The Missing Cause: An Analysis of Causal Attributions in Reporting on Palestine

Paulina Garcia-Corral
Hertie School
corral@hertie-school.org

Hannah Béchara
Hertie School
bechara@hertie-school.org

Krishnamoorthy Manohara
Hertie School
manohara@hertie-school.org

Slava Jankin
University of Birmingham
v.jankin@bham.ac.uk

Abstract

Missing cause bias is a specific type of bias in media reporting that relies on consistently omitting causal attribution to specific events, for example when omitting specific actors as causes of incidents. Identifying these patterns in news outlets can be helpful in assessing the level of bias present in media content. In this paper, we examine the prevalence of this bias in reporting on Palestine by identifying causal constructions in headlines. We compare headlines from three main news media outlets: CNN, the BBC, and AJ (AlJazeera), that cover the Israel-Palestine conflict. We also collect and compare these findings to data related to the Ukraine-Russia war to analyze editorial style within press organizations. We annotate a subset of this data and evaluate two causal language models (UniCausal and GPT-4o) for the identification and extraction of causal language in news headlines. Using the top performing model, GPT-4o, we machine annotate the full corpus and analyze missing bias prevalence within and across news organizations. Our findings reveal that BBC headlines tend to avoid directly attributing causality to Israel for the violence in Gaza, both when compared to other news outlets, and to its own reporting on other conflicts.

1 Introduction

Media reporting of conflict is often perceived by various stakeholders as biased. Headlines, in particular, are frequently criticized for being misleading, incomplete, or lacking context, which can skew the information presented. Just last year, the BBC reported more than 1,500 complaints over Israel-Palestine coverage, being accused of bias from both sides of the conflict[1]. A significant source of contention is the lack of causal attribution in events. Reports may state that citizens "die" rather than are "killed", or that hospitals are "destroyed" rather than "bombed" by specific actors.

Missing cause bias (Gentzkow and Shapiro, 2006), is a special type of information omission that consistently omits attributing responsibility, placing blame or giving praise to specific acts or actors that caused an event, such as passively describing a violent attack by using sentences that do not contain a subject, or using the passive voice to avoid naming an actor. In the past year, the subject of missing cause bias has become controversial in the media's coverage of the Israel-Palestine conflict. In July 2024, activists decried the BBC's coverage of the death of Mohammad Bhar by using the passive voice and failing to mention the cause of his death[2]. In August 2024, the BBC changed headlines after criticism from the Israeli Foreign Ministry for failing to mention the fact that a bombing was triggered after rockets where allegedly fired from Gaza[3].

In this paper, we propose using a causal relation extraction model to machine annotate online newspaper headlines, to measure missing cause in news media. By using causal relation extraction models to perform span detection of cause and effect (Drury et al., 2022), we can quantify and compare assigned causes and omitted causes after matching articles that cover the same events. The main contributions of this paper include:

- Measure causal headline prevalence
- Measure the prevalence of omitted cause bias
- Control for editorial style by doing a cross

[1] https://www.theguardian.com/media/2023/oct/16/bbc-gets-1500-complaints-over-israel\-hamas-coverage-split-50-50-on-each-side

[2] https://www.newarab.com/news/shameful-bbc-story-israels-killing-disabled-man-revised

[3] https://x.com/EmmanuelNahshon/status/1027440634968326149

Proceedings of the Proceedings of the 1st International Workshop on Nakba Narratives as Language Resources, pages 103–113
January 20, 2025.

comparison of headlines from the Russia-Ukraine war

2 Related Work

2.1 Automatic Bias Detection

While media bias has a long tradition in the realms of social sciences, it remains a relatively young research topic in Natural Language Processing. Common NLP techniques such as sentiment analysis (Lin et al., 2011), topic modeling (Best et al., 2005) and lexical feature analysis (Hube and Fetahu, 2018) have been used to detect media bias. More recently, supervised machine learning classifiers trained on transformer-based models have achieved better results although some underlying issues persist (Rodrigo-Ginés et al., 2024). Altogether, media bias is a a multi-faceted problem with several competing definitions, and bias detection remains a complex task.

Several researchers have addressed the detection of media bias as a classification problem. This classification can be binary (either bias exists or it does not exist), or it can be treated as a multi-classification problem. One such paper, which rates articles by degree of polarization, identifies a set of 130 content-based features that span 7 categories: structure, complexity, sentiment, bias, morality, topic and engagement, and show that they all contribute towards media bias detection (Horne et al., 2018). In contrast, media bias has also been classified by its stance towards an event (Cremisini et al., 2019) or by its place on the political compass (Baly et al., 2020).

