Decoding News Narratives: A Critical Analysis of Large Language Models in Framing Bias Detection

Anonymous ACL submission

Abstract

This work contributes to the expanding research on the applicability of LLMs in social sciences by examining the performance of GPT-3.5 Turbo, GPT-4, and Flan-T5 models in detecting framing bias in news headlines through zero-shot, few-shot, and explainable prompting methods. A key insight from our evaluation is the notable efficacy of explainable prompting in enhancing the reliability of these models, highlighting the importance of explainable settings for social science research on framing bias. GPT-4, in particular, demonstrated enhanced performance in few-shot scenarios when presented with a range of relevant, in-domain examples. FLAN-T5's poor performance indicates that smaller models may require additional task-specific fine-tuning for identifying framing bias detection. Our study also found that models, particularly GPT-4, often misinterpret emotional language as an indicator of framing bias, underscoring the challenge of distinguishing between reporting genuine emotional expression and intentionally use framing bias in news headlines. We further evaluated the models on two subsets of headlines where the presence or absence of framing bias was either clear-cut or more contested, with the results suggesting that these models' can be useful in flagging potential annotation inaccuracies within existing or new datasets. Finally, the study evaluates the models in real-world conditions ("in the wild"), moving beyond the initial dataset focused on U.S. Gun Violence, assessing the models' performance on framed headlines covering a broad range of topics[1].

1 Introduction

In today's digital age, the proliferation of news sources and the rapid dissemination of information highlight the essential need for unbiased reporting. Framing bias, which manipulates news content to potentially shift public perception, presents a formidable obstacle in maintaining a well-informed and unbiased public sphere (Binotto and Bruno, 2018). This form of bias not only influences how events are portrayed but also impacts public attitudes and policy decisions. For instance, the reporting on a government policy change might be emotionally charged in one headline as "Government's heartless cutbacks leave thousands without essential services", whereas another might offer a more neutral perspective: "Government announces reduction in funding for public services".

The concept of framing has received extensive scrutiny in the social sciences, identified as a technique that selectively highlights certain realities to shape audience perceptions and reactions (Goffman, 1974; Entman, 1993). While its significant effect on public communication and media is well-documented, the analysis of framing has historically been challenged by its fragmented nature and reliance on manual, small-scale studies. This situation underscores the necessity for new methods capable of dissecting the complexities of framing in media comprehensively. Amid this landscape, with an overwhelming array of news sources and headlines, it becomes crucial to determine whether they convey biased messages, intentionally or inadvertently. This determination is not merely academic but essential for preserving the integrity of public discourse.

The introduction of large language models (LLMs) opens up novel pathways for identifying framing bias, especially in news headlines. The accessibility and advanced capabilities of models, particularly those within the GPT series, have garnered attention from professionals across various fields, including media analysis and political science, seeking to navigate the extensive corpus of available information. However, despite their impressive success across many applications, the reliability of these models in executing nuanced tasks

[1]All data created in this study will be made publicly available.

like detecting framing bias remains an open question. In this study, we address this gap by providing an in-depth investigation into the strengths and shortcomings of state-of-the-art natural language processing (NLP) models in identifying framing bias. We perform a comprehensive examination of the performance of two top-tier NLP models, GPT-4 and GPT-3.5, alongside the open-source FLAN-T5 model (Wei et al., 2022), which has a reasonable performance across different NLP tasks.

Our findings reveal interesting patterns: while the pre-trained FLAN-T5 struggles with this task, the GPT models show promising results. Our results suggest that prompting the models to explain their decisions results in more reliable performances across different settings. A standout discovery is the enhanced reliability of predictions when models are prompted to explain their reasoning, a technique that also narrows the variability in outcomes across different example sets. To test the real-world applicability of these models, apart from a standard dataset that is annotated by experts, we automatically collect a new evaluation set containing framed headlines from different domains. The GPT models also accurately identifying the majority of framed titles in this dataset.

Our analysis suggests that while the examined GPT models achieve high scores on the examined evaluation sets, there are still areas that they struggle. As an example, GPT-4 often mistook emotional language for framing bias. This insight points to the need for developing more nuanced test sets that can present more complex challenges for evaluation. Additionally, our experiments suggest a potential use case for these models beyond mere bias detection. When GPT models, particularly in a more reliable prompting setting, converge on a prediction that deviates from the established gold standard, it may hint at underlying annotation inaccuracies within the dataset. Our findings suggest that a future direction involves constructing more challenging datasets that mirror the real-world complexities of framing bias, as well as enhancing the performance of smaller and open-source models on this challenging task.

