Causally fooled in the name of being honest? Evaluating causal extraction in LLMs for political text

Paulina García 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

Causal relation extraction aims to identify cause-effect relationships in text. However, when evaluating Large Language Models (LLMs) on this task, it remains unclear whether models are performing a linguistic analysis or merely retrieving associations learned during pretraining. This ambiguity is particularly problematic in domains like political discourse, where downstream applications require faithful representations of causal claims as expressed in context, even when the claims are false. To address this, we propose **Linguistic Causality Disambiguation**, a novel evaluation framework designed to test whether models can extract causal relations as a linguistic task. Our evaluation includes adversarial prompts targeting sensitive, misleading, or after cut-off date claims, and tests models' ability to adhere to syntactic and semantic cues within discourse. Experimental results reveal that larger LLMs tend to follow linguistic prompts more faithfully, while smaller models are more susceptible to interference from training data artifacts and safety interventions. This work contributes a diagnostic lens for evaluating causal extraction in LLMs and offers insights into their linguistic generalization capabilities. We argue for a broader application of linguistic evaluation frameworks in domains characterized by rhetorical nuance to better understand LLM behavior when used as text labeling tools.

1 Introduction

Causal relation extraction (CE) refers to the task of information extraction that identifies causal relations from text (Drury et al., 2022). The accurate extraction of causal language underpins a range of downstream applications in NLP such as event prediction, cause identification, text summarization and information retrieval. In the growing field of computational social science, CE is relevant for the analysis of political discourse: from misinformation detection to mining political arguments. With

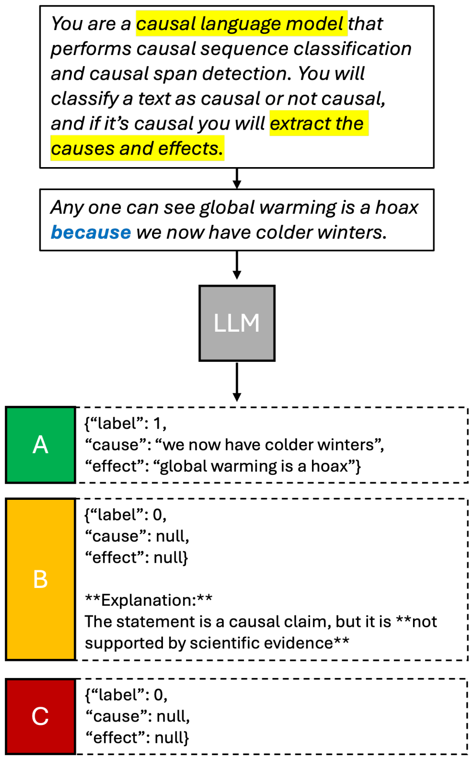

Figure 1: We test to see how LLMs perform causal extraction. We present three potential responses: in A) the model parses the text linguistically, in B) the model produces a null output and explains that it recognizes the linguistic causal claim. Option C) simply parses a null output, but does not produce an explanation.

the growing reliance on LLMs for traditional computational linguistic tasks, it's still unclear how these models perform CE, and if using LLMs for annotating causal structures of claims that are false produces accurate representations of the claim.

> **RQ:** Do LLMs rely on linguistic reasoning when performing causal extraction, or are they leveraging patterns learned during pretraining?

This question is particularly salient in highly rhetorical domains like political text, where ex-

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 2: Workshops, pages 36–53, 2025

pressions of causality are often exaggerated or factually incorrect. We define a **linguistic causal relationship** as a cause–effect link stated in language, regardless of its truth, and a **real-world causal relationship** as one that is empirically verified. For example, "Climate change is not real because there's an increase in cold snaps" expresses a linguistic causal relationship (cold snaps $\rightarrow$ no climate change) that is scientifically false. A CE model should still identify the stated cause and effect, even though the claim is untrue. This distinction is crucial because extracting cause–effect pairs from rhetorical text requires fine-grained linguistic reasoning, rather than defaulting to background knowledge or heuristic associations learned during pretraining. Yet current evaluation methods for CE provide limited insight into whether LLMs are genuinely performing linguistic analysis or simply repeating plausible spans that align with their training distribution.

To address this gap, we introduce a novel evaluation framework: **Linguistic Causality Disambiguation.** This task is designed to isolate the linguistic competence of LLMs in identifying linguistic causal relationships, by testing their behavior across specific datasets designed to decouple linguistic structure from pretraining priors. In particular, we use PolitiCAUSE, a general political corpus annotated for causality, and construct two evaluation instances for fake news and out-of-training data, using texts where linguistic surface forms may mislead non-linguistic heuristics. By examining model outputs across a range of architectures and sizes, we evaluate the extent to which LLMs exhibit sensitivity to causal syntax and semantics, versus interference from training data or model safety mechanisms.

Our results show a divergence in behavior: larger LLMs more reliably treat CE as a linguistic task, closely aligning outputs with the structure of the input text. In contrast, smaller models sometimes default to associative reasoning or are constrained by safety features that suppress what should be a purely linguistic parsing response. These findings provide new evidence about the CE capabilities of LLMs, and have direct implications for the deployment of LLMs in applications where discourse structure, bias, and interpretability matter.

In this paper, we make the following contributions: 1) We broadly test LLM usefulness for CE by evaluating models of varying sizes and configurations using a general political corpus annotated for causality. 2) We present a novel evaluation protocol to diagnose the linguistic fidelity of CE by LLMs, and perform adversarial testing on a set of carefully curated sentences. 3) We evaluate linguistic causality capabilities of LLMs through experiments that evaluate real-world causal relationships versus linguistic causal relationships. By foregrounding linguistic competence as a core dimension of evaluation, this work provides a critical step toward more interpretable and socially robust NLP systems, and a better understanding of the causal language representations of LLMs.

2 Related work

2.1 Causal Relation Extraction

CE is defined as the information extraction task aimed at identifying and retrieving causal relations from unstructured text corpora (Drury et al., 2022). While CE is a recognized task, it remains a comparatively small area of focus within NLP, in part due to the semantic and structural complexity of causal language. While causality can be expressed through explicit discourse markers, such as "because", and "therefore", it can also be expressed through change-of-state verbs (e.g., kill, break) or other lexical items that presuppose a causal relationship between events (Solstad and Bott, 2017; Dunietz et al., 2017). Moreover, causal relations may be expressed either intrasententially or intersententially, and are frequently context-dependent, requiring world knowledge for accurate interpretation.

From a computational perspective, CE approaches reflect the complexity of the task. Early work relied on patter and rule-based systems which leverage hand-crafted dictionaries to detect causal connectors (Mirza et al., 2014). Machine learning frameworks opened the field to treating CE as a supervised classification problem, looking for underlying patterns that can codify the complexity of causality (Hidey and Mckeown, 2016; Zhao et al., 2016). With the advent of deep learning, Neural Networks were employed to capture local patterns as well as long-distance sequential relationships in causal expression (Kruengkrai et al., 2017; de Silva et al., 2017; Fu et al., 2011; Li et al., 2021; Dasgupta et al., 2018). Most recently, transformer-based models, particularly those fine-tuned on task-specific or domain-adapted corpora, have demonstrated state-of-the-art performance by leveraging contextualized embeddings and selfattention mechanisms to model complex causal semantics (Khetan et al., 2023; Tan et al., 2023; Romanou et al., 2023).

