Does Topic Sentiment Cause Perceived Ideology? Comparing Human and LLM Annotations in Political News Articles

Upasana Chatterjee
Department of Computer Science
Columbia University
uc2143@columbia.edu

Abstract

We ask whether topic sentiment has a causal effect on perceived political ideology, and whether the answer depends on who assigns the ideology label. Using articles from AllSides, paired with shared sentiment annotations from Llama-3.3-70b-versatile, we compare ideology labels from expert human annotators, GPT-4o-mini (baseline and finetuned), and Llama-3.3-70B. We apply Double Machine Learning (DML) and community-level mediation analysis across all four annotation paradigms. Human annotations yield no significant causal effects at the community level. Fine-tuned GPT-4o-mini achieves the highest classification accuracy (F1=72.48) and is the only annotator paradigm that produces significant community-level treatment effects and significant natural direct effects (NDEs) in mediation. We interpret this as evidence of shortcut learning: fine-tuning on ideology-labeled data causes the model to internalise a spurious sentiment–ideology coupling not operative in human judgment for this task. This coupling is structurally invisible to F1-based evaluation, with implications for the use of LLM annotations as silver labels and as proxies for human judgment in downstream causal analyses.

1 Introduction

There is growing interest in using large language models (LLMs) as low-cost, scalable surrogates for human participants in social science research (Horton, 2023; Li et al., 2024). This work suggests that LLMs may replicate aspects of human behavioral and judgmental responses in ways that make them useful proxies for study participants.

However, validating that LLM *outputs* correlate with human outputs does not establish that LLMs arrive at those outputs through the same underlying processes. Gao et al. (2025) provide a cautionary empirical demonstration: even in a simple economic game, advanced LLMs fail to replicate human behavior distributions, with failure modes that are inconsistent and difficult to predict. This raises a deeper question about *causal fidelity*: whether the features that causally drive LLM predictions are the same features that causally drive human judgments. Existing validation studies assess output-level agreement; none, to our knowledge, test whether the causal structure of LLM decisions mirrors that of human decisions.

Concretely, focusing on topic-level sentiment as a specific, interpretable feature class, we define a causal question: *What is the causal effect of topic sentiment on perceived ideological stance in news articles?* Following Pearl (2001), we operationalize this question by defining appropriate counterfactuals and using a combination of NLP techniques and causal inference methods. We hold sentiment annotations constant across all experimental conditions and vary only the ideology labels, comparing expert human annotations from AllSides with predictions from GPT-4o-mini (baseline and finetuned) and Llama-3.3-70B. This design isolates differences attributable to the annotation source itself, enabling a direct test of causal fidelity.

We apply our framework to a subset of the AllSides expert-labelled test articles using Double Machine Learning (DML; Chernozhukov et al., 2018) for causal effect estimation and mediation analysis (Pearl, 2001) to decompose direct and indirect pathways. We find that topic sentiment has a significant causal effect on predicted ideology for LLM annotators, particularly the finetuned GPT-4o-mini model, but not for expert human annotators. Mediation analysis reveals significant natural direct effects (NDEs) for finetuned GPT annotations that are absent in the other three paradigms (baseline GPT-4o-mini, Llama-3.3-70B, and expert human annotations), with directional sign reversals in total effects between human and LLM annotators across multiple communities. These findings suggest that, depending on topic groupings, LLM

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 725–740

ideology predictions are more tightly coupled to sentiment-correlated textual features than human judgments, a difference invisible to output-level comparison.

Our contributions are as follows:

1. We propose a causal framework for testing annotation fidelity (comparing human and LLM annotators by holding sentiment constant and varying only ideology labels) that directly probes whether LLMs replicate the *causal structure* of human judgments rather than merely their outputs.

2. We demonstrate that LLM-predicted ideology labels surface significant sentiment-ideology causal effects that are absent in expert human annotations, with implications for the use of LLM-generated silver labels in downstream causal analyses.

2 Related Work

Sentiment and Ideology in Media. The relationship between sentiment, topic coverage, and political ideology has been studied across multiple domains. Smirnova et al. (2017) demonstrate that sentiment patterns in news coverage of specific topics correlate with ideological positioning, while Bhatia and Deepak (2018) show that topic-specific sentiment analysis can help predict political ideology. Bestvater and Monroe (2023) draw a distinction between sentiment and stance, arguing that target-aware classification is necessary for political text analysis. Beyond news media, similar sentiment-ideology relationships have been observed in congressional speech (Gentzkow et al., 2019) and legislator tweets (Spell et al., 2020). Our work differs from these studies in that we move beyond correlational analysis to estimate causal effects of topic sentiment on ideology, and we compare these effects across human and LLM annotation sources.

Causal Inference in NLP. Causal inference methods have been applied to text analysis across a range of settings. Feder et al. (2022) provide a comprehensive survey of causal inference in NLP, covering settings where text serves as treatment, outcome, or confounder. Veitch et al. (2020) propose using text embeddings to adjust for confounding in causal inference, while Keith et al. (2021) develop a framework for text as causal mediator, estimating direct and indirect effects of social group signals through language. Tierney and Volfovsky (2021) apply causal mediation analysis to political polarization through text. In parallel, causal methods have been applied to understand LLM behavior, including gender bias detection via causal mediation analysis (Vig et al., 2020), mechanistic interpretation of arithmetic reasoning (Stolfo et al., 2023), and assessment of LLM comprehension (Han et al., 2025). Our work contributes to this literature by applying causal inference not to explain model internals, but to diagnose systematic differences between human and LLM annotation behavior.

LLMs as Human Surrogates. A growing body of research explores whether LLMs can substitute for human participants in social science tasks (Horton, 2023; Li et al., 2024; Ma et al., 2025; Wu et al., 2023; Strachan et al., 2024). Despite this promise, Gao et al. (2025) provide a cautionary counterpoint: in an economic game requiring strategic reasoning, LLMs consistently fail to replicate the human behavior distribution, with failure modes that are inconsistent across models and input variations. These studies evaluate output-level agreement between LLMs and humans; our work complements them by asking whether the *causal structure* of LLM decisions, specifically which features drive predictions, mirrors that of human annotators.

Double Machine Learning. Our causal estimation relies on Double Machine Learning (DML), introduced by Chernozhukov et al. (2018), which provides root-N consistent estimates of treatment effects in the presence of high-dimensional confounders by using cross-fitting and Neyman-orthogonal score functions. DML is particularly suited to our setting because it accommodates the high-dimensional topic-presence confounders while allowing flexible first-stage models for the treatment and outcome nuisance functions.

3 Experimental Setup

3.1 Dataset and Expert Human Baseline

AllSides tags both news sources and individual articles with categorical ideology ratings (Left, Center, Right); both kinds of labels are assigned through multiple rounds of review and consensus-building among a team of in-house experts[1]. These AllSides article-level expert labels serve as our *expert human baseline* throughout the paper and the same

[1]Details on the AllSides news curation policy can be found at https://www.allsides.com/about/news-curation-principles.

labels are compared against each LLM paradigm under an identical experimental pipeline. Our article corpus consists of expert-labelled articles drawn from AllSides, building on the corpus released by Baly et al. (2020) and collected under the same news-curation protocol. To support reproducibility we release article-level metadata (source URL, outlet, ideology label, topic tags) together with the topic-sentiment annotations used in this paper, covering both the experimental subset and a broader collection of additional articles.[2] Our analyses operate on an $N = 1{,}265$ set of expert-annotated articles. This size reflects API budget constraints on the sentiment extraction and ideology classification pipelines; it was expanded after the original submission (and reflected in the linked dataset), but extension results are not included in this paper.

3.2 Causal Framework

We model the relationships between article content, sentiment, and ideology using the causal diagram in Figure 1, which formalizes our assumptions about the data-generating process and determines which variables must be controlled for to obtain unbiased causal estimates.

