Exploring the Potential of the Large Language Models (LLMs) in Identifying Misleading News Headlines

Md Main Uddin Rony
University of Maryland, College Park

Md Mahfuzul Haque
University of Maryland, College Park

Mohammad Ali
University of Maryland, College Park

Ahmed Shatil Alam
University of Oklahoma

Naeemul Hassan
University of Maryland, College Park

ABSTRACT

In the digital age, the prevalence of misleading news headlines poses a significant challenge to information integrity, necessitating robust detection mechanisms. This study explores the efficacy of Large Language Models (LLMs) in identifying misleading versus non-misleading news headlines. Utilizing a dataset of 60 articles, sourced from both reputable and questionable outlets across health, science & tech, and business domains, we employ three LLMs—ChatGPT-3.5, ChatGPT-4, and Gemini—for classification. Our analysis reveals significant variance in model performance, with ChatGPT-4 demonstrating superior accuracy, especially in cases with unanimous annotator agreement on misleading headlines. The study emphasizes the importance of human-centered evaluation in developing LLMs that can navigate the complexities of misinformation detection, aligning technical proficiency with nuanced human judgment. Our findings contribute to the discourse on AI ethics, emphasizing the need for models that are not only technically advanced but also ethically aligned and sensitive to the subtleties of human interpretation.

1 INTRODUCTION

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 [7]; however, they can still receive misleading information if these headlines do not accurately represent the content. 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.

If the headline is misleading, it may cause a wrong impression, leading to uninformed decision-making [6]. Addressing the issue of Misleading News Headlines is critical to rebuilding trust in journalism and combating misinformation. Manual evaluations, while effective, are impractical due to the sheer volume and speed of news dissemination, necessitating automated solutions. However, constructing such systems presents challenges, particularly the need for extensive, high-quality data. Large-scale, representative datasets are essential for training robust machine learning models. This complexity underscores the significance of leveraging advanced techniques like Large Language Models (LLMs) to detect and classify misleading headlines accurately, ultimately enhancing journalism's credibility and its ability to counteract misinformation.

Recent advances in natural language processing have led to powerful Large Language Models (LLMs) capable of understanding complex languages intricately [8]. These LLMs have been successfully applied to identify and rectify vaccine misinformation, showcasing their potential in public health communication and information validation [3]. However, it's essential to acknowledge that misleading headlines differ from other forms of misinformation. Misleading headlines often straddle a fine line, potentially presenting skewed or exaggerated information without being entirely false. This distinct nature adds complexity to the task of utilizing LLMs to detect and address misleading headlines accurately. In light of this gray area, designing detection mechanisms can be more complex, resulting in the following research question:

[1] https://tinyurl.com/misleading-headline-example1

1st HEAL Workshop at CHI Conference on Human Factors in Computing Systems, May 12, Honolulu, HI, USA
2024.

RQ: To what extent can Large Language Models accurately identify headlines as misleading?

2 RELATED WORK

2.1 Overview of Misleading Headline Detection

Misleading headlines create a disconnect between the title and the article's content, leading to potential misinterpretation by readers. These headlines may present overrated, false, or unsupported information, aiming to attract attention or drive web traffic through exaggerated or sensational content [2, 17]. They often leverage emotional language, biasing readers even before they engage with the article, and may omit key information or emphasize less relevant details, leading to confusion and misinformation [6, 15]. The challenge lies in the headlines' ability to reinforce existing beliefs, making misinformation appear more credible and difficult to correct, thus significantly impacting reader understanding and opinion formation [1, 11]. A key challenge highlighted in the literature for automated misleading headline detection is the inadequacy of existing datasets and NLP methods in capturing incongruence between headlines and article content. This gap necessitates the development of more nuanced datasets and methodologies that go beyond simple agreement or disagreement models [2]. Additionally, the variability in dataset creation methods and the limitations of existing datasets in representing the full scope of misleading headlines pose significant challenges to developing effective automated detection systems [9, 10].

2.2 Overview of LLMs in NLP and Misinformation

The proliferation of Large Language Models (LLMs) in natural language processing represents a significant leap forward, enabling these models to grasp complex language structures with remarkable depth [8]. Demonstrated by their effectiveness in combating vaccine misinformation, LLMs hold promise for enhancing public health communication and ensuring the accuracy of information [3]. Nonetheless, the challenge of misleading headlines, which may convey skewed or exaggerated information without being outright false, underscores a unique dilemma. This subtlety complicates the use of LLMs for detecting and addressing misleading content, revealing a gap in their application. This nuanced challenge underlines the need for research into the capabilities of LLMs to discern and classify misleading headlines, highlighting a critical area for exploration.

