

Matthew Louis Mauriello
Computer and Information Sciences
University of Delaware
Newark, Delaware, USA
mlm@udel.edu

Abstract

Public concern over media-driven polarization and the rise of AI-modified news has sparked interest in how sentiment and framing shape perceptions. This study examines variations in sentiment (neutral vs. extreme) and framing (balanced vs. one-sided) in LLM-transformed news, along with disclosure of LLM involvement, to assess effects on readers' emotions, perceptions, and credibility judgments. In a 2×2 between-subjects experiment (≈180 U.S. participants) plus a baseline control (45), articles were adapted from real news and transformed with LLMs. Results show extreme sentiment worsened outcomes, heightening negative emotions and lowering trustworthiness, while framing exerted more nuanced effects. Balanced news articles with extreme sentiment elicited amplified perceptions of bias and surprise consistent with the Hostile Media Effect, where balanced coverage appears biased due to amplified opposing viewpoints. Disclosure of LLM involvement modestly improved trustworthiness without undermining fairness or credibility. Overall findings highlight the need for transparent, user-facing interventions and editorial oversight in AI-mediated journalism.

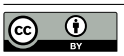
CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; *Empirical studies in HCI*; *HCI theory, concepts and models*; *User studies*; Empirical studies in collaborative and social computing; • **Computing methodologies** → Artificial intelligence;

Keywords

AI-modified news, Media framing, Sentiment, Trust, Media literacy, Human–AI interaction, Political communication

ACM Reference Format:
Prerana Khatiwada, Varun Pappu, Benjamin E. Bagozzi, and Matthew Louis Mauriello. 2026. When AI Rewrites the News: How Sentiment, Framing, and LLM Disclosure Shape Perceptions. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/3772318.3791527



CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791527

1 Introduction

Political polarization increasingly shapes how individuals perceive, interpret, and act upon information, creating significant challenges for news consumption and public discourse [11]. Today's fragmented, algorithmically curated environments often expose readers to content that confirms their existing beliefs, creating echo chambers that reinforce divides and erode trust in news media [22, 102, 104]. Experimental evidence further shows that conflict-oriented news frames, especially those high in intrusiveness or sensationalism, can cumulatively undermine political trust. However, the effect varies by audience traits such as extroversion and political efficacy [16]. Together, these dynamics highlight the importance of understanding framing mechanisms, primarily as Artificial Intelligence (AI) increasingly mediates not just the production of news, but also how it is framed and consumed.

These challenges call for moving beyond structural features of media environments and examining the discursive qualities of news itself [13, 86]. Two key discursive features, sentiment [10] (the emotional tone of language) and framing [28] (how information is selected and presented), play central roles in shaping readers' perceptions [16, 85]. While sentiment analysis typically categorizes content as positive, negative, or neutral [58] and reveals emotional engagement, political contexts require a deeper look at stance—the direction of emotional tone toward specific actors or issues [6, 59]. Meanwhile, framing shapes how audiences interpret facts, assign responsibility, and evaluate fairness [33, 79], adding a crucial dimension to understanding media effects. For example, a protest can have negative sentiment but a supportive stance, or neutral sentiment yet reveal partisanship. Prior work likewise shows that sentiment cannot reliably capture evaluative meaning or political position [14]. This shows why sentiment alone is insufficient; understanding how readers interpret content also requires attending to framing—how issues and perspectives are presented, because framing shapes perceived bias and fairness. Unlike studies that examine these features in isolation, our study manipulates both within the

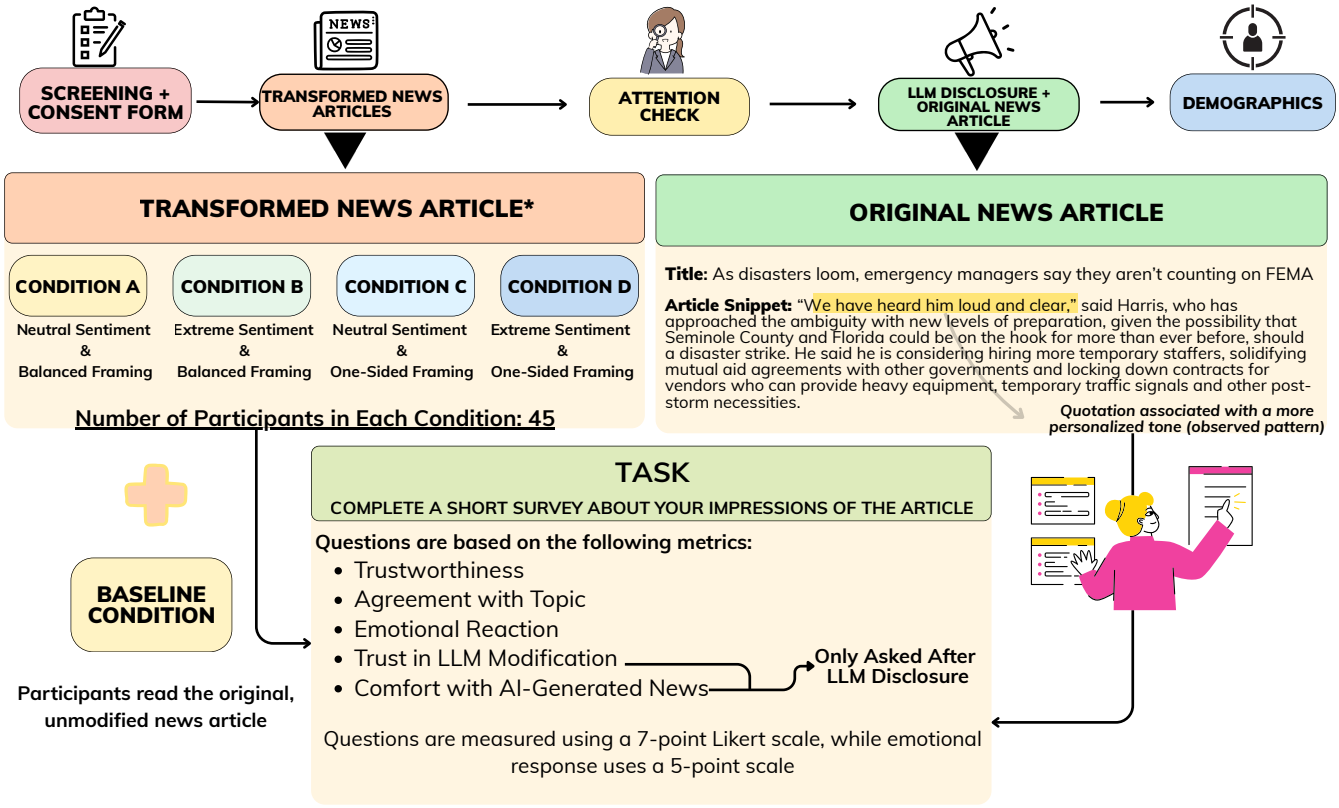


Figure 1: Overview of the main study design and survey flow. The figure illustrates the sequence of experimental components, including article presentation and participant exposure to different transformation conditions. As one example, omission of quotations in some transformed articles reflects the LLM's rewriting behavior rather than an experimental manipulation.

same articles to assess their interactive effects, linking longstanding concerns about polarization and selective exposure with the emerging challenges of AI-assisted journalism, where credibility, fairness, and transparency are increasingly at stake.

The emergence of LLMs (Large Language Models) adds a new layer of complexity to this challenge. Already, newsrooms and media platforms are experimenting with LLMs for tasks ranging from headline generation [29, 76] and article summarization to complete editorial drafting [103, 119]. While these tools can improve efficiency and enhance reader comprehension, they also introduce risks: fluent, persuasive text may obscure underlying bias, blur lines of authorship, or subtly shift tone and framing [48, 85]. Early research highlights both the potential and the pitfalls of LLM-mediated content—on the one hand, LLMs can increase engagement and accessibility; on the other hand, they may amplify polarization or erode trust in journalism if deployed without oversight [75]. Despite the growing adoption of AI-assisted writing, relatively little is known about how algorithmic editorial changes—such as altering sentiment or reframing arguments—affect audience perceptions of fairness, credibility, and trustworthiness. Likewise, the role of disclosure—across different types of content manipulation and varying degrees of extremity—remains underexplored: does acknowledging LLM involvement reassure readers through transparency, or does it instead increase skepticism about content integrity? This study addresses these gaps by experimentally varying sentiment, framing, and disclosure in AI-modified news articles to examine how these discursive features shape perceptions of bias, trust, engagement, and emotional response.

Specifically, we explore: ***(RQ1)*** *How do emotional tone (sentiment) and news framing influence readers' perceptions of bias, trust, and emotional engagement with news?* ***(RQ2)*** *When sentiment is held constant, what role does framing play in shaping perceptions of bias, trust, and balance? And* ***(RQ3)****, extending beyond textual features, how does disclosure that an article was modified by an LLM affect reader perceptions of bias, trust, and engagement?*

To examine these questions, we conducted an online experiment, 2×2 between-subjects with 180 U.S.-based participants recruited via Amazon Mechanical Turk (MTurk)[1]. We also collected data from an additional 45 participants who read the original, unmodified article, serving as a baseline comparison group. Participants read one of four versions of political news articles, systematically manipulated for sentiment (neutral or extreme) and framing (balanced or one-sided), and validated using both computational tools and

[1]https://www.mturk.com/

human reviewers. They then evaluated the articles for perceived bias, trustworthiness, agreement with content, and emotional response. Across conditions, sentiment produced minimal changes, while framing exerted a more substantial descriptive influence on perceived bias and imbalance. Insights from this study can inform adaptive tools that calibrate trust, flag one-sided framing, and introduce alternative perspectives without disrupting narrative coherence or reader agency. For example, if framing is more influential than sentiment, interventions can emphasize missing viewpoints rather than flag emotional tone.

This study makes three key contributions. (1) It provides empirical evidence on how sentiment and framing jointly shape perceptions of bias and trust in political news, extending prior work into the design-oriented concerns of HCI. (2) It introduces a methodological framework for generating and validating controlled sentiment–framing manipulations using multiple LLMs, offering a reusable approach for studying user perceptions of AI-mediated content. (3) It provides design implications for AI-assisted news systems, emphasizing transparency, framing awareness, and sentiment calibration as strategies to reduce perceived bias and support informed engagement.

By situating sentiment–framing interactions within the broader HCI context of AI-mediated information environments, this work advances understanding of how to design news systems that preserve trust and mitigate polarization in an era of increasing AI.

2 Background and Related Work

We review prior research across four interconnected strands. First, we outline theoretical foundations of framing and emotional tone in media effects research, highlighting classic conceptualizations and findings that demonstrate how these discursive features shape audience perceptions. Second, we review how sentiment and framing influence audience interpretation and engagement. Third, we review work on media bias, trust, and technological interventions in digital news environments. Lastly, we examine emerging research on AI- and LLM-mediated content generation and disclosure to situate our study within current debates on algorithmic transparency, credibility, and perception.

2.1 On Framing and Emotional Tone

At the core of media effects research is the idea that the choices surrounding what information is selected and emphasized profoundly shape how audiences interpret events, attribute responsibility, and form judgments about political and social matters [28, 33, 93, 106]. Political communication scholars in this area have increasingly leveraged cognitive and decision-making insights [107] to distinguish issue/emphasis frames (i.e., changes in which considerations matter) from equivalency/valence frames (i.e., presentations of the same facts in more positive vs. negative terms) in efforts to map when frames change opinions versus simply changing awareness [21, 30]. Related research has, in turn, suggested that emotional tone and sentiment in news and political speech operate partly independently of framing and often amplify framing effects [17, 97]. To this end, studies of emotional appeals show that affective cues (e.g., fear, anger, or enthusiasm) alter attention, information processing, and motivation in manners that can magnify or change a frame's directional impact [17, 64].

With regards to research into framing, and equivalency/valence effects more broadly, Hurtíková [44] emphasizes the above-mentioned valence dimension to framing in considering how the positive or negative presentation of issues can significantly influence political preferences. Authors found that this framing dynamic in television news coverage has a stronger attitude-shaping effect among viewers whose prior political attitudes matched the valence of that coverage. Psychological research also helps explain why negative frames, in particular, often produce more substantial evaluative shifts than positive frames, given that the former is commonly processed with heightened attention and perceived as more meaningful by individuals [61]. Political communication research further verifies that framing can alter politicians' policy preferences and evaluations, depending on the strength of the frame, the presence or absence of competing frames, and prior beliefs [20, 30]. And closely related insights into balanced versus one-sided frames further show that explicitly competitive or balanced presentations tend to attenuate single-frame effects. In contrast, sustained one-sided framing can shift issue salience and, at times, attitudes—especially when motivation to counter-argue is low [19, 30]. Classic theories of biased assimilation and attitude polarization further suggest that people interpret information through their prior beliefs (e.g., [62]). The Hostile Media Effect [43, 108] extends this logic: partisans not only assimilate congenial evidence but also perceive even neutral or balanced coverage as biased against their side [42, 108]. Building on research on emotional information processing, which shows that people appraise and respond to emotionally charged content through both deliberate evaluation and fast, bottom-up affective responses [9, 35, 64], we propose that extreme sentiment can heighten affective arousal and prompt rapid evaluative judgments. At the same time, more balanced framing may temper these reactions by signaling informational integrity.

As alluded to further above, research suggests that the emotional tone or sentiment of media also exhibits an important and at times interactive influence on how audiences construe issues [114]. To this end, political communication research has examined emotional tone or sentiment as a critical mechanism influencing audience response. Soroka and colleagues, for example, show that negativity bias in news is associated with stronger emotional reactions, particularly anger, compared to other affective states [95–97]. This body of work highlights sentiment not just as descriptive but as an active driver of political engagement. Experimental research also demonstrates how framing interacts with emotional tone to shape public opinion: early work finds that responsibility frames elicit distinct affective and attitudinal outcomes depending on whether anger or sympathy is evoked [70], while Nelson and Kinder [69] shows that issue frames interact with group-centered predispositions to shape attitudes toward policy. This is in line with Lecheler *et al.'s* research showing that discrete emotional cues in television news mediate associated framing effects [60], reaffirming the notion that emotional tone and framing are distinct but at times interactive dimensions of media and its varied effects on the public. Together, framing and emotional tone accordingly serve as dual discursive mechanisms that not only structure the presentation of information but also

shape how audiences process, evaluate, and respond to political content.

2.2 Audience Effects of Sentiment and Framing

Research shows how sentiment and framing shape audience interpretation, trust, and engagement with news. Khoo *et al.* [53] applied appraisal theory to analyze political news, capturing dimensions such as the appraiser's bias, the type of attitude, and subtle linguistic cues—elements often overlooked by conventional polarity-based methods. Expanding sentiment analysis to differentiate between discrete emotions, Soroka *et al.* [96] found that negativity in news is more strongly associated with anger than fear, underscoring the importance of domain-specific lexicons to improve construct validity. Puschmann and Powell (2018) [89] further contextualize sentiment analysis as a sociotechnical practice, arguing that while sentiment scores are presented as objective measures of public mood, they are constrained by both technical limitations and the cultural framing of computational tools in the media. At the level of audience effects, Kim *et al.* [54] show that anger- versus sadness-inducing news frames elicit distinct processing styles and attitudes, revealing the strategic impact of emotional tone in shaping public responses during crises.

Research also shows how framing and tone shape credibility and trust: Watimin *et al.* [112] demonstrated that economic and responsibility frames in social media posts significantly influenced user sentiment and reactions. Sparks and Hmielowski (2022) [98] found that ideological extremity in U.S. media correlates with simpler and more negative language. Similarly, Ali and Gill (2022) [4] found that Hurricane Harvey coverage varied in tone and framing depending on the actor involved, shaping perceptions of responsibility and trust. Finally, sentiment analysis has emerged as a critical component of misinformation detection, with Mannan *et al.* [63] emphasizing that understanding the emotional intent behind content can enhance the identification of fake news, particularly in highly contextual and socially charged environments.

Together, these computational and empirical studies illustrate that framing and sentiment are interdependent mechanisms shaping how audiences interpret and trust news content. Yet most prior work examines them in isolation or only descriptively, leaving open questions about responses to controlled manipulations in AI-mediated contexts. Our study addresses this by experimentally varying sentiment (neutral vs. extreme) and framing (balanced vs. one-sided) in LLM-assisted news, with and without disclosure, to test their effects on perceived bias, trust, and engagement. These findings offer empirical guidance for designing more transparent and trustworthy AI-mediated news systems.

2.3 Media Bias, Trust, and Tech Interventions

Perceptions of media bias and trust have become central to understanding news use in the digital era. Ardèvol *et al.* [7] demonstrate that while higher trust in social and citizen media predicts increased reliance on social media for news, perceived bias in traditional media correlates with decreased overall news use. Extending this perspective, Strömbäck et al. [101] further reviews how media trust is conceptualized across platforms and calls for more nuanced frameworks to capture its varied effects. On a global scale, Park et al. [83] find that increased use of social media for accessing news is linked to declining trust in news media generally, suggesting a reinforcing cycle of mistrust in the networked environment. Fisher [36] critiques the conceptual ambiguity surrounding "trust" in news media, questioning whether trust is even a desirable outcome in an era marked by uncertainty and contested information. Complementing these findings, Fletcher and Park [37] demonstrate that individuals with low trust in mainstream news gravitate toward non-mainstream sources—such as blogs, digital-born platforms, and engage more actively in online news participation, potentially as a means of seeking alternative perspectives and verifying information credibility.

Alongside these perception-centered accounts, recent HCI and technology-driven interventions have explored ways to make news bias more visible and interpretable to audiences. Spinde et al. [99] showed that explicit visualizations of annotated bias improved readers' comprehension of partisan slant, even if awareness did not always shift perceptions. Similarly, the Media Bias Detector [111] uses LLMs to surface tone, topical emphasis, and editorial choices in real time, enhancing users' awareness of how stories are framed. Comparative work finds that LLM inferences can approximate human-coded framing patterns [5], suggesting that automated cues may help readers recognize subtle narrative differences. More targeted systems, such as BIASsist [73], aim to explain and neutralize biased language, fostering critical engagement. Complementing these efforts, Bianchi et al. [15] proposed a classification-based approach to assess the trustworthiness of online news publishers, leveraging large datasets and external evaluations (e.g., NewsGuard) to provide automated trust scores at the source level. Collectively, prior work shows that bias and trust are not only intertwined social constructs but also emerging design challenges, with efforts ranging from bias visualization to LLM-powered corrections. Yet most studies treat bias descriptively or build mitigation tools without testing how audiences respond to deliberate content manipulations. Our study addresses this gap by examining how sentiment, framing, and disclosure of LLM involvement jointly shape perceptions of credibility, fairness, and trust in AI-mediated news.

