Between Consensus and Ambiguity: Expert Evaluation of LLM Explanations for Misleading Headlines

Md Main Uddin Rony
Bowling Green State University
Bowling Green, Ohio, USA

Naeemul Hassan
University of Maryland
College Park, Maryland, USA

ABSTRACT

As large language models (LLMs) are considered for editorial support, their reliability in subjectively ambiguous tasks remains uncertain. Misleading headline detection exemplifies this challenge, as even human annotators often disagree due to rhetorical nuance or contextual omission. This paper examines how annotator agreement shapes LLM performance and how expert journalists assess model explanations. Using a dataset of 60 real-world headlines annotated for misleadingness, we stratify cases by agreement level and compare LLM predictions with human judgments. We then engage professional journalists to evaluate the quality of LLM-generated explanations along dimensions such as correctness, clarity, and ambiguity awareness. Our results show that LLMs align well with human labels in high-agreement cases but diverge substantially when annotator consensus is low. In contested cases, experts found model explanations overly confident and lacking contextual nuance. These findings highlight the need for ambiguity-aware explanation strategies to better support editorial decision-making.

KEYWORDS

Misleading News Headlines, Large Language Models (LLMs), Annotator Agreement, Explainable AI (XAI), Human-AI Collaboration

1 INTRODUCTION

Large language models (LLMs) have achieved impressive performance across a growing range of natural language processing tasks, including fact-checking [33], summarization [42], and legal reasoning [24]. Their fluency [28] and contextual generalization [10] have led to widespread integration in organizational workflows such as business [7], software engineering [31], journalism [23], and education [13]. This adoption opens new opportunities for Human–AI collaboration, where AI systems support rather than replace expert decision-making.

Effective collaboration depends on explanation quality, since users must interpret AI outputs. Explanations influence trust, reliance, and decision quality [22]. In objective domains such as fraud detection or medical diagnosis, explainable AI (XAI) aids decisions by providing verifiable justifications [15]. But in subjective domains marked by ambiguity or contested knowledge, explanations must also serve argumentative functions, helping users weigh competing interpretations [19]. In such contexts, LLMs risk obscuring uncertainty or fostering unwarranted trust [38].

We focus on misleading news headlines, a subtle but consequential form of misinformation [12]. Misleading headlines distort or omit context, shaping public perception without explicit falsehoods, and often spread more widely than fake news [4]. Detecting them requires sensitivity to framing, omission, and tone [26], yet even experts frequently disagree [40], underscoring their inherently subjective nature. In such cases, explanation quality cannot be assessed by correctness alone but must be judged against professional norms and contextual reasoning [41].

This paper examines how LLMs handle misleading headline detection under conditions of human disagreement. Using 60 real-world headlines annotated for misleadingness [35], we stratify cases by annotator agreement to assess how ambiguity shapes classification accuracy and explanation quality. We further engage professional journalists to evaluate model-generated explanations along dimensions such as correctness, clarity, and ambiguity awareness, with open-ended feedback on divergences from editorial reasoning.

Our study addresses two research questions:

- **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?

By connecting annotator agreement (RQ1) and ambiguity divergence (RQ2), this work shows LLM limits in editorial tasks and points to ambiguity-aware strategies for stronger explanations.

2 RELATED WORK

2.1 Misleading Headlines

Misleading news headlines, sometimes referred to as *incongruent headlines*, distort or misrepresent article content without making explicitly false claims [12]. They often rely on selective framing, emphasis, or omission, guiding readers toward misinterpretation despite factual accuracy. Reader responses to such headlines are highly subjective, shaped by prior knowledge, ideology, and linguistic framing [14]. Research shows that misleading headlines impair inference and memory [18], with effects magnified by the fact that many readers only skim headlines [20]. Yet most misinformation studies continue to emphasize clearly false claims, overlooking the interpretive complexity and social consequences of ambiguous content [5].

CJ Symposium '25, December 11–12, 2025, Miami, FL

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2.2 Annotation Disagreement as a Signal

Crowdsourced misinformation annotation studies consistently show that disagreement is not merely noise but a signal of underlying ambiguity. Approaches such as *CrowdTruth* leverage diverse annotations to capture multiple valid perspectives [17], and aggregated crowd evaluations have been shown to align with expert judgments in contested cases [3]. However, variability in annotator expertise can lower reliability [6], and attempts to enforce consensus often suppress nuance. Despite recognition of disagreement as meaningful, many misinformation datasets still rely on binary true/false labels, limiting their ability to capture interpretive complexity [39, 43]. These practices flatten cultural and contextual variation [8], underscoring the need for evaluation frameworks that embrace disagreement as a feature of social judgment tasks.