3 METHOD

3.1 Data Collection

In our study, we collected news articles from 12 sources, categorized into reliable (e.g., ABC News, NY Times, Washington Post) and unreliable (e.g., Infowars, Lifezette) groups based on assessments from Media Bias/Fact Check (MBFC) [2], a third-party website that evaluates media source credibility. Our focus was on articles within the Health, Science & Tech, and Business domains. Three domain-knowledgeable annotators selected five articles from each domain from four sources, starting from March 31st, 2022, assessing if headlines were misleading by reviewing both the headline and the content. This process yielded a balanced preliminary dataset of 60 articles, comprising 30 misleading and 30 non-misleading headlines.

Table 1: Performance of LLMs in Detecting Misleading News Headlines

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

During annotation, each annotator independently reviewed 40 articles (20 misleading and 20 non-misleading) compiled by the others, without source identifiers to avoid bias. The review process involved three rounds of detailed examination to label articles as misleading or non-misleading, with annotations reflecting varying confidence levels. Consensus was reached on 18 articles being unanimously misleading, while at least two annotators agreed on 27 articles. Given our rigorous criterion that a headline is considered misleading if it could potentially mislead at least one reader, the final dataset consisted of 37 misleading and 23 non-misleading headlines.

3.2 LLM Evaluation

ChatGPT(version 3.5 and 4) and Gemini evaluated the collected headlines for labels and explanations, aiming to understand their capability to identify misleading headlines.

The headlines and relevant news content were submitted to LLMs for assessment. The LLMs determined if the headline is misleading and explain their decisions. The API requests were sent to the LLMs, which evaluated the news content's representation and provided a decision and an explanation.A sample request prompt would be as follows:

> *prompt=“Evaluate if the following headline is misleading based on the news content provided. Headline: [Your Headline Here] News Content: [Your News Content Here] Is this headline misleading? Please explain your decision."*

4 RESULTS

This study aimed to assess the capability of large language models (LLMs) — specifically ChatGPT-3.5, ChatGPT-4, and Gemini — to detect and explain misleading news headlines accurately. Employing a dataset of 60 news articles, where human annotators identified 37 as having misleading headlines, we explored how these LLMs could align with human judgment in identifying misinformation.

4.1 LLM's Classification Performance Analysis

This section presents the findings from evaluating three Large Language Models (LLMs) — ChatGPT-3.5, ChatGPT-4.0, and Gemini based on a binary classification task.

4.1.1 *LLMs' Overall Performance Analysis.*

Each LLM was assessed through precision, recall, f1-score, and overall accuracy metrics, providing insight into their effectiveness in addressing the RQ1.

ChatGPT-3.5 Performance The performance of ChatGPT-3.5 showed a high level of precision in identifying non-misleading

[2] https://mediabiasfactcheck.com/

headlines (precision: 1.00) but with a notably low recall rate (recall: 0.09), indicating a tendency to misclassify non-misleading headlines as misleading. Conversely, for misleading headlines, the model demonstrated a lower precision (0.46) but a perfect recall score (1.00), suggesting it was effective in identifying misleading headlines but with a considerable rate of false positives. The accuracy of ChatGPT-3.5 stood at 48%, with a macro-average f1-score of 0.39, indicating a moderate level of imbalance in its classification capability, skewed towards identifying misleading headlines.

ChatGPT-4.0 Performance ChatGPT-4.0 significantly improved over its predecessor, achieving an accuracy of 88%. It showed high precision and recall in identifying both misleading (precision: 0.95, recall: 0.77) and non-misleading headlines (precision: 0.85, recall: 0.97), reflected in a balanced f1-score for non-misleading (0.90) and misleading (0.85) headlines. The macro and weighted average f1-scores were both close to 0.88, illustrating a robust capability in accurately classifying headlines while maintaining a balanced performance across both classes.

Gemini Performance Gemini's performance presented a balanced approach between the two extremes of ChatGPT-3.5 and ChatGPT-4.0, with an overall accuracy of 67%. It demonstrated moderate precision and recall for non-misleading (precision: 0.68, recall: 0.79) and misleading headlines (precision: 0.65, recall: 0.50), leading to an f1-score of 0.73 and 0.57, respectively. The macro and weighted average f1-scores were 0.65 and 0.66, indicating a reasonable but not optimal balance in classification capability across the two categories.

