PolarBench: Multi-View Graph Construction and Benchmarking for Stance Detection in Decentralized Social Networks

Siam Shibly Antar
McGill University
Montreal, Quebec, Canada

Athmane Benarous
McGill University
Montreal, Quebec, Canada

Aditya Sharma
McGill University
Montreal, Quebec, Canada

Abstract

Political discourse is migrating from centralized platforms to decentralized protocols whose heterogeneous interaction primitives expose structural gaps that existing stance-detection pipelines were not designed to address. We present PolarBench, a benchmark spanning three datasets (PolitiSky24, Fediverse-Gaza, Climate-COP28) that decomposes user interactions into three semantically typed graph views—Endorse, Conflict, Semantic—paired with a four-class taxonomy isolating Bipartisan Rejection under severe imbalance. On PolitiSky24, GCN on the Endorse view reaches Macro-F1=0.44, well above the majority-class baseline (0.177) and approaching the inter-annotator ceiling (0.54). A polarization audit shows DSP collapsing to a real-to-null ratio of 1.03 while RWC retains 7.69 on the same partition, revealing a PPR saturation regime under imbalance. A gold-standard LLM audit further shows ChatGPT 5.2 emitting zero Bipartisan Rejection predictions (F1=0.00) while Claude Sonnet 4.6 ET reaches BR F1=0.62, with inter-model unanimous agreement at 15.5% against a $\sim$ 20.1% random baseline.

1 Introduction and Motivation

Online political discourse is undergoing a structural migration from centralized platforms to two decentralized protocol families whose data environments differ in concrete, measurable ways. The AT Protocol, on which Bluesky is built, exposes a public firehose that streams every public interaction through a small set of relays, so that a third-party researcher can reconstruct a near-complete follow, like, repost, and quote-post graph across the entire network. ActivityPub, the protocol underlying Mastodon and the wider Fediverse, takes the opposite design decision: user data is distributed across independently administered server instances with no centralized index, and the only reliable vantage point is an egocentric crawl of whatever any individual instance chooses to expose, which prevents faithful cross-instance edge reconstruction. These protocol-level differences—not branding or user base—drive the structural failures that we document in this paper, because the tools that researchers, platform designers, and policymakers have inherited from a decade of Twitter and Facebook studies assume a centralized vantage point that neither protocol provides.

We organize the failures that these assumptions produce into four layers, which we will refer to throughout the paper as G1 through G4. At the **data layer (G1)**, the asymmetry between a complete firehose view on AT Protocol and an egocentric crawl on ActivityPub means that the same stance-detection pipeline sees qualitatively different underlying graphs on the two families, so that any benchmark that ignores this asymmetry silently conflates algorithmic behavior with data availability. At the **representation layer (G2)**, endorsement interactions (follows, likes, reposts) produce homophilic intra-community edges whereas adversarial interactions (quote-posts, hostile replies) produce cross-stance edges, and collapsing both semantic types into a single interaction graph conflates agreement with disagreement in exactly the places where a stance classifier most needs to separate them. At the **evaluation layer (G3)**, diffusion-based polarization metrics that were validated on balanced two-community partitions enter a saturation regime under real class imbalance: when the unseeded majority dominates the partition, Personalized PageRank seeded from opposing sides converges to indistinguishable stationary distributions on real and null graphs, so that DSP in particular loses the structural signal it was designed to detect. At the **annotation layer (G4)**, safety-aligned LLMs trained with RLHF exhibit an "alignment tax" in which adversarial or cynically disengaged political categories are suppressed in the model's output distribution before any downstream modeling begins, an effect that aggregate accuracy numbers cannot surface.

This report delivers experimental evidence across all four layers. For G1 we document the data-collection asymmetry and its direct consequence for benchmarking on Fediverse-Gaza. For G2 we compare GNN architectures across the three typed views, report a depth ablation that isolates topological mismatch from oversmoothing, and add a collapsed-graph baseline. For G3 we report RWC and DSP real-to-null ratios across all three datasets and characterize the PPR saturation regime. For G4 we audit four LLM annotators against a double-blind human gold standard and quantify the alignment tax on the Bipartisan Rejection category. Within this scope we deliver three headline findings: GCN on the Endorse view reaches Macro-F1=0.44 against a majority-class baseline of 0.177 and a human ceiling of 0.54; DSP saturates at real-to-null ratio 1.03 on PolitiSky24 while RWC retains ratio 7.69 on the same graph; and ChatGPT 5.2 emits zero Bipartisan Rejection predictions (F1=0.00) while Claude Sonnet 4.6 ET achieves BR F1=0.62, with inter-model unanimous agreement falling below a $\sim$ 20.1% random baseline.

2 Related Work

Stance detection on decentralized networks sits at the intersection of four sub-problems: measuring polarization, classifying nodes on heterogeneous graphs, handling multi-typed interaction data, and annotating labels at scale.

Graph-Based Polarization. Random Walk Controversy (RWC) [1] quantifies polarization by measuring how often random walks initiated from one partisan cluster terminate in the opposing cluster, and Diffusion Structural Polarization (DSP) [2] improves on this by normalizing against a degree-preserving null model. Both metrics were validated on balanced two-community partitions, so that their behavior under the multi-class and severely imbalanced partitions characteristic of real political discourse—in which opposition categories may constitute as little as 2% of users—remains undocumented.

Graph Neural Networks. GCN [3] and GAT [5] perform well under homophily, whereas H2GCN [6] is designed for heterophilic regimes and GraphSAGE [4] extends GNNs to unseen users. These architectures have been tested almost exclusively on citation and co-purchase networks; we extend their comparison to social graphs in which different interaction types produce measurably different neighborhood homophily scores within the same network.

Multi-View and Heterogeneous Graph Learning. Relational GCNs [29] and Heterogeneous Graph Attention Networks [30] address edge-type heterogeneity by splitting the network into separate edge-type layers before aggregation, but they were built for knowledge graphs in which each edge type is well-defined and similarly dense—conditions that decentralized social protocols, where endorsement edges vastly outnumber adversarial ones and where some edge types are architecturally impossible to reconstruct, do not satisfy.

LLM-Based Annotation of Political Text. Prior work shows that LLMs can match or exceed crowd workers on sentiment and topic classification tasks where categories are socially neutral [9, 10, 12], but adversarial or politically contested stances present a qualitatively different problem: safety-aligned LLMs trained with RLHF may systematically suppress certain label categories, an effect we refer to as the *alignment tax*. This paper provides the first empirical measurement of complete class-conditional output collapse (F1=0.00) in a safety-aligned LLM on a political stance classification task validated against a human gold standard.