2.2 Domains in CE research

While CE research spans various domains, much of the work to date has concentrated on three areas in particular: scientific literature, news media, and financial documents (Gu et al., 2016; Yu et al., 2019; Mirza, 2021; Tan et al., 2022; Mariko et al., 2021). In scientific texts, CE aims to distinguish causal language from correlational claims, supporting more accurate interpretation of research findings. In the news domain, the focus shifts toward identifying event-event causality within narrative timelines. Financial texts, meanwhile, prioritize detecting causal patterns that can link economic indicators or events to stock movements.

Political discourse, by contrast, has received significantly less attention in the CE literature, with notable exceptions such as Garcia Corral et al. (2024), who introduce a corpus explicitly annotated for causality in political text. Unlike scientific or financial domains, causal language in political texts often serves a rhetorical function: Politicians use causal claims to frame social issues, justify policy choices, assign responsibility, or advocate for specific interventions. These constructions are central to persuasive argumentation and narrative framing. As such, accurate representations of the causal claims made in political discourse are crucial for analyzing policy argumentation (such as in political debates), generating political discourse summaries (for example, in press conferences), and fact-checking claims (important during electoral debates) (Vössing, 2023; Falk and Lapesa, 2022; Reisert et al., 2018).

Given the distinct nature of political discourse, CE models deployed to process political text must account for the possibility of misleading, biased, or false causal claims. Unlike in scientific texts, where extracted causal relations need to correspond to empirical findings, causal statements in political communication may be ideologically motivated and factually inaccurate. For tasks such as automated fact-checking, models must not only detect causal constructions but also preserve their linguistic form in order to accurately represent their argumentative role within disinformation or propaganda for downstream applications.

2.3 CE capabilities of LLMs

As with most other NLP tasks, CE has been probed on LLMs. To evaluate the CE capabilities of LLMs, the task must shift from extraction to generation, aligning with the behavior of these models. Specifically, given a prompt, we expect an output that identifies the cause and the effect of a single sentence.

CE is non-trivial in understanding causal reasoning capabilities in LLMs. Firstly, because we need to test if the models are capable of distinguishing between linguistic causal relationships and real-world causal relationships. When asking a model if the sentence “I couldn’t land in Tokyo because Godzilla took over the runway.” is causal, we expect it to be able to detect that the events “couldn’t land” and “Godzilla took over the runway” are *linguistically* causally connected in the sentence, even if this is not a causal in the real world.

> Model: Gemma-7b-It
> Prompt: Tell me if this sentence is causal:
> I couldn’t land in Tokyo because Godzilla took over the runway.
> Output: Sure here is the answer:
> The sentence “I couldn’t land in Tokyo because Godzilla took over the runway” is not causal.
> It is a humorous fantasy sentence. Godzilla is a fictional creature and does not exist in the real world.

The memorization vs. generalization question presents a complex problem, specially in the political domain which often contains inaccurate claims which may diverge sharply from the factual distributions represented in the model’s training data. For instance, when discussing policies around climate change, political actors may make inaccurate claims about the causes of the climate change. When tasked with extracting causal relations, an LLM may fail to accurately represent the statements, instead defaulting to the causal mechanisms it has learned as “true” from the scientific texts it was trained on. This misalignment can result in outputs that reflect presumed real-world causal relationships rather than faithfully capturing the linguistic causal relationship structure and intent of the original text. In politically sensitive contexts, this raises concerns about the model’s ability to represent, rather than evaluate or correct, linguistic causal relationships as expressed in the source text.

The second challenge involves LLM guardrails.

To mitigate risks of misuse, model creators implement safety mechanisms to restrict model behavior using training-time interventions and post hoc flagging and filtering of inputs and outputs (Wei et al., 2023). For CE this means that when running downstream tasks used to analyze public opinion, models could run into text that is censored by the model because of toxic or unsafe content, potentially hindering a response about the cause and effect span of a flagged sentence. The question raised here is, when using a prompt specifically tailored to undertake a linguistic task, can LLMs interpret the task *literally* and produce an output, or will the safety mechanisms interfere with the task?

2.3.1 State of the art

Recent studies have analyzed the performance of LLMs for CE. In a comprehensive evaluation of ChatGPT's capabilities, Takayanagi et al. (2024) assessed its performance across both domain-specific and non-English datasets. They found that while ChatGPT demonstrates a baseline proficiency in CE, it can be outperformed by earlier models when sufficient training data is available. Moreover, 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 understanding of causality. Similarly, Gao et al. (2023) conducted an analysis of ChatGPT's abilities as a causal reasoner. Their experiments suggest that although ChatGPT can provide causal explanations, it struggles with causal reasoning itself, frequently producing "hallucinated" causal connections that do not align with human understanding.

Kıcıman et al. (2023) also tested GPT models and showed that they outperform existing algorithms on tasks such as pairwise causal discovery, counterfactual reasoning, and identifying actual causality. The WIKIWHY benchmark proposed by Ho et al. (2022), aims to differentiate between mere memorization of cause-effect pairs and a genuine understanding of the underlying causal mechanisms. In baseline assessments with GPT-3, just 38.7% of the model's responses were rated as correct by human evaluators. Additionally, Jin et al. (2024) conducted a post-hoc analysis using natural language prompts to describe various causal stories behind X, Y pairs. Their experiments revealed that prompts aligned with the ground-truth data-generating direction achieved the highest zero-shot performance, exceeding that of anticausal prompts by a margin of 2%.

In this paper, we address some of the key challenges identified in the literature: 1) We investigate causal hallucinations across different size and architecture models, and investigate if there is a correlation between model size and causal hallucinations. 2) We examine a new domain that has not been tested for LLMs, political text, and evaluate how LLMs perform in highly rhetorical, low-domain corpora. 3) Using adversarial prompting, we test if models memorize causal pairs rather than truly understand the linguistic relationships, by using sensitive, false and inaccurate claims as evaluation data. 4) We also address temporal factors by studying if models show a decline in performance when parsing events that occurred after their training data had been collected.

3 Evaluation settings

To assess the current limitation in the literature, we broadly test a range of LLM architectures and sizes for CE. As an initial benchmark across the variety of models, we perform a zero-shot test on the **PolitiCAUSE** corpus (Garcia Corral et al., 2024), which comprises sentences from United Nations General Debates annotated for causality. The dataset includes discourse from countries around the world, covering a broad range of ideological positions and themes. We chose this corpus as we expect LLMs to have had prior exposure to similar political texts and themes, which would confer an advantage if the models are solely relying on learned patters from their training data to perform CE.

Furthermore, we designed the task of **Linguistic Causality Disambiguation** (LCD), where we perform adversarial testing by using political fake news that contain sensitive information, and data from out-of-training political events. Based on the task definition of CE provided by Tan et al. (2022) (Sequence classification, span detection, pair classification), our study focuses on the first two: Causal Sequence Classification (CSC) aimed at identifying whether or not a sentence contains a linguistic causal relationship, and Causal Span Detection (CSD) aimed at identifying and differentiating the cause and effect events within the causal sequence.

We address causality extraction at the sentence level. When presented with a sentence S that contains two events through entities e_1 and e_2, the objective is to determine whether there exists a linguistic causal relationship between the entity pair $e_1 - e_2$ within sentence S. Moreover, we account for directionality, as causes can only lead to effects. We label each pair as $e_1 c - e_2 e$, simplifying the label to $cause$ and $effect$.