2.4 AI and LLMs in Content Generation

Building on these gaps, it becomes essential to consider the growing role of artificial intelligence (AI) in news production. LLMs such as GPT-4 [78] and Gemini [39] can generate fluent, human-like articles, raising concerns about authenticity and editorial responsibility. While these systems offer efficiency and scale, readers often struggle to distinguish AI-generated content from human-written journalism, fueling concerns about trust, manipulation, and misinformation [90]. Detection tools like GLTR (Giant Language Model Test Room) [115] and OpenAI's classifier remain limited, especially for fine-tuned or lightly edited outputs, and perception studies indicate that although users are generally poor at identifying AI-generated text, they sometimes experience an "uncanny valley" effect when engaging with sophisticated LLM outputs, reflecting ambivalence between fluency and authenticity [90].

The ethics of disclosure, whether and how to inform readers that AI has partially or fully generated content, has become a growing concern in both media ethics and AI policy. Transparency about AI involvement can boost trustworthiness. However, it may also

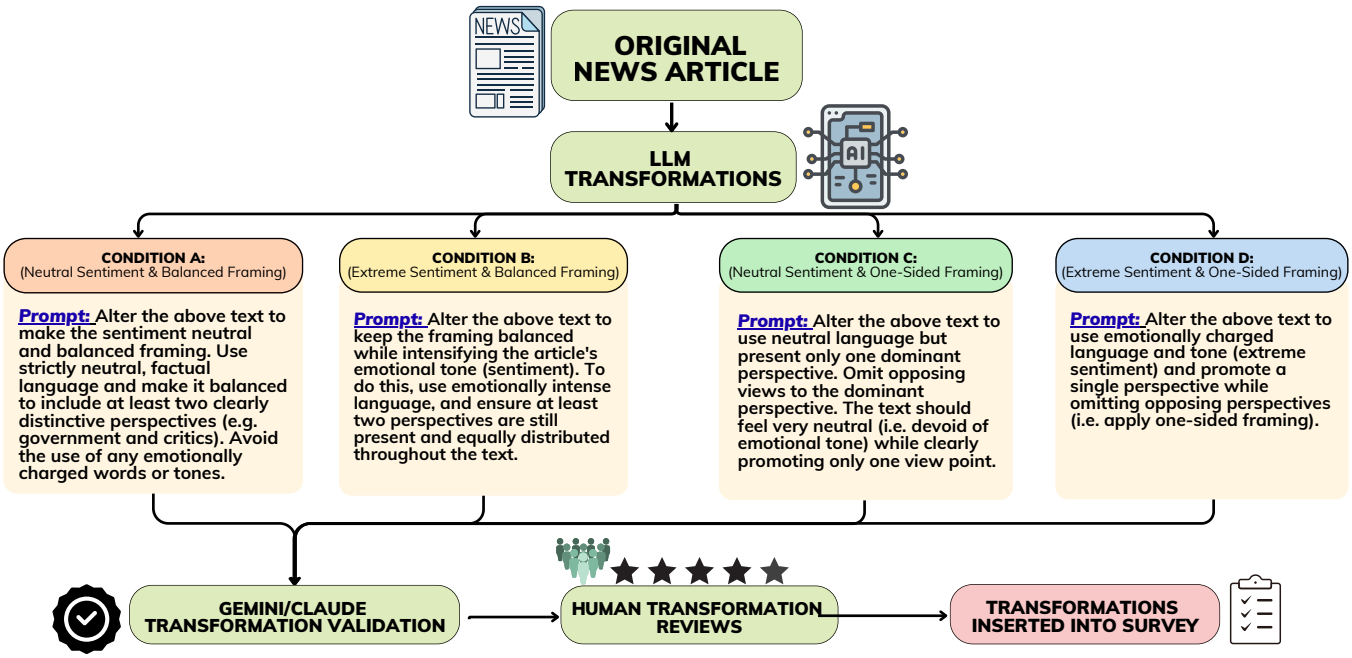


Figure 2: Overall transformation process across conditions. This figure illustrates the sequential stages and variations in transformation as influenced by different experimental or contextual conditions.

prompt skepticism, depending on users' attitudes toward automation. For example, Parshakov et al. [84] shows that users initially prefer LLM-generated content. Yet, disclosure significantly reduces this preference, highlighting the impact of provenance information. Similarly, Zhang and Gosline [121] find that human favoritism, not AI aversion, drives quality evaluations. When users know the same content is human-authored, its perceived quality increases. Disclosure of AI involvement does not reverse this evaluation gap. Other work reports mixed effects. Rossner et al. [92] find no credibility differences in sports journalism after disclosure. Huschens et al. [45] report that users often rate human- and AI-written content as comparably credible. Sometimes, AI-written content is seen as clearer and more engaging.

These mixed outcomes echo broader findings in deliberation research. Prior work on intrapersonal deliberation shows that reflection and knowledge access can act as "double-edged swords"—they can increase attitude certainty and shape willingness to express opinions, but sometimes in unanticipated directions [116]. In both cases—AI disclosure and reflective deliberation—there are parallels: the intervention is not neutral. Instead, it reframes how people interpret information, heightening sensitivity to bias or distortion while also risking unintended skepticism. This dynamic, therefore, mirrors broader debates about LLMs, whose outputs act less as neutral tools than as active interventions shaping perception and judgment.

Recent advances in LLMs have transformed content generation across news, education, and organizational contexts. While LLMs produce fluent, contextually appropriate text, challenges remain around factual accuracy and hallucinations [49, 117]. In the news domain, LLMs are increasingly used to automate multi-source news synthesis and article generation [71], support newsroom automation and authoring workflows [18], curate personalized feeds [25], and reshape consumption practices through speculative design [55], while writing systems such as Friction support reflective revision and iterative feedback [118]. Perceptions of LLM-generated content are shaped not only by quality but also by prior experience and usage patterns in workplaces [2], while in education, structured interventions such as the Educational LLM Framework (ELF)—which evaluates and refines AI-generated content for classroom use—have been shown to improve engagement and comprehension but still require careful validation [105]. Taken together, this body of work shows that disclosure and transparency around AI involvement can shape trust in complex ways, often contingent on context, prior attitudes, and domain of use, and even deceptive explanations can shift beliefs [27]. Yet little is known about how these dynamics unfold when LLMs transform politically charged news. Our study addresses this gap by testing how sentiment, framing, and disclosure jointly influence perceptions of bias, trust, and emotion, providing empirical evidence to inform the design of AI-mediated news systems.

3 Method

Here, we outline our recruitment process, study materials, research design, variables, and execution. The Institutional Review Board of the University of Delaware (IRB 1871618-8) approved this work, and the study complies with all relevant ethical regulations.

3.1 Stimulus Generation & Transformation

Our study materials consisted of news articles on three politically or socially contentious topics (e.g., immigration, climate change, voting laws), selected from credible, fact-based sources such as the New York Times, Reuters, and Associated Press (AP). Articles are chosen based on moderate length (300-650 words) and current political events that happened within the last three years.

Using a structured LLM prompt (see Figure 2 for the complete set of prompts used), we generate four editorially rewritten versions of each original news article using a panel of LLMs, specifically GPT and Grok. Each version reflects a distinct combination of Emotional Tone and Narrative Framing: (i) Neutral Tone with Balanced Framing, (ii) Extreme Tone with Balanced Framing, (iii) Neutral Tone with One-Sided Framing, and (iv) Extreme Tone with One-Sided Framing. The original news article is provided in Supplemental Note 3.

In our design, sentiment refers specifically to the affective valence and intensity of evaluative language, while framing refers to the distribution and visibility of distinct political perspectives within the article.

For sentiment, we instructed the models either to (a) use strictly neutral, descriptive, and policy-focused language (neutral tone) or (b) use highly evaluative, emotionally intense language (extreme tone) while preserving the underlying facts and topic. We did not specify the detailed linguistic strategies (e.g., quote usage, adjective selection, moral qualifiers) in the prompts; rather, these patterns emerged naturally in the transformed articles. Neutral tone outputs tend to avoid affective adjectives (e.g., "outrageous," "devastating," "foreign terrorists"), moralizing qualifiers (e.g., "irresponsible," "shameful"), and evocative metaphors, relying instead on factual, unembellished language and less frequent use of direct quotes. Extreme-tone outputs, by contrast, often retained or amplified direct quotes and incorporated strongly valenced adjectives, moral evaluations, and personalized descriptions of harm or benefit.

We guided the overall direction of sentiment using validated sentiment lexicons, but observed the specific linguistic choices as properties of the model's transformations rather than requirements we imposed. We then further manually verified that the adjustments outlined above were consistent with extant research. For example, variation in direct quote usage outlined above directly aligns with past experimental findings concerning the effects of such quotes from named individuals in news stories eliciting more emotional responses among readers [12, 87, 110].

For framing, we instructed the models either to (a) present multiple perspectives in a balanced way (balanced framing) or (b) foreground a single dominant viewpoint (one-sided framing). Balanced framing requires rewrites to include at least two clearly distinguishable perspectives (e.g., government officials vs. critics) and to articulate both arguments and counterarguments with comparable prominence and space. One-sided framing, in contrast, promotes a single perspective by either omitting opposing views entirely or relegating them to brief, subordinate mentions. Thus, framing is operationalized through the number, distinctness, and relative emphasis of viewpoints in the article rather than purely through word choice.

We iteratively refined our prompts through multiple pilot tests on held-out articles, with validation by two team members and a political science expert, who confirmed that the resulting texts exhibited clear, face-valid differences in tone and framing consistent with established political communication theory.

3.1.1 Multi-Model Validation for Robust Stimulus Generation. ChatGPT[2](powered by GPT-4.1, OpenAI) and Grok[3] (Grok 3, xAI) served as the primary interface for generating the transformed articles, given their widespread use and established capabilities in text generation. Neutral-tone transformations were generally produced with high fidelity; however, attempts to induce highly valenced sentiment frequently led to lexical instability, exaggerated adjectival intensity, or drift into tangential elaboration. To mitigate these limitations and reduce reliance on a single model's internal heuristics, we employed Gemini[4] (Gemini 2.5 Pro, Google) and Claude[5] (Claude Sonnet 4, Anthropic) as independent validators that assessed whether generated outputs conformed to the intended tone and framing specifications. All four systems were accessed through their web-based interfaces, though we also tested with their model APIs and observed comparable results.

Using multiple LLMs for validation is critical because models tend to recognize and systematically prefer their own generations when serving as evaluators, a phenomenon documented across GPT, Llama, and other model families [81, 94]. Such self-preference can distort assessments of sentiment, framing accuracy, and stylistic fidelity, particularly when the same model serves as both generator and evaluator. By validating GPT- and Grok-generated transformations with architecturally distinct models (Claude and Gemini), we separated the evaluatee from the evaluator, reducing the likelihood that model-idiosyncratic phrasing, bias cues, or stylistic fingerprints were mistaken for genuine manipulation effects. This design reflects a growing methodological consensus that reliable LLM evaluation requires cross-model corroboration rather than reliance on a single system; prior work routinely compares multiple models to establish consistency and limit evaluation artifacts [1, 3]. Moreover, research on LLM self-critique shows that models perform significantly worse when evaluating their own outputs than when assessed by independent systems or human experts [100, 109].

Taken together, the convergence across models here provides stronger evidence that transformations reflect the intended framing × sentiment manipulation rather than the idiosyncrasies of a single model's linguistic priors, thereby increasing the robustness and construct validity of our pipeline. All outputs were reviewed by the research team in consultation with political science and domain experts; articles that failed validation were regenerated with revised prompts and reevaluated jointly by two team members to confirm alignment with the targeted manipulation.

3.2 Manual Validation Rubric

After the model-based transformations, three human coders reviewed each article and its four rewritten versions to (i) assess the clarity and effectiveness of the transformations and (ii) select one article and a corresponding set of versions for the experiment. The

[2]https://chatgpt.com/
[3]https://grok.com/
[4]https://gemini.google.com/
[5]https://claude.ai/

Table 1: Rubric for Evaluating News Article Transformations

Metric	Description	Answer
Fluency	Output is grammatically correct, coherent, and natural-sounding.	Likert scale
Faithfulness to Content	Output preserves the core factual meaning of the original, aside from intended framing or tone shifts.	Likert scale
Tone Accuracy	Transformed text clearly and appropriately reflects the target tone (e.g., neutral or extreme).	Likert scale
Framing Clarity	Transformation effectively shifts the perspective or emphasis (e.g., positive vs. critical framing of the same fact).	Likert scale
Overt Bias or Misrepresentation	Transformation introduces unintended bias, exaggeration, or misleading framing.	Likert scale
Conciseness	Output appropriately concise without losing key information or rhetorical intent.	Likert scale
Overall Effectiveness	In your view, how well does the transformation achieve the desired communicative goal?	Open-ended

review was conducted by independent coders with expertise in political science and content analysis: (A) a full professor of political science, (B) a Ph.D. candidate, and (C) a visiting assistant professor.

Each coder independently evaluated all original and transformed articles. Special attention was paid to how generative transformations altered or preserved the characteristics, such as meaning, tone, and framing. All reviews are completed within the same week to ensure temporal consistency and avoid drift in interpretation. Coders are instructed to evaluate each article transformation along a set of qualitative criteria (see Table 1 for the rubric), to take breaks in between evaluations of each article transformation, and to re-read the original article before assessing each transformation. The article receiving the strongest agreement across reviewers and metrics was selected for the survey experiment that follows. Detailed results from the human reviewer evaluation are provided in Section 3.2.1.

After the human evaluation, we conducted a post hoc validation using automated text analysis. VADER[6] [46, 47] was used to assess sentiment alignment, and entropy[7] [66] was estimated as a proxy for framing diversity. Articles were segmented with NLTK[8] , and each sentence was classified as pro, neutral, or con. The stance entropy is calculated using the Shannon entropy formula: $H = -\sum_i p_i \log_2(p_i)$, where p_i is the proportion of article segments classified with stance i (i.e., pro, neutral, or con). The distribution of these stance labels was then used to compute Entropy, with higher values indicating more balanced framing. This process, adapted from prior stance analysis [56], quantified variation in viewpoint expression. Sentiment and stance-entropy results aligned with human evaluations: articles judged more emotionally charged had higher polarity scores, while those seen as balanced showed higher Entropy. For instance, balanced conditions (A: 1.40, B: 1.37) scored higher than one-sided ones (C: 1.30, D: 1.22). These parallels validate our automated measures and provide a robust framework for extending evaluations to dimensions like emotional salience, simplicity, or causal emphasis, while preserving a consistent evaluation pipeline.

3.2.1 Human-Reviewer Article Selection, Content Validity and Evaluation Results. As shown in Table 2, coders rated each transformation on six 7-point scales. Article 2 performed best, with top scores in fluency (6.75) and faithfulness (6.75). It also scored highly on tone accuracy (6.67) and conciseness (6.25), while maintaining relatively low ratings for overt bias or misrepresentation (2.00). By contrast, Article 1 displayed greater variability across reviewers and a higher overt bias rating (2.42, SD = 1.38), while Article 3, despite performing well on tone accuracy (6.83), was consistently judged weaker in framing depth and balance. Notably, Article 2's framing clarity scores showed more variance (5.50, SD = 1.39), but this reflected the intended experimental manipulation of framing rather than an execution flaw.

Table 2: Human evaluation scores across Articles 1–3 (averaged across Readers A–C). Metrics were selected to capture both editorial quality and substantive integrity of LLM-rewritten news articles.

Metric	Article 1		Article 2		Article 3	
	Avg	SD	Avg	SD	Avg	SD
Fluency	6.50	0.50	6.75	0.43	6.42	0.14
Faithfulness	6.42	0.38	6.75	0.00	6.33	0.14
Tone Accuracy	6.67	0.58	6.67	0.14	6.83	0.29
Framing Clarity	6.33	0.52	5.50	1.39	6.42	0.52
Overt Bias / Misrepresentation	2.42	1.38	2.00	1.15	1.58	0.29
Conciseness	5.50	1.09	6.25	0.25	6.08	0.38

The reviewer's open-ended feedback reinforced these quantitative results. All three readers agreed that the conditional transformations of Article 2 were fluent, coherent, and effective in achieving their intended tonal shifts. Inter-rater reliability was excellent for Article 2 (0.84), showing strong agreement among reviewers. Reviewers repeatedly described the Article 2 outputs as "natural" and "writerly," noting that both the neutral and emotional conditions were executed clearly without becoming exaggerated or implausible. One reader remarked that Article 2 *"reads smooth… as if a great writer wrote this piece,"* while another emphasized that it "*did a great job… very good relative to the assignment.*" Importantly, Article 2 produced the most coherent and realistic one-sided neutral narratives, which readers described as "*exactly as tasked,*" and its emotional transformations were consistently seen as "*sufficiently emotional but still balanced.*" Unlike the other articles, it maintained neutrality and emotional intensity without breakdowns in fluency, coherence, or factual clarity. In contrast, reviewers identified substantive issues in Articles 1 and 3. Article 1 raised recurring concerns about factual integrity, with multiple readers flagging

[6]https://pypi.org/project/vaderSentiment/

[7]Entropy is a measure from information theory that quantifies uncertainty or diversity [65]. Compared with simpler measures, entropy accounts for both the number of topics and their relative proportions, making it a preferred choice for assessing balance across content areas.

