

ABSTRACT

Title of Dissertation: **WHEN NEWS HEADLINES GO WRONG:
AN IN-DEPTH ANALYSIS AND
AI-DRIVEN INTERVENTION OF
MISLEADING NEWS HEADLINES**

**Md Main Uddin Rony
Ph.D. Candidate, 2025**

Dissertation Directed by: **Dr. Naeemul Hassan
Philip Merrill College of Journalism and
College of Information Studies**

Misleading news headlines that distort, exaggerate, or omit information without presenting outright falsehoods pose a persistent challenge in the digital news ecosystem. These headlines often exploit commercial and algorithmic pressures, taking advantage of limited reader attention and heuristic processing. Despite their widespread impact, misleading headlines have received limited in-depth investigation in both misinformation research and HCI. This dissertation investigates the issue through a multi-method, three-part inquiry: examining human perceptions and correction practices (Project 1), testing the behavioral effects of headline correction strategies (Project 2), and evaluating large language models' (LLMs) capacity to support editorial reasoning (Project 3).

Project 1 explores how two key stakeholder groups, journalists and news readers, perceive and respond to misleading headlines. Through semi-structured interviews with 12 journalists and

12 readers, the study identifies competing notions of responsibility, with journalists emphasizing audience literacy and readers expecting inherent trustworthiness. The analysis surfaces three key correction strategies that stakeholders independently employ: adding uncertainty cues, restoring critical context, and removing emotional framing. These findings reveal editorial tensions and motivate the need to assess how such strategies function when deployed at scale.

Project 2 builds on these qualitative insights through a between-subjects experiment with 399 participants, testing the effects of the three correction strategies on reader outcomes. The study evaluates six headline versions across engagement, credibility, and interpretation accuracy. Results show that corrections, particularly the removal of emotional language, can significantly enhance perceived credibility and interpretive accuracy without diminishing engagement. The findings challenge the presumed trade-off between truthfulness and reader interest and offer empirical grounding for ethical headline design in journalism and platform interventions.

Project 3 investigates how LLMs such as GPT and Gemini explain misleadingness in headlines under varying levels of annotator agreement. Using a stratified dataset of 60 headlines and explanations generated by three LLMs, the study engages six professional journalists to evaluate explanation quality along editorial dimensions, including correctness, ambiguity awareness, and risk sensitivity. While LLMs align well with human reasoning in high-consensus cases, they often falter in ambiguous ones, failing to surface interpretive complexity or journalistic reasoning. The analysis informs design directions for editorially aligned, expert-in-the-loop AI systems.

Together, the three chapters advance a situated understanding of misleading headlines as a socio-technical problem and offer design-relevant implications for computational journalism, explainable AI, and platform governance. This dissertation highlights the need for editorial

transparency, role-aware collaboration, and systems that support nuanced, context-sensitive decision-making.

WHEN NEWS HEADLINES GO WRONG: AN IN-DEPTH ANALYSIS AND
AI-DRIVEN INTERVENTION OF MISLEADING NEWS HEADLINES

by

Md Main Uddin Rony

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2025

Advisory Committee:

Dr. Naeemul Hassan, Chair/Advisor
Dr. Jordan Boyd-Graber, Dean's Representative
Dr. Cody Buntain
Dr. Giovanni Luca Ciampaglia
Dr. Ronald Yaros

© Copyright by
Md Main Uddin Rony
2025

Dedication

To my beloved daughter, **Rumaaisa Amatullah**— whose first two years were marked by her father’s frequent absence. Your quiet resilience and bright smile have been my greatest motivation.

To my **wife**— whose steadfast support, patience, and unwavering belief made the impossible possible. Your unspoken encouragement carried me through every long night and early dawn.

To my **parents**— whose boundless love, prayers, and innumerable sacrifices lit every step of this journey. Everything I achieve rests on the foundation you built.

Acknowledgments

I would like to take this opportunity to express my heartfelt gratitude to my advisor, research committee members, collaborators, colleagues, friends, and family. Without their inspiration, encouragement, and unwavering support, this dissertation would not have been possible.

First and foremost, I am deeply grateful to my advisor, Dr. Naeemul Hassan, for giving me the opportunity to pursue my Ph.D. and for his steadfast guidance throughout this journey. His continuous encouragement, patience, and belief in my potential have not only shaped my academic trajectory but have also provided critical support during personal challenges. I could not have asked for a better mentor, and I am forever indebted to him for making this journey both meaningful and fulfilling.

I am sincerely thankful to my dissertation committee members, Dr. Cody Buntain, Dr. Giovanni Luca Ciampaglia, Dr. Ronald Yaros, and Dr. Jordan Boyd-Graber, for their insightful feedback, critical comments, and intellectual guidance. Their thoughtful engagement with my work has significantly contributed to improving the quality and clarity of this dissertation.

I would also like to extend my thanks to Dr. Enamul Hoque, Dr. Mohammad Yousuf, and again Dr. Ronald Yaros for their collaboration and mentorship in various research projects during my doctoral program. I have learned immensely from working with them. I am also grateful to my research collaborators Farhana Uddin, Saransh Grover, and Ahmed Shatil Alam,

whose contributions enriched many aspects of my research.

Special thanks go to all the members of the Computational Journalism Lab (CJLab) at the University of Maryland. I am particularly thankful to Mahfuz and Dr. Ali, who have served as informal mentors, motivating and supporting me through numerous hurdles. I am also grateful to my labmate and co-author Sophie, whose collaboration and assistance have been invaluable.

I appreciate the institutional support provided by the College of Information Studies (iSchool), including the Doctoral Student Research Award (DSRA), and the welcoming environment of the Computational Journalism Lab at the Philip Merrill College of Journalism, which became a second home throughout this journey.

This dissertation also benefited from the use of various computational tools and platforms. I acknowledge the support of large language models and digital infrastructure such as OpenAI's GPT, Grammarly, Google Colab, Qualtrics, and several Python and R libraries that facilitated data analysis and writing.

On a personal note, I am forever indebted to my wife and life partner, Khadija, for her boundless patience, love, and sacrifice. Your unwavering support and belief in me, especially while raising our daughter during my long academic hours, made this achievement possible. No words can adequately express the depth of my gratitude. With all my heart, I thank my beloved daughter, Rumaaisa. During the years when you needed me most, I was often absent, and yet you patiently endured. This dissertation belongs to both of you as much as it belongs to me.

I owe my deepest gratitude to my mother, Rokeya Akhter, and my late father, Md. Jaism Uddin, for their countless sacrifices, enduring patience, and unconditional support. I wish my father were here to witness this milestone. I am also thankful to my brother, Md. Mahiuddin, for standing by my family in Bangladesh while I pursued my dreams abroad. Heartfelt thanks

go to my in-laws, whose continued support and understanding meant a great deal throughout this journey.

Above all, I am profoundly thankful to Allah for granting me the strength, patience, and perseverance to complete this work. Alhamdulillah.

Table of Contents

Dedication	ii
Acknowledgements	iii
Table of Contents	vi
List of Tables	x
List of Figures	xii
Chapter 1: Introduction	1
1.1 Conceptualization of Misleading News Headline	2
1.1.1 Defining Misleading News Headline	3
1.1.2 Related concepts of Misleading News Headline	3
1.1.3 Distinguishing Misleading News Headlines from Established Misinformation Types	4
1.2 Structure of the Project	6
1.3 Organization of the Dissertation	7
Chapter 2: Literature Review	8
2.1 Function and Purpose of News Headlines	8
2.1.1 News Headlines as Summarizers and Attention Seekers	8
2.1.2 Balancing Summarization and Attention-Seeking	9
2.1.3 Challenges in Crafting Informative and Engaging Headlines	9
2.2 News Headline and Audience Engagement	10
2.2.1 Psychological and Emotional Drivers of Engagement	10
2.2.2 Impact of Headlines on Reading, Sharing, and Commenting	11
2.3 The Phenomenon of Misleading Headlines	12
2.3.1 Defining and Framing Misleading Headlines	12
2.3.2 Misleading Headlines as Subjective Misinformation	13
2.3.3 Linguistic and Structural Features of Misleadingness	14
2.3.4 Psychological Mechanisms Underlying Misleadingness	15
2.3.5 Impacts of Misleading Headlines	16
2.4 Responding to Misinformation at the Individual Level	17
2.4.1 Cognitive and Behavioral Responses to Misinformation	18
2.4.2 Influence of Beliefs and Ideologies	18
2.5 Correcting Misinformation and Misleading Content	19

2.5.1	Correction of Misleading Headlines	19
2.5.2	Correction in Journalism	20
2.5.3	General Misinformation Correction Strategies	22
2.6	Computational Approaches to Misinformation Mitigation	26
2.6.1	Detection of Misleading Headlines: Computational Approaches	26
2.6.2	Computational Approaches to Fake News, Clickbait, and Other Misinformation	27
2.7	Human–AI Collaboration in Subjective and High-Stakes Domains	29
2.7.1	Domains and Mechanisms of Successful Collaboration	29
2.7.2	Collaboration in Ambiguous or Subjective Contexts	29
2.7.3	Human–AI Collaboration in Misinformation Prevention	30
2.7.4	Human–AI Collaboration in Journalism	30
2.8	Explainable AI in Collaborative and Subjective Contexts	30
2.9	Human Annotation Disagreement as a Signal, Not Noise	31
Chapter 3:	Perceiving and Correcting Misleading Headlines: Perspectives from Journalists and News Readers	33
3.1	Introduction	33
3.2	Method	35
3.2.1	Preliminary Study	35
3.2.2	Semi-Structured Interviews	36
3.3	Findings	38
3.3.1	Understanding Misleading Headlines: Perceptions, Pressures, and Reader Reactions	38
3.3.2	Correction in Practice: How Journalists and Readers Revise Misleading Headlines	42
3.4	Discussion	44
3.4.1	The Need to Recognize Misleading Headlines as a Distinct Misinformation Threat	44
3.4.2	The Conflict Between Journalists and News Readers Over Misleading Headlines	45
3.4.3	Digitalization and the Evolution of Misleading Headlines	45
3.4.4	Editorial Safeguards and the Limits of Manual Oversight	46
3.4.5	Correction as Design: How Journalists and Readers Modify Misleading Headlines	47
3.4.6	Toward Human–AI Collaboration in Editorial Practice	47
3.4.7	Limitations	48
3.4.8	Future Research Directions	49
3.5	Conclusion	50
Chapter 4:	Examining the Behavioral Impact of Correcting Misleading News Headlines	51
4.1	Introduction	51
4.1.1	The Effects of Headline Correction on Reader Engagement	51
4.1.2	Effects of Corrective Strategies on Perceived Credibility	53
4.1.3	Effects of Corrective Strategies on News Headline Interpretation	54

4.2	Methods: Online Experiment	55
4.2.1	Study Design	55
4.2.2	Experimental Conditions: Headline Correction Strategies	55
4.2.3	Variables	58
4.3	Result	60
4.3.1	Reader Engagement (RQ1)	60
4.3.2	Perceived Credibility (RQ2)	64
4.3.3	Interpretation Accuracy (RQ3)	67
4.4	Discussion	70
4.4.1	Summary of the Key Findings	70
4.4.2	Implications	71
4.4.3	Limitations	72
4.5	Conclusion	73
Chapter 5: Human–AI Editorial Collaboration: A Situated Evaluation of LLM Explanations for Misleading News Headlines		74
5.1	Introduction	74
5.2	Method	77
5.2.1	Dataset	77
5.2.2	Expert Evaluation Study Design	80
5.2.3	Comparative Annotation of Human and LLM Explanations	82
5.2.4	Data Analysis	82
5.3	Result	84
5.3.1	Overview of Expert Ratings and Inter-Rater Reliability	85
5.3.2	Impact of Annotator Agreement on LLM Accuracy and Explanation Quality (RQ1)	86
5.3.3	Divergence of LLM Explanations from Expert Reasoning in Ambiguous Cases (RQ2)	88
5.3.4	Predicting Expert Trust and Usefulness from Explanation Characteristics (RQ3)	94
5.3.5	Comparing Human and LLM Explanations Across Editorial Concerns (RQ4)	97
5.4	Discussion	102
5.4.1	Subjectivity, Disagreement, and Situated Evaluation	104
5.4.2	Alignment and Limits of LLM Explanations	105
5.4.3	Design Implication	106
5.4.4	Limitations and Future Directions	112
5.5	Conclusion	113
Chapter 6: Discussion, Limitation and Future Work		114
6.1	Implications	115
6.1.1	Reframing Misinformation: Beyond Factuality Toward Interpretive Misleadingness	115
6.1.2	Rethinking Media Trust and Responsibility in a Platform-Driven Ecosystem	116
6.1.3	The Role of Ambiguity in Human-AI Collaboration	117

6.1.4	Understanding the Impact of AI on News Ecosystems	117
6.1.5	Designing for Editorial Collaboration in the Age of Generative AI	118
6.2	Future Work	119
6.2.1	Expanding Genre and Platform Diversity	119
6.2.2	Editorial Constraints: SEO Optimization and Paywall Design	119
6.2.3	Longitudinal Studies on Trust and Corrective Framing	119
6.2.4	Prototyping and Real-World Testing of Human–AI Editorial Collaboration	120
6.3	Limitations	120
Chapter 7:	Conclusion	124
Appendix A:	Additional details of Chapter 3	125
Appendix A:	Additional details of Chapter 4	129
Appendix A:	Additional details of Chapter 5	136
Bibliography		151

List of Tables

4.1	Descriptive Statistics for Self-Reported Engagement (IRT Score)	61
4.2	Logistic Regression Predicting Behavioral Engagement (Target Selection)	62
4.3	Descriptive Statistics for Perceived Credibility (IRT Score)	65
4.4	ANCOVA Results Predicting Perceived Credibility	65
4.5	Linear Regression Predicting Perceived Credibility	66
4.6	Descriptive Statistics for Interpretation Accuracy	67
4.7	ANCOVA Results Predicting Interpretation Accuracy	68
4.8	Linear Regression Predicting Interpretation Accuracy	68
5.1	Example Headline and Corresponding Labels and Explanations from Human and LLMs	78
5.2	Performance of LLMs in Detecting Misleading News Headlines	79
5.3	Demographic and Professional Backgrounds of Expert Participants	81
5.4	Mean and SD of Expert Ratings across Seven Explanation Quality Dimensions, by LLM and Agreement Category	88
5.5	Mann-Whitney U Test Results Comparing Unanimous vs. Ambiguous Cases by Dimension and LLM	89
5.6	Mean and Standard Deviation of Expert Ratings for Ambiguous Headlines (by LLM)	92
5.7	Wilcoxon signed-rank test results comparing expert evaluations of GPT and Gemini explanations across seven dimensions for ambiguous headlines, with corresponding rank-biserial effect sizes.	93
5.8	Predictive power of editorial dimensions on trustworthiness and actionability across three analysis methods. Boldface indicates top contributors within each method.	96
5.9	Spearman correlations (ρ) between editorial dimensions and expert outcomes, stratified by agreement level.	97
5.10	Mean (M) and standard deviation (SD) of expert ratings across four editorial dimensions, grouped by explanation type.	97
5.11	Mean and standard deviation of expert ratings across four editorial dimensions, stratified by agreement level and explanation type.	99
5.12	Friedman and Wilcoxon test results by agreement level and editorial dimension.	100
5.13	Wilcoxon signed-rank test results comparing explanation types across each editorial dimension. Statistically significant results ($p < .05$) are bolded.	101

5.14	Illustrative contrasts between top- and bottom-rated explanations across dimensions.	103
A.1	News Readers Demographics	127
A.2	Journalist Demographics	128
A.3	Distribution of collected articles by sources	128
A.4	Distribution of news articles by topics	128
A.5	Categories for misleading headlines developed through thematic analysis from articles collected between the authors	129
A.1	Headline Pairs and Correction Strategies	136
A.3	Final Selected Headlines for Expert Evaluation Study. Each row presents one of the 15 headline-article pairs selected for expert analysis, reflecting different combinations of human annotator agreement and LLM classification agreement.	144
A.1	Expert Evaluation Questions and Supplementary Clarifications	146
A.2	Structured Open-Ended Questions for Expert Evaluation	147
A.4	Likert Scale for Coding Journalistic Concerns in Explanations	147
A.5	Spearman correlation coefficients (ρ) and associated p -values between explanation dimensions and expert ratings of trustworthiness and actionability.	147
A.6	Multiple linear regression results predicting expert ratings of trustworthiness and actionability from editorial dimensions. Statistically significant predictors ($p < .05$) are bolded.	148
A.7	Mean and SD of Expert Ratings across Seven Explanation Quality Dimensions, by LLM and Agreement Category	149
A.8	Mann-Whitney U Test Results Comparing Unanimous vs. Ambiguous Cases by Dimension and LLM	150
A.9	Descriptive Statistics for Expert Ratings Across Seven Evaluation Dimensions	150
A.10	Intraclass Correlation Coefficient (ICC(2,1)) for Expert Agreement by Dimension	150

List of Figures

4.1	Self-reported engagement index by version and correction strategy. Error bars indicate 95% confidence intervals.	63
4.2	Perceived credibility (IRT score) by headline version and correction strategy. Error bars represent 95% confidence intervals.	64
5.1	Mean weighted accuracy of Gemini and GPT-4 across two agreement categories: ambiguous and unanimous. Error bars represent ± 1 standard deviation. Both LLMs demonstrate higher alignment with human annotations in the unanimous (high agreement) condition, while performance drops notably under ambiguity. GPT-4 consistently outperforms Gemini across both categories, particularly for unanimous headlines	87
5.2	Paired expert ratings for GPT and Gemini explanations across seven quality dimensions for ambiguous headlines (i.e., headlines with partial human annotator agreement). Each line connects the scores given to both LLMs for the same headline. The visual symmetry and overlapping trends illustrate that experts rated the explanations comparably, with no consistent pattern favoring one model. These findings align with the statistical results indicating no significant differences in explanation quality between GPT and Gemini under ambiguous conditions.	90
5.3	Distribution of difference scores (GPT - Gemini) in expert ratings across seven explanation quality dimensions for ambiguous headlines. Values above zero indicate higher GPT scores; values below zero indicate higher Gemini scores. Most differences cluster around zero, suggesting no systematic model advantage.	91
5.4	Mean expert ratings across editorial dimensions, comparing explanations written by humans, Gemini, and GPT. Error bars are not shown due to overlap in standard deviations.	98
5.5	Mean expert ratings across editorial dimensions for each explanation type, stratified by agreement level. Human explanations are included only for ambiguous headlines, as unanimous cases were exclusively evaluated for LLM outputs.	99
5.6	Proposed Human-AI Collaboration Workflow for Headline Writing and Verification	111
A.1	Behavioral Engagement: Headline Selection at the beginning of the experiment .	131
A.2	Instruction screen if participant selects the targeted headline	131

A.3	Instruction screen if participant does NOT select the targeted headline	132
A.4	Screen of the Engagement measurement (Partial)	133
A.5	Screen of the Perceived Credibility measurement (Partial)	134
A.6	Screen of the Interpretation measurement (Partial)	135
A.1	Expert Evaluation of LLM Explanation: Headline - Content View	140
A.2	Expert Evaluation of LLM Explanation: Explanation View	141
A.3	Expert Evaluation of LLM Explanation: Likert Scale Rating (Sample 1)	142
A.4	Expert Evaluation of LLM Explanation: Likert Scale Rating (Sample 2)	143
A.5	Expert Evaluation of LLM Explanation: Open-ended QA	144

Chapter 1: Introduction

In today's fast-paced digital age, where information is rapidly disseminated, news articles (with headline and body text) are a common form of information delivery. News headlines are crucial, as they play a key role in grabbing readers' attention and summarizing the news content [104, 145]. News headlines are precursors to comprehensive stories and serve as persuasive messages, making their accuracy and authenticity crucial. Gabielkov et al. note that many readers may not proceed beyond the headlines to read the full content; however, they can still receive misleading information if these headlines do not accurately represent the content [136]. We use the term Misleading News Headlines to describe this particular phenomenon. Misleading News Headlines arise when the headline of a news article fails to represent its content accurately. Consider the following example for illustration.

Headline: Hot tea linked to increased risk of esophageal cancer ¹

Content: People who like hot tea may want to wait until it gets cooler before taking that first sip. Drinking more than 700 milliliters of tea at higher than 60 degrees Celsius, or 140 degrees Fahrenheit, was linked to a 90 percent increased risk of esophageal cancer, according to a study ...

“Many people enjoy drinking tea, coffee, or other hot beverages. However, according to our report, drinking very hot tea can increase the risk of esophageal cancer,” said lead author Farhad Islami, a researcher at the American Cancer Society and study lead author, in a news release. ... In 2016, the International Agency for Research on Cancer said that drinking any drink over 65 degrees Celsius makes it a carcinogen or something likely to cause cancer. Other studies have linked drinking hot tea and drinking excessive amounts of alcohol daily to esophageal cancer, as well.

The headline *Hot tea linked to increased risk of esophageal cancer* is misleading because it specifically singles out hot tea, despite the article indicating that the risk is associated with consuming any very hot beverage. This narrow focus on hot tea could lead readers to incorrectly believe that only hot tea poses this cancer risk, potentially causing them to overlook the similar risks associated with other hot beverages. Consequently, readers might make uninformed decisions about their beverage choices, erroneously assuming that switching from hot tea to another hot drink, like coffee, would mitigate their risk of esophageal cancer when the temperature, not the type of beverage, is crucial.

Unlike many studies that frame misinformation as disinformation, deliberate attempts to mislead, this dissertation emphasizes the unintentional or ambiguous nature of misleading headlines. By doing so, it positions Misleading News Headlines as a form of misinformation

¹<https://tinyurl.com/misleading-headline-example1>

rather than disinformation. This framing reflects an intent not to ascribe malice or ill-intent to journalists, but instead to highlight structural, linguistic, and editorial practices that can mislead even when factual accuracy is maintained [418]. For instance, while a headline like “Hot tea linked to increased risk of esophageal cancer” may not stem from deceptive intent, its partial representation can still yield interpretive confusion. While it is acknowledged that misleading headlines can also be crafted with malicious or strategic intent, as in the case of partisan media manipulation or agenda driven distortion, this study intentionally limits its scope to the unintentional creation of misleadingness. The aim is to better understand the subtleties of how editorial and linguistic choices contribute to misrepresentation, without presupposing actor intent.

People often skim the headlines instead of reading the full articles before being drawn to whatever piques their interest [136]. Thus, if the headline is misleading, it may cause a wrong impression, leading to uninformed decision-making. Moreover, Ecker et al. found that even after reading an article in full, a reader is likely to relate to his or her first impression gained by the headline [111]. Therefore, the proliferation of Misleading News Headlines has emerged as a significant concern, potentially leading to misinformation and a distorted understanding of current events [418].

The impact of Misleading News Headlines can be far-reaching, affecting both news producers and consumers and undermining the credibility of journalism as a whole. In addition, the problem has worsened due to the quick growth of digital news platforms [13]. With the advent of social media and the ease of sharing news articles, headlines have become crucial in determining whether readers engage with the full content or merely skim through the headline itself. Moreover, research suggests that individuals tend to rely heavily on headlines when forming judgments about the accuracy and relevance of news content [227]. Studies have also shown that individuals are likelier to engage with content that aligns with their pre-existing beliefs and that misleading headlines can reinforce these biases [28].

Furthermore, online news consumption has brought new challenges in distinguishing reliable sources from unreliable ones [398]. The democratization of news production and the ease of sharing information have allowed misinformation to spread quickly, leading to a decline in trust in the media. As a result, addressing the issue of misleading news headlines has become paramount in restoring public trust in journalism and reducing the spread of misinformation.

This dissertation investigates how news consumers and producers perceive and interpret Misleading News Headlines. It examines the nature of these interactions to uncover how headlines can be improved without compromising their appeal, thereby preserving trust in news content. By exploring the interplay between perceptual accuracy, editorial integrity, and audience engagement, the dissertation provides insights into the design of ethical human–AI collaboration for misinformation mitigation in the digital era.

1.1 Conceptualization of Misleading News Headline

This section aims to conceptually explore Misleading News Headlines, a critical and distinct form of misinformation in today’s media landscape. We will formally define the phenomenon based on existing literature, examine the varied terminologies and conceptual frameworks employed by scholars, and delineate how Misleading News Headlines differ from

established misinformation types(e.g., fake news, clickbait, etc.). This analysis is intended to clarify the unique nature and implications of Misleading News Headlines, highlighting their significant impact on public perception and discourse in the digital era.

1.1.1 Defining Misleading News Headline

Within the information disorder framework, as Wardle et al. outlined, a critical subset is identified as *False Connection*, particularly pertinent to the discussion of Misleading News Headlines [418]. This concept addresses the misalignment between a news headline and its corresponding article content.

Misleading News Headlines, as a specific manifestation of *False Connection*, are headlines that do not faithfully encapsulate the news article’s main content or critical nuances. As exemplified in various hyper-partisan sites, these headlines are crafted to capture immediate attention and evoke strong reactions, often at the expense of accurately representing the article’s content.

Considering this, the following definition can be proposed:

Misleading News Headline refers to a headline that creates a ‘False Connection’ by not accurately representing the corresponding news article’s main content or essential nuances. It is characterized by factors such as, but not limited to, exaggerations, emotional appeals, or distortions, all of which can mislead the reader about the true context or content of the article.

This definition attempts to capture the broad nature of misleading news headlines, acknowledging that the factors causing such misalignment can be varied and complex.

1.1.2 Related concepts of Misleading News Headline

Different researchers have interpreted the concept of *Headline Incongruency* in digital journalism in various ways. Chesney et al. initially defined it as headlines that do not accurately represent the information contained in the article [77]. This broad definition sets the groundwork for subsequent interpretations and applications of the concept.

Piotrkowicz et al. viewed incongruity when the headline presents a biased view of the main content, suggesting that headlines are created subjectively [311]. Brogly et al. then identified headline incongruency as a key indicator for detecting clickbait, suggesting a direct correlation between the allure of a headline and its potentially misleading nature [56]. However, Chesney et al. had already distinguished this issue from clickbait detection, noting that it requires different Natural Language Processing (NLP) approaches [77].

Furthermore, Zhang et al. approached headline incongruency from the perspective of stance detection, framing it as a task of identifying the stance of the headline relative to the article’s content [437]. This approach, however, does not encompass scenarios where the content supports the headline’s claim but remains incongruent in how it is presented or contextualized.

In contrast to these varying interpretations, Wei et al. defined “Misleading Headline” as a headline whose meaning differs from that of the story’s content. [420] This definition addresses a specific issue: the misrepresentation or distortion of content, regardless of topical consistency or stance alignment.

Given these varied perspectives, this thesis focuses on the term Misleading News Headlines. This decision is made to highlight those headlines that, while potentially topically consistent and supportive of the article’s stance, can still mislead by distorting or skewing the content. This redefined focus aims to more accurately reflect the real-world nuances of how headlines can mislead readers beyond the scope of what is traditionally understood as headline incongruity.

1.1.3 Distinguishing Misleading News Headlines from Established Misinformation Types

In this section, we explore and contrast Misleading News Headlines with well-established forms of misinformation, such as fake news and clickbait. This comparison aims to highlight the unique characteristics of Misleading News Headlines, underscoring their distinct nature within the spectrum of misinformation. By delineating these differences, we emphasize the need for specialized attention from the research community to address the challenges posed by Misleading News Headlines, recognizing their specific impact on public discourse and information accuracy.

Misleading News Headlines vs Fake News

Misleading News Headlines significantly differ from Fake News in several key aspects, primarily in their linguistic features and the nature of the content they present.

Fake news typically involves creating and disseminating false information or hoaxes, often designed to deceive or mislead readers [13]. The content, including the headline, in fake news is fabricated or significantly altered from the truth. An example of a fake news headline would be “Heat, garlic, and ginger can prevent COVID-19”² when there is no scientific research to support that claim.

On the other hand, Misleading News Headlines usually involve a twist or distortion of real facts [77]. These headlines are often based on actual events or data but are presented in a way that misrepresents or skews the underlying story. For instance, the headline “Hot tea linked to increased risk of esophageal cancer,” while not entirely false, exemplifies a *Misleading News Headline* as it suggests that hot tea specifically increases cancer risk without mentioning that the risk is associated with all very hot beverages and not solely hot tea. This contrasts Fake News, which typically involves a fabricated story, such as a headline falsely claiming a medical breakthrough that never occurred. Though rooted in factual research, the “hot tea” headline skews the information by not presenting the full context, thereby misleading readers. This distinction is crucial in the exploration of “Fake News vs. Misleading News Headlines,” highlighting the subtler nature of misleading headlines compared to the outright falsehoods often seen in fake news.

Misleading News Headlines vs Clickbait

Misleading News Headlines and Clickbait, while overlapping in their capacity to mislead, differ significantly in their intent, structure, and impact.

²<https://www.rappler.com/newsbreak/fact-check/264621-heat-garlic-ginger-cure-covid-19/>

Clickbait headlines are designed to attract attention and entice readers to click on a link, often using sensationalized, provocative, or extremely intriguing phrases. They typically promise more than the article delivers or use hyperbole to exaggerate its content [42]. For example, a clickbait version of the “hot tea” headline might read: “Doctors Shocked: Drinking Hot Tea Might Kill You!” This headline sensationalizes the information to an extreme degree, aiming to exploit the reader’s curiosity and emotional response.

In contrast, Misleading News Headlines, such as the earlier example, “Hot tea linked to increased risk of esophageal cancer,” may not always be designed to be sensationalist in the same way as clickbait. They can mislead by presenting a skewed interpretation of the content, omitting key details, or exaggerating a part of the truth. The “hot tea” headline, while not employing the hyperbolic or sensational language typical of clickbait, still misleads by implying that the risk is specific to hot tea, whereas the actual research refers to very hot beverages in general.

The crucial distinction lies in the fact that while all clickbait headlines are misleading (due to their exaggerated or sensational nature), not all misleading headlines are clickbait. Many headlines may mislead through subtler means, such as selective presentation of facts or lack of context, without necessarily being sensationalist or designed to maximize clicks [72]

This thesis focuses on encompassing all forms of Misleading News Headlines, regardless of whether they fall under clickbait. This broader approach acknowledges that Misleading News Headlines can arise from various linguistic and structural elements, not limited to the attention-grabbing tactics of clickbait. Therefore, addressing Misleading News Headlines in their entirety, beyond the scope of clickbait, is essential for developing comprehensive strategies to counter misinformation effectively.

Misleading News Headlines vs Other Misinformation Types

Misleading News Headlines, while sharing certain characteristics with other misinformation types, have distinct features that set them apart and necessitate focused attention.

- **Satire or Parody.** Unlike Misleading News Headlines, which may unintentionally distort facts, satire or parody intentionally uses humor or exaggeration for critique or entertainment. Although not meant to deceive, satire can be misconstrued as factual, leading to misinformation [387].
- **Propaganda.** This misinformation is primarily used to influence public opinion or promote a political agenda with a biased or misleading nature [205]. In contrast, Misleading News Headlines may not always have a political motive and can arise from subtler forms of misrepresentation.
- **Rumors.** These typically involve unverified information and are not necessarily linked to a credible source or a news article [100]. Misleading News Headlines, however, are associated with news content and may contain elements of truth, albeit presented in a distorted way.
- **Conspiracy Theories.** Often lacking empirical evidence, conspiracy theories propose explanations based on hidden agendas and powerful actors [381]. In contrast, Misleading

News Headlines usually stem from factual content but are misleading through their presentation or lack of context.

Each type of misinformation contributes uniquely to the spread of false or misleading information. However, Misleading News Headlines demand particular attention due to their ability to subtly distort truth within a factual framework, making them more challenging to identify and correct than more overt forms of misinformation. This subtle distortion can have significant implications, as headlines are often the primary means audiences engage with news content. Understanding and addressing the nuances of Misleading News Headlines are thus crucial in the broader fight against misinformation.

1.2 Structure of the Project

This dissertation investigates the multifaceted problem of misleading news headlines, headlines that distort, exaggerate, or omit key information, through three interconnected research chapters. Each chapter addresses a distinct layer of the issue, progressing from human perception and editorial judgment to experimental validation of corrective strategies and ultimately to the computational evaluation of large language model (LLM) explanations, offering design insights for future human–AI collaboration in editorial contexts.

Chapter 2 presents a comprehensive literature review of the social, cognitive, and technological dimensions of misleading headlines. It synthesizes relevant work in journalism studies, human–computer interaction, misinformation detection, and explainable AI. This chapter identifies gaps in current approaches and motivates the need for a multi-stakeholder, multi-method investigation.

Chapter 3 explores how both news producers (journalists) and consumers perceive misleading headlines. Through qualitative interviews, this chapter examines how stakeholders define misleadingness, identify problematic headlines, and propose corrections. The findings uncover critical tensions between editorial intent and audience trust, and reveal how institutional norms and reader expectations shape headline interpretation.

Chapter 4 expands on these insights by examining the impacts of three corrective strategies: adding cues of uncertainty, providing critical context, and eliminating emotionally charged language, on reader engagement, interpretation accuracy, and perceived credibility. Using a between-subjects experimental design, the chapter demonstrates that these strategies can improve comprehension and trustworthiness without sacrificing appeal, offering a viable path toward ethical and effective headline correction.

Chapter 5 investigates whether large language models can assist in identifying and explaining misleading headlines. Using a stratified dataset of ambiguous and unambiguous headlines, the study evaluates LLM-generated explanations through an expert-in-the-loop protocol involving professional journalists. The findings show that while LLMs can replicate surface-level consensus, they often struggle with editorial nuance, ambiguity handling, and context sensitivity, highlighting both the promise and the limitations of AI in editorial decision-making.

Together, these chapters offer a layered investigation into the nature, correction, and automation of misleading headline detection, contributing to journalism, HCI, and computational misinformation research.

1.3 Organization of the Dissertation

This dissertation begins with an introduction outlining the problem of misleading headlines, their implications, and the overarching research goals. It is followed by Chapter 2, which presents a detailed literature review covering existing work on misinformation, editorial practices, reader cognition, and the use of AI in journalistic contexts. The core of the proposal consists of three empirical chapters: (1) a qualitative study on stakeholder perceptions of misleadingness, (2) an experimental study testing the effectiveness of correction strategies, and (3) an expert-informed evaluation of LLM-generated explanations.

Chapter 2: Literature Review

In this dissertation, we aim to unravel and address the phenomenon of misleading news headlines by examining their creation, perception, and potential mitigation strategies. This exploration navigates the intricate relationship between crafting and interpreting news headlines and the pivotal role of journalistic integrity and technological advancements in shaping public discourse. It emphasizes the crucial need for ethical considerations and accuracy in news production, particularly in the age of rapid digital media consumption and the proliferation of misinformation.

2.1 Function and Purpose of News Headlines

This section will highlight news headlines' fundamental roles and functions within the media landscape. It will explore the dual nature of headlines as both summarizers of content and critical drivers of reader engagement, examining how they navigate the delicate balance between accurately conveying news stories and capturing audience attention. The challenges and intricacies of crafting headlines that fulfill these dual roles effectively will also be critically analyzed.

2.1.1 News Headlines as Summarizers and Attention Seekers

News headlines serve a dual function in contemporary journalism: they summarize complex stories while competing for readers' attention [104]. Often described as the "abstract of an abstract" [27], headlines distill the essence of an article into a brief, digestible form that allows audiences to quickly assess the relevance of a story [27, 284]. In digital environments saturated with information, this summarizing role becomes even more vital, enabling efficient content navigation and consumption [284].

The rise of natural language processing (NLP) has amplified the importance of summarization in headline generation. Recent advancements in extractive and abstractive techniques have enabled automated systems to produce headlines that preserve the core meaning of articles while maintaining fluency and contextual alignment [309]. These approaches, deployed across both high- and low-resource languages, emphasize informativeness and accuracy alongside brevity [284, 309].

Concurrently, headlines are crafted to act as powerful attention-seeking tools. Journalists and editors often employ rhetorical strategies, emotional appeals, or even mild sensationalism to ensure visibility in competitive media spaces [27, 48]. These pragmatic choices can shape reader perception and determine which stories gain traction, particularly on algorithm-driven platforms [48]. However, the pursuit of attention can have consequences: emotionally charged

or provocative language may increase click-through rates but also elevate the risk of spreading misleading or toxic content [135, 325]. Studies show that headlines from less credible sources are more likely to use such language, raising ethical concerns about the balance between engagement and journalistic responsibility [135, 325].

Linguistically, headlines form a distinctive genre characterized by compactness, creativity, and strategic ambiguity [27, 48]. They often rely on ellipsis, punning, or lexical economy to convey maximum impact in minimal space [48]. From a computational perspective, generating headlines that are both succinct and semantically faithful remains a challenge, one that recent language models such as GPT-2 have begun to address with increasing sophistication [116].

2.1.2 Balancing Summarization and Attention-Seeking

Striking a balance between these primary roles has become increasingly complex in the digital news ecosystem, where attention is a scarce resource and algorithms reward engagement over accuracy. A well-crafted headline must convey the article’s core message clearly and truthfully, acting as a concise summary [104, 405]. However, headlines are also designed to maximize clicks, shares, and visibility, leading many to adopt emotionally charged, ambiguous, or sensational language that may distort the article’s content [42, 346].

This tension between informativeness and engagement is at the heart of ongoing research in both journalism studies and computational modeling. For instance, Sepúlveda-Torres et al. developed *HeadlineStanceChecker* to detect inconsistencies between headlines and article bodies, aiming to catch misleading cues used for ideological or commercial purposes [360]. Others, like Omidvar et al., use transformer-based models to generate headline variations and predict their popularity, selecting those that optimize attention without entirely sacrificing accuracy [293]. Meanwhile, Zhan et al. [436] and Jin et al. [199] propose headline generation techniques that incorporate stylistic constraints, allowing for greater control over tone and engagement features.

Although these computational strategies demonstrate promise, they also underscore the ethical tension in headline writing. As emotionally manipulative or misleading headlines can fuel misinformation and erode trust, researchers are increasingly focused on interventions that preserve both fidelity and appeal. This dissertation builds on that tradition by examining whether corrective strategies, such as adding context, softening emotional language, or introducing uncertainty cues, can improve informativeness without reducing engagement. In doing so, it aims to inform both editorial best practices and AI-assisted headline generation models that seek to restore this delicate balance.

2.1.3 Challenges in Crafting Informative and Engaging Headlines

Crafting headlines that are both accurate and compelling is a persistent challenge for journalists, especially in the digital news environment where competition for attention is fierce. Newsrooms face increasing pressure to produce headlines that drive traffic and engagement, often leading to a trade-off between informativeness and click-worthiness. As Liu et al. [177] note, online editors frequently optimize headlines to maximize reader clicks, sometimes at the cost of distorting the article’s actual message or emphasis.

One central challenge lies in the space constraint: headlines must condense complex or nuanced stories into a few words, requiring difficult editorial decisions about which aspects to highlight and which to omit. This often results in partial representations of the article, leading to interpretive ambiguity or overgeneralization [333]. Moreover, while effective for capturing attention, stylistic techniques such as ellipsis, rhetorical questions, and sensational phrasing can mislead readers about the substance or tone of the underlying article [42, 346].

The issue is further complicated by audience segmentation and platform incentives. Headlines are increasingly crafted with platform-specific dynamics in mind; for instance, what works well on social media may not align with what promotes informed reading [289, 424]. Journalists may tailor headlines differently depending on perceived reader preferences, political leanings, or even device type [44]. These strategies, while practical, risk amplifying confirmation bias or selective exposure.

From a computational standpoint, automated headline generation presents additional challenges in balancing informativeness and appeal. Neural models trained to optimize for engagement metrics, such as click-through rates or social media shares, have been shown to prioritize attention-grabbing or emotionally charged elements, sometimes at the expense of factual accuracy or nuance. One well-documented issue is hallucination, where generated headlines include information not supported by the source text, often driven by the model's tendency to favor salience over fidelity when engagement signals dominate the training objective [362]. Similarly, personalization frameworks that tailor headlines to individual readers may overemphasize emotionally salient or confirmatory cues derived from user behavior, reinforcing biased or dramatized narratives [63]. These findings underscore the difficulty of designing systems that produce headlines that are both compelling and journalistically responsible.

2.2 News Headline and Audience Engagement

This section will synthesize various findings to illustrate how news headlines significantly impact audience engagement through psychological drivers, content, style, and platform-specific adaptations.

2.2.1 Psychological and Emotional Drivers of Engagement

Reader engagement with news content is deeply influenced by psychological and emotional mechanisms, including how headlines trigger emotional responses, arouse curiosity, or capture attention. This section reviews two central pathways through which headlines influence engagement: emotional appeal and curiosity-driven behavior.

Role of Emotion in News Engagement

Emotion plays a pivotal role in shaping audience responses to news headlines. Headlines that evoke strong emotions, particularly fear, anger, or moral outrage, are more likely to be clicked, shared, or commented on, especially in the context of polarizing or high-stakes topics like politics and climate change. For example, Xu et al. found that clickbait headlines in

climate reporting often relied on fear-based language to amplify engagement on social media platforms like Facebook and Twitter [429]. Similarly, Oh et al. showed that emotionally negative headlines tend to attract more clicks on news sites, while positive emotions increase sharing on Twitter [289]. These findings underscore the dual nature of emotional engagement: while emotional appeals can drive audience interest, they may also risk sensationalism or misrepresentation.

Importantly, emotionally charged headlines are not only more attention-grabbing but also influence downstream perception and memory of news content. Emotionally misleading headlines, those that exaggerate or distort for emotional impact, can impair interpretation accuracy and damage perceived credibility, as shown in experimental findings on headline corrections [192]. This highlights the ethical challenge of designing emotionally engaging headlines that do not compromise informational integrity.

Attention, Curiosity, and Click Behavior

Headlines also function as cues that activate curiosity and direct reader attention. Attention and curiosity are key precursors to user engagement, often triggered by information gaps or incongruent framing. Loewenstein’s information gap theory suggests that people are motivated to click when headlines evoke a sense of missing information [249]. This mechanism underpins the effectiveness of clickbait-style headlines, which raise questions or make surprising claims to compel readers to seek closure [31, 303].

Studies have shown that headlines using interrogative forms, novelty cues, or teasers are particularly effective in prompting user interaction [315]. At the same time, headlines that manipulate curiosity without providing informative content risk damaging trust and long-term engagement. For instance, Jiang et al. found that excessive use of sensational or vague headlines may lead to click regret, where readers feel misled after engaging [197]. Balancing curiosity with clarity remains a core challenge in headline design.

2.2.2 Impact of Headlines on Reading, Sharing, and Commenting

News headlines significantly influence audience engagement with stories, affecting whether readers choose to read an article, share it with others, or comment on it. This section reviews how specific headline features and social media contexts influence these engagement behaviors.

Headline Features Affecting User Behavior

The linguistic and stylistic features of headlines significantly affect user behavior across digital platforms. Banerjee et al. identified that lexical richness, informativeness, and novelty positively influence click-through rates, while excessive ambiguity or sensationalism may reduce trust [31]. Similarly, negative sentiment in headlines is associated with higher click rates on news websites, whereas positive headlines tend to be shared more frequently on social media [289].

Platform-adaptive editing strategies are increasingly common. Park et al. found that media outlets often rewrite the same headline to fit the engagement norms of different platforms, such as shortening length or increasing emotional tone for social media distribution [299]. These

findings show that even small stylistic changes, such as emotional framing or lexical emphasis, can significantly impact how readers react to a headline.

Furthermore, technical features such as text length and the inclusion of specific trigger words also play a role. Jiang et al. showed that shorter, emotionally intense headlines perform better in crowded news feeds, especially on mobile devices [197]. This suggests a growing tension between concise expression and preserving editorial integrity, particularly as algorithms continue to favor attention-maximizing content.

Social Media Dynamics and Virality

The architecture and algorithms of social media platforms amplify the role of headlines in shaping audience engagement. Headlines are often the only visible part of a story in feeds or trending lists, making their phrasing crucial for initial visibility and virality. Piotrkowicz et al. found that headlines with high news value and distinct linguistic styles are more likely to be shared on platforms like Twitter and Facebook [311, 312].

In social media environments, vividness and post format also matter. Schultz et al. reported that incorporating images or videos alongside headlines enhances user interaction, even compensating for longer headline or post length [353]. Additionally, conflict-based or emotionally divisive headlines are more likely to trigger comment activity, often leading to polarized discussions [109].

News organizations are increasingly adapting their headlines to platform-specific engagement logic. Lamot et al. documented how headline rewrites tailored for Facebook led to higher interaction rates than the original versions published on websites [231]. However, this optimization for virality can unintentionally encourage sensationalism or skewed interpretations, raising concerns about long-term effects on information quality and trust.

2.3 The Phenomenon of Misleading Headlines

This literature review comprehensively examines the multifaceted nature of misleading news headlines. It delves into their definitions, characterizations, underlying psychological mechanisms, societal impacts, and the advancements in automated detection and relevant datasets. The goal is to provide a holistic understanding of how misleading headlines are constructed, perceived, and combated in the digital age. This exploration is crucial for informing strategies to mitigate the spread of misinformation and enhance public media literacy, thereby safeguarding informed societal discourse.

2.3.1 Defining and Framing Misleading Headlines

Misleading headlines are commonly defined as those that distort, exaggerate, or otherwise misrepresent the content of the associated news article. They provide a biased representation of the story, steering readers toward a specific, and often misleading, interpretation while remaining factually plausible. This incongruity between the headline and the body text forms the conceptual basis of misleadingness. Such headlines may omit crucial context, exaggerate causality, or frame an event through emotionally charged or sensationalist language that is not substantiated by the article's content [77, 435].

Several studies refer to these as *incongruent headlines*, highlighting the semantic or pragmatic mismatch between a headline and the full article [267, 297]. In prior research, misleadingness has been examined in relation to headline “hallucination,” particularly in AI-generated content where the headline contains fabricated or irrelevant information that is not supported by the article body [361]. Whether manually written or machine-generated, misleading headlines pose serious interpretive challenges, particularly for readers who consume only the headline without reading the full article [394].

Misleadingness is not just a binary quality; it exists on a spectrum. Some headlines may contain blatant contradictions with the body text, while others may subtly distort meaning through emphasis, omission, or misleading framing techniques. These strategies can influence how readers interpret the information without necessarily violating the literal truth [111, 194]. This blurring of lines between accurate summarization and deceptive framing positions misleading headlines as a subtype of *subjective misinformation*, statements that are technically true but strategically misleading [232].

The framing function of headlines plays a pivotal role here. According to Entman’s framing theory, framing involves selecting certain aspects of reality and making them more salient to promote a specific problem definition, causal interpretation, or moral evaluation [118]. Misleading headlines often exploit this mechanism by emphasizing peripheral or emotionally provocative aspects of a story, thereby shaping initial impressions and constraining subsequent interpretation. As a result, they not only mislead through content but also through the strategic framing of what is omitted, emphasized, or suggested [87, 319].

Given that headlines are frequently the first and sometimes the only part of the news that readers engage with [136], the potential for misleadingness carries significant epistemic and social implications. Misleading headlines can skew public understanding, reinforce stereotypes, and distort issue salience, all while operating within the bounds of technical truth. This makes their detection and mitigation particularly difficult and calls for more nuanced definitions that account for pragmatic effects on readers, not just semantic content.

2.3.2 Misleading Headlines as Subjective Misinformation

Misleading headlines often occupy a gray area between truth and falsehood. Unlike explicit misinformation or fake news, these headlines do not necessarily contain factually incorrect information but instead distort meaning through selective emphasis, omission, or suggestive framing. This subtle manipulation results in what scholars refer to as subjective misinformation, claims that are technically true but likely to mislead based on their interpretation [9, 33]. In this sense, misleading headlines exemplify how truth can be bent without breaking it, creating interpretive discrepancies that can shape public opinion and memory.

A core characteristic of subjective misinformation is that its misleadingness depends on context, prior knowledge, and reader assumptions. For example, headlines that omit key qualifiers or exaggerate causality may not be objectively false but can still lead readers to draw inaccurate conclusions [111, 297]. This subjectivity complicates both computational and human evaluation: a headline judged misleading by one reader may appear truthful to another, particularly when motivated reasoning or ideological bias comes into play [237, 259].

Misleading headlines also challenge conventional binary labeling schemes in misinformation detection research. Most fact-checking systems and automated models focus

on detecting false claims, overlooking headlines that strategically frame true information in misleading ways [232, 394]. This highlights a critical gap in current misinformation taxonomies, which tend to prioritize verifiability over interpretability. Subjective misinformation, by contrast, demands evaluation not just of factuality but of intent, pragmatic effect, and reader response.

Moreover, because headlines are often the only part of an article that readers see or remember [43, 225], subtle distortions in this space carry disproportionate influence. The cumulative effect of repeated exposure to such content can reinforce inaccurate impressions, a process made more potent by psychological mechanisms like the illusory truth effect [124].

In this context, misleading headlines serve as a significant source of misinformation that is challenging to identify, more difficult to regulate, and particularly effective in environments filled with fast-moving, attention-seeking content. Their subjectivity underscores the need for both media literacy and nuanced computational tools that go beyond surface-level fact-checking.

2.3.3 Linguistic and Structural Features of Misleadingness

Misleading headlines often rely on specific linguistic and structural techniques to subtly distort meaning without introducing outright falsehoods. These features, ranging from selective omission to emotional framing, can influence readers' perceptions before they engage with the full article, or even in the absence of reading further. This section explores two major categories of such features: (1) omission, ambiguity, and framing bias; and (2) emotional and sensational language.

Omission, Ambiguity, and Framing Bias

One common technique for misleading readers is strategic omission, which involves leaving out crucial qualifiers, actors, or contextual details that would otherwise alter the interpretation of a headline. For example, omitting attribution (“Police shoot suspect” vs Suspect shot during police confrontation”) can dramatically shift perceived culpability or narrative tone. Such omissions create semantic gaps between the headline and article content, leading to misalignment that encourages misinterpretation [77, 111].

Ambiguity, too, is frequently leveraged. Headlines may use vague references, passive constructions, or lexical ambiguity to suggest a particular interpretation without making definitive claims. This allows for plausible deniability, where the headline appears technically accurate while also implying something misleading [267].

Relatedly, framing bias occurs when the headline selectively emphasizes certain aspects of a story while downplaying or ignoring others. Framing determines which actors appear central, which values are emphasized, and which angles are foregrounded [118]. Even in factually accurate headlines, biased framing can nudge readers toward specific conclusions, especially when paired with ideological or partisan cues [164]. These techniques are especially potent because many readers rely on headlines alone to form judgments [43].

Emotional and Sensational Language

Another hallmark of misleading headlines is the use of emotionally charged or sensational language, often deployed to maximize clicks or provoke strong reactions. These headlines rely

on exaggerated language, moral outrage, or hyperbole to amplify engagement while downplaying nuance or accuracy [30, 406]. Emotionally evocative phrasing can bypass analytic thinking, instead activating fast, intuitive responses that increase the likelihood of belief, sharing, or memory retention [306].

This technique is not exclusive to fringe outlets. Even mainstream media may resort to emotional framing to compete in attention economies [335]. For instance, headlines that invoke threat or outrage (e.g., “X is destroying America”) are more likely to go viral, particularly on platforms like Facebook and Twitter [402, 445]. However, these emotional cues can distort comprehension by prioritizing engagement over clarity or truth.

Studies show that emotionally manipulative headlines reduce reader skepticism and increase the perceived importance of the story, regardless of factual content [232]. This emotional framing contributes to the spread of misleading narratives, especially when combined with ambiguity or sensational structure. As a result, such headlines not only affect engagement metrics but also shape readers’ political attitudes and beliefs [127].

2.3.4 Psychological Mechanisms Underlying Misleadingness

Misleading headlines succeed not only due to structural or linguistic manipulation but also because they exploit core psychological mechanisms that shape how individuals process, remember, and respond to information. These mechanisms include inferential shortcuts, cognitive biases, emotional heuristics, and memory dynamics. Understanding these processes is critical to diagnosing the potency of misleading content and devising effective interventions.

Inferential Reasoning and Priming

Misleading headlines often shape interpretation by activating specific cognitive schemas before readers engage with the full article [111]. For example, when a headline frames an event using emotionally loaded or causal language (“X causes Y”), it can predispose readers to adopt a causal interpretation even when the body text provides more nuanced or contradictory evidence. This mechanism operates through inferential reasoning, where individuals fill in information gaps based on the headline’s cues rather than critically evaluating the complete content [124]. Studies show that readers’ mental models, once primed by a misleading headline, are difficult to correct, even when the article itself presents clarifications [83, 111]. This “sticky” nature of first impressions underlines the risk posed by subtly misleading summaries.

Exploiting Cognitive Biases and Susceptibility

Misleading headlines effectively exploit well-documented cognitive biases, particularly confirmation bias, the tendency to accept information that aligns with pre-existing beliefs, and the fluency effect, where repeated or familiar claims feel more truthful [1, 307]. These heuristics make readers more susceptible to distortion, especially in fast-paced digital environments where headlines serve as primary or sole exposure to news.

Additionally, psychological mechanisms such as motivated reasoning and expressive responding contribute to the acceptance of misleading information. Readers may endorse misleading headlines not because they believe the literal claims, but because those claims

align with their political or social identities, allowing them to express group affiliation [403]. Emotional arousal—particularly when induced by provocative wording that evokes fear, outrage, or moral indignation—can bypass critical reasoning processes and increase both belief in and the likelihood of sharing misinformation [304, 310]. These effects are moderated by individual susceptibility: research shows that lower levels of cognitive reflection, analytical reasoning, and science literacy, as well as certain psychological traits such as grandiose narcissism or anxiety, are correlated with increased vulnerability to misleading or ambiguous content [162, 278, 310].

The Illusory Truth Effect

One of the most persistent psychological phenomena aiding misleading headline belief is the illusory truth effect, the tendency to believe information as true simply through repetition. Even when statements are initially identified as false, repeated exposure increases perceived accuracy due to enhanced cognitive fluency [124, 167].

This effect is particularly problematic for misleading headlines because they are frequently encountered in isolation (e.g., via social media shares) and can be reinforced by subtle rewordings across outlets [143]. Studies show that even minimal exposure to misleading but fluent phrasing can increase later belief, especially when individuals are distracted, tired, or cognitively overloaded [123]. In an age of algorithmic curation, where similar narratives reappear frequently, the illusory truth effect amplifies the influence of misleading headlines over time.

Memory, Impression Formation, and Routine Cognition

Misleading headlines also exploit how memory and impressions are formed. Headlines create a gist memory, a simplified, general idea of the article, that can persist even when specifics are forgotten [83, 327]. Once a reader has formed an initial interpretation, it becomes difficult to revise, even in the presence of contradictory information.

Moreover, under routine cognitive processing, readers tend to default to acceptance rather than skepticism when encountering new information [147]. This automatic processing is particularly prevalent in casual browsing contexts, where cognitive effort is low. As a result, plausible but misleading headlines are often remembered as true or partially true, contributing to long-term misinformation effects even after retraction or clarification [111, 327].