2.3 Explainable AI in Subjective Contexts

Explainable AI (XAI) has been widely studied as a means of making model reasoning transparent [30], especially in high-stakes domains like healthcare [29] and autonomous driving [27]. Explanations help users validate outputs, negotiate responsibility, and calibrate trust, though their effectiveness depends on user expertise and domain norms [2]. Evaluation frameworks emphasize not only correctness but also interactional usefulness, clarity, and contextual relevance [25]. In misinformation detection, explanation techniques such as feature attribution and free-text rationales have shown promise in surfacing misleading strategies [32], and can improve user discernment [38]. Yet challenges remain: many XAI methods inadequately communicate bias or fairness [1], highlighting the need for explanation systems that foreground ambiguity, interpretive pluralism, and accountability in socially contested tasks.

Together, this work highlights gaps in how LLM explanations are evaluated under ambiguity, motivating our focus on annotator disagreement and expert editorial assessment.

3 METHOD

3.1 Dataset

We use the misleading headline dataset introduced by Rony et al. [35], which contains 60 real-world news headlines. Each headline was annotated for misleadingness by three human raters, who also provided free-text justifications. This produced four agreement levels: 18 headlines labeled misleading by all three annotators (3/3), 9 labeled misleading by two (2/3), 10 labeled misleading by one (1/3), and 23 unanimously judged not misleading (0/3). Cases with full agreement (3/3, 0/3) represent consensus, while mixed cases (2/3, 1/3) reflect subjective ambiguity.

Each headline was also classified by three large language models (LLMs)—GPT-3.5, GPT-4, and Gemini 1.0—accompanied by natural language explanations. Following prior validation [35], we focus on GPT-4 (hereafter GPT) and Gemini, which showed stronger alignment with human judgments. GPT-3.5 was excluded due to over-predicting misleadingness. Table 1 presents a sample headline with labels and explanations from both humans and LLMs.

3.1.1 Stratification by Agreement Level. To operationalize ambiguity, we stratified headlines by annotator agreement: unanimous consensus (3/3 or 0/3) versus ambiguous cases (2/3 or 1/3). This allows us to compare model performance under clear versus contested human judgments.

3.1.2 Data Sampling. From the full set, we selected 21 headlines using a 2×2 matrix of human agreement (unanimous vs. ambiguous) and LLM alignment with the human majority (both, one, or neither model aligned). To reduce expert burden, we refined this to 15 headlines while preserving diversity in topic domains (public health, politics, corporate communication) and coverage of each stratification category. This sampling ensured a manageable yet representative set for expert evaluation.

3.2 Expert Evaluation Study Design

To assess the editorial quality of LLM explanations, we conducted an expert-in-the-loop evaluation with professional journalists. This complements human annotations and model outputs by providing a grounded benchmark of explanation adequacy.

3.2.1 Expert Recruitment. We recruited six U.S.-based participants with professional journalism experience, including reporters, graduate students with newsroom training, and researchers. The group was diverse in gender, age, and career stage, spanning early-career to seasoned professionals. All reported moderate to high familiarity with editorial practices such as headline evaluation and framing analysis. Experts were briefed on the study but received no formal training to preserve authentic judgments and were compensated per institutional guidelines (see Table 2 in Appendix).

3.2.2 Evaluation Dimensions. Each explanation was independently rated across seven dimensions: *Correctness*, *Clarity*, *Exhaustiveness*, *Ambiguity Awareness*, *Risk of Harm*, *Trustworthiness*, and *Actionability*. These dimensions build on XAI evaluation frameworks [25, 38] and capture both intrinsic explanation quality and contextual utility in editorial decision-making.

3.2.3 Task Procedure. Experts reviewed 15 sampled headlines (Section 3.1.2), each paired with its full article. To reduce cognitive load while retaining full context, we adopted a highlighting strategy [38], marking article sentences most relevant to the headline and explanation. Experts could consult the complete article as needed.

For each explanation, experts provided Likert ratings (1 = Not at all, 4 = Completely) on the seven dimensions. We used an even-numbered scale to avoid a neutral midpoint, following methodological recommendations that encourage more discriminative judgments [21]. In addition, experts gave open-ended feedback on weaknesses, missing elements, suggested improvements, and handling of ambiguity.