[8]https://www.nltk.org/

the omission of key details—such as the attacker's nationality—as "*misleading*" and producing "*illogical justifications.*"

Reviewers also criticized inconsistent lexical choices (e.g., "reprisal" vs. "retribution") that confused the intended framing. Article 3, while generally fluent, was repeatedly described as "*too short*," "*synthetic*," or *lacking depth*, particularly in its neutral transformations, where the exclusion of alternative perspectives reduced critical engagement and realism. Reviewers noted that its one-sided versions achieved the prompt but did so largely by removing content, making them "a tad less realistic."

Article 2's performance likely reflects its subject matter. Because it focused on Federal Emergency Management Agency (FEMA) restructuring and disaster response, it supported both emotional and neutral framings without becoming overly polarizing or factually unstable. This allowed the transformations to preserve readability, communicative intent, and framing clarity even when content was shortened. Thus, the quantitative scores (Table 2) and the qualitative feedback (Table 3)—underscored by consistent praise identified Article 2 as the most stable, balanced, and analytically defensible choice. We therefore selected it as the stimulus for the survey experiment.

Beyond identifying the strongest candidate, we deliberately used a single, carefully vetted article to maintain methodological rigor and experimental control. A central priority in our design was maximizing content validity, ensuring that the stimulus precisely represented the constructs we aimed to manipulate. While we initially developed a larger pool of manipulated articles, we undertook an extensive screening and pretesting process to identify the stimulus that most cleanly and consistently represented the constructs of interest—sentiment, framing, and AI-generated stylistic variation. Rather than presenting multiple articles that would have varied in topic, tone, or baseline bias, we invested substantial effort upfront, as outlined above, in systematically screening and evaluating all candidate articles.

Moreover, introducing multiple stimuli would have increased noise and introduced topic-level confounds unrelated to our framing and sentiment manipulations, ultimately weakening our ability to isolate the effects of the intervention. By relying on one rigorously validated article, we treated the stimulus as a controlled instrument rather than an uncontrolled source of variability, a deliberate methodological choice that strengthens construct alignment, improves internal validity, and ensures clearer interpretability of effects. Additionally, once an AI disclosure is presented, participants would inevitably interpret subsequent content through that lens, reducing ecological realism and blurring the specific effects of our manipulations. To avoid this contamination and preserve both construct fidelity and experimental precision, we selected a single, carefully validated stimulus.

3.3 Participation and Recruitment

We recruited 180 U.S. MTurk [9] participants in August 2025 for the main 2×2 design and 45 additional unique participants (Details in Section 3.4) who read the original (pre-transformation) article and completed the same outcome survey to provide a baseline for the four conditions (total $N = 225$). We conducted an *a priori* power analysis using G*Power based on 2*2 ANOVA. Using a medium effect size ($f = 0.25$), $\alpha = 0.05$, and a power of 0.80, the analysis indicated a required sample of about 180 participants (45 per condition) across four independent conditions.

We scheduled weekday, business-hours batches (e.g., 45 in Condition A) and used MTurk qualifications to block repeat participation. Running batches on weekdays during business hours helps ensure higher participant availability, faster response times, and more reliable study monitoring. Participants were compensated at a rate of $15 per hour, prorated by the minute (approximately $3.75 per survey response). Default MTurk qualifications included U.S. residency, age 18 or older, and English proficiency.

Two participants responded "no" to the screening question ("Are you 18 years of age or older?" and "Do you consume news online regularly?"), and five others left these questions unanswered. These participants were automatically filtered out by the survey system: answering "no" or leaving the item blank triggered the survey logic that redirected them to the end of the study, preventing them from accessing any subsequent questions. Participants who remained in the study also passed an attention check requiring them to select "Disagree" for a control item; all respondents answered correctly. See Figure 1 for the full study flow.

3.4 Research Execution

The study used a 2×2 between-subjects experimental design, manipulating two independent variables: sentiment (Neutral vs. Extreme) and framing (Balanced vs. One-sided). Upon accepting the task, each participant was first presented with a digital consent form to review and agree to before proceeding. Each participant was shown a news article that had been systematically modified to match the assigned condition.

All participants in the four experimental conditions completed the full survey experience, beginning with an eligibility screen and an initial set of perception measures (bias, trustworthiness, emotional response, and agreement with the content) after reading the transformed article without advanced disclosure[10] of LLM modification. We withheld AI disclosure during the initial evaluation to avoid introducing source- or authorship-based biases. Early disclosure that text is AI modified can shift perceptions of credibility and authenticity: users judge AI-labeled content differently [50, 68], and disclosure alone can alter trust and social evaluations in human–AI interactions [8], thereby compromising the validity of our primary outcome measures. In line with standard ethical guidelines for studies involving undisclosed stimulus manipulations [24], participants were fully debriefed at the end of the study. At that point, we explained that the articles had been modified using LLMs and described the purpose of the manipulation. Next, participants were shown the original version of the article, re-evaluated the same measures, and completed the reflective, scenario-based, and demographic questions. In contrast, the 45 participants in the baseline group read only the original unmodified article and then completed the reflection and demographic questions. This enabled comparisons with participants reviewing the manipulated versions.

[9] Amazon Mechanical Turk (MTurk) is a crowdsourcing marketplace where businesses and researchers (Requesters) can outsource small online tasks that computers cannot perform, and individuals (Workers) can complete these tasks for pay [82].

[10] Materials were transparently disclosed as AI-generated or edited, screened to avoid harmful content, and participants could discontinue at any time

Table 3: Comparative qualitative evaluation of Article 1–3 transformations, summarizing feedback from three independent reviewers.

Dimension	Article 1	Article 2 (Selected)	Article 3
Fluency & Faithfulness to Content	*Reviewer Observations:* More fluent than the original but sometimes awkward or synthetic, with the ending paragraph noted as unnatural. *Quote:* "The very last paragraph sounds like ChatGPT… parts of the text read too good and vague at the same time."	*Reviewer Observations:* Consistently smooth, clear, and human-like, flow improved upon the original. *Quote:* "It reads smooth in comparison to the original article… sounds quite natural as if a great writer wrote this piece… this one is great, it performed according to the prompt quite well."	*Reviewer Observations:* Mixed in quality: some versions were very fluent, while others were overly ornate, with emotional variants overusing sophisticated wording. *Quote:* "This one reads even better than [the] original article," but also "reads a bit artificial with these sophisticated emotional words."
Tone & Framing clarity	*Reviewer Observations:* Emotional shifts were often clear but sometimes confusing or overdone, with word choices occasionally muddling the intended tone. *Quote:* "There is confusing framing… using retribution vs reprisal," and at times "perhaps a little over the top."	*Reviewer Observations:* The emotional tone was stronger but still controlled, highlighting hardships without losing overall balance. *Quote:* "The shift in tone is very clear… sufficiently emotional but more balanced… overall I think it does a good job."	*Reviewer Observations:* The emotional tone was often seen as excessive, and the heavy emotional wording made the text harder to follow. *Quote:* "The emotive language is a bit much here… too overloaded with emotional words that hinder smooth reading… more focused on distinct emotional words than the content."
Bias Presentation & Conciseness	*Reviewer Observations:* Important facts were omitted, creating one-sidedness—for example, removing the bomber's nationality and related political backlash. *Quote:* "It has now obscured the fact that the bomber was Egyptian – not from a country banned… political backlash is completely missing… only presents one perspective by cutting out other content, a tad artificial."	*Reviewer Observations:* Some shortening occurred, but omissions were generally task-aligned and preserved the core meaning, even though some opposition or challenge details were lost. *Quote:* "Fairly dry and shortened, omitting some detail… removing that was needed for the transformation… it does not lose [the] key point because it focuses on just one side… created a one-sided neutral narrative as it was tasked to do."	*Reviewer Observations:* Content was largely preserved but often oversimplified or thinned, with reactions and the Trump–Musk split underplayed or omitted. *Quote:* "Some of the details of the reactions are lost… it sticks to the content but still omits and reframes some parts… oversimplifies it too much or omits some details."
Overall effectiveness	*Reviewer Observations:* Generally successful but weakened by factual gaps and confusing framing, requiring readers to infer missing logic around Egypt and the travel ban. *Quote:* "Overall this does a good job," yet "the onus is on the reader to catch the illogical justification… it obscured key facts."	*Reviewer Observations:* The most consistent and positively evaluated across readers, seen as effective and realistic in both neutral and one-sided forms. *Quote:* "This one is great… performed according to the prompt quite well… it did a great job; it reads smooth… sufficiently emotional but more balanced."	*Reviewer Observations:* Generally effective but more variable in quality, with some transformations praised while others were harder to follow. *Quote:* "A great transformation," but also "too overloaded with emotional words… hard to follow… I do not know if it sounds fully natural or not."

***Note.* While all three articles produced usable transformations, Article 2 received the most consistent praise for balancing fluency, factual clarity, and controlled tonal shifts, motivating its selection as the stimulus in our main study.**

3.4.1 Independent Variables. Our study manipulates two factors that influence news perception: sentiment and framing. Sentiment is manipulated as Neutral (factual, emotionally neutral) or Extreme (strongly emotional, either positive or negative). Framing is manipulated as Balanced (presenting multiple perspectives) or One-sided (emphasizing a single perspective).

3.4.2 Dependent Variables. To assess the impact of these manipulations, we measured four dependent variables: 1. Perceptions of Bias: the extent to which the article seemed partisan or one-sided (e.g., "How biased did this article seem?") 2. Trustworthiness: perceived credibility and reliability (e.g., "How trustworthy did you find this article?") 3. Emotional Response: ratings of anger, happiness, anxiety, and surprise on a 5-point Likert scale, and 4. Agreement with Content: whether arguments appeared to favor one side more than the other. The three evaluative constructs (bias, trustworthiness, and argument imbalance) were each measured using single-item 7-point Likert-style questions with a consistent response format—complete questionnaire in Supplemental Note 2.

3.4.3 Post-Exposure Evaluation. After the main tasks, participants answered post-disclosure questions about whether LLM modification altered meaning, introduced bias, omitted or exaggerated content, or shifted partisanship, tone, or balance. Open-ended items further probed attitudes toward AI in journalism—whether it should replace or support human journalists—and how disclosure shaped trust. This approach builds on prior CHI work using post-exposure questions to examine how perceptions shift after reflective or disclosure based interventions (e.g., [120]). In our case, disclosure of AI involvement served as a reflective stimulus, prompting participants to reconsider judgments of credibility, neutrality, and informational quality. Following Zhang *et al.* [120], who showed that reflection can increase attitude certainty and willingness to express opinions, our design captured how awareness of AI authorship reshaped participants' evaluative and behavioral intentions.

This procedure is also inspired by D. Molina *et al.* [26], who, in their clickbait study, first showed participants a headline and allowed a "Read More" click, thereby ensuring access to the full content to prevent disengagement and confusion—anchoring judgments in the actual article content. By combining reflective probing

post-disclosure with full-content anchoring, our method provides a transparent basis for evaluation and supports more grounded and interpretable assessments.

4 Results

Before presenting findings, we briefly summarize the analytic approach. Closed-ended responses were descriptively and statistically analyzed, with standard assumptions checked (Shapiro–Wilk for normality [38], Levene for homoscedasticity [77]) prior to conducting two-way ANOVAs (Sentiment × Framing) and Bonferroni-adjusted post hoc tests [113].

Open-ended survey responses were analyzed using a light thematic analysis appropriate for brief, single-prompt comments. Because the dataset consisted of short, focused responses, we did not compute formal reliability statistics, consistent with recommendations in applied thematic analysis that interpretive or small datasets, where consensus is readily attainable, are more appropriately coded through collaborative agreement rather than statistical coefficients [41]. To ensure rigor, we followed standard double-coding[11] practice: two coders independently coded all responses, then compared their codes and discussed any disagreements until they reached full agreement. Representative quotes in Sections 4.5.2 and 4.5.3 illustrate key themes and complement the quantitative findings.

Below, we report participant demographics across all conditions, followed by findings from the closed- and open-ended survey questions, with relevant statistical tests. The survey took participants an average of 14.6 minutes to complete.

4.1 Participant Demographics

A total of 225 valid participants were included in the analyses after applying eligibility criteria. These demographic figures reflect the full sample (not broken down by condition): The average age was 33.36 years (SD = 6.78), with participants ranging from 23 to 72 years. The sample skewed slightly male (67.1% male, 32.4% female, 0.4% prefer not to say). In terms of education, the majority held a Bachelor's degree (72.4%), followed by a Master's degree (25.3%), with only a few participants reporting a maximum educational attainment level of a high school diploma (0.9%), a Doctoral/professional degree (0.4%), or less than high school (0.4%). Participants were distributed across political party affiliation, with 51.6% identifying as Democrat, 39.6% as Republican, and 8.4% as Independent—tracking closely with 2025 U.S. population estimates, though with a slight Democratic skew relative to national benchmarks [88]. Participants also showed ample support across the full ideological spectrum, with the largest shares identifying as Extremely Conservative (28.9%) or Moderately Liberal (28.4%), followed by Extremely Liberal (20.0%), Moderately Conservative (19.1%), and Centrists (3.1%). Patterns of political news consumption indicated that most participants were highly engaged: 53.3% reported consuming news daily, 28.0% multiple times per week, 16.0% once per week, and only 2.7% once per month.

Together, this profile reflects a relatively well-educated, politically engaged participant pool with balanced representation across major U.S. political affiliations and ideological orientations. See Supplemental Note 1 for the full demographic breakdown.

[11] Independent double-coding followed by adjudication is well suited to small qualitative samples where consensus and coder agreement are more informative than formal reliability coefficients; in such cases, double-coding primarily serves to improve rigor rather than to produce a statistical reliability metric [80].

4.2 Bias, Trustworthiness, and Balance

Here, we examine how participants' perceptions differ when reading the original news article, LLM-modified news articles before disclosure, and again after disclosure that an LLM had modified the news article content.

We selected three related aspects of news evaluation—bias, trustworthiness, and argument imbalance because they capture complementary dimensions of how readers assess news content [34, 51]. All were measured on the same 1–7 Likert scale (see Section 3.4.2). We first conducted one-way ANOVAs within each construct to test condition effects. We then conducted a two-way mixed-effects ANOVA with construct as a repeated-measures factor and condition as a between-subjects factor to assess whether participants distinguished among evaluative dimensions and whether any construct-by-condition interaction emerged. Although the measures share a common response format, the constructs reflect distinct dimensions; accordingly, this analysis examines differentiation patterns rather than treating the scales as strictly commensurate.

4.2.1 ***Original → Pre-Disclosure.*** Descriptively, participants' ratings were highly stable across baseline (original) and altered articles. Perceived bias ranged from 4.8 (Neutral–Balanced) to 5.36 (Extreme–Balanced), with the baseline article at 5.0. Trustworthiness scores were consistently high, ranging from 5.20 to 5.44, with the baseline (5.38) nearly identical to the altered versions. Argument imbalance also clustered narrowly around 5.27–5.53, again with little departure from the baseline (5.33). In short, sentiment (neutral vs. extreme) and framing (balanced vs. one-sided) produced only small descriptive fluctuations.

A mixed-model ANOVA showed a significant main effect of judgment type (bias, trust, imbalance), $F(2, 440) = 9.31, p < .001$, with a small effect size ($\eta^2 = .01$). Although the overall effect was statistically significant, the magnitude indicates only modest differences between the constructs. Post hoc comparisons reflected this pattern: responses on the bias item differed slightly from responses on the trustworthiness item ($M_{\text{diff}} = -0.28, p = .003$, Cohen's $d \approx 0.11$) and the argument imbalance item ($M_{\text{diff}} = -0.32, p < .001$, $d \approx 0.13$), with both effects being small in magnitude.

Taken together, these results suggest that participants distinguished among the three constructs, but the differences were small: articles were seen as similarly trustworthy and balanced, with perceived bias only marginally lower on average. Crucially, subtle manipulations of sentiment and framing before disclosure did not meaningfully shift these explicit perceptions (see Figure 3).

4.2.2 ***Pre-Disclosure → Post-Disclosure.*** Descriptively, ratings remained clustered around the midpoint to high end of the 7-point scale (overall $M = 5.21$). Among pre-disclosure articles, Extreme Sentiment & Balanced received the highest overall ratings ($M = 5.44$), while Neutral Sentiment & Balanced was lowest ($M = 5.11$). After disclosure, values shifted only slightly: Neutral–Balanced rose to 5.23, while Extreme–One-Sided dropped to 5.07. Notably, trustworthiness showed the clearest descriptive change, dipping

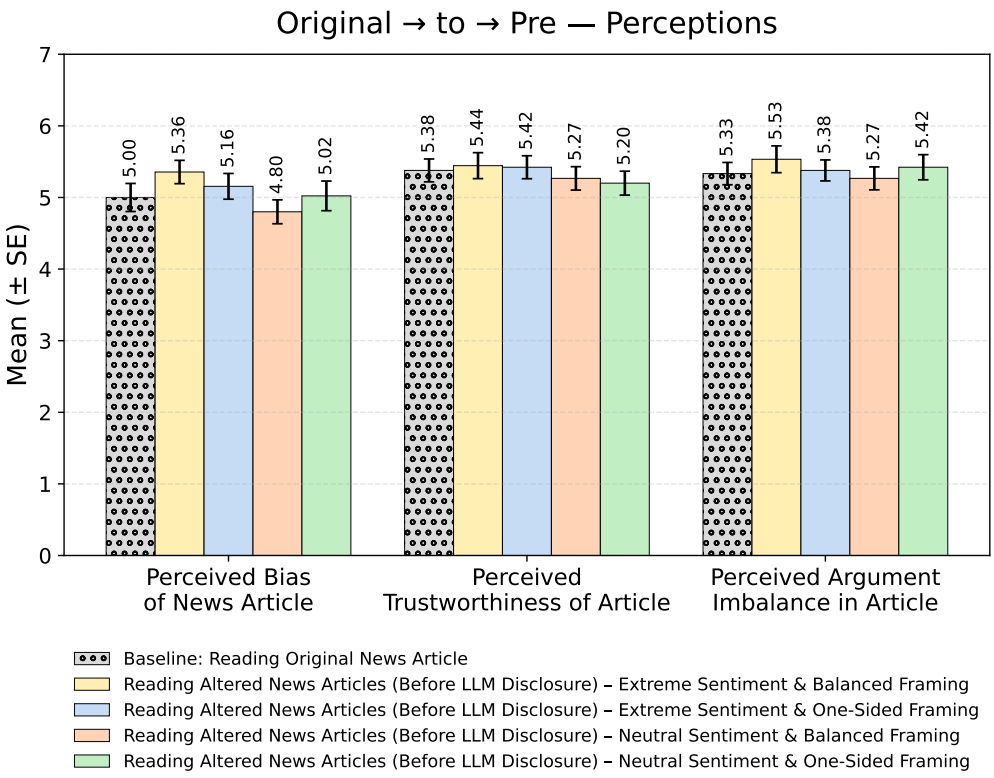


Figure 3: Comparison of baseline (original articles) and altered articles (before LLM disclosure) on perceived bias, trustworthiness, and argument imbalance. Error bars show ± SE. Although participants distinguished between bias, trust, and imbalance overall, conditions (neutral vs. extreme; balanced vs. one-sided) did not significantly affect ratings.

in the post-disclosure evaluations for some one-sided and extreme conditions (e.g., $M = 4.87$).