Given that supervised classification usually uses case-specific annotations, there are some relevant attempts for our case study. To detect media bias using NLP, Al-Sarraj and Lubbad (2018) compared three supervised machine learning algorithms trained on an Israel-Palestine conflict news dataset, a SVM with bio-grams achieved the highest performance of 91.76% accuracy and F1-score of 91.46%. Wei and Santos (2020) collected data from history book excerpts and newspaper articles of the same conflict, and trained a sequence classifiers to predict authorship provenance. Their best model to detect narrative origin achieved an F1-score of 85.10% for history book excerpts and 91.90% for newspaper articles. Additionally, Cremisini et al. (2019) manually classified pro-Russia and pro-Western bias of news articles. Using a baseline SVM classifier with doc2vec embeddings, they achieved an F1-score of 86%. However, their results suggest that models may be learning journalistic styles rather than actually modeling bias. Similarly, Potash et al. (2017) applied a novel methodology to a gold-labeled set of articles annotated for Pro-Russian bias, where a Naive Bayes classifier achieved 82.60% accuracy and a feed-forward neural networks achieved 85.60% accuracy.

2.2 Causal Language Modeling

Causality mining is the task of identifying causal language that connects events in a text. It is a subtask of information extraction, where causal relations are identified and mined from a collection of documents (Drury et al., 2022). Causal language can be expressed explicitly or implicitly. The former tends to use connectors such as "because" or "therefore" to signal a causal relationship between events. Causal verbs can also be used to express causality. Causality can be found inter or intra-sententially. Because causal language is a linguistically, syntactically and semantically complex construction used to express causal reasoning (Solstad and Bott, 2017; Neeleman and Van de Koot, 2012), it tends to rely on contextual information, and yields low inter-annotator agreement when annotated by humans (Dunietz et al., 2017).

Causal language is usually modelled in three steps: 1) causal sequence classification, 2) causal extraction and 3) pair classification (Tan et al., 2023). Causal mining used to be based on pattern matching by identifying causal connectors (Drury et al., 2022). Advances in machine learning allowed for statistical pattern recognition models, such as SVMs and Bayesian models (Hidey and Mckeown, 2016; Zhao et al., 2017). With the introduction of deep learning, a combination of architectures such as CNNs (Kruengkrai et al., 2017; de Silva et al., 2017), CRFs (Fu et al., 2011), or LSTMs (Li et al., 2019; Dasgupta et al., 2018) improved previous results. Moreover, fine-tuning transformer-based models such as BERT significantly improved both classification and extraction capabilities across domains (Yang et al., 2023; Tan et al., 2023; Khetan et al., 2022).

More recently, LLMs have been investigated for their causal extraction capabilities. Takayanagi et al. (2024) assessed the performance of ChatGPT across both domain-specific and non-English datasets. They found that while ChatGPT demonstrates a baseline proficiency in causal text mining, it can be outperformed by earlier models when suf-

ficient training data is available. Similarly, Hobbhahn et al. (2022) explored GPT-3's capacity to identify causes and effects. Their results emphasize the significance of prompting, which suggests that GPT-3's predictions may be influenced more by the form of the input than by its content, raising questions about the model's true comprehension of causality. Luo et al. (2024) designed an LLM implementation that modifies causal datasets to optimize Event Causality Extraction. Experiments on both Chinese and English event causality extraction datasets achieved a 92% and 93% accuracy after using their proposed framework.

Furthermore, Causal language modeling has been tested on diverse news corpora. For example, Gusev and Tikhonov (2022) introduced HeadlineCause, a dataset annotated for implicit causal relations between paired news headlines in English (5,000 pairs) and Russian (9,000 pairs), annotated via crowdsourcing. Their XLM-RoBERTa-based model achieved 83.5% accuracy in English and 87.9% in Russian. Similarly, Tan et al. (2022) annotated protest-related news articles to create the Causal News Corpus, containing 3,559 sentences. They achieved an 81.20% F1-score on the test set and 83.46% in five-fold cross-validation. Additionally, Mariko et al. (2022) introduced FinCausal, a dataset designed to detect causal relationships in financial news. Lastly, Garcia Corral et al. (2024) developed a dataset to benchmark causal language detection that included data fro political press conferences.

Recent advancements in language modeling have enabled significant progress in causal event relation extraction. These advancements have, in turn, translated into significant improvements in downstream tasks that rely on causality mining to derive meaning from textual data. For example, Sun et al. (2024) achieved state-of-the-art results on the *Choice of Plausible Event in Sequence* (COPES) dataset. Their approach led to a 3.6%–16.6% improvement in correlation with human ratings in downstream narrative quality evaluation tasks, highlighting the importance of causality in computational narrative understanding. Similarly, Hosseini et al. (2019) demonstrated that causally and semantically coherent documents are more likely to be shared on social media, finding that coherence strongly influences online sharing behavior. These findings highlight how causal event detection can be leveraged for understanding textual organization and extracting key insights, such as underlying bias or positionality.