2 Related Work

2.1 Automatic Framing Detection

Various NLP methodologies have been used to automate framing bias detection. Early studies in framing detection primarily utilized Topic Modeling, Structural Topic Modeling, and Hierarchical Topic Modeling to uncover themes and topics within large datasets (DiMaggio et al., 2013; Nguyen et al., 2015; Gilardi et al., 2021). These methods, while effective in identifying "what" is being discussed, often fall short in revealing "how" information is framed. Latent Dirichlet Allocation (LDA) Topic Modeling, for instance, served as a starting point for creating lists of frames deductively in tools like the one presented by Bhatia et al. (2021) for computational framing analysis. However, as noted by Ali and Hassan (2022), the emphasis in such approaches remains on detecting topics rather than the nuanced framing of those topics. The focus on topics also derives from the connections between agenda setting and framing strategies in computational social sciences, with studies analysing these two phenomena together (Field et al., 2018). Further analyses have included pragmatics cues, examining how specific word choices, like the use of "again" in "Again, Dozens of Refugees Drowned", subtly influence reader perception (Yu, 2022). This shift towards granular analysis is complemented by advanced models, including Neural Network and deep learning techniques, which offer refined tools for detecting framing nuances (Burscher et al., 2016; Card et al., 2015; Liu et al., 2019; Mendelsohn et al., 2021). Tourni et al. (2021) demonstrated that combining transformers models for processing news headlines with residual network models to process news lead images could lead to accurate framing bias detection. Similarly, Naderi and Hirst (2017) explored the use of various deep neural networks, such as LSTMs, BiLSTMs, and GRUs, for frame prediction at the sentence level using the Media Frame Corpus (MFC) (Card et al., 2015). Building on this foundation, Liu et al. (2019) and Akyürek et al. (2020) fine-tuned BERT (Devlin et al., 2019) to predict frames in news headlines. Their work resulted in the creation of the Gun Violence Frame Corpus (GVFC), a benchmark dataset for framing analysis which will be further discussed in section 2.1.1. More recently, in a novel approach to automatic framing detection, Lai et al. (2022) introduced an unsupervised learning method leveraging Wikipedia's category system to identify frames within news articles, using the GVFC to evaluate this method.

2.1.1 Datasets

Media Frame Corpus (MFC) MFC (Card et al., 2015) is a collection of annotated U.S. newspaper articles on topics like immigration, smoking, and same-sex marriage, analyzed for framing. Utilising the Policy Frames Codebook (PFC) by Boydstun et al. (2014), the MFC adopts 14 frame dimensions such as "security and defense" and "cultural identity" for categorising policy discourse. Despite achieving an inter-coder reliability (ICR) of 0.60, critiques, particularly from Ali and Hassan (2022), argue that the PFC's broad dimensions conflate topics with frames, potentially missing nuanced strategic framing. Moreover, MFC categorises content into wide-ranging dimensions that might not always precisely capture the specific framing intended by a news headline. This categorisation can make it difficult to directly identify whether and how a headline is framed without a deeper, nuanced analysis.

Gun Violence Frame Corpus (GVFC) Another significant dataset in the field of framing analysis is the Gun Violence Frame Corpus (GVFC), introduced by Liu et al. (2019). This dataset concentrates on the issue of Gun Violence in the U.S. The creation process began with defining nine distinct "frames" related to the topic, drawing from existing literature and a preliminary data analysis. A specialised codebook was then developed, serving as a training tool for annotators along with annotation guidelines. GVFC is made of 2990 news headlines, with 1,300 headlines specific to the issue of Gun Violence in the United States. All the headlines are coded to have a primary frame, while only 319 have 2 frames. For instance, the headline "It's Time to Hand the Mic to Gun Owners" is annotated with "Public opinion" as the first frame and "2nd Amendment" as the second frame. Similarly, "Trevor Noah: The Second Amendment Is Not Intended for Black People" is annotated with "2nd Amendment" and "Race/Ethnicity" frames Liu et al., 2019, p. 507.