A mixed-model ANOVA showed that participants distinguished between bias, trust, and imbalance ($F(2, 704) = 6.38$, $p = .002$), although the effect size was small ($\eta^2 = .01$). Disclosure, however, had no reliable impact across conditions, with neither a main effect nor an interaction.

Post hoc comparisons indicated small differences between responses to the bias item and the argument imbalance item ($M_{\text{diff}} = -0.25$, $p < .001$, Cohen's $d \approx 0.09$). Responses to the bias and trustworthiness items and to the trustworthiness and argument imbalance items did not differ reliably. Descriptively, minor trends were observed (e.g., a slight post-disclosure decrease in trust and marginally higher bias in extreme-sentiment variants), but these were not statistically reliable. Overall, participants' evaluations of credibility, bias, and argument balance were highly stable and did not meaningfully change after disclosure (see Figure 4).

To unpack these patterns further, we broke down the results for trustworthiness and bias across the four article conditions and compared before vs. after disclosure (see Figure 5).

Although articles with neutral and balanced framing showed slightly higher mean trustworthiness ratings than the other conditions, the differences across conditions were not statistically significant. By contrast, timing showed a small but reliable effect: trust ratings were slightly higher before disclosure and dropped afterward (overall $\Delta \approx 0.26$ points). An independent-samples t-test confirmed that this drop was statistically significant, $p = .042$, but the effect was small (Cohen's $d = 0.21$). In other words, disclosure made people less trusting, regardless of article style.

Articles with extreme sentiment were generally seen as more biased than neutral ones, especially when paired with balanced framing. In contrast, one-sided vs. balanced framing made little difference, and disclosure had no clear impact on bias ratings. Statistically, there was a modest effect of condition on perceived bias ($p = .046$, $\eta^2 = .02$), and no effect of disclosure. Post-hoc comparisons reinforced this uncertainty: after correcting for multiple comparisons, none of the pairwise differences was significant (all Bonferroni-adjusted $p > .05$). The only descriptive trend—slightly higher bias ratings for extreme-sentiment articles with balanced framing—did not survive correction ($p \approx .054$–$.056$).

Overall, bias perceptions were shaped more by tone than by disclosure. While disclosure slightly reduced trust, it did not meaningfully shift bias judgments, highlighting the limited impact of current transparency strategies.

4.2.3 ***Original → Post-Disclosure.*** When comparing the original articles to the AI-modified ones after disclosure, perceptions stayed very close overall, with means around 5 on the 7-point scale.

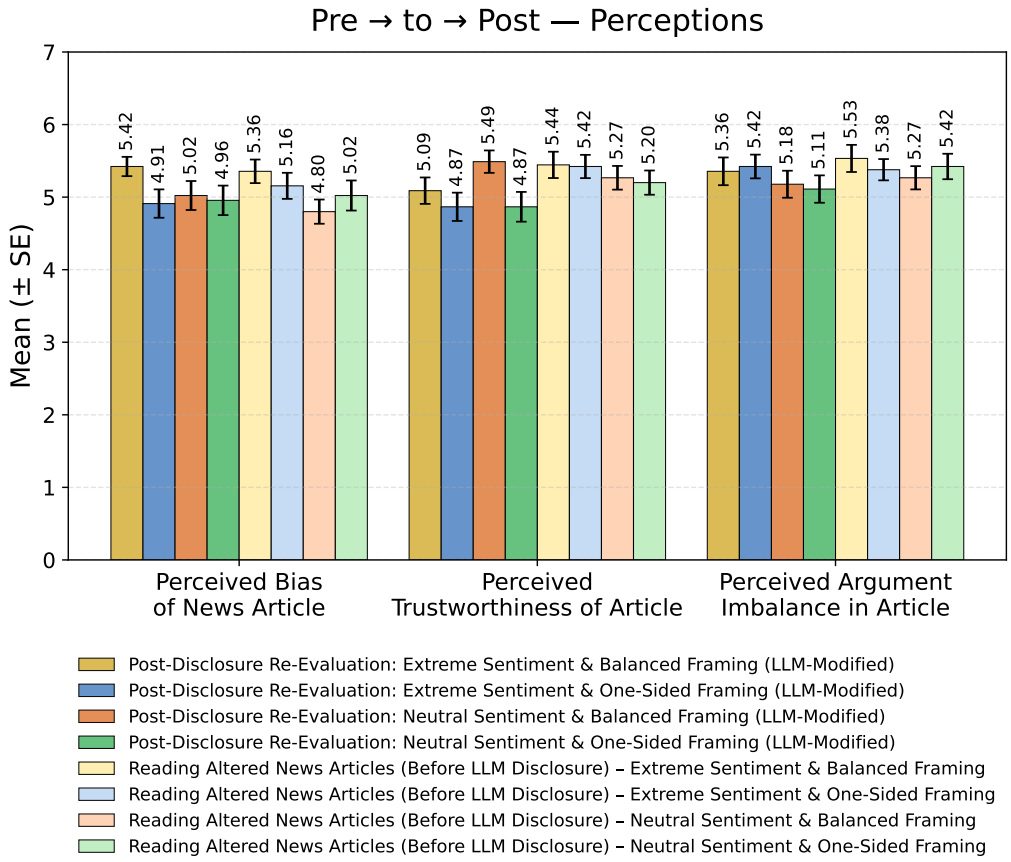


Figure 4: Comparison of pre-disclosure evaluations and post-disclosure re-evaluations on perceived bias, trustworthiness, and argument imbalance. Error bars show ± SE. Disclosure did not significantly alter ratings, though descriptively trust appeared to dip slightly after disclosure.

For **bias**, the baseline article was rated at 5.0, while post-disclosure articles ranged from about 4.9 (Neutral + One-Sided, Extreme + One-Sided) up to $M = 5.42$ ($\Delta = +0.42$) (Extreme + Balanced), suggesting that extreme sentiment made articles appear slightly more biased. For **trustworthiness**, the baseline was 5.38, with Neutral + Balanced showing the highest post-disclosure score ($M = 5.49$, $\Delta = +0.11$) and both extreme/one-sided conditions dipping to the lowest ($M = 4.87$, $\Delta = -0.51$), indicating that disclosure may have reduced trust when articles were already extreme or one-sided. For **argument imbalance**, ratings were stable across conditions (≈ 5.1–5.4), nearly identical to the baseline (5.33).

Statistically, there was a significant interaction with judgment type, meaning disclosure affected some perceptions (e.g., trust) more than others. Post hoc tests confirmed that bias was rated lower than imbalance ($p = .024$).

Taken together, these results suggest perceptions remained largely unchanged from original to post-disclosure, with the main exception that trustworthiness declined selectively in articles that were already extreme or one-sided (see Figure 6). Thus, readers do not overhaul their views when told an article was AI-modified, but disclosure appears most likely to erode trust in articles that already look extreme or one-sided.

4.3 Emotional Reactions: Original vs. Altered

We compare participants' emotional reactions when reading the original news article and altered articles without LLM disclosure. Participants then re-evaluated those articles after disclosure that an LLM had modified the news. This provides insight into whether LLM manipulations or disclosure selectively amplify negative or positive emotional responses.

4.3.1 ***Original → Pre-Disclosure.*** Participants reported moderate emotional reactions overall, near the midpoint of the 5-point scale ($M = 3.25$). Relative to baseline ($M = 3.08$), altered articles—particularly those with extreme sentiment elicited stronger negative emotions. Disgust increased from $M = 2.62$ to $M = 3.36$ ($\Delta = +0.74$), anxiety rose from $M = 2.84$ to $M = 3.53$ ($\Delta = +0.69$), and anger showed a smaller but consistent increase from $M = 2.93$ to $M = 3.51$ ($\Delta = +0.58$). In contrast, happiness remained stable across conditions, while surprise declined under extreme sentiment (from $M = 3.73$ to $M = 3.02$, $\Delta = -0.71$).

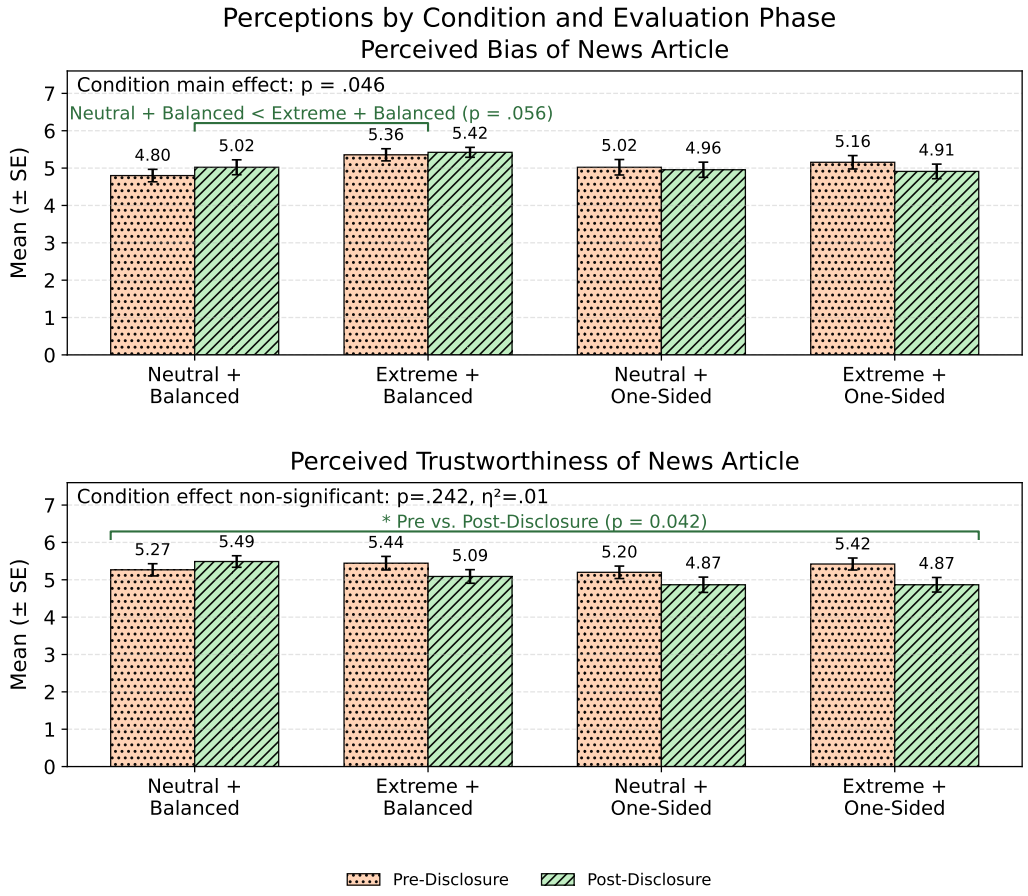


Figure 5: Trustworthiness and bias are shown by condition (Neutral/Extreme × Balanced/One-Sided) and timing (pre vs. post disclosure); error bars represent ±SE. For trustworthiness, the overall timing effect (pre > post) was statistically significant when averaged across all conditions. The significance bracket for trustworthiness reflects the overall main effect of Timing (Pre vs. Post disclosure) across all conditions. Simple pre–post contrasts within individual conditions were also tested, but none reached significance. Bias shows the expected descriptive pattern of higher ratings for extreme-sentiment articles, but no significant timing effect was observed.

Statistically, one-way ANOVAs indicated significant but modest effects of condition for disgust ($F = 3.23$, $p = .013$, $\eta^2 = .06$), anxiety ($F = 2.55$, $p = .040$, $\eta^2 = .04$), and surprise ($F = 3.43$, $p = .010$, $\eta^2 = .06$). Post hoc Bonferroni comparisons showed that baseline articles were rated significantly lower in disgust than both extreme-sentiment conditions, lower in anxiety than the Extreme + One-Sided condition, and higher in surprise than the Extreme + One-Sided condition.

Taken together, extreme sentiment amplified negative emotions most strongly, particularly disgust and anxiety, and reduced surprise in one-sided contexts. Extreme Sentiment & Balanced Framing tended to produce slightly higher anger, disgust, resentment, and anxiety compared to baseline, while Extreme Sentiment & One-Sided Framing also heightened resentment and anxiety but showed the lowest happiness ratings ($M \approx 2.9$). Neutral manipulations were closer to baseline. Negative emotions shifted more noticeably across conditions, whereas positive emotions such as happiness and surprise remained comparatively stable (see Figure 7).

4.3.2 ***Pre-Disclosure → Post-Disclosure.*** Participants' emotional responses stayed broadly similar before and after disclosure, but a few shifts stand out. Negative emotions (anger, disgust, resentment, anxiety) were largely stable, though disgust rose slightly in the Extreme + Balanced condition from $M = 3.36$ pre to $M = 3.47$ post ($\Delta = +0.11$). The largest descriptive movement was seen in surprise, which remained high for Extreme + Balanced ($M = 3.82$ pre, $M = 3.71$ post). Still, it was much lower for Extreme + One-Sided ($M = 3.02$ pre, $M = 3.04$ post), producing a spread of about $\Delta \approx 0.7$ between the two article types. Happiness also showed a slight post-disclosure increase in Extreme + Balanced (from $M = 3.58$ to $M = 3.56$), while Neutral + One-Sided dipped from $M = 3.47$ pre to $M = 3.27$ post ($\Delta = -0.20$).

Statistically, one-way ANOVAs indicated significant but modest effects of condition for disgust ($F = 2.34$, $p = .024$, $\eta^2 = .04$), surprise ($F = 3.16$, $p = .003$, $\eta^2 = .06$), and happiness ($F = 2.27$, $p = .029$, $\eta^2 = .04$). However, Bonferroni post hoc comparisons showed

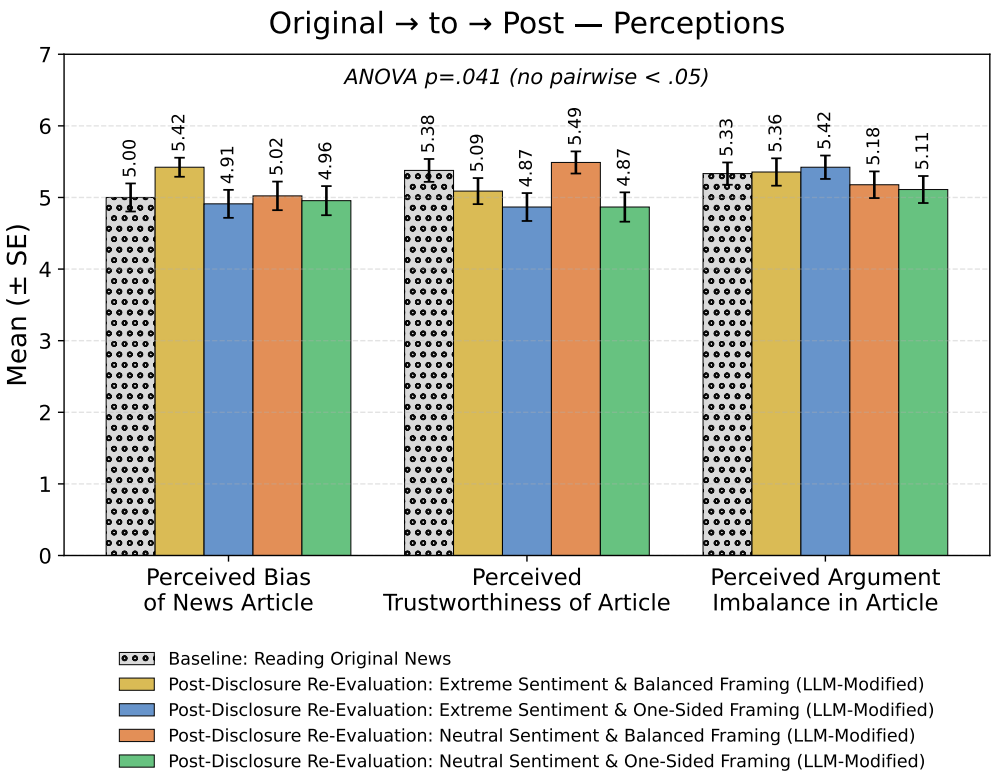


Figure 6: Comparison of baseline (original articles) and post-disclosure re-evaluations (LLM-modified articles) on perceived bias, trustworthiness, and argument imbalance. Error bars show ± SE. Ratings were largely stable, though disclosure selectively reduced trustworthiness in some extreme and one-sided conditions.

that most pairwise differences were not significant. The clearest pattern appeared for surprise: participants rated Extreme + Balanced articles as more surprising than Extreme + One-Sided, both before disclosure and relative to the post-disclosure re-evaluations of the Extreme + One-Sided condition.

Overall, disclosure itself did not meaningfully alter emotional intensity. Instead, the content style (e.g., balanced vs. one-sided; neutral vs. extreme) appeared to shape which emotions were evoked. These patterns should be interpreted cautiously given the small effect sizes and limited post-hoc significance (see Figure 8).