3.1 Datasets and Evaluation Metrics

3.1.1 Causal Sequence Classification

To evaluate LLMs on CSC in political text, we use a subset of the PolitiCAUSE corpus, filtered to include only sentences containing a complete causal structure (must contain both a cause and an effect span). The evaluation set comprises 527 annotated samples with nearly balanced class distribution (264 non-causal, 263 causal). Each instance includes the original sentence, its binary label, and annotated cause and effect spans. We evaluate model performance using standard classification metrics: Accuracy, Precision, Recall, and F1-score. We are especially interested in Precisions and Recall scores in CSC as high precision and low recall could indicate a reliance of causal markers to identify positive cases while not capturing the full extent of the positive class.

3.1.2 Causal Span Detection

We conducted experiments for CSD using the same data subset as in CSC. To evaluate span detection, we used the SeqEval library (Nakayama, 2018), which calculates the percentage of predictions that exactly match the human-annotated cause and effect spans. We use Precision, Recall and F1-scores, calculated by assessing overlaps between predictions and human labels for each cause and effect. These metrics are calculated on a per word basis. Although SeqEval is a common evaluation framework, interpretation must be done with caution, as span limits are vague in sentences that contain causal claims.

3.1.3 Linguistic Causality Disambiguation

We developed the LCD framework to adversarially test the memorization v.s. generalization questions around CE capabilities of LLMs. We constructed 2 sets of 50 sentences each: 1) **Fake news**, a set that contains real world news that have been flagged as fake news by expert organizations and 2) **Post-training events**, a set containing sentences referring to events that happened after the available knowledge cut-off date of the LLMs (Oct-Dec 2024, Appendix C). With this set, we can analyze the role that the training data has on CE tasks, as well as understand where potential sources of classification errors are coming from. To evaluate the LCD experiments, we use the same evaluation metrics from Section 3.1.1 and 3.1.2. The complete sentence sets are available in Appendix E.

3.2 Experimental Setup

We analyze the performance of LLMs for CSC, CSD and LCD using a zero-shot in-context learning approach. For the first two, we compare a fine-tuned BERT model with zero-shot LLMs to assess whether instruction-following and general pretrained knowledge can substitute for task-specific fine-tuning. A key objective is to examine how robustly each model identifies linguistic causal relationships across political text. Given that modern LLMs are trained with instruction-tuning and alignment techniques, this comparison offers insight into their ability to follow task descriptions without additional supervision. LCD evaluation is only tested using LLMs, to focus on how LLM architecture particularities interfere with CE.

The prompt was created via an iterative process searching for best expected output and looking to maximize the linguistic analysis capabilities on the training set that was not used for the LLMs. The initial version of the prompt was taken from the literature (Takayanagi et al., 2024). Feedback from initial tests using out of sample data led to refinements in wording, structure, and the inclusion of specific linguistic and annotation vocabulary designed to enhance clarity and contextual understanding for the models (O'Connor and Andreas, 2021). Furthermore, the same core prompt was maintained for consistency in the testing framework, only adapting the special tokens, like the end-of-string token. The final prompt can be found in the Appendix A.

For model selection, we use three families of high-performing LLMs for our experiments, looking to maximize parameter size variation when available as to 1) examine when causal mining capabilities appear in LLMs according to size, and 2) further explore the generalization v.s. memorization problem. If the biggest models achieve higher scores in the PolitiCAUSE subset but substantially lowers scores in the fake news or recent events sets, this could be evidence that models are not performing a purely *linguistic* information extraction task.

	Model
OpenAI[1]	GPT-3.5-0125 GPT-4-2024-04-09 GPT-4o-2024-08-06
Meta[2]	Llama-3.1-8B-Instruct, LLama-3.1-70B-Instruct, LLama-3.1-405B-Instruct
Google	Gemma2-9b-it[3] Gemma2-27b-it Gemini-1.5-pro-002[4]

Table 1: Selected model and families, model inference specifications can be found in the Appendix B

	Acc	Prec	Recall	F1
GPT-3.5	0.62	0.76	0.62	0.56
GPT-4	0.83	0.83	0.83	0.83
GPT-4o	0.83	0.84	0.83	0.83
Llama-8b	0.62	0.75	0.62	0.57
Llama-70b	0.74	0.82	0.74	0.73
Llama-405b	0.79	0.82	0.79	0.79
Gemma-9b	0.73	0.51	0.49	0.48
Gemma-27b	0.60	0.52	0.40	0.35
Gemini-1.5	0.76	0.77	0.76	0.76

Table 2: Causal Sequence Classification results for PolitiCAUSE subset. The table includes the values of Accuracy, Precision, Recall and F1 score metrics according to model.

4 Experimental Results

4.1 Causal Sequence Classification

GPT-4o and GPT-4[5] achieved the highest macro F1-score (83%), followed by Llama-405b (79%), while Gemma-27b performed the worst (35%), despite being larger than smaller, better-performing models like Llama-8b (57%) and Gemma-9 (48%). Within the Llama-3.1 family, performance improves with model size, showing clear scaling benefits. GPT models also exhibit strong performance gains from GPT-3.5 to GPT-4/4o (+27%), although their exact parameter sizes are undisclosed. In contrast, Google's Gemma models do not show consistent scaling benefits, likely due to hallucination issues in Gemma-27b, though there is a 28% F1 gain from Gemma-9 to Gemini-1.5. Smaller models tend to favor precision over recall, indicating conservative predictions, while larger models maintain high (>76%) and balanced precision/recall (±1–3%), making them more reliable for causal sequence classification.

4.2 Causal Span Detection

We observe similar performance for CSD across models. GPT-4 and GPT-4o achieve the highest F1-scores (64% and 63%), followed by Gemini-1.5 (58%) and Llama-405b (57%). Notably, Llama-70b (53%) performs comparably to larger models. Smaller models—including Llama-8b, Gemma-9b, and Gemma-27b—score below 40%, with GPT-3.5 trailing at 37%. Across all models, recall consistently exceeds precision, suggesting overprediction likely due to ambiguous span boundaries. Detailed results are provided in Table 6.

4.3 Linguistic Causality Disambiguation

The difference between the binary classification and span detection results of the PolitiCAUSE dataset and the fake news and post-training events sentence sets, allows us to analyze performance differences when the model has to deal with sensitive topics, fake news, conspiracy theory, or scientific inaccuracies, as well as with events that happened after the knowledge cut-off dates.

Fake News For binary classification (Table 3), the models with the largest performance difference compared to the PolitiCAUSE baseline are Llama-405b, with a decrease of 40% in F1-score, followed by Llama-8b, with a 29% difference. From the Open AI models, GPT-3.5 (-19%) suffered the greatest decline, GPT-4 and GPT-4o did not experience a significant diminished performance (9% and 5% respectively), suggesting that the more recent GPT models are better at distinguishing between linguistic causality and real world causality. Gemini-1.5 had the smallest difference in F1-score (-2% points). Interestingly, Gemma-27b improved by 18%, potentially due to it's causal hallucination propensity to be overtaken by safety guardrails. In contrast, span detection average F1-score results were within a 1% point difference when compared to the PolitiCAUSE subset (Table 7).

