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

Valeria Pastorino, Jasivan Alex Sivakumar, Nafise Sadat Moosavi
Department of Computer Science
University of Sheffield
United Kingdom
{vpastorino1|jasivakumar1|n.s.moosavi}@sheffield.ac.uk

Abstract

The growing complexity and diversity of news coverage have made framing analysis a crucial yet challenging task in computational social science. Traditional approaches, including manual annotation and fine-tuned models, remain limited by high annotation costs, domain specificity, and inconsistent generalisation. Instruction-based large language models (LLMs) offer a promising alternative, yet their reliability for framing analysis remains insufficiently understood. In this paper, we conduct a systematic evaluation of several LLMs, including GPT-3.5/4, FLAN-T5, and Llama 3, across zero-shot, few-shot, and explanation-based prompting settings. Focusing on domain shift and inherent annotation ambiguity, we show that model performance is highly sensitive to prompt design and prone to systematic errors on ambiguous cases. Although LLMs, particularly GPT-4, exhibit stronger cross-domain generalisation, they also display systematic biases, most notably a tendency to conflate emotional language with framing. To enable principled evaluation under real-world topic diversity, we introduce a new dataset of out-of-domain news headlines covering diverse subjects. Finally, by analysing agreement patterns across multiple models on existing framing datasets, we demonstrate that cross-model consensus provides a useful signal for identifying contested annotations, offering a practical approach to dataset auditing in low-resource settings.

Keywords: Framing, LLMs, Prompting, News Narratives

1. Introduction

In today's digital age, the rapid growth of news sources and the widespread dissemination of information have intensified the need for unbiased and transparent reporting. At the same time, news coverage is often shaped through framing, a communication strategy that selectively emphasizes certain aspects of an issue in order to influence public perception and promote particular interpretations (Binotto and Bruno, 2018). As a result, framing represents a persistent obstacle to maintaining a well-informed public audience. It affects not only how events are perceived and remembered, but also how they are evaluated, discussed, and translated into policy preferences. For instance, a government policy change may be framed as "Government's heartless cutbacks leave thousands without essential services" or, alternatively, as "Government announces reduction in funding for public services".

In social science, framing is understood as a mechanism for guiding audience interpretation through selective emphasis (Goffman, 1974; Entman, 1993). Building on this tradition, we consider a headline to be framed when it selectively emphasizes certain aspects of an event while downplaying others in order to steer readers toward a particular interpretation. This definition distinguishes framing from merely emotional or descriptive language and grounds our computational task in established communication theory.

Figure 1: Example of two different ways of framing the same news.

Despite its importance, empirical studies of framing have traditionally relied on small, manually annotated datasets (Baumer et al., 2015). While

Proceedings of The 3rd Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2026) @ LREC 2026, pages 17–28
16 May 2026, Palma, Mallorca (Spain).

supervised computational approaches have been proposed to scale such analyses, their performance often deteriorates under domain shift, limiting their applicability across diverse news contexts (Sinelnik and Hovy, 2024). As online media continues to expand in volume and scope, there is a growing need for scalable and reliable methods for framing analysis. Large language models (LLMs) offer a promising alternative to traditional approaches. Instruction-following, pre-trained LLMs can be adapted to new tasks with minimal supervision, reducing the cost of domain-specific annotation. However, despite their increasing use in social-science research, their reliability for detecting nuanced framing remains insufficiently understood.

This paper addresses this gap through a systematic investigation of instruction-based LLMs for news framing detection. We evaluate GPT-4, GPT-3.5 Turbo, Llama 3, and FLAN-T5 across zero-shot, few-shot, and explanation-based prompting settings, with an emphasis on robustness, bias, and failure modes. To enable a principled analysis of cross-domain generalisation, we first introduce a new dataset of real-world news headlines spanning diverse topics. Using this resource, we examine how reliably these models detect framing under domain shift and limited supervision, how their predictions vary with prompt design, and what systematic biases arise in practice.