4.3.3 ***Original → Post-Disclosure.*** Compared to the original baseline, emotional responses shifted upward for several negative emotions when participants re-evaluated articles after disclosure. The largest increases were in disgust ($M = 2.62$ baseline → $M = 3.47$ Extreme + Balanced, $\Delta = +0.85$) and anxiety ($M = 2.84$ baseline → $M = 3.60$ Extreme + Balanced, $\Delta = +0.76$). Anger also rose modestly ($M = 2.93$ baseline → $M = 3.36$ Extreme + Balanced, $\Delta = +0.43$). By contrast, surprise remained high across all conditions, with the baseline at $M = 3.73$ and post-disclosure scores ranging from $M = 3.04$ to $M = 3.71$, showing that disclosure did not systematically dampen surprise but widened the spread between conditions. Happiness stayed fairly stable, though Extreme + One-Sided dropped the lowest ($M = 2.87$).

We observed significant but modest effects of condition for disgust ($F = 3.88$, $p = .005$, $\eta^2 = .07$), anxiety ($F = 2.43$, $p = .048$, $\eta^2 = .04$), and surprise ($F = 2.97$, $p = .020$, $\eta^2 = .05$). Consistent with these modest effects, post hoc tests identified only a subset of significant pairwise differences. For disgust, Extreme + Balanced articles elicited higher ratings than the baseline and were also higher than Neutral + One-Sided. For anxiety, Extreme + Balanced was higher than the baseline. For surprise, the baseline was higher than Extreme + One-Sided, and Extreme + Balanced was higher than Extreme + One-Sided.

Disclosure itself did not create broad emotional shifts, but it accentuated differences between content styles: extreme sentiment paired with balance framing evoked the strongest negative reactions, while extreme one-sided framing dulled positive affect (see Figure 9 for a detailed breakdown of emotional responses).

4.4 Perceived Article Modification (Post-Survey)

After reading, participants completed a post-survey assessing *perceived changes* to the article, including whether modifications distorted or omitted information, introduced bias, altered the message, showed partisanship, used manipulative language, exaggerated issues, sought to influence opinions, or raised ethical concerns.

4.4.1 ***Meaning and Tone Across Conditions.*** Descriptive results indicated that across all conditions, participants rated the

Figure 7: Comparison of baseline (original articles) and altered articles (before LLM disclosure) on emotional reactions (angry, disgusted, resentful, anxious, surprised, happy). Error bars show ± SE. Participants reported similar emotional responses across conditions, with extreme sentiment producing slightly stronger negative emotions.

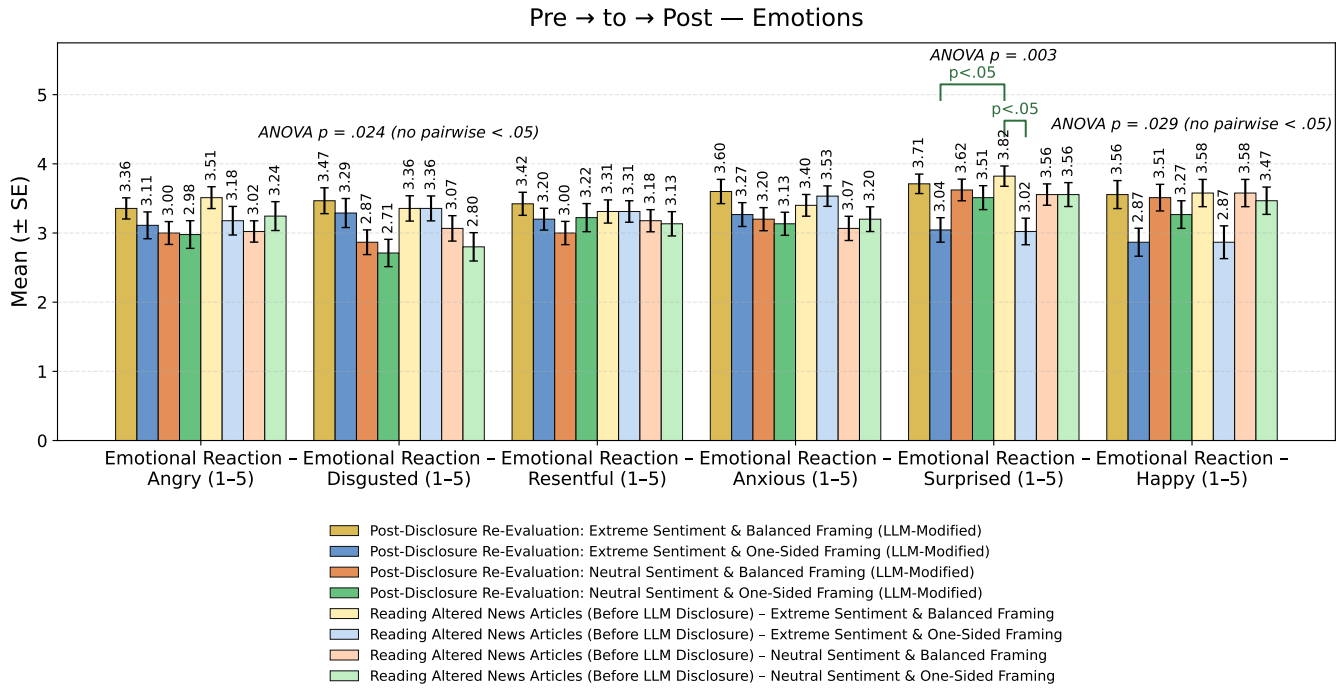


Figure 8: Comparison of pre-disclosure evaluations and post-disclosure re-evaluations on emotional reactions. Error bars show ± SE. Disclosure had little overall effect on emotions, though slight declines in positive emotions (e.g., happiness) were observed.

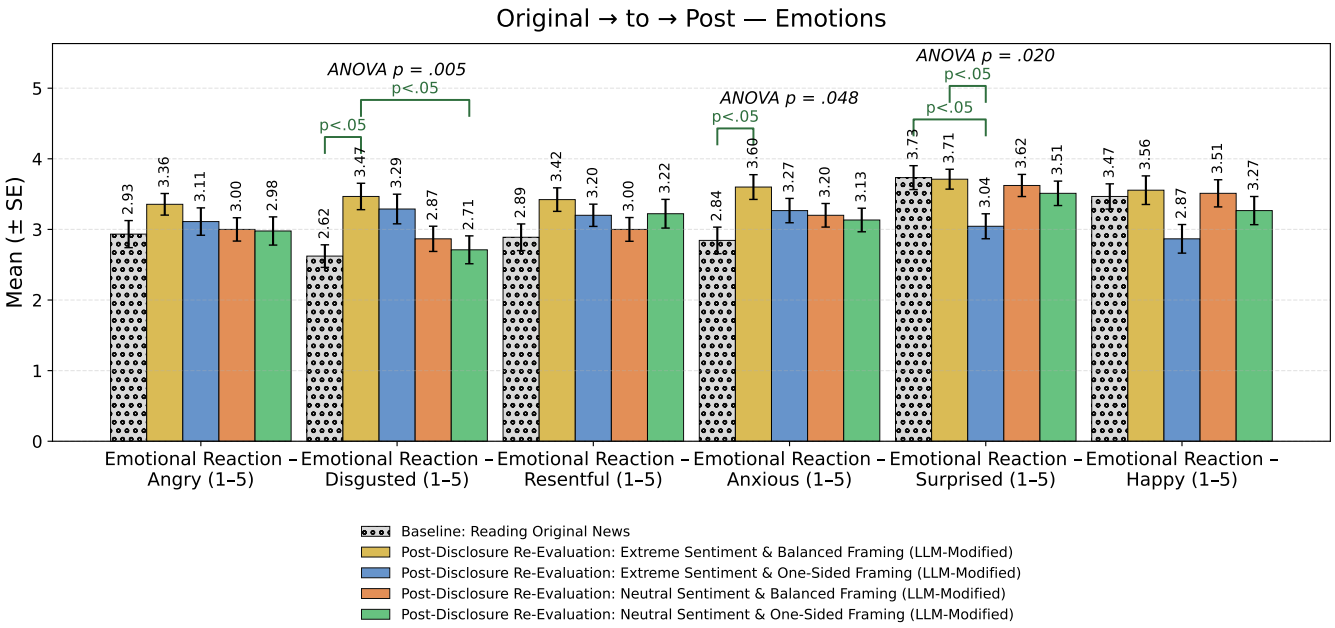


Figure 9: Comparison of baseline (original articles) and post-disclosure re-evaluations (LLM-modified articles) on emotional reactions. Error bars show ± SE. Patterns of emotional response remained relatively stable, with extreme sentiment conditions eliciting somewhat higher anger, disgust, and anxiety than neutral ones.

AI-modified articles as moderately altering the meaning or tone of the original (overall M = 4.53 on a 7-point scale). The highest ratings were observed for omissions of important information (M = 4.67) and introductions of new bias (M = 4.61), while misrepresentation (M = 4.31) and exaggeration (M = 4.48) received slightly lower agreement. Looking by condition, the Neutral Sentiment & Balanced Framing group tended to report the highest overall distortion (M = 4.64), while Neutral Sentiment & One-Sided Framing reported the lowest (M = 4.35). The Extreme Sentiment & One-Sided Framing group showed relatively higher perceptions of omission (M = 4.84) and bias (M = 4.87), whereas Extreme Sentiment & Balanced Framing scored highest on exaggeration (M = 4.84).

A mixed-model ANOVA showed a small main effect of distortion type ($F(4, 704) = 4.24$, $p = .002$, $\eta^2 = .01$) and a sentiment × framing interaction ($F(12, 704) = 2.08$, $p = .017$, $\eta^2 = .01$), with no significant condition differences. After correction, no pairwise comparisons were significant, and descriptive differences were minimal—for instance, neutral sentiment was rated only slightly more distorting than extreme sentiment under both framing types.

Overall, results indicate modest, pattern-level differences in how distortion is perceived across conditions, with framing exerting a somewhat stronger influence than sentiment. Importantly, ratings across all conditions remained above the midpoint of the scale, suggesting a general baseline skepticism among participants that AI modification may introduce some degree of distortion, even when editorial changes are minimal (see Table 4, Item A).

4.4.2 ***Partisanship and One-Sidedness Across Conditions.*** Participants generally viewed the AI-modified articles as somewhat partisan or one-sided (overall $M = 4.73$ on a 7-point scale). Across items, the strongest perceptions were that the modified versions appeared politically motivated (M = 4.79) and emotionally charged (M = 4.69), while exaggeration (M = 4.72) and bias compared to the original (M = 4.75) scored slightly lower. Ratings of balance (reverse-coded) were comparable (M = 4.68). Across conditions, mean differences were modest: Neutral Sentiment & Balanced Framing showed slightly higher partisanship ratings ($M = 4.85$), whereas Neutral Sentiment & One-Sided Framing was lowest ($M = 4.60$).

A mixed-model ANOVA revealed no significant main effects or interactions. Mean differences across the four sentiment–framing conditions were small (all ≤ 0.25) and not statistically significant. For example, Neutral Sentiment & Balanced Framing showed slightly higher partisanship ratings than Neutral Sentiment & One-Sided Framing (mean diff. = 0.25), but this difference was not inferentially meaningful. Accordingly, any descriptive differences should be interpreted cautiously.

Overall, the pattern shows only subtle variation across conditions: neutral sentiment with balanced framing received slightly higher partisanship ratings than more overtly extreme or one-sided versions. Rather than indicating causal effects, these trends may suggest that subtle framing shifts can paradoxically increase suspicion, perhaps because "balanced" AI-modified articles appear less transparent in intent.

In contrast, strongly partisan or one-sided outputs may be easier for participants to recognize and discount as biased. Taken together, these findings suggest that perceptions of AI-introduced partisanship are shaped less by overt cues and more by subtle modifications that undermine trust (see Table 4-Item B).

Table 4: Participant Ratings of AI-Modified News Articles with Statistical Results. Participants consistently perceived AI-modified articles as introducing some degree of distortion, with framing exerting a somewhat stronger influence than sentiment. Articles were rated as moderately partisan overall, with *Neutral Sentiment & Balanced Framing* descriptively associated with higher perceptions of bias, political motivation, and ethical concern than other conditions. In contrast, *Neutral Sentiment & One-Sided Framing* received the lowest ratings across these dimensions.

Note: Partial eta squared (η_p^2) values below .01 indicate negligible effects; values around .01 indicate small effects.

Item	Mean ± SD	Statistical Results
A. How Much Did the AI-Modified Version Change the Meaning or Tone?		
The AI-modified version misrepresented or distorted the original article.	4.31 ± 1.42	**Item main effect:** $p = .002$, $\eta_p^2 = .01$
The AI-modified version exaggerated aspects of the original article.	4.48 ± 1.43	**Item × Condition:** $p = .017$, $\eta_p^2 = .01$
The modified version omitted important information found in the original.	4.67 ± 1.35	**Post-hoc (Bonferroni):**
The AI-modified version introduced bias that wasn't present in the original.	4.61 ± 1.34	Omitted info > Misrepresented ($p = .002$)
The main message of the original article was lost or altered in the modified version.	4.60 ± 1.38	Message lost > Misrepresented ($p = .034$)
B. How Much Did You Notice Partisanship or One-Sidedness in AI-Modified News Articles?		
The modified version was more biased than the original.	4.75 ± 1.24	**Item main effect:** $p = .784$, $\eta_p^2 < .01$
The article presented a balanced view of the issue.	4.68 ± 1.31	**Condition:** $p = .701$, $\eta_p^2 < .01$
The modified article felt politically motivated.	4.79 ± 1.30	**Item × Condition:** $p = .075$, $\eta_p^2 = .01$
The language used in the modified article was emotionally charged.	4.69 ± 1.32	No significant pairwise contrasts
The article exaggerated positive or negative aspects of the issue.	4.72 ± 1.37	
C. How Much Did the AI-Modified Article Seem Designed to Influence Your Perception, in Terms of Intent and Ethics?		
The AI-modified article seemed designed to influence readers' opinions.	4.66 ± 1.33	**Item main effect:** $p = .133$, $\eta_p^2 < .01$
The changes made by AI felt ethically questionable.	4.68 ± 1.43	**Condition:** $p = .242$, $\eta_p^2 = .02$
The modified article felt manipulative.	4.66 ± 1.38	**Item × Condition:** $p = .256$, $\eta_p^2 = .01$
It was clear that the article had been altered to push a certain agenda.	4.84 ± 1.19	No significant pairwise contrasts

4.4.3 ***Intent and Ethical Concerns Across Conditions.*** Participants generally viewed the AI-modified articles as moderately concerning in terms of intent and ethics (overall $M = 4.72$). Perceptions were highest for "pushing an agenda" ($M = 4.84$) and ethical questionableness ($M = 4.68$), with similar ratings for manipulativeness and emotional tone ($M = 4.66$). Across conditions, mean differences were modest: Neutral Sentiment & Balanced Framing showed the greatest concern ($M = 4.98$), while Neutral Sentiment & One-Sided Framing showed the lowest ($M = 4.49$).

A mixed-model ANOVA revealed no significant main effects, interactions, or pairwise contrasts. At a high level, Neutral Sentiment & Balanced Framing tended to receive slightly higher concern ratings than the one-sided conditions (mean diffs. ≤ 0.49), though these patterns were not statistically reliable. Rather than indicating meaningful causal effects, they may reflect a general tendency for "balanced" AI-modified articles to appear more ethically questionable or manipulative than overtly extreme versions. This implies that subtle modifications may paradoxically heighten suspicion, perhaps because they obscure intent in ways that overtly partisan outputs do not.

Taken together, these findings suggest that concerns about AI's ethical influence on news are not limited to overt or easily identifiable bias cues but may also be exacerbated by more understated changes (see Table 4, Item C).

4.5 Post-Survey Reflections on Comfort, Trust, and the Future of AI in News

Responses from all participants to the post-survey questions highlighted broader concerns about the role of AI in news production and consumption. We did not split these by condition, as the aim was to capture general reflections rather than group-wise differences. When asked how comfortable they were with AI-generated or modified news articles being published in public media without explicit labeling, participants reported moderate levels of comfort on average ($M = 4.76$, $SD = 1.68$ on a 7-point scale). Responses spanned the full range of the scale (min = 1, max = 7), indicating substantial variability in individual attitudes.

Participants were also presented with a scenario-based question: *"Imagine an LLM summarizes news about a personal area of concern (e.g., war, terrorism, discrimination) because you enabled this feature to avoid an emotional or psychological trigger. The LLM does not alter the facts or narrative of the article; rather, it alters the language to be less distressing or inflammatory to you, the reader. Would you feel comfortable relying on this type of AI-generated summary for staying informed on issues you care about but may find distressing?"* Responses revealed cautious openness to this possibility. Roughly one-third indicated they would be comfortable with such summaries only if they were clearly labeled (33.8%), while nearly as many said they would be entirely comfortable without qualifications

(29.3%). A further 24.9% expressed conditional comfort, emphasizing that human oversight would be necessary. By contrast, a smaller proportion reported being not comfortable at all (8.9%), and a few were unsure (2.2%).

Taken together, these results suggest that while participants recognized the potential benefits of AI-generated summaries in sensitive contexts, labeling and human involvement were seen as critical safeguards to ensure trust and credibility.

After rating their comfort with AI-generated or modified news articles being published without labeling, participants were also asked a series of follow-up and open-ended questions to probe the reasoning behind their choices and their views on specific scenarios. These included: (1) explaining their comfort rating, (2) reflecting on whether they would rely on an LLM-generated summary that softened language for sensitive topics (while preserving facts), (3) discussing whether AI should fully replace human journalists in a hypothetical future of unbiased, perfectly accurate AI, and (4) describing how disclosure of AI involvement would affect their trust in news content.

4.5.1 ***Filtering and Quality Control of Qualitative Data.*** In line with prior findings that MTurk participants can often produce low-quality, inattentive, or automated responses to open-ended questions—while still yielding generally reliable closed-form survey data [31]—we retained all closed-form responses but applied stricter quality screening to open-ended text. Of 225 total entries, 115 were excluded, leaving 110 valid responses. Excluded responses fell into common categories identified in prior work: non-answers (e.g., restating the prompt or providing generic definitions), article summaries or copy-pasted content, clearly AI-generated responses (including explicit mentions of "Ask GPT"), off-topic material (e.g., hotel reviews or unrelated personal anecdotes), incoherent statements, and entries under 50 characters. This approach aligns with established recommendations for MTurk-based qualitative research and preserves the integrity of the quantitative analysis while ensuring that only substantive qualitative data are interpreted.