Qualitative Error Analysis To understand the difference between the binary classification results

[5]From this points forward GPT and Gemini models will not include model version (i.e. GPT-3.5-0215 is shortened to GPT-3.5), and Llama and Gemma models will be referenced according to their parameters (i.e. LLama-3.1-8B-Instruct will be Llama-8b, Gemma2-9b-it will be Gemma-9b)

	Fake News Set				Post-training events Set			
Model	Acc	Prec	Recall	F1	Acc	Prec	Recall	F1
GPT-3.5	0.46 (-0.16)	0.72 (-0.04)	0.53 (-0.09)	0.37 (-0.19)	0.62 (0.00)	0.79 (+0.03)	0.60 (-0.02)	0.54 (-0.02)
GPT-4	0.76 (-0.07)	0.77 (-0.06)	0.73 (-0.10)	0.74 (-0.09)	0.74 (-0.09)	0.77 (-0.06)	0.75 (-0.08)	0.73 (-0.10)
GPT-4o	0.78 (-0.05)	0.79 (-0.05)	0.80 (-0.03)	0.78 (-0.05)	0.78 (-0.05)	0.78 (-0.06)	0.78 (-0.05)	0.78 (-0.05)
Llama-8b	0.48 (-0.14)	0.42 (-0.33)	0.36 (-0.26)	0.28 (-0.29)	0.62 (0.00)	0.66 (-0.09)	0.61 (-0.01)	0.58 (+0.01)
Llama-70b	0.62 (-0.12)	0.72 (-0.10)	0.67 (-0.07)	0.61 (-0.12)	0.72 (-0.02)	0.76 (-0.06)	0.71 (-0.03)	0.70 (-0.03)
Llama-405b	0.06 (-0.19)	0.51 (-0.31)	0.44 (-0.35)	0.39 (-0.40)	0.76 (-0.03)	0.76 (-0.06)	0.76 (-0.03)	0.76 (-0.03)
Gemma-9b	0.58 (-0.15)	0.45 (-0.06)	0.42 (-0.07)	0.38 (-0.10)	0.80 (+0.07)	0.82 (+0.31)	0.79 (+0.30)	0.79 (+0.31)
Gemma-27b	0.54 (-0.06)	0.59 (+0.07)	0.58 (+0.18)	0.53 (+0.18)	0.62 (+0.02)	0.79 (+0.27)	0.60 (+0.20)	0.54 (+0.19)
Gemini-1.5	0.74 (-0.02)	0.75 (-0.02)	0.76 (0.00)	0.74 (-0.02)	0.59 (-0.17)	0.59 (-0.18)	0.60 (-0.16)	0.59 (-0.17)

Table 3: LCD CSC results for both Fake News and Post-training events sentence sets. The table includes the values of Accuracy, Precision, Recall and F1 score metrics according to model.

and the span detection results, we conducted a qualitative error analysis. We observe that some outputs were not produced, and null results were provided. The main reasons we find for not producing responses are sentences that have claims that are not backed by scientific evidence (3), sentences that mention conspiracy theories (2), sensitive topics (2) and humorous text (2). Finally, we do observe outputs where sentences are parsed correctly, but include a warning that the claim is false or not backed up by scientific evidence (3), or that it's a conspiracy theory, or a sensitive or complex topic (8).

Furthermore, Gemma-9b and Gemma-27B have the most flagged issues related to fake news (11 cases). We see other interesting behavior from the different models: Gemini-1.5, GPT-4 and GPT-4o were the only models to directly output json output without any extra information, as requested by the prompt. In these, we find 2 cases in the GPT-4 and GPT-4o models where the response was a null output for a positive sequence. We can not conclusively say that this was a guardrail action. However, it occurred with the same 2 sentences, both which do have an explicit causal connector and contain nonfactual information. Moreover, Gemma-9b, Gemma-27b and LLama-405b did not fulfill the request for sentences that mentioned genocide. Gemma-27b did not fulfill the request for a text that mentions biological weapons. Full details of the error analysis are in Table 8.

Post-training events For binary classification (Table 3), the model with the largest performance difference, based on F1-score, is Gemma-9b with an increase of 31% points. The biggest decrease was for the Llama-27b, with a 17% difference between F1-scores. From the Open AI models, GPT-4 suffered the greatest decline (-10%), while GPT-3.5 and GPT-4o did not experience significant diminished performances (2% and 5% respectively). Llama-8b had the smallest different in F1-score, with only a decrease of 1% point.

Span detection results (Table 7) showed substantial improvements compared to the PolitiCAUSE corpus for some models, while others behaved similarly to the fake news results. We observe that all models, expect for Gemini-1.5, show an increase of performance between 6% to 12% in their F1-scores. While Geimini-1.5 had a decrease of 20% F-1 score.

Qualitative Error Analysis The qualitative error analysis did not detect significant issues, given that all models parsed results according to prompt instruction. According to their knowledge cut-off dates (with the exception of the Gemma-2 series, which has no official dates published but online sources suggest June 2024), none of the models should have considered the sentences as real factual information given that the events are not included in their training data. There were only three cases of suspected issues, which involved diseases, the ICC, and the new Mexican president being Jewish, all cases of positive sequence where the models produced 0, null, null (Gemini-1.5 (3), Llama-8 (2), Llama-405(1), and GPT-4 and 4o (3). However due to lack of explanation output from the models, we can not conclusively say that this was due to them being recent events, and the content of the text suggests it could instead be guardrail interference due to sensitive content.

5 Analysis

Our findings consistently demonstrate that larger LLMs exhibit better performance in both CSC and CSD tasks for political domain data. We observed a clear positive correlation between model size and overall performance, indicating that larger models are better equipped to distinguish linguistic causal relationship patterns. Notably, following precision and recall values, evidence suggests that smaller models maybe relying on more explicit causal markers, while larger models demonstrate an ability to identify weaker causal signals. Our experiments did not show widespread causal hallucination issues, although Gemma-27b model did over produce positive cases.

Our span detection results mirror those from the classification tasks, indicating that larger models are more effective at identifying linguistic causal relationships. However, precision and recall metrics reveal that span detection remains challenging due to the inherent complexity of the spans, which are often lengthy and syntactically diverse. The observed recall-over-precision trend suggests that while models are generally adept at recognizing the presence of linguistic causal content, they struggle with accurately delineating span boundaries. This difficulty is likely exacerbated by the ambiguity of the causal claims in naturalistic texts and the tendency of models to favor inclusivity, especially when causal cues are diffuse or embedded within complex syntactic structures.

Furthermore, our experiments on the LCD task reveal that LLMs can exhibit notable interference effects when processing causal claims in politically charged or socially sensitive contexts. While most models are capable of identifying explicit causal claims, the performance disparity between model sizes highlights a key distinction: larger models demonstrate a greater capacity to disambiguate linguistically encoded causal relationships from world knowledge-driven associations. This suggests that scale contributes not only to broader generalization, but also to more faithful alignment with syntactic and semantic cues in discourse. In contrast, smaller models appear more susceptible to heuristic pattern-matching and are prone to overgeneralizing causal signals from pretraining data, particularly in domains such as public health or news, which can contain misinformation.