Our analysis shows that although LLMs, especially GPT-4, exhibit stronger cross-domain generalisation than fine-tuned models, their predictions remain highly sensitive to prompt design and prone to systematic errors on inherently ambiguous cases. In particular, we identify a consistent tendency of GPT-4 to conflate emotional language with framing, leading to recurrent false positives. Finally, by analysing patterns of agreement and disagreement across models on existing framing datasets, we show that cross-model consensus provides a useful signal for identifying contested or potentially problematic annotations, offering a practical tool for dataset auditing.[1]

2. Related Work

2.1. Automatic Framing Detection

A wide range of NLP methods has been proposed for automatic framing detection. Early studies primarily relied on topic modeling approaches, including Topic Modeling, Structural Topic Modeling, and Hierarchical Topic Modeling, to uncover themes in large document collections (DiMaggio et al., 2013; Nguyen et al., 2015; Gilardi et al., 2021). While effective for identifying what is discussed, these methods often provide limited insight into how issues are 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 events. The focus on topics could also derive 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 transformer models for processing news headlines with residual network models to process news lead images could improve the accuracy of framing detection. Similarly, Naderi and Hirst (2017) explored the use of various 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 2.2.

In contrast to the supervised methodologies outlined above, which often struggle with generalisation due to their reliance on domain-specific training data (Ali and Hassan, 2022), our work explores an alternative approach using instruction-following language models, which potentially offer a more flexible and scalable solution for detecting framing in a broad array of news contexts.

2.2. 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, analysed 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

[1] For reproducibility and future research, our dataset is available https://github.com/vpastorino/ITW-dataset.

identity" for categorising policy discourse. Despite achieving an inter-coder reliability (ICR) of 0.60, critiques, particularly 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 (i.e., politics, economic, etc.) that might not always precisely capture the specific framing intended by a news headline (Ali and Hassan, 2022). This categorisation can make it difficult to directly identify if 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 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 2616 headlines specific to the issue of Gun Violence in the United States. All the in-domain headlines are coded to have a primary frame, while only 319 have two 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).

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 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).

2.3. Evaluating LLMs in Social Science

Applying LLMs 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 of the examined language models 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.

3. Experimental Setup

3.1. Data

For our evaluation, we select GVFC (Liu et al., 2019), motivated by its comprehensive coverage of U.S. Gun Violence framing as well as the high ICR met in the annotation process. In our experiments, we have excluded the headlines that are not relevant to Gun Violence and hence, are not annotated with framing information. 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. We consider headlines tagged with any of the above categories as framed, and all others as not framed. Further, we exclude 22 relevant headlines in order to use them for few-shot in domain prompting, leaving 2594 relevant headlines for our analysis. Of these, 1293 are framed and 1301 are not framed. Hence, the majority label is Not Framed and the majority baseline is 50.15%.

3.2. Models

To evaluate the performance of contemporary large language models in framing detection, we consider two widely used closed-source models: GPT-4[2] (OpenAI, 2023) and GPT-3.5-Turbo[3] (Ye et al., 2023). These models are known for their strong performance across a wide range of NLP tasks and their ability to follow complex instructions. We select them due to their widespread adoption

[2]GPT-4-0613. Temperature parameter set to 0.

[3]GPT-3.5-turbo-0613. Temperature parameter set to 0.

and practical accessibility for social science researchers, as they can be used via paid APIs without requiring specialised hardware or extensive technical setup.

In addition, we include two open-source alternatives: Llama 3 (Dubey et al., 2024) and FLAN-T5 (Wei et al., 2022), which have demonstrated competitive performance across diverse benchmarks (Chung et al., 2022). We evaluate the 8B version of Llama 3[4] and multiple variants of FLAN-T5, including small (77M parameters), base (248M parameters), and large (783M parameters), to examine the effect of model scale on framing detection.

All models are evaluated using a unified set of prompts under three experimental conditions: (1) a zero-shot setting, assessing models' baseline capabilities; (2) a few-shot setting, in which a small number of examples are provided; and (3) an explanation-based prompting setting, where models are asked to justify their predictions.