4.5.2 ***Human Oversight, Trust, and Transparency.*** When asked how comfortable they felt about AI-generated or modified news being published without explicit labeling, most participants emphasized that transparency is foundational to trust. Many expressed discomfort, stressing that undisclosed AI involvement could mislead readers about authorship, accuracy, or intent. As one participant noted, *"Without labeling, readers may assume the content was created entirely by human journalists, which can be misleading. This is especially important in journalism, where credibility and accountability are crucial [P2]."* Others worried about accountability gaps—human bylines create responsibility, while anonymous AI output leaves no one to challenge if errors or biases arise—and some described an "uncanny valley of journalism *[P6]*," where machine-written prose that feels human produces unease and skepticism. At the same time, a smaller group saw benefits when AI preserved factual accuracy or improved clarity; as one put it, *"If the core facts remain unchanged, I'm still getting the essentials—just without the emotional spike. [P79]"* Even among the more comfortable respondents, clear labeling and easy access to the original were treated as non-negotiable, with light human oversight preferred to catch nuance loss or bias.

When asked whether perfectly accurate, unbiased AI should replace journalists or still require human oversight, participants favored a hybrid model: AI for speed and scale, and humans for ethics, context, and accountability. Respondents emphasized that journalism is not just about facts but also moral judgment, cultural nuance, investigative instincts, and public responsibility. As one participant noted, *"Facts aren't enough—journalism is also about judgment, ethics, and context. [P96]"* Even in a *"perfect AI"* future, participants argued that humans must still decide what counts as news, weigh consequences, and provide empathy in sensitive reporting.

Finally, when asked how knowing an article was AI generated would affect their trust, most participants said it would make them more cautious rather than distrustful. Transparency again emerged as the decisive factor: clear labeling of AI involvement, disclosure of whether the system summarized, edited, or wrote content, and evidence of human review all increased confidence. By contrast, hidden or unexplained AI use eroded credibility, raising concerns about subtle bias, loss of nuance, and accountability. Overall, participants supported AI as a valuable journalistic tool—but only when paired with human oversight and transparent practices that preserve trust, accountability, and ethical responsibility.

4.5.3 ***Guardrails for Emotion-Aware AI.*** When asked whether they'd rely on AI summaries that maintain facts intact but soften distressing language, most participants expressed cautious comfort, welcoming a balance between staying informed and protecting mental health, if core safeguards are in place. Clear labeling and access to the original were non-negotiable (*"I want to know the tone was adjusted, not the truth. [P21]"*). Many favored light human oversight to catch nuance loss or bias (*"Have a person sanity-check so nothing serious gets sanitized [P27]"*). Supporters said softer wording sustains engagement with complex topics without overwhelming; skeptics warned of *"tone vs. truth [P6]"* drift, hidden assumptions in what counts as *"less distressing [P13, P14]"*, and over-reliance that could replace full articles. Overall, conditional acceptance prevailed: label it, preserve fidelity, show the original, and add optional human review so readers get the facts with fewer triggers and without dulling the gravity of events.

5 Discussion

In this study, we examined how emotional tone (sentiment), news framing (balanced vs. one-sided), and LLM disclosure relate to readers' perceptions of bias, trust/credibility, balance, and emotion using a 2×2 between-subjects design with pre- and post-disclosure evaluations. Overall, we observed several descriptive trends across conditions, though many effects were non-significant or marginal.

Briefly, we found that sentiment is the strongest lever: with respect to *RQ1*, articles with a more extreme tone were *associated with* higher perceived bias and elicited stronger negative emotional responses (anger, disgust, anxiety), while happiness remained relatively stable across conditions. Framing was generally influential, although instances of "balanced but extreme" article content were sometimes linked to increased surprise and suspicion. Considering *RQ2*, when sentiment was held constant, framing-related differences were smaller and less consistent. Neutral + Balanced was often viewed to be most credible, but paradoxically, balanced framing was sometimes seen as more manipulative than overtly extreme

versions, suggesting that subtle balance cues can undermine trust. Regarding *RQ3*, disclosure showed selective and modest effects. While overall credibility and balance judgments remained relatively stable, trustworthiness tended to decline slightly following disclosure, particularly for articles that were already extreme or one-sided. Rather than indicating uniform or strong causal effects, these patterns suggest that readers did not overhaul their judgments but became more cautious after disclosure. This reinforces the idea that transparency alone may erode trust unless paired with sufficient contextual information and warrants further testing in higher-powered, longitudinal studies.

5.1 Limits of Disclosure and Power of Sentiment

The results further reveal a nuanced picture. First, disclosure selectively reduced trustworthiness: readers who encountered post-disclosure articles rated them as slightly less trustworthy than those in baseline conditions (Section 4.2.3), particularly when news articles were already extreme and one-sided. Although the effect size was modest, this finding underscores that AI labels can dampen confidence in news content. Importantly, however, disclosure had little effect on perceived credibility or argument balance, suggesting that audiences continue to rely on traditional markers of journalistic quality such as clarity, evidentiary support, and rhetorical fairness when judging article's believability. In other words, transparency about AI involvement can influence trust, but perceptions of article content are more immutable.

Interestingly, political science experts reviewing the transformations noted that the articles largely "did what they were supposed to do"—the framing and sentiment shifts were clearly present and purposeful. Yet, despite recognizing these intended changes, the experts acknowledged that the overall effects on key perception metrics were modest or inconsistent. This reveals an important tension: while experts—who are trained to notice subtle discursive cues and shifts—recognized that the LLMs achieved the desired editorial changes, these changes did not produce significant differences in the metrics we tracked across conditions, indicating that average readers did not respond with the same level of sensitivity or altered judgments. In other words, the interventions achieved formal discursive variation without producing proportionate perceptual effects. Rather than a shortcoming, this highlights the complexity of audience responses and suggests that achieving meaningful shifts in perceptions of bias and trust through LLM modifications alone is a challenging task.

For lay readers, exposure to a single manipulated article may not be enough to alter judgments, either because the differences are too subtle to register or because readers rely on more stable heuristics (e.g., source cues, prior beliefs) that override textual variations. Alternatively, readers may engage superficially with the text, noticing tone but not recalibrating deeper trust assessments. From this perspective, the limited effects we observe should not be interpreted simply as null results, but rather as evidence of the difficulty of shifting perceptions through micro-level textual interventions alone. Consistent with this, perceived bias was driven more by extreme sentiment (Section 4.2.1 and Section 4.2.2), while balance ratings remained close to baseline (Section 4.2.3). This suggests that sentiment, more than framing, drove perceptions of bias. This finding aligns with research on misinformation detection, where sentiment analysis has been shown to play a central role in distinguishing fake from legitimate content [63]. Just as fake news detection benefits from sentiment cues, our participants' bias perceptions were most strongly influenced by emotional tone rather than structural framing shifts. Future efforts to minimize perceived bias in AI-generated or modified content may therefore benefit from careful attention to sentiment.

5.2 Affective Responses to AI-Modified News

While emotional outcomes showed stronger effects, not all emotional reactions were equally sensitive to our experimental manipulations of sentiment and framing. Most notably, extreme sentiment when paired with balanced framing elicited heightened surprise (Section 4.3.3), with ratings increasing by roughly 0.5 points on a 5-point scale compared to other conditions. Readers appeared attuned to this dissonance: a neutral-looking frame carrying emotionally charged language creates an unexpected and jarring experience. In theoretical terms, this aligns with, but also extends, the Hostile Media Effect (HME) [108], which suggests that readers tend to perceive balanced news coverage as biased by overweighting opposing frames relative to those that align with their views. Our results extend this framework by showing that such perceptions can arise not only from reader ideology but also from the way content is constructed—specifically, when balanced framing is paired with extreme sentiment. Under these conditions, negative reactions become more acute, highlighting the role of sentiment in amplifying perceived bias.

Beyond surprise, extreme sentiment consistently worsened emotional outcomes overall (Section 4.3). Participants reported higher levels of anger, disgust, resentment, and anxiety, while happiness remained comparatively stable and difficult to elicit. This asymmetry echoes prior findings that negativity is more psychologically potent than positivity in news contexts [96], with large-scale evidence that negative words in news headlines significantly increase consumption and engagement [91]. Related work further emphasizes that negativity in news is not monolithic: anger and fear, for instance, emerge as distinct emotional registers with different political and psychological consequences [96]. Our findings resonate with this distinction—anger and disgust were much more responsive to sentiment manipulations than fear or happiness—suggesting that AI-modified news may accentuate specific moral-emotional pathways rather than a generalized negativity. In practical terms, this asymmetry between negative and positive emotions highlights a vulnerability in how audiences respond to AI-modified or sentiment-laden news. Such content disproportionately evoked moral-emotional reactions (e.g., disgust) and cognitive-discrepant states (e.g., surprise, anxiety), underscoring how extreme tone can unintentionally fuel outrage. This aligns with recent evidence that emotionally charged content—especially anger and out-group animosity—is systematically amplified by ranking algorithms, intensifying polarization and distrust [67].

At the same time, our results suggest a critical interaction between framing and sentiment. While balanced coverage is generally thought to reduce perceived bias, it may have the opposite effect when paired with extreme sentiment—potentially intensifying,

rather than easing, hostile reactions. Further, "balanced" framing is not a safe default when both perspectives are expressed in extreme affective language; rather than neutralizing perceptions, this pairing can backfire by heightening surprise and perceptions of bias. Post hoc contrasts revealed that extreme sentiment, combined with one-sided framing, was the most consistent driver of heightened reactions (Section 4.3), producing significant differences relative to both the baseline and balanced conditions. Yet some omnibus differences (e.g., disgust, happiness shifts pre- to post-disclosure) did not yield reliable pairwise contrasts, underscoring the fragility and contextual dependence of these effects. This resonates with theoretical and empirical understandings of the HME, where partisans perceive relatively balanced news coverage as biased against their side [108]. It also aligns with subsequent HME work showing that such perceptions depend on format and audience expectations [42]. Yet our results extend these and related HME insights by showing that hostile reactions can arise not only from variations in audience ideology, but also from variation in content construction—specifically, when balanced framing is paired with extreme affective language. This interpretation is also consistent with Affective Intelligence Theory [64], which argues that heightened emotional cues—particularly those signaling threat or anxiety—can intensify audience scrutiny alongside concomitant hostile reactions to ostensibly balanced content.

Overall, these results reinforce that disclosure mechanisms matter most for trust, that sentiment is the strongest lever for bias and emotion, and that subtle "balanced but extreme" manipulations can backfire by increasing adverse perceptions of news media content. For HCI, this highlights that designing tools with stylistic neutrality alone is insufficient: tools that prioritize surface-level balance but fail to moderate sentiment risk worsening audience distrust and emotional volatility. In short, emotional responses, rather than credibility judgments alone, are central to understanding the role of LLMs in journalism. If left unchecked, emotionally manipulative interventions may sharpen morally charged responses, amplify destabilizing uncertainty, and worsen audience distrust, polarization, and democratic discourse. This calls for interactive systems that not only disclose AI's role in shaping content but also proactively support reflective, emotionally aware engagement with news.

5.3 Opacity and Skepticism in AI Rewrites

Beyond credibility and the emotional dynamics, participants' post-survey responses revealed a counterintuitive paradox in perceptions of AI-modified articles. Across multiple measures—including distortion, partisanship, and ethical concern, Neutral Sentiment & Balanced Framing was often rated more negatively than overtly extreme or one-sided versions. This paradox suggests that when AI modifications present themselves as "balanced," participants may actually become more suspicious of hidden agendas. This echoes Radivojevic et al. [90] findings that humans often perform poorly at detecting AI-generated content but nonetheless sense discomfort—an uncanny valley effect in text—which may explain why ostensibly balanced rewrites triggered suspicion despite being less overtly biased. In contrast, overtly extreme or one-sided modifications were easier for participants to recognize and discount as biased. Thus, subtle, ostensibly neutral modifications may undermine trust more than overt cues of partisanship, highlighting that attempts to temper sentiment and balance framing may backfire if the nature of the modification remains opaque. Qualitative findings (Section 4.5.2) echoed this skepticism, as participants emphasized that transparency is essential for trust and warned that undisclosed AI edits could mislead readers about the intent or accountability. This mirrors the study by Rossner et al. [92] on AI sports journalism, where disclosure had little effect on credibility, and the experiment by Parshakov et al. [84], which showed that disclosure reduced the preference for AI content. Our results extend this tension—when articles appear neutral but lack clear disclosure, suspicion intensifies. Transparency must therefore go beyond stating origin to revealing how changes were made. Participants' concerns map directly onto the paradox we observed: subtle changes risk being perceived as covert manipulation, precisely because they are harder to detect and explain.

Across measures of distortion, partisanship, and ethical concern, participants expressed that emotionally manipulative combinations of sentiment and framing can intensify reactions such as disgust and anxiety, while even neutral-seeming outputs may backfire if readers cannot discern how or why the article was changed. As one participant summarized, *"Facts aren't enough—journalism is also about judgment, ethics, and context."*

This lived skepticism aligns with broader concerns in media effects research, which suggests that hidden or understated biases can be more damaging to trust than overt partisanship [32, 40, 72]. Together, these findings highlight two critical risks associated with AI-mediated news modification. First, emotionally manipulative combinations of sentiment and framing can intensify destabilizing reactions such as disgust and anxiety. Second, even neutral-seeming outputs may backfire by heightening suspicion if readers cannot discern how and why the article was changed. Both patterns suggest that the risks of AI rewriting lie not only in overt distortion but also in subtle opacity, which can undermine reader trust even when sentiment is moderated.

For LLM-based interventions, the implication is that transparency must go beyond simple labels: tools should make visible not only that AI was involved, but also *how* text was changed, so that "balanced" rewrites do not inadvertently undermine credibility. Qualitative responses echoed this: participants emphasized that clear labeling, access to the original, and human oversight were "non-negotiable." Editorial dashboards that surface omissions, exaggerations, or agenda-pushing cues could help mitigate this risk, while provenance trails and side-by-side comparisons would allow readers to see what was altered. In short, readers are alert to distortion and manipulation even when subtle, and AI-mediated balance may backfire if not paired with robust transparency and oversight.

These findings result in several theoretical contributions by advancing debates about AI in journalism in three key ways. First, *transparency without penalty,* our results suggest that fears of an "AI penalty" are overstated: disclosure can enhance trust without undermining credibility or balance. Second, *sentiment as a central lever:* While framing effects were conditional and subtle, sentiment reliably shaped both perceptions and emotions, underscoring its outsized role in news design. Third, *extending hostile media theory:* we demonstrate that perceived bias can arise not only from audience ideology but also from pairing balanced framing with an extreme tone—an issue relevant to LLM-assisted writing. These

findings point to key implications for technological interventions: beyond transparency, tools must also support tone management and editorial oversight.

5.4 Design Recommendations

Guided by the findings, we outline several design directions below.

5.4.1 Contextualized Disclosure. News platforms should move beyond simple "AI-generated" badges toward richer process-oriented disclosures. For example: "This article was AI-modified for style; human editors verified the facts." Our findings showed that disclosure reduced trustworthiness primarily for extreme or one-sided content (Section 4.2.2), suggesting that readers penalize AI labels most when the article is already perceived as biased or emotionally charged. By providing contextual details about *what* was changed (e.g., style vs. substance) and *who* reviewed it (e.g., human editor verification), disclosures can mitigate unnecessary erosion of trust while preserving transparency. To address the "neutral balanced" paradox (Section 5.3), solutions should move beyond simply disclosing AI modifications. A core issue appears to be readers' cognitive tendency to downplay content that aligns with their views and overemphasize opposing perspectives—even in balanced articles. Interfaces could help mitigate this by explicitly signaling the presence of multiple viewpoints. For example, visual indicators or balance scores could make the article's framing more transparent, prompting readers to recognize its dual-perspective structure rather than perceiving it as biased or one-sided. Importantly, while some degree of skepticism is likely inevitable once readers learn content was AI-modified, our findings suggest that well-designed, process-oriented disclosures can shape *how* that skepticism is directed, encouraging critical evaluation of content quality rather than blanket distrust driven solely by the presence of AI.

5.4.2 Emotion-Aware Highlighting. Interfaces should incorporate sentiment and framing analysis to flag emotionally charged or one-sided passages. Interfaces could allow toggling between "original" and "neutralized" versions, or visually highlight language that exceeds affective thresholds or reflects an imbalance in arguments. Our results showed that extreme sentiment reliably elevated bias perceptions and amplified negative emotions such as anger, disgust, and anxiety, while positive emotions such as happiness were relatively stable (Section 4.3.1). In particular, "balanced but extreme" articles triggered increased surprise and suspicion, affecting their intended neutrality. Emotion-aware highlighting would allow readers to see where and how affective intensity is shaping the text, supporting reflective engagement rather than passive consumption. Building on this, systems could escalate from simple highlighting to adaptive framing alerts when extreme sentiment and one-sided framing co-occur—a pairing that most consistently provoked disproportionate disgust, anxiety, and surprise. Such layered cues empower readers to calibrate their own engagement with tone, reducing the risk of hostile media perceptions, and align with prior work on recommending in-situ warning methods for inflammatory or sensational language in online news [52].

5.4.3 Risk Aware Moderation and Signaling. LLM pipelines could automatically detect and soften extreme language while preserving factual accuracy. For example, extreme or one-sided articles could be sentiment-moderated and paired with a prominent disclosure ("AI-modified; may use heightened language; editor-verified"), while neutral pieces would undergo lighter adjustments and carry only a subtle badge. Interventions should be adaptive: stronger disclosures and editorial nudges may be needed in high-risk contexts (e.g., polarizing stories). At the same time, lighter-touch strategies can help avoid user fatigue in more routine coverage. This dual strategy addresses our findings that extreme sentiment amplified perceptions of bias and negative emotion, and that disclosure penalties were most pronounced in already extreme or one-sided conditions (Section 4.2.2). By combining moderation with adaptive transparency, tools can prevent "balanced but extreme" coverage from backfiring while preserving trust and usability. This direction aligns with recent work on *BIASsist*, which demonstrated that LLM-driven systems can go beyond flagging to actively identify, explain, and neutralize multiple forms of bias in news articles [74]. Extending such approaches to emotion-aware moderation would enable pipelines that not only reduce inflammatory tone but also help readers recognize how bias enters text and encourage more reflective engagement.