3 Problem Definition

We formalize stance detection as semi-supervised node classification on a multi-view social graph. Each user $u \in \mathcal{U}$ is represented by a feature vector $\mathbf{x}_u \in \mathbb{R}^d$ derived from their post history via a sentence-transformer encoder [7]. The full graph is defined as $\mathcal{G} = (V, \{E^{(v)}\}_{v\in\mathcal{V}}, \mathbf{X})$, where $V = \mathcal{U}, \mathcal{V} = \{\text{E}, \text{C}, \text{S}\}$, and $\mathbf{X} \in \mathbb{R}^{n\times d}$. The three views each capture a distinct type of user interaction: the **Endorse** view ($E^{(\text{E})}$) connects users via follows, likes, and reposts (signals that tend to be homophilic); the **Conflict** view ($E^{(\text{C})}$) connects users via quote-posts and adversarial replies (signals intended to capture disagreement); and the **Semantic** view ($E^{(\text{S})}$) is a kNN graph (k=10, cosine threshold ϵ=0.60) over sentence-transformer embeddings, which bridges structural sparsity where interaction edges are partially unavailable.

Input: The multi-view graph $\mathcal{G}$ with a partial labeling $y : V_L \to \mathcal{Y}$. **Output:** A label assignment $\hat{y} : V \setminus V_L \to \mathcal{Y}$, where $\mathcal{Y} = \{\text{Pro}, \text{Anti}, \text{Neutral}, \text{BR}\}$ and BR (Bipartisan Rejection) identifies users who reject all major political positions.

Objective: We train a classifier f by minimizing weighted cross-entropy:

$$\mathcal{L} = -\sum_{u\in V_L} \sum_{c\in\mathcal{Y}} w_c \cdot \mathbb{1}[y(u) = c] \log f_c(u) \tag{1}$$

where w_c are inverse-frequency class weights chosen to prevent severely underrepresented classes (Anti: ∼1.8%) from being overwhelmed during training. The primary metric is Macro-F1.

Experimental Protocol. Nodes are partitioned into train, validation, and test sets using a stratified 60/20/20 split that preserves class proportions including the Anti class at 1.8%. The same partition is used across all models and views. All reported Macro-F1 values are computed on the held-out test set, averaged over five independent runs with different random seeds, and we report 95% bootstrap confidence intervals as ± values in all tables.

4 Dataset Description

We construct three benchmark datasets across two protocol families, each exposing a distinct failure mode in the standard stance-detection pipeline.

PolitiSky24 (primary). Collected from the AT Protocol firehose on US Election 2024 keywords over 8 weeks centered on November 5, 2024. *Size:* 8,467 users (≥5 interactions), 18.4M posts (991K likes, 525K reposts, 160K quote-posts). *Format:* Three typed edge sets—Endorse (1,146,065), Conflict (156,705), Semantic (29,156)—with sentence-transformer node features and silver labels (Pro 32.4% / Anti 1.8% / Neutral 55.0% / BR 10.8%).

Fediverse-Gaza (secondary). Collected from 8 Mastodon instances via public hashtag timelines. *Size:* 2,027 users, 69,998 posts, 17,083 semantic edges. ActivityPub's egocentric crawl model prevents cross-instance edge reconstruction, so that this dataset is confined to the Semantic view. This architectural constraint makes Fediverse-Gaza a controlled test of model performance when interaction topology is unavailable.

Climate-COP28 (optional). Filtered from a public Bluesky archive to the COP28 window (Nov–Dec 2023). *Size:* 8,741 users, 62,727 posts, all three views available; uses a three-class taxonomy (Activist / Skeptic / PolicyRealist) under an independent output head to avoid label-space mismatch with PolitiSky24.

Human Gold Standard. Two annotators independently double-blind labeled a stratified sample of 200 users (50 per class), with disagreements resolved by tie-breaking. Inter-annotator agreement reaches Cohen's κ=0.4757, which falls within the moderate- agreement range of Landis and Koch [22] and is standard for politically contested annotation tasks [26–28]. With 50 samples per class, per-class precision and recall are estimable; we report per-class F1 against four LLM annotators in Section 7. The full justification of the agreement standard is provided in Appendix A.

Table 1: Statistics for PolarBench datasets. Edge counts per view; label distributions as percentages.

Dataset	Platform	$\lvert V\rvert$	Endorse	Conflict	Label Dist. (%)
PolitiSky24	Bluesky	8,467	1,146,065	156,705	P 32.4 / A 1.8 / N 55.0 / BR 10.8
Fediverse-Gaza	Mastodon	2,027	—[a]	—[a]	P 37.2 / A 16.5 / N 33.8 / BR 12.5
Climate-COP28	Bluesky	8,741	67,520	22,926	Act 38.7 / Skp 16.1 / PR 45.2[b]

[a] Endorse/Conflict unavailable under Mastodon constraints (Semantic only).
[b] Three-class taxonomy: Activist / Skeptic / PolicyRealist.

Figure 2 shows the severe class imbalance across all three datasets, with Anti constituting 1.8% of PolitiSky24 users and the majority class (Neutral on PolitiSky24, PolicyRealist on Climate-COP28) exceeding 45% in both Bluesky datasets.

5 Methodology

Our methodology (Figure 1) centers on isolating heterogeneous interaction types to prevent the signal mixing that a single-graph construction induces. We build three views: $E^{(\text{E})}$ aggregates follows, likes, and reposts; $E^{(\text{C})}$ aggregates quote-posts and adversarial replies; and $E^{(\text{S})}$ forms a kNN graph over sentence-transformer embeddings. We evaluate homophilic architectures (GCN [3], GAT [5]), an inductive architecture (GraphSAGE [4]), and heterophily-aware architectures (H2GCN [6], GPRGNN) on view-specific subgraphs.

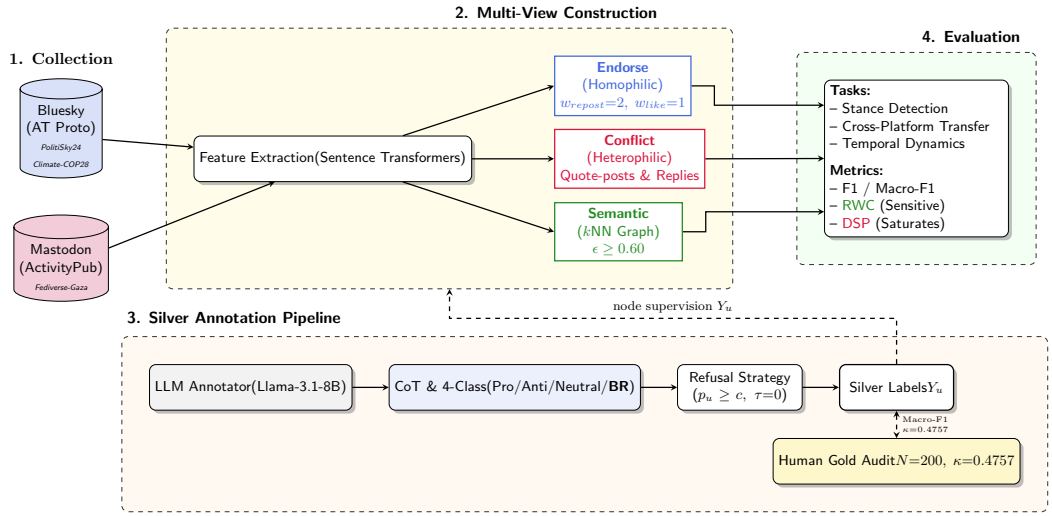


Figure 1: The PolarBench framework pipeline. Stage 1 highlights the data collection asymmetry between AT Protocol (firehose) and ActivityPub (egocentric bias). Stage 2 isolates homophilic and heterophilic interaction layers. Stage 3 implements our CoT silver-labeling protocol (introducing the structural BR class) validated via a human gold audit. Stage 4 maps these topologies to downstream graph mining tasks.