Additionally, source credibility plays a role. When misleading headlines are published by reputable outlets, they gain added legitimacy, reinforcing belief and reducing scrutiny [327]. These factors collectively reveal how routine cognition facilitates the persistence of misleading interpretations, especially when readers rely solely on headlines.

2.3.5 Impacts of Misleading Headlines

Misleading headlines exert significant influence on both individual cognition and societal discourse. They distort readers' understanding of news events, erode trust in journalism, and contribute to the amplification of misinformation at scale. This section examines their consequences at two critical levels: the interpretive and trust-related effects on readers, and the broader societal and political ramifications.

Reader Misinterpretation and Trust

One of the most immediate consequences of misleading headlines is reader misinterpretation. When headlines are incongruent with the article body, readers often form impressions based on inaccurate cues, leading to flawed judgments, false memories, or incomplete understanding [111, 435]. Because many readers skim headlines without engaging with full articles, the misleading framing becomes the dominant takeaway, one that can persist even when the article content provides more nuance or correction.

This misalignment undermines epistemic trust, especially when readers later realize that the full article contradicts the headline's implication. Over time, repeated exposure to misleading headlines can diminish overall trust in news organizations and journalism as a profession [267]. The credibility of even reputable outlets is threatened when editorial decisions prioritize engagement or sensationalism over accuracy [77]. Moreover, misleading headlines shared on social media, often divorced from article context, further skew public perception by reinforcing biases and contributing to confirmation-based reading [375].

In science journalism, for instance, misleading headlines that oversimplify or exaggerate scientific claims can generate public misunderstanding of scientific consensus, erode trust in researchers, and inflate skepticism toward legitimate findings [271].

Broader Societal and Political Effects

The effects of misleading headlines extend far beyond individual readers. At a macro level, they contribute to infodemics, defined as the overabundance of information, both accurate and not, that spreads rapidly during major events like pandemics [434]. By distorting facts or exaggerating claims, misleading headlines accelerate the dissemination of half-truths that clutter the media environment and hinder informed decision-making.

In the political realm, headlines from hyperpartisan or ideologically driven outlets are particularly damaging. These headlines often shape issue salience and political beliefs by framing stories in a way that reinforces group identity or vilifies outgroups [364]. This can foster polarization, deepen distrust in democratic institutions, and distort public discourse by encouraging oversimplified or emotionally reactive responses to complex issues [291].

The economic consequences are also notable. Misleading headlines can influence investor sentiment [88, 371], consumer behavior [4], health crises [16], or public policy [367]. Erroneous conclusions drawn from such headlines have the potential to create instability in markets or fuel public panic.

In sum, misleading headlines contribute to a media ecosystem that prioritizes virality over veracity. Their influence spans cognitive, emotional, and behavioral dimensions of information processing, and their societal consequences reinforce the urgency of intervention strategies that promote responsible headline writing and consumption.

2.4 Responding to Misinformation at the Individual Level

Understanding how individuals cognitively and behaviorally respond to misinformation is crucial for designing effective interventions. This section synthesizes research on perception,

belief formation, correction behaviors, and individual moderators that shape how people engage with misleading or false information, including news headlines.

2.4.1 Cognitive and Behavioral Responses to Misinformation

Individuals process misinformation through a combination of cognitive biases and heuristic-based behaviors. Common biases, such as confirmation bias, trust bias, and the bandwagon effect, lead users to accept information aligning with their pre-existing beliefs or circulating widely within their social circles, often without verification [1, 7, 144]. The time-saving heuristic also discourages critical engagement with digital content, especially when readers are overloaded or skimming through news feeds [144].

Behaviorally, misinformation prompts a range of reactions, from passive acceptance to active sharing, driven by source credibility, emotional salience, and social relevance [2]. Social media platforms exacerbate this dynamic by rewarding immediate engagement, reducing the likelihood of careful verification [409]. Selective exposure and echo chambers further reinforce misinformation as individuals preferentially engage with content aligned with their identity or worldview [355].

2.4.2 Influence of Beliefs and Ideologies

Misinformation, including misleading headlines, rarely spreads purely due to informational deficits; instead, its resonance and reach are deeply embedded in pre-existing belief systems, political ideologies, and identity-based cognition. Individuals are not neutral information processors; they interpret content through cognitive frameworks shaped by their social identities, ideological commitments, and affective predispositions.

Motivated Reasoning and Identity-Protective Cognition

One of the primary mechanisms through which beliefs shape misinformation engagement is motivated reasoning, a process wherein individuals selectively interpret information to confirm their prior beliefs and avoid cognitive dissonance [385]. This tendency becomes particularly pronounced in politically polarized environments, where information that aligns with partisan values is more readily accepted, even when it's inaccurate, while disconfirming evidence is scrutinized or dismissed [244].

Motivated reasoning is often intensified by identity-protective cognition, where individuals reject factual corrections or inconvenient truths not because they are misinformed per se, but because accepting such information would threaten their social or political identity [207]. This explains why fact-checking and corrective interventions often backfire or fail to produce attitude change among strongly partisan individuals [383].

Role of Partisanship and Ideological Congruence

Ideological congruence significantly influences the perceived credibility of news content, including misleading headlines. Research by Pennycook et al. [305] shows that individuals are more likely to rate false headlines as accurate when they align with their political orientation.

Similarly, Allcott et al. [13] found that during the 2016 U.S. election, fake news stories were disproportionately shared within ideologically homogeneous networks, amplifying their reach through affective polarization.

This dynamic is not limited to the American context. Studies across different countries have shown that individuals with stronger partisan affiliations exhibit selective exposure and selective acceptance of information, engaging more with ideologically reinforcing headlines and avoiding or distrusting counter-attitudinal news [157, 204].

Framing Effects and Ideological Filters

The way in which information is framed also interacts with belief systems. News frames that emphasize conflict, morality, or partisanship can trigger affective responses that short-circuit critical evaluation, especially among ideologically aligned readers [105, 334]. Misleading headlines that invoke politically salient cues, such as references to immigration, crime, or government overreach, are particularly effective at eliciting strong emotional reactions and belief persistence [1].

2.5 Correcting Misinformation and Misleading Content

Correcting misinformation and misleading content is a critical step toward restoring public trust and promoting informed discourse. This section reviews journalistic and computational correction strategies, examines individual and societal responses to such interventions, and evaluates factors that influence their effectiveness and ethical implications.

2.5.1 Correction of Misleading Headlines

The correction of misleading headlines is a vital but underexplored frontier in combating subtle misinformation. Unlike overtly false claims, misleading headlines often operate within factual boundaries while distorting reader interpretation through selective emphasis, omission, or sensational framing. This makes correction especially complex: it requires not just detection but interventions that preserve reader trust and maintain engagement.

Existing research has largely focused on the automated detection of misleading headlines using computational techniques such as headline-body incongruity modeling, stance detection, and adversarial learning. Notable approaches include BaitWatcher and MuSEM, which identify headlines that diverge from the content’s primary claim or frame it misleadingly for attention purposes (e.g., exaggeration, clickbait) [267, 297]. However, these studies often fail to assess whether such detections result in effective corrections, meaning changes in how readers interpret the news or their attitudes towards the content.

Recent work has begun addressing this correction gap more directly. Shen et al. propose a framework for detecting hallucinations in news headlines, claims that are not supported by the article body, and generate natural language explanations to justify the labeling [361]. Their approach combines entailment-based verification with explainable AI, improving detection accuracy and interpretability. A follow-up study by Shen et al. extends this framework into multilingual contexts by introducing a fine-grained hallucination dataset across five languages

and proposing prompt-based techniques for few-shot headline verification [362]. These studies suggest that explainable detection systems can help identify misleading elements in headlines, though they stop short of actively revising or correcting them.

Complementing this detection-centric perspective, Jahanbakhsh et al. explore the opposite end of the spectrum: user-driven headline correction. In their browser extension study, participants were invited to rewrite headlines to better reflect the article content. While users primarily modified headlines for clarity or humor, rather than introducing misinformation, the edits frequently aligned with the users' ideological perspectives, sometimes reinforcing polarization [191]. This raises important questions about the trade-offs between editorial authority and reader empowerment in the correction process.

Meanwhile, emerging NLP methods are experimenting with automated headline generation that explicitly seeks to reduce misleadingness. For example, studies have proposed transformer-based models to generate alternative headlines that vary in tone or emotional salience while remaining faithful to article content [293, 436]. While promising, these systems are rarely evaluated through user studies, leaving open questions about how such corrections influence trust, engagement, or comprehension.

Some experimental work suggests that incorporating uncertainty cues, contextual elaboration, or the removal of emotionally charged language into headline corrections may reduce misinterpretation while preserving engagement [1, 17, 318]. However, the psychological and behavioral consequences of headline corrections remain insufficiently explored, especially compared to broader misinformation debunking studies.

Overall, the correction of misleading headlines requires bridging computational detection, editorial judgment, and reader-centered evaluation. As these recent studies show, effective correction is not only about rewriting for accuracy but also about anticipating how readers interpret, trust, and emotionally react to those changes. This dissertation contributes to that goal by not only modeling the mechanics of misleading headline detection but also testing correction strategies grounded in both user cognition and journalistic values.

2.5.2 Correction in Journalism

While considerable attention has been devoted to correcting misinformation in general, the journalistic process of acknowledging and correcting errors holds unique ethical and practical significance. This section examines how journalistic corrections are implemented, the ethical principles that guide them, and how audiences perceive and respond to such corrections.

Practices and Ethics in Journalistic Corrections

Journalistic corrections function as a vital self-regulatory mechanism, upholding transparency and editorial accountability in a fast-paced and often polarized media environment. Their implementation reflects a negotiation between professional ethics, practical newsroom realities, and reputational considerations.

Correction Practices. Scholars have identified a spectrum of correction formats tailored to the severity and visibility of the error. Standard corrections address minor factual mistakes, such as typographical errors, misspelled names, or inaccurate dates, and are typically placed

in designated correction sections (in print) or appended as footnotes in digital articles [21]. In contrast, “major mea culpas” are reserved for significant breaches of professional norms, such as fabrication, plagiarism, or misreporting with widespread consequences. These involve lengthy editorials, formal apologies, and institutional reflection aimed at rebuilding trust and reaffirming journalistic values [339].

In digital journalism, correction practices have adapted to the demands of real-time publishing. These practices include live updates during breaking news coverage, revision histories that clearly document editorial changes, and the controversial option of completely removing content when errors are deemed too significant to correct within the original article [134]. While unpublishing is framed as damage control, critics argue it may “memory-hole” errors without accountability, undermining the very transparency corrections are meant to preserve.

Ethical Frameworks. International codes of journalistic ethics uniformly stress the importance of corrections. The transparency imperative requires that errors be acknowledged promptly, prominently, and unambiguously, especially when they cause harm or mislead audiences [21]. The proportionality principle dictates that the visibility and gravity of the correction should correspond to the original error’s potential impact: a brief footnote may suffice for a typographical slip, but not for factual inaccuracies that shaped public discourse. Furthermore, stronger correction mandates tend to correlate with higher levels of press freedom, while authoritarian regimes often sidestep public acknowledgments of error altogether [21].

Ethical Tensions. Even in democracies, correction practices are constrained by competing priorities. News outlets must balance institutional credibility against the potential reputational damage of publicly admitting fault. This tension is particularly acute in “major mea culpa” cases, where corrections are not just about setting the record straight but navigating journalistic identity and authority [339]. In the digital domain, these tensions are amplified: archiving errors risks prolonging misinformation exposure, while deleting them may obscure accountability. These trade-offs require ongoing ethical reflection within journalism as a profession.

Public Reactions to News Corrections

Audience responses to news corrections are shaped by a complex interplay of trust, prior beliefs, perceived transparency, and emotional orientation toward media institutions. Readers who possess high trait forgiveness or elevated levels of news media literacy are generally more receptive to corrections and more likely to retain trust in the news outlet, particularly when the correction is timely and appears sincere [171]. However, broader dynamics often complicate this positive reception.

One major determinant of correction effectiveness is media credibility. When trust in the media is high, factual corrections are more readily accepted; when media are viewed with suspicion or negativity, even accurate corrections may be dismissed or ignored [254]. This creates a vulnerability where the corrective process itself can be undermined by preexisting cynicism, limiting its potential to restore truth.

Corrections also carry a negative spillover effect in some contexts. Repeated exposure to warnings about misinformation, common in the current information landscape, can increase overall skepticism, reducing trust in both inaccurate and accurate news alike [377, 404]. Similarly, peer-led social corrections on platforms like Twitter or Facebook may reduce belief in misinformation but also erode perceived accuracy of true information, operating as a double-edged sword [377]. Despite long-standing fears, backfire effects, where corrections inadvertently reinforce false beliefs due to familiarity, are relatively rare. Recent replication studies have found little consistent evidence to support this concern, especially when corrections are carefully worded and not overly repetitive of the original falsehood [320].

Lastly, the framing of corrections and their alignment with readers' prior views significantly influence their reception. When corrections resonate with existing opinions or are delivered by trusted sources, they are perceived as more truthful and are more likely to shift beliefs. Conversely, corrections that contradict identity-driven narratives or political leanings may be met with resistance or dismissal [243].

Together, these findings underscore that the success of journalistic corrections hinges not only on their factual content but also on the psychological and contextual landscape into which they are received.

2.5.3 General Misinformation Correction Strategies

Misinformation correction strategies encompass a wide range of interventions aimed at preventing belief in false or misleading content, mitigating its effects, and restoring accurate understanding. This subsection reviews major intervention types, evaluates their effectiveness and constraints, and explores ethical and psychological considerations that shape their real-world impact.

Types and Timing of Interventions

The correction of misinformation involves diverse intervention strategies that vary in both type and timing. These strategies are typically classified into three broad categories: prebunking (preemptive), debunking (reactive), and nudging (behavioral), each of which corresponds to a different stage in the misinformation lifecycle. Understanding the design and deployment of these interventions is critical to developing effective misinformation mitigation systems.

Direct corrections are explicit statements that refute or clarify false claims and are typically disseminated by credible sources such as journalists, scientists, or public institutions. These corrections may target either the content of the misinformation (message-focused) or the originator's credibility (source-focused). While direct corrections alone are effective, their impact is significantly enhanced when combined with source discreditation strategies [113].

Source discreditation aims to reduce the influence of misinformation by undermining trust in its source. Experimental studies have demonstrated that this approach is particularly effective when paired with factual refutations, as it not only counters the false content but also addresses the credibility of its origin [113].

Another important distinction lies in the format of corrections. Corrections can be framed as narratives or statistical statements. Narrative corrections, which contextualize information through storytelling, have been found to be more persuasive when emphasizing collective

interests. In contrast, statistical corrections, which present numerical evidence, are more effective when highlighting individual-level implications [428].

Preemptive interventions, often based on inoculation theory, aim to build cognitive resistance to misinformation before exposure occurs. These interventions typically involve media literacy education, critical thinking exercises, or exposure to weakened forms of misinformation, thereby preparing individuals to recognize and reject false information upon future encounters. Empirical evidence supports the efficacy of inoculation-based strategies in various national and cultural contexts [35, 239].

Social corrections refer to instances in which peers challenge misinformation in online environments. Although peer-based corrections can be effective, their application is often limited by social and psychological constraints. Individuals may hesitate to engage in corrective behavior due to fear of confrontation or reputational harm, despite recognizing misinformation [161].

Nudging interventions are designed to subtly influence users' behavior without necessarily altering their beliefs. These include accuracy nudges, which prompt individuals to reflect on the truthfulness of content before sharing it, and social norm cues, which indicate what information is accepted or trusted by others in a given network or community. Both techniques have demonstrated effectiveness in reducing the spread of misinformation in controlled and real-world settings [62, 202, 218, 336].

The timing of interventions is a critical determinant of their success. Immediate corrections, delivered shortly after misinformation is encountered, are typically more effective in preventing the consolidation of false beliefs [52, 332]. Preemptive corrections, such as inoculation, have been shown to produce more durable resistance by equipping individuals with tools to critically evaluate future misinformation [35, 239]. In contrast, delayed corrections often suffer from diminished impact due to memory decay or the entrenchment of initial impressions, underscoring the importance of timeliness in misinformation interventions [52, 332].

Taken together, these findings suggest that combining multiple types of interventions at different temporal stages—before, during, and after misinformation exposure—offers the most comprehensive and effective approach to mitigating the influence of misinformation.

Effectiveness and Limitations

The effectiveness of misinformation correction strategies is highly context-dependent and shaped by the nature of the intervention, the credibility of the source, the audience's psychological predispositions, and the topic of misinformation. While various interventions, including fact-checking, inoculation, and accuracy nudging, have demonstrated success in mitigating belief in false information, each approach has important limitations that constrain its utility across contexts.

Preemptive interventions, such as psychological inoculation, have consistently shown promise in improving resistance to misinformation prior to exposure. Interactive tools like Bad News ¹ or media literacy campaigns significantly reduce belief in and willingness to share misinformation, especially in politically polarized environments [35, 239]. These interventions work by simulating misinformation tactics in controlled settings, helping individuals recognize rhetorical manipulation in real-world scenarios. However, inoculation requires continuous

¹<https://www.getbadnews.com/>

engagement and periodic reinforcement to maintain its protective effect over time [239].

Reactive interventions, such as fact-checks and corrective statements, are most effective when delivered promptly, clearly, and by trusted sources [182, 246]. Source credibility plays a pivotal role in determining whether audiences accept or resist corrections; individuals are more likely to reject corrections issued by sources they perceive as ideologically misaligned or untrustworthy [86, 440]. Additionally, cognitive limitations, such as belief perseverance and the familiarity effect, can reduce the impact of post hoc corrections, particularly when misinformation has already been internalized [114].

Nudging interventions, including accuracy prompts and social norm cues, offer a low-friction mechanism for reducing misinformation dissemination without requiring individuals to change their beliefs. These techniques have been shown to decrease sharing intentions in online environments with minimal disruption to user experience [62, 336]. Nevertheless, nudging may yield only short-term behavioral changes and may not be sufficient to alter entrenched beliefs, especially among individuals with strong prior attitudes or partisan commitments [202, 218].

A notable challenge across all interventions is the phenomenon of generalized skepticism. Some studies indicate that exposure to frequent corrections or misinformation warnings may erode trust not only in false claims but also in accurate information and reputable news outlets [377, 404]. This negative spillover effect underscores the need for balance: while it is crucial to alert the public about misinformation, doing so excessively or without precision may contribute to information fatigue or cynicism.

Another concern is the topic-specific variability in correction outcomes. Corrections are more likely to be accepted in less identity-relevant domains (e.g., health, science) compared to politically or ideologically charged issues, where corrections may provoke defensiveness or motivated reasoning [86, 408]. Additionally, although backfire effects, where corrections lead to increased belief in misinformation, were once seen as a significant risk, recent replication studies have shown that they are rare or negligible when corrections are appropriately designed. [320].

In summary, while correction interventions can mitigate misinformation effects under certain conditions, their success is influenced by timing, source credibility, psychological receptivity, and the sociopolitical context of the misinformation. Acknowledging these limitations is essential for designing more targeted and sustainable corrective strategies.

Ethical and Psychological Considerations

The correction of misinformation, while necessary for preserving informational integrity, is fraught with ethical and psychological complexities. From the ethics of data handling to the nuances of belief revision, interventions must navigate a landscape shaped by regulatory constraints, algorithmic biases, cognitive tendencies, and public trust.

Ethical Considerations. Ethical concerns frequently arise in misinformation research and intervention, particularly when researchers engage with social media data. Violations such as unauthorized data scraping, identity misrepresentation, and neglect of informed consent are not uncommon, especially when researchers without expertise in health data ethics enter sensitive domains, leading to blurred boundaries between public and private digital spaces [78]. These issues highlight the pressing need for more rigorous ethical standards and domain-specific guidelines.

Interventions to correct misinformation also confront the challenge of balancing harm reduction with the protection of free speech. While many individuals favor moderation policies that mitigate harm without suppressing expression, political ideology influences these preferences. Moreover, algorithmic moderation mechanisms introduce further ethical dilemmas, particularly when efforts to suppress harmful falsehoods risk overstepping into legitimate discourse [219].

Misinformation detection systems, especially those relying on automation, are susceptible to informational injustices. These include representational gaps due to the underrepresentation of marginalized communities in training datasets, credibility disparities that mislabel content from minority voices, and the disproportionate targeting or surveillance of vulnerable populations [285]. Such systemic biases raise fundamental questions about fairness, inclusivity, and the distribution of burdens in misinformation mitigation efforts.

Amid concerns about censorship, scholars and policymakers increasingly emphasize regulatory approaches grounded in individual empowerment. Rather than imposing top-down restrictions, strategies such as media literacy training and community-based fact-checking offer more ethically tenable solutions. However, if poorly implemented, even these approaches risk amplifying public distrust, especially in politically polarized environments [257, 382].

Psychological Considerations. Psychological factors significantly influence the reception and effectiveness of corrections to misinformation. Preemptive strategies, such as psychological inoculation, where individuals are exposed to weakened versions of misleading claims, have demonstrated substantial benefits. These interventions bolster credibility judgments, reduce belief in falsehoods, and decrease the likelihood of sharing misinformation [250, 263].

The temporal dimension of correction is critical. Refutational narratives that challenge misinformation work best prior to exposure, while straightforward factual rebuttals tend to be more effective post-exposure. However, delayed corrections often contend with entrenched beliefs and cognitive inertia, leading to diminished effectiveness [162, 265]. The persistence of misinformation despite corrections, known as the continued influence effect, demonstrates how false information can linger in memory and shape reasoning even after it is discredited [238, 391].

Motivational drivers also shape belief in misinformation. Individuals are more likely to accept false claims that align with their ideological views, offer entertainment value, or resonate with existing skepticism toward institutions [138, 265]. Age-specific vulnerabilities further complicate intervention design. Older adults, who often face declines in cognitive speed and digital literacy, are more susceptible to misinformation but remain underserved by current correction strategies despite their political and social influence [138].

Finally, while misinformation corrections often improve belief accuracy, they rarely translate into behavioral change, such as vaccine uptake or adherence to public health guidelines. Trust in the source of correction mediates its success; when corrections come from low-credibility sources, they are far less likely to alter beliefs or actions [250, 391].

These ethical and psychological considerations highlight the complexity of designing effective, fair, and trustworthy interventions against misinformation. They emphasize the necessity of interdisciplinary approaches that blend technical innovation with human-centered values and behavioral science.

2.6 Computational Approaches to Misinformation Mitigation

In response to the growing scale and sophistication of misinformation, computational approaches have emerged as essential tools for automated detection, classification, and correction. These methods leverage natural language processing, machine learning, and increasingly, large language models to identify misleading headlines, fake news, and deceptive content across platforms. This section surveys the current landscape of computational techniques, available datasets, and their applicability to addressing different forms of misinformation.

2.6.1 Detection of Misleading Headlines: Computational Approaches

A growing body of computational research has focused on detecting incongruent or misleading news headlines by leveraging linguistic, semantic, and contextual features. These approaches vary in architecture, data scale, modality, and explainability, yet most share the core goal of identifying headline-body mismatches or hallucinations that may mislead readers.

Wei and Wan [420] proposed a dual-classification framework that separately identifies ambiguous headlines, those containing vague or unclear language, and misleading headlines that contradict body text. They utilized class sequential rules for ambiguity detection and features measuring headline-body congruence for misleadingness. Their work was among the first to formally separate ambiguity from misinformation in headline assessment.

Deep learning architectures have since advanced the state of the art in headline incongruity detection. Yoon et al. [435] introduced a hierarchical LSTM encoder that models document-level context by augmenting training data at the paragraph level. Their system achieved 89% classification accuracy on a million-scale dataset annotated with fine-grained incongruity labels, highlighting the potential of hierarchical encoders to detect nuanced headline-body mismatches.

Other researchers have explored multilingual and cross-linguistic detection. Sepúlveda-Torres et al. [359] created *ES_Headline_Contradiction*, a Spanish-language dataset comprising 3,600 headline-body pairs labeled as compatible, contradictory, or unrelated. Fine-tuned BERT models reached an F1-score of 91%, demonstrating transferability of contradiction modeling to non-English news corpora.

Recent approaches have incorporated explanation generation to improve interpretability. Shen et al. [361] introduced *ExHalder*, a model combining natural language inference (NLI) with rationales that justify headline hallucination predictions (e.g., “Headline claims X, but article states Y”). Trained on over 6,000 multilingual pairs, *ExHalder* achieved 85% accuracy and generated human-readable justifications that enhance transparency in headline assessments.

Graph-based models have also gained traction. Yoon et al. [434] employed a graph neural network (GNN) to model semantic relationships between headline and body sentences. This model outperformed text-only baselines by 7.2% in F1-score, though it required substantial computational resources due to graph construction and propagation steps.

Lightweight tools for real-world deployment have also emerged. Park et al. [297] developed *BaitWatcher*, a browser plugin designed to detect headline incongruence in real time. Their attention-based model prioritized speed and efficiency but traded off model depth, making it suitable for casual users rather than professional journalists.

Cross-domain robustness has been a persistent challenge. Jang et al. [193] improved generalizability by incorporating auxiliary textual data, such as reader comments and related

articles, into a BiLSTM framework. While this approach enhanced performance in diverse news domains, its reliance on external data sources limits scalability and applicability to low-resource contexts.

Emerging efforts have also tackled multimodal misleading content. Sung et al. [380] introduced a video-headline mismatch detection dataset containing annotated pairs of news videos and their associated headlines. They benchmarked multimodal baselines for detecting when headlines misrepresent video content, extending the misleadingness detection paradigm beyond text.

Despite these advances, current computational techniques exhibit notable limitations. First, most large-scale datasets are synthetically generated or weakly labeled, failing to capture the subtle, context-dependent techniques used in real-world misleading headlines. Consequently, existing models are adept at identifying clear-cut incongruence but struggle with subjective or editorially nuanced misleadingness. Second, few systems provide explanations aligned with journalistic principles or human reasoning, limiting their usefulness for editorial workflows or public transparency. These gaps motivate the need for human-in-the-loop, explanation-aware systems that better reflect the subjective nature of headline misleadingness.

2.6.2 Computational Approaches to Fake News, Clickbait, and Other Misinformation

The broader misinformation detection landscape includes a wide array of computational approaches targeting fake news, clickbait, and sensational content. These techniques range from traditional linguistic models to multimodal architectures and knowledge-aware systems. While these methods provide valuable insights, they often fall short in addressing the more nuanced and subjective nature of misleading headline detection.

Fake news detection has received extensive attention, with numerous systems trained to classify article-level credibility using text, user interaction data, and source metadata. For instance, Shu et al. provided a foundational review of machine learning pipelines combining content-based and context-based signals across various modalities [366]. Similarly, Zhou et al. categorized detection methods by data input type (text, network, temporal) and emphasized the role of user engagement patterns [443]. On social platforms, models have incorporated tweet propagation trees and thread dynamics to detect deceptive information early [61]. Fact-checking datasets like *ClaimBuster* [168] and *CLEF CheckThat!* [10] have supported the development of claim-level verification tools that integrate retrieval-based and natural language inference techniques.

Clickbait detection, though narrower in scope, focuses on identifying headlines crafted to mislead or sensationalize in order to capture user attention. Chakraborty et al. employed linguistic patterns and curiosity-inducing features to distinguish clickbait headlines from neutral ones [67]. Rony et al. applied deep learning methods to detect clickbait and analyzed a large-scale dataset to demonstrate that clickbait practices are also prevalent among reputable news sources [335]. More recent approaches incorporate transformer-based classifiers and attention mechanisms to enhance accuracy while reducing false positives [324].

Sensationalism detection has received increasing scholarly attention, with several computational frameworks proposed to identify exaggerated or emotionally charged headlines.

Frameworks such as *SENS-HEAD* identify sensationalist headlines using features like sentiment polarity, punctuation density, syntactic complexity, and stop-word ratios [69]. Maximum entropy models have been applied to assess sensationalism and scientific accuracy in health-related articles [176]. However, heavy dependence on manual feature extraction continues to constrain scalability.

For bias and propaganda detection, studies have deployed both linguistic and visual analysis. Ibrahim et al. [185] conducted a large-scale analysis of racial and gender representation in over 441,000 articles, uncovering systematic underrepresentation of minorities. Word-embedding-based dictionaries enable sentence-level bias detection with high accuracy [91], while transformer-based models have proven effective in identifying ideological slants and hyperpartisanship [342]. Nakov and Martino [277] emphasized how propaganda techniques such as emotional appeals blended medical and political misinformation during the COVID-19 pandemic.

Multimodal misinformation detection techniques have also emerged, such as *SAFE*, which integrates image-text coherence to detect fake news involving misleading visuals [442]. These techniques are particularly effective in social media environments where visual content strongly influences perception. Meanwhile, *DeClarE* leverages both content and metadata attention to align claims with external evidence from trusted sources [313]. A growing body of work suggests that combining text, visual information, social context, and propagation patterns significantly improves detection accuracy [159, 329]. Recent advances in large vision-language models (LVLMs), such as those introduced in the *LEMMA* framework, further enhance cross-modal reasoning for complex verification tasks [430]. When augmented with external fact-checking knowledge bases, LVLMs demonstrate improved performance in identifying multimodal misinformation [431].

In terms of datasets, several benchmarks have guided progress in this domain. The *FEVER* dataset provides over 185,000 claims annotated against Wikipedia evidence and is widely used for training natural language inference-based fact-checkers [394]. The *LIAR* dataset features over 12,000 short political statements annotated with truthfulness labels such as “true,” “pants-on-fire,” and “half-true,” based on expert fact-checks [415]. The FakeNewsNet repository aggregates news content, publisher metadata, and social context from *BuzzFeed* and *PolitiFact* sources, enabling multimodal and contextual analysis [365]. *FbMultiLingMisinfo* expands the scope by providing a large-scale, multilingual benchmark composed of fact-checked URLs from Facebook and their propagation patterns on Twitter, supporting misinformation detection across multiple languages and platforms [34]. *VERITE* addresses limitations in prior multimodal datasets by offering balanced image-text pairs specifically curated to reduce unimodal bias, thereby enhancing the robustness of misinformation detection models [295]. Other widely used multimodal datasets include *GossipCop* and *PolitiFact*, which contain paired news articles and images annotated with fact-check labels, and have become standard benchmarks for evaluating visual-textual misinformation detection methods [264]. The *NELA* dataset family contributes large-scale, multi-year corpora of news articles labeled with source-level veracity information, offering valuable ground truth for studying misinformation, propaganda, and media bias at scale [156].

Despite these advances, several limitations persist. Many models rely on fully-formed articles or structured claims rather than isolated headlines. Moreover, most fake news datasets consist of relatively clear-cut or overtly false claims, lacking the subtlety, ambiguity, or reader-dependent interpretations common in misleading headlines. Clickbait

detection, while structurally closer, focuses on sensationalism rather than factual distortion. Furthermore, multimodal methods often target visual deception or conspiracy content, which differs significantly from headline-level inference.

While these computational techniques have propelled misinformation research forward, they do not adequately address the linguistic, contextual, and subjective complexity of misleading headline detection. This gap provides a key motivation for the dissertation’s focus on modeling and evaluating misleadingness at the headline level.

2.7 Human–AI Collaboration in Subjective and High-Stakes Domains

Human–AI collaboration has shown promise across a range of domains, particularly when each partner contributes distinct strengths to the task. When properly designed and deployed, such collaborations can yield superior outcomes compared to either humans or AI working alone [187]. However, this synergy is not guaranteed: empirical studies find that while human–AI teams often outperform unaided humans, they do not always surpass AI-only systems. The complementary advantage appears strongest when the human has a relative edge in domain-specific knowledge or interpretive judgment [350, 401].

2.7.1 Domains and Mechanisms of Successful Collaboration

Collaboration between humans and AI has been successful in various application areas. In medicine, large language models (LLMs) and diagnostic AI systems have reached accuracy levels as high as 92%, demonstrating their utility in supporting clinical decisions [150]. In creative tasks, human–AI partnerships have produced more compelling artistic outputs: human-in-the-loop poetry was rated as more beautiful [314], and collaborative systems fostered greater creativity in visual art generation [173].

Beyond outcomes, scholars have examined the mechanisms that make collaboration successful. Shifting from automation to partnership requires both organizational and technical changes [422]. Effective AI collaborators tend to decompose complex tasks [170], provide justifiable reasoning [330], and allow humans to calibrate when to rely on AI suggestions [169]. These systems do not merely provide answers, but actively support human reflection and deliberation, particularly in ambiguous or high-stakes tasks.

2.7.2 Collaboration in Ambiguous or Subjective Contexts

Successful human-AI collaboration is especially crucial in subjective or ambiguous decision-making contexts. For instance, in the moderation of hate speech, where ground truth is often socially contested, LLMs can propose initial labels, but humans are needed to apply contextual nuance and resolve interpretive uncertainty [296]. Research shows that humans selectively rely on AI, increasing trust when the AI is correct but demonstrating caution in ambiguous scenarios [330]. Human involvement in the early stages of creative or interpretive tasks also helps retain originality and guard against generic outputs [181]. Moreover, AI explanations that are clear and well-contextualized help reduce user uncertainty [184], and can foster more effective collaboration, if paired with thoughtful interface design that avoids automation bias [74].

2.7.3 Human-AI Collaboration in Misinformation Prevention

Human–AI collaboration is increasingly central to combating misinformation, combining computational scalability with human judgment. While AI systems can rapidly detect patterns, inconsistencies, or manipulative cues across large datasets, including text, images, and deepfakes, human experts play a vital role in contextualizing findings, verifying sources, and interpreting intent [326, 426]. Collaborative strategies such as MisinfoEval leverage generative AI to simulate disinformation scenarios, helping humans proactively anticipate and counter emerging threats [137]. Experimental studies show that dialoguing with AI can improve human discernment of visual misinformation [326], but trust hinges on explanation quality; poorly designed or inaccurate rationales can mislead users and degrade performance [119, 374]. In effective hybrid workflows, AI might draft counter-messaging while humans ensure alignment with cultural and ethical standards [137]. As the threat landscape evolves, future systems must integrate human discretion with AI speed to track disinformation across platforms and support trustworthy information ecosystems [261, 300].

2.7.4 Human–AI Collaboration in Journalism

In the journalism domain, AI systems are increasingly used to support editorial workflows. They assist journalists in monitoring news sources, extracting data, detecting emerging trends, and even drafting content [101, 323]. In investigative journalism, AI has been employed to automate data extraction and cross-referencing, enabling deeper reporting with reduced manual effort [80]. These tools streamline routine processes such as transcription, summarization, and headline generation, allowing journalists to focus on storytelling and editorial judgment.

However, the integration of AI into journalism raises significant concerns. AI-generated content can introduce errors, exaggeration, or fabrication, posing risks to accuracy, a foundational value of journalistic ethics [245]. The blurred accountability in co-produced content further complicates editorial oversight. Additionally, covert use of AI tools may erode public trust, underscoring the need for transparency around AI involvement. To mitigate these risks, best practices emphasize editorial review: journalists must validate, edit, and contextualize AI outputs [245, 397]. Collaborative tool design, bringing together journalists, technologists, and data scientists, can also help ensure that AI aligns with newsroom values and professional norms [427].

2.8 Explainable AI in Collaborative and Subjective Contexts

Explainable AI (XAI) seeks to make the reasoning processes of machine learning models transparent and intelligible to human users [274, 345]. Within human–AI collaboration settings, explanations play a pivotal role in enabling users to understand, verify, and act upon model outputs. Particularly in high-stakes domains such as healthcare [338] and autonomous driving [229], interpretable explanations have been shown to enhance trust and safety by contextualizing model behavior in human-understandable terms.

In collaborative workflows, XAI can improve task performance and support complementary decision-making. Prior work has shown that clear explanations enable humans to more effectively validate AI predictions [209, 282], negotiate responsibility [330], and modulate

system autonomy in real time [151]. These mechanisms support the broader CSCW goal of designing cooperative systems that preserve human discretion while enhancing system efficacy. Senoner et al. [358], for example, found that explanation quality directly impacts collaboration fluency and shared task awareness.

However, explainability also introduces trade-offs. Post hoc XAI methods can inadvertently reinforce biases present in training data [258], and the push for transparency can come at the cost of predictive accuracy in complex models [229]. Additionally, the effectiveness of an explanation is not uniform across user types: explanations that are helpful to clinicians may confuse patients [14]. This suggests a need for situated, user-centered design that accounts for domain-specific norms and varying levels of expertise.

Efforts to evaluate explanation quality have historically lacked standardization. Kim et al. [213] identify methodological fragmentation across 73 studies, noting that only a fraction apply shared evaluation frameworks. Their taxonomy of “meaningful explanations” emphasizes not just explanation correctness, but also its role in fostering productive interaction and supporting decision-making. Hoffman et al. [174] similarly argue for metrics that assess practical understanding, while Sovrano et al. [372] propose a model-agnostic “Degree of Explainability” metric combining informativeness, clarity, and relevance. Mohseni et al. [268] advance this agenda by aligning XAI evaluation methods with the needs of diverse stakeholder groups across HCI and machine learning.

These evaluation needs are particularly salient in misinformation detection, an area marked by subjectivity, ambiguity, and social consequence. XAI has emerged as both a technical and sociotechnical response to the challenges of fake news and disinformation. Recent work highlights how explanation techniques such as feature attribution and multi-modal rationale surfacing help expose misleading claims and platform-specific manipulation [203, 226, 287]. In disinformation detection tasks, Schmitt et al. [351] found that free-text explanations improved layperson discernment, while structured rationales better supported experts, suggesting that explanation modality should be tailored to user roles.

Nonetheless, challenges remain. Alikhademi et al. [12] observe that many widely used XAI tools lack the capacity to communicate bias, and few support fairness-oriented interpretation. These findings underscore the CSCW principle that explainability must go beyond model transparency to consider social context, interpretive pluralism, and accountability in collaborative systems.

Together, this body of work positions XAI as a core mechanism for supporting collaborative sensemaking in sociotechnical systems, especially where ambiguity and subjectivity are central. For tasks like editorial decision-making and misinformation detection, explanation systems must be not only technically sound but also context-aware and designed for interaction across diverse stakeholders.

2.9 Human Annotation Disagreement as a Signal, Not Noise

Crowd-sourced misinformation annotation studies have illuminated the inherent subjectivity and complexity of evaluating misinformation, emphasizing that disagreement among annotators is not merely noise but a meaningful reflection of ambiguity. Prior works [107] advocate for interpreting disagreement as a signal of semantic nuance, showing how approaches

like *CrowdTruth* leverage diverse annotations to capture multiple valid perspectives rather than forcing consensus. Similarly, Allen et al. demonstrate that aggregated crowd evaluations align well with expert judgments, suggesting that collective interpretations can effectively surface contested or ambiguous content [15]. However, studies like Bak-Coleman et al. ([29]) and Hoff and Preminger ([106]) highlight challenges stemming from annotators' varying expertise, cautioning that crowdsourced labels can fluctuate in quality without mechanisms to address variability. These findings challenge the assumption of a single, objective ground truth in misinformation detection and underscore the need for AI evaluation frameworks that embrace disagreement as a feature of social judgment tasks.

Despite this recognition, most AI evaluation pipelines still prioritize consensus and factual verification, overlooking situated, subjective interpretations that characterize misinformation in practice. Many misinformation datasets rely on binary true/false labels, ignoring ambiguity and leading to unreliable results [393]. For example, models evaluated on FEVER often exploit surface-level cues rather than understanding evidence [354]. Detection systems also struggle with AI-generated misinformation that mimics credibility signals, revealing blind spots to contextual subjectivity [441]. Efforts to increase annotator agreement often impose rigid labeling rules, sacrificing nuanced interpretations of author intent or cultural framing [41]. Additionally, large-scale fact-checking frameworks overlook shifting community contexts, such as evolving health narratives [153]. Together, these challenges point to the need for AI evaluations that embrace disagreement and situated interpretation rather than forcing binary, consensus-driven outcomes.

Chapter 3: Perceiving and Correcting Misleading Headlines: Perspectives from Journalists and News Readers

In this chapter, we conducted a qualitative study to understand how journalists and news readers perceive misleading headlines, examining their definitions, reactions, and proposed correction strategies. By analyzing the tensions and mismatches between these two stakeholder groups, the study revealed divergent expectations around headline accuracy and laid the groundwork for designing more transparent and accountable news communication systems. This work was conducted in collaboration with Saransh Grover, Farhana Uddin, Yoo Yeon Sung, Mohammad Ali, and Dr. Naeemul Hassan, and portions of the findings were presented at the Computation + Journalism Symposium 2024.

3.1 Introduction

In today’s digital ecosystem, news consumption is increasingly shaped by algorithmic feeds [206], engagement metrics [222], and headline-driven attention flows [311]. While the internet has democratized access to information, it has also amplified the spread of misinformation in subtler forms, such as misleading headlines. A misleading news headline occurs when the title of a news article misrepresents, exaggerates, or omits crucial details from the body content, often distorting a reader’s interpretation. For example, the headline:

*Amazon Workers Left In Tears After Boss Lays Everyone Off Because He’s Moving To A New State!*¹

leads readers to believe Amazon itself is conducting layoffs. However, the article clarifies that the workers were employed by a third-party Amazon Delivery Service Partner (DSP), not Amazon directly. Without that context, the headline conveys a misleading impression.

This phenomenon is especially problematic given that many readers engage with headlines alone, without reading the full article [136]. Even when they do, first impressions formed by headlines tend to persist and influence later recall and judgment [111]. As digital headlines increasingly serve as standalone artifacts, circulating through social media and mobile news apps, the potential for misleading impressions becomes a serious civic concern.

Although CHI and CSCW researchers have extensively studied misinformation, much of this work has focused on fabricated content, platform moderation, or user-side detection [235, 357, 376]. Comparatively, little attention has been paid to misleading headlines despite evidence that headlines from mainstream outlets can cause more harm on social media than outright fake news [16]. Prior work has explored various aspects of news manipulation, including clickbait

¹<https://www.comicsands.com/amazon-delivery-drivers-laid-off>

detection [335], linguistic bias [72], and the effects of misleading headlines on memory [111]. However, these studies rarely examine how different stakeholders, such as journalists and readers, conceptualize and respond to misleadingness in context.

To understand and address the problem of misleading headlines more holistically, we argue it is critical to examine both sides of the news ecosystem: journalists, who craft headlines while navigating institutional norms, commercial incentives, and platform-driven pressures, and readers, who often encounter these headlines in isolation and rely on them to form impressions and make decisions. Prior work has shown that headline consumption frequently occurs without reading the full article, making readers particularly vulnerable to misinterpretation or misinformed judgments when headlines are misleading [136]. At the same time, journalists must often balance accuracy with attention capture in fast-paced digital environments, where editorial discretion may be constrained by algorithmic visibility demands [98].

This dual-sided approach builds on prior CSCW and HCI work examining stakeholder tensions in contested or asymmetric information environments, such as misinformation ecosystems [165] and algorithmically mediated labor platforms [166]. These studies have demonstrated that aligning responsibilities and surfacing interpretive asymmetries across roles (e.g., between fact-checkers and users or between platform workers and clients) is essential for building shared accountability and promoting trust. We extend this framing to the journalism domain, where similar role-based conflicts shape how information is framed, interpreted, and contested. By foregrounding the perspectives of both headline producers and consumers, our study aims to unpack how misleadingness is perceived, rationalized, and corrected across professional and public contexts.

In this study, we conduct semi-structured interviews with journalists and news readers to investigate how misleading headlines are defined, perceived, and corrected. We ask:

RQ1a: *How do journalists, as news creators, perceive misleading headlines?*

RQ1b: *How do news readers perceive misleading headlines?*

RQ2a: *How do journalists attempt to correct headlines they perceive as misleading?*

RQ2b: *How do readers attempt to correct headlines they perceive as misleading?*

Our findings contribute to CSCW and HCI scholarship in the following ways:

- We provide a grounded understanding of how misleading headlines are often conflated with clickbait, bias, and sensationalism, revealing inconsistencies in both editorial practice and audience interpretation.
- We surface the role-based conflict between journalists and readers over who is responsible for headline integrity, exposing mismatched expectations around credibility, interpretation, and trust.
- We provide design-related insights for computational journalism, including the creation of assistive tools that identify potential sources of misleading information in headlines. Additionally, we propose editorial workflows that support collaboration between journalists and AI, aiming to clarify editorial intent, promote headline–content alignment, and enhance transparency and accountability in news presentation.

3.2 Method

We conducted semi-structured interviews with journalists and general news readers and performed a thematic analysis to answer the research questions. This approach allowed us to explore in-depth perceptions and corrections of misleading news headlines from both groups.

3.2.1 Preliminary Study

We performed a preliminary study to collect and analyze instances of misleading headlines to build an initial dataset. The misleading instances we identified helped to uncover certain textual, pragmatic, or linguistic factors that make a headline misleading. This analysis aimed to categorize these factors and select representative news articles from each category for the interviews. During the interviews, participants examined two misleading and two non-misleading news headlines to identify and analyze them. Collecting misleading and non-misleading samples through this preliminary study was essential for the subsequent thematic analysis.

Article Selection

For the preliminary study, news articles were gathered from a total of 12 sources (Table A.3 in Appendix), which included six reputed to be reliable and six considered unreliable. These articles spanned three domains: Health, Science & Tech, and Business. Each of the three authors was tasked with examining articles from four distinct sources, selecting five articles from each category in a chronological manner starting from March 31st, 2022. The evaluation process entailed reading the headline, interpreting its meaning, perusing the entire content, and then determining if the headline was indeed misleading. If deemed misleading, the reasons were duly noted. From each author's compilation, 10 misleading and 10 non-misleading articles were identified, leading to a preliminary dataset of 30 misleading and 30 non-misleading articles. Thus, the final dataset comprised 60 news articles, split evenly between misleading and non-misleading headlines. Distributions of the explored articles by sources can be found in Table A.3 under Appendix.

Article Annotation

In the annotation phase, each author was tasked with reviewing a set of 20 articles deemed misleading and another 20 deemed non-misleading, all previously compiled by the other two authors. These articles were provided in text format, and stripped of any source-related information to preclude potential biases towards specific sources. The approach to review mirrored the method detailed in the earlier section. During this exercise, every headline underwent three rounds of examination, which led to labeling them as either misleading or non-misleading, accompanied by varying degrees of confidence. Upon collating annotations and reviews, a consensus emerged that 18 articles were unanimously seen as misleading by all authors. Concurrently, 27 articles received agreement from at least two authors, while a minimum of one author found 37 articles to have misleading headlines. Distributions of the misleading and non-misleading headlines by sources can be found in Table A.3 under Appendix. Table A.4 also shows the distribution of misleading and non-misleading articles by topic.

Content Analysis

For the headlines identified as misleading by at least one author post-annotation, a thematic analysis was undertaken. The rationale offered by the authors during their reviews served as a foundation for this content analysis. We adhered to the six-step thematic analysis approach outlined by Braun and Clarke [53]. As a starting point, the authors collectively analyzed twelve articles, leading to the preliminary identification of seven codes. Subsequently, each author took on the individual task of examining the remaining twenty-five articles, performing open coding to pinpoint new codes. After a thorough discussion and reconciliation of discrepancies, a consolidated list of eight distinct codes emerged (shown in Table A.5 under Appendix section), delineating the various categories of misleading headlines evident in our dataset.

3.2.2 Semi-Structured Interviews

To investigate the research questions presented, we carried out semi-structured interviews with both journalists and everyday online news readers. This approach allowed us to gather insights into their perceptions, responses, and experiences when encountering misleading headlines.

Questionnaire Development

We developed two separate sets of questionnaires for the expert (journalists) and general (news readers) participants. While developing the questionnaires, we took into consideration relevant literature related to journalism practices [140, 386], the perceived credibility of news [131, 201], and user engagement with news content [395, 410]. The questionnaires were collaboratively developed by four authors of this paper.

First, the questionnaire for expert participants was designed to understand their perspectives on crafting potentially misleading headlines, their reactions to misleading headlines, and the types of corrections they provide. To develop this questionnaire, we reviewed the literature on journalists' headline-creation processes and the impact of economic pressures on journalistic standards [280, 386]. We piloted the questionnaire with two experienced journalists. Based on their feedback, we revised several questions to ensure clarity and relevance and then finalized the questionnaire.

Second, to develop the questionnaire for general participants, we reviewed related literature about how news consumers perceive and interact with news headlines [131, 201]. This questionnaire was designed to gather data on their news consumption habits, their ability to identify misleading headlines, and their reactions to such headlines. We piloted the questionnaire with two frequent news readers to finalize it, incorporating their feedback to improve the wording and flow of the questions.

The Institutional Review Board (IRB) of the authors' institutions examined and approved the questionnaires. In the appendix section, we have included the interview protocols.

Participants

We recruited 24 participants for this study, including 12 journalists ('J') and 12 general news readers ('P'). The study required participants with professional experience in news reporting. Due to the challenge of recruiting participants with the desired experience, we included some Journalism graduate students who have some professional experience (e.g., internship, part-time) as well. Researchers recruited professional participants through their personal and professional connections by sending invitations over email and posting on social media. For general news readers, researchers recruited adults who habitually read news articles online, posting recruitment notices on social media and in public places like libraries and canteens.

Eligibility criteria required participants to be over the age of 18 and reside in the United States. Journalists needed some professional experience in writing news articles or to be graduate students in the Journalism discipline. General participants needed to regularly read English news articles. Each participant took part in a one-hour interview.

The 12 journalists (J1-J12) included 9 females and 3 males, ranging in age from 18 to 75 years (Mean = 33.0, SD = 16.97). They held diverse roles, including journalists, editors, public radio producers, and journalism students with internship experience, representing a mix of early-career, mid-career, and senior professionals in the field (Details in the Appendix section [A.2](#)). The 12 general news readers (P1-P12) included 4 females and 8 males, ranging in age from 18 to 65 years (Mean = 27.9, SD = 11.48) [Details can be found in Appendix [A.1](#)]. They all reported a consistent habit of reading news articles online.

Procedure

This study involved two distinct groups of interviewees: experts (journalists) and general users (news readers). After securing participation consent, each interviewee was provided with a set of four articles, two misleading and two non-misleading, from our curated collection, without disclosing which articles were misleading. Before the interview, participants read the articles and recorded their observations in a shared Excel sheet, noting their expectations based on the headline, whether those expectations were met, any discrepancies encountered during reading, their perception of whether the article was misleading, their emotional reaction to misleading headlines, and suggestions for correcting the misleading headlines. During expert interviews, sessions began with questions about their professional background, preferred news topics, and perspectives on misleading headlines, followed by queries on their methods for identifying misleading headlines, the frequency of encountering such headlines, and their reactions to them. In the latter part of the interview, experts analyzed the pre-selected articles, explaining misleading scenarios from a journalistic standpoint. Interviews with general users followed a similar structure, with the final segment focusing on their experiences with the pre-selected articles and describing misleading scenarios from a reader's perspective. All interviews were audio-recorded with participant consent. Upon completion of the interview, expert participants were compensated with a \$30 Amazon gift card, while general participants received a \$20 Amazon gift card.

Data Analysis

We transcribed the interview audio using OpenAI’s open-source automatic speech recognition system, Whisper ². The resulting transcripts were manually reviewed and corrected by the authors, who divided the verification workload among themselves to ensure accuracy and consistency. We then applied the thematic analysis method to analyze the interview data. Following the methodological guidelines of [57] on thematic analysis, three authors coded the interview transcripts of journalists and news consumers. Initially, each author separately read the transcripts multiple times to identify recurring themes and supporting evidence. Specifically, the authors transformed the raw interview data into brief codes. Next, we looked for connections among these codes, adhering to axial coding principles [378]. For instance, we identified journalists’ and news consumers’ responses related to their perceptions of misleading headlines. After that, we analyzed the codes, identified patterns among them, and developed an initial set of themes that captured the underlying structure of experiences and processes evident in the raw data. We conducted within-narrative and between-narrative comparisons while developing these themes. The three authors then discussed the themes together until consensus was reached. To ensure the confidentiality and safety of the interviewees, we kept their identities anonymous when reporting the findings.

3.3 Findings

Drawing on interviews with journalists and news readers, the findings explore how individuals define, respond to, and correct misleading news headlines. Organized around two research questions, the analysis reveals important contrasts in how professional and everyday news consumers interpret headline credibility, experience emotional impact, and navigate correction strategies.

3.3.1 Understanding Misleading Headlines: Perceptions, Pressures, and Reader Reactions

This section synthesizes how journalists and readers perceive and respond to misleading headlines, highlighting the interplay between individual interpretation, institutional constraints, and emotional engagement. These findings offer insight into the human and systemic factors that shape trust and meaning-making in news consumption.

How Do People Perceive Misleading Headlines?

Participants across both groups shared varied and sometimes conflicting interpretations of what makes a headline misleading. This subsection explores their definitions, highlighting where journalists and readers align and diverge in conceptualizing misleadingness and assigning responsibility.

²<https://openai.com/index/whisper/>

Blurring the Lines: Conceptualizing Misleading Headlines Both journalists and readers struggled to distinguish between misleading headlines and concepts such as clickbait, sensationalism, and bias. While participants agreed that misleading headlines distort the reader's expectations, their definitions varied in scope and precision.

Journalists described misleading headlines as those that lack context, exaggerate minor details, or do not align with the article's main content. J2 (female, 18-23), a journalism intern, explained, "*You find an article, you read the headline, but whenever you read the content, it is different.*" Similarly, J6 (female, 18-23), a student journalist, noted, "*You end up being surprised by the content of the story; it's not what you expected from the headline.*" Others highlighted how even slight emphasis shifts can distort meaning. J1 (female, 18-23), a journalism student, remarked, "*They briefly mention that part in the article, but the real article is about something slightly different, that's a little misleading.*"

Several journalists also acknowledged the overlap between misleadingness and clickbait. As J4 (male, 60–65), a senior editor, reflected, "*A misleading headline can be clickbait-y... You can't say that all misleading headlines are clickbait, or all clickbait headlines are misleading.*" Sensationalism was also cited as a contributing factor. J10 (male, 70-75), a retired journalist, commented, "*The headline was a little more sensational than the actual article.*"

Readers echoed similar sentiments, often conflating misleadingness with clickbait, emotional manipulation, or partisan bias. P5 (female, 24-29) suggested, "*Misleading articles would be the category, and clickbait would be a subcategory... the reader is still being led to believe something else.*" Others emphasized interpretive distortion over factual inaccuracy. P1 (male, 18–23) explained, "*A headline might be misleading if it's biased—written in a way that favors one side.*" Despite this ambiguity, some readers offered clearer definitions. P8 (male, 24-29) stated, "*If the topic or content of the article is different from what the headline has stated, that could be misleading.*"

These accounts suggest that the perceived misleadingness of headlines is shaped not only by journalistic practices, but also by reader expectations, emotional framing, and prior knowledge. The conceptual blurring between clickbait, bias, and misleadingness reveals a challenge for both editorial consistency and reader discernment.