Our evaluation framework focuses on assessing model behaviour without task-specific fine-tuning, reflecting realistic conditions in social science research, where annotated datasets are often limited. This design allows us to examine the robustness and generalisability of LLMs for framing detection under low-supervision settings.

3.3. Zero-Shot Prompting

In the zero-shot setting, we evaluate model performance using two prompt variants. In the first, models are asked to determine whether a headline is framed without any additional context or examples.[5]

In the second variant, we augment the prompt with an explicit definition of framing to guide the model's decision. Specifically, framing is defined as "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".[6] Models are then asked to classify headlines according to this definition. This setting allows us to assess whether conceptual guidance improves framing detection.

[4]Meta-Llama-3-8B-Instruct. Temperature parameter set to 0.1.

[5]GPT and Llama prompt: "Decide whether this claim is framed".
FLAN-T5 prompt: "Is this claim framed? OPTIONS Yes | No". The FLAN-T5 prompt is designed to align with its instruction-based pre-training format.

[6]This definition is grounded in established accounts of framing (Goffman, 1974; Entman, 1993, 2007).

3.4. Few-Shot Prompting

In the few-shot setting, we extend the zero-shot setup by providing a small number of labeled examples of framed and non-framed headlines within the prompt. These experiments are designed to examine how example-based supervision influences model behaviour and performance.

Impact of Example Quantity To examine how the number of examples affects model performance, we evaluate two few-shot configurations. The first includes a minimal set of two GVFC examples, one framed and one not framed, serving as a baseline for assessing the effect of example-based prompting. The second includes eight examples, of which four are framed, allowing us to analyse the impact of increased supervision.

Relevance of Examples We further investigate how the relevance of examples to the test headlines, when available a priori, influences framing detection. We consider four settings:

- **Focused in-domain**: In this setting, all the eight examples are relevant to gun violence, with all four framed headlines addressing a single aspect: health.
- **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 performance on topics not known in advance, we use examples from completely different domains, such as immigration.
- **Mixed Domain**: Combining in-domain and cross-domain examples, this scenario includes two framed instances related to gun violence and two from unrelated areas.

3.5. Explanation-Based Prompting

To further analyse model behaviour, we revisit both zero-shot and few-shot settings by requiring models to provide an explicit rationale for their predictions. Specifically, we append the instruction "then give an explanation for your response" to the original prompts, prompting models to justify their framing decisions alongside label predictions.

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

		GPT-3.5 Turbo				GPT-4				Llama3 8B			
				Explainable				Explainable				Explainable	
		F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.
ZS	No Def.	20.48	52.70	61.43	53.28	64.96	58.40	63.75	60.41	51.73	57.19	58.59	55.81
	+Def.	51.55	59.14	59.03	58.79	65.36	63.84	64.64	60.99	52.66	53.55	57.42	50.35
FS	2Ex.	26.83	54.16	**65.84**	61.68	58.25	65.84	64.23	64.92	57.23	59.08	56.93	56.18
	8Focus.	40.72	58.25	65.32	61.57	64.50	63.30	66.44	63.96	52.29	61.47	55.39	60.12
	8Varied	46.36	60.22	64.40	60.79	**68.51**	65.38	**70.41**	**66.92**	**58.58**	61.94	**61.66**	59.08
	8Cross	41.22	57.67	60.76	62.08	59.42	**66.04**	59.09	64.61	47.84	**63.05**	53.40	**60.49**
	8Mixed	**56.93**	**63.49**	64.60	**62.99**	63.33	64.37	63.87	63.76	58.32	62.21	57.98	59.72

Table 1: Comparative performance of GPT and Llama3 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.

	GPT-3.5 Turbo				GPT-4				Llama3 8B			
			Explainable				Explainable				Explainable	
	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.
No Definition	20.48	52.70	61.43	53.28	64.96	58.40	63.75	60.41	51.73	57.19	58.59	55.81
Diff. Wording	25.39	53.55	59.04	54.64	47.40	60.56	45.55	57.76	51.65	58.44	56.68	56.18
+Definition	51.55	59.14	59.03	58.79	65.36	63.84	64.64	60.99	52.66	53.55	57.42	50.35
Diff. Wording	58.64	58.16	63.56	59.44	57.03	62.63	59.02	61.07	50.80	54.53	57.06	49.71

Table 2: Comparative performance of GPT-3.5 Turbo, GPT-4 and Llama3 in Zero-Shot settings with varied prompt wordings.