We define the following variables:

- X: (text) the full article text.
- T: (topic tags) AllSides topic tags (e.g. Politics, Government Efficiency, Foreign Affairs).
- F: (sentiment) sentiment toward each topic, inferred from article text using Llama. Details are provided in Section 3.4.
- Y: (perceived ideology) the ideology label assigned by the annotator (human and LLM).

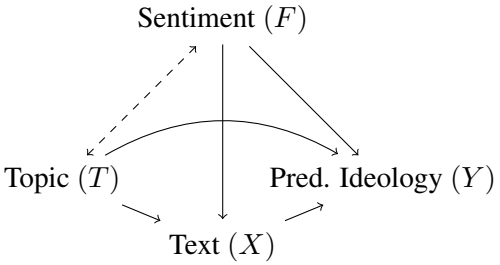

Figure 1: Complete causal diagram for article ideology classification. Note, we do not include article text in our investigations.

[2]Available at https://huggingface.co/datasets/upasanachatterjee/AllSides-sentiments

Each arrow represents a hypothesized causal direction. The arrow from X to Y reflects that article content affects perceived ideology. The arrow from F to X reflects that sentiment toward a topic influences the written text. The bidirectional dashed arrow between T and F captures the mutual relationship between topics and sentiment: certain topics tend to evoke particular sentiments, while prevailing sentiment patterns influence which topics receive coverage.

In practice, each node expands into many variables. Figure 2 illustrates this: each sentiment node F corresponds to a distinct topic, and sentiment toward one topic (F_T, the treatment) may influence sentiment toward other topics ($F_{M_1}, F_{M_2}, \ldots$, the mediators), which in turn affect ideology classification Y. The structural relationships *between* variable types remain as in Figure 1; what changes is the dimensionality *within* each node.

The diagram represents a *partial model of the labeling process*: for human annotators, an abstraction of one cognitive channel (sentiment toward salient topics) used in arriving at an ideology label; for LLM annotators, an abstraction of one channel of the labeling pipeline. We apply the same SCM to both paradigms and compare them through this shared structural lens. That a single structure fits both is itself an assumption, and one of the things this paper effectively probes. Many additional features plausibly enter both judgments (illustrated in Figure 2); we deliberately scope to topic-level sentiment to keep the experimental design tractable and the estimands interpretable.

Applying a single SCM to both paradigms means we are comparing per-annotator conditional distributions $p(Y \mid X, \text{annotator})$ under a shared structural assumption rather than claiming the underlying processes are the same. Human annotators bring deliberation, prior beliefs, source/outlet priors, and between-annotator consensus dynamics that the diagram does not encode; LLM annotators carry training-data composition, instruction tuning, architecture, and hyperparameter choices that are likewise absent. Both sets of factors are absorbed into the per-annotator estimator for simplicity, at the cost of losing some nuance. As a result, divergences in estimated effects across annotators reflect the combined contribution of these unmodeled factors operating *through* the sentiment channel we do model. A more complete account would require separate causal probing of each annotator's data-generating process, which we treat as future

work.

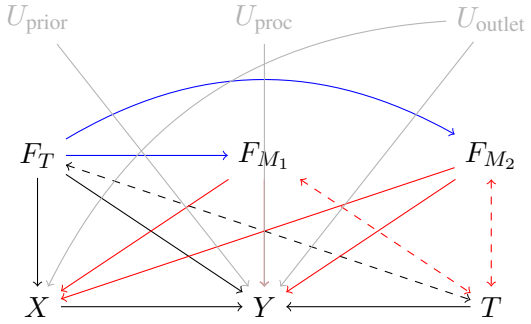

Figure 2: Expanded setup: sentiment variables as mediators. Blue arrows show treatment-to-mediator paths; red shows mediator-to-outcome paths. Grey nodes denote labeller-side factors absorbed into the per-annotator estimator and not modelled in this experiment.

The roles of each variable in the estimation framework are:

Treatment Variable (F_T): Sentiment toward a specific topic of interest (e.g., sentiment toward "Immigration" or "Healthcare"). This is the variable whose causal effect on ideology we seek to estimate.

Outcome Variable (Y): The ideology label (Left/Center/Right) assigned to the article.

Mediator Variables ($F_{M_1}, F_{M_2}, \ldots$): Sentiment toward all other topics present in the article. These lie on the causal pathway between treatment sentiment and ideology: sentiment toward one topic may influence sentiment toward related topics, which in turn affects ideology classification. Blue arrows in Figure 2 show these paths.

Confounder Variables: T (topic presence) is the primary confounder, since the presence of a topic influences both sentiment toward it and the overall ideological character of the article.

We additionally restrict the analysis to articles containing the treatment topic, preventing invalid comparisons between articles that discuss a topic and those that do not. This design supports two complementary families of causal estimands. The Average Treatment Effect (ATE) captures the overall causal impact of topic-community sentiment on ideology prediction, averaged across all articles in a community. Mediation analysis (Pearl, 2001) decomposes this total influence into a Natural Direct Effect (NDE), the portion of sentiment's impact that operates independently of other topic sentiments, and a Natural Indirect Effect (NIE), the portion that is mediated through co-occurring topic sentiments; the two sum to the Total Effect (TE). Together, these estimands allow us to ask not only whether sentiment toward a topic is causally associated with ideology prediction, but also through which pathways that association operates. Comparing human and LLM-derived estimates across both ATE and mediation quantities can reveal differences in how annotation sources represent these causal structures.

Two experimental notes that we want to make explicit:

- F is observational. The treatment is a sentiment summary statistic extracted from X, not a manipulation of the article itself. Identification leans on conditional ignorability given T together with DML's nuisance modelling, not a natural-experiment design.

- Under greedy decoding, Y is a deterministic function of X for a fixed annotator, so reported variability is across-article rather than within-article.

3.3 LLM Selection and Performance

LLM	F1-Macro
Finetuned GPT-4o-mini	72.48
GPT-4o-mini	50.07
Llama-3.3-70B-versatile	54.61

Table 1: Baseline LLM performance on AllSides political ideology classification.

Table 1 shows the best model performance, evaluated against the expert human baseline as the ground truth. LLM evaluation setup details are provided in Appendix A.1.

We tested a few different large language models from the GPT and Llama families on the ideology prediction task and selected the `gpt-4o-mini-2024-07-18` model snapshot for additional finetuning. This model was selected as it offered the most favorable balance between classification accuracy (as demonstrated in our baseline experiments) and computational cost among available OpenAI models. Details about the finetuning process and hyperparameters are provided in Section A.4.

3.4 Topic Sentiment Extraction

To extract sentiment variables (F), we apply a two-step process to the full article text (X): first, named entities and key concepts are extracted and assigned sentiment polarity on a continuous scale from -1.0 (strongly negative) to $+1.0$ (strongly positive); second, each entity is associated with predefined topic tags assigned by the AllSides team, and entity sentiment scores are aggregated by topic to produce a per-topic sentiment profile for each article. Full model and implementation details are provided in Appendix A.2.

3.5 Community Detection

To identify the topical structure of the article corpus, we apply the Louvain community detection algorithm to a co-occurrence graph of topic tags, where nodes are tags and edge weights reflect how often two tags appear together in the same article. Louvain is well-suited to this setting because it maximises modularity directly on the co-occurrence graph, grouping topics that frequently co-appear into the same community. Since co-occurrence is a direct measure of how strongly two topic treatment columns move together across articles, the resulting communities correspond to natural clusters of correlated treatments. Graph construction details are provided in Appendix A.3.

For our community-level multi-treatment DML experiments, we limited our analysis to communities where each topic has at least 5 articles, leaving nine communities.

3.6 Experimental Settings

We use `LinearDML` (Chernozhukov et al., 2018) with PCA-reduced confounders and continuous sentiment scores, with bootstrap confidence intervals ($B = 2000$). Full experimental settings are provided in Appendix A.3.

4 Experiments and Results

All experiments are conducted across four annotation paradigms: the LLMs introduced in Table 1 and expert human-annotators. This multi-way comparison allows us to examine whether causal relationships are stable across annotation approaches or vary systematically between human judgment and large language models[3].