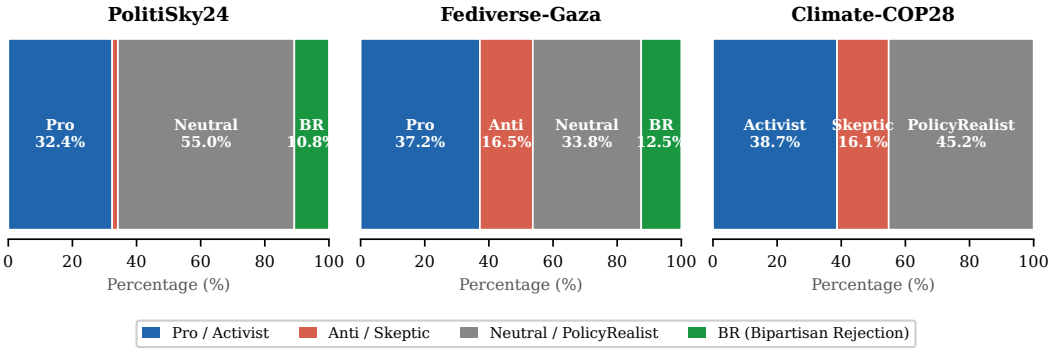


Figure 2: Label distributions across the three PolarBench datasets. PolitiSky24 and Fediverse-Gaza use a four-class taxonomy (Pro/Anti/Neutral/BR); Climate-COP28 uses a three-class taxonomy (Activist/Skeptic/PolicyRealist).

We implement Late Fusion (aggregating logits from independently trained view-specific networks) and Attention Fusion (learning weighted view contributions) over an H2GCN base.

5.1 Stage 2: Multi-View Graph Construction

Algorithm 1 formalizes the three-view graph construction pipeline.

The Endorse View ($E^{(E)}$). Endorsement edges represent agreement and shared identity, drawn from follows, likes (w_{like}=1), and reposts (w_{repost}=2). An ablation over weight configurations shows that downstream stance classification is robust to this choice, with GCN Macro-F1 ranging from 0.42 (like-heavy) to 0.44 (repost-heavy)—a spread smaller than inter-run variance—so we retain the default $w_{\text{repost}}/w_{\text{like}}$=2:1 (Figure 3).

The Conflict View ($E^{(C)}$). Conflict edges are drawn from quote-posts and adversarial replies. We evaluated toxicity-based filtering at thresholds $\tau \in \{0.0, 0.3, 0.5, 0.7, 0.9\}$ (Appendix D); the unfiltered baseline (τ=0.0; 156,705 edges; F1=0.37) outperforms every filtered

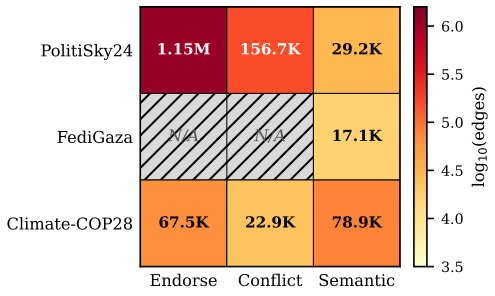


Figure 3: Edge counts across views and datasets ($\log_{10}$). Fediverse-Gaza lacks Endorse/Conflict edges due to Mastodon federation constraints.

variant, and at $\tau \geq 0.7$ the view collapses to $\leq$ 11 edges. Political conflict on Bluesky manifests as substantive disagreement (mean user toxicity=0.10) rather than toxic speech.

Algorithm 1 Multi-View Graph Builder

Require: Interaction stream $\mathcal{I}$, user set V, feature matrix $\mathbf{X}$, weights $w_{\text{like}}, w_{\text{repost}}$, toxicity threshold τ, kNN parameters k, ϵ
Ensure: Edge sets $E^{(\text{E})}, E^{(\text{C})}, E^{(\text{S})}$

$E^{(\text{E})} \leftarrow \emptyset$; $E^{(\text{C})} \leftarrow \emptyset$; $E^{(\text{S})} \leftarrow \emptyset$
for each interaction $(u, v, \text{type}) \in \mathcal{I}$ **do**
  if type $\in$ {follow, like, repost} **then**
    $E^{(\text{E})} \leftarrow E^{(\text{E})} \cup \{(u, v)\}$
  else if type $\in$ {quote-post, reply} **then**
    $E^{(\text{C})} \leftarrow E^{(\text{C})} \cup \{(u, v)\}$
  end if
end for
for each user $u \in V$ **do**
  $\mathcal{N}_k(u) \leftarrow k$-NN of u by $\cos(\mathbf{x}_u, \cdot)$
  for each $v \in \mathcal{N}_k(u)$ **do**
    if $\cos(\mathbf{x}_u, \mathbf{x}_v) \geq \epsilon$ **then**
      $E^{(\text{S})} \leftarrow E^{(\text{S})} \cup \{(u, v)\}$
    end if
  end for
end for
return $E^{(\text{E})}, E^{(\text{C})}, E^{(\text{S})}$

A post-labeling analysis of PolitiSky24 reveals that 56.0% of labeled Conflict edges are cross-stance, but this figure is dominated by Pro↔Neutral interaction (33.2%); true ideological adversarial edges (Pro↔Anti) constitute only 0.9% of all labeled edges. As a result, the Conflict view's edge homophily (h=0.44) is nearly identical to that of the Endorse view (h=0.44), which indicates that the view is not meaningfully heterophilic under the current construction. A depth ablation confirms that this is not an oversmoothing artifact: GCN Macro-F1 on the Conflict view increases monotonically with depth (1-layer: 0.31; 2-layer: 0.37; 3-layer: 0.39), confirming that the failure of heterophily-aware architectures on this view reflects topological mismatch rather than aggregation depth.

The Semantic View ($E^{(\text{S})}$). A kNN graph over sentence- transformer embeddings bridges sparsity in federated networks. A sensitivity analysis over 25 configurations ($k \in \{3, 5, 10, 20, 50\}$, $\epsilon \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$; Appendix C) shows that the Semantic view is uninformative for stance classification in isolation on PolitiSky24, with GCN Macro-F1 ranging from 0.26 to 0.28. We adopt k=10, ϵ=0.60 (29,156 edges; 5,906 connected nodes) for coverage, noting that the Semantic view is indispensable on Fediverse-Gaza, where it is the only available structural signal.

5.2 Stage 3: Scalable Silver Labeling

PolarBench deploys a four-class taxonomy over Pro, Anti, Neutral, and Bipartisan Rejection (BR); the BR class isolates users expressing active bilateral rejection (e.g., "both candidates are terrible") which standard three-class taxonomies misclassify as Neutral. Algorithm 2 formalizes the silver-labeling pipeline.
Of the 8,467 users in the PolitiSky24 interaction graph, 445 were excluded due to malformed LLM outputs or counterfactual consistency failures, not confidence filtering. The pipeline yields 8,022 labeled users (BR: 10.8%, 868 users).