Whose Responsibility? Navigating Misleading Headlines as a Shared Burden Both journalists and readers reflected on the challenges of identifying misleading headlines, revealing differing views on where responsibility lies. Journalists acknowledged that subtle framing choices can make misleadingness difficult for the average reader to detect, particularly when translating complex issues like scientific research into accessible language. J1 (female, 18-23), a journalism student, explained, "*Misleading headlines aren't necessarily easily identifiable for people who don't have knowledge about it.*" J6 (female, 18-23), a student journalist, added, "*Scientists write very wordy, confusing papers, and then regular people try to translate it, it's just not understandable. That's where misinformation comes from.*"

Some journalists further emphasized that reader interpretation also plays a role in perceived misleadingness. J9 (male, 18-23), a student with internship experience, remarked, "*Sometimes people read things into the headline; they project assumptions, and that's not necessarily the headline's fault.*" While a few participants highlighted the importance of skepticism and selective reading as tools for navigating ambiguity (J3, female, 24-29; J4, male, 60-65), others stressed that

journalists must ultimately uphold headline accuracy.

Readers, however, placed primary responsibility on news creators. They expressed frustration that headlines often required additional scrutiny to assess their accuracy. P12 (female, 18-23) stated, *“It shouldn’t be on me to figure out if a headline is misleading. If I see a headline, I should be able to trust it tells me what the story is about.”* Similarly, P7 (male, 30-35) critiqued headlines that frame interpretation, noting, *“You read a headline and it kind of pushes you to think one way, even if the article doesn’t really say that.”* These responses suggest that while journalists see interpretive work as shared, many readers expect headlines to be inherently trustworthy and straightforward, indicating a potential mismatch between editorial practice and user expectations.

Journalists’ Perspectives: Institutional Pressures and Safeguards

Purpose Behind Misleading Headlines: Attracting Clicks, Driving Revenue, and Shaping Narratives Journalists often attributed misleading headlines to the pressures of digital media economics. Many acknowledged that headlines are crafted to emphasize the most attention-grabbing elements, even if those are not central to the article. J1 (female, 18-23), a journalism student, noted, *“They focus on a really interesting piece, even if it’s not the main part of the article.”*

Economic incentives were a recurring concern. J2 (female, 18-23), a journalism intern, explained, *“Newsrooms rely on advertising and clicks... it’s not necessarily to mislead, but to exaggerate facts so they’re more outrageous.”* Similarly, J5 (female, 24-29), a working journalist, said, *“You’re only going to get people to subscribe if your stories are compelling.”*

Some journalists also highlighted ideological motives. J3 (female, 24-29) remarked, *“It’s not just about clicks, it’s about pushing a message. The right wording gets people to support a cause.”* These accounts suggest that misleading headlines often arise not from negligence but from deliberate strategies shaped by economic and narrative goals.

Digital Pressures and Misleading Headlines: SEO, Social Media, and Paywalls Journalists described how digital infrastructures have reshaped headline production, introducing pressures that affect both accuracy and audience perception. Unlike print-era headlines, digital versions must optimize for algorithms, audience engagement, and platform constraints.

Search Engine Optimization (SEO) was framed as both a tool for accessibility and a source of distortion. J4 (male, 60-65), a senior editor, emphasized ethical SEO practice: *“It’s the antithesis of clickbait; you want both Google and your readers to recognize that what you think you’re getting is what you’re going to get.”* In contrast, others acknowledged how SEO can be exploited. J3 (female, 24-29) remarked, *“Some journalists will just put hot-button words in the headline... to be misleading.”* J6 (female, 18-23) added that keyword stacking can produce incoherent or misleading phrasing: *“It kind of becomes word salad... all the words technically fit the article, but they don’t convey the full context.”*

Social media platforms introduced further constraints. J12 (female, 35-40) noted that headlines are often accompanied by blurbs crafted to maximize attention: *“The blurb is the part that’s supposed to grab so much attention.”* These compressed formats incentivize emotional or provocative framings that can distort the story’s focus.

Finally, some journalists raised concerns about paywalls limiting access to full content. J9 (male, 18-23), a student journalist, observed, *“People just are not subscribed, so they can’t read*

it. It's very important information, but it's locked behind a paywall." This structural barrier may cause readers to rely solely on headlines, increasing the risk of misinterpretation.

Together, these insights reveal how digital distribution mechanisms (SEO, platform formatting, paywalls, etc.) create structural tensions between clarity, discoverability, and engagement, complicating the production of accurate yet effective headlines.

Journalistic Safeguards: Practices to Prevent Misleading Headlines Journalists described a mix of structural and individual safeguards against misleading headlines. In well-resourced newsrooms, editorial oversight plays a key role. J10 (male, 70-75), a retired journalist, shared, *"There was a desk that watched over standards and practices... You wouldn't think of exaggerating a headline for clicks."*

Others emphasized collaborative workflows. J5 (female, 24-29) explained, *"Writers suggest headline options and a digital team gives it a final sweep for SEO and clarity."* Still, such layers are not universal, and headline accuracy often depends on pace and capacity.

Certain types of reporting were seen as requiring extra caution in headline writing. For example, journalists covering crime or science topics described stricter norms around accuracy and phrasing to avoid premature or misleading claims. J3 (female, 24-29) noted, *"Even if we believe a death was a homicide, we won't call it that yet in the headline."* Similarly, J7 (male, 24-29) described avoiding exaggeration in climate reporting. These examples show that while policies help, headline integrity often relies on individual judgment and topical sensitivity.

News Readers' Reactions to Misleading Headlines

Readers reported strong emotional responses to misleading headlines, most commonly frustration, disappointment, and annoyance. While some described these reactions as short-lived, dissipating once the article clarified the story, others experienced more persistent irritation. P11 (male, 18-23) noted, *"I feel upset that it was misleading but then indifferent because now I know what happened."* In contrast, P2 (male, 30-35) said, *"I just feel disappointed, and I want to leave. I just leave the page."*

Repeated exposure to misleading headlines often leads to long-term distrust and avoidance. P3 (female, 18-23) explained, *"I'll probably not click on something that has a similar wording in the headlines."* Readers developed personal heuristics to detect unreliable content, relying on past experiences and intuition rather than formal credibility markers. P6 (male, 60-65) shared, *"'Conspiracy theory'; that's a keyword that just means to me that it's not necessarily pure news."*

Reactions also varied by topic. Headlines related to politics or science triggered stronger responses than entertainment news. P5 (female, 24-29) remarked, *"If it's about a celebrity, I kind of roll my eyes and get annoyed-but then I don't really mind."* Even trusted sources were not exempt from scrutiny, though prior favorability could temper long-term reactions. P3 (female, 18-23) reflected, *"If I feel like they're trying to mislead me, I'll probably read less—but maybe after some time, I'll forget this and start following that source again."* This suggests that while misleading headlines may momentarily damage trust, familiarity, and positive prior experiences can restore reader engagement over time.

These findings indicate that misleading headlines affect both immediate engagement and long-term trust. Emotional reactions, especially those linked to feelings of manipulation, can lead

readers to disengage from a news source. However, these effects may be temporary, as existing trust and familiarity with the source can help reduce the long-term impact.

3.3.2 Correction in Practice: How Journalists and Readers Revise Misleading Headlines

Participants across both journalist and reader groups proposed multiple strategies for correcting misleading headlines. Thematic analysis revealed five recurring correction practices: *reinstating missing context*, *removing inflammatory or misleading language*, *clarifying uncertainty*, *balancing curiosity and clarity*, and *envisioning participatory feedback mechanisms*.

Reinstating Context: Repairing Missing or Misleading Framing

A recurring corrective strategy among both journalists and general readers involved reinstating contextual information that had been omitted or oversimplified. Journalists described this process as a form of factual re-grounding, emphasizing the need to clarify causes, timeframes, or geographic scope. For example, a senior editor (J4, male, 60-65) noted, *“It’s missing critical context. . . This gives the unspoken impression that people were being reckless when that wasn’t the case.”* Similarly, P10 (male, 24-29) remarked, *“If the headline says it was ‘due to politics,’ then the article should say something about what those politics were. Since it didn’t, it felt misleading to me.”*

While both groups prioritized contextual integrity, their motivations diverged. Journalists focused on balancing accuracy with discoverability — J4 pointed out, *“It would not be discovered easily on search because it’s missing keywords like COVID.”* In contrast, readers emphasized interpretability and fairness. As P1 (male, 18-23) explained, *“Saying ‘multiple cases’ doesn’t tell you how serious it is. . . Saying ‘0.3% of passengers tested positive’ gives you a better sense of scale.”*

Repairing Language: Removing Inflammatory or Misleading Phrasing

Another prominent correction strategy involved softening emotionally charged or exaggerated language. Journalists described this as “tone calibration,” seeking to align language with the evidentiary weight of the story. J3 (female, 24-29) observed, *“This reads like investigative journalism, when in fact it is a fairly sensational and pointed opinion piece. . . designed to provoke outrage.”* J6 (female, 18-23) added, *“The word ‘warning’ felt like a red flag—it immediately read as sensationalized.”*

Readers likewise responded negatively to exaggerated language, especially when it created misleading expectations. P1 (male, 18-23) critiqued the phrase “wiped out” in a COVID-related headline: *“It makes it sound like protection is completely gone, which isn’t true. ‘Reduced’ would be more accurate.”* P11 (male, 18-23) commented, *“Calling it a ‘magic pill’ made it seem like there was some miracle COVID cure, but that’s not what the article was about.”*

Clarifying Uncertainty: Signaling the Limits of Evidence

Participants frequently corrected headlines by adding cues to reflect uncertainty—such as hedging language or attribution markers. Journalists described these as part of responsible editorial practice. J4 (male, 60-65) explained, “*We add terms like ‘reportedly’ or ‘may’ to avoid locking in something that’s not confirmed. It’s ethical and safe.*” J3 (female, 24-29) noted, “*Putting ‘probably’ in air quotes helps... It shows that this is someone’s interpretation, not an established fact.*”

Readers also valued uncertainty cues, particularly in relation to scientific reporting. P1 (male, 18-23) pointed out, “*The headline makes it seem like the link between COVID and brain damage is proven, but the research is ongoing. It should say ‘possible link’ or something like that.*” He added, “*You can’t make a sweeping claim based on two studies—especially when one only had 60 people.*”

Balancing Curiosity and Clarity: Tensions Between Correction and Engagement

A cross-cutting theme that emerged from both journalists and readers was the inherent tension between crafting a clear, accurate headline and preserving enough intrigue or emotion to drive engagement. This dilemma surfaced in reflections on whether corrected headlines still attracted clicks or prompted article reading. For example, J1 (female, 18–23), a journalism student whose family members had suffered from COVID-19, acknowledged that a less dramatic headline might reduce its immediate appeal, but personal relevance would still motivate her to read: “*Even if it didn’t have strong language, I’d still click it because it’s about COVID, and I want to see what the study says.*” Her comment reflects how engagement is not solely driven by headline sensationalism but also by topic relevance and reader context.

In contrast, some readers admitted that emotionally loaded or curiosity-driven phrases significantly shaped their decision to engage. P11 (male, 18-23) noted, “*I probably wouldn’t have clicked if it didn’t say ‘Bloody Secret’; it made me want to know what they were hiding.*” At the same time, participants expressed sensitivity to misleading cues: P8 (male, 24-29) shared that headlines including terms like “Warning” raised his suspicion even before reading, suggesting a credibility judgment formed at first glance. Journalists, too, acknowledged that excessive headline transparency may leave little incentive to read further. As J7 (male, 24-29), a public radio producer, put it: “*You can’t put everything in the headline, there has to be something in the content to make it worth reading.*” These comments reveal a design tension central to headline production and correction: balancing informative clarity and cognitive/emotional engagement.

Collaborative Correction: Envisioning Participatory Feedback Loops

Some participants, especially readers, proposed more participatory approaches to headline correction. They envisioned systems where audiences could offer feedback or view real-time editorial updates. P10 (male, 24-29) suggested, “*News sites could allow a space for readers to leave comments or suggestions under headlines... That could help writers improve in the future.*” P1 (male, 18-23) added, “*They could update the article later with comments that come in... that would build credibility, not just for the agency, but for the outlet.*”

While journalists were more reserved about public-facing feedback mechanisms, some acknowledged the role of post-publication updates. J4 (male, 60–65) noted, “*You can add an editor’s note when official statements come in later, like from the health department. That helps make the story more complete.*” These comments highlight a shared openness to collaborative correction, pointing toward design opportunities for interactive, transparent feedback systems.

3.4 Discussion

Our findings offer several insights into how journalists and news readers perceive, interpret, and attempt to correct misleading headlines. By examining both professional and audience perspectives, this study surfaces the competing expectations and constraints that shape headline production and consumption. The analysis revealed not only how misleadingness is conceptualized across roles but also the rhetorical and editorial strategies participants use, or wish to see used, for mitigating its effects.

3.4.1 The Need to Recognize Misleading Headlines as a Distinct Misinformation Threat

Participants in our study revealed a persistent difficulty in distinguishing misleading headlines from related forms of low-quality news, such as clickbait and sensationalism. Many readers described misleadingness in terms of emotional language or interpretive framing rather than factual inaccuracy, aligning with prior findings that users often rely on affective and stylistic cues when making credibility judgments [306]. This underscores a limitation of many media literacy interventions, which tend to focus on detecting outright falsehoods while overlooking subtler distortions like contextual omission or exaggerated framing [269].

Journalists, by contrast, articulated a more refined understanding of the boundaries between engagement and distortion, though they too acknowledged the influence of structural pressures. Some viewed attention-grabbing language as necessary for survival in algorithmically curated environments, while others voiced concern that performance metrics often incentivize borderline misleading practices. This reflects a systemic tension: digital news production is increasingly shaped by metrics that reward virality and emotional engagement over clarity and precision [59].

While prior research has examined misinformation through lenses such as community moderation [235], disinformation networks [376], fact-checking workflows [357], and headline-level linguistic analysis [72], most work isolates either the reader or algorithmic effects in abstraction. This study offers a more grounded perspective by bringing together two core stakeholders, journalists and news readers, to directly compare how misleadingness is perceived, justified, and contested. In doing so, we contribute to ongoing HCI and information studies literature that examines how different stakeholders make sense of sociotechnical artifacts in conflicting ways [133, 142, 337]. Our findings show that editorial intentions and public interpretations often diverge, complicating efforts to standardize corrective strategies or assign responsibility, especially in high-ambiguity domains like journalism and algorithmic curation [368].

These dynamics are further shaped by platform design. Headlines with emotionally charged or ambiguous phrasing may be disproportionately amplified by engagement-driven

algorithms [148], favoring attention over accuracy. While emerging approaches have called for recommender systems that consider news quality signals [251], few account for the specific harms posed by technically factual but misleading headlines.

Addressing this challenge will require interventions across both institutional and platform layers. Newsrooms should establish clearer editorial criteria to distinguish persuasive from misleading framing, particularly on sensitive topics. Platforms could experiment with framing disclosures, contextual overlays, or explainable headline histories to support user interpretation [414]. Recommender systems, too, may benefit from multi-objective optimization that balances engagement with epistemic integrity [251].

Overall, our findings suggest that misleading headlines represent a distinct class of misinformation, subtle enough to bypass conventional filters, yet powerful enough to shape perception and trust at scale. As headlines often serve as stand-alone informational artifacts in digital news environments, ensuring their accuracy is not just an editorial task but a civic one.

3.4.2 The Conflict Between Journalists and News Readers Over Misleading Headlines

Our findings highlight a persistent tension between journalists and news readers over who bears responsibility for addressing misleading headlines. Journalists frequently argued that readers should approach the news with greater skepticism, echoing prior work showing that media professionals expect audiences to develop critical literacy skills to navigate today's misinformation-rich environments [158, 223]. In contrast, readers in our study insisted that headlines should be inherently trustworthy and should not require additional interpretive effort. This reflects broader concerns about declining public trust in journalism [130], where even reputable sources are sometimes perceived as deploying strategic framings that manipulate rather than inform [269].

Such role-based conflicts are not unique to journalism. HCI and CSCW literature has long examined how sociotechnical systems can mediate disagreement and misalignment between stakeholder groups with divergent perspectives [165, 166]. Misleading headlines represent one such space of conflict, where journalists' editorial decisions and readers' interpretive expectations often clash in subtle but consequential ways. As headlines increasingly circulate as standalone informational units, particularly in algorithmically curated environments [136], even minor discrepancies between intent and interpretation can lead to erosion of trust and disengagement.

Addressing this divide requires deeper attention to the asymmetries in control, expectation, and interpretive labor between news producers and audiences. Rather than treating misleadingness as solely a content accuracy issue, future work should consider it as a product of relational breakdown, one that calls for better coordination across editorial and reader-facing systems.

3.4.3 Digitalization and the Evolution of Misleading Headlines

The digital transformation of journalism has introduced new structural pressures that shape how headlines are produced and consumed [18, 298]. Journalists in our study described how search engine optimization (SEO) [149], social media engagement metrics [298], and payroll constraints [276] influence headline writing, often incentivizing attention-grabbing language

over contextual accuracy. While SEO was considered essential for discoverability, participants expressed concern that algorithmic optimization encourages emotionally charged or trending keywords that may increase the risk of misleading interpretations [266]. These findings build on prior work examining how algorithmic visibility and audience analytics shape editorial decisions [97, 308], adding insight into how journalists experience these pressures as both strategic tools and ethical challenges.

Social media platforms were perceived as exacerbating this tension. Because headlines are often consumed in isolation, detached from the article body, journalists reported that visibility demands lead to a selective emphasis on provocative story elements. This aligns with research showing how platform design shapes audience attention and reshapes editorial practice [231]. Our study contributes by highlighting journalists' growing awareness of how adapting to these dynamics can unintentionally fuel misinformation.

Paywalls emerged as another structural barrier to informed news consumption. Journalists noted that many readers form judgments based solely on headlines or preview snippets, leading to fragmented or inaccurate understandings. This concern reflects prior work showing that readers frequently rely on headlines without accessing full content [111, 306]. By foregrounding journalist perspectives, our findings extend this literature to emphasize that paywall models affect access and contribute to distorted public interpretation.

These dynamics raise important implications for human-centered news platform design [36, 189]. Algorithmic systems that prioritize engagement must also account for the interpretive harm caused by misleading headlines. Interface-level interventions, such as transparency indicators [39], context-rich previews [412], or explainable ranking [392] may help mitigate distortion. Our findings also inform ongoing efforts to design adaptive paywall models that maintain access to essential information [90] and to optimize SEO in ways that surface credible news over sensational content [384]. As AI becomes more embedded in editorial workflows, future HCI research should explore how human-AI collaboration [247] can support journalistic integrity and mitigate misleadingness at scale.

3.4.4 Editorial Safeguards and the Limits of Manual Oversight

Journalists in our study described a range of safeguards to prevent misleading headlines, including editorial guidelines, newsroom review processes, and personal discretion. While larger news organizations often maintain dedicated standards desks, smaller outlets rely more heavily on individual judgment, making consistency difficult to sustain. Participants emphasized that headline writing is usually a collaborative process, ideally led by the article's author, to ensure alignment with content. Yet practical constraints such as SEO demands and character limits, as well as performance pressures from algorithmic visibility metrics, frequently compromise these intentions [279, 308].

Accuracy standards tend to be stricter in high-stakes domains like crime, health, or climate reporting, where misleading language can have legal or public health consequences. These patterns echo prior research showing that editorial precision is highly context-dependent and shaped by available institutional resources [152, 411]. Still, even in well-resourced environments, the fast pace of digital journalism and fragmented editorial structures can allow distortive headlines to pass through, primarily when engagement metrics compete with ethical clarity.

These findings point to the growing need for computational support tools to augment,

not replace, editorial judgment. As AI systems increasingly participate in content generation, ranking, and recommendation, ensuring headline integrity will require mechanisms for shared oversight [241, 288]. Designing assistive systems that detect rhetorical exaggeration, surface epistemic uncertainty, or offer real-time feedback on headline-article alignment could support more balanced decision-making. Rather than fully automating headline production, future systems must enable collaborative workflows where human expertise and algorithmic cues work in tandem to uphold journalistic standards in high-pressure, digital-first newsrooms.

3.4.5 Correction as Design: How Journalists and Readers Modify Misleading Headlines

Our analysis of how participants approached the correction of misleading headlines revealed three recurring strategies: *adding uncertainty cues*, *providing critical context*, and *removing emotionally charged or exaggerated language*. While journalists emphasized aligning headlines with underlying evidence and study limitations, readers often focused on reducing sensational tone or clarifying implied intent. This overlap suggests a shared but differently prioritized commitment to restoring interpretive accuracy. Prior research has shown that linguistic cues such as hedging and specificity influence perceived credibility and reduce overgeneralization [111], supporting that targeted language adjustments can serve corrective functions without sacrificing readability or engagement.

Importantly, neither group advocated for complete neutrality or oversimplification. Rather, participants sought a balance between informativeness and restraint, preserving clarity while avoiding distortion. Journalists described using quotation marks or explicit “opinion labeling” to signal uncertainty or subjectivity, whereas readers recommended replacing emotionally loaded terms with descriptive, evidence-based alternatives. These findings reinforce the view that headline correction is a form of editorial sense-making [115] shaped by competing goals such as accuracy, engagement, and interpretability rather than a binary fix.

Recent work has begun exploring the use of AI systems to automatically generate [122, 241] and evaluate [294] the quality of news headlines, offering promising directions for editorial support. However, such systems remain limited by a lack of understanding of how different misleading techniques, such as omission of critical context, use of emotion-evoking language, or overstated certainty, affect readers’ interpretive and behavioral outcomes, particularly in ambiguous or emotionally sensitive news contexts.

3.4.6 Toward Human–AI Collaboration in Editorial Practice

As artificial intelligence (AI) becomes increasingly embedded in editorial workflows [97], newsrooms are beginning to rely on automated systems to support tasks such as content generation, personalization & recommendation, and content summarization [288]. However, our findings suggest that integrating AI into these workflows demands more than efficiency or automation; it requires a careful understanding of the interpretive tensions, ethical constraints, and accountability expectations that shape headline production and perception.

This study reveals that misleading headlines are often the result of subtle linguistic strategies, such as omitted context, emotionally evocative language, or overstated certainty, that may be difficult to detect without domain expertise. Readers and journalists diverge not only in

how they identify these issues, but also in who they believe should be responsible for mitigating them. These role-based conflicts reflect broader sociotechnical dynamics that HCI and CSCW research have observed in other high-stakes domains [165, 166], where coordination between differently positioned actors, such as experts and users, can benefit from designed infrastructures for shared responsibility.

Motivated by this, we propose that editorial AI systems should be developed not as replacements for human judgment, but as collaborative agents that support more transparent, responsive headline production. For example, based on the correction strategies and perceived harms surfaced in this study, AI-assisted writing tools could flag potentially misleading phrasings, offer alternative framings with uncertainty cues, or suggest additions of critical context. These tools should prompt editorial reflection rather than prescribe fixed outputs, preserving journalistic discretion while scaffolding interpretive clarity.

This vision lays the groundwork for future work on human-AI collaboration in editorial settings [396], especially when the issue is subjective in nature. Subsequent investigations could explore how professional editors engage with AI-generated justifications or how algorithmic suggestions affect headline trustworthiness and reader comprehension. By translating patterns identified here into computational interventions, such work can deepen our understanding of how to design AI systems that align with the ethical, communicative, and editorial values of journalism.

3.4.7 Limitations

This study has several limitations that shape how the findings should be interpreted.

First, the sample size of both general readers and expert participants was relatively small, which may limit the generalizability of the results. While efforts were made to include diverse voices, the sample was drawn primarily from English-speaking participants with internet access, which may exclude important demographic or cultural perspectives. Although the insights gained are valuable, a larger and more diverse sample could provide a broader understanding of perceptions and reactions to misleading headlines across different sociocultural and media contexts.

Second, the study relies on self-reported data from interviews, which may be subject to biases such as social desirability or recall bias. Participants might have provided responses they believed to be socially acceptable or might not have accurately recalled their experiences and perceptions. In particular, when discussing sensitive topics like news credibility or trust, participants may underreport their own susceptibility to misinformation or overstate their media literacy.

Third, the study focused on a specific set of articles and headlines, which may not capture the full spectrum of misleading headline phenomena across different news organizations and media platforms. The selected headlines leaned toward soft news and general-interest topics, which may not evoke the same interpretive tensions as politically polarized or high-stakes headlines. Future research could benefit from analyzing a more extensive range of headlines and articles, including partisan, scientific, or crisis-oriented coverage, to understand the variability in misleading practices better.

Fourth, the rapidly evolving nature of digital media and news consumption means that findings may quickly become outdated. New technologies, changing algorithms on social media

platforms, and shifting consumer behaviors could alter the landscape of headline crafting and dissemination. For instance, generative AI is already being used in newsrooms to produce or tweak headlines, which may introduce new forms of ambiguity, framing, or bias that differ from human-generated practices. Continuous research is necessary to keep pace with these changes and to develop relevant strategies for addressing misleading headlines.

Fifth, while the study surfaced meaningful distinctions between how readers and journalists interpret headline responsibility, it did not formally assess these interpretations using controlled comparisons or experimental manipulation. Thus, causal claims about how role, expertise, or media literacy influence headline perception remain speculative. A follow-up study could incorporate factorial designs or role-switching experiments to isolate these effects more rigorously.

Lastly, the thematic coding process, though rigorous, was conducted by researchers with their own interpretive lenses, and subjective judgment may have influenced how themes were grouped or prioritized. While intercoder discussion and alignment were conducted, the lack of formal intercoder reliability metrics may raise concerns about replicability or researcher bias. Future studies might include multiple coders and use statistical measures (e.g., Cohen's κ) to increase transparency in the thematic analysis process.

3.4.8 Future Research Directions

Building on this study, future work can more systematically examine how specific headline correction strategies influence reader behavior, particularly in terms of trust, engagement, and interpretive accuracy. While our findings revealed a shared interest among both journalists and readers in restoring headline clarity through strategies such as adding uncertainty cues, removing emotionally charged language, and providing critical context, empirical evaluations are needed to understand the cognitive and behavioral consequences of these interventions at scale [192, 306]. Experimental and longitudinal studies could help quantify how different correction techniques shape perception in various news domains, especially in emotionally sensitive or ambiguous contexts.

There is also a growing need to explore how computational systems can support editorial decision-making without undermining journalistic autonomy. As AI-driven tools are increasingly used to generate [122, 241] and assess [294] headlines, future research should investigate how these systems can assist in surfacing potential misleadingness, identifying rhetorical exaggeration, or aligning headlines with article content. Designing assistive systems that enable human-AI collaboration, rather than full automation, offers a promising direction for supporting both efficiency and accountability in fast-paced newsrooms.

Finally, future studies should investigate the role of search engine optimization (SEO) and paywall infrastructures in shaping misleading headline practices. While SEO is often framed as a tool for discoverability, our findings suggest that its misuse, such as prioritizing trending or emotionally charged keywords, can inadvertently contribute to distorted framing. In-depth analyses of SEO strategies and their influence on headline integrity could inform more ethical optimization practices that preserve accuracy while maintaining visibility. Similarly, research is needed to examine how paywall models affect reader interpretation, especially when only headline-level content is accessible. Designing paywalls that balance financial sustainability with transparency and contextual access may help mitigate the unintended effects of headline-driven

misinformation.

3.5 Conclusion

This study advances our understanding of how misleading headlines are perceived and corrected by both journalists and readers, revealing deep-rooted tensions around responsibility, interpretability, and trust. While journalists face institutional and algorithmic pressures to craft attention-grabbing headlines, readers expect transparency and alignment with article content. These competing expectations underscore the broader challenge of balancing engagement with accuracy in digital news environments.

Our findings suggest that headline correction is not a simple act of clarification, but a site of editorial negotiation, shaped by structural constraints, audience interpretation, and linguistic nuance. Addressing these challenges will require not only clearer editorial standards but also design interventions that support mutual accountability. Future work should explore how correction strategies influence information behavior and how computational systems can assist in preserving headline integrity without diminishing human editorial judgment.

Ultimately, fostering trustworthy news environments will depend on collaborative efforts across journalism, platform governance, and HCI, where editorial expertise, reader needs, and algorithmic systems are thoughtfully aligned.

Chapter 4: Examining the Behavioral Impact of Correcting Misleading News Headlines

In Chapter 3, we explored how journalists and readers perceive and correct misleading headlines, identifying recurring patterns such as omission of context, emotionally charged language, and overstated certainty. These findings revealed not only the nuanced strategies that shape headline misleadingness but also the interpretive disconnect between producers and consumers of news. Building on these insights, this chapter investigates how specific correction strategies derived from participant-driven suggestions affect readers' interpretation accuracy, perceived credibility, and engagement. Through an experimental study, we examine whether such corrections can mitigate the harm of misleading headlines without sacrificing audience attention. This work was conducted in collaboration with Dr. Ronald Yaros and Dr. Naeemul Hassan and has been accepted for publication in the proceedings of the 17th International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2025), Multidisciplinary Track.

4.1 Introduction

Headline transparency is critical in shaping how audiences interpret and trust news content [39, 215]. While Chapter 3 examined how journalists and readers perceive misleading headlines and identified key correction strategies such as contextual clarification, tone reduction, and the addition of uncertainty cues, this chapter transitions from qualitative diagnosis to quantitative evaluation. Specifically, it empirically tests whether such corrections can improve readers' comprehension, credibility assessments, and engagement with news headlines.

In contemporary media environments, where headlines often circulate independently of full articles [136], the challenge of designing accurate and attention-grabbing headlines becomes increasingly salient [294]. This chapter addresses that challenge by investigating whether corrections derived from stakeholder-informed strategies can mitigate misleadingness without diminishing reader interest. We evaluate the effects of corrected headlines on three reader-centered outcomes: perceived credibility, interpretation accuracy, and engagement intentions.

4.1.1 The Effects of Headline Correction on Reader Engagement

In digital news environments, readers often rely on headlines alone to decide whether to engage with full articles [136]. As such, headlines are crafted not just to inform but to attract attention, frequently drawing on emotional language, ambiguity, or curiosity gaps to provoke engagement [42, 225, 410]. This focus on engagement, however, raises ethical concerns: when attention-grabbing strategies distort the meaning of the underlying article, headlines risk misleading readers.

As described in Chapter 3, both journalists and readers proposed correction strategies that adjust headline language to improve clarity, reduce emotional manipulation, and signal uncertainty. These include (a) adding hedging language to reflect the tentative nature of scientific claims, (b) restoring omitted context or source attribution, and (c) removing exaggerated or emotionally charged terms. While these strategies are promising for improving transparency, they may affect how engaging the headline appears to readers. This chapter empirically evaluates whether such corrective interventions reduce readers' curiosity to explore full news content.

RQ1: *What is the effect of headline correction on readers' curiosity to engage further with the news content?*

Grounding for H1.a (Uncertainty Correction). Uncertainty cues, aka Hedging, typically expressed through modal verbs or cautious phrasing (e.g., “may,” “possibly,” “suggests”), is a common feature in scientific reporting [26]. While earlier work suggested that hedging might weaken reader confidence [211], more recent studies argue that such language can increase perceptions of credibility, particularly when the uncertainty is attributed to experts or research findings [195]. Banerjee and Urminsky [32] further show that hedging does not inherently reduce engagement and may, in some cases, signal intellectual honesty, which sustains curiosity. Therefore, we propose:

- **H1.a:** *Correcting a headline by adding uncertainty cues **will not** significantly reduce engagement.*

Grounding for H1.b (Context Restoration). Context omission is a common strategy in misleading headlines, where headlines withhold critical information such as the source, scope, or reason behind a claim [73, 321]. However, studies have shown that readers value transparency and specificity, particularly when evaluating credibility [111]. Providing contextual cues such as naming the poll, clarifying causality, or including relevant statistics can preserve engagement while enhancing interpretive clarity [99]. Thus, we hypothesize:

- **H1.b:** *Correcting a headline by providing critical context **will not** significantly reduce engagement.*

Grounding for H1.c (Emotion Removal). Emotionally charged headlines are consistently shown to drive higher engagement by triggering affective arousal [50, 410]. Words like “shocking,” “heartbreaking,” or “explosive” increase the likelihood of clicks and shares, especially on social media [273]. While these terms may reduce accuracy, they enhance salience and user interest. Removing such language is likely to dampen emotional response and thus reduce engagement. Therefore:

- **H1.c:** *Correcting a headline by removing emotionally charged language **will** significantly reduce engagement.*

4.1.2 Effects of Corrective Strategies on Perceived Credibility

In digital journalism, credibility is a foundational concern for both readers and news organizations. Research has shown that headlines, often the primary point of engagement, play a disproportionate role in shaping readers' credibility judgments [111, 136]. Misleading headlines that omit context, exaggerate claims, or use emotionally charged language can trigger skepticism, erode trust, and reduce the perceived reliability of the broader news source [73, 273, 321]. Even when an article itself is accurate, incongruent or manipulative headlines can damage overall credibility [77, 398].

This study investigates whether correcting such misleading strategies can help restore credibility, leading to the following research question:

RQ2: *How does the correction of misleading headline techniques impact readers' perceptions of news credibility?*

Grounding for H1.a (Uncertainty Correction). A growing body of work has examined the effects of hedging, expressions of uncertainty or caution, in science and news writing. While hedges may suggest lower certainty, readers often interpret them as signs of transparency and intellectual honesty, especially when attributed to experts or sources [32, 195]. Rather than undermining credibility, hedging can signal responsible communication and acknowledgment of complexity.

H2.a: *Correcting a headline by adding uncertainty cues **will** increase perceived credibility.*

Grounding for H1.b (Context Restoration). Context omission is another common tactic in misleading headlines. Prior work shows that readers are sensitive to missing or ambiguous details and that such gaps can trigger doubt about the credibility of both the headline and the outlet [99, 321]. Headlines explicitly stating key information or attribution are perceived as more trustworthy, particularly in science and political news [73].

H2.b: *Correcting a headline by providing critical context **will** increase perceived credibility.*

Grounding for H1.c (Emotion Removal). Emotionally charged headlines, although effective for attracting clicks, often raise concerns about manipulation. Studies consistently show that sensational or emotionally arousing headlines are perceived as less credible than neutral ones [186, 373, 410]. This effect is especially pronounced in the news about controversial or serious issues, where readers expect objectivity and restraint [273].

H2.c: *Correcting a headline by removing emotionally charged language will increase perceived credibility.*

4.1.3 Effects of Corrective Strategies on News Headline Interpretation

Headlines are often the first and sometimes only piece of information that readers encounter when engaging with the news [104, 136]. They set the framing and tone for how a news story is interpreted, often exerting influence disproportionate to their length. Misleading headlines can distort interpretation by misrepresenting the core message, omitting key details, or using vague or emotionally charged phrasing that biases the reader's expectations [111, 395]. These distortions are especially concerning given the "primacy effect" in cognition, where initial impressions can shape subsequent understanding and recall [24, 275]. In the digital era, where headlines circulate widely on social platforms, such misinterpretations can scale rapidly, making headline clarity a vital issue for public understanding and discourse [136].

To investigate this, we pose the following research question:

RQ3: *What is the impact of correcting misleading techniques on readers' interpretation of news headlines?*

Grounding for H1.a (Uncertainty Correction). Ambiguous or hedged headlines have been shown to introduce multiple possible interpretations, which may either promote critical thinking or confuse the reader [32, 195]. However, when uncertainty is clearly attributed or contextualized (e.g., "scientists suggest," "early results indicate"), readers are better equipped to interpret the claim accurately. Rather than simply reducing certainty, such phrasing can clarify interpretive boundaries. This leads to the following hypothesis:

H3.a: *Correcting a headline by adding uncertainty cues will improve interpretation accuracy.*

Grounding for H1.b (Context Restoration). Providing missing context has also been shown to aid interpretation. Headlines that omit relevant detail, such as sample size, timeframe, or causal conditions, leave readers to fill in gaps, which can foster misinterpretation [321, 363]. Context-rich headlines support more accurate mental models of the underlying event or study, particularly for complex or technical topics [111]. This leads to the following hypothesis:

H3.b: *Correcting a headline by providing critical context will improve interpretation accuracy.*

Grounding for H1.c (Emotion Removal). Emotionally charged language may heighten attention but can also distort comprehension by steering readers toward pre-existing beliefs or exaggerated conclusions [186, 273]. Sensational or emotionally evocative headlines often prioritize affective arousal over semantic precision, leading to biased or oversimplified readings of the story [389]. Hence, we propose the following hypothesis:

H3.c: *Correcting a headline by removing emotionally charged language will improve interpretation accuracy.*

Although our hypotheses are grounded in prior literature that identifies how linguistic features like emotionality, uncertainty cues, and contextual framing shape engagement,

credibility, and interpretation (e.g., [32, 195, 410]), most existing studies either rely on large-scale clickstream or social media data [42, 225], or focus on isolated effects of headline style on perception [73, 273]. Few have systematically tested how corrective strategies, applied to originally misleading headlines, affect reader-centered outcomes in a controlled setting. By directly manipulating headline language based on real-world correction patterns and measuring their impact across engagement, credibility, and interpretation, this study provides empirical validation for design choices that are often assumed but rarely evaluated. This experimental approach bridges the gap between linguistic theory and applied news interface design, offering practical insight into how accuracy-enhancing strategies affect user behavior in real time.

4.2 Methods: Online Experiment

To address our research questions, we conducted an online experiment to examine how different correction strategies applied to misleading news headlines influence participants' interpretation accuracy, engagement intentions, and perceived credibility.

4.2.1 Study Design

The experiment employed a 3 (Correction Strategy: Critical Context, Stress-Word Removal, Uncertainty Cues) \times 2 (Version: Corrected vs. Misleading) between-subjects design.

4.2.2 Experimental Conditions: Headline Correction Strategies

Each condition presented three target news headlines that reflected one of the three editorial strategies. The goal was to evaluate how different correction approaches affect participants' credibility perception, engagement intention, and headline interpretation.

Below, we describe the design principles underlying each correction strategy and how they were applied in both the corrected and misleading versions of the headlines.

Providing Critical Context

This condition tested whether supplying additional context important to understanding the event could improve headline interpretation, credibility judgments, and audience engagement. Corrected headlines in this group added factual context that qualified broad or vague claims.

- **Misleading:** *Sleeping with the light on may be harmful to you*
- **Corrected:** *Sleeping with moderate light on may be harmful to you*

The corrected version specifies the conditions (i.e., moderate light) under which the potential harm was observed, providing necessary contextual boundaries.

Removing Stress-Evoking Words

This condition focused on reducing emotional arousal by eliminating charged or sensational terms while preserving the factual core. Emotional stressing words often exaggerate the seriousness of a news event, which can lead readers to misunderstand or misinterpret the actual situation. The goal of this correction strategy was to examine whether removing such emotionally loaded phrasing could sustain audience engagement, enhance perceived credibility, and support more accurate interpretation of the headline content.

- **Misleading:** *Bloody secret revealed: Blood of unborn cows used to create lab-grown meat*
- **Corrected:** *Blood of unborn cows used to create lab-grown meat*

The corrected headline removes the emotionally provocative phrase “Bloody secret revealed,” which may provoke unnecessary shock or outrage, while preserving the factual information about the ingredient source.

Adding Uncertainty Cues

This condition introduced epistemic markers (e.g., “possible,” “pending confirmation”) to reflect ongoing scientific investigation or lack of final conclusions. It evaluated whether introducing such uncertainty altered how participants judged the credibility or clarity of a claim.

- **Misleading:** *Scientists believe they’ve detected mysterious radio signal from alien planet*
- **Corrected:** *Scientists believe they’ve detected mysterious radio signal from alien planet, pending confirmation*

The corrected version communicates scientific caution by acknowledging the preliminary nature of the finding.

Participants were randomly assigned to one of six experimental conditions. Each condition presented three target news headlines manipulated according to the assigned correction strategy and version. Participants were not informed about which version (misleading or corrected) or which correction strategy (e.g., critical context, stress word removal, or uncertainty cues) they were viewing. This was done to minimize potential demand characteristics and ensure naturalistic responses. The full task took approximately 15 minutes to complete. The study protocol was reviewed and approved by the authors’ Institutional Review Board (IRB) in accordance with ethical guidelines for human subjects research.

Stimuli and Pretest

The experimental stimuli consisted of nine news headlines selected to represent three distinct correction strategies for mitigating misleading content: (1) *Providing Critical Context*, (2) *Removing Stress-Evoking Words*, and (3) *Adding Uncertainty Cues*. Each strategy included three unique headline samples, resulting in a total of nine base headlines. These were drawn from two sources: a dataset previously curated and annotated in Chapter 3 of this dissertation

and an independent online blog¹ that tracks and reports misleading headlines in real-world media coverage. To minimize the likelihood of strong identity-based or emotionally charged responses that could overshadow the headline correction effects, we selected headlines that predominantly focused on lifestyle, health, and science/technology topics [421]. The complete set of news headlines used in the study is provided in the Appendix.

For each of the nine misleading headlines, one of the authors generated a corrected version using the assigned strategy. Importantly, each headline was paired with only one correction strategy, and the author made minimal word-level changes to preserve the overall semantic structure and avoid introducing unintended variables into the manipulation. This resulted in 18 total headlines (9 misleading, 9 corrected), with each correction carefully crafted to isolate the strategic element being tested.

To validate the effectiveness of the manipulations, we conducted a small-scale pretest with 20 participants in a between-subjects design. Ten participants were randomly assigned to read the corrected versions of the headlines, while the remaining ten reviewed the misleading versions. Each participant answered a set of factual multiple-choice questions designed to assess their interpretation of the headline's intended meaning. These questions were developed to capture the degree to which participants inferred accurate information from the headlines alone without seeing the article content. Based on the pretest results, three headline pairs fell short of the ideal benchmark of at least 70% correct responses in the corrected condition and a meaningful improvement over the misleading version (defined here as a ≥ 20 percentage point gap). Based on this feedback, these headline pairs were reworded to reflect the intended correction strategy more clearly before being included in the main study.

Recruitment, Screening, and Compensation

Sample Size and Power Analysis An a priori power analysis [120] was conducted to determine the minimum sample size required to detect medium-sized effects ($d = 0.4$) with 80% power ($1 - \beta = 0.80$) and $\alpha = .05$. This effect size threshold was chosen based on prior work in news credibility and misinformation correction research, which often reports medium-sized effects for misinformation intervention strategies [236]. The analysis indicated that a minimum of 64 participants per group would be sufficient. To accommodate potential attrition, technical exclusions, or non-compliance, we increased the target sample size by approximately 10%, aiming for 70 participants per group. With six between-subject experimental conditions, this yielded a planned sample size of 420 participants.

Eligibility Criteria and Recruitment Procedure Recruitment was conducted through Amazon Mechanical Turk (MTurk) using built-in prescreening tools to ensure participant eligibility and demographic diversity. Eligible participants were required to (1) be at least 18 years old, (2) reside in the United States, (3) be fluent in English, and (4) have a HIT approval rate of 90% or higher. Compensation for participation was set at 2.15 for an estimated 15-minute task duration, in accordance with U.S. federal minimum wage guidelines.

Prior to launch, two volunteer pilot participants completed the full study and provided qualitative feedback on the survey structure, timing, and item clarity. Their completion times

¹<https://www.tomliberman.com/category/misleading-headlines>

averaged approximately 12 minutes and informed final adjustments to task design and timing benchmarks.

Screening and Data Quality Procedures To ensure data quality, two exclusion criteria were applied. First, responses from participants who completed the task in less than 10 minutes were removed, as this fell well below the average completion time observed during pilot testing. Second, three attention-check questions were embedded throughout the study to identify inattentive or automated responses. Participants who failed any one of these checks were excluded from the final dataset. After applying these exclusion criteria, a final analytic sample of 399 participants remained.

4.2.3 Variables

Manipulated Variables

The study employed a between-subjects design that manipulated two independent variables: the correction strategy used to revise misleading headlines and the version of the headline presented. The correction strategy variable had three levels, reflecting distinct editorial approaches: providing critical context, removing stress-evoking words, and adding uncertainty cues. The version variable had two levels, where participants were either exposed to corrected headlines or to their original misleading counterparts.

Dependent Variables

Behavioral Engagement (Headline Selection) To capture a behavioral indicator of curiosity and interest, participants completed a headline selection task prior to rating any individual headline. At the beginning of the study, participants were shown a set of four headlines and asked: “*If you were browsing these headlines, which one would you be most likely to click on?*” One of the four randomly positioned headlines was the target headline that the participant would later evaluate in detail. This design ensured that participants made their selection without knowing which headline they would subsequently engage with, preserving the validity of the behavioral measure.

To avoid drawing undue attention to the target headline, all four headlines in the selection set were carefully matched in both topical domain and manipulation type. For example, when the target headline used emotion-evoking phrasing, the three distractor headlines were also reworded to include emotionally charged language. This approach helped reduce saliency bias and ensured that the selection reflected participants’ genuine interest rather than headline distinctiveness. The selection of the target headline was coded as a binary variable (1 = selected, 0 = not selected) and served as a behavioral proxy for engagement. This method aligns with prior work leveraging headline choice to assess real-world curiosity and interest [192].

Perceived Credibility Participants rated the credibility of each headline using three items assessing perceived accuracy, authenticity, and believability. These items were adapted from Appelman et al. [22], who developed a validated scale to capture message credibility across media contexts. An example item includes: “*The headline appears to be factually correct and*

reasonable to me.” Each item was rated on a 5-point Likert scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree), and participants completed these items for three headlines in their assigned condition.

To construct an overall credibility measure, all nine credibility-related responses (3 items \times 3 headlines) were combined using a graded response model (GRM [341]). The GRM estimates a latent trait score for each participant by modeling both item discrimination and threshold parameters. The resulting credibility index represents an underlying judgment of perceived credibility across all headlines viewed. Scores were centered ($M = 0, SD \approx 1$) and ranged from -2.16 to 1.89 . Higher values reflect stronger perceptions of credibility. This approach aligns with prior research using latent trait modeling to evaluate headline credibility [192]. If corrections improve credibility, we expect participants in corrected headline conditions to report higher scores on this index than those in misleading conditions.

Engagement Intentions To assess participants’ intent to engage with the news content, we asked how likely they were to (1) read, (2) share, or (3) ignore each headline. These items have been widely used to capture curiosity-driven or behavioral engagement with digital news [192]. Participants rated each behavior on a 5-point Likert scale (1 = Not at all likely, 5 = Extremely likely) for all three headlines they evaluated.

All nine responses (3 items \times 3 headlines) were used to estimate a unidimensional engagement index using a graded response model [341]. Before modeling, the “ignore” items were reverse-coded so that higher values uniformly indicated higher engagement. The GRM-based index provides a latent score that reflects a participant’s overall propensity to engage with the headline content across multiple headlines and behaviors. Scores were standardized ($M = 0, SD \approx 1$), ranging from -2.51 to 2.21 , with higher scores indicating stronger behavioral interest in the headline. This latent modeling approach helps isolate general engagement tendencies while accounting for variability across items.

Interpretation Accuracy To assess participants’ comprehension of the headline’s meaning relative to the article content, we asked three multiple-choice questions per participant, each mapped to one of the three headlines they evaluated. These questions were adapted from established approaches to testing factual comprehension of news [389]. Each headline was paired with (1) a manipulation check question verifying whether the participant noticed the intended correction strategy, (2) a neutral question that both groups were expected to answer correctly, and (3) a discriminatory question designed to differentiate comprehension based on headline version.

Only the discriminatory questions were used to calculate the interpretation accuracy score. Each response was binary (correct = 1, incorrect = 0), and scores across the three headlines were summed to create an interpretation index ranging from 0 to 3. Higher scores indicate more accurate understanding of the article content. We expected participants exposed to corrected headlines to demonstrate higher interpretation accuracy than those shown misleading headlines.

Measured Variables

In addition to experimental manipulations, we collected several individual difference measures to explore how demographic and cognitive characteristics might influence responses

to misleading and corrected headlines. These included participants' age, gender, education level, political affiliation, news consumption habits, and media literacy.

Participants reported their age on a 5-point scale ($M = 2.22$, $SD = 0.54$), where higher values indicate older age brackets. Gender was reported as either male ($n = 316$, 79.2%) or female ($n = 83$, 20.8%). Education level ranged from some high school (2) to graduate degree (7), with most participants holding at least a college degree ($M = 5.17$, $SD = 0.59$). Participants also rated their political affiliation on a 5-point scale from very liberal (1) to very conservative (5), with the distribution skewed toward the conservative end ($M = 4.03$, $SD = 1.28$).

News consumption frequency was assessed on a 5-point scale (1 = Never, 5 = Multiple times daily), with participants reporting moderate to frequent consumption ($M = 3.61$, $SD = 1.02$).

Finally, to assess participants' ability to evaluate news and media content critically, we included a brief media literacy assessment using a validated short-form scale adapted from Maksl et al. [256]. The scale included three Likert-style items measuring behaviors like fact-checking, evaluating source credibility, and cross-referencing information. Participants responded using a 5-point scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree). To compute a latent measure of media literacy, we used a graded response model (GRM [341]), consistent with our scoring approach for credibility and engagement. The resulting media literacy index provided a latent trait score for each participant, reflecting their general capacity to engage with media critically. Scores were standardized ($M \approx 0$, $SD \approx 1$) and ranged from -2.15 to 1.85 . This index was used as a continuous predictor in subsequent models to examine whether media literacy moderated the effects of headline version and correction strategy on perceived credibility, engagement intentions, and interpretation accuracy.

In subsequent analyses, media literacy, political affiliation, and news consumption are included as covariates or moderators where relevant, particularly in models predicting perceived credibility and interpretation accuracy. These variables help contextualize how individual predispositions shape information behavior and responsiveness to headline corrections.

4.3 Result

4.3.1 Reader Engagement (RQ1)

This section addresses RQ1: *How does correcting misleading headlines influence reader engagement?* Engagement was measured in two ways: a latent self-reported engagement index derived from Likert-scale responses, and a behavioral proxy measuring whether participants selected the target headline to read.

Self-Reported Engagement

Engagement in this context was operationalized as a composite score derived from participant responses to three Likert-scale items measuring their likelihood to (1) read, (2) share, and (3) comment on each of the three target headlines they were shown. These items were adapted from prior studies on news interaction and were aggregated into a continuous latent engagement index using Item Response Theory (IRT), which accounts for variability in individual response patterns.

Table 4.1: Descriptive Statistics for Self-Reported Engagement (IRT Score)

Version	Strategy	Mean	SD	N
Correct	Context	-0.108	0.812	62
Correct	Emotion	0.098	0.899	63
Correct	Uncertainty	0.153	0.980	70
Misleading	Context	-0.095	0.783	65
Misleading	Emotion	0.089	0.888	69
Misleading	Uncertainty	-0.143	0.863	70

The experimental manipulation included two types of headline versions: participants either viewed the original misleading headlines or revised, corrected versions. These corrected headlines were modified using one of three strategies. The context strategy involved restoring omitted background information; the emotion strategy removed emotionally charged or stress-inducing language; and the uncertainty strategy introduced hedging cues or language that conveyed interpretive ambiguity.

Table 4.1 reports the descriptive statistics for the IRT-based engagement scores across all six experimental conditions. While the highest average engagement occurred in the Correct–Uncertainty condition ($M = 0.153$) and the lowest in the Misleading–Uncertainty condition ($M = -0.143$), the overall variation across conditions was modest.

To examine the effects of headline version and correction strategy on self-reported engagement, we conducted a two-way ANCOVA, controlling for participant gender and media trust, both known to influence patterns of news consumption.

The results revealed no significant main effect of headline version ($F = 1.59, p = .209$), suggesting that participants did not report lower engagement with corrected headlines compared to misleading ones. The main effect of strategy was also non-significant ($F = 1.42, p = .243$), indicating that none of the three correction approaches produced systematically different engagement outcomes. Similarly, the interaction between version and strategy was not significant ($F = 1.37, p = .256$), implying that the effectiveness of a correction strategy did not depend on whether the headline was misleading or corrected.

These findings suggest that corrections, whether achieved by adding context, softening emotional tone, or introducing uncertainty, do not reduce readers’ reported interest in engaging with the content.

Behavioral Engagement (Headline Selection)

In addition to self-reported ratings, we assessed engagement behaviorally by observing which headline participants chose to read first. At the start of the task, each participant was presented with a set of four headlines and asked to select one to read in full before providing any ratings. One of these was the target headline—manipulated either in its original misleading form or a corrected version using one of three strategies (restoring omitted context, removing emotionally charged phrasing, or adding uncertainty cues). The remaining three headlines served as distractors and were drawn from similar topics.

We used logistic regression to model the binary outcome of whether or not participants selected the target headline. Predictor variables included headline version (misleading vs.

corrected), strategy (context, emotion, or uncertainty), and their interaction. We also included covariates such as gender, political affiliation, media literacy, news consumption habits, and education level to account for potential individual-level differences in engagement behavior.

Table 4.2: Logistic Regression Predicting Behavioral Engagement (Target Selection)

Predictor	Coefficient	<i>p</i> -value
Intercept	2.269	.051
Version (Misleading vs. Correct)	0.156	.682
Strategy: Emotion (vs. Context)	0.793	.056
Strategy: Uncertainty (vs. Context)	0.365	.340
Version × Emotion	-0.574	.308
Version × Uncertainty	-0.189	.723
Media Literacy Index	0.026	.861
Political Affiliation	0.085	.333
News Consumption	-0.097	.394
Education	-0.329	.096
Gender (Male = 0)	0.041	.886

The results showed no statistically significant main effects of either headline version or correction strategy on headline selection behavior. However, the Emotion strategy demonstrated a marginally significant increase in the likelihood of selection compared to the Context condition ($b = 0.793$, $p = .056$). This suggests that headlines corrected by removing emotionally charged language might still draw attention or even slightly more than their misleading or context-corrected counterparts. No significant interaction effects were observed.

Hypothesis Interpretation

To assess whether the three headline correction strategies influenced engagement, we tested the following hypotheses using both self-reported and behavioral outcomes. The version variable distinguished whether participants saw the original misleading headline or a corrected version, while the strategy variable referred to the type of correction: context restoration, emotional tone removal, or uncertainty cue insertion.

H1.a: *Uncertainty correction will not decrease engagement significantly.*

Supported. Participants shown uncertainty-corrected headlines reported engagement levels similar to those exposed to misleading versions. The mean self-reported engagement score for the *Corrected–Uncertainty* condition was $M = 0.15$ ($SD = 0.98$), compared to $M = -0.14$ ($SD = 0.86$) in the *Misleading–Uncertainty* group. The interaction effect in the ANCOVA was non-significant ($F(2, 393) = 1.37$, $p = .256$), as was the behavioral interaction in the regression model ($b = -0.189$, $p = .723$). These results suggest that signaling uncertainty does not diminish engagement, supporting H1.a.