[3]Code available at https://github.com/upasanachatterjee/causal-inference-on-text

Estimation Details. The outcome variable Y is encoded as an ordinal (Left $= 0$, Center $=$ 1, Right $= 2$) and treated as continuous, with `RandomForestRegressor` as the nuisance model for both outcome and treatment. This encoding assumes equal spacing between adjacent ideology categories: a unit ATE of $+1.0$ corresponds to a one-category rightward shift in expected ideology (e.g., ATE $= -0.108$ represents a 10.8% shift toward Left). EconML's `LinearDML` handles cross-fitting of nuisance models internally.

Significance is determined by whether the 95% confidence interval (CI) excludes zero, equivalent to a two-sided test at $\alpha = 0.05$. To construct these intervals, we use non-parametric bootstrap percentile CIs ($B = 2{,}000$).

We focus on community-level aggregation of topics, which increases articles per treatment unit and yields more reliable estimates than topic-level analysis. Topic-level results are presented in Appendix C.

4.1 Community-level Aggregation

Instead of treating each topic individually, we aggregate sentiments within detected topic communities. The community sentiment F_C is the mean sentiment over member topics present in an article, and equals zero when no member topic appears. Let $\mathbf{F}_C = (F_{C_1}, \ldots, F_{C_L})$ be the community-aggregated treatment matrix. The marginal ATE for each community C_ℓ is

$$ATE_{C_\ell} = E\left[\frac{\partial Y}{\partial F_{C_\ell}}\right], \quad \forall\, \ell = 1, \ldots, L, \quad (1)$$

estimated via a single `LinearDML` model with $\mathbf{F}_C$ as the treatment matrix and $\tilde{\mathbf{T}}$ as confounders, following the same structure as the topic-level ATE.

The human and Llama annotations yield no significant results. The finetuned GPT annotations reveal significant effects for Communities 3, 9, and 11 and baseline GPT shows significant effects for Community 1, as seen in Figure 3. We present and discuss the full community-level ATE results in Appendix B.2.

The topics within each significant community reveal the following groupings and effect directions:

- Community 1 (Electoral Politics: topics include Bernie Sanders, presidential elections, mail-in voting): Positive sentiment shifts baseline GPT ideology predictions left (ATE $= -0.015$; $\approx$1.5% of one category step toward Left per unit sentiment increase)

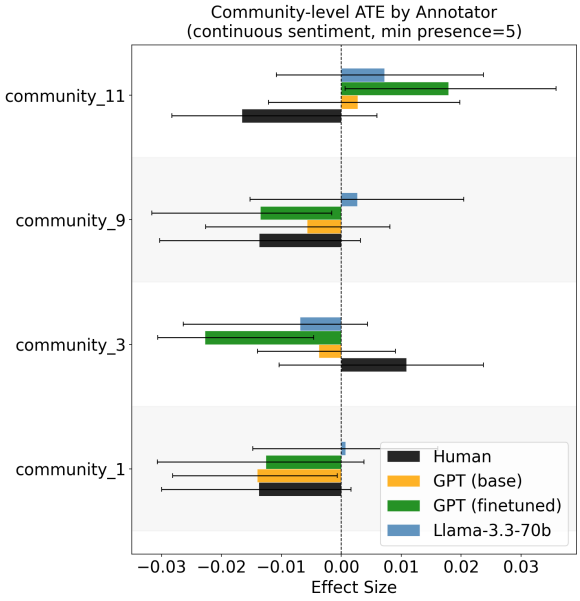

Figure 3: Community-level ATE.

- Community 3 (Social Issues: topics include immigration, LGBTQ, abortion, religion): Positive sentiment shifts fine-tuned GPT ideology predictions left (ATE $= -0.023$; $\approx$2.3% of one category step toward Left)

- Community 9 (Geopolitics and Foreign Affairs: topics include foreign policy, climate, global affairs): Positive sentiment shifts fine-tuned GPT ideology predictions left (ATE $= -0.013$; $\approx$1.3% of one category step toward Left).

- Community 11 (Trump Presidency: topics include January 6, Trump, impeachment hearings): Positive sentiment shifts fine-tuned GPT ideology predictions right (ATE $= +0.018$; $\approx$1.8% of one category step toward Right)

4.2 Community-level Mediation

For community-mode mediation, let C be the treatment community and C' the mediator community. The community sentiment score F_C is the mean sentiment over member topics present in an article; it equals zero when no member topic appears. Let $f_1 = Q_{75}$ and $f_0 = Q_{25}$ be the interquartile contrast values of F_C. The estimands are:

$$\text{TE}(f_1, f_0;\, Y) = E[Y_{f_1}] - E[Y_{f_0}], \tag{2}$$

$$\text{NDE}(f_1, f_0;\, Y) = E\left[Y_{f_1,\, F_{C'}(f_0)}\right] - E[Y_{f_0}], \tag{3}$$

$$\text{NIE}(f_1, f_0;\, Y) = \text{TE}(f_1, f_0;\, Y) \tag{4}$$

$$- \text{NDE}(f_1, f_0;\, Y), \tag{5}$$

where $F_{C'}(f_0)$ is the natural value of the mediator community sentiment when $F_C = f_0$. Of the communities with significant ATEs, only the $(9 \to 11)$ and $(11 \to 9)$ pathways had sufficient article overlap for valid mediation analysis; we report these as our main results. See Appendix D for a full account of pathway selection.

The multi-treatment ATE from Section 1 is estimated over the full analysis subset. For Community 9 (Geopolitics and Foreign Affairs), approximately 98.8% of articles have $F_C = 0$ because no Community 9 topic is present. Although the DML confounder matrix W includes community-presence indicators, the estimation is dominated by the mass of zero-sentiment observations.

Mediation analysis, by contrast, restricts to the $N = 335$ articles where Community 9 is actually present. Within this subgroup, the $Q_{25} \to Q_{75}$ contrast spans a sentiment shift from -0.3 to 0.0 (a 0.3-unit change). The total effect of the finetuned GPT model within this subgroup is TE $= -0.108$, approximately $-0.108/0.3 \approx -0.36$ per unit—roughly 28$\times$ larger than the global ATE on a per-unit basis. The ATE averages a concentrated subgroup effect across the articles where Community 9 is never discussed; the mediation analysis isolates that effect within the engaged subpopulation.

Figure 4 summarizes results for Community 9 (Geopolitics and Foreign Affairs) as treatment and Community 11 (Trump Presidency) as mediator, for the $Q_{25} \to Q_{75}$ IQR contrast (sentiment $-0.3 \to 0.0$), using $B = 2{,}000$ bootstrap resamples. Full numerical results are in Appendix D.

Only the finetuned GPT model reaches significance: TE $= -0.108$ (approximately 5.4% of one category step toward Left) and NDE $= -0.105$. The NIE is negligible (-0.003, 3.2% of TE), ruling out the cascading interpretation: Community 9 sentiment does not shift ideology *through* changes in Community 11 sentiment. The effect operates through a direct pathway, showing how an annotator frames geopolitics and climate content independently predicts ideology classification, regardless of how they frame domestic political content. All

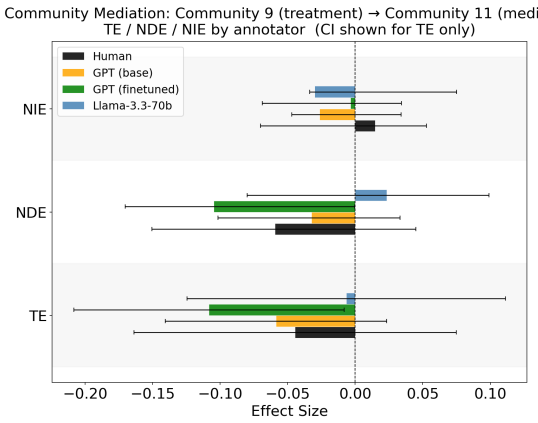

Figure 4: Community 9 → 11 mediation results. Error bars show 95% bootstrap CIs; filled markers indicate significance.