The evaluation was administered online, designed for completion in multiple sittings, and required 75–90 minutes on average. Instructions emphasized assessing explanations on how accurately and completely they addressed the misleading potential of each headline in relation to the article content.

3.3 Data Analysis

We analyzed LLM performance using a mixed-methods approach that integrated human annotations, model predictions, and expert evaluations. For RQ1, we measured alignment between LLM outputs and human judgments by computing *weighted accuracy*, which accounts for partial agreement in ambiguous cases (1/3 or 2/3). We compared accuracy between unanimous and ambiguous conditions using Mann–Whitney U tests. Expert ratings of explanation quality were aggregated across seven dimensions (Correctness, Clarity, Exhaustiveness, Ambiguity Awareness, Risk of Harm, Trustworthiness, Actionability) and similarly compared by agreement level to assess whether perceived quality varied with human consensus.

For RQ2, we examined model behavior in ambiguous cases through both quantitative and qualitative lenses. Paired expert ratings for GPT and Gemini explanations were compared using Wilcoxon signed-rank tests with effect size estimation. To complement these scores, we conducted a thematic analysis of expert open-ended feedback on weaknesses, missing information, suggested improvements, and handling of ambiguity. This enabled us to identify recurring divergences between LLM rationales and professional editorial reasoning, particularly in cases of interpretive uncertainty.

4 FINDINGS

This section presents findings from our evaluation of LLM-generated explanations for misleading news headlines, focusing on the first two research questions. RQ1 examines how levels of human annotator agreement relate to both LLM classification accuracy and expert evaluations of explanation quality. RQ2 investigates ambiguous cases (1/3 or 2/3 agreement) to understand how LLM rationales align with or diverge from expert editorial reasoning. We report both quantitative comparisons of model performance and expert ratings, as well as qualitative insights from journalist feedback.

4.1 Overview of Expert Ratings and Reliability

Experts evaluated LLM-generated explanations across seven dimensions using a 4-point scale. Average scores ranged from 2.17 for *Risk of Harm* to 3.46 for *Clarity*, suggesting explanations were generally understandable but often lacked depth in addressing ambiguity or completeness. *Correctness* (3.13) and *Actionability* (2.83) were moderate, while *Exhaustiveness* and *Ambiguity Awareness* were lower, pointing to limitations in how models captured nuance (see Table **??** in the Appendix).

Inter-rater reliability, assessed via Intraclass Correlation Coefficients (ICC), ranged from 0.34 to 0.53, indicating fair to moderate consistency across dimensions. Agreement was strongest for *Ambiguity Awareness* (0.53) and *Correctness* (0.46), suggesting shared standards for identifying caution and factual alignment. Lower values for *Clarity* (0.38) and *Actionability* (0.34) reflect interpretive variation in how these qualities were judged. Overall, the results justify the use of aggregated expert scores while underscoring the value of qualitative feedback to capture editorial nuance.

4.2 Annotator Agreement and LLM Accuracy

We examined how human annotator agreement shaped LLM classification accuracy and expert evaluations of explanation quality (RQ1).

4.2.1 Classification Accuracy. Both models performed significantly better on unanimous cases compared to ambiguous ones. Gemini achieved a mean weighted accuracy of 0.68 ($SD = 0.47$, $n = 41$) on unanimous headlines and 0.51 ($SD = 0.17$, $n = 19$) on ambiguous cases ($U = 532.0$, $p = 0.016$). GPT followed the same pattern, with 0.90 ($SD = 0.30$) for unanimous and 0.53 ($SD = 0.17$) for ambiguous cases, a highly significant difference ($U = 703.0$, $p < 0.001$). As shown in Figure 1, GPT consistently outperformed Gemini, approaching near-perfect alignment in high-agreement cases.

4.2.2 Explanation Quality. Expert evaluations revealed more stability across agreement levels. For *correctness*, GPT maintained consistent scores ($M = 3.33$) in both unanimous and ambiguous cases, while Gemini's ratings rose slightly from 2.73 to 3.14, though not significantly ($p = 0.449$). Ratings of *ambiguity awareness* improved modestly for both models in ambiguous cases (Gemini: 2.48 $\rightarrow$ 3.00; GPT: 2.62 $\rightarrow$ 2.95), suggesting greater hedging or nuance when annotators disagreed, though differences were not significant ($p > 0.26$).