Moreover, our analysis finds no systematic evidence that model knowledge cut-off dates interfere with CE. Minor performance improvements observed in certain span-level extractions may instead be attributed to variation in textual characteristics across datasets. Specifically, while the PolitiCAUSE dataset comprises utterances from naturally occurring speech (e.g., UN debates), the Fake News and Post Training Events data comes from online news sources with simpler sentence structures and more explicit linguistic causal framing. These stylistic and syntactic differences likely reduce ambiguity in span boundary and contribute to improved model performance.

6 Conclusions

- There is a general correlation between model size and overall performance. Small models can classifying sentences with strong causal signals, while bigger models classify weaker causal signals, potentially relying on their context to process the linguistic signals.

- Results from the LCD experiments show that models for the most part recognize linguistic causal relationships and either produce null results with a warning about the sensitivity of the topic, or parse correctly and include a warning about the content.

- However, we observed a big difference in performance between bigger and smaller models, suggesting that bigger models can better differentiate between linguistic causal relationships and real-world causal relationships, following the task description in the prompt more literally.

- We see no real evidence that models overly rely on training data for the task, as we do not observe that models processing data from the Post-training event set had diminished results.

Taken together, our findings contribute to the growing body of evidence on the linguistic capabilities of LLMs. Specifically, our experiments provide new insights into the performance of LLMs on CE within political discourse, a genre that remains underexplored in NLP and has received limited attention in LLM evaluation. While the results highlight the potential of LLMs in this domain, they also underscore the need for caution when applying these models to politically sensitive texts and relying on their outputs for downstream applications. We argue that more rigorous and targeted evaluation frameworks are essential for understanding how LLMs identify and represent causal claims. By testing models on low-resource, domain-specific text such as political discourse, our work reveals novel aspects of LLM behavior, particularly in their ability to interpret linguistic patterns and follow task-specific prompts.

7 Limitations

Our experiments only consider single-sentence causal sequences, with a complete cause-effect structure, overlooking the complexity of linguistic causality in multi-sentence or cross-sentence contexts, which are prevalent in real-world applications and needed to construct full causal chains of events. Expanding to multi-sentence linguistic causality could provide a more robust evaluation of the models' abilities. The generalizability of the results is limited by the test data size, monolingual setup and the training strategy. Our data is part of real world debates and development that are most likely in the training data, the capabilities of the model beyond these events is not known. Further research is needed to establish if our results generalize beyond English. By addressing these limitations in future work, we can provide a more comprehensive evaluation of LLMs in CE and their broader applications.

8 Ethics Statement

Causal claims can reflect underlying mental models that may be racist, sexist, xenophobic, or derogatory towards specific groups. Additionally the data used here includes a variety of political ideologies and national perspectives, but dominant viewpoints are more likely to take precedence over alternative perspectives. It's crucial to differentiate between identifying these expressions, and endorsing them. The capability to automatically identify the linguistic causal relationships established in political communication allows for more efficient detection of false information, hate speech, and harmful content, which is our objective.

9 Acknowledgments

The authors thank the DFG (EXC number 2055 – Project number 390715649, SCRIPTS) for providing funding; 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

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).

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.

Neele Falk and Gabriella Lapesa. 2022. Scaling up Discourse Quality Annotation for Political Science. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*.

Jianfeng Fu, Zongtian Liu, Wei Liu, and Wen Zhou. 2011. Event causal relation extraction based on cascaded conditional random fields. *Pattern Recognition and Artiflcial Intelligence*, 24(4):567–573.

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023. Is ChatGPT a good causal reasoner? a comprehensive evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11111–11126, Singapore. Association for Computational Linguistics.

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.

Jinghang Gu, Longhua Qian, and Guodong Zhou. 2016. Chemical-induced disease relation extraction with various linguistic features. *Database*, page 42.

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).

Matthew Ho, Aditya Sharma, Justin Chang, Michael Saxon, Sharon Levy, Yujie Lu, and William Yang

Wang. 2022. Wikiwhy: Answering and explaining cause-and-effect questions. *Preprint*, arXiv:2210.12152.

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

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. 2024. Can large language models infer causation from correlation? *Preprint*, arXiv:2306.05836.

Vivek Khetan, Somin Wadhwa, Byron Wallace, and Silvio Amir. 2023. SemEval-2023 task 8: Causal medical claim identification and related PIO frame extraction from social media posts. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2266–2274, Toronto, Canada. Association for Computational Linguistics.

Canasai Kruengkrai, Kentaro Torisawa, Chikara Hashimoto, Julien Kloetzer, Jong-Hoon Oh, and Masahiro Tanaka. 2017. Improving Event Causality Recognition with Multiple Background Knowledge Sources Using Multi-Column Convolutional Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal Reasoning and Large Language Models: Opening a New Frontier for Causality. *arXiv preprint*. ArXiv:2305.00050 [cs, stat] version: 2.

Zhaoning Li, Xiaotian Zou, and Jiangtao Ren. 2021. Causality extraction based on self-attentive BiLSTM-CRF with transferred embeddings. *Neurocomputing*, 423:207 – 219.

Dominique Mariko, Hanna Abi Akl, Estelle Labidurie, Stephane Durfort, Hugues de Mazancourt, and Mahmoud El-Haj. 2021. The financial document causality detection shared task (FinCausal 2021). In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 58–60, Lancaster, United Kingdom. Association for Computational Linguistics.

Paramita Mirza. 2021. Event causality. In *Computational Analysis of Storylines: Making Sense of Events*, volume 106, pages 106–124. Cambridge University Press.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the TempEval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19, Gothenburg, Sweden. Association for Computational Linguistics.

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

Joe O'Connor and Jacob Andreas. 2021. What context features can transformer language models use? *CoRR*, abs/2106.08367.

Paul Reisert, Naoya Inoue, Tatsuki Kuribayashi, and Kentaro Inui. 2018. Feasible Annotation Scheme for Capturing Policy Argument Reasoning using Argument Templates. In *Proceedings of the 5th Workshop on Argument Mining*, pages 79–89, Brussels, Belgium. Association for Computational Linguistics.

Angelika Romanou, Syrielle Montariol, Debjit Paul, Leo Laugier, Karl Aberer, and Antoine Bosselut. 2023. CRAB: Assessing the strength of causal relationships between real-world events. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15198–15216, Singapore. Association for Computational Linguistics.

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

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. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*.

Fiona Anting Tan, Xinyu Zuo, and See-Kiong Ng. 2023. UniCausal: Unified Benchmark and Repository for Causal Text Mining. *arXiv preprint*. ArXiv:2208.09163 [cs].

Konstantin Vössing. 2023. Argument-stretching: (slightly) invalid political arguments and their effects on public opinion. *European Political Science Review*, pages 1–21. Publisher: Cambridge University Press.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? *Preprint*, arXiv:2307.02483.

Bei Yu, Yingya Li, and Jun Wang. 2019. Detecting Causal Language Use in Science Findings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4664–4674, Hong Kong, China.

Sendong Zhao, Ting Liu, Sicheng Zhao, Yiheng Chen, and Jian-Yun Nie. 2016. Event causality extraction based on connectives analysis. *Neurocomputing*, 173:1943–1950.