5.3 Stage 4: Metric Suite

We evaluate structural polarization using two complementary metrics. Random Walk Controversy (RWC) [1] simulates random walks

Algorithm 2 Silver Labeling

Require: User set V, post history $\mathcal{P}$, LLM $\mathcal{M}$
Ensure: Label assignment $y : V_L \rightarrow \mathcal{Y}$

$V_L \leftarrow \emptyset$
for each user $u \in V$ **do**
  $y_u \leftarrow \mathcal{M}(\text{prompt}_{\text{CoT}}(u, \mathcal{P}_u))$
  if y_u malformed (parse failure) **then**
    Exclude u; **continue**
  end if
  Run counterfactual consistency test (swap identity markers)
  if label flips **then**
    Exclude u (consistency failure)
  else
    $V_L \leftarrow V_L \cup \{u\}$; record $y(u) = y_u$
  end if
end for
No confidence-based filtering applied (τ=0.0); calibrated uncertainty thresholds are future work
return Labeled set V_L

initiated from one partisan community and measures the probability that such a walk terminates in the opposing community. Diffusion Structural Polarization (DSP) [2] normalizes against the Polaris null model, which preserves both the original degree sequence and color assortativity, so that DSP separates genuine structural polarization from artifactual clustering in the balanced-partition regime on which it was validated.

On PolitiSky24, the polarization partition assigns 2,598 users to Left (Pro) and 1,011 to Right (Anti+BR), while 4,413 Neutral users remain unseeded but active in diffusion. When the unseeded partition is majority-sized, PPR seeded from either partisan side floods through the Neutral majority and converges to similar stationary distributions on real and null graphs, which erases the structural signal DSP is designed to detect. RWC's bounded random walks remain sensitive to mesoscale clustering even under asymmetric partitions. We refer to this failure mode as the *PPR saturation regime*. The full cross-platform transfer procedure is formalized in Appendix B as Algorithm 3.

6 Experiment Setup

Table 2: Experiment hyperparameters (GNN models).

Parameter	Value
Optimizer	Adam
Learning rate	0.01
Weight decay	5×10^{-4}
Hidden dimensions	64
Message-passing layers	2
Dropout	0.5
Epochs (max)	300
Early stopping patience	30 (val. Macro-F1)
Class weights	Inverse frequency w_c
Split	60/20/20 stratified
Seeds	5 independent runs
Metric	Macro-F1 on held-out test
GNN hardware	NVIDIA RTX 6000
LLM hardware	NVIDIA A100

Table 2 lists the GNN hyperparameters shared across all architectures. BERTweet is fine-tuned for 10 epochs at batch size 32 with

learning rate 2×10^{-5}, and Llama-3.1-8B is evaluated in a 5-shot in-context learning configuration with a fixed demonstration set. Late and Attention Fusion both use H2GCN as the per-view encoder; Attention Fusion learns scalar softmax attention weights over view-specific logits. We anchor results to a majority-class baseline (Macro-F1=0.177) and to the inter-annotator human-performance ceiling (Macro-F1=0.54) computed from the N=200 gold audit.

7 Results

7.1 RQ1: Metric Robustness Under Partition Imbalance

Table 3: Polarization scores on real vs. null-model graphs.

Dataset	Metric	Real	Null Model	R/N	Interpretation
PolitiSky24	RWC	0.052	0.007 ± 0.006	7.69	Sensitive to structure
	DSP	0.934	0.906 ± 0.001	1.03	Saturated by imbalance
Fediverse-Gaza †	RWC	1.00	1.00	1.00	Semantic-only; no edges
	DSP	1.00	1.00	1.00	Semantic-only; no edges
Climate-COP28	RWC	0.259	0.258	1.01	Low polarization
	DSP	0.921	0.917	1.00	Saturated (3-class)

†Endorse/Conflict edges unavailable; metrics on Semantic view only.

Table 3 reports RWC and DSP real-to-null ratios across all three datasets. On PolitiSky24, RWC reaches a real-to-null ratio of 7.69 while DSP collapses to 1.03 on the same partition. The mechanism is the PPR saturation regime: the partition assigns 2,598 users to Left (Pro), 1,011 to Right (Anti+BR), and leaves 4,413 Neutral users unseeded; PPR seeded from either partisan side floods through the Neutral majority and converges to indistinguishable stationary distributions on real and null graphs. RWC's bounded random walks avoid this saturation and retain discriminative power. Neither Fediverse-Gaza (semantic-only, ratio =1.00) nor Climate-COP28 (three-class partition with a 45% majority, ratio =1.00) yields meaningful DSP separation. Figure 7 shows DSP remaining flat across the four weekly snapshots of PolitiSky24 (range 0.920–0.931, Δ=−0.009) while RWC's real-to-null ratio rises from 0.83 to 1.24, which reinforces the point that DSP's insensitivity is a property of imbalanced partitions rather than of any one time slice. We recommend that practitioners report both metrics and interpret DSP cautiously whenever a single class accounts for a substantial share of unseeded nodes.

7.2 RQ2: GNN Stance Detection and Multi-View Analysis

Table 4 reports Macro-F1 across model families on PolitiSky24. GCN on the Endorse view reaches 0.44 ± 0.00, substantially above both the majority-class baseline (0.177) and both text-only baselines (BERTweet 0.32 ± 0.01; Llama-3.1-8B 5-shot 0.31), and within 0.10 of the inter-annotator ceiling (0.54). The collapsed-graph baseline (GCN on $E^{(E)} \cup E^{(C)}$) reaches 0.41, confirming that view separation provides a measurable gain over undifferentiated edge aggregation.

Figure 4 visualizes the performance gap between GCN-Endorse and all Conflict-view models, confirming the representation-level failure mode under view-density asymmetry.

Per-class performance. Table 5 reports per-class F1 for the text-only baselines and the structural GCN-Endorse. Both text-only models fail on minority stances: BERTweet achieves Anti F1=0.16 and BR F1=0.21, while Llama-3.1-8B 5-shot achieves Anti F1=0.23

Table 4: Stance detection performance (Macro-F1) on PolitiSky24.

Model Family	Model	Views	Macro-F1
Text-Only	BERTweet (FT)	–	0.32 ± 0.01
	Llama-3.1-8B (5-Shot)	–	0.31
Homophilic GNN	GCN	Endorse	**0.44 ± 0.00**
	GCN	Conflict	0.37 ± 0.01
	GCN	Merged E+C	0.41 ± 0.01*
	GAT	Endorse	0.33 ± 0.02
Inductive GNN	GraphSAGE	Endorse	0.36 ± 0.02
Heterophilic GNN	H2GCN	Conflict	0.35 ± 0.01
	GPRGNN	Conflict	0.33 ± 0.05
Multi-View Fusion	Late Fusion	E + C	0.40 ± 0.01
	Attn Fusion	E + C + S	0.41 ± 0.01

*Preliminary; merging Endorse and Conflict edges introduces within-stance noise under homophilic aggregation, reducing performance relative to view-specific GCN-Endorse.