3 Methodology

To measure and analyze missing cause bias, we prepared a data selection and model evaluation pipeline to machine-annotate newspaper headlines at scale. We focused on headlines, as opposed to full articles, as they are optimized for contextual effect and processing effort, while directing readers to construct the optimal context for interpretation (Dor, 2003). This aligns with recent research that has shown that people can make inference from causal explanations (Kirfel et al., 2022). In other words, headlines give just enough information for readers to reconstruct the news story via inference, and omitting or including causal attributions in the headlines directly allows for implicit biases to be communicated without further information. Our data analysis and evaluation strategy can be divided into the following steps:

- Step 1: Data collection – We collected 4,993 headlines from AJ, the BBC and CNN, between May 2023 to February 2024. We scraped the Middle East and Europe sections of their online webpages, and filtered for relevant articles by searching for keywords around the Israel-Palestine and the Russia-Ukraine war.
- Step 2: Human Annotation – we labeled a subset of 541 random sentence to obtain a human "gold standard" for evaluation.
- Step 3: Model Comparison – We compared two models, one Bert-based and one LLM model, evaluated against the gold standard built in step 3.
- Step 4: Machine-Annotate Corpus – Using the best-performing model, we annotated the selected headlines for causal labels and causal spans.
- Step 5: Compare Explicit Cause Presence – Finally, after matching events across press organizations, we compared explicit cause presence across the different conflicts holding the event constant.

3.1 Data Collection

Our data collection process consisted of selecting three global news media outlets from different regions of the world, looking to maximize coverage

diversity. We chose Al-Jazeera (AJ), British Broadcasting News (BBC) and Cable Network News (CNN). AlJazeera English is an English-language news channel headquartered in the Middle East and funded in part by the Qatari government. The BBC (British Broadcasting Corporation) is a British public service broadcaster, the oldest and largest in the United Kingdom, and is funded principally by a license fee charged by the British Government. Finally, CNN is a multinational news channel and website operating out of the USA. Both BBC and AJ can be considered "state media" and mainstream of their respective governments, for the purposes of this study. To include a third English language organization from a different region, we included CNN, a private American news broadcasting agency.

In order to establish that the differences in the prevalence of causal headlines and the causal attributions are not merely stylistic choices, we selected two on-going conflicts in two regions of the world. We collected data from the Ukraine-Russia war, and the Israel-Palestine war. We scraped the online web sections of AJ Ukraine- Russia war, and AJ Israel-Palestine Conflict, BBC Middle East, BBC Ukraine, and CNN-Europe and CNN-Middle East between 17/05/2023 and 17/02/2024. We filtered out any articles that made no mention of "Israel", "Palestine", "Russia" and "Ukraine" to create the final dataset of headlines. Table 1 describes the composition of our dataset, listing the number of articles and their proportion to the total dataset (N). Table 2 shows a cross-section headlines from each region and source. All the data will be available in our repository.

Region	Source	N
Pal	AJ	1,251 (0.48)
Pal	BBC	792 (0.30)
Pal	CNN	567 (0.22)
Ukr	AJ	1,018 (0.43)
Ukr	BBC	784 (0.33)
Ukr	CNN	581 (0.24)

Table 1: Statistics for the corpus of all collected news articles. "Region" is where the conflict is occurring, "Source" refers to the news organization (AJ, BBC or CNN), and N refers to the total number of articles with the relative proportion between parenthesis.

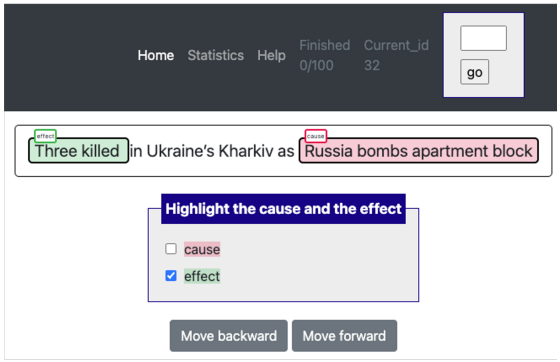

Figure 1: Potato annotation software example screen. We can see one headline from the Ukraine-Russia conflict with two spans selected for "cause" and "effect".

3.2 Corpus Annotation

In order to evaluate the selected model's accuracy in both the binary cause identification and span detection tasks, we created a gold standard of human-annotated data. We randomly selected 541 random headlines weighted across region and media outlets. We used a combination of Prolific[4] and Potato (Pei et al., 2022), a freely available web-based annotation tool which integrates with Prolific (See Figure 1 as reference). We hired and trained 8 students to annotate using the *Bring your own participants* (BYOP) option in Prolific. They annotated each headline's cause and effect from the subset we described in Section 3.1. The full statistics are detailed in Table 3. The human-annotated data has a distribution of 61% to 39% not causal to causal proportion, which is higher than what is expected from natural occurring (Dunietz et al., 2017). This might be due to a headline effect, where newspapers use causality more often than in natural occurring text, or a stylistic choice according to media organization, where click bait has heavily influence headlines style, such as phrasing headlines as questions. Alternatively, it could be an annotator drop-out rate effect where all the data was not consistently annotated across the weighted preselection.