4.1.2 LLM's Performance by Consensus Level. The efficacy of LLMs in identifying misleading content was examined in contexts of unanimous consensus by annotators versus mixed consensus (Majority and Minority Misleading) (See in Table 2).

Unanimous Consensus In scenarios where human annotators unanimously agreed on the nature of the headlines (either misleading or not misleading), ChatGPT-4 exhibited the highest performance, accurately classifying misleading headlines with an accuracy of 83.3% and non-misleading headlines with 95.7%. Gemini followed with 61.1% accuracy for misleading and 73.9% for non-misleading headlines. ChatGPT-3.5 showed a topmost accuracy, with 94.4% for misleading but performed poorly for non-misleading headlines with 8.7% accuracy. These results indicate a potential alignment between advanced LLM judgments and unanimous human consensus.

Mixed Consensus (Majority & Minority Misleading)

- Majority Misleading: When a majority (but not all) of the annotators identified headlines as misleading, ChatGPT-4's performance significantly decreased to 33.33% accuracy for misleading headlines. While Gemini experienced a more pronounced drop to 22.2%, ChatGPT-3.5 demonstrated a better performance with an accuracy rating of 88.9%, which is generally due to the tendency to misclassify non-misleading headlines as misleading. The results of this study suggest that there may be challenges in cases where there is less clear-cut human agreement.
- Minority Misleading: For headlines deemed misleading by a minority of annotators, ChatGPT-4's accuracy was 20%. Although Gemini exhibited the same accuracy as ChatGPT-4, ChatGPT-3.5 performed significantly better (90%) than its counterpart models, which underscores the difficulty LLMs have when there is a lack of strong human consensus.

Table 2: LLMs' Performance by Human Consensus Level

Consensus level	Model	Is Misleading?	# of Headlines
Unanimous Not Misleading	Gemini	Yes	6
		No	17
	ChatGPT-4	Yes	1
		No	22
	ChatGPT-3.5	Yes	21
		No	2
Minority Misleading	Gemini	Yes	2
		No	8
	ChatGPT-4	Yes	2
		No	8
	ChatGPT-3.5	Yes	9
		No	1
Majority Misleading	Gemini	Yes	2
		No	7
	ChatGPT-4	Yes	3
		No	6
	ChatGPT-3.5	Yes	8
		No	1
Unanimous Misleading	Gemini	Yes	11
		No	7
	ChatGPT-4	Yes	15
		No	3
	ChatGPT-3.5	Yes	17
		No	1

5 DISCUSSION

The evaluation of Large Language Models (LLMs) in distinguishing misleading news headlines reveals essential insights into the intersection of artificial intelligence and media integrity. This discussion delves into the implications of our findings within the broader context of human-centered evaluation and auditing methods for LLMs, highlighting the nuanced role these models play in supporting stakeholders across the digital information landscape.

5.1 Integrating Human-Centered Evaluation in LLM Auditing

As presented in our findings, the exploration of the effectiveness of large language models (LLMs) in discerning misleading news headlines emphasizes the imperative for incorporating human-centered evaluation and auditing frameworks. This approach not only benchmarks the performance of LLMs against human judgment but also aligns with the broader discourse on enhancing AI interpretability and reliability in media contexts [4].

5.2 LLM Performance and Human Consensus

5.2.1 Alignment with Unanimous Consensus. The better performance of ChatGPT-4 in instances of unanimous human consensus on misleading headlines highlights the advancements in AI's capability to parallel human reasoning in clear-cut scenarios. This observation resonates with the literature emphasizing the need for AI systems to understand and replicate human-like judgment in tasks requiring nuanced interpretation [13]. Such alignment is crucial for stakeholders, including media professionals and content moderators, who rely on AI to filter through vast amounts of data for potential misinformation.

5.3 Navigating Mixed Consensus

The nuanced challenge presented by mixed human consensus highlights a frontier in AI development. The differential performance of LLMs, particularly in majority and minority misleading scenarios, reflects the complexity of human cognition and the subjective nature of misinformation. This observation aligns with the AI ethics community's push for models that are not just technically advanced but also attuned to the nuances of human thinking and ethical concerns [12, 16].

5.4 Implications for Stakeholders

The practical implications of these findings are manifold. For journalists and media outlets, the deployment of LLMs that accurately identify misleading headlines could represent a significant step forward in maintaining informational integrity. For developers and AI researchers, our study highlights the importance of embedding human-centered design principles in the development of LLMs, ensuring these tools are both effective and ethically aligned with societal norms [14].