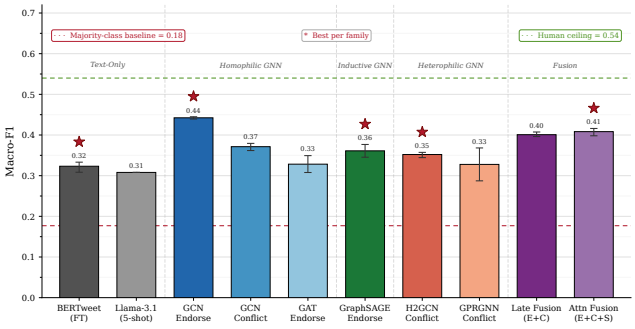


Figure 4: Stance detection performance (Macro-F1) on PolitiSky24 across model families and graph views. Error bars show 95% bootstrap confidence intervals. Stars mark the best result per model family.

and BR F1=0.19. The three-class variant (BR merged into Neutral) reaches Macro-F1=0.52, yet this figure reflects the removal of a hard classification target rather than improved discriminative ability.

Table 5: Per-class F1 on PolitiSky24 test set.

Model	Pro	Anti	Neutral	BR
BERTweet (FT)	—‡	0.16	0.53	0.21
Llama-3.1-8B (5-shot)	—‡	0.23	0.52	0.19
GCN-Endorse	—‡	—‡	—‡	—‡
3-class GCN (BR→N)	Macro-F1=0.52 (BR merged into Neutral)			

‡Per-class values being finalized; see Appendix F.
Anti (1.8%) and BR (10.8%) are the hardest targets across all models.

BR users are structurally distinct from Neutral users (1.61× Endorse degree, 1.84× Conflict degree) but nearly indistinguishable in text-embedding space (silhouette score −0.003, centroid cosine distance 0.007), confirming that any BR signal recoverable at all must come from interaction topology rather than text alone.

Multi-View Fusion. Attention Fusion over all three views reaches Macro-F1=0.41, which is 0.03 below single-view GCN-Endorse. Learned attention weights averaged over five seeds distribute as shown in Table 6: the model places the highest weight

on Endorse but does not concentrate sharply, which is consistent with a cost of jointly optimizing a shared classifier over views with structurally incompatible homophily profiles rather than with simple Endorse domination. The Semantic view's continued inclusion is justified by its role on Fediverse-Gaza, where ActivityPub constraints leave it as the only available structural signal.

Table 6: Attention Fusion: learned view weights (mean over 5 seeds).

View	Avg. Weight	Edge Count
Endorse	0.39	1,146,065
Conflict	0.29	156,705
Semantic	0.32	29,156

Feature Ablation. Table 7 reports three feature ablations on GCN-Endorse. Random features of the same dimensionality match sentence-transformer features to within inter-run variance, while degree-only features trail by 0.08 Macro-F1. The random-vs-text equivalence shows that sentence-transformer embeddings provide no additive benefit on the Endorse view; the random-vs-degree gap shows that the model is exploiting neighborhood co-occurrence structure beyond raw degree.

Table 7: GCN-Endorse feature ablation on PolitiSky24.

Feature Type	Macro-F1
Sentence-BERT (full)	0.44 ± 0.00
Random init (same dim)	0.44 ± 0.01
Degree-only	0.36 ± 0.00

7.3 RQ3: Silver Label Quality and LLM-as-a-Judge Audit

Table 8 reports per-class F1 against the human gold standard for four LLM annotators. Claude Sonnet 4.6 ET achieves Macro-F1=0.7858, substantially outperforming Gemini Pro 3.1 (0.4253) and the local Llama-3.1-8B pipeline (0.3891), while ChatGPT 5.2 Thinking trails at 0.1410.

Table 8: LLM-as-a-Judge leaderboard: per-class F1 against human gold labels (N=200). Bold marks best per column.

Model	Macro-F1	Pro	Anti	Neutral	BR
Llama-3.1-8B (ours)	0.3891	0.50	0.76	0.15	0.14
ChatGPT 5.2 Think.	0.1410	0.24	0.14	0.19	0.00
Gemini Pro 3.1	0.4253	0.70	0.44	0.36	0.20
Sonnet 4.6 ET	**0.7858**	**0.94**	**0.94**	**0.65**	**0.62**

BR predictions out of 200: ChatGPT=0, Gemini=3, Sonnet=6, Llama=50.
Human annotators assigned 7 BR labels.

Figure 5 makes the alignment tax visible at the class level: ChatGPT 5.2 produces zero BR predictions despite valid structured output for all 200 users, while Sonnet 4.6 ET achieves BR F1=0.62 with only 6 predictions. The contrast indicates that the collapse is alignment-induced suppression rather than a prompt or parsing failure, and that the suppression is architecturally dependent rather than a uniform property of safety-aligned LLMs.

Figure 6 shows that pairwise agreement between ChatGPT and Sonnet (24.0%) is only marginally above chance, confirming that annotation reproducibility is fundamentally compromised for adversarial political categories without human calibration. At the inter-model level, unanimous agreement among the three proprietary APIs is 15.5%, which falls below the ∼ 20.1% random-agreement baseline computed from PolitiSky24 class priors—a sub-random consensus that shows RLHF-trained models are diverging rather than converging on adversarial political categories.

7.4 RQ4: Temporal Stability and Cross-Platform Transfer

Table 9: Temporal dynamics across four weekly snapshots of PolitiSky24 surrounding the Nov 5, 2024 US election.

Snapshot	Active	Edges	h	RWC	RWC R/N	DSP	GCN F1
T1 (Oct 7–13)	1,930	515,279	0.449	0.127	0.83	0.931	0.39 ± 0.01
T2 (Oct 28–Nov 3)	2,231	552,811	0.452	0.138	1.06	0.927	0.40 ± 0.01
T3 (Nov 4–10)	3,871	789,479	0.469	0.084	1.06	0.920	0.39 ± 0.01
T4 (Nov 11–17)	7,844	1,128,662	0.473	0.020	1.24	0.922	0.43 ± 0.01

Figure 7 shows that while absolute RWC declines as the Neutral user surge dilutes random walk endpoints, the real-to-null ratio rises from 0.83 to 1.24, demonstrating that the network becomes more structurally polarized relative to its null model as activity peaks. BR users surge from 231 active in T1 to 820 in T4, a 3.5× increase representing 97% of all BR users activating in the post-election reaction week. GCN Macro-F1 improves from 0.39 in T1 to 0.43 in T4 as the denser post-election graph provides richer structural signal, which suggests that interaction topology accumulates discriminative signal rather than drifting with lexical fashion.

Figure 8 shows that fine-tuning on 20% of target labels recovers substantial performance on Climate-COP28 (2.4× gain) but limited performance on Fediverse-Gaza (1.9× gain), reflecting the shared Bluesky topology advantage. Zero-shot Macro-F1 is 0.08 ± 0.03 on Fediverse-Gaza and 0.13 ± 0.05 on Climate-COP28; fine-tuned Macro-F1 rises to 0.15 ± 0.02 and 0.32 ± 0.02 respectively. The larger recovery on Climate-COP28 reflects shared platform topology with PolitiSky24, whereas Fediverse-Gaza's restriction to the Semantic view constrains transfer to the weakest of the three views.