We aggregated the causal label (0,1) by majority voting, and the causal spans using Overlap-Based Consensus, as we expect spans may vary slightly. To quantify the agreement between spans from different annotators, we used Intersection over Union (IoU) and a threshold of $\tau = 0.5$. Table 3 shows the descriptive statistics of human annotations.

[4] https://www.prolific.com/data-annotation

Date	Text	Region	Source
2024/02/12	Israel kills dozens in Rafah strikes, frees two captives	Palestine	AJ
2023/06/13	'Massive' Russian missile attack on Ukraine's Kryvyi Rih city	Ukraine	AJ
2024/01/26	Gaza war: ICJ to rule on call for Israel to stop military action - BBC News	Palestine	BBC
2024/01/28	Ukraine says it has uncovered major arms corruption - BBC News	Ukraine	BBC
2024/02/15	Israeli special forces raid largest functioning hospital in Gaza	Palestine	CNN
2023/08/24	Ukraine says it landed troops on the shores of Russian-occupied Crimea	Ukraine	CNN

Table 2: Example of headlines collected for the headline corpus from AJ, the BBC and CNN, for Middle East and Europe conflicts.

Region	Source	Causal	N	Perc
Palestine	Al Jazeera	0	73	0.549
		1	60	0.451
	BBC	0	53	0.602
		1	35	0.398
	CNN	0	43	0.597
		1	29	0.403
Ukraine	Al Jazeera	0	77	0.687
		1	35	0.313
	BBC	0	48	0.686
		1	22	0.314
	CNN	0	38	0.576
		1	28	0.424

Table 3: Human annotated subset and label statistics according to Region, Source and Causal headline label. "Perc" refers to the percentage of causal v. not causal headline in relation to the Region, Source and Causal label.

3.3 Model evaluation

To generate the machine annotations, we performed classification and causal span detection on all the collected headlines using two models. For both models, we ran inference with the out-of-the-box versions and did not perform fine-tuning with our human labeled data.

1 **UniCausal** (Tan et al., 2023), a BERT (Devlin et al., 2018) based causal language model fine-tuned on six, high-quality human-annotated corpora for causality. UniCausal is especially well-suited for our task as five of these six causal corpora include newspaper text. UniCausal achieved a 70.10% Binary F1-score for sequence classification, and a 52.42% F1-score on span detection on the overall corpus.

2 **GPT-4o** (OpenAI), a multilingual, multimodal generative pre-trained transformer developed by OpenAI. Model hyper-parameters and prompt are included in the Appendix (c.f. Section A.2).

Model	Accuracy	Prec	Recall	F1
GPT-4o	0.746	0.649	0.751	0.696
Unicausal	0.712	0.685	0.478	0.563

Table 4: GPT-4o and Unicausal model results for causal sequence classification against human labeled data

Table 4 and 5 report the results of GPT-4o and UniCausal on both sequence and spans detection. In both tasks, we see better performance from GPT-4o, which achieved an overall F1-score of 70% on binary sequence classification, with a high accuracy of 75%. Meanwhile, Unicausal achieved an F1-score of 56%, with a with a high accuracy of 71% but a low recall value of 48%, highlighting the model's difficulty in distinguishing between classes. For causal span extraction, evaluation is based on the exact match between predicted and human labeled entities. We used Seqeval (Nakayama, 2018) library for evaluation metric computing. The difference between model performance is underscored even more in causal extraction. While GPT-4o achieves an overall F1-score of 42% in causal labeling (43% for Cause and 40% for Effect), Unicausal dramatically underperforms with a score of 9% overall F1-score (9.5% for Cause and 8% for Effect). In line with previous related work, our results demonstrate that for smaller, domain specific datasets, LLMs can outperform causal sequence identification and span extraction when tested against out-of-the-box, not fine-tuned smaller models. We include confusion matrices to analyze classification error type in Figures 2 and 3.

Model	Span	Accuracy	Precision	Recall	F1
GPT-4o	Cause		0.375	0.521	0.436
	Effect		0.372	0.448	0.406
	Overall	0.755	0.374	0.481	0.420
Unicausal	Cause		0.107	0.084	0.095
	Effect		0.100	0.070	0.083
	Overall	0.714	0.106	0.076	0.089

Table 5: Using Seqeval, reported metrics for GPT-4o and Unicausal causal span detection against human labeled data

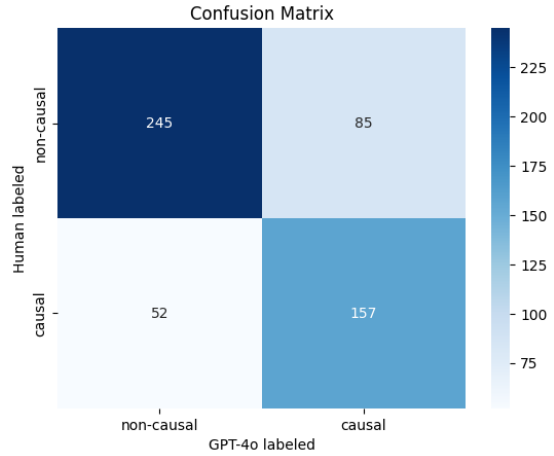

Figure 2: GPT-4o confusion matrix for sequence evaluation. The model achieves a reasonably high recall (75%), capturing most of the true causal instances. However, the model produces a fair number of false positives (85).