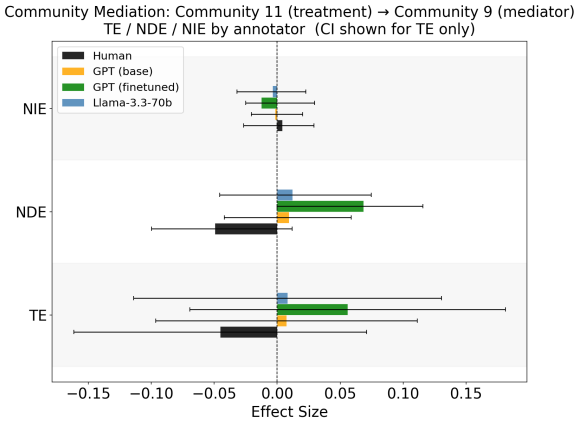

Figure 5: Community 11 → 9 mediation results. Error bars show 95% bootstrap CIs; filled markers indicate significance.

five annotators show a negative direction (positive geopolitics/climate sentiment → left-leaning prediction), confirming directional consistency; the four non-significant annotators reflect a power limitation at $N = 335$ rather than directional disagreement.

For the reverse direction (Community 11 as treatment, $Q_{25} \to Q_{75}$ contrast from sentiment -0.4 to 0.1), no annotator produces a significant TE. The finetuned GPT TE $= +0.056$ is the largest in magnitude but falls short of significance. The asymmetry between $9 \to 11$ (significant) and $11 \to 9$ (null) is structurally coherent: Community 9's effect on ideology is concentrated and direct; Community 11 is a larger community whose ideological influence is diffuse across many mediators. Full tabular results are in Appendix D.

5 Discussion

Our results reveal a consistent pattern: causal analysis applied to LLM-predicted ideology labels surfaces statistically significant sentiment-ideology relationships that are absent when the same analysis is applied to expert human annotations. As only the ideology labels vary across the datasets, these divergences reflect systematic differences in how LLMs and humans assign ideological stance to article text.

5.1 Fine-tuning Accuracy and Shortcut Learning

Prior work establishes that human annotators draw on metalinguistic framing cues, including sentiment toward salient entities and topics, when assigning ideological stance to article text (Nagy, 2007; Entman, 2010; Smirnova et al., 2017). Whether LLM annotators rely on sentiment in the same way is an open question: output-level metrics cannot tell us whether a model uses such cues at all, and if it does, whether it does so on the same topics and to a comparable extent as human annotators.

Our causal analysis asks (i) whether sentiment causally influences LLM ideology predictions, (ii) whether the topics on which this coupling appears match those for human annotators, and (iii) whether the effect sizes and the balance of direct versus mediated pathways are comparable across annotators.

Read in these terms, the Community 9→11 configuration (Figure 4) shows a near-total direct effect for finetuned GPT (NDE $= -0.105$, NIE $= -0.003$, 3.2% mediated): all five annotators are negative in direction, but only finetuned GPT reaches significance, and the human point estimate is consistent with a weaker but non-zero coupling that we are underpowered to detect at this per-community sample size. The substantive divergence is thus in the strength and directness of the sentiment-to-ideology pathway rather than in its presence: finetuned GPT routes sentiment to the label almost entirely through a direct edge, with negligible mediation through downstream framing.

This pattern is consistent with shortcut learning (Yuan et al., 2024; Zhou et al., 2024), where fine-tuning on AllSides ideology labels exposes the model to a training distribution in which framing cues are statistically correlated with ideology, and the model amplifies a coupling that humans appear to use in weaker, more mediated form. The validity criterion this implies for silver labelling is alignment, not accuracy: an LLM annotator is a

defensible stand-in for human annotators in downstream causal analysis to the extent that its causal effects (sign, magnitude, and mediation profile) mirror those derived from human labels.

Crucially, fine-tuning is not uniformly harmful on this criterion. The community-level ATEs in Appendix B (Table 5) show the effect of fine-tuning on human-LLM alignment is community-dependent. For Community 1 (Electoral Politics), zero-shot GPT produces a significant effect (-0.015^{*}) that fine-tuning attenuates (-0.013, n.s.) toward the human point estimate (-0.014); for Community 9 (Foreign Affairs and Climate), the zero-shot effect is muted (-0.006) and fine-tuning increases its magnitude (-0.013^{*}) into close agreement with the human estimate (-0.014). By contrast, fine-tuning produces sign reversals relative to humans in Communities 3 and 11, where the shortcut interpretation is most apt. Fine-tuning can therefore tighten or loosen alignment with the human causal structure depending on the topic domain, and the Community 9→11 mediation result should be read as a localised diagnostic of overblown direct routing.

5.2 Implications for LLM as Human Annotator Replacement

A growing line of work explores the potential of LLMs as replacements for human study participants (Horton, 2023; Ma et al., 2025; Gao et al., 2025). Our findings caution against this substitution in annotation contexts where downstream causal analyses are planned. A fine-tuned LLM annotator may encode a particular pattern of ideology assignment that does not reflect the full range of human judgments. The fine-tuned GPT model has internalised a specific annotator profile: one that more strongly associates positive social/cultural sentiment with left-leaning ideology than human annotators do.

This matters because if LLM annotations are used as silver-standard training data or as inputs to causal analyses, the spurious pathways learned during fine-tuning will propagate. A researcher who trusts the fine-tuned labels would conclude that geopolitics/foreign affairs framing (Community 9) causally shifts ideology perception; a researcher using human labels would find no such effect. The finding that higher classification accuracy does not guarantee preservation of causal structure should inform how researchers evaluate LLM annotators for tasks where causal or correlational downstream analyses are planned.

5.3 Differences Across LLM Families

The three LLM annotators exhibit markedly different causal profiles despite all predicting ideology from the same raw article text, suggesting that architecture and training data differences shape which causal pathways models rely on during inference.

Baseline GPT and finetuned GPT share the same underlying architecture, so the divergence between them is attributable entirely to fine-tuning. baseline GPT produces a significant ATE only for Community 1 (Electoral Politics; ATE $= -0.015$; $\approx$1.5% of one category step toward Left), with no significant effects for Communities 9 or 11 and no mediation effects. Fine-tuning corrects some of this centre bias but introduces the spurious community-level pathways described above. Llama-3.3-70B shows no significant effects despite its pronounced leftward prediction distribution (Figure 6), consistent with its lower F1 (54.61) and broader prediction variance.

The systematic divergences across LLM families, even when controlling for the same input text and outcome task, suggest that ideology annotation by LLMs is not a uniform operation but varies with model-specific feature reliance patterns.

5.4 Mediation Analysis as a Diagnostic Framework

A central challenge in deploying LLMs as annotators is that standard output-level metrics do not expose *why* a model produces a given label. Moraffah et al. (2020) argue that causal interpretability, understanding decisions in terms of which inputs *caused* which outputs, is essential for building trustworthy ML systems, precisely because correlation-based evaluation leaves spurious feature reliance invisible. Jin et al. (2024) show that LLMs perform near-randomly on tasks requiring pure causal inference from correlational evidence, and Joshi et al. (2024) demonstrate that LLMs are prone to post hoc and ordering fallacies when reasoning about causation. Taken together, these results imply that LLMs cannot be relied upon to self-report or self-correct the causal basis of their annotations; external causal analysis is required.

Our mediation framework provides one such approach. By holding the input text and causal structure fixed and varying only the annotation paradigm, we isolate differences that are at-

tributable to the annotator rather than the data. Applied to annotation auditing, mediation analysis can surface which textual features drive predictions through direct versus mediated pathways, offering a principled basis for trustworthiness assessment that F1 and agreement metrics cannot provide.