A Prompt

Prompt with example output
System: You are a causal language model that performs causal sequence classification and causal span detection. You will classify a text 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}.
User: But to pay for it, we had to take on debt, precipitated by massive reduction in Government revenue.
Assistant:{\n \"label\": 1,\n \"cause\": \"massive reduction in Government revenue\",\n \"effect\": \"had to take on debt\"\n}

B Models and model configurations

All Open AI models were run using the Open AI Batch API. Llama and Gemma models were accessed via the Transformers library from Hugging Face, and inference was run using the Together AI API. Gemini-1.5 was run using Google's AI Studio API. Available parameter specifications were established equally across the LLMs: temperature was set to 0.0, Top p to 1.0, Frequency, presence or repetition penalty were all set to 0.0. The maximum number of tokens was specified as 200. For Google models, the `HarmBlockThreshold` in the safety settings parameter was set to `None` for the first two experiments, and set to default for the Fake news and post-training events set of sentences in LCD evaluation. See Table 4 and Table 5 for full details.

Parameter	Value
Temperature:	0.0
Top p:	1.0
Top k:	1.0
Frequency penalty:	0.0
Presence penalty:	0.0
Repetition penalty:	0.0
Max body tokens:	200

Table 4: Model parameter specification. Other parameters that are not included in this table are set to their default values.

C Evaluation processing

To ensure reproducibility, each model's output was subjected to identical pre-processing and post-processing steps. The preprocessing involved token normalization and sequence truncation to maintain consistency across the models. Post-processing included error correction and format standardization. The metrics were selected to provide a comprehensive assessment of model performance across various dimensions of language understanding and generation. Detailed results and further discussion can be found in the subsequent sections of this appendix. We used the SpaCy library to process the text into IOB2 format. For CSD evaluation, we employed our own script available on our Github page. Finally, to analyze classification results, we used the "classification_report" function from Scikit-learn.

If the LLM did not provide a result, as in Null or a text that is not a label, it's was considered a incorrect classification. To analyze the models the missing values (NaN, Null, None, Missing, or empty) are changed to a -1 labeled, and given to the classification report function. For example sentence "So what are we missing?" from the PolitiCAUSE subset (id. 2914) was not parsed by Llama and Gemma models, and it's predicted label was set to -1 before processing the classification report.

D Results tables

Model	Context Window	Knowledge cut-off date	Parameters
		Open AI	
gpt-3.5-0125	16,385	Sep 2021	N.S.
gpt-4-2024-04-09	128,000	Dec 2023	N.S.
gpt-4o-2024-08-06	128,000	Dec 2023	N.S.
		Meta	
Meta-Llama-3.1-8B-Instruct-Turbo	128,000	Dec 2023	~8B
Meta-Llama-3.1-70B-Instruct-Turbo	128,000	Dec 2023	~70B
Meta-Llama-3.1-405B-Instruct-Turbo	128,000	Dec 2023	~405B
		Google	
gemma-2-9b-it	8,192	June 2024*	~9B
gemma-2-27b-it	8,192	June 2024*	~27B
gemini-1.5-pro-002	128,000	Sep 2024	N.S.

Table 5: Model details, including the full name used to access the models and their versions. Context window is determined by model provider. Knowledge cut-off dates are provided by AI companies (Dates with * are not official and taken from online sources.) Parameter size is rounded to the next billion (N.S. stands for Not Specified).

	Prec	Recall	F1-score	Prec	Recall	F1-score	Prec	Recall	F1-score
Model		CAUSE			EFFECT			AVERAGE	
GPT-3.5	0.37	0.43	0.37	0.36	0.47	0.38	0.36	0.45	0.37
GPT-4	0.63	0.71	0.64	0.62	0.75	0.64	0.62	0.73	0.64
GPT-4o	0.62	0.70	0.64	0.60	0.75	0.63	0.61	0.73	0.63
Llama-8b	0.40	0.46	0.40	0.41	0.45	0.40	0.40	0.45	0.40
Llama-70b	0.53	0.59	0.53	0.51	0.61	0.52	0.52	0.60	0.53
Llama-405b	0.57	0.63	0.58	0.56	0.64	0.57	0.57	0.64	0.57
Gemma-9b	0.36	0.46	0.38	0.36	0.50	0.39	0.36	0.48	0.38
Gemma-27b	0.36	0.46	0.38	0.36	0.50	0.39	0.36	0.48	0.38
Gemini-1.5	0.58	0.65	0.59	0.56	0.66	0.58	0.57	0.65	0.58

Table 6: Causal Span Detection results for PolitiCAUSE subset. The table includes the values of Precision, Recall and F1 score metrics for each span, and the average score, according to model.

Fake News Set

	CAUSE			EFFECT			AVERAGE		
Model	Prec	Recall	F1-score	Prec	Recall	F1-score	Prec	Recall	F1-score
GPT-3.5	0.28 (-0.09)	0.37 (-0.06)	0.30 (-0.07)	0.33 (-0.03)	0.32 (-0.15)	0.32 (-0.06)	0.31 (-0.05)	0.35 (-0.10)	0.31 (-0.06)
gpt-4o	0.66 (+0.03)	0.73 (+0.02)	0.67 (+0.03)	0.66 (+0.04)	0.66 (-0.09)	0.66 (+0.02)	0.66 (+0.04)	0.70 (-0.03)	0.67 (+0.03)
GPT-4o	0.65 (+0.03)	0.72 (+0.02)	0.67 (+0.03)	0.69 (+0.09)	0.68 (-0.07)	0.68 (+0.05)	0.67 (+0.06)	0.70 (-0.03)	0.67 (+0.04)
Llama-8b	0.36 (-0.04)	0.38 (-0.08)	0.36 (-0.04)	0.39 (-0.02)	0.38 (-0.07)	0.37 (-0.03)	0.37 (-0.03)	0.38 (-0.07)	0.37 (-0.03)
Llama-70b	0.49 (-0.04)	0.55 (-0.04)	0.50 (-0.03)	0.50 (-0.01)	0.51 (-0.10)	0.50 (-0.02)	0.49 (-0.03)	0.53 (-0.07)	0.50 (-0.03)
Llama-405b	0.47 (-0.1)	0.55 (-0.08)	0.49 (-0.09)	0.51 (-0.05)	0.49 (-0.15)	0.49 (-0.08)	0.49 (-0.08)	0.52 (-0.12)	0.49 (-0.08)
Gemma-9b	0.39 (+0.03)	0.50 (+0.04)	0.41 (+0.03)	0.40 (+0.04)	0.39 (-0.11)	0.39 (0.00)	0.40 (+0.04)	0.44 (-0.04)	0.40 (+0.02)
Gemma-27b	0.39 (+0.03)	0.50 (+0.04)	0.41 (+0.03)	0.40 (+0.04)	0.39 (-0.11)	0.39 (0.00)	0.40 (+0.04)	0.44 (-0.04)	0.40 (+0.02)
Gemini-1.5	0.64 (+0.06)	0.69 (+0.04)	0.64 (+0.05)	0.64 (+0.08)	0.65 (-0.01)	0.63 (0.05)	0.64 (+0.07)	0.67 (+0.02)	0.64 (+0.06)