Non-English Data Expanding the scope of framing analysis to non-English content, Akyürek et al. (2020) introduced a multilingual extension of the Gun Violence Frame Corpus, which encompasses news headlines in German, Turkish, and Arabic, focusing on U.S. gun violence. This extension involved training two native speakers per language to annotate headlines—350 in German, 200 in Turkish, and 210 in Arabic.

Piskorski et al. (2023b) presented an annotated dataset made of articles spanning though nine languages: English, French, German, Georgian, Greek, Italian, Polish, Russian, and Spanish. This dataset addresses a variety of topics including the COVID-19 pandemic, abortion-related legislation, migration, Russo-Ukrainian war, and various parliamentary elections. The annotation process made use of the PFC codebook, using the 15 dimensions as frames (Piskorski et al., 2023a).

While our study focuses on the English language, exploring framing bias detection in less resourced languages presents a promising future direction. By establishing a foundation in English, we lay the groundwork for future research to investigate the strengths and shortcomings of framing bias detection in languages beyond English.

2.2 Evaluating LLMs in Social Science

Applying Large Language Models to social science tasks, such as evaluating sociability (Choi et al., 2023), morality (Abdulhai et al., 2023), and controversial issues and bias (Sun et al., 2023), has received increasing interest, showcasing a wide range of strengths and limitations unique to each task. This diversity stems from the specific challenges and nuances of social phenomena. Although LLMs excel in generating and understanding human-like text, the complex requirements of social science tasks necessitate a detailed, task-specific examination of their performance and reliability.

In this work, we contribute to the expanding research on the applicability of LLMs in social sciences by specifically investigating their reliability in detecting framing bias. This exploration not only aims to identify the strengths and shortcomings of state-of-the-art models in a crucial area of framing analysis but also to extend our comprehension of how these models can be optimised for the nuanced task of framing detection.

3 Experimental Setup

3.1 Data

For our evaluation, we selected the Gun Violence Framing Corpus (GVFC) dataset (Liu et al., 2019), motivated by its comprehensive coverage of U.S. Gun Violence framing through 2900 annotated newspaper headlines as well as the high ICR met in the annotation process.[2] The dataset's annotations identify whether each headline reflects any of nine critical aspects of gun violence framing: gun rights, gun control, politics, mental health, public/school safety, race/ethnicity, public opinion, social/cultural issues, and economic consequences. In our study, headlines tagged with any of these framing aspects were categorised as framed, while all others were classified as not framed.

3.2 Models

To assess the capabilities of state-of-the-art NLP models in identifying framing bias, our study focuses on two primary categories of models: the GPT-4[3] (OpenAI, 2023) and GPT-3.5-Turbo[4] (Ye et al., 2023), known for their state-of-the-art performance across a wide range of NLP benchmarks, and the open-source FLAN-T5 (Wei et al., 2022), which, despite its accessibility, demonstrates significant efficacy in various tasks (Chung et al., 2022). Additionally, we examined various sizes of the FLAN-T5 model, i.e., small (77M parameters), base (248M parameters), and large (783M parameters), to explore how model scale influences bias detection capabilities.

Our evaluation methodology employs a uniform series of prompts applied across all models under three distinct experimental conditions: (1) a zero-shot setting, probing the models' inherent knowledge on framing bias detection; (2) a few-shot scenario, wherein models are provided with a small set of examples to inform their bias detection process; and (3) an explainability-focused approach, where models are asked to explain their rationale behind their decisions aside from label prediction.

Our evaluation framework is designed to thoroughly evaluate the models' adeptness at navigating tasks without being finetuned on framing bias datasets. This aspect is particularly important within the domain of social science research, where specialised training datasets are often scarce.

3.2.1 Zero-Shot Prompting

In zero-shot settings, we deploy two distinct approaches to evaluate the models' effectiveness in detecting framing bias. The first approach involves presenting the models with a straightforward task: determining if a headline is framed, without any additional context or examples.[5]

In the enhanced second zero-shot setting, we further add a specific definition to the input prompt to guide the task of framing detection. We use the "a communication strategy often used in journalism and political language, where certain aspects of an issue are highlighted while others are minimised or ignored, thereby promoting a particular interpretation of that issue" definition for this purpose.[6] The model is then tasked with deciding if a given claim is framed based on this definition. This additional guidance aims to refine the models' analytical lens, providing a clearer framework for identifying and evaluating framing within texts.