4. Results and Analysis

We organise our analysis around four central dimensions highlighted in the Introduction: reliability under prompt variation, generalisation under domain shift, systematic failure modes, and cross-model agreement. In particular, we examine how prompting strategies affect prediction stability, how models generalise beyond the gun-violence domain, where systematic errors arise, and whether model agreement can provide insight into annotation quality.

4.1. Reliability and Prompt Sensitivity

We first examine the stability of model predictions under different prompt formulations. Table 1 reports performance under standard zero-shot and few-shot prompting with and without definitions, while Table 2 isolates the effect of alternative prompt wording in zero-shot settings.

Together, these tables reveal substantial sensitivity to prompt design. For example, as shown in Table 2, GPT-4's F_1 score drops from 64.96 in the standard zero-shot setting without a definition to 47.40 under alternative wording. A similar pattern is visible in Table 1, where zero-shot performance varies considerably across prompt variants.

Including an explicit framing definition reduces this variability. In both Table 2 and Table 1, GPT-4's F_1 scores remain more stable when a definition is provided. Explanation-based prompting further improves consistency, as reflected in reduced variance across explainable and non-explainable conditions. In contrast, Llama 3 exhibits comparatively smaller fluctuations across wording variants, suggesting greater robustness to prompt variation.

4.2. Generalisation under Domain Shift

We next investigate how well models generalise across domains and supervision regimes. We first analyse few-shot performance on the GVFC dataset, where topical information is available. Results are reported in Table 1. GPT-4 and Llama 3 achieve their highest F_1 scores in the 8-varied in-domain setting (68.51 and 58.58, respectively), indicating that diverse in-domain examples support framing detection when domain cues are accessible.

To evaluate performance under more realistic conditions, where headline topics are varied and unknown in advance, we construct a new In-the-Wild (ITW) dataset composed of real-world news headlines covering diverse subjects. Framed headlines are collected from a website dedicated to highlighting news framing[8], while non-framed instances are sampled from mainstream outlets, including the BBC, *The Guardian*, and *Daily Mail*, and selected to match framed headlines in terms of publication period and topical diversity. This sampling strategy

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

	GPT-3.5-Turbo		GPT-4		Llama3 8B		FLAN-T5 Large fine-tuned	
	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.
GVFC	67.39	65.06	69.08	67.18	60.98	62.93	**76.33**	**78.83**
ITW	75.68	71.34	**82.49**	**80.25**	76.14	73.25	69.18	68.78

Table 3: Performance on GVFC test set (FS varied in domain) and ITW dataset (FS cross-domain).

aims to reduce potential source and temporal biases. The resulting dataset contains 157 headlines spanning topics such as weather, public health, migration, and European affairs, including 83 framed and 74 non-framed instances, corresponding to a majority baseline accuracy of 52.86%. All headlines were annotated by a domain expert following the framing definition introduced in Section 1.

Evaluation results on the ITW dataset are reported in Table 4. In the few-shot cross-domain setting, GPT-4 achieves its highest performance, reaching 80.25% accuracy, demonstrating that instruction-following LLMs can generalise to heterogeneous real-world data when appropriately prompted. Nevertheless, performance remains more variable than in-domain evaluations, indicating persistent challenges under domain shift.

We further contrast pre-trained and fine-tuned models using this evaluation set. While fine-tuned FLAN-T5 Large performs strongly on GVFC, its performance deteriorates substantially on ITW, falling behind pre-trained GPT and Llama 3 models, as shown in Table 3. This contrast highlights the limited transferability of task-specific fine-tuning and supports the use of pre-trained LLMs in low-resource, cross-domain settings.[9]

4.3. Systematic Failure Modes

Beyond aggregate performance, we examine recurring error patterns that limit model reliability. As illustrated by examples in Table 6, GPT-4 frequently misclassifies emotionally charged headlines as framed, contributing to systematic false positives. This behaviour suggests that emotional intensity is often treated as a proxy for framing, even when no strategic framing is present.