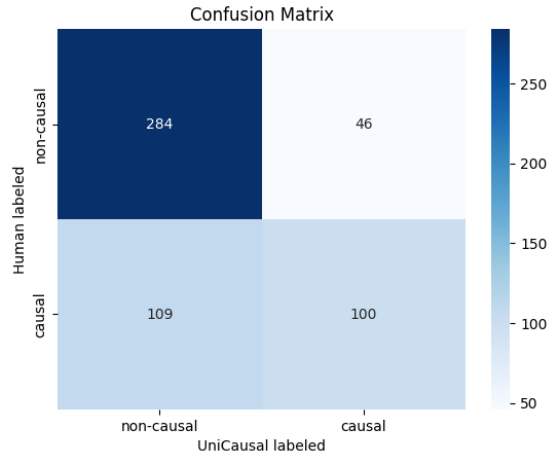

Figure 3: UniCausal confusion matrix for sequence evaluation. The model has a low recall (47%), failing to capture true causal instances. However, the model has a better precision compared to GPT-4o, better discerning true causal instances (100).

3.4 Matching Events with Cosine Similarity

To identify related articles across the three news organizations, and qualitatively analyze cause and effect spans, we matched headlines that refer to the same event across the news media outlets. To this end, we employed a temporal matching algorithm that linked articles from one source to another based on publication dates that fell within ±1 day of each other. Then, to evaluate the semantic alignment between matched pairs, we utilized a sentence-transformer model to compute cosine similarity scores between article headlines. Finally, all semantic similarity scores above 0.70 where qualitatively analyzed to confirm that the headlines were referring to the same event, and considered a complete match. Table 6 shows a sample of these aligned headlines.

4 Results

4.1 Sequence Classification

Based on our results, we compared the spans annotated by GPT-4o on all the collected headlines. We found that while 50% of AJ's headlines pertaining to Israel and Palestine are marked as causal, the same is true for only 35% of the BBC's headlines. CNN's headlines, on the other hand, were closer to AJ's in that 48% of headlines were marked as having a causal construction. The discrepancy between AJ and BBC diminishes when we look at headlines pertaining to the Ukraine-Russia conflict, where 38% of AJ's headlines are causal and 40% of BBC's headlines are causal. CNN's headlines, on the hand, do not vary greatly between regions. These results are summarized in Table 8. While these results give us a superficial look at how causality varies between regions and outlets, some of these differences can be attributed to editorial styles of each outlet. We therefore take a closer look at what is being left out.

4.2 Cause Identification

For a more fine-grained analysis, we take a closer look at the "cause" and "effect" spans annotated in the data. We selected for sentences in which the "effect" span includes references to violent acts as they tend to be contested in the context of conflict reporting. This search was based on relevant keywords. These keywords include the words: kill, murder, destroy, burn, dead/die, shot/shoot, strike, bomb, and attack.

AJ	BBC	CNN
Senior Hamas official Saleh al-Arouri killed in Beirut suburb	Hamas deputy leader Saleh al-Arouri killed in Beirut blast - BBC News	Senior Hamas leader killed in Beirut blast, heightening fears of wider regional conflict
Israel and Hamas agree to extend truce by two days, Qatar says	Israel-Hamas truce in Gaza extended for two days, Qatar says - BBC News	Deal reached to extend Israel-Hamas truce by two days, Qatar says
Gaza authorities say hundreds killed in Israeli air raid on hospital	Hospital blast in Gaza City kills hundreds - health officials - BBC News	Between 100 and 300 believed killed in Gaza hospital blast, according to preliminary US intelligence assessment
Family of Al Jazeera Gaza bureau chief killed in Israeli air raid	Wael Al-Dahdouh: Al Jazeera reporter's family killed in Gaza strike - BBC News	Journalist's family was killed in Gaza strike, says Al Jazeera

Table 6: A Sample of Aligned Headlines Using Temporal algorithm and Cosine Similarity scores to match.

Region	Source	Causal	N	Perc
Palestine	Al Jazeera	0	618	0.494
		1	633	0.506
	BBC	0	508	0.641
		1	284	0.359
	CNN	0	292	0.515
		1	275	0.485
Ukraine	Al Jazeera	0	630	0.619
		1	388	0.381
	BBC	0	465	0.593
		1	319	0.407
	CNN	0	284	0.489
		1	297	0.511

Table 7: GPT-4o machine labeled data contingency table by location, source and causal distribution of headlines.