6 Future Work

Our framework treats each topic sentiment as a separate treatment or mediator. Modeling multiple topic sentiments as co-treatments would capture cross-topic interactions, and incorporating the article-level features discussed in our limitations (outlet, author, date, length, lexical complexity, framing) as additional treatments, mediators, or confounders would yield a richer multi-feature SCM that better isolates which textual dimensions drive the human-LLM divergences we observe.

A complementary direction is to *probe each annotator's data-generating process separately* rather than under a shared SCM: a human-specific diagram would represent deliberation, prior beliefs, and consensus dynamics, while an LLM-specific diagram would represent training-data and fine-tuning effects. Comparing the two paradigm-specific causal profiles would more directly characterize *where* and *why* human and LLM annotations diverge.

Our mediation-as-diagnostic framework could be applied beyond political ideology to other annotation tasks where LLM labels are used as silver-standard training data, such as sentiment analysis, stance detection, or toxicity classification. Comparing mediation pathways across human and LLM annotators in these domains would help determine whether the sentiment-sensitivity patterns we find are specific to ideology or reflect a more general tendency of LLMs.

Our current study is limited to two LLM families (GPT-4o-mini and Llama-3.3-70B). Expanding the investigation to a broader set of models, particularly those with reported ideological tendencies such as Grok (xAI, 2025), would allow a more systematic characterisation of how baseline ideological priors interact with fine-tuning to shape causal pathways. A model with a reported rightward lean, for instance, may exhibit sign reversals in different community configurations than a centrist baseline model, or may differ in its susceptibility to acquiring spurious sentiment-ideology couplings during fine-tuning. Comparing causal profiles across a wider model zoo would help disentangle the contributions of pre-training ideology, instruction tuning, and task-specific fine-tuning to the divergences we observe.

7 Limitations

Sentiment is operationally an LLM construct. Throughout the paper, what we call "sentiment" (F) is operationally Llama-3.3-70B's sentiment estimate, not a human-validated construct: we did not human-annotate any subset of the extractions. The treatment variable should therefore be read as "Llama-extracted sentiment" rather than "sentiment, full stop," and the entire analysis is conditional on sentiment that is legible to Llama-3.3-70B. Cases where a human reader would perceive sentiment but the extractor returns nothing (or misses the entity-topic association) are systematically absent from F, biasing estimates toward LLM-legible features and likely understating the role of more implicit or contextual cues that human annotators could still use.

Beyond presence/absence, both the *direction* and *magnitude* of an extracted sentiment score are annotator-dependent artifacts. Direction (positive vs. negative) is plausibly more stable across annotators, but magnitude can vary substantially between annotators (human or model) even when they agree on sign. Because our DML and mediation estimates are driven by variation in the magnitude of F, annotator-specific scale and calibration differences propagate directly into the estimated effects.

Unmodeled article-level features. Our causal framework controls for topic presence but does not encode outlet identity, author, publication date, article length, or lexical complexity. Outlets in particular may carry systematic sentiment patterns that confound the sentiment-ideology relationship; controlling for outlet identity would address this but would also substantially reduce the variation available for estimation given the correlation between outlet and ideology.

Statistical power. With our 1,265-article analysis subset distributed across topics and communities, some analyses operate on small effective sample sizes (as few as 21 articles per community after topic filtering). This limits statistical power, particularly for the community-level and mediation analyses. Null findings for human and Llama annotations should accordingly be interpreted with cau-

tion, as they may reflect insufficient power rather than true absence of effects.

Bounding the shared-SCM claim. Section 3 frames the same SCM as a deliberately simplified, sentiment-only model of the labeling process for both human and LLM annotators, with per-annotator factors absorbed into the per-annotator estimator. Reported effects should accordingly be read as effects propagating through the topic-level sentiment channel, conditional on the topics actually present in each article, rather than as claims about the underlying generative processes of human or LLM judgment.

References

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, Online. Association for Computational Linguistics.

Samuel E Bestvater and Burt L Monroe. 2023. Sentiment is not stance: Target-aware opinion classification for political text analysis. *Political Analysis*, 31(2):235–256.

Sumit Bhatia and P Deepak. 2018. Topic-specific sentiment analysis can help identify political ideology. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 79–84.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Robert M. Entman. 2010. Media framing biases and political power: Explaining slant in news of campaign 2008. *Journalism*, 11(4):389–408.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10.

Yuan Gao, Dokyun Lee, Gordon Burtch, and Sina Fazelpour. 2025. Take caution in using llms as human surrogates. *Proceedings of the National Academy of Sciences*, 122(24):e2501660122.

Matthew Gentzkow, Jesse M. Shapiro, and Matt Taddy. 2019. Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica*, 87(4):1307–1340.

Yujin Han, Lei Xu, Sirui Chen, Difan Zou, and Chaochao Lu. 2025. Beyond surface structure: A causal assessment of LLMs' comprehension ability. In *The Thirteenth International Conference on Learning Representations*.

John J Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.

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? In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Nitish Joshi, Abulhair Saparov, Yixin Wang, and He He. 2024. LLMs are prone to fallacies in causal inference. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10553–10569, Miami, Florida, USA. Association for Computational Linguistics.

Katherine Keith, Douglas Rice, and Brendan O'Connor. 2021. Text as causal mediators: Research design for causal estimates of differential treatment of social groups via language aspects. In *Proceedings of the First Workshop on Causal Inference and NLP*, pages 21–32, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peiyao Li, Noah Castelo, Zsolt Katona, and Miklos Sarvary. 2024. Frontiers: Determining the validity of large language models for automated perceptual analysis. *Marketing Science*, 43(2):254–266.

Bolei Ma, Berk Yoztyurk, Anna-Carolina Haensch, Xinpeng Wang, Markus Herklotz, Frauke Kreuter, Barbara Plank, and Matthias Aßenmacher. 2025. Algorithmic fidelity of large language models in generating synthetic German public opinions: A case study. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1785–1809, Vienna, Austria. Association for Computational Linguistics.

Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. 2020. Causal interpretability for machine learning - problems, methods and evaluation. *SIGKDD Explor. Newsl.*, 22(1):18–33.

William Nagy. 2007. Metalinguistic awareness and the vocabulary-comprehension connection. *Vocabulary acquisition: Implications for reading comprehension*, pages 52–77.

Judea Pearl. 2001. Direct and indirect effects. *Probabilistic and Causal Inference*.

Anastasia Smirnova, Helena Laranetto, and Nicholas Kolenda. 2017. Ideology through sentiment analysis: A changing perspective on russia and islam in nyt. *Discourse & Communication*, 11(3):296–313.

Gregory Spell, Brian Guay, Sunshine Hillygus, and Lawrence Carin. 2020. An Embedding Model for Estimating Legislative Preferences from the Frequency and Sentiment of Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 627–641, Online. Association for Computational Linguistics.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052, Singapore. Association for Computational Linguistics.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, and 1 others. 2024. Testing theory of mind in large language models and humans. *Nature human behaviour*, 8(7):1285–1295.

Graham Tierney and Alexander Volfovsky. 2021. Sensitivity analysis for causal mediation through text: an application to political polarization. In *Proceedings of the First Workshop on Causal Inference and NLP*, pages 61–73, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Victor Veitch, Dhanya Sridhar, and David Blei. 2020. Adapting text embeddings for causal inference. In *Conference on uncertainty in artificial intelligence*, pages 919–928. PMLR.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Patrick Y Wu, Jonathan Nagler, Joshua A Tucker, and Solomon Messing. 2023. Large language models can be used to estimate the latent positions of politicians.

xAI. 2025. Grok: A large language model by xAI.

Yu Yuan, Lili Zhao, Kai Zhang, Guangting Zheng, and Qi Liu. 2024. Do llms overcome shortcut learning? an evaluation of shortcut challenges in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12188–12200.

Yuhang Zhou, Paiheng Xu, Xiaoyu Liu, Bang An, Wei Ai, and Furong Huang. 2024. Explore spurious correlations at the concept level in language models for text classification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 478–492.