To further analyse model behaviour on ambiguous instances, we conduct a manual review of 1,300 headlines sampled from the GVFC dataset. This analysis reveals discrepancies in the original annotations, including potentially incorrect labels and cases in which multiple framing interpretations are plausible. Based on this review, we identify 134 *contested* instances. These correspond to headlines whose framing status cannot be determined unambiguously or for which reasonable alternative interpretations exist. Examples include "Live: Trump visits Pittsburgh after synagogue shooting" and "Shopify bans sale of certain firearms, accessories", both annotated as framed in GVFC despite the absence of clearly identifiable framing strategies. At the same time, not all contested cases reflect erroneous annotations. Some headlines, such as "Thousands gather to honor victims of the mass shooting with tears, candlelight, and song", present more nuanced challenges in framing detection, where emotional content and descriptive language complicate interpretation. Within this subset, 63.4% of headlines are annotated as framed.

For comparison, we also construct a *clear* subset of 134 headlines for which the presence or absence of framing is readily identifiable based on the annotation guidelines. Representative examples include "Two dead including shooter at Florida yoga studio" (non-framed) and "Parkland school shooter blames massacre on a 'demon' voice" (framed), both of which exhibit unambiguous framing status and align with their original annotations.

These two subsets enable a controlled analysis of model performance under conditions of low and high interpretive ambiguity.

Performance on these subsets is reported in Table 5. While all models perform well on the clear subset, performance deteriorates sharply on contested cases, with GPT-4's F_1 score dropping to 3.60. Similar degradation is observed for GPT-3.5 Turbo and Llama 3. These results indicate that current LLMs struggle when framing judgments require nuanced interpretation and contextual reasoning, even when they perform reliably on unambiguous instances.

4.4. Cross-Model Agreement and Annotation Quality

Finally, we examine whether patterns of agreement across models can provide insight into annotation quality. To this end, we analyse how often GPT-4, GPT-3.5 Turbo, and Llama 3 produce identical predictions on the clear and contested subsets. Agreement statistics are reported in Table 7.

On the contested subset, the three models reach unanimous agreement in 46.62% of cases. However, only 12.90% of these agreed-upon predictions match the original gold labels. In contrast, agreement on the clear subset is substantially higher

[9]The full results of FLAN-T5 models are available in appendix A and B.

	GPT-3.5 Turbo		GPT-4		Llama3 8B	
	F_1	Acc.	F_1	Acc.	F_1	Acc.
ZS + definition (S)	74.16	70.70	78.61	72.61	60.34	54.78
ZS + definition (E)	74.15	66.24	78.00	71.97	63.27	54.14
FS8 cross-domain (S)	67.50	66.88	81.11	78.34	69.33	70.70
FS8 cross-domain (E)	75.68	71.34	**82.49**	**80.25**	76.14	73.25

Table 4: Performance on the ITW Dataset, using the settings for Zero-Shot+Definition, Few-shot 8 examples cross domain. (S) stands for the standard setting, while (E) stands for the explainable setting.

	Llama3 8B		GPT-3.5		GPT-4	
	F_1	Acc.	F_1	Acc.	F_1	Acc.
Clear	76.60	75.00	71.43	70.15	75.81	77.61
Contested	21.31	27.82	41.43	38.81	3.60	20.15

Table 5: Comparative performance of the GPT and Llama3 models on the clear and contested framing subsets using the models' best performing settings.