3.2.2 Few-Shot Prompting

In few-shot experiments, which build upon the zero-shot setting, we provide specific examples of both framed and not-framed headlines in the input prompt. These experiments are designed to investigate different factors influencing the model's performance with example-based prompts:

Examining the Impact of Example Quantity: To understand how the number of examples affects the model's accuracy, we evaluate few-shot models in two different configurations. The first configuration includes a minimal set of two examples, one framed and one not framed. This setup helps us measure the baseline impact of including examples on model performance. The second configuration includes eight examples, from which four examples are framed headlines.

Assessing the Relevance of Examples: We assess four scenarios to determine how the relevance of examples to the headline topics affects the model's ability to detect framing:

- **Focused in-domain**: In this setting, all the eight examples are relevant to gun violence, with a particular focus where all four framed headlines specifically address one aspect of gun violence news, which is health.

[2] In our experiments, we have excluded the headlines that are not relevant to Gun Violence. Furthermore, we have excluded 22 relevant headlines in order to use them for Few-Shot in domain prompting, leaving 2594 relevant headlines for our analysis.

[3] GPT-4-0613

[4] GPT-3.5-turbo-0613

[5] This involves using the "Decide whether this claim is framed:" prompt for GPT models, and "Is this claim framed? OPTIONS Yes | No" for FLAN-T5. We chose the FLAN-T5 model's prompt to ensure it aligns with the instructions that this model encountered during pretraining.

[6] This definition is grounded on the existing definitions of framing by Goffman (1974); Entman (1993, 2007).

- **Varied in-Domain**: In this setting framed examples cover four diverse aspects of gun violence[7] to test the model's adaptability to a range of in-domain cues.

- **Cross-Domain**: To evaluate the model's generalisation skills across topics, we use examples from completely different domains, such as immigration, in the few-shot prompts.

- **Mixed Domain**: Combining in-domain and cross-domain examples, this scenario includes two framed instances related to gun violence and two from unrelated areas.

3.2.3 Explainable Prompting

To enhance our analysis, we revisited both the zero-shot and few-shot settings, introducing an additional requirement: the model must not only determine if a claim is framed but also provide an explanation for its decision. To achieve this, we appended the instruction "then give an explanation for your response" to the prompts.

4 LLMs' Reliability in Detecting Framing Bias

The evaluation results for GPT and FLAN-T5 models across different settings are presented in Table 1 and Table 2, respectively. We can summarise the notable findings of the results of as follows:

Explainable Prompting Enhances Reliability: Explainable prompting consistently yielded more reliable outcomes in both zero- and few-shot variations, as shown by the reduced variance in accuracy and F_1 scores. This underscores the importance of explainable settings for social science research on framing bias.

Optimal Performance in Few-Shot with Diverse Examples: The GPT-4 model achieved the highest accuracy and F_1 scores in few-shot scenarios with a wide range of in-domain examples (8 varied in-domain). This finding suggests that incorporating diverse, relevant in-domain examples significantly improves model performance when the news headline domain is known.

Tendency Towards Framing Bias Classification: In explainable settings, both GPT-3.5 and GPT-4 models exhibit a higher tendency to classify headlines as framed. This is indicated by higher F_1 scores for framed headline detection alongside lower overall accuracy, suggesting a bias towards classifying headlines as framed when prompted to identify framing.

Challenges with Cross-Domain Examples: F_1 scores dropped in cross-domain settings without explainable prompts, falling below those in zero-shot settings with definitions. This highlights the challenges of applying advanced models to new or mixed domains, where headline topics vary widely. Therefore, for new domain applications, we suggest to either employ a zero-shot setting with task definition, or opting for explainable prompts when examples are used in the input prompts. However, for large-scale data analysis, opting for a zero-shot approach without explanations could be more efficient, albeit at the potential cost of reduced reliability compared to the explainable setting.