Other dimensions, such as *risk of harm*, were consistently rated low across conditions, with GPT slightly lower overall. Taken together, these findings indicate that while LLM classification accuracy drops sharply under ambiguity, expert perceptions of explanation quality remain relatively stable, with a slight tendency toward higher ambiguity awareness in contested cases.

4.3 Explanations Under Ambiguity

We examined expert evaluations of LLM explanations for ambiguous headlines (1/3 or 2/3 annotator agreement) to answer RQ2.

4.3.1 Quantitative Comparison. Across the seven evaluation dimensions, GPT tended to score slightly higher than Gemini on most measures (e.g., correctness: 3.33 vs. 3.14; trustworthiness: 2.93 vs. 2.76), while Gemini was marginally higher on clarity (3.45 vs. 3.40). However, Wilcoxon signed-rank tests revealed no statistically significant differences between the two models (all $p > 0.05$), and effect sizes were small to moderate but inconsistent in direction (Table 3).

Difference-score distributions (Figure 2) further confirmed the absence of systematic preference: most medians clustered around zero, and paired scores showed high overlap. While some dimensions (e.g., correctness, trustworthiness) displayed greater variability across headlines, neither model consistently outperformed the other. These results suggest that in contexts of interpretive ambiguity, experts rated GPT and Gemini explanations as comparably adequate, or comparably limited.

4.3.2 Expert Diagnoses of Explanation Shortcomings. Thematic analysis of expert feedback on ambiguous cases (1/3 or 2/3 misleading votes) revealed six recurring shortcomings in LLM explanations. Even when factually accurate, rationales often failed to meet standards of editorial judgment.

1. Missing Context and Scientific Precision. Experts repeatedly flagged absent qualifiers in health and science headlines. For example, E1 noted that a Gemini explanation of vaccine effectiveness "*failed to mention that the study was focused on older adults and did not clarify whether it was peer-reviewed or observational.*" Similarly, E6 emphasized that GPT's explanation of a health study "*should have clarified the difference between statistical association and causality—without this, readers could be misled into thinking the study proved more than it did.*"

2. Emotional Framing Overlooked. Journalists also highlighted that explanations often ignored rhetorical or emotional language. E1 criticized a Gemini explanation of a headline using the word "illegal," observing that "*it doesn't acknowledge how the term carries a strong moral and legal charge, beyond what the article itself substantiates*." Likewise, E6 wrote that a GPT explanation of a vaccine booster headline "*could go further by pointing out the declarative framing ('I'm Done') and how this may come across as dismissive or hostile*."

3. Policy vs. Opinion Confusion. Experts expected precise attribution in political coverage. E6 critiqued GPT's explanation of a Florida vaccine policy headline for "*missing the nuance between the chief health official's personal stance and the broader state policy*." E2 made a similar point about Gemini: "*The explanation only mentioned the health official, not the governor, and therefore failed to convey that the statement was part of a state-level recommendation*."

4. Surface-Level Factual Matching. Explanations were often seen as overly literal. E3 wrote of a Gemini explanation about a video game hacking headline: "*It fails to differentiate between the video game and the underlying blockchain bridge—most readers will take it literally, which changes the story*." E1 similarly critiqued GPT: "*It focuses on factual matching without addressing how the headline steers interpretation, especially on sensitive issues like vaccines*."

5. Weak Ambiguity Handling. Experts felt ambiguity was under-addressed. E5 bluntly stated of a Gemini explanation: "*The explanation does not acknowledge any ambiguity—it assumes a straightforward relationship between the headline and article*." E1, reflecting on a GPT case, wrote: "*It identified jurisdictional issues but didn't fully address how the public might misinterpret the headline*." Even when complexity was noted, it was often underdeveloped; as E6 put it, GPT "*gestures toward complexity but stops short of analyzing how the absence of context could mislead or polarize readers*."

6. Suggestions for Improvement. Experts offered actionable recommendations, including headline rewrites and more precise contextualization. For instance, E6 suggested reframing a crypto hacking headline as "*Hackers Steal Over $600M in Crypto from Blockchain Linked to Popular Game Axie Infinity*" to avoid misleading readers into thinking the game itself was hacked. Others urged LLMs to explicitly identify whether an article was news or opinion, with one journalist noting: "*If the piece is an opinion column, the explanation should flag that—otherwise, readers might take it as reporting fact*."