A Methodology and Reproducibility Details

A.1 LLM Evaluation Setup

Llama models were evaluated with temperature = 1.0 to maintain output diversity, while GPT models employed default temperature settings. For computational efficiency, we conducted preliminary screening on a 100-article subset to identify the most promising GPT variants before running the full-scale evaluation. Details about the GPT-4o-mini finetuning process and hyperparameters are provided in Section A.4.

Ideology Classification Prompt. All LLMs (baseline GPT-4o-mini, Llama variants) were prompted with the following system instruction for ideology classification:

> You are a political bias classifier specific to U.S. politics. Given the user's INPUT TEXT, return only valid JSON of the form `{"bias":"left|center|right"}`. No extra text.

The article text was provided as the user message. For fine-tuned GPT-4o-mini, the same prompt structure was used during both training and inference.

A.2 Topic Sentiment Extraction Details

Sentiment extraction uses two separate LLM calls per article:

1. **Entity and Sentiment Analysis:** The `Llama-3.3-70b-versatile` model extracts named entities and key concepts from the article text and assigns each a continuous sentiment polarity score between -1.0 (strongly negative) and $+1.0$ (strongly positive). This model was selected for its sensitivity to contextual nuance relative to smaller alternatives.

2. **Entity-to-Topic Association:** The `Llama3-70b-8192` model quantifies how strongly each extracted entity relates to predefined AllSides topic tags. The model handles ambiguity by distinguishing between a broad concept and specific incidents, for example, correctly separating terrorism as a general category from a specific event like the 2019 Christchurch shooting when assigning the entity "shooting" to a topic.

Entity sentiment scores are then aggregated by topic tag to produce a per-topic sentiment profile for each article.

3. **Top 5 Topic Selection:** The top 5 topics with the highest absolute sentiment scores are selected for each article. As the scores were aggregated at the previous step, this means that the final topic sentiment scores are no longer limited to having an absolute value ≤ 1.

To limit the risk of LLMs inferring bias from explicit sentiment signals, we used distinct sessions for the sentiment analysis and the ideology prediction steps.

A.3 Community Detection and Experimental Settings

The co-occurrence graph for Louvain community detection was constructed by filtering to edges with a minimum weight of five, meaning topic tag pairs that co-occurred in fewer than five articles were excluded. This threshold removes noise from outlier or one-off co-occurrences that may not reflect stable topical relationships. Self-loops allow single-tag articles to form their own communities. The algorithm yields a modularity score of 0.557 across 15 communities, with community sizes ranging from 1 to 90 tags.

Parameter	Setting
Estimator	LinearDML
Sentiment Treatment Encoding	Continuous
PCA Threshold	0.95
Minimum Topic Presence	50
Bootstrap Iterations	2000
Bootstrap Random Seed	42
Baseline GPT Model	`gpt-4o-mini-2024-07-18`

Table 2: Experimental settings for causal analysis.

A.4 GPT Finetuning

Hyperparameter	Values
Epochs	3, 10
Learning Rate Multiplier	1.8
Batch Size	4

Table 3: GPT-4o-mini fine-tuning hyperparameters.

A.4.1 Experimental Variables and Configurations

Table 3 summarizes the hyperparameters used for fine-tuning the GPT-4o-mini model. We experimented with two different epoch counts (3 and 10) to evaluate the impact of training duration on model performance. We used the OpenAI recommended settings for learning rate multiplier (1.8) and batch size (4). We evaluate fine-tuning performance using 150, 300, 1,000, and 2,000 labeled examples, representing different points on the data efficiency curve. These sample sizes span the range from few-shot learning (150) to moderate supervision (2,000), enabling analysis of diminishing returns in training data scaling.

A.4.2 Fine-tuning Protocol

Input formatting follows the standard conversational template required by OpenAI models, with political bias labels converted to structured JSON responses to enable systematic evaluation. Each training example consists of the article text as user input and the corresponding political orientation (left/center/right) as the assistant response.

Configuration	F-1 Macro
Zero-shot Baseline	51.86
3 epochs, 150 samples	33.37
3 epochs, 300 samples	25.04
3 epochs, 1,000 samples	69.23
3 epochs, 2,000 samples	**69.37**
10 epochs, 150 samples	28.03

Table 4: GPT-4o-mini fine-tuning performance across different training configurations.

We used the highest performing configuration (3 epochs, 2,000 samples) for the main causal analysis experiments. Table 4 provides a detailed breakdown of performance across all configurations, demonstrating the impact of training set size and input length on classification metrics.

B Community-level ATE breakdown

B.1 Ideology Prediction Breakdown for Significant Communities

This section provides a breakdown of the ideology prediction distributions for the significant communities identified in the ATE analysis. For each community, we show the proportion of Left, Center, and Right predictions across the four annota-

tion paradigms (Human, GPT base, GPT finetuned, Llama). We select the topics with the highest absolute sentiment scores, up to a maximum of five per article, and limit our analysis to these matched articles.

Figure 6 shows the overall prediction distribution across the full dataset. GPT base exhibits a pronounced centrist tendency, GPT finetuned and Llama both skew leftward, and human annotations are the most right-leaning overall, reflecting the ideological composition of the AllSides corpus.

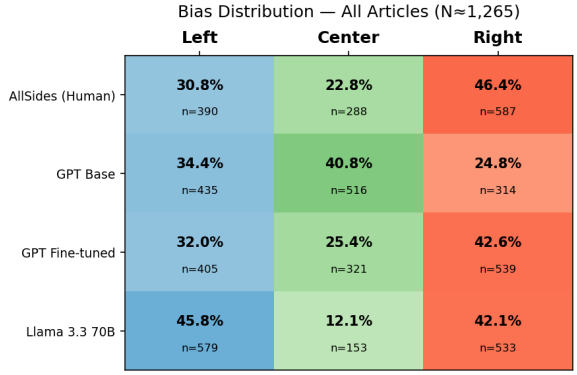

Figure 6: Overall ideology prediction distribution across all 1,265 articles, by annotation paradigm.

Figures 7 to 10 show the per-community distributions for the four communities with significant ATE results.

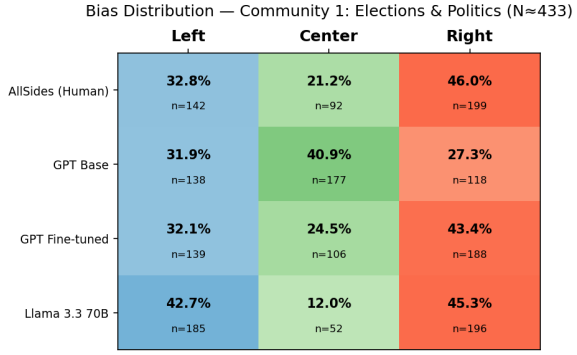

Figure 7: Ideology prediction distribution for Community 1 (Electoral Politics).

B.2 Complete Community-level ATE breakdown

Table 5 presents the average treatment effect of topic-community sentiment on ideology classification across all nine communities that met the minimum presence threshold. We compare estimates derived from human annotations against three LLM-based annotation pipelines: GPT-4o-mini (base), GPT-4o-mini (finetuned), and Llama-3.3-70b.

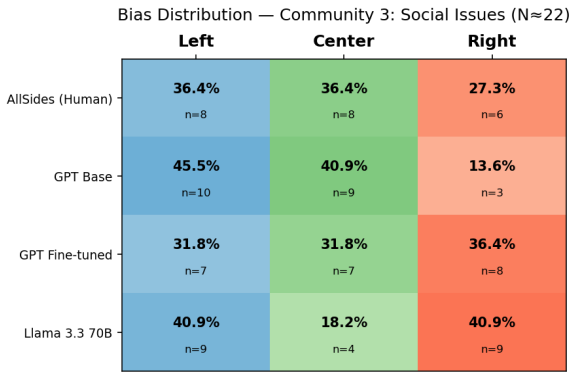

Figure 8: Ideology prediction distribution for Community 3 (Social Issues).