H1.b: *Context correction will not decrease engagement significantly.*

Supported. The Context strategy, used as the reference group, yielded no significant differences in engagement when comparing corrected vs. misleading versions. The mean self-reported engagement scores were nearly identical: $M = -0.11$ for the corrected version

and $M = -0.10$ for the misleading version. The logistic regression analysis also showed no significant differences in selection behavior. This provides consistent support that restoring omitted context does not reduce reader engagement.

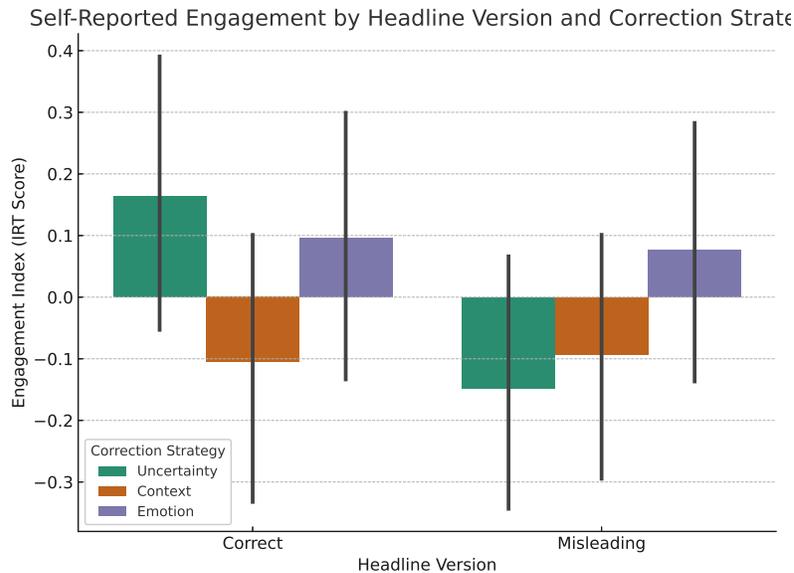


Figure 4.1: Self-reported engagement index by version and correction strategy. Error bars indicate 95% confidence intervals.

H1.c: *Emotion correction will decrease engagement significantly.*

Not supported. Contrary to this hypothesis, emotionally corrected headlines did not reduce engagement. Instead, the *Corrected-Emotion* condition produced a slightly higher mean self-reported engagement score ($M = 0.10$, $SD = 0.90$) than its misleading counterpart ($M = 0.09$, $SD = 0.89$). In the behavioral analysis, participants were marginally more likely to select emotionally corrected headlines ($b = 0.793$, $p = .056$). This challenges the assumption that emotional language is necessary for engagement, suggesting that ethical corrections may maintain or even enhance appeal.

Taken together, the results strongly support H1.a and H1.b: uncertainty and context corrections do not reduce reader engagement. H1.c is not supported, as emotionally corrected headlines did not suppress engagement and may have even increased behavioral selection rates. These findings challenge the assumption that emotionally manipulative wording is necessary to capture attention and instead suggest that transparent corrections can maintain or even boost engagement.

Correcting misleading headlines does not negatively affect engagement. Participants were equally likely, and in some cases even more likely, to engage with corrected headlines. This finding suggests that responsible headline correction can maintain audience attention without relying on misleading cues.

4.3.2 Perceived Credibility (RQ2)

This section addresses RQ2: *How does headline correction affect perceptions of the credibility of the news content?* Credibility was operationalized as a composite score derived from participant responses to nine Likert-scale items assessing dimensions such as accuracy, authenticity, and believability of each headline. These items were aggregated using a graded response model (IRT), generating a continuous credibility index suitable for statistical comparisons.

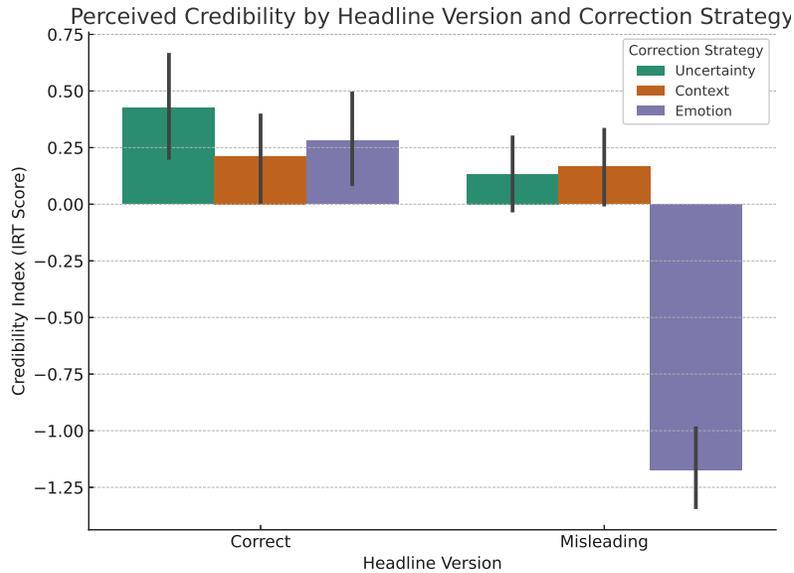


Figure 4.2: Perceived credibility (IRT score) by headline version and correction strategy. Error bars represent 95% confidence intervals.

As with the previous analysis, the version variable refers to whether the headline was shown in its corrected form or in its original misleading version. The strategy variable indicates which type of correction was applied, providing omitted context, removing emotionally charged language, or adding uncertainty cues.

Descriptive Statistics

Table 4.3 presents descriptive statistics of the IRT-based credibility scores across the six experimental conditions. Participants rated corrected headlines as more credible overall, especially when corrected using uncertainty cues. The Corrected–Uncertainty group had the highest average credibility ($M = 0.43$, $SD = 0.91$), while the Misleading–Emotion group had the lowest ($M = -1.17$, $SD = 0.76$), indicating that emotionally misleading headlines were perceived as the least trustworthy.

Inferential Analysis

A two-way ANCOVA tested the effects of headline version and correction strategy on perceived credibility, controlling for media literacy, political affiliation, news consumption,

Table 4.3: Descriptive Statistics for Perceived Credibility (IRT Score)

Version	Strategy	Mean	SD	N
Correct	Context	0.213	0.757	62
Correct	Emotion	0.281	0.853	63
Correct	Uncertainty	0.425	0.913	70
Misleading	Context	0.166	0.660	65
Misleading	Emotion	-1.172	0.759	69
Misleading	Uncertainty	0.132	0.727	70

education, and gender. Results are shown in Table 4.4. There were significant main effects for version ($F = 50.48, p < .001$), strategy ($F = 56.19, p < .001$), and their interaction ($F = 49.43, p < .001$). Media literacy was a strong positive predictor ($F = 133.62, p < .001$); other covariates were not significant.

Table 4.4: ANCOVA Results Predicting Perceived Credibility

Predictor	F	p-value
Version	50.48	< .001
Strategy	56.19	< .001
Version \times Strategy	49.43	< .001
Media Literacy Index	133.62	< .001
Political Affiliation	2.75	.098
News Consumption	0.31	.577
Education	0.45	.502
Gender	2.62	.106

Linear Regression

A linear regression model further explored these effects (Table 4.5). The *Version \times Emotion* interaction was significant and large ($b = -1.444, p < .001$), indicating a substantial drop in credibility for emotionally misleading headlines. Neither the *Version \times Uncertainty* nor *Version \times Context* interactions were significant. Media literacy remained a strong positive predictor ($b = 0.512, p < .001$).

Hypothesis Testing

To evaluate how each correction strategy influenced perceived credibility, we examined interaction effects between headline version (corrected vs. misleading) and strategy type (context, emotion, uncertainty). These interactions reveal whether credibility perceptions significantly differ based on whether a headline was corrected using a particular approach.

First, we hypothesized that *uncertainty-based corrections would significantly increase credibility* (H2.a). The data does not support this hypothesis. Although the *Corrected–Uncertainty* group had the highest mean credibility score ($M = 0.43, SD = 0.91$),

the interaction term between version and uncertainty strategy was not statistically significant in the regression model ($b = -0.097, p = .548$). This indicates that while participants rated these headlines slightly more credible, the correction did not produce a statistically distinguishable increase over the misleading version.

Table 4.5: Linear Regression Predicting Perceived Credibility

Predictor	Coefficient	<i>p</i> -value
Intercept	0.141	.674
Version (Misleading)	0.045	.701
Strategy: Emotion	0.007	.955
Strategy: Uncertainty	0.065	.577
Version × Emotion	-1.444	< .001
Version × Uncertainty	-0.097	.548
Media Literacy Index	0.512	< .001
Political Affiliation	0.044	.098
News Consumption	0.019	.577
Education	-0.038	.502
Gender (Male = 0)	0.137	.106

Second, we hypothesized that *context-based corrections would significantly increase credibility* (H2.b). This hypothesis is similarly rejected. The *Corrected–Context* group showed a modest increase in perceived credibility ($M = 0.21, SD = 0.76$), but no significant interaction was observed in the model, as context served as the baseline condition and no comparative strategy × version interaction approached significance. Thus, context corrections offered small gains but were not statistically meaningful.

Finally, we hypothesized that *emotion-based corrections would significantly increase credibility* (H2.c). This hypothesis is clearly supported. Emotionally misleading headlines received the lowest average credibility rating of all conditions ($M = -1.17, SD = 0.76$), and the *Version × Emotion* interaction was large and statistically significant in the regression model ($b = -1.444, p < .001$). This result indicates that correcting emotionally manipulative headlines had a strong positive effect on perceived credibility relative to leaving them uncorrected.

These results offer partial support for Hypothesis 2. Emotionally misleading headlines significantly reduced perceived credibility, and corrections targeting emotional language yielded the strongest and most statistically robust improvements. In contrast, corrections using context or uncertainty cues produced only modest gains and did not result in significant differences compared to misleading versions. These findings suggest that while emotion-based corrections are especially effective in restoring trust, other strategies may require more nuanced application or contextual tailoring. Additionally, media literacy emerged as an important moderating factor, with more literate participants consistently rating corrected headlines as more credible. Together, these results underscore that not all correction strategies are equally effective and highlight the need to match interventions to the type of misleadingness present.

4.3.3 Interpretation Accuracy (RQ3)

This section addresses RQ3: *How does correcting misleading headline techniques influence readers' interpretation accuracy of the underlying news content?* Participants responded to three factual multiple-choice questions per headline (one per article), which were scored as either correct (1) or incorrect (0). The final interpretation score ranged from 0 to 3 per participant.

The version variable indicates whether participants viewed misleading or corrected headlines. The strategy variable refers to the type of correction applied to misleading headlines, providing missing context, removing emotionally charged words, or adding uncertainty cues.

Descriptive Statistics

Table 4.6 shows mean interpretation accuracy across experimental conditions. Participants who viewed corrected headlines, especially those corrected using uncertainty cues, demonstrated substantially higher comprehension. The Corrected–Uncertainty group scored highest on average ($M = 2.03$, $SD = 0.62$), while the Misleading–Uncertainty group scored lowest ($M = 0.78$, $SD = 0.55$), illustrating the considerable impact of uncertainty-based corrections on interpretive understanding.

Table 4.6: Descriptive Statistics for Interpretation Accuracy

Version	Strategy	Mean	SD	N
Correct	Context	1.968	0.940	62
Correct	Emotion	1.822	0.831	63
Correct	Uncertainty	2.029	0.617	70
Misleading	Context	0.877	0.740	65
Misleading	Emotion	0.812	0.753	69
Misleading	Uncertainty	0.781	0.545	70

Inferential Analysis

We ran a two-way ANCOVA to test the effects of headline version (Corrected vs. Misleading), correction strategy (Context, Emotion, Uncertainty), and their interaction on interpretation accuracy. The model controlled for media literacy, political affiliation, news consumption, education, and gender. As shown in Table 4.7, there were significant main effects of version and strategy, as well as a strong interaction between them.

To explore these effects further, we ran a linear regression model including main effects and interactions (Table 4.8). The most notable result was a large and significant interaction between version and the uncertainty strategy ($b = -1.211$, $p < .001$), confirming that correction significantly enhanced comprehension. Emotion-based corrections also significantly improved interpretation accuracy ($b = 0.947$, $p < .001$).

Table 4.7: ANCOVA Results Predicting Interpretation Accuracy

Predictor	F	<i>p</i>-value
Version	251.51	< .001
Strategy	52.00	< .001
Version × Strategy	71.22	< .001
Media Literacy Index	1.21	.272
Political Affiliation	0.68	.409
News Consumption	0.19	.664
Education	3.33	.069
Gender	0.84	.360

Table 4.8: Linear Regression Predicting Interpretation Accuracy

Predictor	Coefficient	<i>p</i>-value
Intercept	2.634	< .001
Version (Misleading)	-1.757	< .001
Strategy: Emotion	-1.682	< .001
Strategy: Uncertainty	0.004	.983
Version × Emotion	0.947	< .001
Version × Uncertainty	-1.211	< .001
Media Literacy Index	-0.050	.272
Political Affiliation	-0.030	.409
News Consumption	0.026	.664
Education	-0.119	.069
Gender (Male = 0)	0.095	.360

Hypothesis Interpretation

H3.a: *Uncertainty correction will increase interpretation accuracy significantly.*

Supported. Participants exposed to headlines corrected by adding uncertainty cues demonstrated the highest comprehension performance across all conditions. The Corrected–Uncertainty group scored an average of $M = 2.03$ ($SD = 0.62$) out of 3 on factual interpretation questions, compared to only $M = 0.78$ ($SD = 0.55$) in the Misleading–Uncertainty group. This stark difference suggests that introducing uncertainty language, such as qualifiers and cautionary phrasing, enhances readers’ ability to extract accurate meaning from the news content.

The statistical evidence supports this conclusion robustly. A significant version \times strategy interaction was observed in both the ANCOVA ($F = 71.22$, $p < .001$) and linear regression model, where the interaction term *Version* \times *Uncertainty* yielded a large negative coefficient ($b = -1.211$, $p < .001$), indicating a sharp drop in accuracy when uncertainty cues were removed. In other words, the absence of uncertainty phrasing in otherwise identical content substantially reduced interpretive clarity. Together, these results provide strong empirical support for the efficacy of uncertainty-based corrections in restoring accurate understanding.

H3.b: *Context correction will increase interpretation accuracy significantly.*

Supported. When key contextual information, previously omitted in misleading versions, was restored in the corrected headlines, participants’ comprehension significantly improved. The Corrected–Context group achieved a mean accuracy of $M = 1.97$ ($SD = 0.94$), while the Misleading–Context group averaged only $M = 0.88$ ($SD = 0.74$), revealing a gap of over one full point on a 3-point scale. This large discrepancy indicates that the omission of critical background or explanatory content in headlines can lead to substantial misunderstanding of the article’s core message.

This interpretation is supported by significant main effects for both version and strategy in the ANCOVA model (Version: $F = 251.51$, $p < .001$; Strategy: $F = 52.00$, $p < .001$), and a highly significant interaction term ($F = 71.22$, $p < .001$). The regression model also confirmed this pattern, with the corrected version providing significant gains in comprehension across all strategy types, including context. While the interaction effect for *Version* \times *Context* was not explicitly labeled in the regression table, the improvement is clearly reflected in the magnitude of the main effect and post hoc contrasts. Overall, these findings validate that restoring missing context is a powerful corrective strategy that meaningfully improves reader interpretation.

H3.c: *Emotion correction will increase interpretation accuracy significantly.*

Supported. Emotionally charged headlines that use dramatic or sensational language significantly impair reader comprehension. The Misleading–Emotion group scored only $M = 0.81$ ($SD = 0.75$), while the Corrected–Emotion group improved to $M = 1.82$ ($SD = 0.83$), nearly doubling the mean interpretation score. This suggests that emotionally evocative phrasing can distract readers from the article’s factual core or bias their interpretation, whereas more neutral phrasing helps focus attention on content.

Statistical results reinforced this interpretation. In the linear regression model, the interaction term *Version* \times *Emotion* was highly significant and positive ($b = 0.947$, $p < .001$), indicating that removing emotional language substantially improved comprehension relative to the misleading version. While the misleading headlines may have generated more curiosity or emotional arousal, these effects appear to come at the cost of factual understanding. Thus,

corrections targeting emotional manipulation offer a dual benefit: preserving attention while enhancing interpretive clarity.

All three hypotheses under RQ3 were supported. Each correction strategy, such as adding uncertainty, restoring context, and removing emotionally manipulative language, led to significant improvements in readers' ability to accurately interpret the underlying news content. Among the three, uncertainty-based corrections had the largest impact, suggesting that hedging language plays a particularly important role in aligning reader expectations with journalistic intent. These results emphasize the potential of linguistically grounded correction techniques to counteract subtle forms of misinformation without sacrificing reader comprehension.

4.4 Discussion

4.4.1 Summary of the Key Findings

This study examined how different strategies for correcting misleading headlines affect readers' engagement, perceived credibility, and interpretation accuracy. Using a between-subjects experimental design, we tested three correction strategies, *Providing Context*, *Removing Emotion-Evoking Language*, and *Adding Uncertainty Cues*, across six conditions involving corrected and misleading versions of news headlines. Three primary research questions guided this investigation, each targeting a distinct but interrelated aspect of user response.

Engagement

Contrary to concerns that removing emotionally charged or curiosity-driven language may reduce reader interest, our findings show that headline correction does not diminish engagement. Both self-reported engagement scores and behavioral headline selection rates remained statistically consistent across all conditions. Notably, emotionally corrected headlines exhibited a marginal increase in behavioral engagement, suggesting that transparency and factual clarity may retain or even enhance click-through interest. These findings offer evidence against the presumed trade-off between accuracy and engagement that has historically underpinned sensational headline design [42, 225].

Perceived Credibility

Headline corrections led to a substantial increase in perceived credibility. Among the three correction strategies, *Removing Emotion-Evoking Language* produced the most pronounced credibility gain, especially when compared to emotionally misleading headlines. While uncertainty and context corrections showed modest improvements, only the emotion strategy demonstrated a statistically robust interaction effect. This pattern aligns with prior research showing that emotionally charged headlines can erode perceived trustworthiness [3, 221], but also suggests that their effects are more malleable—making them especially responsive to corrective interventions.

Interpretation Accuracy

Correcting misleading headlines significantly improved participants' factual understanding of the content. All three strategies were effective in increasing interpretation accuracy, but the strongest gains were observed with the *Adding Uncertainty Cues* strategy. Participants in this condition scored more than two full points higher (on a 0–3 scale) than those exposed to misleading uncertainty-framed headlines. This finding is consistent with prior work showing that uncertainty cues in headlines can cause readers to misremember factual content as speculative [51]. By reframing such cues with clearer language, the correction appears to recalibrate readers' information processing, leading to significantly more accurate interpretations. This supports the idea that lexical markers of uncertainty not only shape individual comprehension but also influence how readers emotionally and socially respond to information [292].

Strategy-Specific Insights

While all three correction strategies were effective to varying degrees, each appeared to exert its strongest influence on a different outcome measure. The emotion strategy yielded the largest gains in credibility, uncertainty correction produced the greatest improvement in interpretation, and all strategies maintained engagement. These findings suggest that correction strategies are not universally interchangeable; rather, their efficacy depends on the communicative goal, be it restoring trust, improving comprehension, or maintaining interest. Taken together, the results support the broader argument that headline corrections can promote informed news consumption without sacrificing audience engagement. These findings contribute empirical evidence to ongoing debates in journalism and information science around balancing attention economics with accuracy and call for a more nuanced understanding of how readers process headline-level misinformation [154, 388].

4.4.2 Implications

This study offers key contributions to the field of information behavior by demonstrating how subtle linguistic changes in headlines shape readers' evaluation and understanding of news [46, 111, 253]. The findings challenge the assumption that emotionally toned or ambiguity-driven headlines are necessary to sustain attention [42]. Engagement remained stable across all correction strategies, suggesting that curiosity and relevance, not just emotional manipulation, drive news interaction. Furthermore, credibility judgments reflected sensitivity to affective cues, with emotionally misleading headlines substantially lowering trust but recovering when corrected. The most striking impact was observed in interpretation accuracy, particularly for uncertainty corrections, which yielded the largest comprehension gains. These results highlight the role of lexical framing [118] in how readers construct meaning and underscore that strategic correction [112, 306] can promote better interpretive alignment, supporting more informed news consumption behaviors.

Beyond information behavior, this research carries broader implications for journalism practice and digital communication. First, it challenges editorial norms that prioritize engagement metrics over clarity by demonstrating that accuracy need not compromise audience interest.

Second, it offers evidence-based guidance for journalists and platform designers on how correction strategies can vary in their effectiveness depending on whether the goal is to improve trust or comprehension. Finally, these findings contribute to ongoing conversations about mitigating subtle misinformation [16, 111], showing that headline corrections can serve as lightweight but powerful interventions, preserving trust and enhancing public understanding in an age where readers often consume content without clicking through.

4.4.3 Limitations

Several limitations should be noted in interpreting the findings of this study.

While the study was powered to detect medium-sized effects, it may have missed more subtle differences across subgroups or strategies. In particular, effects related to participant traits such as political ideology, media literacy, or news consumption habits may interact with correction strategies in nuanced ways that this design was not sufficiently powered to detect. Future studies should consider larger samples or stratified designs to examine interaction effects more precisely.

Moreover, interpretation accuracy was measured using factual multiple-choice questions, which may not fully capture the nuance of readers' inferential or contextual understanding. While this approach ensured objective scoring, it may oversimplify how readers extract meaning or identify bias, especially in ambiguous headlines. Open-ended or justification-based comprehension measures may offer richer insights into interpretive accuracy.

The study focused on soft news topics to avoid ideological bias; thus, results may not generalize to more polarized or politically charged content. Corrections to politically loaded headlines may provoke stronger resistance or motivated reasoning from readers, altering the impact of the correction strategies observed here. Future work should test whether the findings hold across domains such as political news, public health, or science misinformation. Participants evaluated headlines in isolation, without accompanying article text, which mirrors real-world browsing but may limit contextual interpretation. This approach enhances ecological validity for fast-paced digital environments, such as social media feeds, but it also constrains the depth of readers' comprehension by omitting fuller article context. Exploring how correction strategies interact with article content, visual framing, or comment sections would provide a more comprehensive understanding.

Although the correction strategies were informed by both empirical interviews (Chapter 3) and prior literature, their implementation relied on the researcher's judgment. This may introduce interpretive bias in how "context," "emotion," or "uncertainty" were operationalized in headline revisions. Future work should involve journalists or editors directly in the correction design process to enhance editorial realism and validity.

The measurement of behavioral engagement had constraints that may limit its ecological validity. Participants were asked to choose one headline to read from a set of four, which may have artificially forced a choice that doesn't reflect real-world browsing behavior, where skipping all headlines or engaging with multiple is often an option. Moreover, the behavioral measure only captured headline selection without accounting for other meaningful engagement actions like sharing, ignoring, or bookmarking. These dimensions are crucial for understanding how readers interact with news in practice.

Lastly, the between-subjects design limited each participant to only one condition, which

avoids fatigue or order effects but also increases individual-level variability. Within-subjects designs or mixed-method approaches could provide more sensitive comparisons by leveraging each participant's baseline engagement or comprehension tendencies across conditions.

4.5 Conclusion

This study examined how three correction strategies shape reader responses to misleading headlines by adding uncertainty cues, providing critical context, and removing emotion-evoking language. Across a controlled experiment with approximately 400 participants, we found that corrections can improve interpretation accuracy and perceived credibility without diminishing engagement. Each strategy demonstrated distinct strengths: uncertainty cues improved factual understanding most effectively, emotional corrections most strongly restored credibility, and all strategies preserved or modestly enhanced interest.

These findings offer practical implications for editorial decision-making and contribute to broader conversations about mitigating subtle misinformation in digital news. Future work can extend this foundation by exploring correction strategies in more interactive, real-world contexts and by applying them across a wider range of news topics. As both human editors and AI systems play growing roles in shaping headline content, this work provides behavioral insights that can inform human-AI collaborative pipelines for identifying and revising misleading headlines, an area further developed in the third study of this dissertation.

Chapter 5: Human–AI Editorial Collaboration: A Situated Evaluation of LLM Explanations for Misleading News Headlines

In Chapter 3, we examined how journalists and news readers perceive and attempt to correct misleading headlines, uncovering tensions in interpretive expectations and editorial responsibility. These findings highlighted the nuanced strategies used to convey or obscure meaning, such as omission of context, exaggerated language, or overstated certainty, and raised important questions about how such ambiguity might be addressed through computational support. Building on this foundation, this chapter explores the role of large language models (LLMs) in editorial workflows by (a) generating explanations for misleading headlines and (b) evaluating the quality of these explanations through expert journalist assessments, with attention to correctness, ambiguity sensitivity, and editorial utility. This work was conducted in collaboration with Md Mahfuzul Haque, Mohammad Ali, Ahmed Shatil Alam, and Dr. Naeemul Hassan. A portion of this research has been accepted for presentation at the CHI 2025 Workshop on Human-centered Evaluation and Auditing of Language Models (HEAL@CHI’25).

5.1 Introduction

Large language models (LLMs) have achieved impressive performance across a growing range of natural language processing tasks, including fact-checking [322], summarization [248, 438], scientific writing [75], and legal reasoning [210]. Their capacity for fluency [234], contextuality [20], and generalization [60] has led to widespread integration in organizational workflows such as business [38], enterprise [444], software engineering [281], journalism [65, 196], and education [79]. This increasing adoption has opened up new opportunities for Human-Artificial Intelligence (AI) collaboration, where AI systems support rather than replace expert decision-making.

In human-AI collaboration, a critical prerequisite for effective partnership is that human users can interpret and make sense of AI-generated outputs. This need has led to growing interest in AI explainability, where systems are designed to provide transparent rationales that help users understand how outputs are derived [23, 110, 242]. Researchers have actively examined how these explanations affect the collaboration process [407, 416], particularly their influence on trust [128, 175], decision quality [417], and user reliance [71].

In objective domains such as fraud detection [81] or medical diagnosis [11], explainable AI (XAI) has been shown to enhance human decision-making by offering clear, verifiable justifications. Explanations in these settings enable users to calibrate trust, interrogate model reasoning, and assess the actionability of predictions. However, in more subjective decision-making tasks, explanations serve additional purposes. They can support argumentative reasoning by presenting multiple rationales, helping users weigh alternative interpretations or outcomes

[126, 230]. Yet, the effectiveness of explanations becomes more complex in domains characterized by ambiguity, contested knowledge, or interpretive nuance. In such contexts, recent studies have found that LLM-generated explanations, despite their fluency and surface-level plausibility, can mislead users, obscure uncertainty, or foster unwarranted trust in incorrect outputs [96, 352]. These risks raise an important challenge for human-AI collaboration: when and how should users trust AI explanations, particularly in settings characterized by subjective ambiguity, and to what extent are these explanations sufficiently actionable to support decision-making?

This work explores the challenges of trusting and acting on AI-generated explanations in the context of professional editorial practice, particularly under conditions of subjective ambiguity. We focus on the task of detecting misleading news headlines, a subtle but consequential form of misinformation [420]. Misleading headlines typically distort, exaggerate, or omit key contextual details, shaping public perception in ways that diverge from the article’s actual content [72, 111]. While they may not contain explicit falsehoods, their societal impact can be just as serious as that of overtly false information. For instance, during the COVID-19 pandemic, misleading headlines circulated more widely than flagged fake news and played a significant role in fostering vaccine skepticism [16]. Given their reach and influence, addressing misleading headlines is not merely a matter of technical accuracy but a core responsibility of journalistic integrity, central to the profession’s role in preventing the spread of misinformation and maintaining public trust [49, 70]. Unlike binary fact-checking, detecting misleading headlines requires understanding framing [217], context omission [82], and tone [72]. These are rhetorical elements that are open to interpretation [48]. As a result, even trained human annotators frequently disagree on whether a headline is misleading [380]. Such disagreement should not be treated as noise but rather as evidence of the subjective nature of the task [343].

In such subjectively complex domains, the quality of AI-generated explanations cannot be judged solely by correctness or consistency with a presumed ground truth [370]. Instead, it must be evaluated through a situated lens that reflects professional norms, editorial judgment, and contextual reasoning [399]. This study investigates how AI, particularly large language models (LLMs), can assist expert editorial decision-making in identifying ambiguous forms of misinformation, such as misleading headlines, which serve as a compelling testbed for examining the potential and limitations of human-AI collaboration in subjectively interpretive settings. These cases involve contested meanings, rhetorical framing, and selective omission, challenges that require human expertise to interpret. While human-AI collaboration has been explored in other subjective domains such as medical data analysis [348], hate speech detection [125], and hiring decisions [286], its application within journalism, particularly in editorial tasks involving misinformation, remains relatively underexamined.

To address this gap, we leverage the subjective nature of misleading headline detection by operationalizing ambiguity through varying levels of agreement among human annotators. Using the dataset we curated in Chapter 3, we analyze 60 real-world headlines that were annotated for misleadingness by three human raters and explained by three large language models (LLMs): GPT-3, GPT-4, and Gemini. By stratifying these headlines based on annotator agreement, distinguishing between high-agreement (unanimous) and low-agreement (ambiguous) cases, we explore how subjective ambiguity affects the interpretive validity of LLM-generated explanations. Specifically, we ask whether LLMs align more closely with human reasoning in cases of high consensus or whether ambiguity introduces failure modes

such as overgeneralization or misplaced confidence. Also, we introduce an expert-in-the-loop evaluation protocol in which six professional journalists assess the quality of LLM explanations using a structured rubric. This rubric includes seven dimensions: correctness, clarity, exhaustiveness, ambiguity awareness, risk of harm, trustworthiness, and actionability, which are adapted from prior research on explainable AI (XAI) [213, 352] to reflect editorial reasoning standards. In addition to these dimensions, experts also responded to four open-ended prompts capturing perceived weaknesses, missing elements, suggested improvements, and the explanation’s handling of ambiguity. This framework allows us to investigate how human disagreement moderates the alignment between LLM rationales and expert journalistic judgment, as well as which explanation characteristics most influence expert trust and perceived editorial usefulness.

Additionally, we conduct a post hoc comparative analysis between LLM- and human-generated explanations to assess how well models surface key editorial concerns. Drawing from journalism literature, we focus on framing bias [118, 152], omission of context [82, 387], and emotional manipulation [260], elements frequently implicated in misleading headlines. This analysis offers insight into the editorial completeness and interpretive sophistication of AI-generated rationales.

Overall, our study adopts a triangulated evaluation approach [94] that integrates three perspectives: human annotations, LLM explanations, and expert evaluations. This design enables a more comprehensive and situated understanding of AI system behavior, particularly in socially complex tasks where no single ground truth suffices [93].

The study is structured around four interlinked research questions that collectively examine when LLMs align with expert reasoning, how their explanations diverge in ambiguous cases, and what characteristics most influence expert trust and collaborative potential in editorial contexts.

- **RQ1:** *How does the level of human annotator agreement on the misleadingness of news headlines relate to the classification accuracy and expert-perceived quality of LLM-generated explanations?*
- **RQ2:** *In cases of partial or low agreement among annotators, how do LLM-generated explanations diverge from expert journalistic reasoning?*
- **RQ3:** *What explanation characteristics, such as correctness, exhaustiveness, ambiguity awareness, and risk sensitivity, most influence expert trust and editorial utility?*
- **RQ4:** *How do human- and LLM-generated explanations differ in surfacing editorial concerns such as framing bias, omission of context, and emotional language?*

We combine quantitative analysis of model accuracy and expert ratings with qualitative insights from journalist commentary. Our findings show that LLMs perform well in high-agreement cases but falter under ambiguity, often failing to surface context or qualify uncertainty. Expert trust was most strongly associated with explanation correctness, while actionability was better predicted by ambiguity awareness and risk of harm sensitivity. Human-written explanations more reliably surfaced in editorial concerns than those generated by LLMs.

This study presents several significant contributions. First, we introduce a stratified evaluation framework for assessing LLM explanations that consider the varying levels

of agreement among annotators, addressing the subjective ambiguity often present in misinformation detection. Second, we adopt an expert-focused rubric from the explainable AI (XAI) literature to evaluate the quality of explanations in the context of editorial decision-making. Third, we perform a triangulated analysis that compares LLM-generated rationales with both human-sourced annotations and evaluations from professional journalists. Fourth, we provide empirical insights into which characteristics of explanations, such as correctness, awareness of ambiguity, and sensitivity to harm, influence expert trust and perceived actionability. Finally, we outline design guidelines for human-AI collaboration systems aimed at enhancing editorial transparency, accountability, and context-sensitive decision-making.

5.2 Method

This study adopts a mixed-methods, expert-in-the-loop evaluation framework to assess how large language models (LLMs) generate explanations for misleading news headlines under conditions of subjective ambiguity. We build on an existing dataset of real-world headlines previously annotated by human workers for misleadingness and stratify the data based on annotator agreement to operationalize ambiguity. To assess the editorial quality of model-generated explanations, we engage professional journalists as expert raters and apply a structured rubric adapted from explainable AI (XAI) literature. Our analysis integrates both quantitative ratings and qualitative feedback to surface strengths, weaknesses, and failure modes of LLM-generated rationales across varying ambiguity conditions. This approach reflects a situated evaluation strategy grounded in editorial practice and collaborative decision-making.

5.2.1 Dataset

We use the misleading news headline dataset introduced in Chapter 3, which comprises 60 real-world news headlines, each annotated by three human annotators for misleadingness and accompanied by free-text justifications. Annotators classified each headline as either misleading or not misleading, yielding four agreement levels across the dataset: 18 headlines received unanimous misleading labels (3/3), 9 were labeled misleading by two annotators (2/3), 10 by one annotator (1/3), and 23 were unanimously labeled as not misleading (0/3). These patterns of agreement provide a natural proxy for subjective ambiguity, with partial agreement cases reflecting higher interpretive uncertainty.

Each headline in the dataset was independently classified by three large language models (LLMs): GPT-3.5, GPT-4, and Gemini 1.0. Among them, GPT-4 and Gemini 1.0 represent the state-of-the-art LLMs available at the time of the study. Each model generated a binary classification (misleading or not misleading) along with a free-text explanation justifying its decision. To facilitate this, the headline and its corresponding news content were submitted to each LLM via an API request.

The models were prompted to assess whether the headline accurately represented the underlying article and to explain their reasoning. All LLMs were used with their default API configurations, such as default temperature, token limits, and decoding parameters, without any fine-tuning or iterative prompt adaptation. A sample prompt used in the request is as follows:

prompt= "Evaluate if the following headline is misleading based on the news content

Table 5.1: Example Headline and Corresponding Labels and Explanations from Human and LLMs

Headline	Pandemic deaths among pregnant Americans are now ‘surging’	
Source	Label	Explanation
Human (3/3)	Misleading	I was expecting to see the percentage of pregnant women dying because of COVID-19 increase in the news content. Although the article mentioned that the number of pregnant women dying is increasing (around 40%), it did not show/claim how many of them are directly caused by COVID. So, the death toll is rising due to COVID is unproven here.
Gemini	Misleading	The headline does not provide any context or explanation for the surge in deaths. The content of the article makes it clear that the surge is due to the Delta variant of COVID-19, which is more transmissible and can cause more severe symptoms in pregnant women. The article also explains that the surge in deaths is happening in Republican-voting states and counties where pandemic safety measures like masking, social distancing, and vaccinations have been mocked or intentionally blocked by Republican leaders.
GPT	Not Misleading	The headline accurately reflects the main points of the article content, where it is mentioned that COVID-19 has led to an increase in deaths among pregnant Americans. It is also highlighted that significant resistance to safety measures like masking and vaccination are contributing to these fatalities.

provided. *Headline: [Your Headline Here] News Content: [Your News Content Here] Is this headline misleading? Please explain your decision.*”

To evaluate model performance, we compared each LLM’s output against human annotations in a binary classification task. We adopted a conservative labeling rule: a headline was considered “misleading” if at least one of the three human annotators marked it as such; it was considered “not misleading” only if all three agreed it was not misleading.

We assessed each model using standard classification metrics-precision, recall, F1-score, and overall accuracy. Table 5.2 presents the comparative results. GPT-4 and Gemini 1.0 achieved relatively strong performance, whereas GPT-3.5 performed poorly, particularly due to its extremely low recall for non-misleading headlines.

Table 5.2: Performance of LLMs in Detecting Misleading News Headlines

Model	Non-misleading			Misleading			Accuracy
	Precision	Recall	F1	Precision	Recall	F1	
GPT-3.5	1.00	0.09	0.16	0.46	1.00	0.63	0.48
GPT-4	0.85	0.97	0.90	0.95	0.77	0.85	0.88
Gemini 1.0	0.68	0.79	0.73	0.65	0.50	0.57	0.67

Due to GPT-3.5’s comparatively poor classification performance, particularly its tendency to over-predict misleadingness, we excluded it from our analysis. We focus on GPT-4 (hereafter GPT) and Gemini 1.0 (hereafter Gemini) as the two representative LLMs, both of which demonstrated promising results in the classification task. A sample headline with corresponding human and LLM-generated labels and explanations is shown in Table 5.1 to illustrate the structure and content of the dataset.

5.2.1.1 Stratification by Agreement Level

To operationalize subjective ambiguity, we stratified the dataset based on levels of human annotator agreement:

- **3/3 misleading:** All three annotators labeled the headline as misleading.
- **2/3 misleading:** Two labeled it as misleading, one as not misleading.
- **1/3 misleading:** One labeled it as misleading, two as not misleading.
- **0/3 misleading:** All three labeled it as not misleading.

We treat 3/3 and 0/3 agreement as cases of *unanimous* consensus and 2/3 or 1/3 agreement as *ambiguous* cases, reflecting varying levels of interpretive subjectivity.

5.2.1.2 Data Sampling

To enable expert evaluation while maintaining diversity across ambiguity and LLM model alignment conditions, we sampled a subset of 21 headline-article pairs from the full dataset.

The selection was guided by a 2×2 stratification matrix: human agreement level (unanimous vs. ambiguous) and LLM agreement with the human majority (both models agree, only one agrees, or neither agrees).

To reduce the cognitive load on expert raters while preserving analytical coverage, we further refined the sample to 15 headlines. The final set was selected to ensure at least one representative sample from each key stratification category, while maintaining diversity in topic domains such as public health, political discourse, and corporate communication. Preference was given to headlines that raised substantive questions around subjective misleadingness, while thematically peripheral headlines (e.g., those lacking relevance to public trust or interpretive framing) were prioritized. This sampling strategy preserved representation across both human and model agreement levels, enabling situated analysis of LLM behavior under varying conditions of ambiguity and editorial salience.

5.2.2 Expert Evaluation Study Design

To evaluate the alignment of LLM-generated explanations with editorial standards, we conducted an expert-in-the-loop evaluation. Given the subjectivity of misleading headline detection, domain expertise is essential for assessing whether explanations meet professional expectations in terms of accuracy, framing, and practical usefulness. This expert evaluation complements crowd annotations and model outputs by providing a grounded perspective on explanation quality.

5.2.2.1 Expert Recruitment

We recruited six expert participants through personal networks, applying eligibility criteria that required participants to be 18 years or older, currently residing in the U.S., and have professional experience in journalism, particularly in English-language newsrooms. Although recruitment was informal, we ensured diversity in gender, age range, and level of professional experience. All participants were briefed on the study and introduced to the concept of misleading news headlines, but no formal training was provided in order to preserve the authenticity and independence of their editorial judgments. Experts were compensated at an hourly rate in accordance with institutional guidelines for external professional evaluators.

Table 5.3 summarizes the demographic and professional backgrounds of the six experts, including age range, gender, education level, familiarity with editorial tasks, and years of journalism experience. The group reflects a well-rounded mix of professional journalists, graduate students with professional experience, and researchers, covering a range of experience levels from early-career to seasoned professionals with over a decade in the field. Notably, all experts reported moderate to high familiarity with core editorial practices such as headline evaluation, framing identification, and media literacy, demonstrating strong alignment with the study's evaluative goals.

5.2.2.2 Evaluation Dimensions

Each LLM explanation was independently evaluated across seven dimensions: Correctness, Clarity, Exhaustiveness, Ambiguity Awareness, Risk of Harm, Trustworthiness, and Actionability

Table 5.3: Demographic and Professional Backgrounds of Expert Participants

ID	Age	Gender	Education	Familiarity with Editorial Tasks	Journalism Experience
E1	35–44	Male	4-year degree	Headline writing (Extremely); Framing (Very); Media Literacy (Very)	>10 years
E2	18–24	Female	4-year degree	Headline writing (Very); Framing (Extremely); Media Literacy (Extremely)	4–6 years
E3	25–34	Male	Professional degree	Headline writing (Moderately); Framing (Moderately); Media Literacy (Very)	1–3 years
E4	25–34	Female	Professional degree	Headline writing (Extremely); Framing (Very); Media Literacy (Extremely)	7–10 years
E5	35–44	Male	4-year degree	Headline writing (Very); Framing (Very); Media Literacy (Very)	7–10 years
E6	35–44	Male	Doctorate	Headline writing (Extremely); Framing (Extremely); Media Literacy (Extremely)	7–10 years

(i.e., its usefulness for editorial or fact-checking decisions). These dimensions draw from established human–AI explanation evaluation frameworks [352] and align with the “Contextualized Quality” category in Kim et al.’s systematic review of human-centered XAI metrics [213]. This design ensures our evaluation captures not only surface-level explanation quality but also its contextual relevance and utility in editorial decision-making.

5.2.2.3 Task Procedure

Each expert participated in an evaluation study consisting of two phases. In the first phase, experts were presented with a series of news headlines, each paired with its corresponding full article. To reduce cognitive load while preserving full contextual access, we adopted a contextual highlighting strategy inspired by Schmitt et al [352]. Specifically, one of the authors highlighted or underlined key sentences within each article, those directly related to the claims made in the headline and referenced in the human annotator or LLM-generated explanations. Experts were encouraged to use the highlighted sections as guidance to locate the most relevant information, but were also given access to the complete article text should they wish to review it more extensively.

Experts rated each explanation across seven evaluation dimensions (mentioned in 5.2.2.2) using a 4-point Likert scale (1 = Not at all, 4 = Completely). We intentionally employed an even-numbered scale to eliminate a neutral midpoint and encourage experts to take a directional stance (positive or negative) on each evaluation dimension. This design choice follows established methodological recommendations that suggest removing midpoints can reduce satisficing behavior and promote more engaged, discriminative responses, particularly in tasks requiring professional judgment [141, 200, 317]. To complement these quantitative assessments, experts responded to four open-ended prompts capturing perceived weaknesses, missing elements, suggested improvements, and the explanation’s handling of ambiguity.

The entire evaluation study was administered online, with instructions emphasizing that experts should prioritize evaluating explanations based on how accurately and completely they addressed potential misleading aspects of the headline in relation to the article content. Participation was designed to be completed in multiple sittings, with an estimated total time commitment of approximately 75–90 minutes. All study materials and evaluation protocols are

detailed in the Appendix.

5.2.3 Comparative Annotation of Human and LLM Explanations

To address RQ4, we conducted a comparative analysis of explanations authored by humans and two large language models (LLMs): Gemini and GPT. To reduce stylistic disparities between human-written and machine-generated content, the original human explanations were paraphrased using ChatGPT-4, ensuring a more consistent format for fair comparison. We evaluated four well-grounded editorial concerns, each reflecting a key aspect of journalistic judgment: Framing Bias [118, 152], Omission of Context [82, 387], and Emotional Manipulation [260] and Actionability [233]. These dimensions reflect core aspects of editorial reasoning, such as identifying misleading narrative structure, surfacing omitted context, detecting manipulative tone, and judging practical editorial utility. Two raters who were not involved in the expert evaluation study independently rated each explanation. Both raters were male, aged 25–34, and graduate students in journalism recruited through personal networks and compensated at an hourly rate. Ratings were assigned on a 3-point scale: 0 = Not at all, 1 = Partially, and 2 = Clearly. A score of 0 indicates that the explanation either failed to address the concern or incorrectly identified it, while a score of 2 reflects that the concern was clearly identified and explained, or, if absent, was accurately ruled out with justification. The dataset included 15 headlines (same data sampling used in Section 5.2.1.2), each accompanied by one human explanation and two LLM-generated explanations (45 explanations total). One author introduced the raters to the task and rubric, and they jointly rated five explanations to calibrate their interpretations. Each rater then independently evaluated a set of 15 randomly selected explanations, and we calculated inter-rater agreement using Cohen’s κ : 0.58 for Framing Bias, 0.71 for Omission of Context, 0.69 for Emotional Language, and 0.63 for Actionability. Additional discussion was held to clarify rating discrepancies, which approached the threshold of acceptable agreement. The remaining 25 explanations were double-coded, and the final scores were computed as the average of the two raters’ ratings.

5.2.4 Data Analysis

To understand how large language models (LLMs) align with human reasoning in the context of misleading headline detection, we adopted a triangulated evaluation approach [93, 94] that integrates three perspectives: crowd annotations, LLM predictions, and expert assessments of explanation quality. This design enabled us to analyze model behavior across multiple axes of comparison: labeling accuracy, explanation quality, and editorial reasoning alignment. Our analysis proceeded in three parts: we first examined labeling alignment between humans and LLMs, then conducted thematic comparisons of human- and AI-generated explanations, and finally identified which explanation dimensions most strongly predicted expert trust and editorial usefulness. This multi-perspective strategy provided a situated understanding of LLM performance in ambiguous and interpretively rich decision-making contexts.

5.2.4.1 Quantitative Analysis

To investigate the performance and explanatory quality of LLM-generated rationales across multiple conditions and research questions, we employed a set of quantitative methods grounded in both classification metrics and expert evaluation scores. This subsection outlines the analytical strategies applied for each research question.

Expert Ratings and Inter-Rater Agreement Experts evaluated each explanation across seven dimensions: Correctness, Clarity, Exhaustiveness, Ambiguity Awareness, Risk of Harm, Trustworthiness, and Actionability. To assess the reliability of these evaluations, we calculated Intraclass Correlation Coefficients (ICC) for each dimension, providing a robust measure of consistency across expert ratings in small, subjective evaluation panels.

RQ1: Agreement-Level Effects on LLM Performance and Explanation Quality To examine how annotator agreement influences model behavior, we computed *weighted accuracy* for each LLM prediction by comparing it to the proportion of human annotators, labeling the headline as misleading.

For each headline, human annotators provided binary labels (misleading or not misleading), and the total number of annotators who labeled the headline as misleading was tallied (ranging from 0 to 3 out of 3). Weighted accuracy was then computed as follows:

- If the LLM predicted *misleading*, the weight was calculated as the fraction of annotators who also labeled the headline as misleading.
- If the LLM predicted *not misleading*, the weight was calculated as the fraction of annotators who labeled the headline as not misleading (i.e., one minus the fraction of annotators labeling it as misleading).

This approach captures partial alignment in ambiguous cases and full alignment in unanimous ones. We then compared weighted accuracy scores between unanimous and ambiguous conditions using the Mann–Whitney U test [262] to assess statistical significance.

We also analyzed expert evaluations of explanation quality by grouping ratings according to agreement level (unanimous vs. ambiguous) for both GPT and Gemini. Mean scores were computed across all seven dimensions, and Mann–Whitney U tests were applied to determine whether expert assessments significantly differed between agreement conditions.

RQ2: Divergence of Model Explanations in Ambiguous Cases To evaluate how perceived explanation quality varies between models in ambiguous cases, we computed the mean and standard deviation of expert ratings for each model across all dimensions. We then conducted Wilcoxon signed-rank tests to assess whether paired expert ratings for GPT and Gemini explanations differed significantly. Rank-biserial correlation (r) [85] was calculated to estimate the effect size of these differences and identify the magnitude of model-level preference.

RQ3: Predictors of Expert Trust and Actionability To explore which explanation characteristics most influence expert judgments of trust and editorial usefulness, we examined the relationship between five independent dimensions (Correctness, Clarity, Exhaustiveness, Ambiguity

Awareness, and Risk of Harm) and two dependent outcomes: Trustworthiness and Actionability. We used Spearman’s rank correlation [356] to assess bivariate associations, followed by multiple linear regression to determine the most predictive dimensions while controlling for others. To validate and supplement these results, we also used a random forest regressor to rank feature importance for both dependent variables.

RQ4: Comparative Evaluation of Human and LLM Explanations To assess how human- and LLM-generated explanations differ in surfacing key editorial concerns, we conducted Friedman tests across four dimensions: Framing Bias, Omission of Context, Emotional Language, and Actionability. These dimensions were derived from journalism literature and reflect critical aspects of editorial reasoning. Post-hoc Wilcoxon signed-rank tests [425] were used for pairwise comparisons between the three explanation types (Human, Gemini, GPT), and effect sizes (r) were reported to gauge practical significance.

5.2.4.2 Qualitative Analysis

To address RQ2, we conducted a qualitative thematic analysis of expert feedback on four evaluation dimensions: weaknesses, missing information, suggestions for improvement, and handling of ambiguity. We focused specifically on ambiguous headlines, defined as those with 1/3 or 2/3 misleading votes from human annotators. These cases offer a valuable opportunity to examine how large language models (LLMs) respond to interpretive uncertainty and whether their explanations align with expert editorial reasoning in such contexts.

The thematic analysis was conducted by one author following Braun and Clarke’s widely used six-phase approach to reflexive thematic analysis [53, 54]. Using a single coder is common and well-accepted in qualitative research, particularly within interpretive traditions, where meaning is constructed through deep engagement with the data rather than coder agreement [302].

The analyst began by immersing in the dataset, reading and re-reading the expert comments to develop familiarity and sensitivity to emerging patterns. Through iterative coding, salient segments of text were tagged with inductive codes, capturing expert concerns related to explanation quality and editorial reasoning. These codes were then grouped into higher-order themes that characterized common shortcomings or strengths across multiple explanations. Themes were refined and reviewed in light of the full dataset to ensure coherence and distinctiveness. Each theme was then defined and named to clearly convey its relevance to the research question. This process enabled us to surface six overarching themes that captured recurring limitations in LLM-generated rationales, particularly in ambiguous cases, highlighting how these explanations, even when factually correct, often fail to meet the standards of expert editorial judgment.

5.3 Result

This section presents findings from our multi-phase evaluation of LLM-generated explanations for misleading news headlines. We report results across four research questions, each addressing a distinct aspect of model performance, explanation quality, and human-AI alignment. RQ1 examines how varying levels of human annotator agreement relate to both

LLM classification accuracy and expert evaluations of explanation quality. RQ2 focuses on ambiguous cases to understand how LLM-generated rationales align or diverge from expert editorial reasoning. RQ3 identifies which explanation characteristics most strongly influence expert trust and actionability judgments. Finally, RQ4 compares human- and LLM-generated explanations regarding their ability to surface core editorial concerns such as framing bias, omission of context, and emotional language. Each subsection details both quantitative and qualitative analyses to provide a comprehensive view of LLM behavior in subjective, real-world editorial settings.

5.3.1 Overview of Expert Ratings and Inter-Rater Reliability

5.3.1.1 Descriptive Statistics

We computed descriptive statistics for expert evaluations of LLM-generated explanations across seven key dimensions: *Correctness*, *Clarity*, *Exhaustiveness*, *Ambiguity Awareness*, *Risk of Harm*, *Trustworthiness*, and *Actionability*. Each dimension was rated on a 4-point Likert scale, and the analysis included 30 total ratings.

The mean scores ranged from 2.17 for *Risk of Harm*, indicating relatively low perceived potential for harm, to 3.46 for *Clarity*, suggesting that explanations were generally easy to understand. *Correctness* and *Actionability* also received moderately high average scores (3.13 and 2.83, respectively), whereas *Exhaustiveness* and *Ambiguity Awareness* were somewhat lower, pointing to room for improvement in how comprehensively and cautiously the explanations addressed headline ambiguity. The standard deviations across dimensions were moderate (ranging between 0.39 and 0.69), reflecting some variability in expert judgments. The details of the results can be seen in Table [A.9](#) in the Appendix.

5.3.1.2 Inter-Rater Agreement Analysis

To assess the consistency of expert judgments across multiple evaluation dimensions, we calculated the Intraclass Correlation Coefficient (ICC), using a two-way random-effects model with absolute agreement (ICC(2,1)). This model is particularly well-suited for measuring agreement among multiple raters when evaluating ordinal-scale data, such as Likert responses. The analysis yielded ICC values ranging from 0.34 to 0.53 across the seven dimensions, indicating a generally moderate level of agreement.

Among the dimensions, *Ambiguity Awareness* achieved the highest ICC score of 0.53. This suggests that experts were relatively consistent in identifying whether the LLM-generated explanations acknowledged potential ambiguity or multiple interpretations within the headlines. An ICC above 0.50 is commonly interpreted as moderate to substantial agreement, indicating that this dimension may be conceptually more concrete or recognizable among trained evaluators.

Correctness (0.46), *Exhaustiveness* (0.42), *Risk of Harm* (0.42), and *Trustworthiness* (0.40) also demonstrated moderate levels of inter-rater reliability. These scores suggest that while some subjectivity was present in the interpretation of these qualities, experts shared a reasonably common framework for evaluating whether explanations were factually accurate, sufficiently comprehensive, ethically responsible, and trustworthy. These results are meaningful given that

such assessments often involve both subjective interpretation and application of journalistic principles.

On the lower end, *Clarity* (0.38) and *Actionability* (0.34) exhibited slightly lower ICC values, bordering on fair rather than moderate agreement. This outcome may reflect the more interpretive nature of these dimensions: clarity can be influenced by individual preferences for language and structure, while actionability may be understood differently depending on how a rater conceptualizes the reader’s role or the goal of the explanation.

These findings indicate that while expert evaluations of LLM-generated explanations exhibit some inherent variability, particularly regarding more subjective aspects, the levels of agreement observed are generally within acceptable ranges for studies evaluating human judgment. The results justify the use of aggregated expert ratings for downstream analysis and support the validity of using expert consensus to assess the performance of AI-generated journalistic reasoning. However, the moderate agreement also underscores the need to complement quantitative scores with qualitative feedback to fully capture the nuances in expert reasoning and interpretation.

5.3.2 Impact of Annotator Agreement on LLM Accuracy and Explanation Quality (RQ1)

This analytical approach explores how the level of human annotator agreement affects LLM classification accuracy and the expert-evaluated quality of their explanations. We compare model performance and explanation ratings across unanimous and ambiguous headline cases.

5.3.2.1 Weighted Accuracy Analysis by Agreement Category

Gemini achieved a mean weighted accuracy (mentioned in 5.2.4.1) of 0.68($SD = 0.47, n = 41$) on unanimous cases and 0.51($SD = 0.17, n = 19$) on ambiguous cases. A Mann-Whitney U test revealed that this difference was statistically significant ($U = 532.0, p = 0.016$), indicating that Gemini performed better when human annotators were in full agreement. GPT followed a similar pattern, achieving 0.90($SD = 0.30, n = 41$) for unanimous cases and 0.53($SD = 0.17, n = 19$) for ambiguous cases, with the difference being highly significant ($U = 703.0, p < 0.001$).