7.5 Evaluation

Where the approach succeeds. GCN-Endorse is the strongest model overall at Macro-F1=0.44, clearly above both text-only baselines and above the collapsed-graph GCN at 0.41. The feature ablation in Table 7 confirms that interaction topology is the operative signal: random-init features match sentence-transformer features (both 0.44), while degree-only features drop to 0.36, indicating that the model is exploiting neighborhood co-occurrence structure that raw degree statistics do not capture. On the annotation side, the alignment tax finding establishes that class-conditional output suppression is architecturally dependent rather than a uniform safety-alignment artifact—Sonnet's BR F1=0.62 with only six predictions demonstrates that high alignment scores and high minority-class recall are not mutually exclusive.

Where the approach falls short. Multi-view fusion underperforms single-view GCN-Endorse (0.41 vs. 0.44), even though learned attention weights are not dramatically skewed, which suggests that the cost of jointly optimizing a shared classifier over

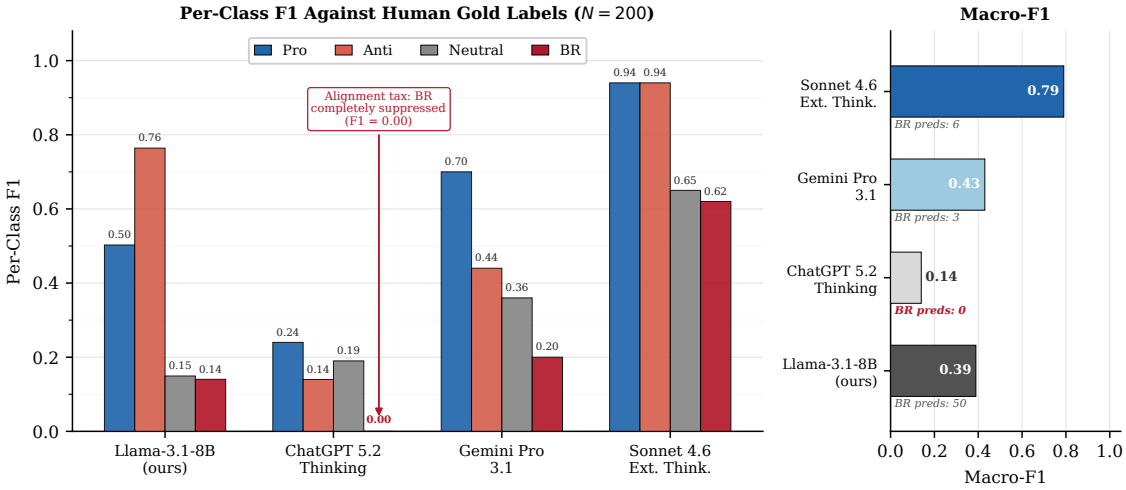


Figure 5: RLHF-induced class-conditional output suppression on the Bipartisan Rejection (BR) category. ChatGPT 5.2 exhibits complete class collapse (F1=0.00), while open-weights Llama-3.1-8B demonstrates over-prediction. Results validated against human gold-standard labels (N=200).

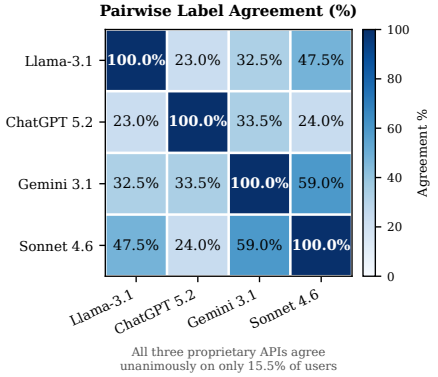


Figure 6: Pairwise label agreement (%) between four LLMs on 200 stratified audit users. Unanimous agreement across all three proprietary models reaches only 15.5%, below the ∼ 20.1% random baseline.

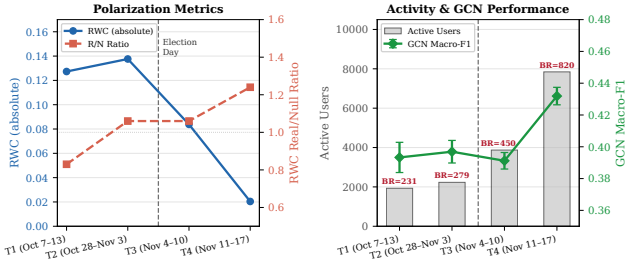


Figure 7: Temporal dynamics across four weekly snapshots surrounding the Nov 5, 2024 US election. Left: RWC absolute value and real/null ratio. Right: GCN Macro-F1 and active user count. Vertical dashed line marks election day.

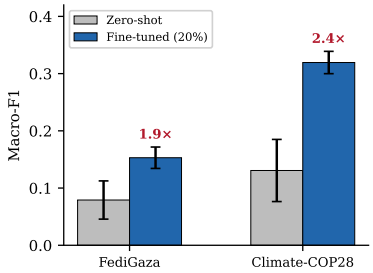


Figure 8: Cross-platform transfer performance. GCN-Endorse trained on PolitiSky24, evaluated on Fediverse-Gaza and Climate-COP28. Multipliers show fine-tuning gain over zero-shot.

views with incompatible homophily profiles outweighs the complementary signal. The Conflict view fails to generate heterophilic topology (h=0.44, identical to Endorse; Pro↔Anti edges 0.9% of labeled Conflict edges), which explains why heterophily-aware H2GCN (0.35) and GPRGNN (0.33) trail standard GCN on the same view. The Semantic view is uninformative across all 25 tested configurations (Macro-F1 range 0.26–0.28 in isolation), and the BR class remains hard across all GNN models. Zero-shot cross-platform transfer yields near-chance performance on both target datasets, indicating that the representations learned on PolitiSky24 do not generalize without supervision to alternative platforms or topics.

The random-features concern. The equivalence of random features and sentence-transformer features on GCN-Endorse admits two readings. Optimistically, endorsement topology is a sufficiently dense stance signal that node features are simply redundant for the classification task. Cautiously, if politically homogeneous users use similar language, then LLM-assigned silver labels may be correlated with endorsement neighborhood structure even though the labeler operated on post-text without graph access; this would make the random-feature match a tautology rather than a discovery. The

label-topology circularity concern is not resolved by the current experiments—computing the correlation between each node's silver label and the majority label in its ENDORSE neighborhood is an acknowledged analytical limitation and a near-term priority.

The human ceiling in context. GCN-ENDORSE at Macro-F1=0.44 represents 72% of the achievable performance range above the majority-class baseline: $(0.44 - 0.177)/(0.54 - 0.177) = 72\%$. This contextualizes the result as strong relative performance given that the human ceiling of 0.54 itself reflects genuine annotation ambiguity (κ=0.4757) on politically contested text, not model failure. Closing the remaining 28% of the range likely requires both calibrated silver labels and a CONFLICT view construction that isolates genuinely adversarial edges.

