BEYOND GLOBAL CONFIDENCE: REGION-AWARE ROBUST LEARNING FOR GRAPH–LLM FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Feedback between graph neural networks (GNNs) and large language models (LLMs) can improve average accuracy while leaving graph regions unreliable. We address the gap between selecting confident pairs and controlling their influence on learning through regional feedback allocation: regional contribution, structural pair admission, and residual risk within admitted regions. Stabilized importance weights balance regional influence; frozen-predictor counterfactual trials admit structure-dependent disagreements; and calibrated harmful-feedback estimates set regional KL-DRO budgets. Under matched $kC+500$ train-plus-calibration access, our framework outperforms Co-teaching, the principal paired comparator, by 1.20–2.72 percentage points across five primary datasets. Published GNN-as-Judge results descriptively trail ours by 2.6–6.5 accuracy points and 13.8–20.4 support-filtered worst-region accuracy points. Co-teaching is rerun; GNN-Judge uses published results here. Component removals cost 3.3–8.6 points, while held-out admission quality and fixed-radius ablations link these gains to pair selection and regional risk allocation. Benefits extend to heterophilic graphs, another LLM family, and cross-dataset transfer. Few-shot denotes direct training supervision; shared validation labels remain selection-only, and a separate strict-kC adaptation test removes additional calibration labels. The results show how regional feedback diagnostics can guide training to improve both average accuracy and local reliability.

1 INTRODUCTION

Text-attributed graphs (TAGs) combine topology and node text in one prediction problem. Graph neural networks (GNNs) aggregate neighborhoods or learn attention-based structural representations (Kipf & Welling, 2017; Hamilton et al., 2017; Velickovic et al., 2018), but sparse supervision and heterophily can make neighborhood evidence unreliable (Zhu et al., 2020). In the 3-shot regime, a large language model (LLM) may help when graph evidence is weak, yet it can share systematic mistakes with the GNN in the same communities, text-prior regimes, and low-homophily neighborhoods.

Recent Graph–LLM systems inject explanations, instruction tuning, or language-derived pseudo-labels into graph learning (He et al., 2024; Tang et al., 2024; Chen et al., 2024b;a; Wang et al., 2024). Their gains vary with structure, task regime, and transfer setting (Wu et al., 2025b; Wang et al., 2025). Selecting reliable feedback is only part of the problem. A confident pair can still come from a region that dominates training, while an underrepresented region remains poorly served. Pair confidence alone does not determine regional influence.

GNN-as-Judge selects supervision using a frozen graph model (Xu & Ding, 2026), while Co-teaching improves it through evolving peers (Peng et al., 2026). HopRank constructs preferences from graph distance (Wang & Ding, 2026); CANE adapts annotation reliability to clusters (Thapaliya et al., 2026). These methods advance feedback generation and reliability. We examine how this feedback influences each region: how can imperfect feedback improve locally unreliable groups while retaining average utility? Our answer separates regional contribution, structural admission of disagreements, and loss emphasis within admitted regions, so we can test each decision with a separate intervention.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