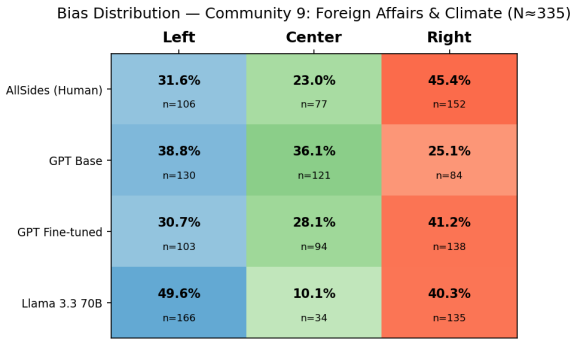

Figure 9: Ideology prediction distribution for Community 9 (Geopolitics and Foreign Affairs).

Most communities yield small, statistically non-significant ATEs across all annotation sources, with confidence intervals spanning zero. Given the modest analysis-subset size, wide confidence intervals are expected and the lack of significance should not be interpreted as sufficient evidence of null effects overall. With larger corpora, additional communities may well yield significant estimates. Nevertheless, the relative patterns across annotation sources remain informative: several communities reveal systematic divergences between human and LLM-derived causal estimates, raising concerns about the potential of using LLM annotations as drop-in replacements for human labels in causal inference pipelines.

The most striking divergence occurs in Community 3 (social issues, immigration, and the courts). Human annotations yield a positive estimate (0.011), while all three LLMs estimate negative effects. GPT (finetuned) produces the largest magnitude estimate (-0.023), the only statistically significant result in this community. This sign reversal suggests that LLMs encode a qualitatively

Community	Human	GPT (base)	GPT (finetuned)	Llama-3.3-70b	Num. Articles
Comm. 0	$\mathbf{-0.009}$ $[-0.024, 0.014]$	$\mathbf{-0.003}$ $[-0.015, 0.009]$	$\mathbf{0.008}$ $[-0.009, 0.021]$	$\mathbf{0.002}$ $[-0.017, 0.018]$	218
Comm. 1	$\mathbf{-0.014}$ $[-0.030, 0.003]$	$\mathbf{-0.015}^{*}$ $[-0.029, -0.001]$	$\mathbf{-0.013}$ $[-0.031, 0.003]$	$\mathbf{0.000}$ $[-0.015, 0.015]$	433
Comm. 2	$\mathbf{-0.001}$ $[-0.021, 0.016]$	$\mathbf{0.006}$ $[-0.008, 0.019]$	$\mathbf{0.003}$ $[-0.008, 0.025]$	$\mathbf{-0.007}$ $[-0.025, 0.007]$	216
Comm. 3	$\mathbf{0.011}$ $[-0.010, 0.024]$	$\mathbf{-0.004}$ $[-0.014, 0.009]$	$\mathbf{-0.023}^{*}$ $[-0.030, -0.005]$	$\mathbf{-0.007}$ $[-0.026, 0.004]$	22
Comm. 4	$\mathbf{-0.005}$ $[-0.025, 0.011]$	$\mathbf{0.008}$ $[-0.001, 0.028]$	$\mathbf{-0.004}$ $[-0.016, 0.021]$	$\mathbf{0.011}$ $[-0.011, 0.024]$	228
Comm. 5	$\mathbf{0.004}$ $[-0.016, 0.021]$	$\mathbf{-0.003}$ $[-0.013, 0.009]$	$\mathbf{-0.001}$ $[-0.018, 0.018]$	$\mathbf{-0.003}$ $[-0.017, 0.014]$	97
Comm. 7	$\mathbf{0.006}$ $[-0.013, 0.023]$	$\mathbf{0.004}$ $[-0.012, 0.018]$	$\mathbf{0.003}$ $[-0.020, 0.021]$	$\mathbf{-0.004}$ $[-0.018, 0.016]$	21
Comm. 9	$\mathbf{-0.014}$ $[-0.030, 0.003]$	$\mathbf{-0.006}$ $[-0.023, 0.008]$	$\mathbf{-0.013}^{*}$ $[-0.032, -0.002]$	$\mathbf{0.002}$ $[-0.015, 0.021]$	335
Comm. 11	$\mathbf{-0.017}$ $[-0.029, 0.006]$	$\mathbf{0.003}$ $[-0.012, 0.020]$	$\mathbf{0.018}^{*}$ $[0.000, 0.036]$	$\mathbf{0.007}$ $[-0.011, 0.024]$	691

Table 5: Community-level ATE by dataset (min presence=50). $^{*}p < 0.05$.

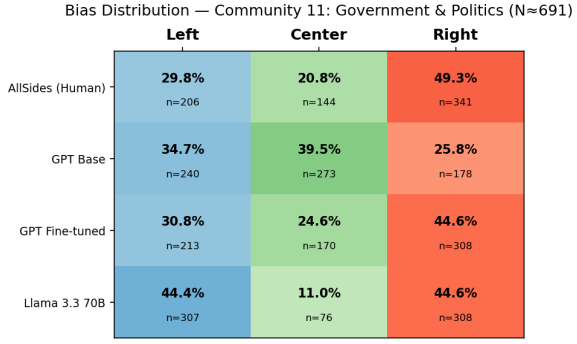

Figure 10: Ideology prediction distribution for Community 11 (Trump Presidency).

different mapping between sentiment toward socially contentious topics and perceived ideological leaning.

A similar pattern emerges in Community 11 (Trump, Congress, and domestic governance), the largest community by article count. Human annotators estimate a negative effect (-0.017), whereas all LLMs estimate positive effects, with GPT (finetuned) reaching borderline significance (0.018, 95% CI $[0.000, 0.036]$). That the direction of the estimated causal effect reverses entirely when substituting LLM for human annotations, on a community dominated by highly polarized political figures, underscores the sensitivity of downstream causal conclusions to annotation source.

In contrast, Community 1 (elections and campaigns) exhibits the strongest human-LLM alignment. Human (-0.014), GPT base (-0.015), and GPT finetuned (-0.013) produce nearly identical estimates, with GPT base achieving statistical significance (95% CI $[-0.029, -0.001]$). Notably, Llama-3.3-70b diverges here (0.000), detecting no effect whatsoever. A comparable pattern holds for Community 9 (foreign affairs and climate), where human and GPT finetuned estimates are closely aligned (-0.014 and -0.013, respectively), while Llama again attenuates toward zero.

These results point to two model-family-specific tendencies. First, GPT-based models, particularly after finetuning, tend to produce estimates that are either well-calibrated to human judgements or sharply divergent, with little middle ground. Second, Llama-3.3-70b consistently compresses effect sizes toward zero across communities, suggesting a general attenuation bias that may stem from more conservative or hedged sentiment predictions. Neither failure mode is uniformly preferable: sign reversals risk misleading causal conclusions, while systematic attenuation risks obscuring genuine effects. We note that the limited sample size constrains statistical power throughout this analysis; the directional disagreements reported here may represent conservative lower bounds on the true extent of human-LLM divergence, as larger datasets would narrow confidence intervals and potentially surface additional significant discrepancies.

C Topic-Level Results

Topic Tag	Articles	Community
Politics	594	11
Donald Trump	233	11
Media Bias	115	4
Race and Racism	104	5
Defense and Security	76	7

Table 6: Top five topic tags by frequency in the test set.

In this section, we present results from our topic-level analyses, which estimate the causal effect of sentiment toward individual topics on ideology. We focus on the five most frequent topic tags in the test set (Table 6), which provide the greatest statistical power for estimation while spanning multiple thematic communities. We first present Average Treatment Effect (ATE) estimates for each topic, then examine the Donald Trump$\rightarrow$ Politics mediation pathway in detail.