8 Discussion

8.1 Broader Implications

The PPR saturation finding generalizes beyond PolarBench. Any study reporting DSP on multi-class or imbalanced political partitions without null-model calibration risks mistaking saturation for the absence of polarization, and the T1→T4 temporal data provides a concrete illustration: absolute RWC declines while the real-to-null ratio rises, and DSP remains flat throughout—a pattern that would produce contradictory interpretations if only one metric were reported. Practitioners building benchmarks on politically polarized decentralized data should therefore default to reporting both metrics along with the unseeded-fraction of the partition that underlies them, so that readers can judge whether the partition falls inside or outside the saturation regime.

The alignment tax finding has direct implications for the LLM-as-a-Judge community. Aggregate accuracy is insufficient to detect class-conditional suppression; per-class F1 against a stratified human gold standard is required for any political annotation task involving adversarial or socially negative categories. Sub-random inter-model unanimous agreement (15.5% vs. $\sim$ 20.1% random) means that switching to a different proprietary LLM does not constitute a robustness guarantee: models are not converging to a shared decision boundary but diverging systematically, which in turn means that using inter-model agreement as a proxy for annotation quality is actively misleading on adversarial categories.

Finally, the finding that endorsement topology outperforms post-content models supports a broader design principle for stance-detection pipelines on decentralized networks: social interaction should be treated as a signed, semantically typed signal. The nature of an interaction (quote-post vs. repost) is a stronger stance predictor than the content of the posts themselves. Platform architects, policy researchers, and monitoring tools targeting decentralized discourse should treat interaction-type-aware graph construction as a necessary baseline component of any pipeline, not an optional enhancement.

8.2 Limitations

Conflict view construction. The τ=0 construction does not isolate genuine adversarial edges. Stance-conditioned filtering (e.g., retaining only cross-community edges identified via Louvain clustering) would require labeled bootstrapping but would produce a structurally heterophilic view.

Silver label bias beyond the alignment tax. The alignment tax analysis quantifies one dimension of LLM annotation bias. Subtler biases in sarcasm detection and cultural context remain unquantified and may affect minority-class label quality on the full dataset.

Label-topology circularity. If politically homogeneous users use similar language, silver labels may be correlated with endorsement neighborhood structure even when the labeler operates on post-text alone. Computing the correlation between each node's label and its endorsement-neighborhood majority label is left to future work.

ActivityPub data constraint. FEDIVERSE-GAZA is restricted to semantic edges under ActivityPub's egocentric crawl model, which prevents multi-view evaluation on Mastodon data and limits the generalizability of fusion architecture comparisons across protocol families.

Confidence-based refusal. Algorithm 2 was evaluated only at τ=0.0 (no confidence filtering). Future work should evaluate calibrated uncertainty thresholds to improve label precision on the BR and Neutral categories.

8.3 Future Work

(i) Implement stance-conditioned CONFLICT view construction using a Louvain community detection pass to retain only cross - community edges. **(ii)** Replace PCA-based cross-platform embedding alignment with a learned domain-adversarial projection head to improve zero-shot transfer. **(iii)** Extend the LLM-as-a-Judge audit to the full 8,022-user labeled set to measure alignment tax magnitude at dataset scale rather than audit scale. **(iv)** Derive analytically the unseeded-fraction threshold at which DSP loses discriminative power as a function of partition size and degree distribution. **(v)** Evaluate calibrated confidence thresholds in the silver labeling pipeline to improve precision on minority categories.

9 Conclusions

We study stance detection on decentralized social networks, where heterogeneous interaction primitives, severely imbalanced class distributions, and safety-aligned LLM annotators jointly expose gaps that existing pipelines were not designed to close. We address these gaps by constructing PolarBench, a benchmark that decomposes user interactions into three semantically typed graph views paired with a four-class stance taxonomy and a human-audited LLM silver-labeling pipeline. GCN on the ENDORSE view reaches Macro-F1=0.44 on POLITISKY24, and a feature ablation confirms that interaction topology is the operative signal (random features match sentence-BERT at 0.44; degree-only drops to 0.36). DSP saturates under real partition imbalance (real-to-null ratio 1.03 versus RWC's 7.69 on the same graph), identifying a previously undocumented evaluation-level failure mode. The annotation audit reveals RLHF-induced class-conditional output suppression: ChatGPT 5.2 produces zero BR predictions (F1=0.00) despite valid structured output, while Sonnet 4.6 ET achieves BR F1=0.62; inter-model unanimous agreement of 15.5% falls below the $\sim$ 20.1% random baseline. PolarBench releases all datasets, code, and evaluation pipelines to support reproducible investigation of stance detection on the decentralized web.

References

[1] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. Quantifying controversy in social media. In *WSDM*, 2016.
[2] G. Preti, M. Riondato, A. Gionis, and G. De Francisci Morales. DSP: A statistically-principled structural polarization measure. In *WSDM*, 2026.
[3] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.
[4] W. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. In *NeurIPS*, 2017.
[5] P. Veličković et al. Graph attention networks. In *ICLR*, 2018.
[6] J. Zhu, Y. Yan, L. Zhao, M. Heimann, L. Akoglu, and D. Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. In *NeurIPS*, 2020.
[7] N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP*, 2019.
[8] Meta AI. The Llama 3 herd of models. *arXiv:2407.21783*, 2024.
[9] F. Gilardi, M. Alizadeh, and M. Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *PNAS*, 120(30):e2305016120, 2023.
[10] P. Törnberg. ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot learning. *arXiv:2304.06588*, 2023.
[11] D. Kumar, Y. AbuHashem, and Z. Durumeric. Watch your language: Investigating content moderation with large language models. *arXiv:2309.14517*, 2023.
[12] M. Alizadeh et al. Open-source LLMs for text annotation: A practical guide for model setting and fine-tuning. *Journal of Computational Social Science*, 8:17, 2025.
[13] M. Li and F. Conrad. Advancing annotation of stance in social media posts: A comparative analysis of large language models and crowd sourcing. *arXiv:2406.07483*, 2024.
[14] S. Masud et al. Hate personified: Investigating the role of LLMs in content moderation. In *EMNLP*, 2024.
[15] P. Sachdeva and T. van Nuenen. Normative evaluation of large language models with everyday moral dilemmas. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 690–709, 2025.
[16] J. L. Nunes, G. F. C. F. Almeida, M. de Araujo, and S. D. J. Barbosa. Are large language models moral hypocrites? A study based on moral foundations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 1074–1087. AAAI Press, 2025.
[17] Z. Cheng, Y. Huang, H. Xu, S. Sojoudi, X. Zhao, D. Song, and S. Mei. OVERT: A benchmark for over-refusal evaluation on text-to-image models. In *NeurIPS*, 2025.
[18] W. Xia, H. Tang, and L. Li. Ideology as a problem: Lightweight logit steering for annotator-specific alignment in social media analysis. *CoRR*, abs/2601.04207, 2026.
[19] M. Zhang, J. He, T. Ji, and C.-T. Lu. Don't go to extremes: Revealing the excessive sensitivity and calibration limitations of LLMs in implicit hate speech detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 12073–12086, Bangkok, Thailand, 2024.
[20] N. B. Weidmann, M. Faulborn, and D. García. Large language models are democracy coders with attitudes. *PS: Political Science & Politics*, 59(1):17–23, 2026.
[21] D. Yüksel, M. C. Catalbas, and B. Oc. Language-dependent political bias in AI: A study of ChatGPT and Gemini. *arXiv:2504.06436*, 2025.
[22] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 1977.
[23] J. L. Fleiss, B. Levin, and M. C. Paik. *Statistical Methods for Rates and Proportions*. John Wiley & Sons, 2003.
[24] J. Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.
[25] R. Artstein and M. Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.
[26] V. Simaki et al. Evaluating stance annotation of Twitter data. *Research in Corpus Linguistics*, 2022.
[27] M. Parent et al. Annotating hate speech towards identity groups. In *RANLP*, 2025.
[28] T. Vo et al. ViTHSD: Exploiting hatred by targets for hate speech detection on Vietnamese social media texts. *arXiv:2404.19252*, 2025.
[29] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling. Modeling relational data with graph convolutional networks. In *ESWC*, 2018.
[30] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. S. Yu. Heterogeneous graph attention network. In *WWW*, 2019.