5.4.4 Distortion and Intent Auditing. Equip newsroom editors and staff with tools that automatically flag potential omissions, exaggerations, or agenda-pushing cues in AI-modified articles. Our results showed that participants perceived moderate levels of distortion and ethical concern (Section 4.4), with "balanced" versions sometimes seen as more manipulative than overtly extreme ones (Section 4.4.2). By surfacing these risks through automated auditing dashboards, editors can intervene before publication, ensuring that subtle manipulations do not erode credibility. Such auditing not only complements disclosure but also provides accountability within hybrid AI–human workflows.

5.5 Limitations and Future Work

We recognize some important considerations in our study. First, our reliance on the MTurk convenience sample may limit generalizability, as participants were relatively well-educated, politically engaged, slightly skewed toward Democrats, and not fully representative of the broader news-consuming public. This homogeneity is notable because trust in AI systems and media credibility varies across cultural, national, and partisan contexts; thus, our findings may reflect attitudes specific to U.S. news readers rather than more universal patterns. We did not analyze results by demographic subgroups, but doing so could provide valuable insights in future work, especially given that MTurk samples remain only a proxy for larger, more diverse populations. Future studies should incorporate cross-national or field-based samples to better capture how cultural, political, and contextual factors shape responses to AI-mediated news and to enhance external validity. Second, while we experimentally manipulated sentiment and framing, real-world news involves many other factors—such as source credibility, multimedia, and social context—that were beyond the scope of this study.

Third, although our study relied on a single article, constraining topic breadth—this was a deliberate design choice informed by extensive pretesting to ensure strong content validity and control over topic-level variation. Prior work cautions that single-stimulus

designs can be more susceptible to topic-specific effects, limit generalizability across topics, and may understate cross-topic variability, even as they offer important advantages in statistical power, interpretability, and experimental control when topic similarity is high [23]. Accordingly, our findings should be interpreted with appropriate caution when generalizing beyond the studied topic, and future work should balance these trade-offs by examining multiple topics while carefully managing topic-level confounds. Fourth, as with most surveys, responses may reflect social desirability bias [57]. Finally, although we included an AI disclosure, we did not test different formats or levels of transparency, an area future work could explore to better understand how disclosure shapes trust and engagement.

Building on these insights, future work should explore more ecologically valid interventions in dynamic, real-world environments. For example, AI-driven systems could flag articles with extreme sentiment or one-sided framing, or provide multi-perspective summaries that counterbalance perceived bias. Observed skepticism following AI disclosure may also reflect reactions to non-human authorship rather than to specific rewriting strategies; future work should disentangle attribution effects from transformation effects. The LLM-based manipulation pipeline introduced in this study can be scaled to larger datasets and applied to news feeds or social media contexts, enabling in situ evaluation of such interventions. Importantly, future studies should test the effects of varying disclosure styles, such as explicit labeling, inline cues, or source-level indicators, on trust, credibility, and willingness to engage. Together, these directions point toward adaptive, LLM-driven tools that surface persuasive techniques and provide context-aware support to help readers interpret AI-mediated news.

6 Conclusion

This work demonstrates how design decisions regarding language, framing, and disclosure directly impact user trust and engagement with AI-mediated content by combining controlled sentiment–framing manipulations with post-survey reflections. Through this approach, we contribute to HCI scholarship by revealing the effects of these variables on interaction outcomes.

Across conditions, participants consistently perceived AI-modified articles as introducing some distortion, with omissions and bias rated higher than exaggeration. Importantly, our results demonstrate that whereas framing played a more conditional role, with potential backfire effects when paired with extremes—for instance, neutral sentiment in a "balanced" frame often heightened suspicion of partisanship more, and neutral sentiment in a "balanced" frame often raised greater suspicion of partisanship than overtly extreme versions. Extreme sentiment manipulations strongly amplified negative emotions, whereas positive emotions (happiness and surprise) remained comparatively stable. Finally, disclosure effects were selective rather than global, reducing trust mainly for articles that were already perceived as extreme or one-sided. While participants recognized efficiency and supportive use cases for AI involvement (e.g., reducing distressing language), they expressed strong preferences for clear labeling and human oversight as safeguards.

Together, these findings highlight not only the importance of transparency and human oversight but also the urgent need for emotion-aware, sentiment-moderation tools that can temper the amplification of outrage and suspicion in AI-mediated news. Subtle design choices in language and framing are never neutral; they shape trust, amplify or dampen polarization, and directly influence how people exercise agency in information ecosystems. By making these dynamics visible, this work invites the HCI community to more directly engage with the sociotechnical implications of AI-driven content systems and to design interventions that support informed, reflective engagement rather than undermine it.

Acknowledgments

We thank members of the Sensify Lab and the greater Department of Computer and Information Sciences community at the University of Delaware (UD) for their valuable feedback and support throughout this project. We also thank our volunteer coders from the UD Department of Political Science and International Relations for their valuable time. We also thank the UD Undergraduate Research Program for supporting Varun Pappu's participation via the Summer Scholars program.

References

[1] Mayank Agarwal, Priyanka Sharma, and Pinaki Wani. 2025. Evaluating the Accuracy and Reliability of Large Language Models (ChatGPT, Claude, DeepSeek, Gemini, Grok, and Le Chat) in Answering Item-Analyzed Multiple-Choice Questions on Blood Physiology. *Cureus* 17, 4 (2025). doi:10.7759/cureus.81871

[2] Alexandre Agossah, Frédérique Krupa, Matthieu Perreira Da Silva, and Patrick Le Callet. 2023. LLM-Based Interaction for Content Generation: A Case Study on the Perception of Employees in an IT Department. In *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences* (Nantes, France) *(IMX '23)*. Association for Computing Machinery, New York, NY, USA, 237–241. doi:10.1145/3573381.3603362

[3] Rohan Alexander, Lindsay Katz, Callandra Moore, Michael Wing-Cheung Wong, and Zane Schwartz. 2023. Evaluating the Decency and Consistency of Data Validation Tests Generated by LLMs. https://api.semanticscholar.org/CorpusID:263605445

[4] SM Asger Ali and Duane A Gill. 2022. Media Framing and Agenda Setting (Tone) in News Coverage of Hurricane Harvey: A Content Analysis of the New York Times, Wall Street Journal, and Houston Chronicle from 2017 to 2018. *Weather, Climate, and Society* 14, 2 (2022), 637–649. doi:10.1175/WCAS-D-21-0009.1

[5] David Alonso del Barrio and Daniel Gatica-Perez. 2023. Framing the News: From Human Perception to Large Language Model Inferences. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval* (Thessaloniki, Greece) *(ICMR '23)*. Association for Computing Machinery, New York, NY, USA, 627–635. doi:10.1145/3591106.3592278

[6] Ron Aminzade, Doug McAdam, et al. 2001. Emotions and contentious politics. *Silence and voice in the study of contentious politics* (2001), 14–50. doi:10.1017/CBO9780511815331.003

[7] Alberto Ardèvol-Abreu and Homero Gil de Zúñiga. 2017. Effects of editorial media bias perception and media trust on the use of traditional, citizen, and social media news. *Journalism & mass communication quarterly* 94, 3 (2017), 703–724. doi:10.1177/1077699016654684

[8] Zahra Ashktorab, Q. Vera Liao, Casey Dugan, James Johnson, Qian Pan, Wei Zhang, Sadhana Kumaravel, and Murray Campbell. 2020. Human-AI Collaboration in a Cooperative Game Setting: Measuring Social Perception and Outcomes. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 96 (Oct. 2020), 20 pages. doi:10.1145/3415167

[9] Elizabeth J. Austin. 2005. Emotional intelligence and emotional information processing. *Personality and Individual Differences* 39, 2 (2005), 403–414. doi:10.1016/j.paid.2005.01.017

[10] Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik van der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. 2010. Sentiment Analysis in the News. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias (Eds.). European Language Resources Association (ELRA), Valletta, Malta. https://aclanthology.org/L10-1623/

[11] Matthew A. Baum and Tim Groeling. 2008. New Media and the Polarization of American Political Discourse. *Political Communication* 25, 4 (2008), 345–365. doi:10.1080/10584600802426965 arXiv:https://doi.org/10.1080/10584600802426965

[12] Charlie Beckett and Mark Deuze. 2016. On the Role of Emotion in the Future of Journalism. *Social Media + Society* 2, 3 (2016), 2056305116662395. doi:10.1177/2056305116662395 arXiv:https://doi.org/10.1177/2056305116662395

[13] Monika Bednarek and Helen Caple. 2014. Why do news values matter? Towards a new methodological framework for analysing news discourse in critical discourse analysis and beyond. *Discourse & Society* 25, 2 (2014), 135–158. doi:10.1177/0957926513516041

[14] Samuel E Bestvater and Burt L Monroe. 2023. Sentiment is not stance: Target-aware opinion classification for political text analysis. *Political Analysis* 31, 2 (2023), 235–256. doi:10.1017/pan.2022.10

[15] John Bianchi, Manuel Pratelli, Marinella Petrocchi, and Fabio Pinelli. 2024. Evaluating Trustworthiness of Online News Publishers via Article Classification. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing* (Avila, Spain) *(SAC '24)*. Association for Computing Machinery, New York, NY, USA, 671–678. doi:10.1145/3605098.3636044

[16] Madalina Botan, Nicoleta Corbu, and Dani Sandu. 2016. The complicated relation between news frames and political trust: A case study of Romania. *Středoevropské politické studie/Central European Political Studies Review* 18, 2-3 (2016), 122–140. doi:10.5817/CEPSR.2016.23.122

[17] Ted Brader. 2006. *Campaigning for Hearts and Minds: How Emotional Appeals in Political Ads Work*. University of Chicago Press, Chicago.

[18] David Caswell. 2024. Audiences, automation, and AI: From structured news to language models. *AI Mag.* 45, 2 (June 2024), 174–186. doi:10.1002/aaai.12168

[19] Dennis Chong. 2019. Competitive Framing in Political Decision Making. In *Oxford Research Encyclopedia of Politics*. Oxford University Press. https://api.semanticscholar.org/CorpusID:211387162

[20] Dennis Chong and James N. Druckman. 2007. Framing Public Opinion in Competitive Democracies. *American Political Science Review* 101, 4 (2007), 637–655. doi:10.1017/S0003055407070554

[21] Dennis Chong and James N Druckman. 2007. Framing theory. *Annu. Rev. Polit. Sci.* 10, 1 (2007), 103–126. https://doi.org/10.1146/annurev.polisci.10.072805.103054

[22] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118, 9 (2021), e2023301118. https://www.pnas.org/doi/abs/10.1073/pnas.2023301118

[23] Scott Clifford and Carlisle Rainey. 2025. The limits (and strengths) of single-topic experiments. *Political Analysis* 33, 2 (2025), 164–170. doi:10.1017/pan.2024.20

[24] AE Code and Part LXIII Psychologists. 2017. Ethical principles of Psychologist and code of conduct. *Published online* (2017).

[25] Erica Coppolillo, Federico Cinus, Marco Minici, Francesco Bonchi, and Giuseppe Manco. 2025. Engagement-Driven Content Generation with Large Language Models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2* (Toronto ON, Canada) *(KDD '25)*. Association for Computing Machinery, New York, NY, USA, 369–379. doi:10.1145/3711896.3736932

[26] Maria D. Molina, S. Shyam Sundar, Md Main Uddin Rony, Naeemul Hassan, Thai Le, and Dongwon Lee. 2021. Does Clickbait Actually Attract More Clicks? Three Clickbait Studies You Must Read. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 234, 19 pages. doi:10.1145/3411764.3445753

[27] Valdemar Danry, Pat Pataranutaporn, Matthew Groh, and Ziv Epstein. 2025. Deceptive Explanations by Large Language Models Lead People to Change their Beliefs About Misinformation More Often than Honest Explanations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 933, 31 pages. doi:10.1145/3706598.3713408

[28] Claes H De Vreese. 2005. News framing: Theory and typology. *Information design journal+ document design* 13, 1 (2005), 51–62. doi:10.1075/idjdd.13.1.06vre

[29] Zijian Ding, Alison Smith-Renner, Wenjuan Zhang, Joel Tetreault, and Alejandro Jaimes. 2023. Harnessing the power of LLMs: Evaluating human-AI text co-creation through the lens of news headline generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 3321–3339. doi:10.18653/v1/2023.findings-emnlp.217

[30] James N. Druckman. 2004. Political Preference Formation: Competition, Deliberation, and the (Ir)relevance of Framing Effects. *American Political Science Review* 98, 4 (2004), 671–686. doi:10.1017/S0003055404041413

[31] Marc Dupuis, Karen Renaud, and Rosalind Searle. 2022. Crowdsourcing Quality Concerns: An Examination of Amazon's Mechanical Turk. In *Proceedings of the 23rd Annual Conference on Information Technology Education* (Chicago, IL, USA) *(SIGITE '22)*. Association for Computing Machinery, New York, NY, USA, 127–129. doi:10.1145/3537674.3555783

[32] Nicholas D Duran, Stephen P Nicholson, and Rick Dale. 2017. The hidden appeal and aversion to political conspiracies as revealed in the response dynamics of partisans. *Journal of Experimental Social Psychology* 73 (2017), 268–278. doi:10.1016/j.jesp.2017.07.008

[33] Robert M. Entman. 1993. Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication* 43, 4 (1993), 51–58. doi:10.1111/j.1460-2466.1993.tb01304.x

[34] Carrie Figdor. 2019. Trust Me: News, Credibility Deficits, and Balance. In *Media Ethics, Free Speech, and the Requirements of Democracy*, Carl Fox and Joe Saunders (Eds.). Routledge, 69 – 86. doi:10.4324/9780203702444-5

[35] Marina Fiori, Shagini Udayar, and Ashley Vesely Maillefer. 2022. Emotion information processing as a new component of emotional intelligence: Theoretical framework and empirical evidence. *European Journal of Personality* 36, 2 (2022), 245–264. doi:10.1177/08902070211007672

[36] Caroline Fisher. 2016. The trouble with 'trust'in news media. *Communication Research and Practice* 2, 4 (2016), 451–465. https://doi.org/10.1080/22041451.2016.1261251

[37] Richard Fletcher and Sora Park. 2017. The impact of trust in the news media on online news consumption and participation. *Digital journalism* 5, 10 (2017), 1281–1299. doi:10.1080/21670811.2017.1279979

[38] Elizabeth González-Estrada and Waldenia Cosmes. 2019. Shapiro–Wilk test for skew normal distributions based on data transformations. *Journal of Statistical Computation and Simulation* 89, 17 (2019), 3258–3272.

[39] Google. 2024. Gemini AI Overview. https://gemini.google.com/app. Accessed: November 10, 2025.

[40] Andrew M. Guess, Pablo Barberá, Simon Munzert, and JungHwan Yang. 2021. The consequences of online partisan media. *Proceedings of the National Academy of Sciences* 118, 14 (2021), e2013464118. doi:10.1073/pnas.2013464118

[41] Greg Guest, Kathleen M MacQueen, and Emily E Namey. 2011. *Applied thematic analysis*. sage publications. doi:10.4135/9781483384436

[42] Albert C Gunther and Kathleen Schmitt. 2004. Mapping boundaries of the hostile media effect. *Journal of Communication* 54, 1 (2004), 55–70. doi:10.1111/j.1460-2466.2004.tb02613.x

[43] Glenn J. Hansen and Hyunjung Kim. 2011. Is the Media Biased Against Me? A Meta-Analysis of the Hostile Media Effect Research. *Communication Research Reports* 28 (2011), 169 – 179. https://api.semanticscholar.org/CorpusID:145256611

[44] Hana Hurtíková. 2017. The Importance of Valence-Framing in the Process of Political Communicati on: Effects on the Formation of Political Attitudes among Viewers of Television News in the Czech Republic. *Media Studies* 8, 15 (2017). doi:10.20901/ms.8.15.6

[45] Martin Huschens, Martin Briesch, Dominik Sobania, and Franz Rothlauf. 2023. Do You Trust ChatGPT? - Perceived Credibility of Human and AI-Generated Content. *ArXiv* abs/2309.02524 (2023). https://api.semanticscholar.org/CorpusID:261557351

[46] Clayton J. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media* (2014). https://api.semanticscholar.org/CorpusID:12233345

[47] Mahmud Isnan, Gregorius Natanael Elwirehardja, and Bens Pardamean. 2023. Sentiment Analysis for TikTok Review Using VADER Sentiment and SVM Model. *Procedia Computer Science* 227 (2023), 168–175. doi:10.1016/j.procs.2023.10.514 8th International Conference on Computer Science and Computational Intelligence (ICCSCI 2023).