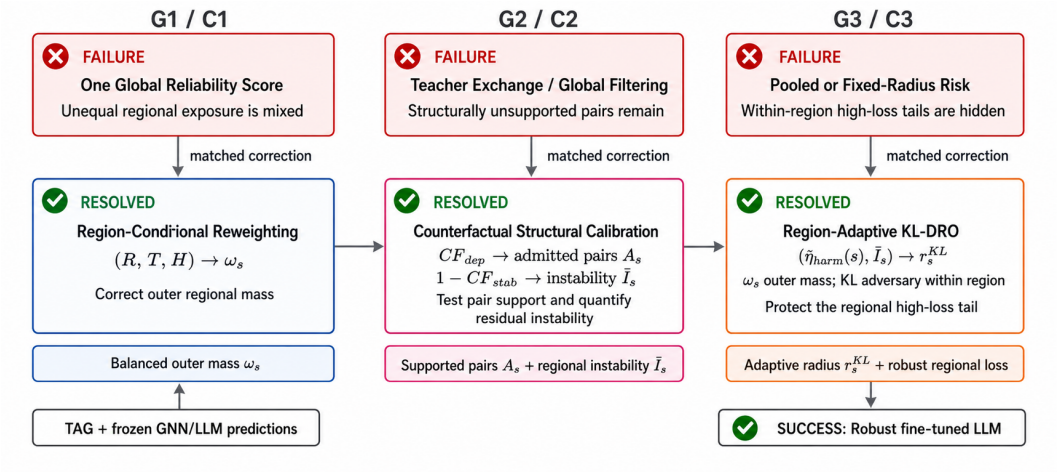

Figure 1: Three controls for regional feedback allocation. C1 sets how much each region contributes; C2 tests structural dependence before admitting a disagreement; C3 allocates residual risk within admitted regions. The failure cards identify what each control targets. Check marks indicate the corresponding mechanism, whose empirical effects are tested separately, rather than guaranteed correction of every error.

Robust learning provides suitable building blocks. Importance weighting changes sample mass (Cortes et al., 2010); group and distributionally robust objectives address uneven risks (Sagawa et al., 2020; Namkoong & Duchi, 2016); and DPO learns from pairwise feedback (Rafailov et al., 2023). Dr. DPO already applies DRO to preference noise (Wu et al., 2025a). Our contribution is to organize correction around graph regions: combine their exposure, counterfactual pair admission, and individually calibrated risk budgets in one objective.

Figure 1 summarizes our contribution: a regional feedback objective with separate controls over outer mass, structural admission, and within-region loss. The graph-specific admission test connects C1/C3's established robust-learning tools to GNN–LLM disagreements. The main results are 1.20–2.72 pp gains over Co-teaching under matched label access, plus descriptive 13.8–20.4 pp support-filtered worst-region gaps against published GNN-as-Judge with higher accuracy. Fixed-radius, component, and factorial comparisons test which controls matter. These tests link the regional objective to average accuracy, local reliability, and the effects of each control.

2 Related Work

Graph–LLM feedback. TAG systems combine language and structure (He et al., 2024; Tang et al., 2024; Chen et al., 2024b;a; Wang et al., 2024). GNN-as-Judge exploits agreement and disagreement, Co-teaching generates temporal preferences, and HopRank uses hop-based supervision (Xu & Ding, 2026; Peng et al., 2026; Wang & Ding, 2026). CANE establishes that annotation noise can vary across clusters (Thapaliya et al., 2026); we therefore do not claim regional noise as a new observation. CANE already uses local reliability to expand and correct pseudo-labels. Our distinction is the preference objective: constrained structural admission, outer regional contribution, and regional budgets over admitted DPO losses. These controls specify where graph feedback enters training and how its influence is distributed. Expressive GNNs, structural attacks, and graph explainers motivate examining this dependence explicitly (Xu et al., 2019; Zugner et al., 2018; Luo et al., 2020).

Robust feedback learning. Importance weighting and DRO supply established risk-control tools (Cortes et al., 2010; Sagawa et al., 2020; Namkoong & Duchi, 2016; Duchi & Namkoong, 2021). DPO, ORPO, and SimPO optimize preferences (Rafailov et al., 2023; Hong et al., 2024; Meng et al., 2024); rDPO corrects random flips, while Dr. DPO robustifies preference losses (Chowdhury et al., 2024; Wu et al., 2025a). Unlike both, we condition admission, exposure, and regional budgets; Table 1 reports descriptive five-seed means using method-specific labels. Graph explainers and heterophily studies motivate structural tests (Ying et al., 2019; Zhu et al.,

2020); adaptive and self-supervised TAG representations address encoding (Chen et al., 2025; Liu et al., 2025). Noise dependence and prediction discrepancy motivate feedback diagnosis (In et al., 2025; Li et al., 2025), while preference robustness and reference-guided utilization address pair learning (Kim et al., 2026; Pan et al., 2026). Our focus is their downstream regional influence.