These results show that both models aligned more closely with human annotations when human agreement was strong, while their alignment weakened for ambiguous headlines. Notably, GPT consistently outperformed Gemini across both categories, particularly on unanimous cases where it approached near-perfect accuracy. Figure 5.1 visualizes these findings, illustrating the weighted accuracy drop under ambiguity for both LLMs.

5.3.2.2 Explanation Analysis by Agreement Category across Quality Dimensions

For the *correctness* dimension, GPT achieved consistent scores across agreement categories ($M = 3.33$ for both ambiguous and unanimous cases)(As shown in Table A.7), suggesting that GPT-generated explanations may maintain more stable perceived accuracy regardless of

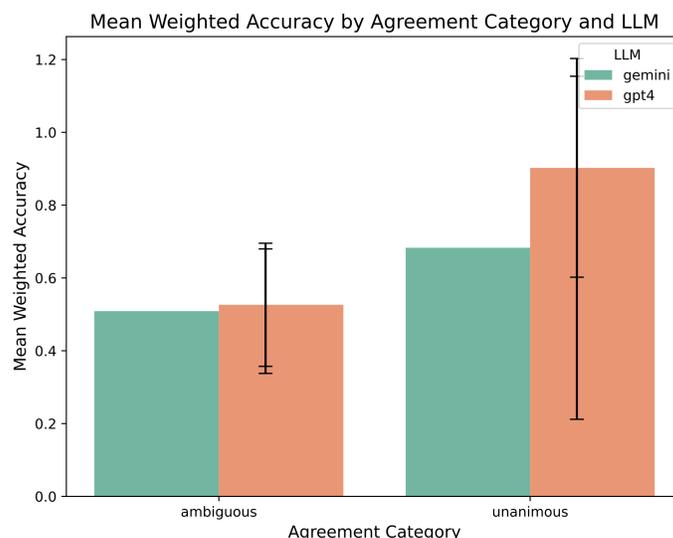


Figure 5.1: Mean weighted accuracy of Gemini and GPT-4 across two agreement categories: ambiguous and unanimous. Error bars represent ± 1 standard deviation. Both LLMs demonstrate higher alignment with human annotations in the unanimous (high agreement) condition, while performance drops notably under ambiguity. GPT-4 consistently outperforms Gemini across both categories, particularly for unanimous headlines

underlying disagreement among annotators. In contrast, Gemini exhibited a modest increase in perceived correctness for ambiguous headlines ($M = 3.14$) compared to unanimous ones ($M = 2.73$), though this difference was not statistically significant ($U = 21.00, p = 0.4491$) (Reported in Table A.8).

Perceptions of *ambiguity awareness*, an essential dimension for interpreting potentially misleading headlines, were also higher in ambiguous cases. For Gemini, the mean score increased from 2.48 (unanimous) to 3.00 (ambiguous), while GPT improved from 2.62 to 2.95. These results suggest that LLMs may adjust their explanations in ways that acknowledge uncertainty or complexity when human disagreement signals contextual ambiguity. Still, as with correctness, these improvements were not statistically significant ($p > 0.26$ for both models).

Interestingly, *risk of harm* was rated relatively low across both agreement conditions and both models, with GPT receiving slightly lower scores overall. This may reflect the generally cautious nature of the explanations or that none of the examples were perceived to introduce major editorial risks.

Overall, the findings suggest a weak but consistent trend where ambiguous cases receive slightly more favorable ratings across several explanation dimensions, possibly due to LLMs adopting more hedged or nuanced reasoning under uncertainty.

Table 5.4: Mean and SD of Expert Ratings across Seven Explanation Quality Dimensions, by LLM and Agreement Category

Dimension	LLM	Agreement	Mean	SD
Correctness	Gemini	Unanimous	2.73	0.88
Correctness	Gemini	Ambiguous	3.14	0.56
Correctness	GPT	Unanimous	3.33	0.31
Correctness	GPT	Ambiguous	3.33	0.53
Clarity	Gemini	Unanimous	3.29	0.44
Clarity	Gemini	Ambiguous	3.45	0.36
Clarity	GPT	Unanimous	3.69	0.14
Clarity	GPT	Ambiguous	3.40	0.52
Exhaustiveness	Gemini	Unanimous	2.48	0.83
Exhaustiveness	Gemini	Ambiguous	2.98	0.33
Exhaustiveness	GPT	Unanimous	2.92	0.53
Exhaustiveness	GPT	Ambiguous	2.95	0.47
Ambiguity awareness	Gemini	Unanimous	2.48	0.83
Ambiguity awareness	Gemini	Ambiguous	3.00	0.37
Ambiguity awareness	GPT	Unanimous	2.62	0.85
Ambiguity awareness	GPT	Ambiguous	2.95	0.52
Risk of harm	Gemini	Unanimous	2.54	0.79
Risk of harm	Gemini	Ambiguous	2.17	0.52
Risk of harm	GPT	Unanimous	1.90	0.44
Risk of harm	GPT	Ambiguous	2.07	0.58
Trustworthiness	Gemini	Unanimous	2.38	0.80
Trustworthiness	Gemini	Ambiguous	2.76	0.50
Trustworthiness	GPT	Unanimous	2.85	0.42
Trustworthiness	GPT	Ambiguous	2.93	0.57
Actionability	Gemini	Unanimous	2.52	0.63
Actionability	Gemini	Ambiguous	2.76	0.48
Actionability	GPT	Unanimous	3.06	0.47
Actionability	GPT	Ambiguous	2.98	0.47

5.3.3 Divergence of LLM Explanations from Expert Reasoning in Ambiguous Cases (RQ2)

To address RQ2, we analyzed expert evaluations of LLM-generated explanations for ambiguous headlines, defined as those receiving partial agreement (1/3 or 2/3 misleading votes) from human annotators. These cases present interpretive uncertainty, offering a valuable lens into how language models reason in contexts where even humans diverge. Our goal was to examine whether and how GPT and Gemini explanations differ in quality and alignment with expert journalistic judgment under such conditions. We focused both on quantitative ratings across seven explanation dimensions and qualitative comments capturing perceived weaknesses, omissions, and ambiguity handling.

Table 5.5: Mann-Whitney U Test Results Comparing Unanimous vs. Ambiguous Cases by Dimension and LLM

Dimension	LLM	U	p-value
correctness	Gemini	21.00	0.4491
correctness	GPT	24.00	0.6800
clarity	Gemini	22.00	0.5203
clarity	GPT	35.00	0.4328
exhaustiveness	Gemini	20.00	0.3807
exhaustiveness	GPT	24.50	0.7244
ambiguity_awareness	Gemini	18.00	0.2655
ambiguity_awareness	GPT	22.00	0.5199
risk_of_harm	Gemini	33.50	0.5607
risk_of_harm	GPT	23.50	0.6416
trustworthiness	Gemini	21.00	0.4495
trustworthiness	GPT	25.00	0.7701
actionability	Gemini	23.00	0.5992
actionability	GPT	32.50	0.6395

5.3.3.1 Divergence in Expert Ratings for Ambiguous Headlines

Table 5.6 presents the mean and standard deviation of expert ratings for each LLM across all dimensions. While GPT tended to score slightly higher than Gemini on most dimensions, the differences were modest (e.g., clarity: $GPT = 3.40$ vs. $Gemini = 3.45$; correctness: $GPT = 3.33$ vs. $Gemini = 3.14$).

To further investigate whether expert ratings favored one LLM over another in cases of ambiguous human agreement, we conducted within-headline comparisons of explanation quality across seven dimensions (e.g., correctness, clarity, exhaustiveness). Using the Wilcoxon signed-rank test, we tested whether the paired ratings for GPT and Gemini explanations differed significantly across ambiguous headlines. As shown in Table 5.7, none of the dimensions exhibited statistically significant differences, with all p-values well above the conventional 0.05 threshold.

To assess practical divergence, we also report the rank-biserial effect size (r), which quantifies the magnitude of model-level preference. Most values fell in the range of small effects ($r \approx \pm 0.1$ to ± 0.3), with exhaustiveness showing the largest effect ($r = -0.50$), favoring Gemini in a small number of cases. Still, the directionality was inconsistent across dimensions, and the paired dot plots (Figure 5.2) show high visual overlap, reinforcing the overall finding that no systematic preference emerged for either model in the context of interpretive ambiguity.

We also calculated difference scores for each explanation quality dimension by subtracting Gemini ratings from GPT ratings on a per-headline basis (i.e., $Difference = GPT - Gemini$). This analysis helps visualize the distribution and variability of expert preferences within each dimension rather than relying solely on summary statistics or test-level significance. Figure 5.3 presents a boxplot of these difference scores across the seven dimensions. Most median differences hovered around zero, with no consistent directionality favoring one model. This aligns with the Wilcoxon test results, which also failed to detect statistically significant

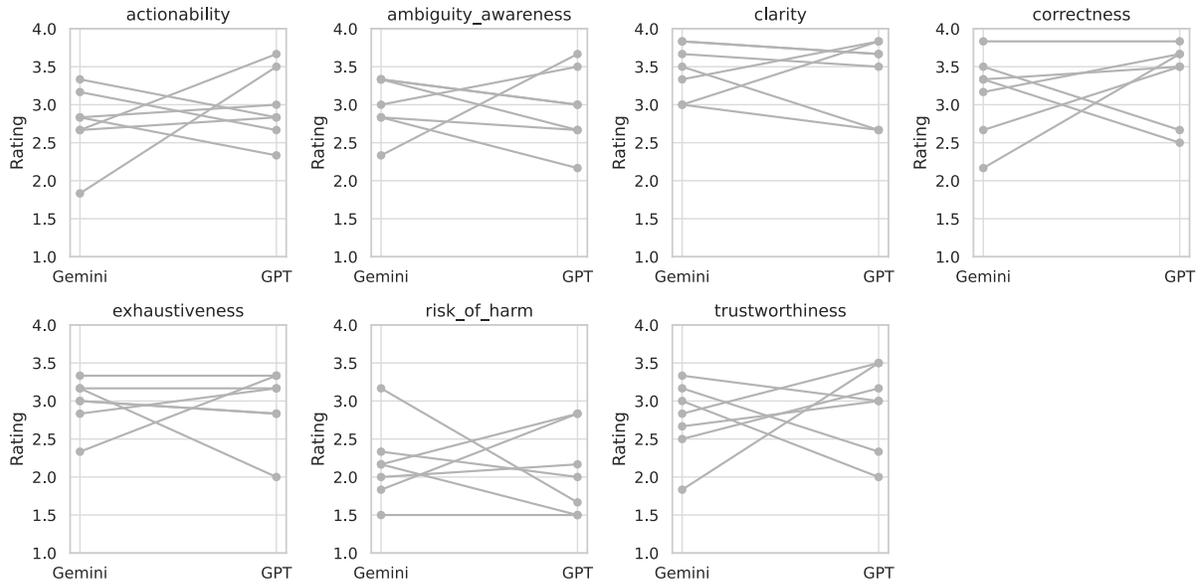


Figure 5.2: Paired expert ratings for GPT and Gemini explanations across seven quality dimensions for ambiguous headlines (i.e., headlines with partial human annotator agreement). Each line connects the scores given to both LLMs for the same headline. The visual symmetry and overlapping trends illustrate that experts rated the explanations comparably, with no consistent pattern favoring one model. These findings align with the statistical results indicating no significant differences in explanation quality between GPT and Gemini under ambiguous conditions.

differences between models. Notably, dimensions such as correctness and trustworthiness showed slightly wider interquartile ranges, suggesting more inter-headline variability in expert preferences. However, there was no dimension with a systematic positive or negative shift, reinforcing the interpretation that both models produce comparably rated explanations in ambiguous headline contexts.

These findings support the hypothesis that while ambiguity introduces interpretive variation, both LLMs tend to produce comparably plausible or flawed rationales, leading experts to rate them similarly. This highlights a potential limitation of current LLMs in surfacing and addressing ambiguity with discernible quality differences in expert-facing scenarios.

5.3.3.2 Expert-Centered Diagnoses of Explanation Shortcomings in Ambiguous Cases

Thematic analysis of expert feedback on four evaluation dimensions: weaknesses, missing information, suggestions for improvement, and handling of ambiguity, focused on ambiguous cases (i.e., headlines with 1/3 or 2/3 misleading votes), revealed key limitations in LLM-generated explanations. Even when technically accurate, these rationales often fell short of supporting robust editorial judgment. Six overarching themes emerged, highlighting recurring gaps across dimensions.

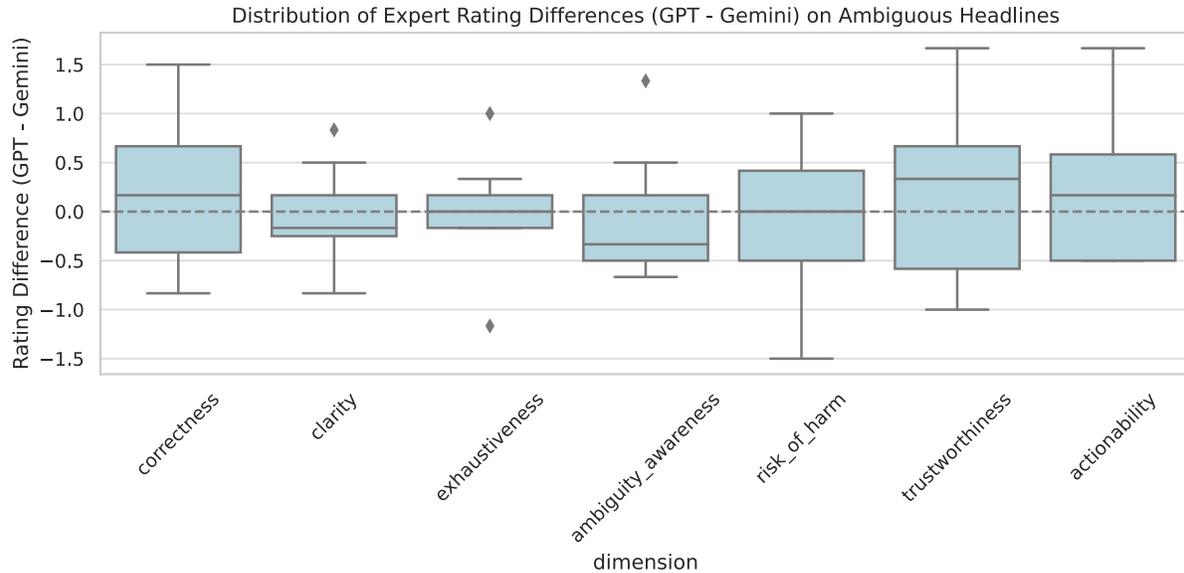


Figure 5.3: Distribution of difference scores (GPT - Gemini) in expert ratings across seven explanation quality dimensions for ambiguous headlines. Values above zero indicate higher GPT scores; values below zero indicate higher Gemini scores. Most differences cluster around zero, suggesting no systematic model advantage.

Lack of Contextual and Scientific Precision: Across the explanations reviewed, journalists frequently flagged the absence of important contextual qualifiers, particularly in health and science reporting. When reviewing a Gemini explanation for a headline about vaccine effectiveness, one expert (E1) emphasized that the summary failed to note the study’s focus on older adults and omitted whether it was peer-reviewed or merely observational. Another expert (E6), who held a doctoral degree and had extensive professional experience, evaluated a GPT explanation for a headline with partial agreement (2/3 misleading votes) and emphasized the need to clarify the distinction between statistical association and causal inference, an editorially critical nuance for ensuring transparency.

Similarly, expert E4, critiquing a Gemini explanation about a neurological study on menopause, noted that it glossed over scientific nuance, writing, “*It did not mention that the study was not published officially,*” and therefore left the interpretation vulnerable to overgeneralization. This pattern of omitting methodological details was common across explanations and was seen to compromise credibility, particularly in complex scientific contexts.

Emotional Framing Left Unaddressed: Another major concern was the failure of LLM explanations to interrogate emotionally charged or rhetorically exaggerated headlines. This was particularly evident in cases where headlines employed sensationalist terms like “*illegal*” or declarative phrasing like “*I’m Done.*” E1, an expert with substantial professional experience in journalism, critiqued a Gemini explanation for a headline with partial agreement, noting that it failed to address the emotional framing of the term “*illegal*”, which could imply scandal or criminality beyond what the article substantiated.

Table 5.6: Mean and Standard Deviation of Expert Ratings for Ambiguous Headlines (by LLM)

Dimension	LLM	Mean	SD
Correctness	Gemini	3.14	0.56
	GPT	3.33	0.53
Clarity	Gemini	3.45	0.36
	GPT	3.40	0.52
Exhaustiveness	Gemini	2.98	0.33
	GPT	2.95	0.47
Ambiguity Awareness	Gemini	3.00	0.37
	GPT	2.95	0.52
Risk of Harm	Gemini	2.17	0.52
	GPT	2.07	0.58
Trustworthiness	Gemini	2.76	0.50
	GPT	2.93	0.57
Actionability	Gemini	2.76	0.48
	GPT	2.98	0.47

In another instance, a GPT explanation addressing a headline about booster skepticism was seen as underplaying the declarative, opinion-driven framing. E6 noted that the explanation “*could go further by pointing out the emotional and declarative framing (‘I’m Done’) and how this might be perceived as dismissive or hostile to vaccines.*” These critiques reflect an expectation that explanations should not only assess factual alignment but also explore how tone and framing may mislead readers.

Failure to Distinguish Policy from Personal Opinion: When headlines reported on politically contentious topics, such as COVID-19 policy recommendations, experts expected the explanations to differentiate between an individual’s opinion and state-level policy. This was particularly relevant in evaluating GPT-generated explanations of a headline where Florida’s top health official recommended against vaccinating healthy children. E6, who claimed to be highly familiar with media literacy, pointed out that “*it missed the nuance between the chief health official’s stance and the state’s policy,*” leading to a misleading interpretation of the headline’s scope.

This critique was not isolated. Another expert (E2) reviewing a Gemini explanation of the same headline (rated under a partial agreement) similarly noted that the explanation “*only mentioned the health official, not the governor,*” failing to clarify that the recommendation was part of a broader state policy. These findings underscore the importance of precise attribution in politically sensitive reporting.

Surface-Level Analysis and Overemphasis on Factual Matching: Experts also critiqued explanations for relying too heavily on factual matching without engaging with the subtleties of journalistic judgment. Several noted that explanations treated the task as a binary classification problem, true or false, rather than grappling with interpretive ambiguity. For instance, E3, a

Dimension	Wilcoxon Statistic	p-value	Rank-biserial r	n (paired)
Correctness	7.5	0.5282	-0.464	7
Clarity	11.5	0.6875	-0.179	7
Exhaustiveness	7.0	0.8923	-0.500	7
Ambiguity Awareness	11.0	0.6875	-0.214	7
Risk of Harm	9.5	0.8335	-0.321	7
Trustworthiness	12.0	0.8125	-0.143	7
Actionability	12.0	0.8125	-0.143	7

Table 5.7: Wilcoxon signed-rank test results comparing expert evaluations of GPT and Gemini explanations across seven dimensions for ambiguous headlines, with corresponding rank-biserial effect sizes.

reviewer with moderate professional experience in journalism, critiqued a Gemini explanation for a headline about video game hacking, noting that “*It fails to differentiate between the video game and the underlying blockchain bridge.*” This indicated that the explanation overlooked how general audiences might interpret the phrase “*hackers hit popular video game*” in an overly literal way.

In a similar vein, E1, examining a GPT explanation, wrote that it “*focuses on factual accuracy without examining how headlines shape reader interpretation, especially on sensitive topics like vaccines.*” These comments point to a common failure to explore interpretive consequences, even when factual discrepancies were minimal.

Ambiguity Poorly Addressed or Ignored: Ambiguity handling was a core area of weakness in many explanations, even when experts acknowledged partial success. When reviewing a Gemini explanation of a legally nuanced headline about COVID immunity policy, E5 stated bluntly that “*the explanation does not acknowledge any ambiguity,*” instead assuming a “*straightforward relationship between the headline and article.*” Another expert (E1), reviewing a GPT explanation of the same headline, noted that while it correctly identified jurisdictional issues, it “*did not fully address ambiguity in public interpretation or the headline’s potential to mislead.*”

Even when ambiguity was acknowledged, it was often only gestured toward. One expert (E6), for example, wrote that a GPT explanation “*gestures toward complexity... but stops short of analyzing how the absence of context... could mislead or polarize readers.*” These limitations suggest that explanations need not only to identify ambiguity but also to contextualize it in terms of audience understanding and editorial responsibility.

Suggestions Emphasizing Specificity, Editorial Judgment, and Attribution: When asked how explanations could be improved, experts offered clear and actionable recommendations. These included suggestions for revising headline language (e.g., “*Hackers Steal Over \$600M in Crypto from Blockchain Linked to Popular Game Axie Infinity*”), calls to present contexts more precisely, and requests to clarify whether data was preliminary or causal. There were also proposals to better reflect opinion vs. reporting tone, as in the case of vaccine skepticism headlines, where experts advised noting when an article was an opinion piece rather than hard news.

Additionally, experts wanted explanations to serve editorial needs. One expert (E6) noted that editorial decisions would benefit from the explanation “*suggesting a revised headline that reflects the article’s actual position,*” particularly in polarized topics such as vaccine policy or public health.

The qualitative analysis of feedback highlights key limitations in LLM-generated explanations for ambiguous headlines. Journalists expect not only factual accuracy but also interpretive rigor, emotional framing awareness, scientific precision, and editorial relevance. The ability to handle ambiguity effectively is not merely about noting uncertainty; it involves recognizing how headlines can frame narratives in ways that require critical interrogation. Addressing these shortcomings will be essential for designing LLM systems that meaningfully support expert editorial judgment in complex media contexts.

5.3.4 Predicting Expert Trust and Usefulness from Explanation Characteristics (RQ3)

Through different statistical analyses, RQ3 seeks the most strong predictors of expert trust and perceived editorial usefulness.

5.3.4.1 Identifying Editorial Predictors of Trustworthiness and Actionability

We examined which editorial dimensions best predict expert ratings of explanation *trustworthiness* and *actionability* by analyzing the relationship between independent dimensions (e.g., correctness, clarity, ambiguity awareness) and two dependent outcomes: **trustworthiness** and **actionability**.

Correlation Analysis. As shown in Table 5.8, for trustworthiness, the strongest positive association was observed with *correctness* ($\rho = 0.92$, $p < .00001$), indicating that explanations rated as factually accurate were more likely to be trusted by experts. Other dimensions showing substantial positive associations included *exhaustiveness* ($\rho = 0.80$) and *ambiguity awareness* ($\rho = 0.71$), suggesting that experts also valued completeness and sensitivity to interpretive uncertainty when forming trust judgments. Notably, *risk of harm* demonstrated a strong negative correlation with trust ($\rho = -0.85$, $p < .00001$), indicating that explanations perceived as potentially misleading or harmful significantly undermined expert confidence.

For actionability, which captures how useful the explanation would be to a journalist deciding whether to revise or flag a headline, the pattern was slightly different. The strongest predictor was again *risk of harm*, showing a large negative correlation ($\rho = -0.89$), followed by positive associations with *correctness* ($\rho = 0.83$), *exhaustiveness* ($\rho = 0.79$), and *ambiguity awareness* ($\rho = 0.77$).

These results suggest that explanations which are factually accurate, contextually complete, and cognizant of uncertainty, while avoiding overstated or misleading claims, are more likely to be both trusted by experts and seen as editorially useful.

Regression Analysis. To assess the relative influence of explanation quality dimensions on expert judgments, we conducted two multiple linear regression models: one predicting

trustworthiness and the other *actionability*.

The model predicting *trustworthiness* accounted for a substantial proportion of variance ($R^2 = .91$), indicating that the included predictors explained 91% of the variability in expert trust ratings. Among the predictors, *correctness* emerged as the only statistically significant factor ($\beta = 0.83, p < .001$), strongly suggesting that experts' trust in LLM-generated explanations was primarily driven by their factual accuracy. *Ambiguity awareness* showed a positive but marginally significant effect ($\beta = 0.21, p = .074$), indicating that, when *correctness* is accounted for, attentiveness to uncertainty might still play a supporting role. All other variables, including *risk of harm*, *clarity*, and *exhaustiveness*, were not statistically significant when modeled alongside *correctness*.

In contrast, the model predicting *actionability* also explained a high proportion of variance ($R^2 = .89$) but identified a different set of predictors. Here, *ambiguity awareness* was a significant positive predictor ($\beta = 0.30, p = .012$), and *risk of harm* was a significant negative predictor ($\beta = -0.56, p = .019$). Interestingly, *correctness*, while correlated with *actionability* in earlier analyses, did not significantly predict it in the multivariate model ($p = .293$), suggesting that factual accuracy alone is insufficient to make an explanation practically useful for editorial decision-making.

Feature Importance Analysis. To complement the linear regression models and capture potential non-linear relationships among variables, we trained two Random Forest regressors to predict expert ratings of *trustworthiness* and *actionability*. Feature importance scores were extracted to assess the relative contribution of each explanation dimension in predicting the outcome variables.

For *trustworthiness*, the most influential predictor was *risk of harm*, with a feature importance score of 0.44, followed closely by *correctness* (0.39). Other dimensions such as *exhaustiveness* (0.07), *clarity* (0.05), and *ambiguity awareness* (0.04), contributed relatively little to the model. This finding aligns with prior analyses in emphasizing *correctness* as a key driver of trust while also highlighting the considerable role of perceived potential harm in undermining expert confidence.

In predicting *actionability*, *risk of harm* again emerged as the most important feature (0.52), reinforcing its role as a central factor in determining whether an explanation is editorially useful. Interestingly, *ambiguity awareness* was the second most important feature (0.26), exceeding *correctness* (0.15), suggesting that in practical editorial settings, experts prioritize an explanation's ability to handle uncertainty and highlight potential misinterpretation risks. Features such as *clarity* and *exhaustiveness* remained low in importance (0.02 and 0.06, respectively).

5.3.4.2 Exploring Trustworthiness and Actionability Across Agreement Levels

To understand whether the editorial criteria influencing expert judgments shift depending on the level of annotator agreement about a headline's misleadingness, we repeated our RQ4 analysis stratified by agreement level (unanimous vs. ambiguous). We examined how each explanation dimension predicted expert ratings of *trustworthiness* and *actionability* using correlation, regression, and feature importance analyses.

Table 5.8: Predictive power of editorial dimensions on trustworthiness and actionability across three analysis methods. Boldface indicates top contributors within each method.

Dimension	Spearman ρ		Regression β		RF Importance
	Trust	Action	Trust	Action	Avg. Rank
Correctness	0.92	0.83	0.83***	0.19	2 (Trust), 3 (Action)
Clarity	0.53	0.60	0.01	0.01	Low
Exhaustiveness	0.80	0.79	-0.14	-0.23	Low
Ambiguity Awareness	0.71	0.77	0.21 [†]	0.30*	3 (Trust), 2 (Action)
Risk of Harm	-0.85	-0.89	-0.04	-0.56*	1 (Trust & Action)

* $p < .05$, ** $p < .01$, *** $p < .001$, [†] $p < .10$

Columns show Spearman correlation coefficients, standardized regression betas, and Random Forest feature importance ranks for predicting trustworthiness and actionability.

Unanimous Headlines. For headlines with high annotator agreement, expert trust in LLM explanations was most strongly correlated with *correctness* ($\rho = .93$, $p < .001$), *exhaustiveness* ($\rho = .83$), and was negatively associated with *risk of harm* ($\rho = -.87$). Actionability followed a similar pattern, showing strong correlations with *exhaustiveness* ($\rho = .93$), *correctness* ($\rho = .89$), and *risk of harm* ($\rho = -.92$). Multiple linear regression confirmed *correctness* as a significant predictor of trust ($\beta = 0.79$, $p = .019$), while Random Forest models identified *risk of harm* and *correctness* as top features for both outcomes.

Ambiguous Headlines. When headlines were interpretively ambiguous, the expert criteria shifted. While *correctness* remained strongly correlated with trust ($\rho = .88$), both *ambiguity awareness* ($\rho = .73$) and *risk of harm* ($\rho = -.83$) also emerged as key predictors. For actionability, *ambiguity awareness* ($\rho = .73$) and *risk of harm* ($\rho = -.86$) were most influential. Regression results showed that all three dimensions: *correctness*, *ambiguity awareness*, and *risk of harm*, significantly predicted trust (all $p < .05$), and *ambiguity awareness* was marginally significant for actionability ($p = .055$). Random Forest results confirmed *ambiguity awareness* as the most important predictor of actionability in ambiguous cases.

Table 5.9 summarizes these differences. While *correctness* consistently drives expert trust regardless of headline ambiguity, *ambiguity awareness* becomes more important to editorial usefulness in ambiguous settings. Additionally, *risk of harm* plays a crucial inverse role in both trust and actionability across conditions.

Compared to the overall RQ3 analysis, the agreement-level breakdown reveals important nuance. In general, the predictive role of *correctness* on trust remains robust across both unanimous and ambiguous headlines. However, when focusing specifically on ambiguous cases, *ambiguity awareness* becomes notably more influential for actionability than in the overall sample. This suggests that when experts face interpretively uncertain headlines, they assign greater value to explanations that explicitly acknowledge uncertainty and clarify potential sources of misunderstanding. In contrast, under unanimous agreement, editorial usefulness is more heavily driven by completeness and clarity of detail. These distinctions highlight the importance

of tailoring LLM explanations to the level of interpretive complexity embedded in a headline.
 graphicx

Table 5.9: Spearman correlations (ρ) between editorial dimensions and expert outcomes, stratified by agreement level.

Dimension	Trust (Unanimous)	Trust (Ambiguous)	Actionability (Unanimous)	Actionability (Ambiguous)
Correctness	$\rho = .93$	$\rho = .88$	$\rho = .89$	$\rho = .82$
Ambiguity Awareness	$\rho = .71$	$\rho = .73$	$\rho = .80$	$\rho = .73$
Risk of Harm	$\rho = -.87$	$\rho = -.83$	$\rho = -.92$	$\rho = -.86$

5.3.5 Comparing Human and LLM Explanations Across Editorial Concerns (RQ4)

We conducted a comparative analysis of explanations authored by humans and two large language models (LLMs): Gemini and GPT across four editorially relevant dimensions: Framing Bias, Omission of Context, Emotional Language, and Actionability. To address RQ4, we performed both descriptive and inferential statistical analyses comparing the average scores across explanation types.

5.3.5.1 Descriptive Statistics

The descriptive results (shown in Table 5.10 and Figure 5.4) reveal a consistent pattern: human-written explanations outperform both LLMs across all four editorial dimensions. The most substantial differences appear in the categories of *Framing Bias* and *Omission of Context*, where human explanations scored more than twice as high as those generated by Gemini. This indicates that human raters were significantly more effective in identifying ideological slants, misleading emphasis, and missing background information, key editorial judgments often requiring context-sensitive reasoning.

While GPT explanations scored modestly higher than Gemini’s on average, especially in identifying *Omission of Context* ($M = 1.20$ vs. $M = 0.90$), both models struggled with *Actionability*, achieving mean scores below 0.75. This suggests that their outputs, while occasionally accurate, lacked the clarity or specificity necessary for real-world editorial use.

Table 5.10: Mean (M) and standard deviation (SD) of expert ratings across four editorial dimensions, grouped by explanation type.

Explanation Type	Framing Bias	Omission of Context	Emotional Language	Actionability
Human	1.57 ± 0.62	1.70 ± 0.46	0.70 ± 0.59	1.57 ± 0.56
Gemini	0.70 ± 0.86	0.90 ± 0.83	0.37 ± 0.48	0.70 ± 0.75
GPT	0.87 ± 0.92	1.20 ± 0.68	0.43 ± 0.70	0.73 ± 0.70

To further understand how explanation quality varies by the nature of the headline, we stratified our analysis based on agreement level among human annotators, distinguishing between

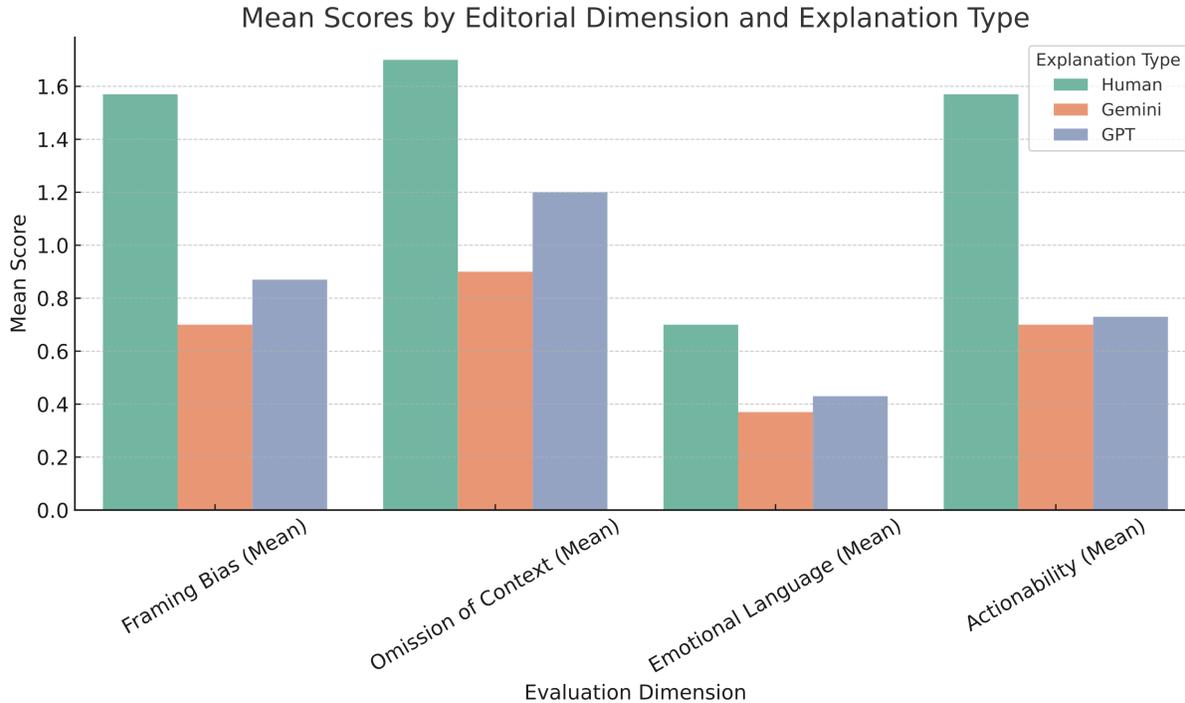


Figure 5.4: Mean expert ratings across editorial dimensions, comparing explanations written by humans, Gemini, and GPT. Error bars are not shown due to overlap in standard deviations.

ambiguous and unanimous headlines. By analyzing results at the agreement level, we aim to uncover whether LLMs are particularly sensitive to ambiguity or consistently underperform across contexts.

As shown in Table 5.11 and Figure 5.5, human explanations consistently received the highest mean scores across all four editorial dimensions in ambiguous and unanimous headline conditions. Humans maintained strong performance even in unanimous cases, with mean scores above 1.3 in *Framing Bias*, *Omission of Context*, and *Actionability*. This suggests that raters found human-generated explanations editorially applicable and contextually appropriate even when the headline meaning was more precise or less disputed.

In contrast, Gemini and GPT explanations scored considerably lower across dimensions, particularly on *Actionability* and *Framing Bias*, regardless of agreement level. This indicates a persistent limitation in the models' ability to provide actionable, bias-aware reasoning. Moreover, both LLMs showed lower ratings in *Emotional Language*, suggesting a shared challenge in identifying affective or sensational language cues.

While GPT slightly outperformed Gemini in *Omission of Context* (1.29 vs. 0.93 in ambiguous, and 1.13 vs. 0.88 in unanimous), neither approached human performance. Importantly, the addition of human data in the unanimous condition reinforces that LLMs do not simply struggle with ambiguity; they underperform even when headline meaning is relatively clear.

These findings suggest that LLMs struggle more with editorially grounded judgments when faced with ambiguous interpretations, while human evaluators can maintain high performance

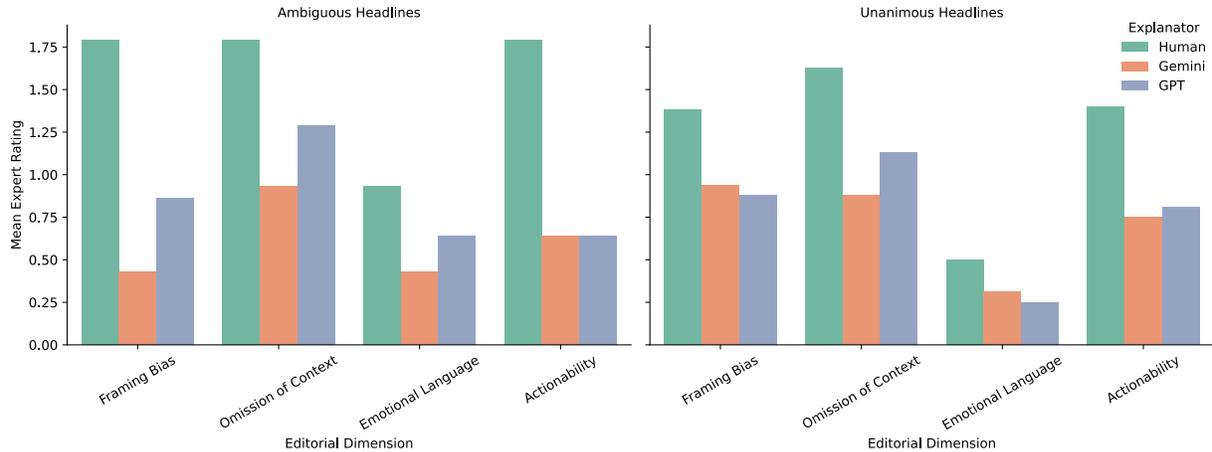


Figure 5.5: Mean expert ratings across editorial dimensions for each explanation type, stratified by agreement level. Human explanations are included only for ambiguous headlines, as unanimous cases were exclusively evaluated for LLM outputs.

even in contexts with low agreement.
 graphicx

Table 5.11: Mean and standard deviation of expert ratings across four editorial dimensions, stratified by agreement level and explanation type.

Agreement Level	Explanator	Framing Bias	Omission of Context	Emotional Language	Actionability
Ambiguous	Human	1.79 ± 0.27	1.79 ± 0.39	0.93 ± 0.61	1.79 ± 0.39
	Gemini	0.43 ± 0.61	0.93 ± 0.84	0.43 ± 0.53	0.64 ± 0.69
	GPT	0.86 ± 0.94	1.29 ± 0.49	0.64 ± 0.85	0.64 ± 0.75
Unanimous	Human	1.38 ± 0.79	1.63 ± 0.52	0.50 ± 0.53	1.40 ± 0.64
	Gemini	0.94 ± 1.02	0.88 ± 0.88	0.31 ± 0.46	0.75 ± 0.85
	GPT	0.88 ± 0.95	1.13 ± 0.83	0.25 ± 0.53	0.81 ± 0.70

5.3.5.2 Inferential Analysis

To statistically analyze differences in explanation quality among the three sources: Human, Gemini, and GPT, we conducted Friedman tests for each editorial dimension, followed by Wilcoxon signed-rank tests for post-hoc pairwise comparisons. Effect sizes (r) were approximated for each comparison to assess practical significance. As shown in Table 5.13, human explanations significantly outperformed LLM-generated explanations in dimensions such as *Framing Bias*, *Omission of Context*, and *Actionability*, with no significant differences observed in the emotional language dimension.

Framing Bias The Friedman test revealed a significant difference in how the three explanation types addressed framing bias, $\chi^2(2) = 8.77, p = .0124$. Post-hoc Wilcoxon tests showed that human explanations significantly outperformed both GPT ($p = .0354, r \approx 3.23$) and Gemini

($p = .0174, r \approx 2.32$). No significant difference was found between GPT and Gemini ($p = .5987$).

Table 5.12: Friedman and Wilcoxon test results by agreement level and editorial dimension.

Agreement	Dimension	Friedman	Wilcoxon Comparison	p-value
Unanimous	Framing Bias	$p = .368$	Human vs Gemini	.250
			Human vs GPT	.313
			Gemini vs GPT	1.00
	Omission of Context	$p = .355$	Human vs Gemini	.188
			Human vs GPT	.359
			Gemini vs GPT	.672
	Emotional Language	$p = .646$	Human vs Gemini	.500
			Human vs GPT	.375
			Gemini vs GPT	1.00
	Actionability	$p = .241$	Human vs Gemini	.219
			Human vs GPT	.219
			Gemini vs GPT	1.00
Ambiguous	Framing Bias	$p = .015$	Human vs Gemini	.016
			Human vs GPT	.094
			Gemini vs GPT	.500
	Omission of Context	$p = .048$	Human vs Gemini	.063
			Human vs GPT	.125
			Gemini vs GPT	.563
	Emotional Language	$p = .165$	Human vs Gemini	.125
			Human vs GPT	.500
			Gemini vs GPT	.625
	Actionability	$p = .021$	Human vs Gemini	.063
			Human vs GPT	.031
			Gemini vs GPT	1.00

Omission of Context A significant difference was also found for the omission of context, $\chi^2(2) = 7.39, p = .0249$. Humans again scored significantly higher than both GPT ($p = .0467, r \approx 3.10$) and Gemini ($p = .0087, r \approx 0.52$). The difference between GPT and Gemini was not statistically significant ($p = .3506$).

Table 5.13: Wilcoxon signed-rank test results comparing explanation types across each editorial dimension. Statistically significant results ($p < .05$) are bolded.

Dimension	Comparison	p-value	Effect size (r)
Framing Bias	Human vs GPT	.0354	3.23
	Human vs Gemini	.0174	2.32
	GPT vs Gemini	.5987	2.84
Omission of Context	Human vs GPT	.0467	3.10
	Human vs Gemini	.0087	0.52
	GPT vs Gemini	.3506	8.39
Emotional Language	All comparisons	> .05	—
Actionability	Human vs GPT	.0128	1.94
	Human vs Gemini	.0145	3.62
	GPT vs Gemini	.9522	5.68

Emotional Language No statistically significant differences were found for emotional language ratings, $\chi^2(2) = 3.72, p = .1555$. This suggests a general challenge across all explanation types in identifying or commenting on emotionally charged language.

Actionability The largest divergence appeared in actionability, $\chi^2(2) = 9.88, p = .0072$. Human explanations were rated significantly more actionable than those from GPT ($p = .0128, r \approx 1.94$) and Gemini ($p = .0145, r \approx 3.62$). No significant difference emerged between GPT and Gemini ($p = .9522$).

Similar to the descriptive analysis, we also conducted inferential testing to assess whether explanation quality significantly differed across explanation types within ambiguous and unanimous headline conditions. Friedman tests were used to detect overall differences among the three explanation sources (Human, Gemini, and GPT) for each editorial concern. When significant, we followed up with Wilcoxon signed-rank tests to compare pairs of explanations. The results are summarized in Table ??.

Unanimous Headlines For unanimous headlines, where crowd annotators strongly agreed on misleadingness, the Friedman tests revealed no significant differences among the three explanation types across all four dimensions ($p > .24$). Although human explanations consistently received higher average scores in the descriptive analysis, these differences were not statistically significant under stricter pairwise comparisons. This suggests that in clearer, low-ambiguity cases, LLM-generated explanations were not rated drastically worse despite trailing humans in overall scores.

Ambiguous Headlines In contrast, the analysis of ambiguous headlines paints a sharper picture. The Friedman test was significant for *Framing Bias* ($\chi^2 = 8.43, p = .0147$), *Omission of Context* ($\chi^2 = 6.09, p = .0476$), and *Actionability* ($\chi^2 = 7.75, p = .0208$). These are precisely the

dimensions that require deeper contextual understanding and editorial judgment—areas where LLMs typically struggle.

Post-hoc Wilcoxon comparisons further revealed that humans significantly outperformed Gemini in detecting *Framing Bias* ($p = .0156$) and showed near-significant trends in *Actionability* and *Omission of Context*. GPT performed modestly better than Gemini in ambiguous settings but still fell short of human-level clarity and utility. These findings reinforce the earlier conclusion that LLMs are more prone to failure in ambiguous scenarios, where detecting nuanced context, implicit omissions, or interpretive framing is required.

5.3.5.3 Qualitative Patterns in High- and Low-Rated Explanations

To complement the quantitative results, we examined explanations that received extreme scores (2 = *Clearly*, 0 = *Not at all*) across editorial dimensions. This analysis focused on ambiguous headlines, where interpretive complexity often challenges both human and machine reasoning. By comparing highly rated explanations to poorly rated ones, we observed key differences in reasoning style, specificity, and editorial utility.

Framing Bias. Highly rated human explanations frequently identified both the framing mechanism and its interpretive effect. For example, one human annotator noted: “*The headline is hiding the fact that the study was done for only those over 60. I thought it will reduce death rate for all age group, but the information is true for only senior citizens.*” In contrast, a low-rated GPT explanation simply restated the headline content without evaluating its framing: “*The headline accurately represents the content of the article, which shows a study reveals that a lack of sleep is linked to an increase in unhealthy abdominal fat.*”

Omission of Context. Top-rated explanations clearly pointed out what was missing and why it mattered. For instance, a human explanation stated: “*The headline made me assume that other states were eager to implement it. But the content says it will only work in the Airport.*” This contrasts with low-rated Gemini and GPT responses, which either ignored the omission or inaccurately stated the headline was complete.

Actionability. Explanations rated highly for actionability often provided concrete editorial reasoning, such as: “*Headline does not reveal that the drivers already have valid licenses.*” These responses gave the rater enough information to determine whether the headline should be revised or accepted. By contrast, many low-rated LLM outputs offered only neutral restatements: “*The headline reflects the content about digital driver’s licenses,*” without indicating if any corrective action was warranted.

Table 5.14 summarizes these observed contrasts, emphasizing the LLMs’ tendency to generate surface-level summaries and their limited editorial guidance in ambiguous contexts.

5.4 Discussion

This study investigated how large language models (LLMs) generate explanations for misleading news headlines in the context of subjective ambiguity, where even human annotators

Table 5.14: Illustrative contrasts between top- and bottom-rated explanations across dimensions.

Dimension	Human (Top)	LLM (Bottom)	Qualitative Insight
Framing Bias	Identifies misleading scope and explains consequence	Restates claim without critique	LLMs miss implicit framing mechanisms
Omission of Context	Names specific qualifiers or caveats	Declares accurate omissions	headline despite LLMs lack contextual sensitivity
Actionability	Indicates whether headline should be revised	Offers observation guidance	neutral without LLMs provide minimal editorial utility

diverge in their assessments of misleadingness. Through a multi-layered evaluation approach that integrates human (non-expert) annotations, model outputs, and expert journalistic assessments, we examined not only how well LLMs align with human judgment (RQ1) but also how their explanations fare under interpretive uncertainty (RQ2) what explanation features drive expert trust and editorial utility (RQ3), and how human- versus AI-generated rationales differ in surfacing key editorial concerns (RQ4).

Our findings reveal several important patterns. First, while both GPT and Gemini aligned more closely with human annotations in high-agreement (unanimous) cases, their predictive performance declined in ambiguous cases, underscoring the sensitivity of LLMs to subjectivity in annotation (RQ1). Interestingly, explanation quality, measured through expert evaluations, did not significantly degrade under ambiguity and, in some cases, was rated slightly higher, likely due to the models’ use of hedging or cautionary framing. Second, although experts rated both GPT and Gemini explanations similarly for ambiguous headlines, qualitative feedback surfaced recurring editorial shortcomings, including omissions, failure to surface uncertainty, and a lack of attribution or contextual depth (RQ2). Third, correctness emerged as the strongest predictor of expert trust in LLM-generated explanations, while ambiguity awareness and risk of harm were more predictive of actionability (RQ3). Finally, in benchmarking human and LLM explanations across editorial concerns such as framing bias and omission of context, we found that human explanations consistently outperformed LLMs in surfacing nuanced journalistic heuristics, particularly under ambiguity (RQ4).

These findings carry several implications for the CSCW and CHI communities. At a foundational level, they call for more situated evaluation frameworks that account for human disagreement, contextual subjectivity, and the editorial realities of decision-making in high-stakes domains [19, 379]. Our use of annotator agreement as a stratification mechanism provides a scalable, reproducible methodology for evaluating explanation quality in tasks where there is no single objective ground truth. At a practical level, our results highlight the need for AI systems that are not merely accurate but editorially legible, that is, capable of making their reasoning visible, accountable, and actionable for human collaborators. These insights inform the design of future human-AI systems in domains such as journalism, fact-checking, and scientific communication, where trust calibration, interpretive flexibility, and shared responsibility are essential.

In the sections that follow, we expand on these themes, exploring where LLMs align and diverge from editorial expectations, what explanation dimensions matter most to expert users, and how ambiguity itself can serve as a productive lens for designing more transparent and collaborative AI systems.

5.4.1 Subjectivity, Disagreement, and Situated Evaluation

Tasks such as detecting misleading news headlines inherently involve human interpretation and judgment. In this study, we foregrounded the role of subjectivity by explicitly stratifying headlines based on annotator agreement and analyzing model behavior across both clear and contested cases. This design choice reflects a broader recognition in CSCW and HCI that ambiguity and disagreement are not anomalies to be corrected but sociotechnical signals that can reveal underlying complexities in data, context, and reasoning [5, 45, 349].

Annotator Disagreement Is Not a Defect

Our treatment of annotator disagreement as a methodological signal, rather than a labeling error, aligns with growing calls in the CSCW community to account for diverse perspectives in collaborative data practices. Disagreement often reflects variation in lived experience, epistemological frameworks, or interpretive frames [349], especially in domains like journalism where definitions of “misleading” can depend on political, cultural, or rhetorical context. Prior work has shown that disagreement among workers, domain experts, or the public is not necessarily noise to be resolved algorithmically but rather a meaningful source of insight into social complexity and representational gaps [89, 220].

In our study, ambiguous headlines, defined as those receiving mixed misleadingness ratings, served as test cases for evaluating how LLMs respond when even human annotators diverge. The finding that LLM classification accuracy declines under ambiguity (RQ1) reinforces the challenge of operationalizing subjective misinformation detection without flattening disagreement. Importantly, we found that expert ratings of explanation quality did not significantly degrade in ambiguous cases. This suggests that models may adopt hedging or cautious reasoning strategies in response to uncertainty, offering potential affordances for communicating nuance, though, as our qualitative findings show, these strategies often lack editorial depth or explicit engagement with uncertainty.

Modeling Misinformation Beyond Binary Ground Truths

Much prior work in misinformation detection has relied on binary labels (e.g., true vs. false, misleading vs. accurate), often drawn from fact-checkers or annotation guidelines that assume an objective ground truth [224, 390]. However, our findings support arguments that such binaries may be insufficient for capturing the interpretive complexity of real-world misinformation and fact-checking scenario [25, 393]. In ambiguous cases, disagreement arises not because one side is objectively wrong but because the headline invokes framing strategies [369], omits context, or leaves room for multiple valid readings [108]. These patterns are consistent with a growing body of work that advocates for uncertainty-aware labeling and explanation in AI systems [183, 439].

By modeling ambiguity as a variable rather than a flaw, our approach reflects CSCW’s emphasis on the multiplicity of truths in collaborative systems, particularly in sociotechnical domains where expertise is distributed and contested [6, 8, 316, 376]. We suggest that misinformation detection systems, particularly those used in journalistic or fact-checking workflows, would benefit from surfacing not only whether a headline is likely to be misleading but also *why* and *under what interpretive frame*, a direction that LLMs are only partially equipped to address.

Situated Evaluation Through Triangulation

Our evaluation framework integrates perspectives from three sources: general users (who annotated the misleadingness of headlines), LLMs (which generated classifications and explanations), and professional journalists (who rated and qualitatively assessed the explanations). This triangulated design aligns with CSCW’s call for *situated evaluations* that reflect the lived practices, interpretive frameworks, and institutional norms of the task at hand [283, 399]. Rather than relying solely on accuracy or crowd consensus, we invited domain experts to serve as critical evaluators of AI explanations, using a rubric grounded in editorial heuristics such as correctness, clarity, exhaustiveness, and ambiguity awareness.

Triangulation also allowed us to identify mismatches across perspectives. For example, some explanations that aligned with annotator majority labels were still judged by experts as vague, incomplete, or misleading. Conversely, in ambiguous cases where annotator agreement was low, experts sometimes rewarded explanations that surfaced alternative interpretations or qualified their claims. These insights suggest that *explanation usefulness is not reducible to predictive alignment*, and that expert-in-the-loop evaluation can uncover deeper dimensions of AI reasoning that are often obscured in benchmark-driven assessments.

5.4.2 Alignment and Limits of LLM Explanations

Our study demonstrates that while LLMs such as GPT and Gemini can approximate human reasoning in some editorial contexts, they fall short in more nuanced tasks that require interpretive judgment, contextual reasoning, or editorial sensitivity. This section outlines both the strengths and the limitations of current LLM-generated explanations based on our multi-method analysis across the four research questions.

When LLMs Perform Well

In high-agreement scenarios where crowd annotators consistently labeled a headline as misleading or not, LLMs, particularly GPT, showed strong alignment with those judgments. As reported in RQ1, GPT achieved near-perfect classification accuracy on unanimous headlines, suggesting that current LLMs can reliably capture surface-level patterns that correspond to crowd consensus. Moreover, as revealed in RQ3, expert trust in explanations was strongly associated with the dimension of *correctness*, where GPT again tended to perform well. These findings support prior work showing that LLMs trained on large-scale web data are particularly effective in reproducing majority sentiment or canonical interpretations when ambiguity is low [47, 208, 255].

Such alignment indicates that LLMs may be suitable for low-subjectivity editorial support tasks, such as flagging obviously distorted headlines or assisting with simple factual verification. However, their apparent reliability in clear-cut cases should not be mistaken for generalizable editorial competence.

When Explanations Fall Short

The limits of LLM-generated explanations became more evident in ambiguous headline cases, where human annotators themselves disagreed. Our analysis in RQ2 and RQ4 revealed that although the average expert ratings of GPT and Gemini did not differ significantly across explanation quality dimensions, qualitative feedback from journalists highlighted key editorial shortcomings. These include the failure to provide contextual precision, inadequate engagement with emotional framing, and the tendency to collapse policy statements into personal opinion.

Rather than offering editorially grounded reasoning, LLMs often rely on surface-level textual associations or summarization logic. This results in explanations that may appear syntactically coherent and factually accurate but still fail to engage with the deeper reasoning processes that underlie professional editorial judgment. As prior CSCW work has noted, explanation usefulness is not reducible to linguistic fluency or factuality, but must also be situated within task-specific and user-centered standards of interpretability [102, 252, 370].