Source	Isr-Pal	Rus-Ukr
AJ	50.6%	38%
BBC	35.8%	40%
CNN	48.5%	51%

Table 8: Percentage of positive sequences (causal sentences) in headlines by region and source as annotated by GPT-4o

Cause	AJ	BBC	CNN
Israel	40%	13%	25%
Russia	39%	34%	33%

Table 9: BBC vs AJ Headline Breakdown of Causal Sentences that Reference the Cause in Headlines Covering Violent Deaths

We then queried the cause span of the positive class for actors involved in the conflicts. Our aim was to determine whether or not there is a discrepancy between conflicts and outlets when it comes to directly identifying the actors. The results show a large gap between the number of headlines that explicitly name the cause in the headlines that refer Palestinian deaths. BBC's causal spans include Israel only 13% of the time, as opposed to AJ's 40% of the time and CNN's 25%. The results are summarized in Table 9. This discrepancy, while observable in the Russia-Ukraine headlines, is much less pronounced, suggesting a selective missing cause bias by conflict.

4.3 Direct Headline Comparison

In order to get a more conclusive look at the discrepancy between BBC, CNN and AJ's reporting on the Israel Palestine conflict, we further investigate a subset of sentences aligned using cosine similarity. We matched the headlines across all 3 outlets, leading to a final dataset of 50 headlines matched in this way. This allowed us to ensure that we are looking at headlines that cover the same

event, and rule out the possibility that the missing cause bias that we observe is simply a result of these outlets focusing on different stories. We then filter out these headlines further to isolate only headlines that refer to violent acts, and find that the difference only diminishes very slightly. As seen in Table 10, only 10% of BBC and 17% of CNN's headlines explicitly name Israel as the cause of this violence, a opposed to 32% of AJ's headlines. To extend the generalizability of the findings, we also extended the same methods to the Russia-Ukraine reporting. We aligned the headlines between all three outlets using cosine similarity and once again directly compare direct references to the cause in the headlines. The results are also reported in Table 10, and compared directly to the Israel-Palestine results.

Cause	AJ	BBC	CNN
Israel	32%	10%	17%
Russia	50%	41%	41%

Table 10: AJ vs BBC vs CNN Headline Breakdown of Causal Sentences that Reference the Cause in Headlines Covering Violent Deaths in Aligned Headlines

Overall, our analysis shows that BBC headlines on the Israel-Palestine conflict often avoiding direct attribution of causality to the responsible actors for the deaths and destruction in Gaza. For example, phrases like "reported killed in latest strikes" or "scores were killed in the camp" are used without explicitly identifying Israel as the cause. This is evident in that only 10% of causal sentences that describe violence will attribute the cause directly the responsible party. This tendency reflects an implicit bias through omission or missing cause bias. This difference is highlighted in the example below, which describes the headlines for February 2, 2024. In all three headlines, the cause is emphasized in bold text.

AJ ***Israel*** *kills dozens in Rafah strikes, frees two captives.*

BBC *Israel rescues two hostages in Rafah amid deadly strikes - BBC News.*

CNN *Israeli forces rescue 2 hostages as* ***airstrikes*** *kill around 100 Palestinians in Rafah.*

5 Discussion and Conclusions

In this paper, we explored the use of causal language in media reporting on Israel and Palestine and how its detection can act as an indicator of bias, offering a window into the subtle ways in which narratives are shaped. We compared headlines from three different media outlets, AJ, BBC and CNN, pertaining to their reporting on the escalation of the Israel-Palestine conflict. We directly compared their reporting the Israel-Palestine conflict to their reporting the Russia-Ukraine conflict. Using a state-of-the-art causal extraction method, we automatically classified the headlines as causal and non-causal. We further extracted the cause and effect spans of each of the headlines. A comparison shows a clear bias by omission on the part of the BBC Israel-Palestine reporting, and to a lesser extend to CNN's Israel-Palestine reporting, especially when compared to opposed to AJ's reporting. Furthermore, it showed a clear omission bias when comparing the BBC's reporting to its own reporting on the Russia-Ukraine conflict.

6 Limitations

Our research is not without its limitations. The scope of the study was confined to just three media outlets, which do not represent the entire spectrum of journalistic practices. Further research could expand upon this work and incorporate headlines from different sources, including different languages and from various political leanings. Furthermore, this study focuses on headlines only, as they are crafted to capture the most attention. However, a future avenue of research could also focus on the articles themselves and the causal language and slant present therein.

7 Acknowledgements

The authors thank the DFG (EXC number 2055 – Project number 390715649, SCRIPTS) for providing funding for the annotation efforts. This project has also received funding from the European Union's Horizon Europe research and innovation program under Grant Agreement No 101057131, Climate Action To Advance HeaLthY Societies in Europe (CATALYSE).

References

Wael F. Al-Sarraj and Heba M. Lubbad. 2018. Bias detection of palestinian/israeli conflict in western media: A sentiment analysis experimental study. In *2018 International Conference on Promising Electronic Technologies (ICPET)*, pages 98–103.

Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020. What was written vs. who read it: News media profiling using text analysis and social media context. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3364–3374, Online. Association for Computational Linguistics.

Clive Best, Erik van der Goot, Ken Blackler, Teófilo Garcia, and David Horby. 2005. Europe media monitor. *Technical Report EUR221 73 EN, European Commission*.

Andres Cremisini, Daniela Aguilar, and Mark A. Finlayson. 2019. A challenging dataset for bias detection: The case of the crisis in the ukraine. In *Social, Cultural, and Behavioral Modeling. SBP-BRiMS 2019*, volume 11549 of *Lecture Notes in Computer Science*. Springer, Cham.

Tirthankar Dasgupta, Rupsa Saha, Lipika Dey, and Abir Naskar. 2018. Automatic Extraction of Causal Relations from Text using Linguistically Informed Deep Neural Networks. In *Proceedings of the SIGDIAL 2018 Conference*, pages 12–14, Melbourne, Australia. Association for Computational Linguistics.

Tharini N. de Silva, Xiao Zhibo, Zhao Rui, and Mao Kezhi. 2017. Causal Relation Identification Using Convolutional Neural Networks and Knowledge Based Features. *International Journal of Computer and Systems Engineering*, 11(6).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Daniel Dor. 2003. On newspaper headlines as relevance optimizers. *Journal of Pragmatics*, 35(5):695–721.

Brett Drury, Hugo Gonçalo Oliveira, and Alneu de Andrade Lopes. 2022. A survey of the extraction and applications of causal relations. *Natural Langauge Engineering*.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017. The BECauSE corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 95–104, Valencia, Spain. Association for Computational Linguistics.

Jian-Feng Fu, Zong-Tian Liu, Wei Liu, and Wen Zhou. 2011. Event causal relation extraction based on cascaded conditional random fields. *Pattern Recognition and Artificial Intelligence*, 24(4):567.

Paulina Garcia Corral, Hanna Bechara, Ran Zhang, and Slava Jankin. 2024. PolitiCause: An annotation scheme and corpus for causality in political texts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12836–12845, Torino, Italia. ELRA and ICCL.

Matthew Gentzkow and Jesse M. Shapiro. 2006. Media bias and reputation. *Journal of Political Economy*, 114(2):280–316.

Ilya Gusev and Alexey Tikhonov. 2022. HeadlineCause: A dataset of news headlines for detecting causalities. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6153–6161, Marseille, France. European Language Resources Association.

Christopher Hidey and Kathleen Mckeown. 2016. Identifying Causal Relations Using Parallel Wikipedia Articles. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 3:1424–1433. Publisher: Association for Computational Linguistics (ACL).

Marius Hobbhahn, Tom Lieberum, and David Seiler. 2022. Investigating causal understanding in LLMs. In *NeurIPS 2022 Workshop on Causality for Real-world Impact*.

Benjamin Horne, Sara Khedr, and Sibel Adali. 2018. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

P. Hosseini, M. Diab, and D. A. Broniatowski. 2019. Does causal coherence predict online spread of social media? In *Social, Cultural, and Behavioral Modeling. SBP-BRiMS 2019*, volume 11549 of *Lecture Notes in Computer Science*. Springer, Cham.

Christoph Hube and Besnik Fetahu. 2018. Detecting biased statements in wikipedia. In *Companion proceedings of the the web conference 2018*, pages 1779–1786.

Vivek Khetan, Roshni R. Ramnani, Mayuresh Anand, Shubhashis Sengupta, and Andrew E. Fano. 2022. Causal bert: Language models for causality detection between events expressed in text. In *Intelligent Computing. Lecture Notes in Networks and Systems*, volume vol 283.

Lara Kirfel, Thomas Icard, and Tobias Gerstenberg. 2022. Inference from explanation. *Journal of Experimental Psychology: General*, 151(7):1481–1501.

C. Kruengkrai, K. Torisawa, C. Hashimoto, J. Kloetzer, J.-H. Oh, and M. Tanaka. 2017. Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Zhaoning Li, Qi Li, Xiaotian Zou, and Jiangtao Ren. 2019. Causality extraction based on self-attentive bilstm-crf with transferred embeddings. *ArXiv*, abs/1904.07629.

Yu-Ru Lin, James Bagrow, and David Lazer. 2011. More voices than ever? quantifying media bias in networks. In *Proceedings of the international AAAI*

conference on web and social media, volume 5, pages 193–200.

Kun Luo, Tong Zhou, Yubo Chen, Jun Zhao, and Kang Liu. 2024. Open event causality extraction by the assistance of LLM in task annotation, dataset, and method. In *Proceedings of the Workshop: Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning (NeusymBridge) @ LREC-COLING-2024*, pages 33–44, Torino, Italia. ELRA and ICCL.