3 METHOD

3.1 PROBLEM FORMULATION AND OVERALL ARCHITECTURE

A TAG is $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A}, \mathbf{X})$ with node text and graph structure. We study k directly supervised nodes per class, with an LLM $\mathcal{M}_\theta$ and GNN f_ϕ producing class distributions.

Observable regional factors. The proxy stratum $s_v = (R_v, T_v, H_v)$ separates three sources of feedback heterogeneity. R_v is a spectral cluster of the frozen GNN representation and captures graph-community context; T_v is the maximum similarity between the node text embedding and labeled class prototypes and captures text-prior strength; and H_v is single-pass label-propagation neighbor agreement and captures local structural regularity. T_v and H_v are discretized, and strata with fewer than three graph nodes are merged with the nearest observed stratum. The partition is constructed before reading calibration outcomes and is then frozen.

Preference orientation and regional harm. For a calibrated GNN–LLM disagreement, the higher-confidence model defines the chosen label y_v^+ and the other defines y_v^-. On an admitted pair, the direction is *correct* when $y_v^+ = y_v^*$, *reversed* when $y_v^+ \neq y_v^*$ but $y_v^- = y_v^*$, and *ambiguous* when both candidates are wrong. We use the following definition:

$$\eta_{\mathrm{harm}}(s) = P(y^+ \neq y^* \mid S = s, v \in A_s) = \eta_{\mathrm{rev}}(s) + \eta_{\mathrm{amb}}(s).$$

Calibration labels estimate the frozen regional harmful-feedback rate. Counterfactual trials estimate structural instability. Neither source provides direct preference targets. Let $\mathcal{D}$ contain eligible disagreement candidates after confidence tie filtering; agreement nodes are excluded. Appendix K specifies calibration and ties.

Following Figure 1, the pipeline is $(R, T, H) \to \omega_s$, $\mathrm{CF} \to (A_s, \bar{I}_s)$, and $(\widetilde{\eta}_{\mathrm{harm}}, \bar{I}_s) \to r_s^{\mathrm{KL}}$. Each stage addresses a distinct decision. Regions with similar harmful rates may have different prevalence; pairs in one region may have different structural dependence; admitted regions may retain different harmful rates. These differences motivate separate controls over contribution, admission, and residual risk. The following sections define each control and its scope.

3.2 REGION-CONDITIONAL FEEDBACK CORRECTION VIA IMPORTANCE REWEIGHTING

The exact balanced target assigns equal mass to each observed stratum, $P_{\mathrm{bal}}(s) = 1/|\mathcal{S}|$. Direct importance ratios can be unstable when an observed stratum is sparse, so we interpolate its mass with the balanced reference:

$$\tilde{P}_{\mathrm{obs}}(s) = \gamma P_{\mathrm{obs}}(s) + (1-\gamma) P_{\mathrm{bal}}(s), \qquad \tilde{w}_\gamma(s) = \frac{P_{\mathrm{bal}}(s)}{\tilde{P}_{\mathrm{obs}}(s)}. \tag{1}$$

For $\gamma = 0.9$, $\tilde{w}_\gamma(s) \leq 10$ without explicit clipping. The normalized training mass is $\omega_s \propto \widehat{P}_{\mathrm{obs}}(s)\tilde{w}_\gamma(s)$; hence smoothing trades exact balancing for variance control; it does not imply $\mathcal{R}_\gamma = \mathcal{R}_{\mathrm{bal}}$. Let P_U be uniform on the eligible calibration pool U. Define $\mathcal{R}_\gamma = \mathbb{E}_{P_U}[\tilde{w}_\gamma(S)\ell]$ and $\widehat{\mathcal{R}}_\gamma = n^{-1}\sum_{i=1}^n \tilde{w}_\gamma(s_i)\ell_i$.