5.4.2.1 *Ambiguity-Handling Remains Elusive

A key test of editorial competence lies in how systems handle ambiguity. While we observed that LLMs sometimes adopt hedging strategies in ambiguous cases, using modal verbs (e.g., “may,” “could”) or presenting multiple perspectives, our qualitative analysis suggests that this rhetorical hedging rarely translates into genuine editorial awareness of uncertainty. In RQ2, expert raters repeatedly commented that explanations ignored or downplayed the interpretive challenges posed by the headline, even when human annotators disagreed.

This finding suggests that current LLMs lack mechanisms for surfacing or explicitly reasoning about disagreement or ambiguity, despite being capable of hedged language. As CSCW research on human-AI collaboration emphasizes, ambiguity should not merely be mitigated but surfaced as part of a transparent and adaptive interaction [178, 347]. From this perspective, hedging without contextual explanation amounts to deflection rather than nuance, and risks undermining expert trust in collaborative settings.

Together, these results reinforce the need for LLMs to evolve from general-purpose language predictors to context-sensitive editorial collaborators, capable of understanding, reasoning through, and communicating interpretive uncertainty in ways that support professional judgment rather than mimic surface coherence.

5.4.3 Design Implication

Our findings contribute to emerging conversations in CSCW and HCI about how to design AI systems that do not merely produce plausible outputs but instead support expert interpretation, domain-sensitive judgment, and editorial accountability [76, 216, 423]. While LLMs can generate linguistically fluent and factually correct explanations, these qualities alone

are insufficient for supporting professional editorial decisions. In this section, we offer design-relevant interpretations of our findings and propose directions for creating human-AI systems that are better aligned with journalistic workflows and values.

Editorial Usefulness vs. Explanation Correctness

As shown in our analysis of RQ3, expert trust in LLM-generated explanations was primarily predicted by *correctness*, reflecting a preference for factual accuracy and alignment with article content. However, *actionability*, the extent to which an explanation is useful for making editorial decisions, was shaped by different dimensions: particularly *ambiguity awareness* and *risk of harm*. This suggests that even technically accurate explanations may fail to meet the standards of editorial usefulness if they do not surface uncertainty, qualify interpretive stakes, or signal caution around potentially harmful framing. This distinction supports recent arguments that trust in AI is multi-dimensional and must be disentangled from usefulness in real-world contexts [214, 400]. For decision-support tools in journalism, legal review, or scientific communication, correctness is necessary but not sufficient. Editorial decisions often hinge on recognizing when information is incomplete, when language might be misread, or when a headline might evoke unintended political or emotional implications [155, 272]. LLMs must be designed not only to match facts but also to support these meta-interpretive judgments.

Supporting Expert Oversight, Not Automation

Our expert evaluations (RQ2, RQ4) revealed that journalists frequently identified gaps in LLM explanations, including the omission of source attribution, failure to contextualize scientific claims, and superficial handling of emotionally charged language. These critiques point to a broader limitation: LLMs are not yet capable of performing the kind of layered editorial judgment that experienced human editors apply. Instead of attempting to automate such reasoning, AI systems should be designed to *support* expert oversight by making their reasoning legible, revisable, and context-aware. In line with CSCW's emphasis on co-adaptive systems [121, 331], we advocate for AI explanations that expose their internal assumptions, cite evidence sources when possible, and offer justification pathways that experts can interrogate. Systems might also allow experts to modify or flag explanations, creating a feedback loop that improves future model responses or editorial alignment.

Designing for Human–AI Editorial Collaboration

Our findings contribute to emerging conversations in CSCW and HCI about how to design AI systems that do not merely produce plausible outputs, but instead support expert interpretation, domain-sensitive judgment, and editorial accountability. While LLMs can generate linguistically fluent and factually correct explanations, these qualities alone are insufficient for supporting professional editorial decisions. In this section, we translate our findings into concrete design recommendations for human-AI systems in editorial and fact-checking contexts, where interpretive complexity and subjective reasoning are integral to professional practice.

AI explanations in collaborative, editorial decision-making settings can augment professional judgment by surfacing context, alternatives, and risk. However, to foster productive

human-AI collaboration, these explanations must be clear, trustworthy, and contextually relevant. Drawing on our multi-method analysis across expert ratings and qualitative feedback (RQ2–RQ4), we identify four actionable design guidelines for supporting expert-AI collaboration in editorial settings.

Dimension Highlighting. Explanations should be structured to reflect editorial dimensions such as correctness, ambiguity awareness, and risk of harm. Interfaces can support this by visually tagging sentences or clauses according to these dimensions, enabling experts to filter or prioritize information based on their decision-making goals quickly. Beyond tagging, systems might also provide modular scoring views or dimension-specific summaries that give experts a high-level assessment of how well an explanation performs across editorial dimensions [103]. Prior work shows that users prefer explanations that align with their task-specific goals and support their workflow contexts [212]. Our findings confirm that experts weigh different editorial qualities depending on the outcome of interest, such as correctness when assessing trust or ambiguity awareness when deciding on editorial action (RQ3).

Framing Bias identification. To support the identification of framing bias in human-AI editorial collaboration, we propose several interface-level design ideas that can surface rhetorical emphasis and interpretive slant. First, contrastive framing views could present alternative headline phrasings that emphasize different narrative frames, helping users detect how wording shifts affect interpretation, drawing inspiration from contrastive explanation in NLP [190, 432] and foundational framing theory [118]. Second, systems can highlight ideologically charged or emotionally loaded terms using frame-tagging methods informed by framing lexicons [129] or automated frame detection models [163]. Third, toggle-based designs could let users switch between alternative rhetorical framings [228] (e.g., individual- vs. institution-focused), offering deeper insight into possible bias. Finally, editorial intent elicitation mechanisms could prompt users to articulate their intended framing and compare it against the system’s explanation, highlighting mismatches. These framing-aware designs collectively support critical reflection and empower experts to interrogate model outputs and their alignment with professional editorial standards.

Ambiguity Surfacing. Rather than treating ambiguity as a failure mode, systems should actively surface interpretive uncertainty and plural viewpoints. This can be achieved by offering multiple plausible LLM rationales for a given claim or explicitly flagging underdetermined or debatable aspects of the headline. Such representations can support expert sensemaking [179], particularly in cases of contested interpretation. Prior work shows that making disagreement visible in collaborative systems encourages reflection, deliberation, and more informed decision-making [172]. In our study, experts consistently valued explanations that acknowledged complexity, uncertainty, or alternate readings, especially when evaluating ambiguous headlines (RQ2). This reinforces the need for interfaces that embrace interpretive pluralism in editorial tasks.

Justification Toggling. Editorial professionals operate under varying time pressures and decision needs, ranging from rapid triage to detailed revision. Systems should support these

diverse workflows by enabling users to toggle between concise and elaborated explanations or between assertive and hedged reasoning styles. This allows users to control the granularity and tone of AI-generated justifications based on the editorial context. Prior research in explainable AI and HCI has shown that explanation completeness and tone significantly influence user trust and reliance [117]. In our study, various expert critiques expressed a need for explanations that could adapt to current circumstances, neither overly confident nor excessively vague, highlighting the significance of context-aware presentation in expert-AI collaboration.

Editable Reasoning. In editorial settings, AI-generated explanations should be viewed not as final outputs but as collaborative scaffolds that support expert intervention and refinement. Interfaces should enable experts to annotate, revise, or reframe model rationales, facilitating collaborative auditing [92], documenting editorial judgments, and encouraging reuse in downstream tasks such as fact-checking, corrections, or transparency reporting. This supports a shift from explanation as static output to explanation as a negotiated, co-constructed artifact, aligning with CSCW's emphasis on collaborative sensemaking [180, 301]. Prior work on interactive machine learning and human-AI collaboration stresses the value of user-editable AI systems to ensure trust, oversight, and system adaptability [433]. In our study, journalists repeatedly expressed a desire to retain control over language, nuance, and framing, suggesting that editable explanations are critical for preserving editorial integrity and professional agency in human-AI partnerships.

Collectively, these guidelines move beyond generic interpretability to support a more situated, accountable, and interactive form of human-AI editorial collaboration. Future work should prototype these interface features and test them in journalistic or fact-checking workflows to evaluate their impact on trust calibration, task efficiency, and editorial integrity.

Integrating Human–AI Collaboration into Newsroom Workflows

As generative AI tools become increasingly embedded in journalistic workflows, it is vital to envision how AI–human collaboration, such as the model proposed here, can be functionally integrated into real-world editorial pipelines. While prior tools like The Washington Post's *Heliograf* focused on auto-generating election and sports articles from structured data [132], and Bloomberg's summarization AI supports financial news abstraction [55], these applications primarily center around automation. In contrast, the workflow proposed here foregrounds collaborative interaction, where AI systems assist journalists in reviewing and refining potentially misleading headlines through explainable analysis and revision suggestions.

Figure 5.6 illustrates an enhanced editorial pipeline that integrates AI into the headline review loop, with attention to both textual and visual framing. Each stage of the workflow builds on the previous, supporting traceable, multimodal collaboration.

While this dissertation emphasizes misleading news headlines, incorporating visual elements, especially feature images, is crucial for creating a realistic and effective editorial workflow. In digital news environments, visuals are often the first point of contact between a story and its audience. Feature images accompanying headlines can convey mood, bias, or sensationalism that may reinforce or contradict the headline's framing. Studies have shown that visuals strongly shape reader expectations, attention, and emotional response, sometimes even overriding textual cues [328]. Therefore, evaluating image–headline coherence

is essential to understanding how misleading impressions are formed and how they might be corrected. Incorporating this multimodal component allows the AI-assisted workflow to more accurately reflect how headlines function in the real world, particularly in mobile-first and social media-driven ecosystems.

Step 1: Headline Drafting. At the beginning of the process, journalists and copy editors collaborate on the final draft of the article to create one or more potential headlines. These headlines aim to convey the main message of the article while also drawing in reader interest. At the same time, they choose a selection of feature images, such as hero images or thumbnails, that visually correspond to different aspects or tones of the story. This set may include multiple options, allowing editorial staff or AI systems to suggest the most coherent or effective pairing. The combination of article, headline, and image options forms a cohesive content bundle that is passed on for further review. This bundle serves as the foundational input for AI analysis, and the visual component is especially important, given its powerful influence on headline interpretation in mobile and social media contexts.

Step 2: AI-assisted Editorial Review. Next, a large language model (LLM)-based editorial assistant evaluates the full content bundle, including headline, article, and selected image(s). The system examines whether the headline contains problematic framing, emotionally charged or vague language, or omissions of critical context. It also assesses whether the headline and image are in semantic harmony, ensuring that the visual framing doesn't unintentionally distort the headline's meaning. The AI generates detailed feedback, including highlighted text, revision prompts, explanations for why certain elements may be misleading, and cues that indicate interpretive ambiguity. If multiple images are available, the system may recommend the best visual pairing based on thematic alignment or perceived audience impact. This diagnostic output supports editors in identifying headlines that merit closer human attention and refinement.

Step 3: Human Review and Fact-checking. Informed by the AI-generated suggestions, human editors and fact-checkers then evaluate the content package using professional judgment. They determine whether revisions to the headline are needed, whether the AI's concerns are substantiated, and whether the image reinforces or undermines the headline's clarity. Editors can choose to adopt, reject, or refine the AI's suggestions. This process ensures that final editorial decisions uphold journalistic standards of fairness, accuracy, and clarity while still benefiting from automated triage and explanation tools. If necessary, the headline or image may be adjusted to mitigate the risk of reader misinterpretation.

Step 4: Publishing and Monitoring. Once finalized, the revised headline, article, and image are published through the newsroom's content management system. The publication is then monitored for audience engagement and response. Real-time analytics such as click-through rates, bounce rates, time-on-page, and social sharing behavior offer insights into how the content is being interpreted. Additionally, qualitative signals like reader comments or external complaints can serve as indicators that the headline may still be misleading or misaligned with audience expectations.

Step 5: Feedback Loop. Post-publication performance data feeds back into the editorial process. If a headline underperforms or receives negative feedback, editorial teams can revisit the content, revise the headline or image, and issue corrections as needed. In some cases, AI systems may automatically flag anomalies in audience behavior or language mismatches and suggest follow-up revisions. This loop enables the newsroom to adapt to evolving interpretations, ensure long-term accuracy, and build trust with readers.

Compared to prior uses of AI in journalism, this workflow emphasizes **collaboration**, **interpretability**, and **multimodal sensitivity**. The integration of visual framing into headline analysis acknowledges that misleading cues often arise not just from text, but from image–headline pairings, an especially important factor in algorithmic content distribution and mobile-first news consumption. The ability to trace decisions across stages, supported by explainable AI outputs and human oversight, creates a more transparent and accountable editorial pipeline.

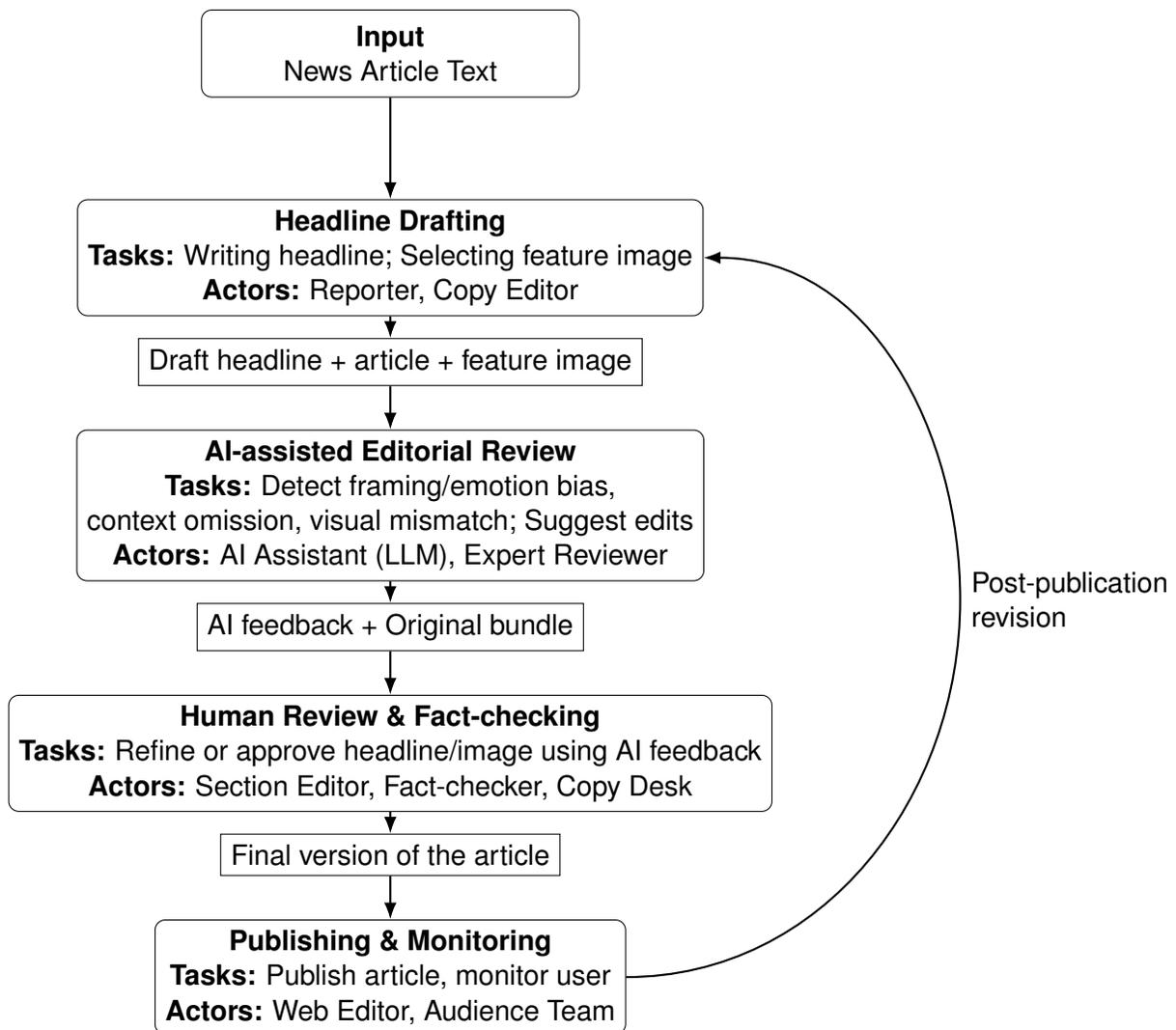


Figure 5.6: Proposed Human-AI Collaboration Workflow for Headline Writing and Verification

5.4.4 Limitations and Future Directions

This study provides valuable insights into the evaluation of explanations generated by large language models (LLMs) in editorial contexts. However, there are several limitations that need to be addressed. Firstly, the sample size for expert evaluations was relatively small. Despite using multiple annotators and models for triangulation, the limited number of headlines ($n = 15$ for in-depth expert analysis) restricts the generalizability of our statistical findings. Future research with a larger and more diverse set of headlines covering various domains such as science, politics, and health would enable a broader understanding of how the quality of explanations and the level of trust differ across different contexts.

Second, our analysis relied on static LLM outputs generated at a single point in time. Given the rapid evolution of LLM capabilities and prompt engineering techniques, the quality and nature of explanations may vary significantly across versions or interaction modalities (e.g., chat vs. batch generation). Incorporating interactive, real-time expert-AI collaboration workflows in future work could better reflect how explanations are judged in practice and how experts shape or adapt them over time.

Third, the LLM configurations used in this study were limited to default settings, such as the default temperature, max tokens, and decoding parameters offered by the APIs (e.g., OpenAI's GPT-4 and Google's Gemini). Additionally, we did not experiment with multi-turn or feedback-driven prompting styles, where expert preferences or critiques could dynamically shape explanation outputs. As a result, the quality and appropriateness of explanations may not represent the full potential of these models under optimized or adaptive prompting strategies. Future work should explore how configuration tuning and interactive refinement affect explanation quality, editorial trust, and usability.

Fourth, there is a potential temporal overlap between the time period of our article dataset and the training cutoff of the LLMs. The headlines used for evaluation were collected during 2022, while the models were queried in 2024. This raises the possibility that some of the article content may have been included in the training data of the LLMs, especially in the case of publicly available web-scraped content. If so, this could bias the explanation quality, as the models may be reproducing previously seen material rather than generating generalizable editorial reasoning. Although we attempted to avoid direct leakage by selecting ambiguous or contested headlines, future studies should explicitly control for data leakage and examine performance on truly out-of-distribution or novel content.

Fifth, while we concentrated on key editorial aspects like correctness, awareness of ambiguity, and potential risk of harm, there are other interpretive factors to consider, such as political bias, epistemic framing, and audience sensitivity that may also influence the usefulness and trustworthiness of explanations. Expanding the evaluation criteria to include these domain-specific concerns would enhance our understanding of expert AI alignment in high-stakes editorial contexts.

Finally, while this study involved professional journalists as expert raters, future research could explore additional stakeholders in collaborative fact-checking ecosystems, including policy analysts, media scholars, or community advocates. Doing so would further diversify the understanding of what constitutes an “actionable” or “trustworthy” explanation depending on role, expertise, and institutional context.

Future research should also explore the implementation of the proposed design guidelines

in working editorial systems. Prototyping explanation interfaces that surface ambiguity, enable dimension-specific reasoning, and support expert edits could provide critical insights into how trust calibration and decision-making unfold in real-world newsroom environments. Longitudinal [146] or in-the-wild studies [68] would be especially valuable for assessing the co-adaptive potential of such systems over time.

5.5 Conclusion

This study investigated how large language models (LLMs) generate and justify predictions about misleading news headlines, particularly in cases characterized by subjective ambiguity and annotator disagreement. By triangulating human annotations, LLM outputs, and expert evaluations, we examined both the alignment of model predictions with human consensus and the editorial adequacy of model-generated explanations. Our stratified, expert-in-the-loop evaluation approach reflects CSCW's commitments to situated reasoning, interpretive pluralism, and task realism. We found that LLMs perform well in high-agreement cases but often fail to surface critical editorial concerns, such as framing bias, omission, and emotional manipulation, in ambiguous contexts. While correctness most strongly predicts expert trust, dimensions like ambiguity awareness and harm sensitivity are more indicative of editorial actionability. Drawing from these findings, we proposed design implications for human-AI collaboration in editorial workflows, including mechanisms for ambiguity surfacing, dimensional transparency, and expert-editable reasoning, advancing the development of context-aware, accountable AI systems in journalism and beyond.

Chapter 6: Discussion, Limitation and Future Work

This dissertation sets out to investigate a pressing but understudied form of misinformation: Misleading News Headlines. Rather than focusing solely on outright falsehoods or deceptive content, the work concentrates on subtle linguistic and contextual manipulations that can distort reader understanding even when the article content is technically accurate. The central goal is to characterize how misleadingness is perceived and produced in contemporary news ecosystems, assess the effectiveness of correction strategies, and explore how emerging AI systems, particularly large language models (LLMs), can assist or hinder editorial reasoning.

To pursue this multifaceted goal, the dissertation adopts a mixed-methods approach across three studies, integrating qualitative, quantitative, and hybrid designs to enable methodological triangulation. Chapter 3 employs qualitative interviews to explore how journalists and readers define, experience, and respond to misleading headlines. Chapter 4 transitions to a quantitative experimental design to measure how specific headline correction strategies influence credibility, engagement, and interpretation outcomes at scale. Chapter 5 brings these strands together by combining rubric-based expert evaluation (qualitative) with descriptive and comparative analyses (quantitative) to assess how LLMs explain misleadingness in both ambiguous and non-ambiguous cases. This integration allows the research to move fluidly between in-depth conceptual insights and broader empirical generalizations, strengthening the validity and applicability of the findings through complementary evidence sources [84] and methodological cross-checks [198].

The research unfolds across three interconnected studies. The first examines how journalists and readers conceptualize misleadingness, what factors lead to misleading headlines, and what strategies they propose for correction. Through interviews with both producers and consumers of news, the study reveals deep-seated tensions between journalistic pressures for engagement and audience expectations for clarity and trust. It highlights not only the institutional and algorithmic constraints shaping headline construction but also the interpretive struggles readers face when headlines omit key context or evoke emotional manipulation. These insights form the foundation for understanding the socio-editorial complexity of the problem.

Building on these findings, the second study moves from diagnosis to intervention. It empirically tests how specific correction strategies, such as adding uncertainty cues, restoring omitted context, and removing emotionally charged language, impact reader outcomes. Using a between-subjects experimental design, the study evaluates whether these ethically motivated corrections come at the cost of user engagement. The results show that corrections do not diminish click-through interest; in fact, some strategies enhance perceived credibility and interpretation accuracy. These findings challenge long-standing assumptions in journalism that accuracy and engagement are in conflict and provide behavioral evidence in support of more responsible headline design.

The final study shifts focus toward the role of AI in supporting editorial decisions. As

LLMs become increasingly integrated into news production pipelines, their ability to detect and explain misleadingness becomes crucial. This chapter introduces a stratified evaluation framework that accounts for human disagreement by categorizing headlines into high- and low-agreement groups. Through expert-in-the-loop assessments using a rubric grounded in editorial reasoning, the study evaluates how GPT and Gemini perform when explaining headline misleadingness. It finds that while models align well with human reasoning in clear-cut cases, they often struggle in ambiguous ones, failing to surface interpretive nuance or justify their decisions with editorial depth. Nevertheless, the framework developed here offers a reproducible, situated method for evaluating explanation quality in tasks where subjectivity plays a central role.

Together, these studies make several core contributions. First, they provide a qualitative account of how misleading headlines are created and contested, uncovering the diverging expectations of journalists and readers. Second, they offer empirical evidence that headline corrections, when done thoughtfully, can improve trust and comprehension without sacrificing engagement. Third, they advance a situated, triangulated framework for evaluating AI-generated explanations in complex editorial domains. Finally, they translate these findings into design-relevant insights for future systems that aim to support human-AI collaboration in newsrooms, balancing algorithmic power with professional judgment and public accountability.

6.1 Implications

6.1.1 Reframing Misinformation: Beyond Factuality Toward Interpretive Misleadingness

A key implication of this dissertation is the need to broaden prevailing conceptualizations of misinformation beyond binary classifications of “true” or “false” content. Across all three chapters, we found that misleadingness in headlines often stems not from outright fabrications but from more nuanced forms of distortion, such as omission of critical context, emotionally manipulative phrasing, or framing bias. These techniques influence reader interpretation, trust, and emotional engagement without violating factual accuracy *per se*.

Existing misinformation research has made significant strides in building automated fact-checking systems [160] and misinformation detection models [188]. However, these systems often rely on rigid ground truths or veracity labels that fail to account for the interpretive nature of much real-world misinformation, especially in journalistic contexts. Our findings support the argument that misleadingness should be treated as an editorial and interpretive concern, not just a factual one, requiring frameworks that account for narrative structure, rhetorical emphasis, and epistemic subtlety [66].

This dissertation contributes to such a reframing by offering empirical and conceptual tools to identify and intervene on interpretive misleadingness. Chapter 3 shows that readers and journalists alike are sensitive to subtleties in framing and emotional tone, but often disagree on what constitutes distortion. Chapter 4 demonstrates that corrections targeting these features can meaningfully improve comprehension and credibility without compromising engagement. Chapter 5 further reveals that current AI systems struggle to detect or explain such subtle manipulations, especially in ambiguous cases. These insights collectively underscore the need for a new class of misinformation interventions that address how information is presented, not

just what is said.

Recognizing misleadingness as an interpretive harm shifts the design goals of misinformation detection tools. Rather than aiming solely for factual correctness, future tools, both human-driven and AI-assisted, should incorporate framing diagnostics, contextual sensitivity, and rhetorical transparency as core features. This reorientation aligns with recent efforts in journalism, psychology, and HCI to study the nuanced ways information can be truthful yet misleading, and to design systems that foreground interpretive clarity over simplistic veracity checks [95, 413].

6.1.2 Rethinking Media Trust and Responsibility in a Platform-Driven Ecosystem

Another central implication of this dissertation concerns how trust in news is formed, eroded, and repaired in the face of headline-level distortion. Our findings, especially from Chapters 3 and 4, show that even subtle headline manipulations, those that may not cross into overt falsehood, can provoke reader frustration, disengagement, and longer-term distrust toward news outlets. Importantly, this erosion of trust often occurs even when the article content is accurate, highlighting the disproportionate weight that readers place on headlines as stand-alone informational artifacts.

This challenges the dominant framing of misinformation as a consumer-side literacy problem alone. While many journalism and media literacy initiatives continue to emphasize improving readers' critical reasoning skills, our work suggests that producers' responsibilities must be considered just as seriously. News consumers in our interviews consistently placed the burden of clarity and fairness on journalists and platforms, expecting headlines to be self-contained and trustworthy without requiring vigilant decoding. Meanwhile, journalists themselves often defended ambiguous or emotionally framed headlines as necessary for SEO, visibility, and audience retention, underscoring the commercial and algorithmic pressures embedded within headline production.

This disconnect reflects a structural dilemma in today's information ecosystem, where headline design is shaped not only by editorial norms but by opaque platform logics [270] and attention-driven metrics [335]. Our findings reveal that readers feel deceived when their expectations, formed by the headline, are not met by the article, whether due to omitted context, exaggerated claims, or emotionally loaded language. Such misalignments not only weaken trust in individual articles but contribute to broader [290].

Addressing this tension requires moving beyond blame-based models of misinformation to shared accountability frameworks [165]. This dissertation argues that media trust must be co-constructed: journalists, platforms, and readers all have roles to play, but their responsibilities differ. Platforms should support transparency mechanisms (e.g., headline audit trails, framing indicators), while newsrooms should adopt clearer headline standards and review practices. At the same time, readers should be given interpretive tools, not just fact-checking cues, to help them assess how headlines are structured and why certain framings are used.

By foregrounding the tension between producer intent and audience interpretation, this work highlights the need for computational and design interventions that go beyond veracity checks. Future systems must support mutual trust by making the editorial process more

transparent and by aligning platform incentives with credibility, not just engagement.

6.1.3 The Role of Ambiguity in Human-AI Collaboration

A recurring insight across Chapters 3 and 5 is that ambiguity is not merely an obstacle in information work; it is a structural feature of editorial reasoning that must be surfaced and negotiated, not suppressed. In many cases, headlines provoked differing interpretations among both annotators and experts, particularly when they involved rhetorical framing, omitted qualifiers, or emotionally ambiguous language. Our evaluation of LLM-generated explanations showed that while models can hedge with modal verbs (“may,” “could”), they often fail to articulate why ambiguity exists or how it affects interpretation, offering surface-level caution without deeper editorial awareness.

This has significant implications for how AI systems are integrated into collaborative, high-stakes domains like journalism. In ambiguous cases, accuracy alone is insufficient; what experts value most is the model’s ability to acknowledge uncertainty, surface plausible alternatives, and clearly state interpretive stakes. Our findings demonstrate that trust and editorial actionability are more strongly predicted by explanation characteristics like ambiguity awareness and risk sensitivity than correctness alone.

To support such nuanced tasks, human-AI systems must treat ambiguity not as a failure mode but as a design signal. Interfaces should be capable of highlighting contested interpretations, presenting multiple rationales, and allowing expert revision of machine-generated outputs. These strategies not only improve expert oversight but align with ongoing HCI/CSCW efforts to design co-adaptive systems that embrace pluralism and resist the flattening of subjectivity in complex decision-making contexts [139, 347].

6.1.4 Understanding the Impact of AI on News Ecosystems

As generative AI becomes increasingly integrated into journalistic workflows, its influence extends beyond productivity gains to shaping the epistemic, ethical, and institutional dimensions of news production. While this dissertation primarily explores how AI can assist in evaluating and correcting misleading headlines, the findings also raise broader questions about the impact of AI on editorial reasoning, labor, and public trust in the media.

One significant implication concerns the shifting role of editorial labor. Tasks once considered highly interpretive, such as detecting framing bias, assessing ambiguity, or explaining misleadingness, are now partially automatable through large language models (LLMs). While this opens up possibilities for decision support and editorial triage, it also risks deskilling editorial staff, delegitimizing human expertise, and transferring normative authority to algorithmic systems trained on opaque, large-scale corpora [59, 97]. In such settings, AI does not merely assist editorial reasoning but may begin to influence what is considered newsworthy, misleading, or appropriate for correction.

This shift introduces concerns around value alignment and institutional influence. As Chapter 5 illustrates, LLMs perform reasonably well when explaining misleading headlines but falter when interpretive ambiguity is involved. Yet, even in these uncertain contexts, AI-generated explanations may be taken at face value, particularly when they are confidently presented or integrated into automated workflows. Without safeguards, this could lead to premature editorial

closure or the normalization of AI-centric standards for judgment [40, 242]. These risks are especially pressing in domains where political, cultural, or epistemic nuance is central to headline interpretation.

Moreover, using LLMs trained on predominantly Western or mainstream news sources raises questions about the homogenization of editorial norms. If AI tools are deployed widely across global newsrooms without localization or human contextualization, they may inadvertently flatten culturally specific norms around accuracy, tone, or audience sensitivity [340]. This standardization may reduce editorial pluralism and marginalize perspectives from underrepresented communities, counteracting efforts to promote diversity and inclusivity in news coverage.

Finally, the deployment of AI in editorial contexts demands stronger ethical guardrails. While this dissertation contributes one approach, expert-in-the-loop evaluation guided by domain-specific rubrics, there remains a broader need for transparency, accountability, and human oversight in the design and deployment of AI systems in journalism. This includes disclosing when AI systems are used, surfacing model uncertainty or disagreement, and allowing human editors to contest or override AI-generated judgments. Without such affordances, AI tools may reproduce or even amplify the same editorial harms they are meant to mitigate [37, 344].

Taken together, these insights suggest that AI's impact on news ecosystems is not merely technical or operational, but deeply normative. Designing editorial AI requires more than optimization; it requires critical engagement with how journalistic values are interpreted, encoded, and enforced by machines.

6.1.5 Designing for Editorial Collaboration in the Age of Generative AI

As generative language models become increasingly embedded in editorial workflows, from automated headline generation to content summarization and verification [240, 288], the question is no longer whether AI will assist in journalism, but rather how it will do so. Across this dissertation, we have shown that while LLMs can replicate consensus-driven reasoning in clear-cut cases, they fall short in ambiguous scenarios that demand contextual awareness, framing sensitivity, and editorial discretion. These gaps suggest a shift in framing is needed: from AI as a standalone decision-maker to AI as a collaborative agent [64] that supports, but does not replace, human judgment.

Our findings underscore the need for tools that foreground editorial dimensions like ambiguity, context omission, and potential harm, not just factual correctness. Effective human-AI collaboration in editorial settings should go beyond explanation plausibility toward systems that expose reasoning steps, justify interpretive choices, and allow journalistic users to annotate, revise, or override AI outputs. Such a design orientation centers human expertise and professional responsibility while enabling AI to serve as a scaffold for critical reflection, not a final authority.

To that end, this dissertation contributes not only empirical insights but also conceptual foundations for building assistive, editable, and context-aware explanation systems in journalism. In doing so, it invites the HCI and CSCW communities to rethink editorial support not as a matter of prediction accuracy alone, but as a co-constructed process of interpretive alignment, transparency, and shared accountability between human and machine agents [419].

6.2 Future Work

While this dissertation offers a multi-perspective understanding of misleading headlines, from perception and correction to LLM-based explanation, several important avenues remain unexplored. Future work can expand the scope of this research by addressing genre and platform diversity, developing interactive systems for collaborative news correction, and investigating the long-term impacts of editorial strategies on audience trust and behavior. These directions aim to build on the current findings while pushing toward more practical, scalable, and context-sensitive solutions for responsible news consumption and AI-supported journalism.

6.2.1 Expanding Genre and Platform Diversity

This dissertation primarily focused on textual news headlines from general-interest news outlets, analyzed within structured experimental and evaluative settings. However, the ecosystem of misleading information spans far beyond this context. Future work should explore how misleading headline strategies manifest across diverse journalistic genres (e.g., opinion, science, tabloid, and local news) and on varied digital platforms such as YouTube thumbnails, Instagram posts, or TikTok captions. Each medium imposes different stylistic constraints and audience expectations, which may affect both how misleading content is perceived and how correction strategies should be designed. Integrating multimodal content (e.g., images, video, or audio) may also present new challenges for LLM explanation and editorial accountability that remain underexplored.

6.2.2 Editorial Constraints: SEO Optimization and Paywall Design

Future work should investigate how structural and commercial constraints, such as SEO optimization practices and paywall configurations, shape the production and perception of misleading headlines. While this dissertation surfaced these factors through qualitative insights, more targeted studies are needed to systematically evaluate how algorithmic visibility incentives influence headline composition across different newsroom contexts. Experimental work could assess how SEO-driven word choices affect perceived credibility or click behavior, and whether more transparent or journalist-informed SEO guidelines could mitigate distortion.

Additionally, the role of paywalls in fragmenting reader access remains underexplored. Many participants noted that paywalls leave readers with only the headline and subheading, potentially exacerbating the spread of misleading interpretations. Future work could explore alternative paywall strategies (e.g., soft paywalls, credibility previews, or headline-context bundles) that balance financial sustainability with access to contextually rich information. Design interventions in this space could help news organizations preserve audience trust without compromising revenue models.

6.2.3 Longitudinal Studies on Trust and Corrective Framing

This dissertation's findings highlight how emotional and contextual corrections affect immediate perceptions of credibility and interpretation. However, we lack understanding of how such corrections influence long-term trust in news sources, headline literacy, and behavior

change. Future work should adopt longitudinal study designs to track whether repeated exposure to corrected vs. misleading headlines shapes audience trust, recall, or skepticism over time. These studies could also test whether personalized corrective framing (e.g., tailored to users' political orientation or prior knowledge) enhances interpretive alignment without introducing new risks of confirmation bias or ideological filtering. Such work would provide deeper insight into the cumulative impact of corrective editorial strategies and their integration into broader media literacy interventions.

6.2.4 Prototyping and Real-World Testing of Human–AI Editorial Collaboration

While this dissertation establishes the conceptual and empirical groundwork for Human–AI collaboration in editorial tasks, future work should move toward building a functional prototype that operationalizes these insights. Such a system could integrate LLM-generated explanations, editorial dimension tagging (e.g., ambiguity awareness, framing bias), and expert-editable reasoning to support real-time newsroom decision-making.

A key direction would involve deploying this prototype within active newsroom settings to evaluate how journalists interact with, modify, or override AI-generated justifications. Observational studies and in-the-wild deployments would help assess usability, trust calibration, and workflow integration in practice. This work would also surface the organizational and cultural dynamics that shape how AI tools are adopted or resisted in professional journalism environments. By embedding collaborative AI systems into real editorial contexts, future research can close the gap between theoretical affordances and practical utility, enabling scalable, trustworthy, and accountable news production.

6.3 Limitations

This dissertation presents a comprehensive exploration of misleading news headlines; however, several limitations should be acknowledged.

This dissertation conceptualizes misleading headlines as a form of misinformation, focusing specifically on unintentional or ambiguous distortions rather than deliberate deception. In doing so, it diverges from studies that frame such issues within the lens of disinformation, defined as the intentional spread of false or misleading information. This choice reflects the aim to explore how structural, linguistic, and editorial practices can inadvertently produce misleadingness even when the article content is factually accurate [418]. While this framing avoids ascribing malicious intent to journalists, it also means that the dissertation does not address intentionally misleading headlines crafted for strategic or ideological manipulation. As such, this constitutes a limitation of the current scope. However, the linguistic and editorial features examined, including contextual omission, emotional framing, and ambiguity, are not limited to unintentional scenarios. These same mechanisms are often leveraged in disinformation campaigns, particularly in partisan or manipulative media environments. Therefore, the frameworks and findings developed in this dissertation can serve as a foundation for future research that explicitly targets the disinformation variant of misleading headlines, including efforts to detect, mitigate, or trace intentional influence operations.

The scope of the headline samples in this study was deliberately limited to ensure analytical feasibility and thematic coherence. This dissertation specifically concentrated on soft news areas, including science, technology, and lifestyle topics, in order to minimize the confounding effects of political polarization. This focus allowed for a more controlled investigation of the characteristics of misleading headlines. While this topical focus supported comparative analysis across studies, it also limited the generalizability of findings to more contentious domains such as politics or crime reporting, where audience interpretations and editorial strategies may differ significantly. Sample sizes were modest, ranging from a small number of articles in Chapter 3 interviews to 12 headlines in the Chapter 4 experiment and 60 in the Chapter 5 evaluation. This modest size may further limit the range of misleading techniques captured. Future research should expand headline sampling across both topic areas and media ecosystems to explore genre-specific patterns of misinformation and examine whether correction strategies and explanatory tools transfer across domains.

Participant representation across the three studies presented several limitations that constrain the generalizability of findings. In Chapter 3, although the study engaged both general readers and professional journalists, the sample sizes were small and lacked sufficient demographic and geographic diversity. The general reader participants shared similar sociocultural backgrounds, and the expert journalist sample did not reflect the full range of newsroom roles, regional experiences, or editorial philosophies, limiting the diversity of perspectives on misleadingness and correction strategies. In Chapter 4, participants were recruited through Amazon Mechanical Turk (MTurk), which, despite being a widely used platform for behavioral research, poses notable challenges. Certain demographic groups, such as older adults, individuals from underrepresented racial or geographic communities, and those with varying levels of education or religiosity, are often underrepresented on MTurk, leading to potential sampling bias. Self-selection bias, variable motivation, and the limitations of self-report measures also affect the interpretability of key constructs like credibility and engagement. In Chapter 5, the expert evaluations of LLM-generated explanations were conducted with a limited number of journalists, primarily from similar editorial environments. This lack of geographic and institutional diversity may have introduced shared biases or normative assumptions, constraining the breadth of feedback on LLM explanation quality and its alignment with broader journalistic standards. Collectively, these limitations highlight the need for future work to engage more demographically and professionally diverse samples across regions, media systems, and stakeholder roles to build a more inclusive and externally valid understanding of headline misleadingness, correction strategies, and human-AI collaboration in editorial settings.

Another key limitation of this dissertation is its primary focus on the linguistic dimensions of misleading headlines, such as ambiguity, contextual omission, emotionality, and vagueness. These features were chosen because they are measurable and directly tied to the textual framing of news content, making them amenable to both human and AI-assisted evaluation and intervention. However, this linguistic focus necessarily excludes broader socio-technical factors that also shape the production and circulation of misleading headlines. Important influences such as newsroom economic pressures, algorithmic curation, editorial time constraints, and platform-level incentive structures remain outside the scope of this study [58, 97]. These structural forces play a significant role in shaping both editorial decision-making and audience exposure. As such, future work should move toward a more integrated analysis that combines linguistic analysis with an examination of the socio-technical ecosystem in which headlines are produced and consumed.

Doing so would offer a more comprehensive understanding of how and why misleadingness emerges and how to design systemic interventions to mitigate it.

Regarding model usage, large language models (LLMs) were employed in zero-shot or few-shot prompting settings, without fine-tuning or adaptive feedback loops. As a result, the explanations generated may not reflect the full potential of LLM performance in more interactive or human-in-the-loop workflows. Furthermore, the analysis in Chapter 5 focused on static, single-turn LLM outputs, which do not reflect the iterative, dialogic nature of editorial decision-making. Given the rapid evolution of LLM capabilities, future work should continue to reassess their alignment with journalistic standards.

Although the correction strategy experiment in Chapter 4 measured engagement, credibility, and interpretation accuracy, it was limited to short-term effects. The durability of these corrections on memory, trust, or reading habits remains unknown. Longitudinal research would be valuable to examine the temporal dynamics of correction efficacy and audience adaptation. In addition, the study relied on self-reported measures and a behavioral proxy (forced-choice selection), which may not fully capture real-world engagement behaviors such as attention, sharing, or commenting.

The dissertation foregrounds editorial concerns such as framing bias, ambiguity, and contextual completeness, but these constructs were operationalized through specific annotation rubrics informed by journalism and XAI literature. While these operationalizations provide a practical foundation, they may not encompass the full nuance of editorial concerns in professional settings. Judgments involving ideological implications, epistemic stance, or audience sensitivity were not directly measured but likely influence how misleadingness is perceived and addressed in practice.

From a stakeholder perspective, the work primarily centers on journalists and news consumers, excluding other critical actors such as platform moderators, fact-checkers, policymakers, or underrepresented communities. These groups may hold distinct definitions of what makes a headline misleading or a correction effective. Integrating their perspectives in future work could inform the development of more inclusive, context-sensitive editorial AI systems.

Another consideration involves the nature of human evaluation in Chapter 5. While rubric-based expert assessments provided structured insight into LLM explanation quality, such evaluations inherently involve subjective interpretation. Variability across expert judgments despite shared professional backgrounds highlights the complexity of editorial reasoning. Future work could benefit from additional calibration procedures or triangulation with other data sources to increase consistency.

Moreover, although the dissertation identifies several design implications for human-AI editorial collaboration, these remain conceptual and untested in real-world deployment. Interface-level recommendations, such as ambiguity surfacing or editable reasoning, require prototyping and evaluation in operational newsroom settings. Without such validation, their practical feasibility and adoption remain uncertain.

An added limitation relates to audience effects, which refers to how individual differences among readers affect their interpretation of misleading headlines. While this dissertation accounts for general audience perceptions through survey-based and experimental methods, it does not systematically analyze how factors such as political orientation, cognitive style, media literacy, or prior beliefs modulate the perception of misleadingness. Prior research has shown that

audiences do not passively receive media messages but actively interpret them through the lens of their identities, experiences, and expectations [111, 398]. For instance, readers with strong partisan identities may perceive the same headline as misleading or trustworthy depending on its ideological alignment. This variability in reception limits the generalizability of any singular evaluation of headline misleadingness and highlights the need for future work that incorporates audience-centered designs, such as segmentation analyses, personalized feedback, or interactive correction systems that adapt to user traits.

Finally, the dissertation's analytical focus is limited to textual content, specifically, headlines and their accompanying explanations. In today's news ecosystem, visual cues such as images, video thumbnails, and layout design play a substantial role in shaping audience interpretation. Multimodal analysis would provide a more ecologically valid understanding of how misinformation operates and how corrections might be perceived across formats.

Together, these limitations highlight important directions for enhancing the scope, inclusivity, and practical relevance of future research at the intersection of misinformation detection, human-centered AI, and editorial decision-making. Despite these boundaries, the dissertation provides a robust empirical and conceptual foundation for advancing responsible media AI design and collaborative editorial transparency.

Chapter 7: Conclusion

This dissertation investigated the complex problem of misleading news headlines through a multi-stage, human-centered, and computational lens. Across three interlinked studies, it explored how misleadingness is perceived, produced, corrected, and explained by journalists, readers, and large language models. Each chapter addressed a critical dimension of this phenomenon: the subjective and institutional tensions around misleading headlines (Chapter 3), the effects of corrective strategies on audience response (Chapter 4), and the capabilities and limitations of AI-generated explanations in editorial tasks (Chapter 5).

At its core, this dissertation challenges the binary framing of misinformation detection and instead treats misleadingness as a situated, interpretive construct, shaped by competing stakeholder goals, editorial pressures, and linguistic nuance. The first study revealed how both journalists and readers struggle with headline integrity under the pressures of digital visibility, showing that institutional and audience expectations frequently diverge. The second study demonstrated that correcting misleading headlines can improve credibility and interpretation without sacrificing engagement, offering empirical support for intervention strategies that realign accuracy with attention. The final study critically evaluated the role of LLMs in supporting editorial judgment, showing that while models perform well under clear-cut scenarios, they struggle to surface ambiguity and nuanced context in contested cases.

Together, these findings advance a more socio-technical understanding of misinformation by emphasizing ambiguity, stakeholder conflict, and explanatory transparency. Methodologically, this work contributes a triangulated evaluation framework that bridges qualitative insights, experimental testing, and expert assessments. Practically, it proposes design principles for human-AI collaboration in journalism, calling for explanation systems that are not only linguistically fluent but also editorially meaningful, revisable, and context-aware.

Ultimately, this dissertation argues that ensuring headline integrity is not merely a matter of computational detection or behavioral correction, but a collaborative editorial challenge. To address this challenge, we need systems that support human reasoning rather than replace it, especially in the high-stakes and interpretively complex field of journalism. By foregrounding human perspectives and expert judgment, this work lays the groundwork for more transparent, accountable, and trustworthy information ecosystems.

Appendix A: Additional details of Chapter 3

Interview Protocol

Expert Participants

- Please specify your Name, occupation, gender, and education. Could you tell us about your news consumption habit?
- How much do you spend each day reading the news online?
- How many articles mainly do you read?
- How often do you read headlines online without reading the content?
- How do you decide which article to read when you come across them online?
- When you come across a news headline that piques your interest, but you're unsure about its meaning or details, how do you determine if the content is worth reading?
- Could you define, according to you, what makes a news headline misleading?
- Can you recall an event where you bumped into a news headline that you found misleading?
- What made you recognize it as such?
- What factors do you consider when deciding whether a headline might be potentially misleading?
- How often do you come across misleading content when researching or browsing through articles?
- Is there a topic in particular that you see frequently have misleading headline?
- What are your reactions after finding that the headline you read was actually misleading?
- Are there certain topics you are more likely to be frustrated with when you see a misleading headline?
- Could you tell us a little about your journalism and news media background?
- How many years of experience do you have?
- Is there a subfield of journalism that you have focused on the most?

- Can you describe your experience and role in journalism, particularly in relation to headline creation?
- Are there any standard practices you have seen in your experience that have led to creating misleading headlines?
- If yes, could you give any examples you can think of?
- (If nothing is mentioned) What about Search Engine Optimization (SEO)? Does your organization consider it when crafting headlines? How does it impact the quality of the headlines?
- Who is responsible for writing headlines in the newsroom in your experience?
- How do you approach the process of crafting headlines for news stories?
- Are there any internal guidelines or checks in place at your organization to prevent misleading headlines?
- Can you describe any specific instances where you had to modify or retract a headline as it seems misleading?
- What are your thoughts on the balance between creating engaging headlines and ensuring they accurately represent the news content?
- How do you balance the need for attention-grabbing headlines with the need for accuracy and credibility?

At this point, participants will be shown FOUR news articles. They will be asked to read out the headlines loud first. Then, they will describe their initial thoughts about the headline and news event (e.g., how they are interpreting the headline, what their expectations are from the content, etc.) Next, they will read the content, and another set of questions will be asked. This process will be done iteratively for the four articles.

After reading the headline

- Did you have any prior knowledge of the topic described in the article?
- After reading the headline, What were your initial impressions/interpretation of the news event?
- Did you assume any information about the news event that is not mentioned while reading the headline?

After reading the news content

- Did you think that the content matched the headline?
- Did you think the headline made you misinterpret the content?
- If you had not read the content, would the headline have misled or misinformed you?

- (If Found Misleading) Could you please elaborate on why you find it misleading?
- (If Found Misleading) In your opinion, what specific elements or characteristics of the misleading headline led you to expect something different from the content?
- (If Found Misleading) Would you click on such an article while browsing online?
- If yes, what would your reaction be knowing that it is misleading?
- (If Found Misleading) Please describe a headline correction or edit that you believe would have improved the accuracy or clarity of the misleading headline you mentioned earlier.
- What do you consider the most important information that should be conveyed in this news headline?
- How would you expect the correction or edit to impact your decision to read the full news article?
- (If Found NOT Misleading) Do you think this headline was an example of a good headline? Why or why not?

At the end of the interview

- What improvements would you suggest to reduce the occurrence of misleading headlines from a journalistic point of view?
- Do you think any automated tools or focused training in identifying and correcting misleading headlines would be beneficial for journalists?
- Is there anything else you would like to share about misleading headlines that you find important?

Table A.1: News Readers Demographics

ID	Age	Gender	Education Level	# of Articles Read/Day
P1	18-23	Male	High School	1-5
P2	30-35	Male	Undergrad	1-5
P3	18-23	Female	High School	1-5
P4	24-29	Male	Graduate	6-11
P5	24-29	Female	Undergrad	1-5
P6	60-65	Male	Graduate	6 -11
P7	30-35	Male	High School	12-17
P8	24-29	Male	Undergrad	6-11
P9	18-23	Female	High School	1-5
P10	24-29	Male	Undergrad	6-11
P11	18-23	Male	High School	1-5
P12	18-23	Female	High School	12-17

Table A.2: Journalist Demographics

ID	Age	Gender	Occupation
J1	18-23	Female	Student with Internship Experience
J2	18-23	Female	Journalism Intern/Student
J3	24-29	Female	Journalist
J4	60-65	Male	Senior Editor
J5	24-29	Female	Journalist
J6	18-23	Female	Student with Part-time Journalist
J7	24-29	Male	Public Radio Producer
J8	24-29	Female	Journalist
J9	18-23	Male	Student with Internship Experience
J10	70-75	Male	Retired Journalist
J11	35-40	Female	Journalist
J12	35-40	Female	Journalist

Table A.3: Distribution of collected articles by sources

Source	Source Type	# of articles explored	# of Articles with Potential Misleading Headlines
ABC News	Reliable	11	3
Washington Post	Reliable	15	0
USA Today	Reliable	4	0
NPR	Reliable	7	0
New York Times	Reliable	14	4
Wall Street Journal	Reliable	15	0
Veterans Today	Unreliable	2	2
Infowars	Unreliable	15	3
Lifefzette	Unreliable	15	7
Health.News	Unreliable	6	6
Before It's News	Unreliable	10	4
Alternet	Unreliable	1	1

Table A.4: Distribution of news articles by topics

Topic	Potentially misleading	Potentially non-misleading
Health	19	15
Technology	7	11
Business	4	4

Table A.5: Categories for misleading headlines developed through thematic analysis from articles collected between the authors

Category Name	Examples	Number of Articles
Cherry-Picking	Study reveals insufficient sleep linked to increase in unhealthy abdominal fat	3
Exaggeration	COVID-19 infections linked to brain damage, studies find	8
Omitting Critical Context	Israeli Study: 2nd Vaccine Booster Significantly Lowers COVID Death Rate	14
False Claim	Biden reverses Mask Policies Due To Politics	3
Impersonation	Wall Street analyst says covid “vaccines” are the “greatest fraud in history”	1
Lack of Supporting Evidence	Doctors warn many patients in hospitals have COVID jab-induced AIDS	26
Hiding Uncertainty	Sweeteners Linked to Increased Cancer Risk – New Research	5
Ambiguous	Lawsuit: Amazon’s unrealistic demands caused driver to crash	5

Appendix A: Additional details of Chapter 4

Online Experiment Questions (Based on Sample Headline)

Sample Headline: *Scientists believe they’ve detected mysterious radio signal from alien planet*

Engagement Measures

Evaluate the following statements based on the headline above:

- This headline makes me want to read the full news article.
- This headline makes me want to share the article with others.
- This headline makes me ignore the article after reading just the headline.

Credibility Measures

Evaluate the following statements based on the headline above:

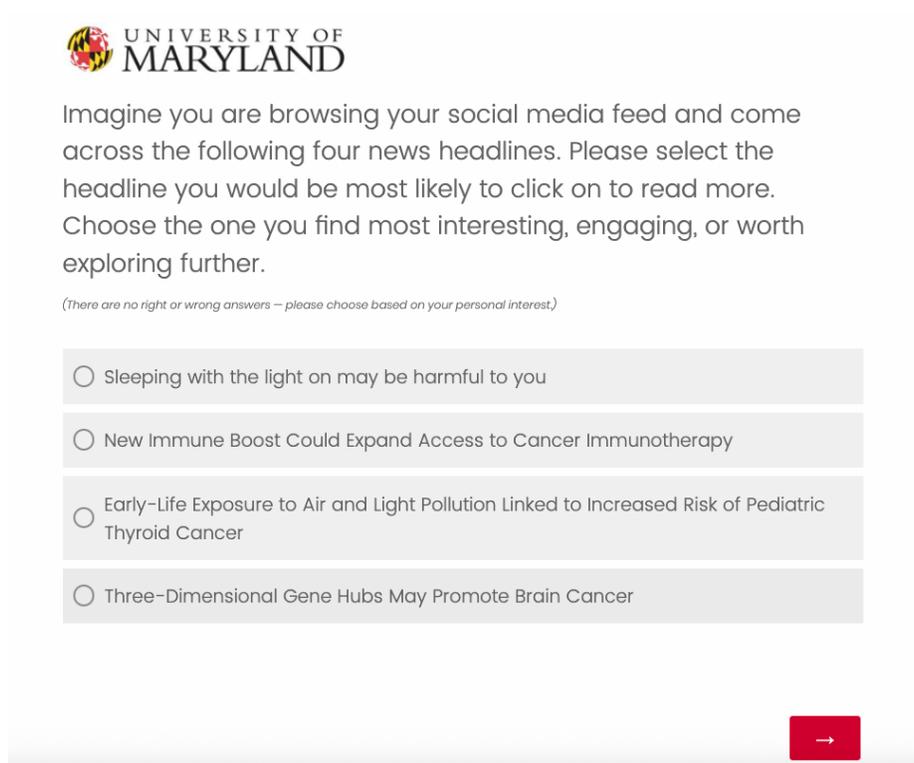
- This headline appears factually correct and reasonable based on the way it is written.
- This headline feels authentic and transparent for the event it describes.
- This headline seems believable to me based on how it is written.
- Regardless of the headline’s message, please select ‘Somewhat disagree’ for this question.