C.1 Topic-Level ATE

We simultaneously estimate the average treatment effect (ATE) of each topic's sentiment on ideology, controlling for all other topics as confounders via PCA-reduced topic-presence indicators. Let $\mathbf{F} = (F_{T_1}, \ldots, F_{T_K})$ be the full treatment matrix of topic sentiment scores and $\tilde{\mathbf{T}}$ the PCA-reduced topic-presence confounders. The marginal ATE for each topic T_k is

$$ATE_{T_k} = E\left[\frac{\partial Y}{\partial F_{T_k}}\right], \quad \forall k = 1, \ldots, K, \quad (6)$$

estimated jointly via a single `LinearDML` model with $\mathbf{F}$ as the treatment matrix and $\tilde{\mathbf{T}}$ as nuisance-only confounders (W), yielding one constant marginal ATE per topic. From the comparison seen in Figure 11, we see that human annotations yield no significant effects. The ATE of 0.022 for Politics is significant in both the finetuned and baseline GPT variants.

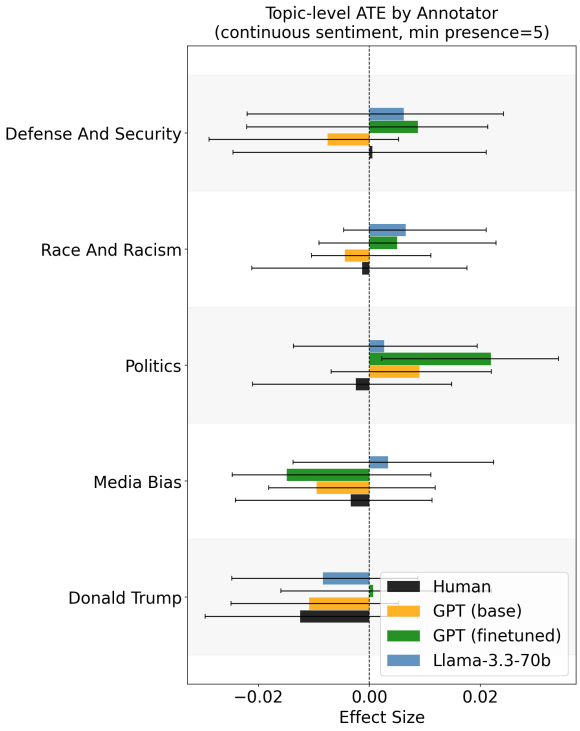

Figure 11: ATE for top five topics.

The topic-level dataset is substantially smaller than the community-level dataset, as individual topics appear in far fewer articles than aggregated communities. This limits statistical power and makes it difficult to draw reliable conclusions from topic-level estimates alone. With a larger corpus covering more diverse topics and time periods, topic-level analysis would become more feasible and could yield more stable, interpretable estimates.

C.2 Donald Trump/Politics Mediation Pathway

We examine the Donald Trump→Politics mediation pathway, estimating TE, NDE, and NIE using the same IQR contrast methodology as the community-level analysis (Section 3). Figure 12 shows that no statistically significant effects are found for any annotator paradigm, including for these two most frequently occurring topics in the dataset (Politics: N=594; Donald Trump: N=233).

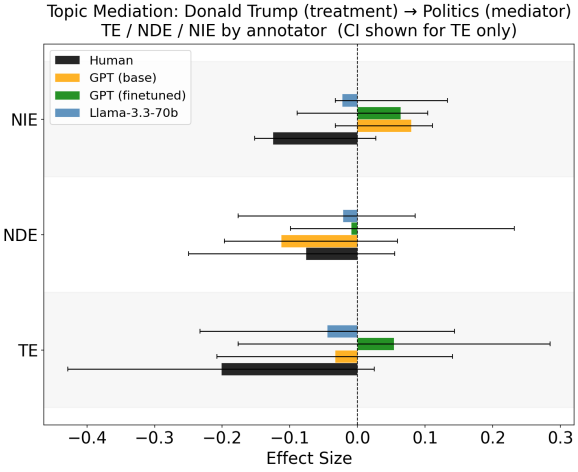

Figure 12: Donald Trump → Politics Mediation Effects

Given the small article counts, it is not possible to determine whether the absence of significant effects reflects a genuine lack of causal coupling or insufficient statistical power. A larger topic-level corpus would be required to distinguish these explanations.

D Community-level Mediation Results

Six treatment-mediator pairs were evaluated across communities $\{3, 9, 11\}$, but only $9 \to 11$ and $11 \to 9$ yielded non-degenerate mediation decompositions. The remaining four pairs degenerated for two distinct reasons. *Community 3 as treatment*: all confounder and mediator candidates were eliminated by the minimum topic co-occurrence filter (≥ 50 articles), reducing the model to an intercept-only specification with no mediation output. *Community 3 as mediator* (pairs $11 \to 3$ and $9 \to 3$): Community 3 failed the co-occurrence filter with both treatment communities; $9 \to 3$ collapsed to Community 11 as the sole surviving mediator (duplicating $9 \to 11$), while $11 \to 3$ substituted un-

Annotator	TE	95% CI	NDE	95% CI	NIE	% med.
Community $9 \to 11$ *(*$N = 335$*)*						
Human	-0.045	$[-0.164, +0.075]$	-0.059	$[-0.151, +0.045]$	$+0.015$	—
GPT (finetuned)	$\mathbf{-0.108}$	$[\mathbf{-0.208}, \mathbf{-0.008}]$	$\mathbf{-0.105}$	$[\mathbf{-0.170}, \mathbf{0.000}]$	-0.003	3.2%
GPT (base)	-0.059	$[-0.141, +0.023]$	-0.032	$[-0.102, +0.033]$	-0.026	—
Llama-3.3-70B	-0.007	$[-0.148, +0.134]$	$+0.023$	$[-0.080, +0.099]$	-0.030	—
Community $11 \to 9$ *(*$N = 691$*)*						
Human	-0.045	$[-0.162, +0.071]$	-0.049	$[-0.100, +0.012]$	$+0.004$	—
GPT (finetuned)	$+0.056$	$[-0.069, +0.181]$	$+0.069$	$[0.000, +0.116]$	-0.013	—
GPT (base)	$+0.007$	$[-0.097, +0.111]$	$+0.009$	$[-0.042, +0.059]$	-0.002	—
Llama-3.3-70B	$+0.008$	$[-0.114, +0.130]$	$+0.012$	$[-0.046, +0.075]$	-0.004	—

Table 7: Mediation results across annotators ($Q_{25} \to Q_{75}$ contrast). Significant results (95% CI excludes zero) are marked with *.

related mediators rather than testing the intended pathway. The six-pair matrix therefore reduces to the two unique, well-specified analyses below in Table 7.

D.1 Results

Table 7 reports total effects (TE), natural direct effects (NDE), and natural indirect effects (NIE) for the $Q_{25} \to Q_{75}$ contrast under each annotation paradigm.

For the $9 \to 11$ contrast, GPT (finetuned) is the only annotator to produce a significant total effect (TE $= -0.108^*$), with nearly all of it attributable to the direct pathway (3.2% mediated). Human and other LLM annotators show point estimates in the same direction but with wide, zero-crossing intervals. For the $11 \to 9$ contrast, no annotator reaches significance, and GPT (finetuned)'s effect reverses sign relative to the $9 \to 11$ direction.

D.2 Confounder Reduction

Confounders consist of all community sentiment scores except those of the treatment and mediator communities, reduced in two sequential steps: (1) communities whose sentiment is non-zero in fewer than 50 articles in the treatment subsample are dropped (`min_presence`$= 50$); (2) the remaining matrix is reduced via PCA retaining 95% of variance. Because this reduction depends only on the treatment subsample—not on ideology labels—it is identical across all four annotation paradigms for a given treatment community.

All four treatment communities begin with 11 potential confounders. For Communities 1, 3, and 11, the presence filter removes 3, leaving 8; PCA further reduces these to 7 components. Community 9 loses one additional community (4 removed, leaving 7), and PCA retains all 7 since they are collectively needed to reach the variance threshold. All configurations therefore yield 7 final confounder components; the choice of mediator community has no effect on this reduction.