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, **Llama3:** 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. **Llama3 Explanation:** The headline does not explicitly promote a particular perspective or interpretation of the issue. It simply presents the information in a neutral manner, without highlighting any specific aspect of the issue or promoting a particular agenda.
Headline: 'Devastated' Tamera Mowry Confirms Niece Alaina Housley Died In Thousand Oaks Shooting **Gold:** Not Framed, **GPT-4:** Framed, **GPT-3.5:** Not Framed, **Llama3:** Not Framed **GPT-4 Explanation:** The claim is framed as it emphasizes the emotional state of Tamera Mowry (being 'devastated') while confirming the death of her niece in the shooting. This can elicit an emotional response from the audience and influence their perception of the issue. **GPT-3.5 Explanation:** The claim is not framed because it is simply reporting the confirmation of Tamera Mowry about the death of her niece in the Thousand Oaks shooting. It does not select any specific aspects or favor a certain interpretation of the issue. **Llama3 Explanation:** This headline does not frame the issue by selecting certain aspects of the event, but rather reports the fact in a neutral way. It does not use emotive language, make value judgments, or highlight specific aspects of the event, but rather presents the information in a straightforward manner.

Table 6: Examples of the impact of emotional language on GPT-4.

and more frequently aligned with the annotations. These results indicate that strong cross-model consensus on contested cases often reflects disagreement with the existing labels rather than model error alone.

For example, all three models classify the headline "Muslim Americans raise more than $200,000 for those affected by Pittsburgh synagogue shooting" as not framed, despite its annotation as framed in GVFC. Manual inspection suggests that this headline does not exhibit an obvious framing strategy, supporting the models' shared interpretation.

Having manually reviewed the contested subset, we find that many instances with strong model agreement correspond to genuinely ambiguous or potentially problematic annotations. This suggests that cross-model consensus, particularly when it diverges from existing labels, can serve as a useful signal for identifying instances that merit further review. Overall, these findings indicate that ensembles of LLMs may provide a practical tool for flagging potential annotation inaccuracies in existing or newly constructed datasets, especially in low-resource settings where large-scale re-annotation is infeasible.

5. Conclusions

In this work, we conducted a systematic investigation of instruction-following large language models for detecting framing in news headlines. Through extensive experiments across zero-shot, few-shot, and explanation-based settings, as well as in-domain and out-of-domain evaluations, we assessed model reliability, generalisation, and failure modes in a task characterised by inherent ambiguity and limited supervision. Our results show that pre-trained LLMs, particularly GPT-4, can achieve

	Clear		Contested	
	Broad	Strict	Broad	Strict
Agr.	100.00	54.55	100.00	46.62
Agr. GL	80.30	48.48	21.80	12.90

Table 7: Clear vs. Contested Annotations: Percentages of agreement between the predictions of GPT-3.5, GPT-4, and Llama, as well as their alignment with the gold label. "Broad Agreement" includes cases where at least two out of three models (2/3 or 3/3) agree on a prediction and match the Gold Label. "Strict Agreement" refers to cases where all three models (3/3) agree, and their prediction matches the gold label. The "Agreement" row indicates the percentage of headlines where at least two models agreed on a prediction, while the "Agreement GL" row shows the percentage of these agreed predictions that align with the gold label.

strong performance when supported by carefully designed prompts and diverse few-shot examples. Explanation-based prompting improves prediction stability, while heterogeneous in-domain examples substantially enhance performance when topical information is available. At the same time, we identify persistent limitations, most notably a systematic tendency to conflate emotional language with framing and a pronounced performance collapse on contested cases requiring nuanced interpretation. We further demonstrate that domain-specific fine-tuning, although effective in controlled settings, does not reliably transfer to heterogeneous real-world data. In contrast, pre-trained models exhibit greater adaptability under domain shift, highlighting their practical value for framing analysis in low-resource and rapidly evolving media environments. Beyond model performance, our study highlights the importance of annotation quality in framing research. By analysing patterns of agreement across multiple models, we show that strong cross-model consensus can serve as a useful signal for identifying ambiguous or potentially problematic annotations. This provides a scalable mechanism for dataset auditing in contexts where expert annotation is costly or scarce. Taken together, our findings underscore both the promise and the current limitations of LLM-based framing detection. While these models offer a flexible and scalable alternative to traditional supervised approaches, their reliability remains constrained by prompt sensitivity, annotation uncertainty, and the inherent complexity of framing. Our analysis and the datasets introduced in this work provide a foundation for future research in this direction.