Overall, experts stressed that effective editorial explanations require more than factual correctness. They must interrogate emotional framing, clarify attribution, and contextualize ambiguity in ways that align with journalistic standards. As one journalist summarized: "*The models can sound accurate, but unless they grapple with how headlines mislead in practice, they're not helping us make editorial judgments*."

5 DISCUSSION

This study examined how LLMs generate explanations for misleading news headlines under conditions of both consensus and disagreement. Our findings show that while models align well with human judgments in clear-cut cases (RQ1), their explanations often fall short of expert editorial standards in ambiguous contexts, where missing context, weak ambiguity handling, and superficial framing analysis limit their usefulness (RQ2). We discuss these results in light of subjectivity in misinformation detection, the limits of current explanations, and design implications for supporting human–AI editorial collaboration.

5.1 Subjectivity, Disagreement, and Evaluation

Detecting misleading headlines is inherently subjective, and our findings underscore the importance of treating annotator disagreement as a meaningful signal rather than error [37]. By stratifying cases into unanimous and ambiguous categories, we showed that LLM classification accuracy drops sharply when human judgments diverge (RQ1), highlighting the limits of binary ground-truth approaches in misinformation detection. Yet, expert ratings of explanation quality did not decline under ambiguity, suggesting that models sometimes adopt hedging strategies, even if these lacked editorial depth. Qualitative feedback revealed why: explanations often missed key context, failed to interrogate emotional framing, blurred distinctions between opinion and policy, and inadequately engaged with ambiguity. Our triangulated framework, combining human annotations, LLM outputs, and expert evaluations, reflects a situated evaluation approach [41], showing that explanation usefulness cannot be reduced to predictive alignment alone but must be assessed in context and with attention to disagreement.

5.2 Alignment and Limits of LLM Explanations

Our findings show that while LLMs such as GPT and Gemini can approximate human reasoning in clear-cut editorial contexts, their limitations become evident when interpretive nuance is required. In unanimous cases, particularly for GPT, classification accuracy was high and expert ratings of correctness remained stable, suggesting that models capture surface-level patterns aligned with consensus (RQ1). This echoes prior work showing that LLMs excel at reproducing majority sentiment or canonical interpretations when ambiguity is low [9]. However, in ambiguous cases where annotators disagreed, experts judged explanations as falling short of editorial standards despite their fluency: they often lacked contextual precision, ignored emotional framing, or blurred distinctions between personal opinion and policy (RQ2). Although models sometimes employed hedging language, experts noted that this rhetorical caution rarely reflected genuine engagement with uncertainty, aligning with broader concerns that LLMs obscure rather than surface ambiguity [16, 38]. As prior research emphasizes, ambiguity should not be treated as noise but as a signal to be surfaced and contextualized in collaborative systems [36]. Taken together, these results highlight that explanation usefulness cannot be reduced to factual alignment or linguistic fluency; effective editorial support requires systems that explicitly reason through ambiguity and contextualize uncertainty in ways that align with professional journalistic judgment.

5.3 Design Implications

Our findings suggest that while LLMs can generate fluent and factually accurate explanations, these qualities alone are insufficient for professional editorial decision-making, underscoring the need for systems that prioritize interpretability, accountability, and domain-sensitive judgment [11]. Expert critiques highlighted omissions of attribution, lack of contextual precision, and superficial handling of rhetorical framing (RQ2), underscoring that AI should support rather than automate editorial reasoning. In line with co-adaptive system design [34], we propose five complementary directions for human–AI editorial collaboration: (1) *dimension highlighting*, where explanations are structured around editorial qualities such as correctness, ambiguity awareness, and harm; (2) *framing bias identification*, using contrastive or frame-tagged views to reveal rhetorical slant; (3) *ambiguity surfacing*, which presents multiple plausible rationales or flags contested claims, supporting sensemaking under uncertainty; (4) *justification toggling*, enabling users to switch between concise vs. elaborated or assertive vs. hedged reasoning styles to match editorial contexts; and (5) *editable reasoning*, where explanations serve as collaborative scaffolds that experts can annotate, refine, or repurpose for downstream transparency and accountability. Collectively, these recommendations move beyond generic interpretability toward interactive, situated systems that make AI reasoning legible, flexible, and accountable to editorial expertise.