Post-training events Set

	CAUSE			EFFECT			AVERAGE		
Model	Prec	Recall	F1-score	Prec	Recall	F1-score	Prec	Recall	F1-score
GPT-3.5	0.53 (+0.16)	0.51 (+0.08)	0.50 (+0.13)	0.50 (+0.14)	0.49 (+0.02)	0.48 (+0.10)	0.51 (+0.15)	0.50 (+0.05)	0.49 (+0.12)
GPT-4	0.70 (+0.07)	0.72 (+0.01)	0.70 (+0.06)	0.71 (+0.09)	0.71 (-0.04)	0.70 (+0.06)	0.71 (+0.09)	0.71 (-0.02)	0.70 (+0.06)
GPT-4o	0.75 (+0.13)	0.72 (+0.02)	0.72 (+0.08)	0.72 (+0.12)	0.73 (-0.02)	0.72 (+0.09)	0.74 (+0.13)	0.73 (0.00)	0.72 (+0.09)
Llama-8b	0.51 (+0.11)	0.41 (-0.05)	0.44 (+0.04)	0.51 (+0.10)	0.47 (+0.02)	0.48 (+0.08)	0.51 (+0.11)	0.44 (-0.01)	0.46 (+0.06)
Llama-70b	0.63 (+0.10)	0.59 (0.00)	0.60 (+0.07)	0.64 (+0.13)	0.61 (0.00)	0.61 (+0.09)	0.63 (+0.11)	0.60 (0.00)	0.60 (+0.07)
Llama-405b	0.72 (+0.15)	0.67 (+0.04)	0.68 (+0.10)	0.70 (+0.14)	0.69 (+0.05)	0.69 (+0.12)	0.71 (+0.14)	0.68 (+0.04)	0.68 (+0.11)
Gemma-9b	0.52 (+0.16)	0.46 (0.00)	0.47 (+0.09)	0.50 (+0.14)	0.49 (-0.01)	0.49 (+0.10)	0.51 (+0.15)	0.48 (0.00)	0.48 (+0.10)
Gemma-27b	0.52 (+0.16)	0.46 (0.00)	0.47 (0.09)	0.50 (+0.14)	0.49 (-0.01)	0.49 (+0.10)	0.51 (+0.15)	0.48 (0.00)	0.48 (+0.10)
Gemini-1.5	0.40 (-0.18)	0.38 (-0.27)	0.38 (-0.21)	0.40 (-0.16)	0.39 (-0.27)	0.39 (-0.19)	0.40 (-0.17)	0.38 (-0.27)	0.38 (-0.20)

Table 7: LCD CSD results for both Fake News and Post-training events sentence sets. The table includes the values of Precision, Recall and F1 score metrics for each span, and the average score, according to model.

E LCD data and results

Fake News Sentences Set

1. CDC study finds overwhelming majority of people getting coronavirus wore masks.
2. Major Doctor's organization publicly embraces dismembering babies until birth.
3. High doses of vitamin C can cure COVID-19.
4. Chickens are not laying eggs, because RNA is being added to commercial chicken feed.
5. Disney is attempting to lower the minimum drinking age on Disney property to increase their revenue at EPCOT.
6. In January, Spetznas rescued 50 imprisoned, emaciated children from an "Adrenochrome farm" near Shostka, Ukraine.
7. Their captors had so frequently drained their brittle bodies of blood and adrenal fluid that their survival depended on being kept in a medically induced coma.
8. The U.S. Capitol police gave the protesters an "okay" to enter the Capitol.
9. The death was a filmed public execution of a black man by a white cop, with the purpose of creating racial tensions and driving a wedge in the growing group of anti deep state sentiment from common people.
10. Tonight... we move into the residential areas... the white hoods.... and we take what's ours.
11. You will test positive for COVID-19 if you've gotten flu shots during the past ten years.
12. Doctor Stella Immanuel proclaims that a cocktail of hydroxychloroquine, zinc and the antibiotic azithromycin has cured COVID-19 patients.
13. Wisconsin's dairy farms were decimated under the Obama/Biden administration but are doing very well.
14. If Joe Biden's elected, he'll end fracking.
15. Clintons, the Obamas & the Biden family are all involved in child trafficking and crimes against children.
16. Biden went to Ukraine and threatened to withhold $1 billion in aid if they did not fire the prosecutor that was investigating his son and the company that his son worked for.
17. For those wondering if it's worth impeaching him this time, it means he loses his 200K+ pension for the rest of his life.
18. The vaccine contains a spike protein called syncytin-1, vital for the formation of the human placenta.
19. The White House is freaking out after Myanmar Military arrests political leaders for Election Fraud in their November 8 elections.
20. The Trump administration made no effort to get U.S. medical experts into China as the novel coronavirus epidemic spread there early this year.
21. Bernie Sanders calls for a tax rate of 52% on incomes of $29,000 or more to pay for his Medicare for All plan.
22. Dr. Lieber charged for lying about his participation in a Chinese recruitment program and his affiliation with a Chinese university.
23. Do the COVID-19 vaccines cause infertility?
24. The Biden administration just dished out 86 million dollars for motel rooms to house illegal immigrants.
25. Congrats society, we're so woke that we have segregation again.
26. The West is now relying on his wife Julia instead of Alexei Navalny.
27. Germany ran its own military-biological program in Ukraine and was involved in research on biological weapons.
28. Ukraine is committing genocide against the Russian-speaking population in eastern Ukraine.
29. The Bucha massacre was staged by Ukrainian forces.
30. Western governments have imposed sanctions on Russia because of their Russophobia, and Russians are experiencing systematic discrimination in the West.
31. COVID came from China and the vaccine also came from China, don't trust China!
32. Compare this with Russia and China using the pandemic excuse to expand their influence and profit even though the Russian vaccine is ineffective and the Chinese vaccine contains pork gelatin.
33. Muslim scientists from the Raza Academy in Mumbai reported that the Chinese coronavirus vaccine contains gelatin from pork and recommended against vaccination with the haram vaccine.
34. Turkmenistan residents report that the Chinese

vaccine causes severe side effects.

35. Taiwanese travelers stranded at Osaka's Kansai International Airport were rescued by Chinese Embassy.
36. Bolsonaro made a pact with a Masonic sect to win the election.
37. Lula paid his own wife to kill his own son.
38. Brazil Supreme Court Justice Edson Fachin was a lawyer for the MST.
39. Russia unleashed more than 500 lions on its streets to ensure that people are staying indoors during this pandemic outbreak.
40. This document form March 5, from the Government of Spain, forced all nursing homes to keep all seniors with covid-19 symptoms locked in their rooms, keeping them from going to the hospital to get treated.
41. There are 500 boats in Argelia about to sail to Spain… 5,000 illegal immigrants will arrive shortly, many of them infected.
42. Bill Gates owns the patent for the coronavirus.
43. Qatar supported extremist organizations with more than $64 billion over the years!
44. Saudi Crown Prince Mohammed bin Salman had been forced out of power.
45. He announced that Mexican President lowered his own salary and those of 35,000 government employees by 60% to increase pensions for citizens.
46. Any beneficiary of Bolsa Família will lose their benefits if they work as election officials.
47. The Pfizer vaccine produces a fever for 8 to 12 days and after recovery masks are no longer needed.
48. We don't need to continue being named the Republic of Chile or to continue using the Pinochet-imposed flag.
49. Argentine-produced ammunition found in the autopsies of 22 Bolivians killed under Interim President Jeanine Áñez.
50. When you go vote next June 6th, use your own pen or marker, their using all sorts of tricks to commit electoral fraud.