[48] Bowen Jiang, Yangxinyu Xie, Zhuoqun Hao, Xiaomeng Wang, Tanwi Mallick, Weijie J Su, Camillo Jose Taylor, and Dan Roth. 2024. A Peek into Token Bias: Large Language Models Are Not Yet Genuine Reasoners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 4722–4756. doi:10.18653/v1/2024.emnlp-main.272

[49] Kai Jiang, Qilai Zhang, Dongsheng Guo, Dengrong Huang, Sijia Zhang, Zizhong Wei, Fanggang Ning, and Rui Li. 2024. AI-Generated News Articles Based on Large Language Models. In *Proceedings of the 2023 International Conference on Artificial Intelligence, Systems and Network Security* (Mianyang, China) *(AISNS '23)*. Association for Computing Machinery, New York, NY, USA, 82–87. doi:10.1145/3661638.3661654

[50] Yongnam Jung, Peixin Hua, Jiaqi (Agnes) Bao, and S. Shyam Sundar. 2025. AI-Generated or AI-Modified? User Reactions to Labeling AI Use in Social Media Posts. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 57, 7 pages. doi:10.1145/3706599.3720264

[51] Dimitri Kelly. 2019. Evaluating the news:(Mis) perceptions of objectivity and credibility. *Political Behavior* 41, 2 (2019), 445–471. doi:10.1007/s11109-018-9458-4

[52] Prerana Khatiwada, Luke Halko, Nabiha Syed, Ashrey Mahesh, Aneseh Alvanpour, and Matthew Louis Mauriello. 2025. Spotting Online News: A Mixed Method Study of Online News Engagement and Perceptions on Misinformation Interventions. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW173 (May 2025), 30 pages. doi:10.1145/3711071

[53] Christopher S. G. Khoo, Armineh Nourbakhsh, and Jin-Cheon Na. 2012. Sentiment analysis of online news text: a case study of appraisal theory. *Online Inf. Rev.* 36 (2012), 858–878. https://api.semanticscholar.org/CorpusID:15553484

[54] Hyo J Kim and Glen T Cameron. 2011. Emotions matter in crisis: The role of anger and sadness in the publics' response to crisis news framing and corporate crisis response. *Communication Research* 38, 6 (2011), 826–855. doi:10.1177/0093650210385813

[55] Joel Kiskola, Henrik Rydenfelt, Thomas Olsson, Lauri Haapanen, Noora Väntinen, Matti Nelimarkka, Minna Vigren, Salla-Maaria Laaksonen, and Tuukka Lehtiniemi. 2025. Generative AI and News Consumption: Design Fictions and Critical Analysis. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 250, 18 pages. doi:10.1145/3706598.3713804

[56] Sergei Koltcov, Vera Ignatenko, and Olessia Koltsova. 2019. Estimating topic modeling performance with sharma–mittal entropy. *Entropy* 21, 7 (2019), 660. doi:10.3390/e21070660

[57] Ivar Krumpal. 2013. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & quantity* 47, 4 (2013), 2025–2047. doi:10.1007/s11135-011-9640-9

[58] Akshi Kumar, Teeja Mary Sebastian, et al. 2012. Sentiment analysis: A perspective on its past, present and future. *International Journal of Intelligent Systems and Applications* 4, 10 (2012), 1–14. doi:10.5815/ijisa.2012.10.01

[59] Andreas Langlotz and Miriam A Locher. 2012. Ways of communicating emotional stance in online disagreements. *Journal of pragmatics* 44, 12 (2012), 1591–1606. doi:10.1016/j.pragma.2012.04.002

[60] Sophie Lecheler, Andreas R. T. Schuck, and Claes H. de Vreese. 2013. Dealing with feelings: Positive and negative discrete emotions as mediators of news framing effects. *Communications - The European Journal of Communication Research* 38, 2 (2013), 189–209. doi:doi:10.1515/commun-2013-0011

[61] Irwin P. Levin, Sandra L. Schneider, and Gary J. Gaeth. 1998. All Frames Are Not Created Equal: A Typology and Critical Analysis of Framing Effects. *Organizational Behavior and Human Decision Processes* 76, 2 (1998), 149–188. doi:10.1006/obhd.1998.2804

[62] Charles G. Lord, Lee Ross, and Mark R. Lepper. 1979. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology* 37, 11 (1979), 2098–2109. doi:10.1037/0022-3514.37.11.2098

[63] Ishrar Mannan and Sifat Nawrin Nova. 2023. An Empirical Study on Theories of Sentiment Analysis in Relation to Fake News Detection. *2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)* (2023), 79–83. https://api.semanticscholar.org/CorpusID:265055181

[64] George E. Marcus, W. Russell Neuman, and Michael MacKuen. 2000. *Affective Intelligence and Political Judgment*. University of Chicago Press, Chicago.

[65] Lesedi Masisi, V Nelwamondo, and Tshilidzi Marwala. 2008. The use of entropy to measure structural diversity. In *2008 IEEE International Conference on Computational Cybernetics*. IEEE, 41–45. doi:10.1109/ICCCYB.2008.4721376

[66] Daniel G McDonald and John Dimmick. 2003. The conceptualization and measurement of diversity. *Communication Research* 30, 1 (2003), 60–79. doi:10.1177/0093650202239026

[67] Smitha Milli, Micah Carroll, Sashrika Pandey, Yike Wang, and Anca D. Dragan. 2023. Twitter's Algorithm: Amplifying Anger, Animosity, and Affective Polarization. *ArXiv* abs/2305.16941 (2023). https://api.semanticscholar.org/CorpusID:263866336

[68] Sophie Morosoli, Emma van der Goot, Valeria Resendez, Claes de Vreese, and Natali Helberger. 2025. The Transparency Dilemma: An Experiment on How AI Disclosures Affect Credibility Perceptions and Engagement Across Topics. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 8. 1748–1757. doi:10.1609/aies.v8i2.36671

[69] Thomas E. Nelson and Donald R. Kinder. 1996. Issue Frames and Group-Centrism in American Public Opinion. *The Journal of Politics* 58 (1996), 1055 – 1078. https://api.semanticscholar.org/CorpusID:154579037

[70] Thomas E. Nelson, Zoe M. Oxley, and Rosalee A. Clawson. 1997. Toward a Psychology of Framing Effects. *Political Behavior* 19 (1997), 221–246. https://api.semanticscholar.org/CorpusID:15874936

[71] Amul Neupane, Kapalik Khanal, Nitesh Nepal, and Naseeb Dangi. 2025. Advanced News Aggregation and Content Generation Using LLMs and NLP Algorithms. *European Journal of Applied Science, Engineering and Technology* 3, 2 (2025), 295–303. doi:10.59324/ejaset.2025.3(2).24

[72] Nic Newman and Richard Fletcher. 2017. *Bias, Bullshit and Lies: Audience Perspectives on Low Trust in the Media*. Technical Report. Reuters Institute for the Study of Journalism, University of Oxford. doi:10.2139/ssrn.3173579 Based on qualitative analyses of the 2017 Reuters Institute Digital News Report.

[73] Yeo-Gyeong Noh, MinJu Han, Junryeol Jeon, and Jin-Hyuk Hong. 2025. BIASsist: Empowering News Readers via Bias Identification, Explanation, and Neutralization. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 248, 24 pages. doi:10.1145/3706598.3713531

[74] Yeo-Gyeong Noh, MinJu Han, Junryeol Jeon, and Jin-Hyuk Hong. 2025. BIASsist: Empowering News Readers via Bias Identification, Explanation, and Neutralization. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 248, 24 pages. doi:10.1145/3706598.3713531

[75] Elie Ofek, Hema Yoganarasimhan, et al. 2025. Balancing Engagement and Polarization: Multi-Objective Alignment of News Content Using LLMs. *arXiv preprint arXiv:2504.13444* (2025).

[76] Amin Omidvar. 2025. Assessing And Enhancing The Quality Of News Headlines Using Machine Learning. (2025).

[77] Michael E O'Neill and Ky L Mathews. 2002. Levene tests of homogeneity of variance for general block and treatment designs. *Biometrics* 58, 1 (2002), 216–224. doi:10.1111/j.0006-341X.2002.00216.x

[78] OpenAI. 2024. GPT-4 Overview. https://platform.openai.com/docs/models/gpt-4. Accessed: November 10, 2025.

[79] Zoe M. Oxley. 2020. Framing and Political Decision Making: An Overview. *Oxford Research Encyclopedia of Politics* (2020). https://api.semanticscholar.org/CorpusID:225873926

[80] Cliodhna O'Connor and Helene Joffe. 2020. Intercoder reliability in qualitative research: Debates and practical guidelines. *International journal of qualitative methods* 19 (2020), 1609406919899220. doi:10.1177/1609406919899220

[81] Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. In *Proceedings of the 38th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '24)*. Curran Associates Inc., Red Hook, NY, USA, Article 2197, 31 pages.

[82] Gabriele Paolacci, Jesse J. Chandler, and Panagiotis G. Ipeirotis. 2010. Running Experiments on Amazon Mechanical Turk. *Behavioral & Experimental Economics eJournal* (2010). https://api.semanticscholar.org/CorpusID:14476283

[83] Sora Park, Caroline Fisher, Terry Flew, and Uwe Dulleck. 2020. Global Mistrust in News: The Impact of Social Media on Trust. *International Journal on Media Management* 22 (2020), 83 – 96. https://api.semanticscholar.org/CorpusID:221538940

[84] Petr Parshakov, Iuliia N. Naidenova, Sofia Paklina, Nikita Matkin, and Cornel Nesseler. 2025. Users Favor LLM-Generated Content - Until They Know It's AI. *ArXiv* abs/2503.16458 (2025). https://api.semanticscholar.org/CorpusID:277244763

[85] Valeria Pastorino and Nafise Sadat Moosavi. 2025. Frame In, Frame Out: Do LLMs Generate More Biased News Headlines than Humans? *arXiv preprint arXiv:2505.05406* (2025).

[86] Mia Perlina. 2019. DISCURSIVE STRATEGIES OF NEWS PRESENTATION IN THE SELECTED ONLINE NEWSPAPERS. *Lexeme: Journal of Linguistics and Applied Linguistics* 1, 1 (2019), 8. doi:10.32493/ljlal.v1i1.2478

[87] Chris Peters. 2011. Emotion aside or emotional side? Crafting an 'experience of involvement'in the news. *Journalism* 12, 3 (2011), 297–316. doi:10.1177/1464884910388224

[88] Pew Research Center. 2025. Party Affiliation Fact Sheet (NPORS). https://www.pewresearch.org/politics/fact-sheet/party-affiliation-fact-sheet-npors/. Accessed: [insert the date you accessed it].

[89] Cornelius Puschmann and Alison Powell. 2018. Turning Words Into Consumer Preferences: How Sentiment Analysis Is Framed in Research and the News Media. *Social Media + Society* 4 (2018). https://api.semanticscholar.org/CorpusID:149855156

[90] Kristina Radivojevic, Matthew Chou, Karla Badillo-Urquiola, and Paul Brenner. 2024. Human perception of llm-generated text content in social media environments. *arXiv preprint arXiv:2409.06653* (2024). doi:10.48550/arXiv.2409.06653

[91] Claire E Robertson, Nicolas Pröllochs, Kaoru Schwarzenegger, Philip Pärnamets, Jay J Van Bavel, and Stefan Feuerriegel. 2023. Negativity drives online news consumption. *Nature human behaviour* 7, 5 (2023), 812–822. doi:10.1038/s41562-023-01538-4

[92] Alexander Rossner, Marie Cassel, and Martin Huschens. 2024. Do Users Really Care? Evaluating the User Perception of Disclosing AI-Generated Content on Credibility in (Sports) Journalism. In *Proceedings of Mensch Und Computer 2024* (Karlsruhe, Germany) *(MuC '24)*. Association for Computing Machinery, New York, NY, USA, 413–418. doi:10.1145/3670653.3677490

[93] Dietram A Scheufele. 2000. Agenda-setting, priming, and framing revisited: Another look at cognitive effects of political communication. *Mass communication & society* 3, 2-3 (2000), 297–316. doi:10.1207/S15327825MCS0323_07

[94] Shreya Shankar, J.D. Zamfirescu-Pereira, Bjoern Hartmann, Aditya Parameswaran, and Ian Arawjo. 2024. Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 131, 14 pages. doi:10.1145/3654777.3676450

[95] Stuart Soroka. 2012. The Gatekeeping Function: Distributions of Information in Media and the Real World. *The Journal of Politics* 74 (2012), 514–528. https://api.semanticscholar.org/CorpusID:153963180

[96] Stuart Soroka, Lori Young, and Meital Balmas. 2015. Bad news or mad news? Sentiment scoring of negativity, fear, and anger in news content. *The ANNALS of the American Academy of Political and Social Science* 659, 1 (2015), 108–121. doi:10.1177/0002716215569217

[97] Stuart N Soroka. 2006. Good news and bad news: Asymmetric responses to economic information. *The journal of Politics* 68, 2 (2006), 372–385. doi:10.1111/j.1468-2508.2006.00413.x

[98] Jessica F. Sparks and Jay D. Hmielowski. 2022. At the Extremes: Assessing Readability, Grade Level, Sentiment, and Tone in US Media Outlets. *Journalism Studies* 24 (2022), 24 – 44. https://api.semanticscholar.org/CorpusID:253651746

[99] Timo Spinde, Felix Hamborg, Karsten Donnay, Angelica Becerra, and Bela Gipp. 2020. Enabling News Consumers to View and Understand Biased News Coverage: A Study on the Perception and Visualization of Media Bias. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (Virtual Event, China) *(JCDL '20)*. Association for Computing Machinery, New York, NY, USA, 389–392. doi:10.1145/3383583.3398619

[100] Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. 2024. On the self-verification limitations of large language models on reasoning and planning tasks. *arXiv preprint arXiv:2402.08115* (2024).

[101] Jesper Strömbäck, Yariv Tsfati, Hajo Boomgaarden, Alyt Damstra, Elina Lindgren, Rens Vliegenthart, and Torun Lindholm. 2020. News media trust and its impact on media use: Toward a framework for future research. *Annals of the International Communication Association* 44, 2 (2020), 139–156. doi:10.1080/23808985.2020.1755338

[102] Natalie Jomini Stroud. 2010. Polarization and partisan selective exposure. *Journal of communication* 60, 3 (2010), 556–576. doi:10.1111/j.1460-2466.2010.01497.x

[103] Faria Sultana, Md. Tahmid Hasan Fuad, Md. Fahim, Rahat Rizvi Rahman, Meheraj Hossain, M. Ashraful Amin, A. K. M. Mahbubur Rahman, and Amin Ahsan Ali. 2024. How Good are LM and LLMs in Bangla Newspaper Article Summarization?. In *Pattern Recognition: 27th International Conference, ICPR 2024, Kolkata, India, December 1–5, 2024, Proceedings, Part XX* (Kolkata, India). Springer-Verlag, Berlin, Heidelberg, 72–86. doi:10.1007/978-3-031-78498-9_6

[104] C Sunstein. 2001. Republic. com Princeton, NJ: Princeton Univ.

[105] Kehui Tan, Jiayang Yao, Tianqi Pang, Chenyou Fan, and Yu Song. 2025. ELF: Educational LLM Framework of Improving and Evaluating AI-generated Content for Classroom Teaching. *J. Data and Information Quality* 17, 3, Article 14 (Sept. 2025), 23 pages. doi:10.1145/3712065

[106] James W Tankard Jr. 2001. The Empirical Approach to the Study of Media Framing. In *Framing public life*. Routledge, 111–121. https://api.semanticscholar.org/CorpusID:140918491

[107] Amos Tversky and Daniel Kahneman. 1981. The Framing of Decisions and the Psychology of Choice. *Science* 211, 4481 (1981), 453–458. doi:10.1126/science.7455683

[108] Robert P. Vallone, Lee D. Ross, and Mark R. Lepper. 1985. The hostile media phenomenon: biased perception and perceptions of media bias in coverage of the Beirut massacre. *Journal of personality and social psychology* 49 3 (1985), 577–85. https://api.semanticscholar.org/CorpusID:16311781

[109] Karthik Valmeekam, Matthew Marquez, and Subbarao Kambhampati. 2023. Can Large Language Models Really Improve by Self-critiquing Their Own Plans? *ArXiv* abs/2310.08118 (2023). https://api.semanticscholar.org/CorpusID:263909251

[110] Dara M. Wald, Erik W. Johnston, Ned Wellman, and John Harlow. 2021. How Does Personalization in News Stories Influence Intentions to Help With Drought? Assessing the Influence of State Empathy and Its Antecedents. *Frontiers in Communication* 5 (2021), 588978.

[111] Jenny S Wang, Samar Haider, Amir Tohidi, Anushkaa Gupta, Yuxuan Zhang, Chris Callison-Burch, David Rothschild, and Duncan J Watts. 2025. Media Bias Detector: Designing and Implementing a Tool for Real-Time Selection and Framing Bias Analysis in News Coverage. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 790, 27 pages. doi:10.1145/3706598.3713716

[112] Nurul Hidayah Watimin, Hasmah Zanuddin, Mohamad Saleeh Rahamad, and Elaheh Yadegaridehkordi. 2023. Content framing role on public sentiment formation for pre-crisis detection on sensitive issue via sentiment analysis and content analysis. *Plos one* 18, 10 (2023), e0287367. doi:10.1371/journal.pone.0287367

[113] Eric W Weisstein. 2004. Bonferroni correction. *https://mathworld. wolfram. com/* (2004).

[114] Werner Wirth and Holger Schramm. 2005. Media and emotions. *Communication research trends* 24, 3 (2005), 1.

[115] Lucía Yan Wu and Isabel Segura-Bedmar. 2025. AI-generated Text Detection with a GLTR-based Approach. *arXiv preprint arXiv:2502.12064* (2025). doi:10.48550/arXiv.2502.12064

[116] Tian Yang, Yang Wang, and Weiyu Zhang. 2021. Effects of knowledge and reflection in intrapersonal deliberation. *Journal of Deliberative Democracy* 17, 1 (2021). doi:10.16997/10.16997/jdd.964

[117] Shunyu Yao, Qingqing Ke, Kangtong Li, Qiwei Wang, and Jie Hu. 2024. News GPT: A Large Language Model for Reliable and Hallucination-Controlled News Generation. In *Proceedings of the 2024 3rd International Symposium on Robotics, Artificial Intelligence and Information Engineering* (Singapore, Singapore) *(RAIIE '24)*. Association for Computing Machinery, New York, NY, USA, 113–119. doi:10.1145/3689299.3689320

[118] Chao Zhang, Kexin Ju, Peter Bidoshi, Yu-Chun Grace Yen, and Jeffrey M. Rzeszotarski. 2025. Friction: Deciphering Writing Feedback into Writing Revisions through LLM-Assisted Reflection. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 935, 27 pages. doi:10.1145/3706598.3714316

[119] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics* 12 (2024), 39–57. doi:10.1162/tacl_a_00632

[120] Weiyu Zhang, Tian Yang, and Simon Tangi Perrault. 2021. Nudge for Reflection: More Than Just a Channel to Political Knowledge. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 705, 10 pages. doi:10.1145/3411764.3445274

[121] Yunhao Zhang and Renée Gosline. 2023. Human favoritism, not AI aversion: People's perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation. *Judgment and Decision Making* 18 (2023), e41. doi:10.1017/jdm.2023.37