Inherent Challenge in Framing Bias Detection: Across both models, the highest F_1 scores observed were in the range of 60-70 points, highlighting the inherent difficulty of detecting framing bias. Framing detection requires understanding subtle language nuances, contextual cues, and the intended message framing, which are complex and context-dependent tasks. Furthermore, the performance between GPT-3.5 and GPT-4 did not significantly differ, particularly in more realistic use cases like zero-shot and cross-domain settings. Considering the higher cost of using GPT-4's API, GPT-3.5 might be a more cost-effective choice for framing bias detection in large scale analysis.

FLAN-T5 Performance Lag: The results from Table 2 show that FLAN-T5 models significantly underperform compared to GPT models, with the best F_1 score reaching only 51 points. This indicates that smaller models like FLAN-T5 may not yet be viable for complex tasks such as framing bias detection without additional task-specific fine-tuning or more extensive training data.

5 Analysis

In this section, we perform a set of additional analysis and experiments to identify potential biases and areas for future research enhancements.

The Impact of Emotional Language on Framing Bias Detection Our analysis of GPT-4's errors reveals a consistent pattern: the model frequently

[7] I.e., politics, public/school safety, race/ethnicity, and social/culture.

		GPT-3.5 Turbo				GPT-4			
				Explainable				Explainable	
			F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1
Acc.									
Zero-Shot	No Definition	20.48	52.70	61.43	53.28	64.96	58.40	63.75	60.41
	+Definition	51.55	59.14	59.03	58.79	65.36	63.84	64.64	60.99
Few-Shot	2 examples	26.83	54.16	**65.84**	61.68	58.25	65.84	64.23	64.92
	8 (focused in-domain)	40.72	58.25	65.32	61.57	64.50	63.30	66.44	63.96
	8 (varied in-domain)	46.36	60.22	64.40	60.79	**68.51**	65.38	**70.41**	**66.92**
	8 (cross domain)	41.22	57.67	60.76	62.08	59.42	**66.04**	59.09	64.61
	8 (mixed domain)	**56.93**	**63.49**	64.60	**62.99**	63.33	64.37	63.87	63.76

Table 1: Comparative performance of GPT models showing "zero-shot" results with and without task definition, "few-shot" results with 2 or 8 examples, and "Explainable" results when models explain predictions. "Acc." columns report overall accuracy, and "F_1" reports detection of framed headlines.

		FLAN-T5 Small				FLAN-T5 Base				FLAN-T5 Large			
				Explainable				Explainable				Explainable	
		F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.
Zero-Shot	No Definition	0.46	49.06	**17.96**	49.21	0	0	**25.86**	**52.74**	34.33	46.19	51.01	**49.06**
	+Definition	0.15	50.13	1.21	50.10	0	0	0.61	50.06	**42.27**	48.06	37.63	43.96
Few-Shot	2 examples	0.31	50.06	0.31	50.06	0	0	3.12	49.98	**42.27**	48.06	37.63	43.96
	8 (focused in-domain)	0.15	50.17	3.84	50.10	0	0	4.83	50.17	16.78	47.53	36.85	44.69
	8 (varied in-domain)	0	0	0.15	49.94	0	0	1.66	49.94	22.78	45.96	35.03	45.27
	8 (cross domain)	**5.30**	49.37	2.95	49.60	0	0	4.99	50.33	40.23	45.11	**51.64**	46.88
	8 (mixed domain)	0.31	**50.21**	1.51	50.02	0	0	3.84	50.13	34.39	44.58	44.05	43.43

Table 2: Comparative performance of FLAN-T5 models using different prompt configurations.