5.4 Limitations and Future Directions

This study has several limitations. Our expert evaluation relied on a relatively small sample of headlines ($n = 15$), which constrains the generalizability of the statistical findings. LLM outputs were generated at a single point in time using default configurations, meaning that improvements from newer models, alternative prompting strategies, or interactive feedback were not captured. In addition, because our dataset overlapped with the likely training period of GPT-4 and Gemini, some explanations may have reflected memorized rather than generalizable reasoning. Finally, our evaluation focused on core editorial dimensions such as correctness, ambiguity awareness, and risk of harm, but other factors like political bias or audience sensitivity also warrant attention. Future work should expand the dataset, explore adaptive and interactive workflows, and involve a broader range of stakeholders to assess how AI explanations can meaningfully support editorial practice in diverse, real-world contexts.

6 CONCLUSION

This study showed that LLMs align well with human judgments in clear-cut cases but struggle to provide adequate explanations in ambiguous contexts, where experts noted missing context, weak ambiguity handling, and shallow engagement with framing. By combining annotator agreement with expert evaluation, we highlighted both the strengths and limits of current models in supporting editorial reasoning. Future work should expand datasets, explore interactive workflows, and develop systems that surface ambiguity and allow expert oversight to ensure explanations are trustworthy and actionable in journalistic settings.

REFERENCES

[1] Kiana Alikhademi, Brianna Richardson, Emma Drobina, and Juan E Gilbert. 2021. Can explainable AI explain unfairness? A framework for evaluating explainable AI. *arXiv preprint arXiv:2106.07483* (2021).

[2] Ben Allen. 2024. The promise of explainable AI in digital health for precision medicine: a systematic review. *Journal of personalized medicine* 14, 3 (2024), 277.

[3] J. Allen, A. A. Arechar, G. Pennycook, and D. G. Rand. 2021. Scaling up fact-checking using the wisdom of crowds. *Science Advances* 7 (2021). Issue 36. https://doi.org/10.1126/sciadv.abf4393

[4] Jennifer Allen, Duncan J Watts, and David G Rand. 2024. Quantifying the impact of misinformation and vaccine-skeptical content on Facebook. *Science* 384, 6699 (2024), eadk3451.

[5] S. Altay, M. Berriche, H. Heuer, J. Farkas, and S. Rathje. 2023. A survey of expert views on misinformation: definitions, determinants, solutions, and future of the field. *HKS Misinfo Review* (2023). https://doi.org/10.37016/mr-2020-119

[6] J. B. Bak-Coleman. 2023. Limiting factors in the effectiveness of crowd-sourced labeling for combating misinformation. (2023). https://doi.org/10.31235/osf.io/ahm27

[7] Niklas Bergman Larsson and Jonatan Talåsen. 2024. Applying Large Language Models in Business Processes: A contribution to Management Innovation.

[8] Austin Blodgett, Claire Bonial, Taylor Hudson, and Clare Voss. [n. d.]. Combined Annotations of Misinformation, Propaganda, and Fallacies Identified Robustly and Explainably (CAMPFIRE). ([n. d.]).

[9] Ljubiša Bojić, Olga Zagovora, Asta Zelenkauskaite, Vuk Vuković, Milan Čabarkapa, Selma Veseljević Jerković, and Ana Jovančević. 2025. Comparing large Language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. *Scientific reports* 15, 1 (2025), 11477.

[10] Mikhail Budnikov, Anna Bykova, and Ivan P Yamshchikov. 2025. Generalization potential of large language models. *Neural Computing and Applications* 37, 4 (2025), 1973–1997.

[11] Ben Chester Cheong. 2024. Transparency and accountability in AI systems: safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics* 6 (2024), 1421273.

[12] Sophie Chesney, Maria Liakata, Massimo Poesio, and Matthew Purver. 2017. Incongruent headlines: Yet another way to mislead your readers. In *Proceedings of the 2017 EMNLP workshop: natural language processing meets journalism*. 56–61.

[13] Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jinheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S Yu, et al. 2025. Llm agents for education: Advances and applications. *arXiv preprint arXiv:2503.11733* (2025).

[14] Aaron Chuey, Yiwei Luo, and Ellen M. Markman. 2024. Epistemic language in news headlines shapes readers' perceptions of objectivity. *Proceedings of the National Academy of Sciences* 121, 20 (2024), e2314091121. https://doi.org/10.1073/pnas.2314091121 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2314091121

[15] Douglas Cirqueira, Markus Helfert, and Marija Bezbradica. 2021. Towards design principles for user-centric explainable AI in fraud detection. In *International Conference on Human-Computer Interaction*. Springer, 21–40.