Interpretation Measures

Answer the following questions based on the headline above:

- Summarize the main claim of this headline in no more than 200 characters.
- Did the headline suggest that the discovery has been confirmed?
- According to the headline, what have scientists detected?
- According to the headline, which of the following best describes the scientific status of the discovery?

Online Experiment Interface Sample



 UNIVERSITY OF MARYLAND

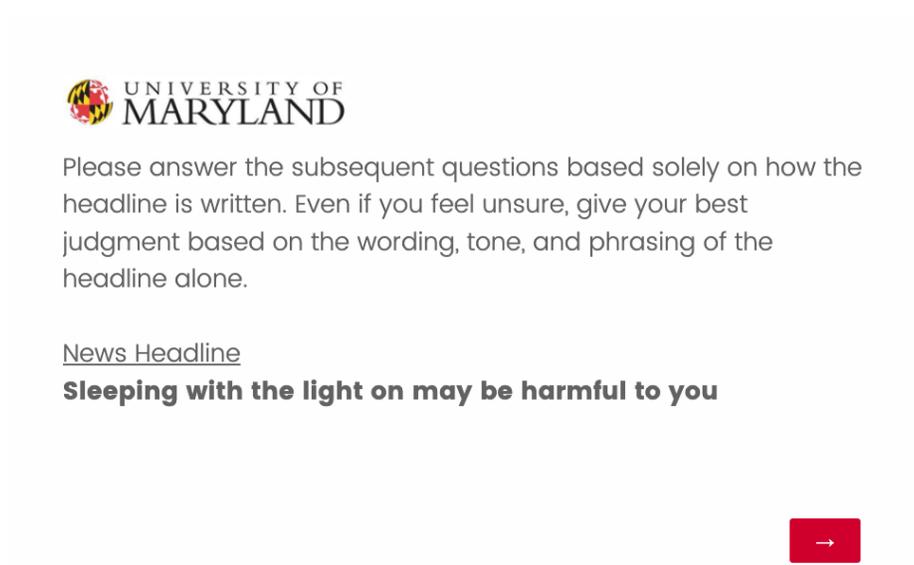
Imagine you are browsing your social media feed and come across the following four news headlines. Please select the headline you would be most likely to click on to read more. Choose the one you find most interesting, engaging, or worth exploring further.

(There are no right or wrong answers – please choose based on your personal interest.)

- Sleeping with the light on may be harmful to you
- New Immune Boost Could Expand Access to Cancer Immunotherapy
- Early-Life Exposure to Air and Light Pollution Linked to Increased Risk of Pediatric Thyroid Cancer
- Three-Dimensional Gene Hubs May Promote Brain Cancer



Figure A.1: Behavioral Engagement: Headline Selection at the beginning of the experiment



 UNIVERSITY OF MARYLAND

Please answer the subsequent questions based solely on how the headline is written. Even if you feel unsure, give your best judgment based on the wording, tone, and phrasing of the headline alone.

News Headline
Sleeping with the light on may be harmful to you



Figure A.2: Instruction screen if participant selects the targeted headline



Even if you didn't select this headline earlier, imagine it was the only one you saw. Please answer the subsequent questions based only on how the headline is written. If you're unsure, give your best judgment based on its wording, tone, and phrasing of the headline alone.

News Headline

Early Morning Exercise Could Actually Be Counterproductive



Figure A.3: Instruction screen if participant does NOT select the targeted headline

News Headline

**Early Morning Exercise Could Actually Be
Counterproductive**

Evaluate the following statements based on the headline above.

This headline makes me want to share the article with others.

- Strongly disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree
- Strongly agree

This headline makes me ignore the article after reading just the headline.

- Strongly disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree

Figure A.4: Screen of the Engagement measurement (Partial)



News Headline

Sleeping with the light on may be harmful to you

Evaluate the following statements based on the headline above.

This headline appears factually correct and reasonable based on the way it is written.

- Strongly disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree
- Strongly agree

This headline seems believable to me based on how it is written.

- Strongly disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree
- Strongly agree

Figure A.5: Screen of the Perceived Credibility measurement (Partial)

News Headline

Sleeping with the light on may be harmful to you

Evaluate the following statements based on the headline above.

This headline makes me want to share the article with others.

- Strongly disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree
- Strongly agree

This headline makes me ignore the article after reading just the headline.

- Strongly disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree
- Strongly agree

This headline makes me want to read the full news article.

- Strongly disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree
- Strongly agree

Figure A.6: Screen of the Interpretation measurement (Partial)

Study Materials

Table A.1: Headline Pairs and Correction Strategies

Correction Strategy	Misleading Version	Corrected Version
Providing Context	Sleeping with the light on may be harmful to you	Sleeping with moderate light on may be harmful to you
Providing Context	Hackers hit popular video game, stealing more than \$600 million in cryptocurrency	Hackers hit popular video game's blockchain, stealing more than \$600 million in cryptocurrency
Providing Context	Early Morning Exercise Could Actually Be Counterproductive	Cutting Sleep to Early Morning Exercise Could Actually Be Counterproductive
Removing Emotion-Evoking Words	Bloody secret revealed: Blood of unborn cows used to create lab-grown meat	Blood of unborn cows used to create lab-grown meat
Removing Emotion-Evoking Words	WARNING: Gene-edited meat products coming to supermarkets near you	Gene-edited meat products coming to supermarkets near you
Removing Emotion-Evoking Words	Big Pharma Now Coming For Your Babies: Science Experiment On Toddlers And Infants Begins	COVID Vaccine Experiment On Toddlers And Infants Begins
Adding Uncertainty Cues	Scientists believe they've detected mysterious radio signal from alien planet	Scientists believe they've detected mysterious radio signal from alien planet, pending confirmation
Adding Uncertainty Cues	Sweeteners Linked to Increased Cancer Risk – New Research	Sweeteners may be Linked to Increased Cancer Risk – New Research
Adding Uncertainty Cues	Brain Implant Allows Fully Paralyzed Patient to Communicate	Brain Implant Allows Fully Paralyzed Patient to Communicate — A possible Breakthrough

Appendix A: Additional details of Chapter 5

Expert Study

Instructions (shown to participants)

You will be shown a news headline along with its associated article excerpt.

Key sentences in the article are highlighted to help you quickly identify the most relevant information.

You will then see two AI-generated explanations, one at a time, each providing a reasoning for whether the headline might be misleading or not.

After reading each explanation, you will rate it across several dimensions.

Finally, you will compare the two explanations and provide your preference and comments.

Headline Example

Florida’s top health official says healthy children should not get coronavirus vaccine

Article Content with Highlighted Sentences

The policy pronouncement by the state’s surgeon general was condemned as ‘reckless’ and ‘dangerous’ by medical experts.

By Lori Rozsa WEST PALM BEACH, Fla.

— Florida’s governor and chief health official announced a new state policy Monday that will recommend against giving a coronavirus vaccine to healthy children, regardless of their age — a policy that flies in the face of recommendations by every medical group in the nation.

The announcement came at the conclusion of a 90-minute forum that Gov. Ron DeSantis (R) hosted in West Palm Beach. “The Curtain Close on COVID Theater” was live-streamed from a studio with hundreds of participants appearing on a towering screen behind the panelists.

Florida Surgeon General Joseph Ladapo prefaced the change by deriding the school closures and mask and vaccine mandates issued by many states since the start of the pandemic as “terrible, harmful policies.” “Florida is going to be the first state to officially recommend against the covid-19 vaccination for healthy children,” Ladapo said.

National data from the Centers for Disease Control and Prevention show that nearly 1,600 youths from birth to age 17 have died of covid-19. More than 40 of those deaths were in Florida.

The CDC recommends coronavirus vaccination for children 5 and older, saying the shots are safe and effective against serious covid-19 disease. More than 25 million children in the United States have received at least one dose.

Infectious-disease experts quickly denounced Florida’s shift as “reckless” and “dangerous.”

“It’s terrible,” said Peter Hotez, a professor of pediatrics and molecular virology and microbiology at Baylor College of Medicine and a leading expert on the virus. “To be at such distinct variance from the hundreds of physicians and scientists at the CDC and the FDA is reckless at best and dangerous at worst.”

DeSantis, who is running for reelection and likely weighing a bid for the 2024 GOP nomination for president, said there is a lack of trust in those agencies. He repeated his frequent criticism of Anthony S. Fauci, chief medical adviser to President Biden and director of the National Institute of Allergy and Infectious Diseases. There is “a gap between the Faucian pronouncement of what needed to be done and the actual data that supported or did not support those policies,” the governor said.

One of the panelists was Joseph Fraiman, a New Orleans emergency-room doctor who claims that no healthy children have died of covid-19 — a false assertion. “The majority of studies

haven't been able to find a single healthy child who died from covid," he repeated Monday. "If you have a healthy child, the chances of that child dying are incredibly low, essentially close to zero, if not zero."

Yet Hotez at the Baylor College of Medicine said studies supporting the efficacy of vaccines for children are readily available. "You don't have to look very hard to get the actual data," he said in an interview from Houston. "But you do have to work hard to cherry-pick it in order to give misleading information."

During his confirmation hearing in January, Ladapo, the Florida surgeon general, was repeatedly asked whether he believed vaccines were effective against the coronavirus. He never gave a direct answer. "I would say that the question is a scientific one, and it's one that is answered with data," Ladapo said. "The question is informed by data on specific outcomes and specific therapies. So that's the scientific question." DeSantis has repeatedly faulted the Biden administration's efforts to end the pandemic, with mask requirements among his top targets. On Wednesday, he scolded a group of high school students for wearing masks at an event at the University of South Florida. "You do not have to wear those masks. I mean, please take them off," he told the teens. Though more than 70,000 people in Florida have died of covid-19, DeSantis said Monday that the state has handled things correctly by avoiding mask and vaccine mandates. "We always sided with the data and rejected the narrative," he said. "And our state is kind of like the center of the world. Everyone wants to come here." The governor sponsored the roundtable as lawmakers in Tallahassee are grappling with some of the most contentious issues in the legislative session, which is scheduled to end Friday. In recent weeks, DeSantis has been holding sometimes daily news conferences, traveling the state and presenting checks to local communities.

Though some doctors on Monday's panel cited studies showing vaccine immunity wanes in children as it does in adults, Hotez countered that the shots still "give high levels of protection against hospitalization" that will be increased when boosters are approved for children.

"The omicron wave is dying down, but it's unlikely to be our last one," Hotez said. "And we've seen terrible epidemics across the Southern states, including Florida and Texas, in the summer of 2020 and the summer of 2021, and we have to anticipate that's going to be true of this summer as well. So now is the time you want to vaccinate."

LLM Explanations

LLM A

The headline "Florida's top health official says healthy children should not get coronavirus vaccine" is misleading without the context of the article. In the article, it is revealed that the announcement by Florida's governor and chief health official contradicts recommendations from medical experts and organizations. The article presents a more nuanced view of the situation, as it includes criticism from infectious-disease experts who call the policy "reckless" and "dangerous."

LLM B

The headline accurately represents the content of the article, which outlines the stance of Florida's chief health official recommending against vaccination for "healthy" children. The

article does present contrasting views and highlights criticism from health experts who consider this recommendation reckless and dangerous, but the headline is factual based on the official's statement.

Evaluation Questions

Explanation Evaluation

- For each of the following statements (Table [A.1](#)), please rate the explanations on a 4-point scale: (1 = Not at all, 4 = Completely)
- Please briefly reflect on the explanation's weaknesses, missing points, possible improvements, and handling of ambiguity. (Table [A.2](#))

Expert Evaluation Interface Sample



Please read the news headline and article carefully. Key sentences are highlighted to help you quickly find the most critical information related to the headline's claim and the LLM explanations. You may also review the full article if you need more context before evaluating the explanations.

Headline: Arizona Offers Driver's Licenses on iPhones. Other States Want to Be Next.

Content:

Digital copies of licenses and state-issued IDs, which can also be stored on Apple Watches, will only work at security checkpoints at a Phoenix airport for now, officials said.

It started as a digital catchall for credit cards and concert tickets, enabling anyone with an iPhone to be waved through checkout lines and turnstiles.

The technology then expanded to vaccine passport records during the pandemic. And this week, the Apple Wallet, an app for iPhones and Apple Watches that stores payment information and QR codes, added driver's licenses for the first time.

On Wednesday, Arizona became the first state to offer digital copies of driver's licenses and state identification cards as part of a sweeping partnership with Apple that was announced last year.

Figure A.1: Expert Evaluation of LLM Explanation: Headline - Content View



Please read the explanation carefully. It represents the AI model's reasoning about whether and how the headline may be misleading, based on the article content. After reading, you will be asked to evaluate the explanation across several dimensions.

Headline

Arizona Offers Driver's Licenses on iPhones. Other States Want to Be Next.

Content

[Use right-click to open on a separate tab](#)

Explanation

The headline suggests that all drivers in Arizona can now use digital copies of their licenses on their iPhones and that other states are eager to adopt the same system. However, the article reveals that at this point, the digital licenses are only being accepted at specific airport security checkpoints in Phoenix, and not for identification purposes at traffic stops or other routine checks. So, the reach and usability of these digital licenses are currently quite limited.

Figure A.2: Expert Evaluation of LLM Explanation: Explanation View

Please rate the explanation based on the following statements and questions.

The explanation is clear, easy to read, and easy to understand.

(That is, the explanation is logically structured, free of confusing language, and understandable without requiring external knowledge)

Strongly disagree

Somewhat disagree

Somewhat agree

Strongly agree

The explanation correctly interprets the relationship between the headline and the article content.

(That is, the explanation accurately reflects whether and how the headline is misleading, exaggerated, or supported by the article)

Strongly disagree

Figure A.3: Expert Evaluation of LLM Explanation: Likert Scale Rating (Sample 1)

The explanation thoroughly covers all major aspects relevant to whether the headline is misleading.

(That is, the explanation addresses all important factors – such as omission, exaggeration, emotional framing – that affect the headline’s interpretation)

Strongly disagree

Somewhat disagree

Somewhat agree

Strongly agree

I trust the explanation as a basis for editorial or fact-checking judgment.

(That is, based on this explanation alone, I would feel confident making an editorial decision about whether the headline is misleading, needs correction, or can be published as is)

Strongly disagree

Somewhat disagree

Figure A.4: Expert Evaluation of LLM Explanation: Likert Scale Rating (Sample 2)

What are the major weaknesses or limitations you observed in this explanation?

What important details or context were missing that should have been included?

How could this explanation be improved to better support editorial or fact- checking decisions?

Figure A.5: Expert Evaluation of LLM Explanation: Open-ended QA

Study Materials

Table A.3: Final Selected Headlines for Expert Evaluation Study. Each row presents one of the 15 headline-article pairs selected for expert analysis, reflecting different combinations of human annotator agreement and LLM classification agreement.

#	Human Agreement	LLM Agreement	Headline
1	3/3	2/2	Lawsuit: Amazon’s unrealistic demands caused driver to crash
2	3/3	2/2	Follow the money: Pfizer paid FDA millions of dollars for COVID jab approval
3	3/3	1/2	Pandemic deaths among pregnant Americans are now ‘surging’
4	3/3	0/2	Google, facing antitrust scrutiny, says it will let apps like Spotify offer alternate billing systems
5	0/3	1/2	The COVID pill is a game-changer, but for some it might be a danger: Experts
6	0/3	1/2	Arizona joins a growing list of states that have passed a 15-week abortion ban
7	0/3	0/2	COVID-19 infection increases risk of developing diabetes, study finds
8	2/3	2/2	CNN ‘Doctor’: “The Science Has Changed” So Lift COVID Restrictions
9	2/3	2/2	Wall Street analyst says covid “vaccines” are the “greatest fraud in history”
10	2/3	1/2	Shortening Natural Immunity Status After COVID Infection Deemed ‘Illegal’
11	2/3	0/2	Israeli Study: 2nd Vaccine Booster Significantly Lowers COVID Death Rate
12	1/3	2/2	COVID-19 Will Be With Us Forever. I Will NOT Take Vaccine Boosters Forever. I’m Done.
13	1/3	1/2	Big Pharma Now Coming For Your Babies: Science Experiment On Toddlers And Infants Begins
14	1/3	1/2	Florida’s top health official says healthy children should not get coronavirus vaccine
15	1/3	0/2	Brain Implant Allows Fully Paralyzed Patient to Communicate

Study Materials

Codebook for Evaluating Explanations

- Framing Bias: Does the explanation identify and address problematic framing in the headline (e.g., ideological slant, misleading emphasis, oversimplification)?
- Omission of Key Context: Does the explanation highlight the missing context in the headline or article (e.g., study limitations, demographic qualifiers, lack of attribution)?
- Emotional Language: Does the explanation identify the use of emotionally charged or sensational language (e.g., fear-inducing, alarmist, or exaggerated phrasing)?
- Actionability / Editorial Utility: Would this explanation be useful to a journalist deciding whether to flag, revise, or accept the headline?

Table A.1: Expert Evaluation Questions and Supplementary Clarifications

Question	Text and Clarification
Q1: Correctness	<p>The explanation correctly interprets the relationship between the headline and the article content.</p> <p><i>(That is, does the explanation accurately reflect whether and how the headline is misleading, exaggerated, or supported by the article?)</i></p>
Q2: Clarity	<p>The explanation is clear, easy to read, and easy to understand.</p> <p><i>(That is, is the explanation logically structured, free of confusing language, and understandable without requiring external knowledge?)</i></p>
Q3: Exhaustiveness	<p>The explanation thoroughly covers all major aspects relevant to whether the headline is misleading.</p> <p><i>(That is, does the explanation address all important factors, such as omission, exaggeration, emotional framing, that affect the headline’s interpretation?)</i></p>
Q4: Ambiguity Awareness	<p>The explanation appropriately acknowledges ambiguity, uncertainty, or complexity related to the headline and article content.</p> <p><i>(That is, does the explanation highlight if there are nuances, limitations, conflicting evidence, or areas where interpretation is not straightforward?)</i></p>
Q5: Risk of Harm	<p>The explanation could itself cause misunderstanding or mislead readers. <i>*(Reverse scored)*</i></p> <p><i>(That is, does the explanation risk oversimplifying, overstating certainty, or otherwise misleading if used for editorial judgment?)</i></p>
Q6: Trustworthiness	<p>I trust the explanation as a basis for editorial or fact-checking judgment.</p> <p><i>(That is, based on this explanation alone, I would feel confident making an editorial decision about whether the headline is misleading, needs correction, or can be published as is.)</i></p>
Q7: Actionability	<p>The explanation would be actionable for editorial decision-making (e.g., deciding to flag or revise the headline).</p> <p><i>(That is, the explanation provides clear and sufficient reasoning that could directly support taking editorial action, such as flagging the headline for revision, issuing a correction, or approving it.)</i></p>

Table A.2: Structured Open-Ended Questions for Expert Evaluation

Question	Prompt and Clarification
1. Weaknesses	What are the major weaknesses or limitations you observed in this explanation?
2. Missing Information	What important details or context were missing that should have been included?
3. Suggestions for Improvement	How could this explanation be improved to better support editorial or fact-checking decisions?
4. Handling of Ambiguity	Did the explanation appropriately acknowledge uncertainty, complexity, or missing context in the article and headline? <i>“Ambiguity” refers to uncertainty, incomplete evidence, or multiple possible interpretations of the headline’s claim based on the article content.</i>

Table A.4: Likert Scale for Coding Journalistic Concerns in Explanations

Score	Rating Description
0 – Not at all	The explanation fails to address the concern or incorrectly claims the concern is present when it is not. For example, it overlooks key omissions or wrongly suggests emotional framing in a neutral headline.
1 – Partially	The explanation somewhat engages with the concern, but does so superficially, inaccurately, or without clear justification. It may briefly mention the concern without evaluating its relevance or without supporting reasoning.
2 – Clearly	The explanation clearly identifies and explains the concern (if present), <i>or</i> accurately recognizes and justifies its <i>absence</i> . For instance, if no emotional language is present, the explanation earns a 2 if it notes that and provides a clear rationale (e.g., “The headline presents the author’s stance without exaggeration or emotive language.”).

Table A.5: Spearman correlation coefficients (ρ) and associated p -values between explanation dimensions and expert ratings of trustworthiness and actionability.

Dimension	Trustworthiness	Actionability
Correctness	$\rho = 0.92, p < .0001$	$\rho = 0.83, p < .0001$
Clarity	$\rho = 0.53, p = .0028$	$\rho = 0.60, p = .0004$
Exhaustiveness	$\rho = 0.80, p < .0001$	$\rho = 0.79, p < .0001$
Ambiguity Awareness	$\rho = 0.71, p < .0001$	$\rho = 0.77, p < .0001$
Risk of Harm	$\rho = -0.85, p < .0001$	$\rho = -0.89, p < .0001$

Table A.6: Multiple linear regression results predicting expert ratings of trustworthiness and actionability from editorial dimensions. Statistically significant predictors ($p < .05$) are bolded.

Predictor	Trustworthiness		Actionability	
	Coef. (β)	p -value	Coef. (β)	p -value
Intercept	-0.01	0.994	3.22	0.021
Correctness	0.83	.001	0.19	0.293
Clarity	0.01	0.970	0.01	0.938
Exhaustiveness	-0.14	0.462	-0.23	0.217
Ambiguity Awareness	0.21	0.074 [†]	0.30	0.012
Risk of Harm	-0.04	0.877	-0.56	0.019
R^2	0.912		0.893	
Adj. R^2	0.894		0.870	
F-statistic	49.68	$p < .001$	39.86	$p < .001$

[†] $p < .10$, **bold** values indicate statistical significance at $p < .05$.

Table A.7: Mean and SD of Expert Ratings across Seven Explanation Quality Dimensions, by LLM and Agreement Category

Dimension	LLM	Agreement	Mean	SD	n
Correctness	Gemini	Unanimous	2.73	0.88	8
Correctness	Gemini	Ambiguous	3.14	0.56	7
Correctness	GPT	Unanimous	3.33	0.31	8
Correctness	GPT	Ambiguous	3.33	0.53	7
Clarity	Gemini	Unanimous	3.29	0.44	8
Clarity	Gemini	Ambiguous	3.45	0.36	7
Clarity	GPT	Unanimous	3.69	0.14	8
Clarity	GPT	Ambiguous	3.40	0.52	7
Exhaustiveness	Gemini	Unanimous	2.48	0.83	8
Exhaustiveness	Gemini	Ambiguous	2.98	0.33	7
Exhaustiveness	GPT	Unanimous	2.92	0.53	8
Exhaustiveness	GPT	Ambiguous	2.95	0.47	7
Ambiguity awareness	Gemini	Unanimous	2.48	0.83	8
Ambiguity awareness	Gemini	Ambiguous	3.00	0.37	7
Ambiguity awareness	GPT	Unanimous	2.62	0.85	8
Ambiguity awareness	GPT	Ambiguous	2.95	0.52	7
Risk of harm	Gemini	Unanimous	2.54	0.79	8
Risk of harm	Gemini	Ambiguous	2.17	0.52	7
Risk of harm	GPT	Unanimous	1.90	0.44	8
Risk of harm	GPT	Ambiguous	2.07	0.58	7
Trustworthiness	Gemini	Unanimous	2.38	0.80	8
Trustworthiness	Gemini	Ambiguous	2.76	0.50	7
Trustworthiness	GPT	Unanimous	2.85	0.42	8
Trustworthiness	GPT	Ambiguous	2.93	0.57	7
Actionability	Gemini	Unanimous	2.52	0.63	8
Actionability	Gemini	Ambiguous	2.76	0.48	7
Actionability	GPT	Unanimous	3.06	0.47	8
Actionability	GPT	Ambiguous	2.98	0.47	7

Table A.8: Mann-Whitney U Test Results Comparing Unanimous vs. Ambiguous Cases by Dimension and LLM

Dimension	LLM	U	p-value
Correctness	Gemini	21.00	0.4491
Correctness	GPT	24.00	0.6800
Clarity	Gemini	22.00	0.5203
Clarity	GPT	35.00	0.4328
Exhaustiveness	Gemini	20.00	0.3807
Exhaustiveness	GPT	24.50	0.7244
Ambiguity awareness	Gemini	18.00	0.2655
Ambiguity awareness	GPT	22.00	0.5199
Risk of harm	Gemini	33.50	0.5607
Risk of harm	GPT	23.50	0.6416
Trustworthiness	Gemini	21.00	0.4495
Trustworthiness	GPT	25.00	0.7701
Actionability	Gemini	23.00	0.5992
Actionability	GPT	32.50	0.6395

Table A.9: Descriptive Statistics for Expert Ratings Across Seven Evaluation Dimensions

Statistic	Correctness	Clarity	Exhaustiveness	Ambiguity	Risk of Harm	Trustworthiness	Actionability
Count	30	30	30	30	30	30	30
Mean	3.13	3.46	2.82	2.75	2.17	2.72	2.83
Std Dev	0.63	0.40	0.59	0.69	0.62	0.60	0.54
Min	1.00	2.50	1.33	1.00	1.33	1.00	1.50
25th Pctl	2.71	3.33	2.67	2.50	1.71	2.38	2.67
Median	3.33	3.58	2.92	3.00	2.00	2.83	2.83
75th Pctl	3.50	3.79	3.17	3.17	2.46	3.17	3.17
Max	3.83	3.83	3.67	3.83	3.83	3.50	3.67

Table A.10: Intraclass Correlation Coefficient (ICC(2,1)) for Expert Agreement by Dimension

Dimension	ICC(2,1)
Correctness	0.465
Clarity	0.384
Exhaustiveness	0.418
Ambiguity Awareness	0.532
Risk of Harm	0.423
Trustworthiness	0.399
Actionability	0.341

Bibliography

- [1] The psychology of misinformation. *Australasian science*, 36(2):21–, 2015.
- [2] Responses to misinformation, 2022.
- [3] Siti Aishah Chu Abdullah and Azza Azreenas Azman. Analytical clickbait strategies in news headlines: A case study of mstar.
- [4] Mohammed O Abu-Rahme, Nabil A Abu-Loghod, Fandi Omeish, Sager Alharthi, Kinana J Joudeh, and Jamal M Joudeh. Investigating the impact of misleading information via social media platforms on the trust and image of beauty and skincare companies, as perceived by customers. *Journal of Posthumanism*, 5(1):569–584, 2025.
- [5] Mark S Ackerman. The intellectual challenge of cscw: the gap between social requirements and technical feasibility. *Human–Computer Interaction*, 15(2-3):179–203, 2000.
- [6] Mark S Ackerman, Juri Dachtera, Volkmar Pipek, and Volker Wulf. Sharing knowledge and expertise: The cscw view of knowledge management. *Computer Supported Cooperative Work (CSCW)*, 22:531–573, 2013.
- [7] Zhila Aghajari, Eric PS Baumer, and Dominic DiFranzo. Reviewing interventions to address misinformation: The need to expand our vision beyond an individualistic focus. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–34, 2023.
- [8] Leah Hope Ajmani, Jasmine C Foriest, Jordan Taylor, Kyle Pittman, Sarah Gilbert, and Michael Ann Devito. Whose knowledge is valued? epistemic injustice in cscw applications. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–28, 2024.
- [9] Firoj Alam, Julia Maria Struß, Tanmoy Chakraborty, Stefan Dietze, Salim Hafid, Katerina Korre, Arianna Muti, Preslav Nakov, Federico Ruggeri, Sebastian Schellhammer, et al. The clef-2025 checkthat! lab: Subjectivity, fact-checking, claim normalization, and retrieval. In *European Conference on Information Retrieval*, pages 467–478. Springer, 2025.

- [10] Firoj Alam, Julia Maria Struß, Tanmoy Chakraborty, Stefan Dietze, Salim Hafid, Katerina Korre, Arianna Muti, Preslav Nakov, Federico Ruggeri, Sebastian Schellhammer, Vinay Setty, Megha Sundriyal, Konstantin Todorov, and Venktesh V. The clef-2025 checkthat! lab: Subjectivity, fact-checking, claim normalization, and retrieval. In Claudia Hauff, Craig Macdonald, Dietmar Jannach, Gabriella Kazai, Franco Maria Nardini, Fabio Pinelli, Fabrizio Silvestri, and Nicola Tonellotto, editors, *Advances in Information Retrieval*, pages 467–478, Cham, 2025. Springer Nature Switzerland.
- [11] Lamia Alam and Shane Mueller. Examining the effect of explanation on satisfaction and trust in ai diagnostic systems. *BMC medical informatics and decision making*, 21(1):178, 2021.
- [12] Kiana Alikhademi, Brianna Richardson, Emma Drobina, and Juan E Gilbert. Can explainable ai explain unfairness? a framework for evaluating explainable ai. *arXiv preprint arXiv:2106.07483*, 2021.
- [13] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236, 2017.
- [14] Ben Allen. The promise of explainable ai in digital health for precision medicine: a systematic review. *Journal of personalized medicine*, 14(3):277, 2024.
- [15] J. Allen, A. A. Arechar, G. Pennycook, and D. G. Rand. Scaling up fact-checking using the wisdom of crowds. *Science Advances*, 7, 2021.
- [16] Jennifer Allen, Duncan J Watts, and David G Rand. Quantifying the impact of misinformation and vaccine-skeptical content on facebook. *Science*, 384(6699):eadk3451, 2024.
- [17] Sacha Altay. How effective are interventions against misinformation? 2022.
- [18] Bhasake Ambadas. Emerging trends in digital journalism: A comprehensive review. *International Journal for Multidisciplinary Research*, 6(2), 2024.
- [19] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13, 2019.
- [20] Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. Make your llm fully utilize the context. *Advances in Neural Information Processing Systems*, 37:62160–62188, 2024.
- [21] Alyssa Appelman and Kirstie E Hettinga. The ethics of transparency: A review of corrections language in international journalistic codes of ethics. *Journal of Media Ethics*, 36(2):97–110, 2021.

- [22] Alyssa Appelman and S Shyam Sundar. Measuring message credibility: Construction and validation of an exclusive scale. *Journalism & Mass Communication Quarterly*, 93(1):59–79, 2016.
- [23] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012*, 2019.
- [24] Solomon E Asch. Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, 41(3):258, 1946.
- [25] Kevin Aslett, Zeve Sanderson, William Godel, Nathaniel Persily, Jonathan Nagler, and Joshua A Tucker. Online searches to evaluate misinformation can increase its perceived veracity. *Nature*, 625(7995):548–556, 2024.
- [26] Nurul Afiqah Aszeli, Dhiya Alia Jamil, and Nurul Hidayah Rahmat. A study of interactional metadiscourse on news article on the impact of covid-19 on education. *European Journal of Literature, Language and Linguistics Studies*, 4(4), 2021.
- [27] Rana Najji Aziz. Pragma-rhetorical analysis to covid-19 headlines. *Social Sciences & Humanities Open*, 6(1):100340, 2022.
- [28] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018.
- [29] J. B. Bak-Coleman. Limiting factors in the effectiveness of crowd-sourced labeling for combating misinformation. 2023.
- [30] Vian Bakir and Andrew McStay. Fake news and the economy of emotions: Problems, causes, solutions. *Digital journalism*, 6(2):154–175, 2018.
- [31] Akshina Banerjee and Oleg Urminsky. The language that drives engagement: A systematic large-scale analysis of headline experiments. *Available at SSRN 3770366*, 2022.
- [32] Akshina Banerjee and Oleg Urminsky. The language that drives engagement: A systematic large-scale analysis of headline experiments. *Marketing Science*, 2024.
- [33] Marzena Barańska. The subjective dimension of fake news. *Studia Iuridica Lublinensia*, 30(5):53–74, 2021.
- [34] Giorgio Barnabø, Federico Siciliano, Carlos Castillo, Stefano Leonardi, Preslav Nakov, Giovanni Da San Martino, and Fabrizio Silvestri. Fbmultilingmisinfo: Challenging large-scale multilingual benchmark for misinformation detection. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.

- [35] Frederico Batista Pereira, Natalia Bueno, Felipe Nunes, and Nara Pavao. Inoculation reduces misinformation: Experimental evidence from a multidimensional intervention in brazil. *Available at SSRN 4185949*, 2022.
- [36] Eric PS Baumer. Toward human-centered algorithm design. *Big Data & Society*, 4(2):2053951717718854, 2017.
- [37] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.
- [38] Niklas Bergman Larsson and Jonatan Talåsen. Applying large language models in business processes: A contribution to management innovation, 2024.
- [39] Md Momen Bhuiyan, Hayden Whitley, Michael Horning, Sang Won Lee, and Tanushree Mitra. Designing transparency cues in online news platforms to promote trust: Journalists’ & consumers’ perspectives. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–31, 2021.
- [40] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. ‘it’s reducing a human being to a percentage’ perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*, pages 1–14, 2018.
- [41] Austin Blodgett, Claire Bonial, Taylor Hudson, and Clare Voss. Combined annotations of misinformation, propaganda, and fallacies identified robustly and explainably (campfire).
- [42] Jonas Nygaard Blom and Kenneth Reinecke Hansen. Click bait: Forward-reference as lure in online news headlines. *Journal of Pragmatics*, 76:87–100, 2015.
- [43] Pablo J Boczkowski. *News at work: Imitation in an age of information abundance*. University of Chicago Press, 2010.
- [44] Pablo J Boczkowski, Eugenia Mitchelstein, and Mora Matassi. “news comes across when i’m in a moment of leisure”: Understanding the practices of incidental news consumption on social media. *New media & society*, 20(10):3523–3539, 2018.
- [45] Kirsten Boehner and Jeffrey T Hancock. Advancing ambiguity. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 103–106, 2006.
- [46] Ljubisa Bojic, Nikola Prodanovic, and Agariadne Dwinggo Samala. Maintaining journalistic integrity in the digital age: A comprehensive nlp framework for evaluating online news content. *arXiv preprint arXiv:2401.03467*, 2024.
- [47] Ljubiša Bojić, Olga Zagovora, Asta Zelenkauskaitė, Vuk Vuković, Milan Čabarkapa, Selma Veseljević Jerković, and Ana Jovančević. Comparing large language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. *Scientific reports*, 15(1):11477, 2025.

- [48] Alireza Bonyadi and Moses Samuel. Headlines in newspaper editorials: A contrastive study. *Sage Open*, 3(2):2158244013494863, 2013.
- [49] Sandra L Borden and Chad Tew. The role of journalist and the performance of journalism: Ethical lessons from “fake” news (seriously). *Journal of Mass Media Ethics*, 22(4):300–314, 2007.
- [50] William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318, 2017.
- [51] Ann-Kathrin Brand, Hauke S Meyerhoff, Florian Holl, and Annika Scholl. When linguistic uncertainty spreads across pieces of information: Remembering facts on the news as speculation. *Journal of Experimental Psychology: Applied*, 29(1):18, 2023.
- [52] Nadia M Brashier, Gordon Pennycook, Adam J Berinsky, and David G Rand. Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences*, 118(5):e2020043118, 2021.
- [53] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.
- [54] Virginia Braun and Victoria Clarke. One size fits all? what counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, 18(3):328–352, 2021.
- [55] Wesley Bray. Bloomberg launches gen ai summarisation tool for news, 2025. Accessed June 2025.
- [56] Chris Brogly and Victoria L Rubin. Detecting clickbait: Here’s how to do it/comment détecter les pièges à clic. *Canadian Journal of Information and Library Science*, 42(3):154–175, 2018.
- [57] Alan Bryman. *Social research methods*. Oxford university press, 2016.
- [58] Eliane Léontine Bucher, Peter Kalum Schou, and Matthias Waldkirch. Pacifying the algorithm–anticipatory compliance in the face of algorithmic management in the gig economy. *Organization*, 28(1):44–67, 2021.
- [59] Taina Bucher. The algorithmic imaginary: Exploring the ordinary affects of facebook algorithms. In *The social power of algorithms*, pages 30–44. Routledge, 2019.
- [60] Mikhail Budnikov, Anna Bykova, and Ivan P Yamshchikov. Generalization potential of large language models. *Neural Computing and Applications*, 37(4):1973–1997, 2025.
- [61] Cody Buntain and Jennifer Golbeck. Automatically identifying fake news in popular twitter threads. In *2017 IEEE international conference on smart cloud (smartCloud)*, pages 208–215. IEEE, 2017.

- [62] Lucy H Butler, Toby Prike, and Ullrich KH Ecker. Nudge-based misinformation interventions are effective in information environments with low misinformation prevalence. *Scientific Reports*, 14(1):11495, 2024.
- [63] Pengshan Cai, Kaiqiang Song, Sangwoo Cho, Hongwei Wang, Xiaoyang Wang, Hong Yu, Fei Liu, and Dong Yu. Generating user-engaging news headlines. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3265–3280, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [64] José J Cañas. Ai and ethics when human beings collaborate with ai agents. *Frontiers in psychology*, 13:836650, 2022.
- [65] David Caswell. Audiences, automation, and ai: From structured news to language models. *AI Magazine*, 45(2):174–186, 2024.
- [66] Andrew Chadwick and James Stanyer. Deception as a bridging concept in the study of disinformation, misinformation, and misperceptions: Toward a holistic framework. *Communication Theory*, 32(1):1–24, 2022.
- [67] Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. Stop clickbait: Detecting and preventing clickbaits in online news media. In *2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*, pages 9–16. IEEE, 2016.
- [68] Alan Chamberlain, Andy Crabtree, Tom Rodden, Matt Jones, and Yvonne Rogers. Research in the wild: understanding ‘in the wild’ approaches to design and development. In *Proceedings of the designing interactive systems conference*, pages 795–796, 2012.
- [69] Po-Hsuan Chang, Akshi Kumar, and Saurabh Raj Sangwan. Sens-head: A machine learning framework for sensationalism detection in news headlines using linguistic and semantic features. *British Journal of Multidisciplinary and Advanced Studies*, 6(3), 2025.
- [70] Manuel Chavez and Eric Freedman. Misinformation, disinformation, and journalism ethics. In *Encyclopedia of Business and Professional Ethics*, pages 1330–1334. Springer, 2023.
- [71] Valerie Chen, Q Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. Understanding the role of human intuition on reliance in human-ai decision-making with explanations. *Proceedings of the ACM on Human-computer Interaction*, 7(CSCW2):1–32, 2023.
- [72] Yimin Chen, Niall J Conroy, and Victoria L Rubin. Misleading online content: recognizing clickbait as “false news”. In *Proceedings of the 2015 ACM on workshop on multimodal deception detection*, pages 15–19, 2015.
- [73] Yu-Chieh Chen, Pei-Yu Huang, Chun Lin, Yi-Ting Huang, and Meng Chang Chen. Headline diagnosis: Manipulation of content farm headlines. *arXiv preprint arXiv:2204.11408*, 2022.

- [74] Zichen Chen, Yunhao Luo, and Misha Sra. Engaging with ai: How interface design shapes human-ai collaboration in high-stakes decision-making. *arXiv preprint arXiv:2501.16627*, 2025.
- [75] Hu-Zi Cheng, Bin Sheng, Aaron Lee, Varun Chaudhary, Atanas G Atanasov, Nan Liu, Yue Qiu, Tien Yin Wong, Yih-Chung Tham, and Ying-Feng Zheng. Have ai-generated texts from llm infiltrated the realm of scientific writing? a large-scale analysis of preprint platforms. *bioRxiv*, pages 2024–03, 2024.
- [76] Ben Chester Cheong. Transparency and accountability in ai systems: safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics*, 6:1421273, 2024.
- [77] Sophie Chesney, Maria Liakata, Massimo Poesio, and Matthew Purver. Incongruent headlines: Yet another way to mislead your readers. In *Proceedings of the 2017 emnlp workshop: Natural language processing meets journalism*, pages 56–61, 2017.
- [78] Emil Chiauzzi and Paul Wicks. Digital trespass: ethical and terms-of-use violations by researchers accessing data from an online patient community, 2019.
- [79] Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jinheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S Yu, et al. Llm agents for education: Advances and applications. *arXiv preprint arXiv:2503.11733*, 2025.
- [80] Besjon Cifliku and Hendrik Heuer. ” this could save us months of work”–use cases of ai and automation support in investigative journalism. *arXiv preprint arXiv:2503.16011*, 2025.
- [81] Douglas Cirqueira, Markus Helfert, and Marija Bezbradica. Towards design principles for user-centric explainable ai in fraud detection. In *International Conference on Human-Computer Interaction*, pages 21–40. Springer, 2021.
- [82] Michael Conklin. The truth can be deceiving: How criminal justice headlines are misinterpreted. *The Truth Can Be Deceiving: How Criminal Justice Headlines Are Misinterpreted, NE. U. EXTRA LEGAL (Mar. 21, 2020)*, 2020.
- [83] John Cook, Ullrich Ecker, and Stephan Lewandowsky. Misinformation and how to correct it. *Emerging trends in the social and behavioral sciences: An interdisciplinary, searchable, and linkable resource*, pages 1–17, 2015.
- [84] John W Creswell and Vicki L Plano Clark. *Designing and conducting mixed methods research*. Sage publications, 2017.
- [85] Edward E Cureton. Rank-biserial correlation. *Psychometrika*, 21(3):287–290, 1956.
- [86] Yue Dai, Wenting Yu, and Fei Shen. The effects of message order and debiasing information in misinformation correction. *International Journal of Communication*, 15:21, 2021.

- [87] Paul D'angelo. News framing as a multiparadigmatic research program: A response to entman. *Journal of communication*, 52(4):870–888, 2002.
- [88] Nabanita Das, Bikash Sadhukhan, Tanusree Chatterjee, and Satyajit Chakrabarti. Effect of public sentiment on stock market movement prediction during the covid-19 outbreak. *Social network analysis and mining*, 12(1):92, 2022.
- [89] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022.
- [90] Heidar Davoudi, Aijun An, Morteza Zihayat, and Gordon Edall. Adaptive paywall mechanism for digital news media. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 205–214, 2018.
- [91] Erik de Vries. The sentiment is in the details: a language-agnostic approach to dictionary expansion and sentence-level sentiment analysis in news media. *Computational Communication Research*, 4(2):424–462, 2022.
- [92] Wesley Hanwen Deng, Michelle S Lam, Ángel Alexander Cabrera, Danaë Metaxa, Motahhare Eslami, and Kenneth Holstein. Supporting user engagement in testing, auditing, and contesting ai. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, pages 556–559, 2023.
- [93] Remi Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. Whose ground truth? accounting for individual and collective identities underlying dataset annotation. *arXiv preprint arXiv:2112.04554*, 2021.
- [94] Norman K Denzin. *The research act: A theoretical introduction to sociological methods*. McGraw-Hill, 1978.
- [95] Jacob Devasier, Rishabh Mediratta, Akshith Putta, and Chengkai Li. Automatic fact-checking with frame-semantics. *arXiv preprint arXiv:2501.13288*, 2025.
- [96] Matthew R. DeVerna, Han Yu Yan, Kai-Cheng Yang, and Filippo Menczer. Fact-checking information from large language models can decrease headline discernment. *Proceedings of the National Academy of Sciences*, 121(50):e2322823121, 2024.
- [97] Nicholas Diakopoulos. *Automating the news: How algorithms are rewriting the media*. Harvard University Press, 2019.
- [98] Nicholas Diakopoulos. Computational news discovery: Towards design considerations for editorial orientation algorithms in journalism. *Digital journalism*, 8(7):945–967, 2020.
- [99] Nicholas Dias, Gordon Pennycook, and David G. Rand. Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review*, 1(1), 2020.

- [100] Nicholas DiFonzo and Prashant Bordia. *Rumor psychology: Social and organizational approaches*. American Psychological Association, 2007.
- [101] Zijian Ding, Alison Smith-Renner, Wenjuan Zhang, Joel R Tetreault, and Alejandro Jaimes. Harnessing the power of llms: Evaluating human-ai text co-creation through the lens of news headline generation. *arXiv preprint arXiv:2310.10706*, 2023.
- [102] Alan Dix. Computer supported cooperative work: a framework. In *Design issues in CSCW*, pages 9–26. Springer, 1994.
- [103] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*, pages 275–285, 2019.
- [104] Daniel Dor. On newspaper headlines as relevance optimizers. *Journal of pragmatics*, 35(5):695–721, 2003.
- [105] James N Druckman, Erik Peterson, and Rune Slothuus. How elite partisan polarization affects public opinion formation. *American political science review*, 107(1):57–79, 2013.
- [106] M. Duan, E. Hill, and M. White. Generating disambiguating paraphrases for structurally ambiguous sentences. *Proceedings of the 10th Linguistic Annotation Workshop Held In Conjunction With ACL 2016 (LAW-X 2016)*, 2016.
- [107] A. Dumitrache, L. Aroyo, and C. Welty. Capturing ambiguity in crowdsourcing frame disambiguation. *Proceedings of the AAI Conference on Human Computation and Crowdsourcing*, 6:12–20, 2018.
- [108] Anca Dumitrache, Lora Aroyo, and Chris Welty. Capturing ambiguity in crowdsourcing frame disambiguation. In *Proceedings of the AAI Conference on Human Computation and Crowdsourcing*, volume 6, pages 12–20, 2018.
- [109] Anamaria Dutceac Segesten, Michael Bossetta, Nils Holmberg, and Diederick Niehorster. The cueing power of comments on social media: how disagreement in facebook comments affects user engagement with news. *Information, Communication & Society*, 25(8):1115–1134, 2022.
- [110] Rudresh Dwivedi, Devam Dave, Het Naik, Smiiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33, 2023.
- [111] Ullrich KH Ecker, Stephan Lewandowsky, Ee Pin Chang, and Rekha Pillai. The effects of subtle misinformation in news headlines. *Journal of experimental psychology: applied*, 20(4):323, 2014.
- [112] Ullrich KH Ecker, Stephan Lewandowsky, Briony Swire, and Darren Chang. Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic bulletin & review*, 18:570–578, 2011.

- [113] Ullrich KH Ecker, Toby Prike, Antonia B Paver, Rosie J Scott, and Briony Swire-Thompson. Don't believe them! reducing misinformation influence through source discreditation. *Cognitive Research: Principles and Implications*, 9(1):52, 2024.
- [114] Ullrich KH Ecker, Caitlin XM Sharkey, and Briony Swire-Thompson. Correcting vaccine misinformation: A failure to replicate familiarity or fear-driven backfire effects. *Plos one*, 18(4):e0281140, 2023.
- [115] Stephanie Edgerly. Audience sensemaking: A mapping approach. *Digital Journalism*, 10(1):165–187, 2022.
- [116] Emre Egriboz and Mehmet Fatih Amasyalı. Gpt-2 modelleri kullanarak haber özetinden başlık ve başlıktan haber özeti üretimi generation of headline from news summary and news summary from headline using gpt-2 models. In *2024 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–6. IEEE, 2024.
- [117] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–19, 2021.
- [118] Robert M Entman. Framing: Toward clarification of a fractured paradigm. *Journal of communication*, 43(4):51–58, 1993.
- [119] Ziv Epstein, Nicolo Foppiani, Sophie Hilgard, Sanjana Sharma, Elena Glassman, and David Rand. Do explanations increase the effectiveness of ai-crowd generated fake news warnings? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 183–193, 2022.
- [120] Edgar Erdfelder, Franz Faul, and Axel Buchner. Gpower: A general power analysis program. *Behavior research methods, instruments, & computers*, 28:1–11, 1996.
- [121] Mingming Fan, Xianyou Yang, TszTung Yu, Q Vera Liao, and Jian Zhao. Human-ai collaboration for ux evaluation: effects of explanation and synchronization. *Proceedings of the ACM on human-computer interaction*, 6(CSCW1):1–32, 2022.
- [122] Noureen Fatima, Sher Muhammad Daudpota, Zenun Kastrati, Ali Shariq Imran, Saif Hassan, and Nouh Sabri Elmitwally. Improving news headline text generation quality through frequent pos-tag patterns analysis. *Engineering Applications of Artificial Intelligence*, 125:106718, 2023.
- [123] Lisa K Fazio. Repetition increases perceived truth even for known falsehoods. *Collabra: Psychology*, 6(1), 2020.
- [124] Lisa K Fazio, Nadia M Brashier, B Keith Payne, and Elizabeth J Marsh. Knowledge does not protect against illusory truth. *Journal of experimental psychology: general*, 144(5):993, 2015.

- [125] Sharon Ferguson, Paula Akemi Aoyagui, Rimsha Rizvi, Young-Ho Kim, and Anastasia Kuzminykh. The explanation that hits home: The characteristics of verbal explanations that affect human perception in subjective decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–37, 2024.
- [126] Sharon A Ferguson, Paula Akemi Aoyagui, and Anastasia Kuzminykh. Something borrowed: exploring the influence of ai-generated explanation text on the composition of human explanations. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2023.
- [127] Emilio Ferrara, Herbert Chang, Emily Chen, Goran Muric, and Jaimin Patel. Characterizing social media manipulation in the 2020 us presidential election. *First Monday*, 2020.
- [128] Andrea Ferrario and Michele Loi. How explainability contributes to trust in ai. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 1457–1466, 2022.
- [129] Charles J Fillmore and Beryl T Atkins. Toward a frame-based lexicon: The semantics of risk and its neighbors. In *Frames, fields, and contrasts*, pages 75–102. Routledge, 2012.
- [130] Katherine Fink. The biggest challenge facing journalism: A lack of trust. *Journalism*, 20(1):40–43, 2019.
- [131] Andrew J Flanagin and Miriam J Metzger. The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. *New media & society*, 9(2):319–342, 2007.
- [132] Fredrik Filipsson. Ai case study: The washington post’s use of ai for automated content creation, 2025. Accessed June 2025.
- [133] Batya Friedman, Peter H Kahn, Alan Borning, and Alina Huldgtren. Value sensitive design and information systems. *Early engagement and new technologies: Opening up the laboratory*, pages 55–95, 2013.
- [134] JH Fritz. Ethics for digital journalists: emerging best practices. *CHOICE: Current Reviews for Academic Libraries*, 52(10):1647–1648, 2015.
- [135] Luis Guilherme G. Da Fonseca, Carlos Henrique Gomes Ferreira, and Julio Cesar Soares Dos Reis. The role of news source certification in shaping tweet content: Textual and dissemination patterns in brazil’s 2022 elections. In *Proceedings of the 20th Brazilian Symposium on Information Systems*, pages 1–10, 2024.
- [136] Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout. Social clicks: What and who gets read on twitter? In *Proceedings of the 2016 ACM SIGMETRICS international conference on measurement and modeling of computer science*, pages 179–192, 2016.

- [137] Saadia Gabriel, Liang Lyu, James Siderius, Marzyeh Ghassemi, Jacob Andreas, and Asu Ozdaglar. Misinfoeval: Generative ai in the era of” alternative facts”. *arXiv preprint arXiv:2410.09949*, 2024.
- [138] Stefan Gaillard, Zoril A Oláh, Stephan Venmans, and Michael Burke. Countering the cognitive, linguistic, and psychological underpinnings behind susceptibility to fake news: A review of current literature with special focus on the role of age and digital literacy. *Frontiers in Communication*, 6:661801, 2021.
- [139] Lingyu Gao, Aditi Chaudhary, Krishna Srinivasan, Kazuma Hashimoto, Karthik Raman, and Michael Bendersky. Ambiguity-aware in-context learning with large language models. *arXiv preprint arXiv:2309.07900*, 2023.
- [140] José A García-Avilés, Miguel Carvajal-Prieto, Alicia De Lara-González, and Félix Arias-Robles. Developing an index of media innovation in a national market: The case of Spain. *Journalism studies*, 19(1):25–42, 2018.
- [141] Ron Garland. The midpoint on a rating scale: Is it desirable? *Marketing Bulletin*, 2(1):66–70, 1991.
- [142] William W Gaver, Jacob Beaver, and Steve Benford. Ambiguity as a resource for design. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 233–240, 2003.
- [143] Anna Gaysynsky, Nicole Senft Everson, Kathryn Heley, and Wen-Ying Sylvia Chou. Perceptions of health misinformation on social media: cross-sectional survey study. *Jmir Infodemiology*, 4:e51127, 2024.
- [144] Christine Geeng, Savanna Yee, and Franziska Roesner. Fake news on facebook and twitter: Investigating how people (don’t) investigate. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14, 2020.
- [145] John G Geer and Kim Fridkin Kahn. Grabbing attention: An experimental investigation of headlines during campaigns. *Political Communication*, 10(2):175–191, 1993.
- [146] Jens Gerken. Longitudinal research in human-computer interaction. 2011.
- [147] Daniel T Gilbert, Romin W Tafarodi, and Patrick S Malone. You can’t not believe everything you read. *Journal of personality and social psychology*, 65(2):221, 1993.
- [148] Tarleton GILLESPIE. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press, 2018.
- [149] Dimitrios Giomelakis, Christina Karypidou, and Andreas Veglis. Seo inside newsrooms: Reports from the field. *Future internet*, 11(12):261, 2019.
- [150] Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A Cool, Zahir Kanjee, Andrew S Parsons, Neera Ahuja, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Network Open*, 7(10):e2440969–e2440969, 2024.

- [151] Catalina Gomez, Sue Min Cho, Shichang Ke, Chien-Ming Huang, and Mathias Unberath. Human-ai collaboration is not very collaborative yet: A taxonomy of interaction patterns in ai-assisted decision making from a systematic review. *Frontiers in Computer Science*, 6:1521066, 2025.
- [152] Lucas Graves. *Deciding what's true: The rise of political fact-checking in American journalism*. Columbia University Press, 2016.
- [153] Yasmin Green, Andrew Gully, Yoel Roth, Abhishek Roy, Joshua A Tucker, and Alicia Wanless. Evidence-based misinformation interventions: Challenges and opportunities for measurement and collaboration. *Cargenie Endowment for International Peace*. January, 9, 2023.
- [154] Linda Greenhouse. Telling the court's story: Justice and journalism at the supreme court. *Yale LJ*, 105:1537, 1995.
- [155] K. Greškovičová, R. Masaryk, N. Synak, and V. Čavojová. Superlatives, clickbaits, appeals to authority, poor grammar, or boldface: is editorial style related to the credibility of online health messages? *Frontiers in Psychology*, 13, 2022.
- [156] Maurício Gruppi, Benjamin D Horne, and Sibel Adalı. Nela-gt-2020: A large multi-labelled news dataset for the study of misinformation in news articles. *arXiv preprint arXiv:2102.04567*, 2021.
- [157] Andrew Guess, Jonathan Nagler, and Joshua Tucker. Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science advances*, 5(1):eaau4586, 2019.
- [158] Rachel Guldin. *Whose Future? Whose Facts?: A Critical Case Study of News Literacy Education in the United States*. PhD thesis, University of Oregon, 2022.
- [159] Haoming Guo, Tianyi Huang, Huixuan Huang, Mingyue Fan, and Gerald Friedland. A systematic review of multimodal approaches to online misinformation detection. In *2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 312–317. IEEE, 2022.
- [160] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 02 2022.
- [161] Selin Gurgun, Emily Arden-Close, Keith Phalp, and Raian Ali. Online silence: why do people not challenge others when posting misinformation? *Internet Research*, 33(5):1928–1948, 2023.
- [162] Paweł Gwiaździński, Aleksander B Gundersen, Michal Piksa, Izabela Krysińska, Jonas R Kunst, Karolina Noworyta, Agata Olejniuk, Mikołaj Morzy, Rafal Rygula, Tomi Wójtowicz, et al. Psychological interventions countering misinformation in social media: A scoping review. *Frontiers in Psychiatry*, 13:974782, 2023.