Headline: This list of lives lost in mass shootings grows longer. Here are their names. **Gold:** Not Framed, **GPT-4:** Framed, **GPT-3.5:** Not Framed **GPT-4 Explanation:** This claim is framed because it specifically focuses on the tragic aspect of lives lost in mass shootings, evoking empathy and sadness. While it's factual that people have died, the phrasing and focus frame the issue in a particular emotional light. **GPT-3.5 Explanation:** This claim is not framed. It simply states that the list of lives lost in mass shootings is growing longer and presents the names of the victims. It does not provide any particular perspective or agenda.
Headline: Pregnant woman who killed intruder in justified shooting now faces felony gun charge due to previous marijuana conviction. **Gold:** Not Framed, **GPT-4:** Framed, **GPT-3.5:** Not Framed **GPT-4 Explanation:** The claim is framed. It selectively emphasizes on the individual's status as a pregnant woman and the nature of the shooting as justified, while sidelining the legal issue of her previous marijuana conviction which is directly linked to the current felony charge. The framing encourages empathy for the person involved by prominently highlighting her pregnancy and the self-defense aspect, provoking an emotional response. **GPT-3.5 Explanation:** This claim appears to accurately represent the situation. It states that a pregnant woman killed an intruder in a justified shooting and is now facing a felony gun charge due to a previous marijuana conviction. The claim does not seem to contain any misleading or false information.
Headline: Gunman upset by divorce is found dead after manhunt. **Gold:** Not Framed, **GPT-4:** Framed, **GPT-3.5:** Not Framed **GPT-4 Explanation:** This claim can be considered as framed because it specifically points out the gunman's divorce as a possible motive. By doing so, the author is driving the audience's view to place emphasis on the gunman's personal issues, making them potentially more sympathetic to him. Other potential factors or the bigger context are not discussed here. **GPT-3.5 Explanation:** This claim appears to be accurately presented. It provides a straightforward account of events, stating that a gunman who was upset by a divorce was found dead after a manhunt. There is no apparent bias or manipulation in the framing of this claim.

Table 3: Examples of the impact of emotional language on GPT-4.

interprets emotional language as an indicator of framing bias, irrespective of the headline's actual framing. This tendency likely stems from the observation that emotionally charged language often accompanies framed headlines, evoking specific frames and eliciting emotional responses from readers, thereby influencing their perception of the news narrative (Valkenburg et al., 1999). However, this leads to a biased error in the model, mistakenly identifying expressions of emotion as indicative of framing bias. Such a pattern highlights the challenge of distinguishing between genuine emotional expression and intentional framing bias. Making this distinction is essential for enhancing the model's precision in analysing news narratives. Table 3 provides examples of such misclassifications by GPT-4 in the explainable zero-shot setting. On the other hand, GPT-3.5 seems to be less affected by the emotional language.

Clear vs. Contested Cases of Framing Annotations During our initial analysis, we observed

	GPT-3.5		GPT-4		Flan-T5					
					Small		Base		Large	
	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.
Clear	71.43	70.15	75.81	77.61	3.23	42.86	9.23	43.81	49.6	40.00
Contested	41.43	38.81	3.60	20.15	3.03	39.62	0	0	65.60	59.43

Table 4: Comparative performance of the GPT and FLAN-T5 models on the clear and contested framing subsets.

discrepancies in some data annotations, with some annotations being potentially incorrect. Additionally, there were ambiguous cases where multiple labels could be justified based on different interpretations. In this regard, we manually examined half of the headlines in GVFC, i.e., 1300 headlines. Within this subset, we identified 134 contested annotations. Examples include "Live: Trump visits Pittsburgh after synagogue shooting" and "Shopify bans sale of certain firearms, accessories", both annotated as framed. However, not all 134 cases are necessarily wrong annotations. Some of them, such as "Thousands gather to honor victims of the mass shooting with tears, candlelight, and song", present more nuanced challenges in framing detection.[8]

Additionally, we selected another set of 134 headlines where the presence or absence of framing bias was more apparent. For instance, headlines such as "Two dead including shooter at Florida yoga studio" and "Parkland school shooter blames massacre on a 'demon' voice" serve as clear examples of non-framed and framed headlines that are also annotated correctly, respectively. We call this subset the "Clear" subset.

Table 4 reports the results of evaluating our models on these two distinct subsets. As expected, both GPT models exhibited strong performance on the Clear subset. However, they have a considerably low results on the contested subset, with GPT-4's F_1 score dropping to 3.6 points.

We have also calculated how often GPT-4 and GPT-3.5's predictions aligned within these subsets. These agreement ratios are shown in Table 5. We observe that while the models reached an agreement of about 59% on the contested subset, only 16% of these agreed-upon predictions matched the gold labels. This suggests that these models' can be useful in flagging potential annotation inaccuracies within existing or new datasets. As an example, both models identify the " Muslim Americans raise more than $200,000 for those affected by Pittsburgh synagogue shooting" as not framed while it is annotated as framed in the data.

[8]63.4% of headlines in this subset are annotated as framed in the dataset.