6. Limitations

The findings of this study have to be seen in light of some limitations. A significant constraint in the field of framing detection is the scarcity of expert-annotated datasets, which are not always publicly available. Even when such datasets are accessible, they often focus exclusively on framed data without including a balanced mix of framed and non-framed content.

Additionally, 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 and analysing media narratives.

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.

Finally, our binary framing classification (framed vs. not framed) simplifies a complex linguistic phenomenon. While framing often operates on a spectrum, treating it as a binary classification helps establish clearer methodological boundaries in framing detection studies. By enforcing a strict distinction between framed and non-framed content, this approach minimizes interpretative ambiguity, making it easier to compare framing patterns. Furthermore, a binary classification framework aligns with prior computational studies on media bias, allowing for more direct comparisons with existing research and ensuring reproducibility across different framing detection methodologies, and can aid in real-world applications, such as automatically flagging potentially framed news for further review.

7. Acknowledgements

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1].

8. Bibliographical 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.

Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. 2015. Testing and Comparing Computational Approaches for Identifying the Language of Framing in Political News. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1472–1482, Denver, Colorado. 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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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.

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.

Antonina Sinelnik and Dirk Hovy. 2024. Narratives at conflict: Computational analysis of news framing in multilingual disinformation campaigns. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 225–237, Bangkok, Thailand. 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.

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.

Appendix

A. FLAN-T5 Performance Lag

Table 8 shows the results of FLAN-T5 without any specific fine-tuning.

B. Fine-tuned FLAN-T5 and Hyperparameters

Table 9 presents the detailed performance comparison of fine-tuned FLAN-T5 small, base and large on the GVFC test set. The FLAN-T5 models (small, base, and large) were fine-tuned using the following hyperparameters: All models employ a learning rate of 5e-05, with a maximum input length of 70 tokens, a maximum label length of 4 tokens, and a batch size of 64. The number of epochs varies with 10 for the small, 40 for the base, and 17 for the large model, each within a 50-epoch limit and with early stopping set at 5 epochs to prevent overfitting. We used one A100 80G GPU, requiring less than 15 minutes of GPU time per model.

		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.
ZS	No Definition	0.46	49.79	**17.96**	49.21	0	50.10	**25.86**	**52.74**	34.33	46.19	51.01	**49.06**
	+Definition	0.15	50.13	1.21	50.10	0	50.10	0.61	50.06	**42.27**	48.06	37.63	43.96
FS	2 examples	0.31	50.06	0.31	50.06	0	50.10	3.12	49.98	5.74	49.64	8.32	49.64
	8 (focused in-domain)	0.15	50.17	3.84	50.10	0	50.13	4.83	50.17	16.78	47.53	36.85	44.69
	8 (varied in-domain)	0	50.13	0.15	49.94	0	50.13	1.66	49.94	22.78	45.96	35.03	45.27
	8 (cross domain)	**5.30**	49.37	2.95	49.60	0	50.10	4.99	50.33	40.23	45.11	**51.64**	46.88
	8 (mixed domain)	0.31	**50.21**	1.51	50.02	0	50.10	3.84	50.13	34.39	44.58	44.05	43.43

Table 8: Comparative performance of FLAN-T5 models using different prompt configurations.

	FLAN-T5 Small fine-tuned				**FLAN-T5 Base fine-tuned**				**FLAN-T5 Large fine-tuned**			
	F_1	Acc.	Prec.	Rec.	F_1	Acc.	Prec.	Rec.	F_1	Acc.	Prec.	Rec.
GVFC	0	52.29	0	0	79.77	79.50	74.82	85.42	76.33	78.83	80.63	72.46
ITW	2.38	47.70	100	1.20	63.87	56.05	56.48	73.49	69.18	68.78	72.36	66.26

Table 9: Performance of fine-tuned FLAN-T5 Small, Base and Large on the GVFC test set and ITW dataset.