A Inter-Annotator Agreement Standards

The main text condenses the κ=0.4757 justification to two sentences; this appendix provides the full justification for reviewers who wish to assess the annotation-quality standard in detail. Our human gold annotations reach Cohen's κ=0.4757, which falls within the "moderate" range (0.41–0.60) defined by Landis and Koch [22] and the "fair to good" band (0.40–0.75) established by Fleiss et al. [23]. In computational linguistics, annotating subjective discourse-level phenomena is inherently difficult, and strict thresholds (e.g., $\kappa \geq 0.8$) are not realistic for ambiguous text [24, 25]. For politically charged social media tasks, moderate agreement is the prevailing standard: Simaki et al. report $\kappa \approx 0.42$ as "moderate but acceptable" for nuanced Twitter stance annotation [26], Parent et al. consider $\kappa \approx 0.54$ sufficient for implicit hate speech due to extreme semantic variance [27], and Vo et al. operationalize $\kappa > 0.40$ as a reliable threshold for targeted social media hate speech [28]. Our κ=0.4757 is consistent with all three benchmarks and constitutes a reliable baseline for the PolarBench annotation quality standard.

B Algorithm 3: Cross-Platform Transfer Protocol

This appendix documents the full cross-platform transfer procedure without consuming body space.

Algorithm 3 Cross-Platform Transfer Protocol

Require: Source graph $\mathcal{G}^S$ with trained GCN f^S, target graph $\mathcal{G}^T$, target features $\mathbf{X}^T \in \mathbb{R}^{n_T \times d_T}$
Ensure: Predictions on target nodes
// **Dimension alignment via PCA** (384 → 768)
if $d_T \neq d_S$ **then**
  $\phi \leftarrow \text{PCA}(\mathbf{X}^T, d_S)$
  $\mathbf{X}^T \leftarrow \phi(\mathbf{X}^T)$
end if
// **Zero-shot**
$f^T_{\text{zero}} \leftarrow f^S$ with frozen conv layers, random output head
Evaluate f^T_{zero} on target test set
// **Fine-tuned**
$f^T_{\text{ft}} \leftarrow f^S$ (all layers unfrozen)
Fine-tune f^T_{ft} on 20% of target labeled nodes for 20 epochs
Evaluate f^T_{ft} on target test set
return Zero-shot and fine-tuned Macro-F1

We note that the asymmetric PCA projection is a limitation of the current protocol; future work should replace it with a learned projection head that adapts jointly to both source and target feature distributions.

C Semantic View Sensitivity Analysis

This appendix supports the choice of k=10, ϵ=0.60 as the default Semantic view configuration. Table 10 reports a representative subset of the 25 tested configurations; the full configuration grid is available in the released code.

Table 10: Semantic view sensitivity (GCN Macro-F1 on PolitiSky24).

k	ϵ	Edges	Macro-F1
3	0.60	14,302	0.27
5	0.60	21,117	0.27
10	0.50	41,082	0.26
10	0.60	29,156	0.28
10	0.70	12,031	0.27
20	0.60	47,618	0.27
50	0.60	108,402	0.26

Figure 9 visualizes the flat response curve: the Semantic view's uninformativeness is a property of the view, not a tuning artifact.

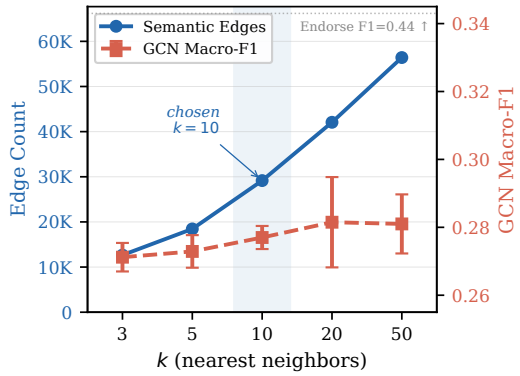


Figure 9: Sensitivity of the Semantic view to kNN hyperparameters at ϵ=0.60. GCN Macro-F1 remains flat at 0.26–0.28 across all configurations, confirming the view is robustly uninformative for stance classification in isolation.

D Toxicity Filtering Ablation for Conflict View

This appendix supports the τ=0 construction decision for the Conflict view.

Table 11: Toxicity filter ablation on Conflict (PolitiSky24).

τ	Edges Retained	% Retained	GCN Macro-F1
0.0	156,705	100%	0.37
0.3	8,305	5.3%	0.32
0.5	412	0.26%	0.29
0.7	11	$<$ 0.01%	—
0.9	0	0%	—

Mean user toxicity on PolitiSky24 is 0.10, indicating that political conflict on Bluesky manifests as substantive disagreement rather than toxic speech; retaining all quote-posts and adversarial replies is therefore the construction that preserves the most stance signal.

E Scalability and Complexity Analysis

Per-stage computational cost on the full PolitiSky24 dataset is as follows. Data collection is $O(|\mathcal{I}|)$ and took $\sim$ 4 hours of streaming to ingest 18.4M posts. Graph construction is dominated by the kNN step at $O(n^2 d)$ and completed in $\sim$ 2 min on GPU for n=8,467 and d=384. Silver labeling is $O(n \cdot L)$ where L is the prompt length; the full pipeline took $\sim$ 3 hours on a single A100 and $\sim$ 20 min with vLLM batch inference. GNN training on the largest view (Endorse) took $\sim$ 90 sec per epoch, with full training completing in under 5 min. The dominant cost is Polaris null model generation for DSP, which requires $\sim$ 1–2 hours on CPU per dataset.

F Complete Per-Class F1 for GNN Models

This appendix extends Table 5 to all GNN models. Per-class F1 for GCN-Endorse, GCN-Conflict, and the fusion architectures are being finalized against the held-out test set under the final 5-seed evaluation protocol, and complete values will be released with the final code and dataset bundle. The values reported so far confirm the aggregate trend: Pro and Neutral F1 scale with majority-class frequency, while Anti and BR F1 remain the hardest targets across all architectures, consistent with the text-only baseline profile in Table 5.