	Clear	Contested
Agreement	69.40	58.95
Agreement GL	58.20	16.45

Table 5: Clear vs. Contested Annotations: Percentages of agreement between GPT-3.5 and GPT-4's predictions, and with the Gold Label. The "Agreement" row indicates the percentage of headlines for which both models provided the same prediction. The "Agreement GL" row reflects the percentage of these agreements where the predicted label matches the gold label.

Evaluation in the Wild Until now, our results and analysis were centered on the GVFC dataset, which is limited to the narrow subject of gun violence within the U.S. To assess the models in a scenario closer to real-world conditions, where the topics of headlines are varied and unknown, we collected a new dataset from a website known for its emphasis on news framing.[9] This collection consists of 130 headlines spanning diverse subjects such as British weather, health, and European matters.

We assessed the models on this novel dataset employing the explainable few-shot setting with 8 cross-domain examples, which resulted in reliable results in Section 4. Table 6 shows the counts of headlines predicted as framed or not framed by each model. Given the dataset's origin from a website known for featuring framed headlines, we anticipated the majority of headlines to be framed. We observe that the GPT-4 model predominantly identified headlines as framed. However, considering GPT-4's tendency towards classifying inputs as framed, a promising avenue for future research involves developing a dataset specifically aimed at challenging the models with not-framed cases. This could include not-framed headlines that still incorporate emotional language, as discussed in the beginning of this section.

[9]https://newsframes.wordpress.com/category/headlines/

	Framed	Not-Framed
GPT-3.5 Turbo	84	38
GPT-4	108	22
Flan-T5 Small	1	129
Flan-T5 Base	0	130
Flan-T5 Large	32	98

Table 6: Evaluation in the wild: number of framed and not-framed predictions for each model.

6 Conclusions

This work advances our understanding of the performance of LLMs such as GPT-3.5 Turbo, GPT-4, and Flan-T5 (small, base, and large) in detecting framing bias within news headlines. Through a comprehensive experimental approach that included zero-shot, few-shot, and explainable prompting settings, we found that GPT-4 excels in few-shot scenarios with a variety of in-domain examples, highlighting its superior capability in recognising framing bias. However, GPT-4 also displayed a tendency to classify headlines as framed, suggesting a valuable line of research could involve testing these models on datasets containing non-framed headlines that potentially also incorporate emotional language, identified as another potential weakness. In fact, our research explores the impact of emotional language on framing bias detection, with GPT-3.5 showing resilience against emotional language.

The study also identifies a performance shortfall in Flan-T5, underscoring the challenges smaller models face in complex detection tasks without specific fine-tuning. Our findings indicate that F1 scores drop in cross-domain scenarios with non-explainable prompts, suggesting the effectiveness of zero-shot approaches with clear definitions or explainable prompts for new domain applications. The study further demonstrates the utility of these models in identifying potential annotation inaccuracies in new or already existing datasets.

7 Limitations

The findings of this study have to be seen in light of some limitations. For example, our evaluation focuses solely on English-language content, leaving space for further investigation on other languages to explore our findings' applicability to non-English contexts. This limitation suggests a need for further investigation into the performance of LLMs across different languages and cultural contexts to fully assess the potential use of these models in social science research for detecting framing bias and analysing media narratives.

Furthermore, two of the five models evaluated in our work are accessible only through OpenAI's API, which is closed-source and subject to changes over time. This could affect the reproducibility of our results with newer versions of API, and they may have their own limitations. Therefore, focusing on improving open-source models emerges as a critical pathway forward, ensuring broader accessibility and reproducibility in research.

References

Marwa Abdulhai, Gregory Serapio-Garcia, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques. 2023. Moral Foundations of Large Language Models. ArXiv:2310.15337 [cs].

Afra Feyza Akyürek, Lei Guo, Randa Elanwar, Prakash Ishwar, Margrit Betke, and Derry Tanti Wijaya. 2020. Multi-Label and Multilingual News Framing Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8614–8624, Online. Association for Computational Linguistics.

Mohammad Ali and Naeemul Hassan. 2022. A Survey of Computational Framing Analysis Approaches. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9335–9348, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vibhu Bhatia, Vidya Prasad Akavoor, Sejin Paik, Lei Guo, Mona Jalal, Alyssa Smith, David Assefa Tofu, Edward Edberg Halim, Yimeng Sun, Margrit Betke, Prakash Ishwar, and Derry Tanti Wijaya. 2021. OpenFraming: Open-sourced Tool for Computational Framing Analysis of Multilingual Data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 242–250, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marco Binotto and Marco Bruno. 2018. Spazi mediali delle migrazioni. framing e rappresentazioni del confine nell'informazione italiana. *Lingue e Linguaggi*, 25(0).