Dominique Mariko, Hanna Abi-Akl, Kim Trottier, and Mahmoud El-Haj. 2022. The financial causality extraction shared task (FinCausal 2022). In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*, pages 105–107, Marseille, France. European Language Resources Association.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Ad Neeleman and Hans Van de Koot. 2012. The Linguistic Expression of Causation. In *The Theta System: Argument Structure at the Interface*. Oxford University Press, Oxford.

OpenAI. Gpt-4o system card. `https://openai.com/index/gpt-4o-system-card/`. Accessed: 2024-11-01.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. Potato: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Peter Potash, Alexey Romanov, Mikhail Gronas, Anna Rumshisky, and Mikhail Gronas. 2017. Tracking bias in news sources using social media: the Russia-Ukraine maidan crisis of 2013–2014. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 13–18, Copenhagen, Denmark. Association for Computational Linguistics.

Francisco-Javier Rodrigo-Ginés, Jorge Carrillo de Albornoz, and Laura Plaza. 2024. A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. *Expert Systems with Applications*, 237:121641.

Torgrim Solstad and Oliver Bott. 2017. 619Causality and Causal Reasoning in Natural Language. In *The Oxford Handbook of Causal Reasoning*. Oxford University Press.

Yidan Sun, Qin Chao, and Boyang Li. 2024. Event causality is key to computational story understanding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3493–3511, Mexico City, Mexico. Association for Computational Linguistics.

Takehiro Takayanagi, Masahiro Suzuki, Ryotaro Kobayashi, Hiroki Sakaji, and Kiyoshi Izumi. 2024. Is chatgpt the future of causal text mining? a comprehensive evaluation and analysis. *Preprint*, arXiv:2402.14484.

Fiona Anting Tan, Ali Hürriyetoğlu, Tommaso Caselli, Nelleke Oostdijk, Tadashi Nomoto, Hansi Hettiarachchi, Iqra Ameer, Onur Uca, Farhana Ferdousi Liza, and Tiancheng Hu. 2022. The causal news corpus: Annotating causal relations in event sentences from news. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2298–2310, Marseille, France. European Language Resources Association.

Fiona Anting Tan, Xinyu Zuo, and See-Kiong Ng. 2023. Unicausal: Unified benchmark and repository for causal text mining. In *Big Data Analytics and Knowledge Discovery - 25th International Conference, DaWaK 2023, Penang, Malaysia, August 28-30, 2023, Proceedings*, volume 14148 of *Lecture Notes in Computer Science*, pages 248–262. Springer.

Jason Wei and Eugene Santos. 2020. Narrative origin classification of israeli-palestinian conflict texts. In *The Thirty-Third International FLAIRS Conference (FLAIRS-33)*.

Jiaoyun Yang, Hao Xiong, Hongjin Zhang, Min Hu, and Ning An. 2023. Causal pattern representation learning for extracting causality from literature. In *Proceedings of the 2022 5th International Conference on Machine Learning and Natural Language Processing*, MLNLP '22, page 229–233, New York, NY, USA. Association for Computing Machinery.

Sendong Zhao, Quan Wang, Sean Massung, Bing Qin, Ting Liu, Bin Wang, and Chengxiang Zhai. 2017. Constructing and Embedding Abstract Event Causality Networks from Text Snippets. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 335–344. ACM.

A Appendix

In Tables 11, 12, 13 we present the collected and filtered headlines that compose out dataset. The first table shows an almost 50-50 distribution of headlines according to conflict type. Second, we see that the majority of the headlines collected are from AJ, then from the BBC and finally from CNN. This is probably due to regional focus each press organization. Finally, we see a cross table comparison were headline count is distributed according to region and press.

A.1 Full corpus statistics

Region	N
Palestine	2,610 (0.52)
Ukraine	2,383 (0.48)

Table 11: Distribution by region of the whole corpus

Media	N
AJ	2,269 (0.45)
BBC	1,576 (0.32)
CNN	1,148 (0.23)

Table 12: Distribution by media outlet of whole corpus

Region	Source	N
Palestine	AJ	1,251 (0.25)
	BBC	792 (0.16)
	CNN	567 (0.11)
Ukraine	AJ	1,018 (0.20)
	BBC	784 (0.16)
	CNN	581 (0.12)
		4,993 (1.00)

Table 13: Corpus distribution per press organization

A.2 GPT-4o parameter and prompt specifications

To machine annotate all the headlines we used batched inference through the OpenAI API. Prompt was based on task standard prompts reported on LLM causal research papers. We selected it to follow convention and allow for cross comparison. Hyper parameter specifications were selected to reduce randomness and optimize for reproducibility.

Prompt:

"You are a causal language model that performs causal sequence classification and causal span detection. You will classify a headline as causal or not causal, and if it's causal you will extract the causes and effects. The output should be a json with label 1 or 0, cause, and effect value such as `{\n \"label\": ,\n \"cause\": ,\n \"effect\": \n}`

Hyperparameter specification

url = /v1/chat/completions

max tokens = 115

model = gpt-4o

temperature = 0.0

top p = 1

frequency penalty = 0

presence penalty = 0