[16] Matthew R. DeVerna, Han Yu Yan, Kai-Cheng Yang, and Filippo Menczer. 2024. Fact-checking information from large language models can decrease headline discernment. *Proceedings of the National Academy of Sciences* 121, 50 (2024), e2322823121.

[17] A. Dumitrache, L. Aroyo, and C. Welty. 2018. Capturing ambiguity in crowdsourcing frame disambiguation. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 6 (2018), 12–20. https://doi.org/10.1609/hcomp.v6i1.13330

[18] U. K. H. Ecker, S. Lewandowsky, E. P. Chang, and R. Pillai. 2014. The effects of subtle misinformation in news headlines. *Journal of Experimental Psychology: Applied* 20 (2014), 323–335. Issue 4. https://doi.org/10.1037/xap0000028

[19] Sharon A Ferguson, Paula Akemi Aoyagui, and Anastasia Kuzminykh. 2023. 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*. 1–7.

[20] Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout. 2016. 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*. 179–192.

[21] Ron Garland. 1991. The Midpoint on a Rating Scale: Is it Desirable? *Marketing Bulletin* 2, 1 (1991), 66–70.

[22] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science* 5 (2023), 1096257.

[23] Gongyao Jiang, Xinran Shi, and Qiong Luo. 2024. Llm-collaboration on automatic science journalism for the general audience. *arXiv preprint arXiv:2407.09756* (2024).

[24] Manuj Kant, Sareh Nabi, Manav Kant, Roland Scharrer, Megan Ma, and Marzieh Nabi. 2025. Towards robust legal reasoning: Harnessing logical llms in law. *arXiv preprint arXiv:2502.17638* (2025).

[25] Doyoung Kim, Sander Maathuis, and Esther-Mirjam Sent. 2024. Human-centered evaluation metrics for explainable AI: A systematic review and taxonomy. *arXiv preprint arXiv:2404.01234* (2024).

[26] Maria Konnikova. 2014. How headlines change the way we think. *The New Yorker* 17 (2014).

[27] Anton Kuznietsov, Balint Gyevnar, Cheng Wang, Steven Peters, and Stefano V Albrecht. 2024. Explainable AI for safe and trustworthy autonomous driving: a systematic review. *IEEE Transactions on Intelligent Transportation Systems* (2024).

[28] Dongyub Lee, Taesun Whang, Chanhee Lee, and Heuiseok Lim. 2023. Towards reliable and fluent large language models: Incorporating feedback learning loops in qa systems. *arXiv preprint arXiv:2309.06384* (2023).

[29] Ramasamy Mariappan. 2025. Extensive Review of Literature on Explainable AI (XAI) in Healthcare Applications. *Recent Advances in Computer Science and Communications* 18, 1 (2025), E200324228159.

[30] Fuseini Mumuni and Alhassan Mumuni. 2025. Explainable artificial intelligence (XAI): from inherent explainability to large language models. *arXiv preprint arXiv:2501.09967* (2025).

[31] Bilal Nasir and Ijaz Hussain. 2024. Automating Agile Workflows: The Role of Multi-Agent LLMs in Modern Software Engineering. (2024).

[32] Mikel K Ngueajio, Saurav Aryal, Marcellin Atemkeng, Gloria Washington, and Danda Rawat. 2025. Decoding Fake News and Hate Speech: A Survey of Explainable AI Techniques. *Comput. Surveys* 57, 7 (2025), 1–37.

[33] Dorian Quelle and Alexandre Bovet. 2024. The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence* 7 (2024), 1341697.

[34] Jeba Rezwana and Mary Lou Maher. 2021. COFI: A Framework for Modeling Interaction in Human-AI Co-Creative Systems.. In *ICCC*. 444–448.

[35] Md Main Uddin Rony, Md Mahfuzul Haque, Mohammad Ali, Ahmed Shatil Alam, and Naeemul Hassan. 2024. Exploring the Potential of the Large Language Models (LLMs) in Identifying Misleading News Headlines. *arXiv preprint arXiv:2405.03153* (2024).

[36] Mike Schaekermann. 2020. Human-ai interaction in the presence of ambiguity: From deliberation-based labeling to ambiguity-aware ai. (2020).

[37] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–19.

[38] Vincent Schmitt, Luis Fernando Villa-Arenas, Niklas Feldhus, Julia Meyer, Rainer P Spang, and Sebastian Möller. 2024. The role of explainability in collaborative human-ai disinformation detection. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2157–2174.