Amber E. Boydstun, Dallas Card, Justin Gross, Paul Resnick, and Noah A. Smith. 2014. Tracking the Development of Media Frames within and across Policy Issues.

Bjorn Burscher, Rens Vliegenthart, and Claes H. de Vreese. 2016. Frames beyond words: Applying cluster and sentiment analysis to news coverage of the nuclear power issue. *Social Science Computer Review*, 34(5):530–545.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The Media Frames Corpus: Annotations of Frames Across Issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs Understand Social Knowledge? Evaluating the Sociability of Large Language Models with SocKET Benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403, Singapore. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. ArXiv:2210.11416 [cs].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Paul DiMaggio, Manish Nag, and David Blei. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of u.s. government arts funding. *Poetics*, 41(6):570–606. Topic Models and the Cultural Sciences.

Robert M. Entman. 1993. Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication*, 43(4):51–58. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1460-2466.1993.tb01304.x.

Robert M. Entman. 2007. Framing Bias: Media in the Distribution of Power. *Journal of Communication*, 57(1):163–173.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and Agenda-setting in Russian News: a Computational Analysis of Intricate Political Strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580, Brussels, Belgium. Association for Computational Linguistics.

Fabrizio Gilardi, Charles R. Shipan, and Bruno Wüest. 2021. Policy diffusion: The issue-definition stage. *American Journal of Political Science*, 65(1):21–35.

Erving Goffman. 1974. *Frame Analysis: An Essay on the Organization of Experience*. Harper colophon books. Harvard University Press.

Sha Lai, Yanru Jiang, Lei Guo, Margrit Betke, Prakash Ishwar, and Derry Tanti Wijaya. 2022. An unsupervised approach to discover media frames. In *Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences*, pages 22–31, Marseille, France. European Language Resources Association.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting Frames in News Headlines and Its Application to Analyzing News Framing Trends Surrounding U.S. Gun Violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514, Hong Kong, China. Association for Computational Linguistics.

Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. Modeling Framing in Immigration Discourse on Social Media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263, Online. Association for Computational Linguistics.

Nona Naderi and Graeme Hirst. 2017. Classifying Frames at the Sentence Level in News Articles. In *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*, pages 536–542. Incoma Ltd. Shoumen, Bulgaria.

Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, and Kristina Miler. 2015. Tea party in the house: A hierarchical ideal point topic model and its application to republican legislators in the 112th congress. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1438–1448, Beijing, China. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Jakub Piskorski, Nicolas Stefanovitch, Valerie-Anne Bausier, Nicolo Faggiani, Jens Linge, Sopho Kharazi, Nikolaos Nikolaidis, Giulia Teodori, Bertrand De Longueville, Brian Doherty, et al. 2023a. News categorization, framing and persuasion techniques: Annotation guidelines. Technical report, Technical Report JRC-132862, European Commission Joint Research Centre.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023b. SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

David Sun, Artem Abzaliev, Hadas Kotek, Christopher Klein, Zidi Xiu, and Jason Williams. 2023. DELPHI: Data for Evaluating LLMs' Performance in Handling Controversial Issues. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 820–827, Singapore. Association for Computational Linguistics.

Isidora Tourni, Lei Guo, Taufiq Husada Daryanto, Fabian Zhafransyah, Edward Edberg Halim, Mona Jalal, Boqi Chen, Sha Lai, Hengchang Hu, Margrit Betke, Prakash Ishwar, and Derry Tanti Wijaya. 2021. Detecting Frames in News Headlines and Lead Images in U.S. Gun Violence Coverage. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4037–4050, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Patti M. Valkenburg, Holli A. Semetko, and Claes H. De Vreese. 1999. The effects of news frames on readers' thoughts and recall. *Communication Research*, 26(5):550–569.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models. ArXiv:2303.10420 [cs].

Qi Yu. 2022. "Again, Dozens of Refugees Drowned": A Computational Study of Political Framing Evoked by Presuppositions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 31–43, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.