Moreover, for policymakers and regulators, understanding the capabilities and limitations of LLMs in identifying misinformation is crucial for crafting guidelines that promote responsible AI use in journalism and beyond. This aligns with ongoing discussions about the regulatory frameworks necessary to govern AI's application in sensitive societal domains [5].

5.5 Future Research Direction

Future research should aim to bridge the gap between LLM performance and the diverse ranges of human judgment, particularly in ambiguous or controversial scenarios. This includes investigating methodologies for incorporating ethical reasoning and bias recognition into LLM training processes. Additionally, expanding the scope of LLM training to encompass multimodal content could enhance their applicability across various media formats, offering a more holistic approach to misinformation detection.

A critical area for future exploration is the examination of explanations generated by LLMs in identifying misleading headlines and how these explanations align with human rationale. Understanding the logic and reasoning behind LLM decisions is essential for improving their reliability and trustworthiness. Analyzing LLM-generated explanations can provide insights into the models' interpretive processes, identifying areas where they may diverge from human thought patterns. This line of inquiry not only contributes to the development of more sophisticated and human-like LLMs but also supports the creation of AI systems whose decision-making processes are transparent, explainable, and, most importantly, aligned with ethical standards and societal expectations.

6 CONCLUSION

Our investigation into the capabilities of Large Language Models (LLMs) to identify misleading news headlines highlights the potential and challenges inherent in aligning AI with human judgment and ethical considerations. The study reveals that while models like ChatGPT-4 show promise in closely mirroring human decisions, particularly in clear-cut cases, discrepancies in performance across varying levels of human consensus highlight the complexity of misleading headline detection. The findings advocate for a human-centered approach in the development and evaluation of LLMs, emphasizing the need for models that are not only technically adept but also sensitive to the nuances of human ethics and reasoning. Future research directions, including examining LLM-generated explanations and expanding training to multimodal content, promise to further bridge the gap between AI and human judgment, paving the way for more reliable, ethical, and effective tools in combating misinformation.

REFERENCES

[1] 2015. The psychology of misinformation. *Australasian science* 36, 2 (2015), 21–.

[2] 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.

[3] Giovanna Deiana, Marco Dettori, Antonella Arghittu, Antonio Azara, Giovanni Gabutti, and Paolo Castiglia. 2023. Artificial Intelligence and Public Health: Evaluating ChatGPT Responses to Vaccination Myths and Misconceptions. *Vaccines* 11, 7 (2023), 1217.

[4] Nicholas Diakopoulos and Michael Koliska. 2017. Algorithmic transparency in the news media. *Digital journalism* 5, 7 (2017), 809–828.

[5] Virginia Dignum. 2019. *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Vol. 2156. Springer.

[6] Ullrich KH Ecker, Stephan Lewandowsky, Ee Pin Chang, and Rekha Pillai. 2014. The effects of subtle misinformation in news headlines. *Journal of experimental psychology: applied* 20, 4 (2014), 323.

[7] 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.

[8] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *Comput. Surveys* 56, 2 (2023), 1–40.

[9] Rahul Mishra, Piyush Yadav, Remi Calizzano, and Markus Leippold. 2020. MuSeM: Detecting incongruent news headlines using mutual attentive semantic matching. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 709–716.

[10] Kunwoo Park, Taegyun Kim, Seunghyun Yoon, Meeyoung Cha, and Kyomin Jung. 2020. 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* (2020), 229–252.

[11] Michal Piksa, Karolina Noworyta, Jan Piasecki, Pawel Gwiazdzinski, Aleksander B Gundersen, Jonas Kunst, and Rafal Rygula. 2022. Cognitive Processes and Personality Traits Underlying Four Phenotypes of Susceptibility to (Mis) Information. *Frontiers in Psychiatry* 13 (2022), 1142.

[12] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson, et al. 2019. Machine behaviour. *Nature* 568, 7753 (2019), 477–486.

[13] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.

[14] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.

[15] Wei Wei and Xiaojun Wan. 2017. Learning to identify ambiguous and misleading news headlines. *arXiv preprint arXiv:1705.06031* (2017).

[16] Jess Whittlestone, Rune Nyrup, Anna Alexandrova, and Stephen Cave. 2019. The role and limits of principles in AI ethics: Towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 195–200.

[17] Seunghyun Yoon, Kunwoo Park, Joongbo Shin, Hongjun Lim, Seungpil Won, Meeyoung Cha, and Kyomin Jung. 2019. Detecting incongruity between news headline and body text via a deep hierarchical encoder. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 791–800.