[39] Tal Schuster, Darsh J Shah, Yun Jie Serene Yeo, Daniel Filizzola, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. *arXiv preprint arXiv:1908.05267* (2019).

[40] Yoo Yeon Sung, Jordan Boyd-Graber, and Naeemul Hassan. 2023. Not all fake news is written: A dataset and analysis of misleading video headlines. *arXiv preprint arXiv:2310.13859* (2023).

[41] Michael Twidale, David Randall, and Richard Bentley. 1994. Situated evaluation for cooperative systems. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*. 441–452.

[42] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics* 12 (2024), 39–57.

[43] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. 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*. 1–20.

A APPENDIX

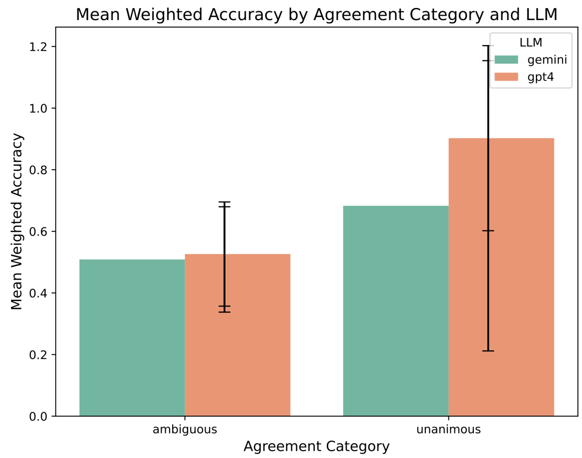

Figure 1: Weighted accuracy of Gemini and GPT-4 across unanimous and ambiguous cases. Error bars show ±1 SD. Both models perform better with unanimous labels, though GPT-4 consistently outperforms Gemini.

Table 1: Example headline with human and LLM labels and explanations.

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.

Table 2: 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

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 3: 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.

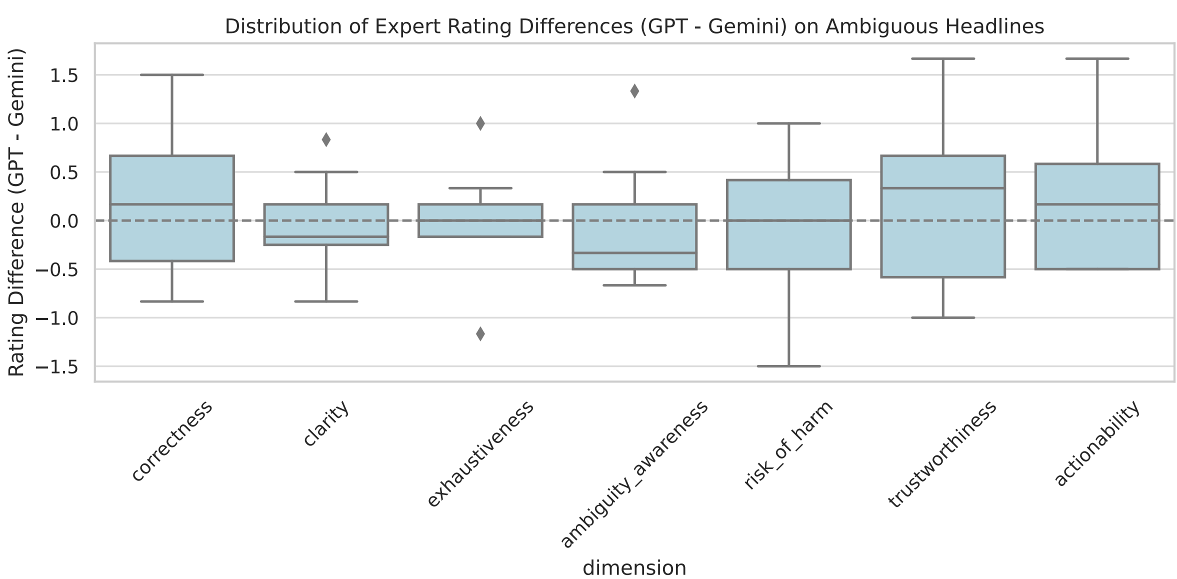

Figure 2: Distribution of difference scores (GPT - Gemini) in expert ratings across seven explanation quality dimensions for ambiguous headlines. Values above zero indicate higher GPT scores; values below zero indicate higher Gemini scores. Most differences cluster around zero, suggesting no systematic model advantage.