- [163] Felix Hamborg. *Revealing media bias in news articles: NLP techniques for automated frame analysis*. Springer Nature, 2023.
- [164] Regula Hänggli and Hanspeter Kriesi. Political framing strategies and their impact on media framing in a swiss direct-democratic campaign. *Political communication*, 27(2):141–157, 2010.
- [165] Md Mahfuzul Haque, Mohammad Yousuf, Ahmed Shatil Alam, Pratyasha Saha, Syed Ishtiaque Ahmed, and Naeemul Hassan. Combating misinformation in bangladesh: Roles and responsibilities as perceived by journalists, fact-checkers, and users. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–32, 2020.
- [166] SM Taiabul Haque, Rayhan Rashed, Mehrab Bin Morshed, Md Main Uddin Rony, Naeemul Hassan, and Syed Ishtiaque Ahmed. Exploring the tensions between the owners and the drivers of uber cars in urban bangladesh. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–25, 2021.
- [167] Aumyo Hassan and Sarah J Barber. The effects of repetition frequency on the illusory truth effect. *Cognitive research: principles and implications*, 6(1):38, 2021.
- [168] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. Claimbuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948, 2017.
- [169] Reese Haydon, Diana R Sanchez, Kathya Garcia, and Monica Tran. Trusting artificial intelligence: The impact of individual predisposition on human-ai interactions.
- [170] Gaole He, Patrick Hemmer, Michael Vössing, Max Schemmer, and Ujwal Gadiraju. Fine-grained appropriate reliance: Human-ai collaboration with a multi-step transparent decision workflow for complex task decomposition. *arXiv preprint arXiv:2501.10909*, 2025.
- [171] Jakob Henke, Stefanie Holtrup, and Wiebke Moehring. Forgiving the news: The effects of error corrections on news users’ reactions and the influence of individual characteristics and perceptions. *Journalism Studies*, 23(7):840–857, 2022.
- [172] Enric Senabre Hidalgo. Dotmocracy and planning poker for uncertainty management in collaborative research: two examples of co-creation techniques derived from digital culture. In *Proceedings of the sixth international conference on technological ecosystems for enhancing multiculturalism*, pages 833–839, 2018.
- [173] Jimpei Hitsuwari, Yoshiyuki Ueda, Woojin Yun, and Michio Nomura. Does human-ai collaboration lead to more creative art? aesthetic evaluation of human-made and ai-generated haiku poetry. *Computers in Human Behavior*, 139:107502, 2023.
- [174] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.

- [175] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance. *Frontiers in Computer Science*, 5:1096257, 2023.
- [176] Steven J Hoffman and Victoria Justicz. Automatically quantifying the scientific quality and sensationalism of news records mentioning pandemics: validating a maximum entropy machine-learning model. *Journal of Clinical Epidemiology*, 75:47–55, 2016.
- [177] Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. Learning to write with cooperative discriminators. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1638–1649, Melbourne, Australia, July 2018. Association for Computational Linguistics.
- [178] L. Homaeian, J. R. Wallace, and S. D. Scott. Joint action storyboards. *Proceedings of the ACM on Human-Computer Interaction*, 5:1–27, 2021.
- [179] Ming-Tung Hong, Jesse Josua Benjamin, and Claudia Müller-Birn. Coordinating agents: Promoting shared situational awareness in collaborative sensemaking. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 217–220, 2018.
- [180] Ming-Tung Hong and Claudia Müller-Birn. Conceptualization of computer-supported collaborative sensemaking. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 199–202, 2017.
- [181] Kartik Hosanagar and Daehwan Ahn. Designing human and generative ai collaboration. *arXiv preprint arXiv:2412.14199*, 2024.
- [182] Yan Huang and Weirui Wang. Message strategies for misinformation correction: Current research, challenges, and opportunities. *Communication and Misinformation: Crisis Events in the Age of Social Media*, pages 145–162, 2025.
- [183] Yingsong Huang, Bing Bai, Shengwei Zhao, Kun Bai, and Fei Wang. Uncertainty-aware learning against label noise on imbalanced datasets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 6960–6969, 2022.
- [184] Zebin Huang, Ziwei Wang, Weibang Bai, Yanpei Huang, Lichao Sun, Bo Xiao, and Eric M Yeatman. A novel training and collaboration integrated framework for human-agent teleoperation. *Sensors*, 21(24):8341, 2021.
- [185] Hazem Ibrahim, Nouar AlDahoul, Syed Mustafa Ali Abbasi, Fareed Zaffar, Talal Rahwan, and Yasir Zaki. A longitudinal analysis of racial and gender bias in new york times and fox news images and articles. *arXiv preprint arXiv:2410.21898*, 2024.
- [186] Elly Ifantidou. Newspaper headlines, relevance and emotive effects. *Journal of Pragmatics*, 218:17–30, 2023.

- [187] Ifrah Ikram and Blnd Othman. Human-ai collaboration: Enhancing creativity and productivity in the digital age. *Baltic Journal of Multidisciplinary Research*, 1(1):1–6, 2025.
- [188] Md Rafiqul Islam, Shaowu Liu, Xianzhi Wang, and Guandong Xu. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, 10(1):82, 2020.
- [189] Dominic Isha. Designing human-centered news recommender systems. *PhD diss., Vrije Universiteit Amsterdam*, 2023.
- [190] Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. Contrastive explanations for model interpretability. *arXiv preprint arXiv:2103.01378*, 2021.
- [191] Farnaz Jahanbakhsh, Amy X Zhang, Karrie Karahalios, and David R Karger. Our browser extension lets readers change the headlines on news articles, and you won't believe what they did! *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–33, 2022.
- [192] Kristina Janét, Othello Richards, and Asheley R Landrum. Headline format influences evaluation of, but not engagement with, environmental news. *Journalism Practice*, 16(1):35–55, 2022.
- [193] Joonwon Jang, Yoon-Sik Cho, Minju Kim, and Misuk Kim. Detecting incongruent news headlines with auxiliary textual information. *Expert Systems with Applications*, 199:116866, 2022.
- [194] Mohammad Hossein Jarrahi, Yuanye Ma, and Cami Goray. An integrative framework of information as both objective and subjective. *Journal of Information Science*, 49(3):582–594, 2023.
- [195] Jakob D. Jensen. Scientific uncertainty in news coverage of cancer research: Effects of hedging on scientists' and journalists' credibility. *Human Communication Research*, 34(3):347–369, 2008.
- [196] Gongyao Jiang, Xinran Shi, and Qiong Luo. Llm-collaboration on automatic science journalism for the general audience. *arXiv preprint arXiv:2407.09756*, 2024.
- [197] Tingting Jiang, Qian Guo, Shunchang Chen, and Jiaqi Yang. What prompts users to click on news headlines? evidence from unobtrusive data analysis. *Aslib Journal of Information Management*, 72(1):49–66, 2020.
- [198] Todd D Jick. Mixing qualitative and quantitative methods: Triangulation in action. *Administrative science quarterly*, 24(4):602–611, 1979.
- [199] Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orii, and Peter Szolovits. Hooks in the headline: Learning to generate headlines with controlled styles. *arXiv preprint arXiv:2004.01980*, 2020.

- [200] Robert Johns. Likert items and scales. *Survey Question Bank: Methods Fact Sheet*, 1, 2005.
- [201] Thomas Johnson and Barbara Kaye. Choosing is believing? how web gratifications and reliance affect internet credibility among politically interested users. *Atlantic Journal of Communication*, 18(1):1–21, 2010.
- [202] Christopher M Jones, Daniel Diethel, Johannes Schöning, Rehana Shrestha, Tina Jahnel, and Benjamin Schüz. Impact of social reference cues on misinformation sharing on social media: Series of experimental studies. *Journal of Medical Internet Research*, 25:e45583, 2023.
- [203] Gargi Joshi, Ananya Srivastava, Bhargav Yagnik, Mohammed Hasan, Zainuddin Saiyed, Lubna A Gabralla, Ajith Abraham, Rahee Walambe, and Ketan Kotecha. Explainable misinformation detection across multiple social media platforms. *IEEE Access*, 11:23634–23646, 2023.
- [204] John T Jost, Sander Van Der Linden, Costas Panagopoulos, and Curtis D Hardin. Ideological asymmetries in conformity, desire for shared reality, and the spread of misinformation. *Current opinion in psychology*, 23:77–83, 2018.
- [205] Garth S Jowett and Victoria O’donnell. *Propaganda & persuasion*. Sage publications, 2018.
- [206] Pascal Jürgens and Birgit Stark. Mapping exposure diversity: The divergent effects of algorithmic curation on news consumption. *Journal of Communication*, 72(3):322–344, 2022.
- [207] Dan M Kahan. Misconceptions, misinformation, and the logic of identity-protective cognition (ssrn scholarly paper id 2973067). *Social Science Research Network*. <https://doi.org/10.2139/ssrn.2973067>, 2017.
- [208] Gaurav Kamath, Sebastian Schuster, Sowmya Vajjala, and Siva Reddy. Scope ambiguities in large language models. *Transactions of the Association for Computational Linguistics*, 12:738–754, 2024.
- [209] Payal Kamboj, Ayan Banerjee, and Sandeep KS Gupta. Expert knowledge driven human-ai collaboration for medical imaging: a study on epileptic seizure onset zone identification. *IEEE Transactions on Artificial Intelligence*, 2024.
- [210] Manuj Kant, Sareh Nabi, Manav Kant, Roland Scharrer, Megan Ma, and Marzieh Nabi. Towards robust legal reasoning: Harnessing logical llms in law. *arXiv preprint arXiv:2502.17638*, 2025.
- [211] Denys Katerenchuk and Rivka Levitan. You should probably read this: Hedge detection in text. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7703–7707. IEEE, 2021.

- [212] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14, 2020.
- [213] Doyoung Kim, Sander Maathuis, and Esther-Mirjam Sent. Human-centered evaluation metrics for explainable ai: A systematic review and taxonomy. *arXiv preprint arXiv:2404.01234*, 2024.
- [214] Sunnie SY Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. Humans, ai, and context: Understanding end-users' trust in a real-world computer vision application. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 77–88, 2023.
- [215] Spiro Kioussis. Explicating media salience: A factor analysis of new york times issue coverage during the 2000 us presidential election. *Journal of communication*, 54(1):71–87, 2004.
- [216] T. Komatsu, M. G. Lopez, S. Makri, C. Porlezza, G. Cooper, A. MacFarlane, and S. Missaoui. Ai should embody our values: investigating journalistic values to inform ai technology design. *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, pages 1–13, 2020.
- [217] Maria Konnikova. How headlines change the way we think. *The New Yorker*, 17, 2014.
- [218] Alex Zhi-Xiong Koo, Min-Hsin Su, Sangwon Lee, So-Yun Ahn, and Hernando Rojas. What motivates people to correct misinformation? examining the effects of third-person perceptions and perceived norms. *Journal of Broadcasting & Electronic Media*, 65(1):111–134, 2021.
- [219] Anastasia Kozyreva, Stefan M Herzog, Stephan Lewandowsky, Ralph Hertwig, Philipp Lorenz-Spreen, Mark Leiser, and Jason Reifler. Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences*, 120(7):e2210666120, 2023.
- [220] Petra Kralj Novak, Teresa Scantamburlo, Andraž Pelicon, Matteo Cinelli, Igor Mozetič, and Fabiana Zollo. Handling disagreement in hate speech modelling. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 681–695. Springer, 2022.
- [221] Gangeshwar Krishnamurthy, Raj Kumar Gupta, and Yinping Yang. Soccogcom at semeval-2020 task 11: characterizing and detecting propaganda using sentence-level emotional salience features. *arXiv preprint arXiv:2008.13012*, 2020.
- [222] Lisa Merete Kristensen. Audience metrics: Operationalizing news value for the digital newsroom. *Journalism Practice*, 17(5):991–1008, 2023.
- [223] Anne Kruger. The future of news literacy in a connected world. 2017.

- [224] W. Krüger, I. Henrico, and H. Smit. The social contract at risk: covid-19 misinformation in south africa. *Jàmbá: Journal of Disaster Risk Studies*, 16, 2024.
- [225] Jeffrey Kuiken, Anne Schuth, Martijn Spitters, and Maarten Marx. Effective headlines of newspaper articles in a digital environment. *Digital journalism*, 5(10):1300–1314, 2017.
- [226] Ajay Kumar and James W Taylor. Feature importance in the age of explainable ai: Case study of detecting fake news & misinformation via a multi-modal framework. *European Journal of Operational Research*, 317(2):401–413, 2024.
- [227] Anna Sophie Kümpel, Veronika Karnowski, and Till Keyling. News sharing in social media: A review of current research on news sharing users, content, and networks. *Social media+ society*, 1(2):2056305115610141, 2015.
- [228] Jim A Kuypers. Framing analysis from a rhetorical perspective. In *Doing news framing analysis*, pages 302–327. Routledge, 2010.
- [229] Anton Kuznietsov, Balint Gyevar, Cheng Wang, Steven Peters, and Stefano V Albrecht. Explainable ai for safe and trustworthy autonomous driving: a systematic review. *IEEE Transactions on Intelligent Transportation Systems*, 2024.
- [230] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q Vera Liao, and Chenhao Tan. Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 1369–1385, 2023.
- [231] Kenza Lamot, Tim Kreutz, and Michaël Opgenhaffen. “we rewrote this title”: How news headlines are remediated on facebook and how this affects engagement. *Social Media+ Society*, 8(3):20563051221114827, 2022.
- [232] Sophie Lecheler, Mario Keer, Andreas RT Schuck, and Regula Hänggli. The effects of repetitive news framing on political opinions over time. *Communication Monographs*, 82(3):339–358, 2015.
- [233] Angela M Lee and Hsiang Iris Chyi. When newsworthy is not noteworthy: Examining the value of news from the audience’s perspective. *Journalism studies*, 15(6):807–820, 2014.
- [234] Dongyub Lee, Taesun Whang, Chanhee Lee, and Heuseok Lim. Towards reliable and fluent large language models: Incorporating feedback learning loops in qa systems. *arXiv preprint arXiv:2309.06384*, 2023.
- [235] Jin Ha Lee, Nicole Santero, Arpita Bhattacharya, Emma May, and Emma S Spiro. Community-based strategies for combating misinformation: Learning from a popular culture fandom. *Harvard Kennedy School Misinformation Review*, 2022.
- [236] Marine Lemaire, Steeven Ye, Lorna Le Stanc, Grégoire Borst, and Mathieu Cassotti. The development of media truth discernment and fake news detection is related to the development of reasoning during adolescence. *Scientific Reports*, 15(1):6854, 2025.

- [237] Neil L Levy and Robert M Ross. The cognitive science of fake news. 2020.
- [238] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, 13(3):106–131, 2012.
- [239] Stephan Lewandowsky and Sander Van Der Linden. Countering misinformation and fake news through inoculation and prebunking. *European review of social psychology*, 32(2):348–384, 2021.
- [240] Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. Large language model agent for fake news detection. *arXiv preprint arXiv:2405.01593*, 2024.
- [241] Zhengpeng Li, Jiansheng Wu, Jiawei Miao, and Xinmiao Yu. News headline generation based on improved decoder from transformer. *Scientific Reports*, 12(1):11648, 2022.
- [242] Q Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the ai: informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–15, 2020.
- [243] Elina Lindgren, Torun Lindholm, Rens Vliegenthart, Hajo G Boomgaarden, Alyt Damstra, Jesper Strömbäck, and Yariv Tsfati. Trusting the facts: The role of framing, news media as a (trusted) source, and opinion resonance for perceived truth in statistical statements. *Journalism & Mass Communication Quarterly*, 101(4):981–1004, 2024.
- [244] Andrew T Little, Keith E Schnakenberg, and Ian R Turner. Motivated reasoning and democratic accountability. *American Political Science Review*, 116(2):751–767, 2022.
- [245] Qingtao Liu. Generative ai and journalism ethics: Controversies over chatgpt. *Journal of Information, Technology and Policy*, pages 1–6, 2025.
- [246] Xingyu Liu. *The Effects of Message, Source, and Audience Features on the Effectiveness of Misinformation Correction*. PhD thesis, University of California, Santa Barbara, 2024.
- [247] Ying Liu and Lei Shen. Consolidating human-ai collaboration research in organizations: A literature review. *Journal of Computer, Signal, and System Research*, 2(1):131–151, 2025.
- [248] Yixin Liu, AR Fabbri, P Liu, D Radev, and A Cohan. On learning to summarize with large language models as references (2023). *arXiv preprint arXiv:2305.14239*.
- [249] George Loewenstein. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin*, 116(1):75, 1994.
- [250] Chang Lu, Bo Hu, Qiang Li, Chao Bi, and Xing-Da Ju. Psychological inoculation for credibility assessment, sharing intention, and discernment of misinformation: Systematic review and meta-analysis. *Journal of Medical Internet Research*, 25:e49255, 2023.

- [251] Feng Lu, Anca Dumitrache, and David Graus. Beyond optimizing for clicks: Incorporating editorial values in news recommendation. In *Proceedings of the 28th ACM conference on user modeling, adaptation and personalization*, pages 145–153, 2020.
- [252] Siwen Luo, Hamish Ivison, Soyeon Caren Han, and Josiah Poon. Local interpretations for explainable natural language processing: A survey. *ACM Computing Surveys*, 56(9):1–36, 2024.
- [253] Bernhard Lutz, Marc Adam, Stefan Feuerriegel, Nicolas Pröllochs, and Dirk Neumann. Which linguistic cues make people fall for fake news? a comparison of cognitive and affective processing. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–22, 2024.
- [254] Ning Ma, Guang Yu, and Xin Jin. Investigation of public acceptance of misinformation correction in social media based on sentiment attributions: Infodemiology study using aspect-based sentiment analysis. *Journal of Medical Internet Research*, 26:e50353, 2024.
- [255] N. Maiorana, S. Marceglia, M. Treddenti, M. Tosi, M. Guidetti, M. F. Creta, T. Bocci, S. Oliveri, F. M. Boneschi, and A. Priori. Is it time for the neurologist to use large language models in everyday practice? 2025.
- [256] Adam Maksl, Seth Ashley, and Stephanie Craft. Measuring news media literacy. *Journal of Media Literacy Education*, 6(3):29–45, 2015.
- [257] João Marecos, Ethan Shattock, Oliver Bartlett, Francisco Goiana-da Silva, Hendramoorthy Maheswaran, Hutan Ashrafian, and Ara Darzi. Health misinformation and freedom of expression: considerations for policymakers. *Health Economics, Policy and Law*, 18(2):204–217, 2023.
- [258] Ramasamy Mariappan. Extensive review of literature on explainable ai (xai) in healthcare applications. *Recent Advances in Computer Science and Communications*, 18(1):E200324228159, 2025.
- [259] Cameron Martel, Steve Rathje, Cory J Clark, Gordon Pennycook, Jay J Van Bavel, David G Rand, and Sander van der Linden. On the efficacy of accuracy prompts across partisan lines: an adversarial collaboration. *Psychological science*, 35(4):435–450, 2024.
- [260] Alice Marwick and Rebecca Lewis. Media manipulation and disinformation online. *New York: Data & Society Research Institute*, 359:1146–1151, 2017.
- [261] Wojciech Mazurczyk, Dongwon Lee, and Andreas Vlachos. Disinformation 2.0 in the age of ai: A cybersecurity perspective. *Communications of the ACM*, 67(3):36–39, 2024.
- [262] Patrick E McKnight and Julius Najab. Mann-whitney u test. *The Corsini encyclopedia of psychology*, pages 1–1, 2010.
- [263] Robert McPhedran, Michael Ratajczak, Max Mawby, Emily King, Yuchen Yang, and Natalie Gold. Psychological inoculation protects against the social media infodemic. *Scientific reports*, 13(1):5780, 2023.

- [264] Sabiha Mehreen, Ahatesham Bhuiyan, and Eftekhar Hossain. An attentive fusion framework for multi-modal misinformation detection. In *2024 27th International Conference on Computer and Information Technology (ICCIT)*, pages 1903–1908. IEEE, 2024.
- [265] Claude H Miller, Elena Bessarabova, Bobi Ivanov, and John A Banas. Motivation-based approaches to countering mass-mediated misinformation, 2024.
- [266] S Milli, M Carroll, Y Wang, S Pandey, and AD Dragan. Engagement, user satisfaction, and the amplification of divisive content on social media. *arxiv*, 2023.
- [267] Rahul Mishra, Piyush Yadav, Remi Calizzano, and Markus Leippold. Musem: Detecting incongruent news headlines using mutual attentive semantic matching. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 709–716. IEEE, 2020.
- [268] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4):1–45, 2021.
- [269] Maria D Molina, S Shyam Sundar, Thai Le, and Dongwon Lee. “fake news” is not simply false information: A concept explication and taxonomy of online content. *American behavioral scientist*, 65(2):180–212, 2021.
- [270] Camila Mont’Alverne, Sumitra Badrinathan, Amy Ross Arguedas, Benjamin Toff, Richard Fletcher, and Rasmus Kleis Nielsen. The trust gap: How and why news on digital platforms is viewed more sceptically versus news in general. 2022.
- [271] Andrew Moore. Bad science in the headlines: Who takes responsibility when science is distorted in the mass media? *EMBO reports*, 7(12):1193–1196, 2006.
- [272] D. A. Morris-O’Connor, A. Strotmann, and D. Zhao. Editorial behaviors for biasing wikipedia articles. *Proceedings of the Association for Information Science and Technology*, 59:226–234, 2022.
- [273] Maria Mousoulidou, Loukia Taxitari, and Anastasia Christodoulou. Social media news headlines and their influence on well-being: emotional states, emotion regulation, and resilience. *European Journal of Investigation in Health, Psychology and Education*, 14(6):1647–1665, 2024.
- [274] Fuseini Mumuni and Alhassan Mumuni. Explainable artificial intelligence (xai): from inherent explainability to large language models. *arXiv preprint arXiv:2501.09967*, 2025.
- [275] Sheila T Murphy and Robert B Zajonc. Affect, cognition, and awareness: affective priming with optimal and suboptimal stimulus exposures. *Journal of personality and social psychology*, 64(5):723, 1993.

- [276] Merja Myllylahti. Newspaper paywalls—the hype and the reality: A study of how paid news content impacts on media corporation revenues. *Digital journalism*, 2(2):179–194, 2014.
- [277] Preslav Nakov and Giovanni Da San Martino. Fake news, disinformation, propaganda, media bias, and flattening the curve of the covid-19 infodemic. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 4054–4055, 2021.
- [278] Xiaoli Nan, Yuan Wang, and Kathryn Thier. Why people believe health misinformation and who are at risk? a systematic review of individual differences in susceptibility to health misinformation. *Social Science & Medicine*, page 115398, 2022.
- [279] Philip Napoli. *Social media and the public interest: Media regulation in the disinformation age*. Columbia university press, 2019.
- [280] Philip M Napoli. *Audience evolution: New technologies and the transformation of media audiences*. Columbia University Press, 2011.
- [281] Bilal Nasir and Ijaz Hussain. Automating agile workflows: The role of multi-agent llms in modern software engineering. 2024.
- [282] Sazid Nazat, Osvaldo Arreche, and Mustafa Abdallah. On evaluating black-box explainable ai methods for enhancing anomaly detection in autonomous driving systems. *Sensors*, 24(11):3515, 2024.
- [283] Dennis C Neale, John M Carroll, and Mary Beth Rosson. Evaluating computer-supported cooperative work: models and frameworks. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 112–121, 2004.
- [284] Ani Nenkova and Kathleen McKeown. A survey of text summarization techniques. *Mining text data*, pages 43–76, 2012.
- [285] Terrence Neumann, Maria De-Arteaga, and Sina Fazelpour. Justice in misinformation detection systems: An analysis of algorithms, stakeholders, and potential harms. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1504–1515, 2022.
- [286] David T Newman, Nathanael J Fast, and Derek J Harmon. When eliminating bias isn’t fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes*, 160:149–167, 2020.
- [287] Mikel K Ngueajio, Saurav Aryal, Marcellin Atemkeng, Gloria Washington, and Danda Rawat. Decoding fake news and hate speech: A survey of explainable ai techniques. *ACM Computing Surveys*, 57(7):1–37, 2025.
- [288] Sachita Nishal, Charlotte Li, and Nicholas Diakopoulos. Domain-specific evaluation strategies for ai in journalism. *arXiv preprint arXiv:2403.17911*, 2024.

- [289] Hyelim Oh, Khim-Yong Goh, and Tuan Q Phan. Are you what you tweet? the impact of sentiment on digital news consumption and social media sharing. *Information Systems Research*, 34(1):111–136, 2023.
- [290] Markus Ojala. Is the age of impartial journalism over? the neutrality principle and audience (dis) trust in mainstream news. *Journalism studies*, 22(15):2042–2060, 2021.
- [291] Femi Olan, Uchitha Jayawickrama, Emmanuel Ogiemwonyi Arakpogun, Jana Suklan, and Shaofeng Liu. Fake news on social media: the impact on society. *Information Systems Frontiers*, pages 1–16, 2022.
- [292] Luciana Oliveira, Paulino Silva, Anabela Mesquita, Arminda Sa Sequeira, and Adriana Oliveira. Sadness, negativity, and uncertainty in education during covid-19 on social media. *International Journal of Online Pedagogy and Course Design (IJOPCD)*, 12(1):1–21, 2022.
- [293] Amin Omidvar and Aijun An. Learning to generate popular headlines. *IEEE Access*, 2023.
- [294] Amin Omidvar, Hossein Poormodheji, Aijun An, and Gordon Edall. Learning to determine the quality of news headlines. *arXiv preprint arXiv:1911.11139*, 2019.
- [295] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C Petrantonakis. Verite: a robust benchmark for multimodal misinformation detection accounting for unimodal bias. *International Journal of Multimedia Information Retrieval*, 13(1):4, 2024.
- [296] Jinkyung Park, Pamela Wisniewski, and Vivek Singh. Leveraging large language models (llms) to support collaborative human-ai online risk data annotation. *arXiv preprint arXiv:2404.07926*, 2024.
- [297] Kunwoo Park, Taegyun Kim, Seunghyun Yoon, Meeyoung Cha, and Kyomin Jung. Baitwatcher: A lightweight web interface for the detection of incongruent news headlines. *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*, pages 229–252, 2020.
- [298] Kunwoo Park, Haewoon Kwak, Jisun An, and Sanjay Chawla. How-to present news on social media: A causal analysis of editing news headlines for boosting user engagement. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 491–502, 2021.
- [299] Kunwoo Park, Haewoon Kwak, Jisun An, and Sanjay Chawla. Understanding effects of editing tweets for news sharing by media accounts through a causal inference framework, 2021.
- [300] Javier Pastor-Galindo, Pantaleone Nespoli, and José A Ruipérez-Valiente. Large-language-model-powered agent-based framework for misinformation and disinformation research: opportunities and open challenges. *IEEE Security & Privacy*, 2024.

- [301] Sharoda A Paul and Meredith Ringel Morris. Cosense: enhancing sensemaking for collaborative web search. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1771–1780, 2009.
- [302] Jessica Pearson. Demystifying the second coder, 2020. Accessed: 2025-05-12.
- [303] Supavich Fone Pengnate, Alex Young, and Jeffrey Chen. New details emerge! revealing the effects of clickbait headlines on user responses. 2018.
- [304] Gordon Pennycook, Adam Bear, Evan T Collins, and David G Rand. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management science*, 66(11):4944–4957, 2020.
- [305] Gordon Pennycook, Tyrone D Cannon, and David G Rand. Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general*, 147(12):1865, 2018.
- [306] Gordon Pennycook and David G Rand. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7):2521–2526, 2019.
- [307] Gordon Pennycook and David G Rand. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50, 2019.
- [308] Caitlin Petre. *All the news that’s fit to click: How metrics are transforming the work of journalists*. Princeton University Press, 2021.
- [309] Yeong Tsann Phua, Kwang Hooi Yew, Mohd Fadzil Hassan, and Matthew Teow Yok Wooi. Malay phoneme-based subword news headline generator for low-resource language. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 13(4):4965–4975, 2024.
- [310] Michal Piksa, Karolina Noworyta, Jan Piasecki, Pawel Gwiazdzinski, Aleksander B Gundersen, Jonas Kunst, and Rafal Rygula. Cognitive processes and personality traits underlying four phenotypes of susceptibility to (mis) information. *Frontiers in Psychiatry*, 13:1142, 2022.
- [311] Alicja Piotrkowicz. *Modelling social media popularity of news articles using headline text*. PhD thesis, University of Leeds, 2017.
- [312] Alicja Piotrkowicz, Vania Dimitrova, Jahna Otterbacher, and Katja Markert. The impact of news values and linguistic style on the popularity of headlines on twitter and facebook. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 767–774, 2017.
- [313] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. Declare: Debunking fake news and false claims using evidence-aware deep learning. *arXiv preprint arXiv:1809.06416*, 2018.

- [314] Brian Porter and Edouard Machery. Ai-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports*, 14(1):26133, 2024.
- [315] Martin Potthast, Tim Gollub, Kristof Komlossy, Sebastian Schuster, Matti Wiegmann, Erika Patricia Garces Fernandez, Matthias Hagen, and Benno Stein. Crowdsourcing a large corpus of clickbait on twitter. In *Proceedings of the 27th international conference on computational linguistics*, pages 1498–1507, 2018.
- [316] Wanda Pratt, Madhu C Reddy, David W McDonald, Peter Tarczy-Hornoch, and John H Gennari. Incorporating ideas from computer-supported cooperative work. *Journal of biomedical informatics*, 37(2):128–137, 2004.
- [317] Carolyn C. Preston and Andrew M. Colman. Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, 104(1):1–15, 2000.
- [318] Clara Pretus, Diána R Hughes, Kobi Hackenburg, Manos Tsakiris, Oscar Vilarroya, and Jay J Van Bavel. The misleading count: An identity-based intervention to counter partisan misinformation sharing. 2022.
- [319] Vincent Price, David Tewksbury, and Elizabeth Powers. Switching trains of thought: The impact of news frames on readers’ cognitive responses. *Communication research*, 24(5):481–506, 1997.
- [320] Toby Prike, Phoebe Blackley, Briony Swire-Thompson, and Ullrich KH Ecker. Examining the replicability of backfire effects after standalone corrections. *Cognitive Research: Principles and Implications*, 8(1):39, 2023.
- [321] H. R. Qasim. Impact of fake news on trust in journalism. *MEDAAD*, 2024:41–45, 2024.
- [322] Dorian Quelle and Alexandre Bovet. The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence*, 7:1341697, 2024.
- [323] Fiqih Rahmawati. Collaboration of artificial intelligence and journalists in online media from the perspective of human-machine communication. *Kalijaga Journal of Communication*, 6(1):103–118.
- [324] Praboda Rajapaksha, Reza Farahbakhsh, and Noel Crespi. Bert, xlnet or roberta: the best transfer learning model to detect clickbaits. *IEEE Access*, 9:154704–154716, 2021.
- [325] Aakansha Ramesh, Gauri Thube, and Swaranjali Jadhav. Realtime news analysis using natural language processing. In *2023 4th International Conference for Emerging Technology (INCET)*, pages 1–6. IEEE, 2023.
- [326] Anku Rani, Valdemar Danry, Andy Lippman, and Pattie Maes. Can dialogues with ai systems help humans better discern visual misinformation? *arXiv preprint arXiv:2504.06517*, 2025.

- [327] David N Rapp. The consequences of reading inaccurate information. *Current Directions in Psychological Science*, 25(4):281–285, 2016.
- [328] Shaina Raza, Caesar Saleh, Emrul Hasan, Franklin Ogidi, Maximus Powers, Veronica Chatrath, Marcelo Lotif, Roya Javadi, Anam Zahid, and Vahid Reza Khazaie. Vilbias: A framework for bias detection using linguistic and visual cues. *arXiv preprint arXiv:2412.17052*, 2024.
- [329] Shaina Raza, Ashmal Vayani, Aditya Jain, Aravind Narayanan, Vahid Reza Khazaie, Syed Raza Bashir, Elham Dolatabadi, Gias Uddin, Christos Emmanouilidis, Rizwan Qureshi, et al. Vldbench: Vision language models disinformation detection benchmark. *arXiv preprint arXiv:2502.11361*, 2025.
- [330] Carlo Reverberi, Tommaso Rigon, Aldo Solari, Cesare Hassan, Paolo Cherubini, and Andrea Cherubini. Experimental evidence of effective human–ai collaboration in medical decision-making. *Scientific reports*, 12(1):14952, 2022.
- [331] Jeba Rezwana and Mary Lou Maher. Cofi: A framework for modeling interaction in human-ai co-creative systems. In *ICCC*, pages 444–448, 2021.
- [332] Patrick R Rich and Maria S Zaragoza. Correcting misinformation in news stories: An investigation of correction timing and correction durability. *Journal of Applied Research in Memory and Cognition*, 9(3):310–322, 2020.
- [333] Julio Rieis, Fabrício de Souza, Pedro Vaz de Melo, Raquel Prates, Haewoon Kwak, and Jisun An. Breaking the news: First impressions matter on online news. In *Proceedings of the international AAAI conference on web and social media*, volume 9, pages 357–366, 2015.
- [334] Joshua Robison and Kevin J Mullinix. Elite polarization and public opinion: How polarization is communicated and its effects. *Political Communication*, 33(2):261–282, 2016.
- [335] Md Main Uddin Rony, Naeemul Hassan, and Mohammad Yousuf. Diving deep into clickbaits: Who use them to what extents in which topics with what effects? In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, pages 232–239, 2017.
- [336] Jon Roozenbeek, Alexandra LJ Freeman, and Sander Van Der Linden. How accurate are accuracy-nudge interventions? a preregistered direct replication of pennycook et al.(2020). *Psychological science*, 32(7):1169–1178, 2021.
- [337] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. The cost structure of sensemaking. In *Proceedings of the INTERACT’93 and CHI’93 conference on Human factors in computing systems*, pages 269–276, 1993.
- [338] Mirka Saarela and Vili Podgorelec. Recent applications of explainable ai (xai): A systematic literature review. *Applied Sciences*, 14(19):8884, 2024.

- [339] Erica Salkin and Kevin Grieves. The “major mea culpa:” journalistic discursive techniques when professional norms are broken. *Journalism Studies*, 23(9):1096–1113, 2022.
- [340] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 315–328, 2021.
- [341] Fumiko Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34(S1):1–97, 1969.
- [342] Javier Sánchez-Junquera, Paolo Rosso, Manuel Montes, and Simone Paolo Ponzetto. Masking and transformer-based models for hyperpartisanship detection in news. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1244–1251, 2021.
- [343] Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Ježek. Why don’t you do it right? analysing annotators’ disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, 2023.
- [344] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, 22(2014):4349–4357, 2014.
- [345] Murat Sariyar and Jürgen Holm. Medical informatics in a tension between black-box ai and trust. In *Informatics and Technology in Clinical Care and Public Health*, pages 41–44. IOS Press, 2022.
- [346] Joshua M Scacco and Ashley Muddiman. The curiosity effect: Information seeking in the contemporary news environment. *New Media & Society*, 22(3):429–448, 2020.
- [347] Mike Schaekermann. Human-ai interaction in the presence of ambiguity: From deliberation-based labeling to ambiguity-aware ai. 2020.
- [348] Mike Schaekermann, Graeme Beaton, Elaheh Sanoubari, Andrew Lim, Kate Larson, and Edith Law. Ambiguity-aware ai assistants for medical data analysis. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14, 2020.
- [349] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–19, 2018.
- [350] Max Schemmer, Andrea Bartos, Philipp Spitzer, Patrick Hemmer, Niklas Kühl, Jonas Liebschner, and Gerhard Satzger. Towards effective human-ai decision-making: The role of human learning in appropriate reliance on ai advice. *arXiv preprint arXiv:2310.02108*, 2023.

- [351] Vera Schmitt, Luis-Felipe Villa-Arenas, Nils Feldhus, Joachim Meyer, Robert P Spang, and Sebastian Möller. The role of explainability in collaborative human-ai disinformation detection. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2157–2174, 2024.
- [352] Vincent Schmitt, Luis Fernando Villa-Arenas, Niklas Feldhus, Julia Meyer, Rainer P Spang, and Sebastian Möller. The role of explainability in collaborative human-ai disinformation detection. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2157–2174, 2024.
- [353] Carsten D Schultz. Driving likes, comments, and shares on social networking sites: how post characteristics affect brand interactions in apparel retailing. In *Proceedings of the 18th Annual International Conference on Electronic Commerce: e-Commerce in Smart connected World*, pages 1–9, 2016.
- [354] Tal Schuster, Darsh J Shah, Yun Jie Serene Yeo, Daniel Filizzola, Enrico Santus, and Regina Barzilay. Towards debiasing fact verification models. *arXiv preprint arXiv:1908.05267*, 2019.
- [355] Lauren Scott, Lynne Coventry, Marta E Cecchinato, and Mark Warner. “i figured her feeling a little bit bad was worth it to not spread that kind of hate”: Exploring how uk families discuss and challenge misinformation. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2023.
- [356] Philip Sedgwick. Spearman’s rank correlation coefficient. *Bmj*, 349, 2014.
- [357] Connie Moon Sehat, Ryan Li, Peipei Nie, Tarunima Prabhakar, and Amy X Zhang. Misinformation as a harm: structured approaches for fact-checking prioritization. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–36, 2024.
- [358] Julian Senoner, Simon Schallmoser, Bernhard Kratzwald, Stefan Feuerriegel, and Torbjørn Netland. Explainable ai improves task performance in human–ai collaboration. *Scientific Reports*, 14(1):31150, 2024.
- [359] Robiert Sepúlveda-Torres, Alba Bonet-Jover, and Estela Saquete. Detecting misleading headlines through the automatic recognition of contradiction in spanish. *IEEE Access*, 11:72007–72026, 2023.
- [360] Robiert Sepúlveda-Torres, Marta Vicente, Estela Saquete, Elena Lloret, and Manuel Palomar. Headlinestancechecker: Exploiting summarization to detect headline disinformation. *Journal of Web Semantics*, 71:100660, 2021.
- [361] Jiaming Shen, Jialu Liu, Dan Finnie, Negar Rahmati, Mike Bendersky, and Marc Najork. “why is this misleading?”: Detecting news headline hallucinations with explanations. In *Proceedings of the ACM Web Conference 2023*, pages 1662–1672, 2023.
- [362] Jiaming Shen, Tianqi Liu, Jialu Liu, Zhen Qin, Jay Pavagadhi, Simon Baumgartner, and Michael Bendersky. Multilingual fine-grained news headline hallucination detection. *arXiv preprint arXiv:2407.15975*, 2024.

- [363] Xinyu Shen. Research of public news headlines' bias in the context of new media: Using "weibo" hot search headlines as a case study. In *SHS Web of Conferences*, volume 178, page 02001. EDP Sciences, 2023.
- [364] Anu Shrestha, Francesca Spezzano, and Indhumathi Gurunathan. Multi-modal analysis of misleading political news. In *Disinformation in Open Online Media: Second Multidisciplinary International Symposium, MISDOOM 2020, Leiden, The Netherlands, October 26–27, 2020, Proceedings 2*, pages 261–276. Springer, 2020.
- [365] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188, 2020.
- [366] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.
- [367] Ergon Cugler de Moraes Silva and Jose Carlos Vaz. How disinformation and fake news impact public policies?: A review of international literature. *arXiv preprint arXiv:2406.00951*, 2024.
- [368] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. Participation is not a design fix for machine learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–6, 2022.
- [369] Tara Slominski, Andrew Fugleberg, Warren M Christensen, John B Buncher, and Jennifer L Momsen. Using framing as a lens to understand context effects on expert reasoning. *CBE—Life Sciences Education*, 19(3):ar48, 2020.
- [370] Kacper Sokol and Julia E Vogt. What does evaluation of explainable artificial intelligence actually tell us? a case for compositional and contextual validation of xai building blocks. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2024.
- [371] Jitendra Soni and Kirti Mathur. Sentiment analysis of news headlines for stock market prediction using vader. In *2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, pages 1215–1222. IEEE, 2023.
- [372] Francesco Sovrano and Fabio Vitali. An objective metric for explainable ai: how and why to estimate the degree of explainability. *Knowledge-Based Systems*, 278:110866, 2023.
- [373] Timo Spinde, Christina Kreuter, Wolfgang Gaissmaier, Felix Hamborg, Bela Gipp, and Helge Giese. Do you think it's biased? how to ask for the perception of media bias. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–10. IEEE, 2021.
- [374] Philipp Spitzer, Joshua Holstein, Katelyn Morrison, Kenneth Holstein, Gerhard Satzger, and Niklas K uhl. Don't be fooled: The misinformation effect of explanations in human-ai collaboration. *arXiv preprint arXiv:2409.12809*, 2024.

- [375] Riktesh Srivastava, Jitendra Singh Rathore, Sachin Kumar Srivastava, and Khushboo Agnihotri. The impact on society of false news spreading on social media with the help of predictive modelling. *International Journal of Knowledge and Learning*, 15(4):307–318, 2022.
- [376] Kate Starbird, Ahmer Arif, and Tom Wilson. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–26, 2019.
- [377] Florian Stoeckel, Sabrina Stöckli, Besir Ceka, Chiara Ricchi, Ben Lyons, and Jason Reifler. Social corrections act as a double-edged sword by reducing the perceived accuracy of false and real news in the uk, germany, and italy. *Communications Psychology*, 2(1):10, 2024.
- [378] Anselm Strauss and Juliet Corbin. Basics of qualitative research techniques. 1998.
- [379] Lucille Alice Suchman. *Plans and situated actions: The problem of human-machine communication*. Cambridge university press, 1987.
- [380] Yoo Yeon Sung, Jordan Boyd-Graber, and Naeemul Hassan. Not all fake news is written: A dataset and analysis of misleading video headlines. *arXiv preprint arXiv:2310.13859*, 2023.
- [381] Cass R Sunstein and Adrian Vermeule. Conspiracy theories: Causes and cures. *Journal of political philosophy*, 17(2):202–227, 2009.
- [382] Daniel Susser. Ethical considerations for digitally targeted public health interventions, 2020.
- [383] Briony Swire-Thompson, Nicholas Miklaucic, John P Wihbey, David Lazer, and Joseph DeGutis. The backfire effect after correcting misinformation is strongly associated with reliability. *Journal of Experimental Psychology: General*, 151(7):1655, 2022.
- [384] Terrence Szymanski, Claudia Orellana-Rodriguez, and Mark T Keane. Helping news editors write better headlines: A recommender to improve the keyword contents & shareability of news headlines. *arXiv preprint arXiv:1705.09656*, 2017.
- [385] Charles S Taber and Milton Lodge. Motivated skepticism in the evaluation of political beliefs. *American journal of political science*, 50(3):755–769, 2006.
- [386] Edson C Tandoc Jr. Journalism is twerking? how web analytics is changing the process of gatekeeping. *New media & society*, 16(4):559–575, 2014.
- [387] Edson C Tandoc Jr, Zheng Wei Lim, and Richard Ling. Defining “fake news” a typology of scholarly definitions. *Digital journalism*, 6(2):137–153, 2018.
- [388] Bolan Tang. Role of language in emotions and bias: A cognitive linguistic perspective of affective political media in the us. *Lecture Notes in Education Psychology and Public Media*, 51:46–51, 2024.

- [389] Percy H Tannenbaum. The effect of headlines on the interpretation of news stories. *Journalism Quarterly*, 30(2):189–197, 1953.
- [390] L. Q. Tay, M. J. Hurlstone, T. Kurz, and U. K. H. Ecker. A comparison of prebunking and debunking interventions for implied versus explicit misinformation. *British Journal of Psychology*, 113:591–607, 2021.
- [391] Li Qian Tay, Stephan Lewandowsky, Mark J Hurlstone, Tim Kurz, and Ullrich KH Ecker. Thinking clearly about misinformation. *Communications Psychology*, 2(1):4, 2024.
- [392] Maartje Ter Hoeve, Anne Schuth, Daan Odijk, and Maarten de Rijke. Faithfully explaining rankings in a news recommender system. *arXiv preprint arXiv:1805.05447*, 2018.
- [393] Camille Thibault, Jacob-Junqi Tian, Gabrielle Péloquin-Skulski, Taylor Lynn Curtis, James Zhou, Florence Laflamme, Yuxiang Guan, Reihaneh Rabbany, Jean-François Godbout, and Kellin Pelrine. A guide to misinformation detection data and evaluation. In *Workshop on Spurious Correlation and Shortcut Learning: Foundations and Solutions*.
- [394] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL-HLT*, 2018.
- [395] Emily Thorson. Belief echoes: The persistent effects of corrected misinformation. *Political Communication*, 33(3):460–480, 2016.
- [396] Daniel Tison. The future of journalism: Opportunities and challenges of human-ai collaboration. 2023.
- [397] Thanasis Troboukis, Kelly Kiki, Antonis Galanopoulos, Pavlos Sermpezis, Stelios Karamanidis, Ilias Dimitriadis, and Athena Vakali. Towards hybrid intelligence in journalism: Findings and lessons learnt from a collaborative analysis of greek political rhetoric by chatgpt and humans. *arXiv preprint arXiv:2410.13400*, 2024.
- [398] Yariv Tsfati and Gal Ariely. Individual and contextual correlates of trust in media across 44 countries. *Communication research*, 41(6):760–782, 2014.
- [399] Michael Twidale, David Randall, and Richard Bentley. Situated evaluation for cooperative systems. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pages 441–452, 1994.
- [400] Takane Ueno, Yuto Sawa, Yeongdae Kim, Jacqueline Urakami, Hiroki Oura, and Katie Seaborn. Trust in human-ai interaction: Scoping out models, measures, and methods. In *CHI conference on human factors in computing systems extended abstracts*, pages 1–7, 2022.
- [401] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. When combinations of humans and ai are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, pages 1–11, 2024.

- [402] Sebastián Valenzuela, Arturo Arriagada, and Andrés Scherman. Facebook, twitter, and youth engagement: A quasi-experimental study of social media use and protest behavior using propensity score matching. *International Journal of Communication*, 8:25, 2014.
- [403] Jay J Van Bavel, Steve Rathje, Madalina Vlasceanu, and Clara Pretus. Updating the identity-based model of belief: From false belief to the spread of misinformation. *Current Opinion in Psychology*, 56:101787, 2024.
- [404] Toni GLA Van Der Meer, Michael Hameleers, and Jakob Ohme. Can fighting misinformation have a negative spillover effect? how warnings for the threat of misinformation can decrease general news credibility. *Journalism Studies*, 24(6):803–823, 2023.
- [405] Teun A Van Dijk. *News as discourse*. Routledge, 2013.
- [406] Chris J Vargo, Lei Guo, and Michelle A Amazeen. The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New media & society*, 20(5):2028–2049, 2018.
- [407] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38, 2023.
- [408] Robert Vidigal and Jennifer Jerit. Issue importance and the correction of misinformation. *Political Communication*, 39(6):715–736, 2022.
- [409] Marco Viviani and Gabriella Pasi. Credibility in social media: opinions, news, and health information—a survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 7(5):e1209, 2017.
- [410] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.
- [411] Emily K Vraga and Melissa Tully. News literacy, social media behaviors, and skepticism toward information on social media. *Information, Communication & Society*, 24(2):150–166, 2021.
- [412] Earl J Wagner, Jiahui Liu, Larry Birnbaum, and Kenneth D Forbus. Rich interfaces for reading news on the web. In *Proceedings of the 14th international conference on Intelligent user interfaces*, pages 27–36, 2009.
- [413] Guan Wang, Rebecca Frederick, Jinglong Duan, William Wong, Verica Rugar, Weihua Li, and Quan Bai. Detecting misinformation through framing theory: the frame element-based model. *arXiv preprint arXiv:2402.15525*, 2024.
- [414] Jenny S Wang, Samar Haider, Amir Tohidi, Anushkaa Gupta, Yuxuan Zhang, Chris Callison-Burch, David Rothschild, and Duncan J Watts. 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*, pages 1–27, 2025.
- [415] William Y Wang. “liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 422–426, 2017.
- [416] Xinru Wang and Ming Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pages 318–328, 2021.
- [417] Xinru Wang and Ming Yin. Effects of explanations in ai-assisted decision making: Principles and comparisons. *ACM Transactions on Interactive Intelligent Systems*, 12(4):1–36, 2022.
- [418] Claire Wardle and Hossein Derakhshan. *Information disorder: Toward an interdisciplinary framework for research and policymaking*, volume 27. Council of Europe Strasbourg, 2017.
- [419] Yashwant Arjunrao Waykar. Human-ai collaboration in explainable recommender systems: An exploration of user-centric explanations and evaluation frameworks. *International Journal of Scientific Research in Engineering and Management*, 7(07):2582–3930, 2023.
- [420] Wei Wei and Xiaojun Wan. Learning to identify ambiguous and misleading news headlines. *arXiv preprint arXiv:1705.06031*, 2017.
- [421] Andreas Widholm and Ester Appelgren. A softer kind of hard news? data journalism and the digital renewal of public service news in sweden. *new media & society*, 24(6):1363–1381, 2022.
- [422] H James Wilson and Paul R Daugherty. Creating the symbiotic ai workforce of the future. *MIT Sloan Management Review*, 61(1):1–4, 2019.
- [423] Maximiliane Windl, Sebastian S Feger, Lara Zijlstra, Albrecht Schmidt, and Pawel W Wozniak. ‘it is not always discovery time’: Four pragmatic approaches in designing ai systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2022.
- [424] Ryan Wold. Segmentation, surveillance, and automation: Practical and ethical considerations for attracting, sustaining, and monetizing audience attention online. *Journal of Business and Technical Communication*, 37(2):103–139, 2023.
- [425] Robert F Woolson. Wilcoxon signed-rank test. *Encyclopedia of biostatistics*, 8, 2005.
- [426] Madelyne Xiao and Jonathan Mayer. The challenges of machine learning for trust and safety: a case study on misinformation detection. *arXiv preprint arXiv:2308.12215*, 2023.

- [427] Qing Xiao, Xianzhe Fan, Felix M Simon, Bingbing Zhang, and Motahhare Eslami. "it might be technically impressive, but it's practically useless to us": Practices, challenges, and opportunities for cross-functional collaboration around ai within the news industry. *arXiv preprint arXiv:2409.12000*, 2024.
- [428] Xizhu Xiao, Porismita Borah, Danielle Ka Lai Lee, Yan Su, and Sojung Kim. A story is better told with collective interests: An experimental examination of misinformation correction during the covid-19 pandemic. *American Journal of Health Promotion*, 37(7):915–923, 2023.
- [429] Zhan Xu, Mary Laffidy, and Lauren Ellis. Clickbait for climate change: comparing emotions in headlines and full-texts and their engagement. *Information, Communication & Society*, 26(10):1915–1932, 2023.
- [430] Keyang Xuan, Li Yi, Fan Yang, Ruochen Wu, Yi R Fung, and Heng Ji. Lemma: towards lvlm-enhanced multimodal misinformation detection with external knowledge augmentation. *arXiv preprint arXiv:2402.11943*, 2024.
- [431] Jingyuan Yi, Zeqiu Xu, Tianyi Huang, and Peiyang Yu. Challenges and innovations in llm-powered fake news detection: A synthesis of approaches and future directions. In *Proceedings of the 2025 2nd International Conference on Generative Artificial Intelligence and Information Security*, pages 87–93, 2025.
- [432] Kayo Yin and Graham Neubig. Interpreting language models with contrastive explanations. *arXiv preprint arXiv:2202.10419*, 2022.
- [433] Seunghyun Yoo. Co-cot: A prompt-based framework for collaborative chain-of-thought reasoning. *arXiv preprint arXiv:2504.17091*, 2025.
- [434] Seunghyun Yoon, Kunwoo Park, Minwoo Lee, Taegyun Kim, Meeyoung Cha, and Kyomin Jung. Learning to detect incongruence in news headline and body text via a graph neural network. *IEEE Access*, 9:36195–36206, 2021.
- [435] Seunghyun Yoon, Kunwoo Park, Joongbo Shin, Hongjun Lim, Seungpil Won, Meeyoung Cha, and Kyomin Jung. Detecting incongruity between news headline and body text via a deep hierarchical encoder. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 791–800, 2019.
- [436] Jiaao Zhan, Yang Gao, Yu Bai, and Qianhui Liu. Stage-wise stylistic headline generation: Style generation and summarized content insertion. *IJCAI*, 2022.
- [437] Qiang Zhang, Shangsong Liang, Aldo Lipani, Zhaochun Ren, and Emine Yilmaz. From stances' imbalance to their hierarchical representation and detection. In *The world wide web conference*, pages 2323–2332, 2019.
- [438] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57, 2024.

- [439] Yanci Zhang and Han Yu. Uncertainty-aware explainable federated learning. *arXiv preprint arXiv:2503.05194*, 2025.
- [440] Yuqi Zhang, Bin Guo, Yasan Ding, Jiaqi Liu, Chen Qiu, Sicong Liu, and Zhiwen Yu. Investigation of the determinants for misinformation correction effectiveness on social media during covid-19 pandemic. *Information Processing & Management*, 59(3):102935, 2022.
- [441] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–20, 2023.
- [442] X Zhou, J Wu, and R Zafarani. Safe: similarity-aware multi-modal fake news detection. arxiv. *arXiv preprint arXiv:2003.04981*, 2020.
- [443] Xinyi Zhou and Reza Zafarani. Fake news: A survey of research, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*, 2, 2018.
- [444] Clara Ziche and Giovanni Apruzzese. Llm4pm: A case study on using large language models for process modeling in enterprise organizations. In *International Conference on Business Process Management*, pages 472–483. Springer, 2024.
- [445] Fabiana Zollo, Petra Kralj Novak, Michela Del Vicario, Alessandro Bessi, Igor Mozetič, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Emotional dynamics in the age of misinformation. *PloS one*, 10(9):e0138740, 2015.