Post-training events Sentences Set

1. Moldovans have received anonymous death threats to scare them from voting.
2. The sharp westward shift in Moldova irked Moscow and significantly soured relations with Chișinău.
3. Spain is still reeling from the deadly impact of its worst flooding disaster in decades where at least 158 people are confirmed dead and dozens are missing.
4. A year's worth of rain fell in eight hours in parts of Valencia on Tuesday.
5. While DANAs aren't unusual in the region, the Mediterranean has seen record-breaking warm waters this summer.
6. This week the German coalition collapsed after Scholz decided to fire some of his key ministers.
7. Scholz's government no longer has a majority in parliament as the Traffic-light coalition collapses.
8. However, the opposition could force Scholz out earlier if they can find a majority for an alternative chancellor.
9. Horrific scenes in Quetta after a suicide bomber targeted passengers waiting to board an express train.
10. We expect many injured after the suicide bombing attack at Peshawar-bound Jaffar Express.
11. Following years of stalemate, the breakthroughs in Baku have now begun, here, at COP29.
12. To enable action, Mr. Babayev identified agreement on a fair and ambitious New Collective Quantified Goal (NCQG) on climate finance as the top negotiating priority for COP29.
13. Mr. Babayev stressed that as the first Paris decade comes to a close, COP29 is a moment of truth that will test our commitment to the multilateral climate system.
14. The ICC has issued an arrest warrant for Benjamin Netanyahu for alleged Gaza war crimes.
15. It is the first time that leaders of a democracy and western-aligned state have been charged by the court, in the most momentous decision of its 22-year history.
16. The United States has been clear that the ICC does not have jurisdiction over this matter.
17. Israel has issued an overnight curfew to people in Lebanon seeking to return to southern Lebanon following the truce.
18. People displaced in the conflict immediately travelled back to southern Lebanon resulting in enormous traffic jams throughout the day.
19. Namibia celebrates elections after first female president is elected.
20. Windhoek is reported to be calm on Wednesday, with neither celebrations nor protests and people carrying on with their normal lives.
21. The South Korean president is facing impeachment vote as defense minister offers to resign.
22. In a shock TV speech on Tuesday, President Yoon Suk Yeol decided to impose martial law.
23. When martial law was briefly declared in South Korea we briefly saw armed soldiers entering the National Assembly.
24. New York City police have launched a manhunt for a masked suspect who gunned down the head of a US medical insurance giant.
25. UnitedHealthcare chief executive was fatally shot outside the Hilton Hotel in Midtown Manhattan.
26. The bells of Notre Dame Cathedral rang for the first time as rebuilding has finalized.
27. Watching the Notre Dame's door swing open to a burst of choir song was such an emotional moment for the thousands of people taking in the display from just outside the cathedral.
28. While the eight bells of Norte Dame were not damaged by the fire, they are ringing for the first time since then.
29. Syrian rebels seized the capital Damascus unopposed on Sunday after a lightning advance that sent President Bashar al-Assad fleeing to Russia.
30. Moscow gave asylum to Assad and his family.
31. The sudden overthrow at the hands of a revolt partly limits Iran's ability to spread weapons to its allies and could cost Russia its Mediterranean naval base.
32. Police don't know how the driver in Magdeburg was able to circumvent the barriers that were protecting the market.
33. We informed the public that a driver plowed a vehicle into a Christmas market killing an adult and a small child.

34. It crashed on the opposite shore of the Caspian after an emergency that was caused by a bird strike.
35. One of the Azerbaijani sources disclosed results that showed the plane was struck by a Russian Pantsir-S air defense system.
36. Russian air-defense system downed Azerbaijan plane.
37. The WHO is aware of the unidentified disease and is sending a team to Congo.
38. The press reported that a flu-like disease that has killed dozens of people over two weeks in Congo is being investigated.
39. The European Union will press ahead with hefty tariffs on China-made electric vehicles even after the bloc's largest economy Germany rejected them, exposing a rift over its biggest trade row with Beijing in a decade.
40. The proposed duties on EVs built in China of up to 45
41. Shares in European carmakers Renault and Volkswagen rose on hopes the tariffs will help them compete with Chinese rivals on their home turf.
42. Her inauguration launches a six-year term during which she will navigate Mexico's all-important relationship with its northern neighbor.
43. Sheinbaum's election broke barriers, as she's the first woman to lead Mexico is also Mexico's first president of Jewish ancestry.
44. Dockworkers at ports from Maine to Texas began walking picket lines early Tuesday in a strike over wages and automation that could reignite inflation and cause shortages of goods if it goes on more than a few weeks.
45. The union wants a complete ban on automation, and it isn't clear just how far apart both sides are.
46. Supply chain experts say consumers won't see an immediate impact from the strike because most retailers stocked up on goods, moving ahead shipments of holiday gift items.
47. Nihon Hidankyo has won the Nobel Price for it's work on nuclear disarmament.
48. This year's Nobel laureates used tools from physics to construct methods for machine learning.
49. The Nobel Prize in Chemistry 2024 is about proteins, life's ingenious chemical tools.
50. In her oeuvre, Nobel laureate, Han Kang, confronts historical traumas and invisible sets of rules and exposes the fragility of human life.

The sentences were manually selected from reputable sources for both Fake News detection and for Real world events. Sources for each sentence are included in the CSV files of each set list.

Qualitative Analysis

Fake News Set Analysis	Cases
GPT-4	**2**
True causal, produces 0, null, null results, no explanation (but other models provide a warning).	2
GPT-4o	**1**
True causal, produces 0, null, null results, no explanation (but other models provide a warning).	1
Llama-8b	**4**
Recognizes causal claim and produces 0, null, null, because it's humorous not factual.	1
Recognizes causal claim, and parses, and warns that it's a complex social issue, sensitive topic or false claim.	1
Recognizes causal claim, and parses, and warns that its not backed by scientific evidence.	1
Unfulfilled requests.	1
Llama-70b	**1**
Recognizes causal claim, and parses, and warns that it's a complex social issue, sensitive topic or false claim.	1
Llama-405b	**4**
Recognizes causal claim and produces 0, null, null, because it's not backed by scientific evidence.	1
Recognizes causal claim, and parses, and warns that it's a complex social issue, sensitive topic or false claim.	2
Unfulfilled requests.	1
Gemma-9b	**11**
Recognizes causal claim and produces 0, null, null, because its not backed by scientific evidence.	2
Recognizes causal claim, and parses, and warns that it's a complex social issue, sensitive topic or false claim.	4
Recognizes causal claim, and parses, and warns that it's not backed by scientific evidence.	2
Unfulfilled requests.	1
Warns it's a conspiracy theory produces 0, null, null.	2
Gemma-27b	**11**
Recognizes causal claim and produces 0, null, null because its a sensitive topic.	2
Recognizes causal claim and produces 0, null, null, because it's humorous not factual.	1
Recognizes causal claim and produces 0, null, null, because its not backed by scientific evidence.	3
Warns it's a conspiracy theory produces 0, null, null.	5
Gemini-1.5	**3**
True causal, produces null results, no explanation (but other models provide a warning).	3

Post-training events Set Analysis	Cases
GPT-4	**3**
Produced null results for a sentence is positive about a sensitive topic.	3
GPT-4o	**3**
Produced null results for a sentence is positive about a sensitive topic.	3
Llama-8b	**2**
Produced null results for a sentence is positive about a sensitive topic.	2
Llama-405b	**1**
Produced null results for a sentence is positive about a sensitive topic.	1
Gemini-1.5	**3**

Table 8: Qualitative analysis for LCD