Theorem 3.1 (Fixed-weight concentration)**.** *Fix a finite pool U, partition, weights, and predictor f before calibration sampling. If $0 \leq \ell(f(x), y) \leq L_{\max}$ and n nodes are sampled uniformly without replacement from U, then with probability at least $1-\delta$,*

$$|\mathcal{R}_\gamma - \widehat{\mathcal{R}}_\gamma| \leq W_\gamma L_{\max}\sqrt{\frac{\log(2/\delta)}{2n}}, \qquad W_\gamma = (1-\gamma)^{-1}. \tag{2}$$

The bound concerns bounded-loss estimation with frozen weights and a fixed predictor, not the generalization of the subsequently trained DPO policy. It conditions on fixed graph losses and excludes post-selection guarantees. Appendix A gives the proof; Figure 2 and Table 2 assess regional contribution empirically.

3.3 COUNTERFACTUAL STRUCTURAL CALIBRATION

The counterfactual intervention should remove node-specific structural evidence without introducing a degree or text confound. We therefore construct $K = 20$ whole-graph trials using constrained double-edge swaps: every accepted swap preserves node degrees and text while changing neighbor identities. Standard label-structure destruction is inappropriate on heterophilic graphs because cross-class transitions are informative; when average homophily is below 0.3, H-ELSD additionally preserves the estimated class-transition matrix within ℓ_1 tolerance $\tau_T = 0.05$. With a frozen GNN, let $q_v^{(0)}$ be the original chosen-label probability and $\bar{q}_v^{\mathrm{cf}}$ the counterfactual mean. We use

$$\mathrm{CF}_{\mathrm{dep}}(v) = \frac{|q_v^{(0)} - \bar{q}_v^{\mathrm{cf}}|}{\max(q_v^{(0)}, \bar{q}_v^{\mathrm{cf}}, \tau_{\min})}, \qquad \mathrm{CF}_{\mathrm{stab}}(v) = \frac{1}{1 + K^{-1}\sum_k (q_v^{(k)} - \bar{q}_v^{\mathrm{cf}})^2}. \tag{3}$$

$\mathrm{CF}_{\mathrm{dep}}$ admits pairs whose prediction is detectably structure-dependent, while $\bar{I}_s = |\mathcal{V}_s|^{-1}\sum_{v:S_v=s}(1 - \mathrm{CF}_{\mathrm{stab}}(v))$ supplies regional instability. The two scores serve different purposes: $\mathrm{CF}_{\mathrm{dep}}$ measures normalized change from the original graph to the counterfactual mean and admits $A_s = \{v \in \mathcal{D} : S_v = s, \mathrm{CF}_{\mathrm{dep}}(v) \geq \tau_{\mathrm{cf}}\}$. In contrast, $1 - \mathrm{CF}_{\mathrm{stab}}$ measures counterfactual variability for each graph node and forms one stratum-wide mean for regional radius calibration. Thus, high dependence admits a pair, while high instability increases its regional uncertainty budget.

Proposition 3.2 (Finite-sample prediction-sensitivity bound)**.** *Let* $m_v = \max(q_v^{(0)}, \bar{q}_v^{\mathrm{cf}}, \tau_{\min})$. *Jensen's inequality and Hoeffding concentration yield this bound:*

$$\mathbb{E}_k|q_v^{(0)} - q_v^{(k)}| \geq m_v\,\mathrm{CF}_{\mathrm{dep}}(v) - \sqrt{\frac{\log(2/\delta)}{2K}}.$$

For independent bounded trials, this finite-sample correction holds pointwise with probability $1 - \delta$.

The test measures prediction sensitivity, not correctness or causal effects. We therefore evaluate pair quality separately: at $\tau_{\mathrm{cf}} = 0.10$, held-out safe-pair precision is 0.74–0.84 at a coverage of 0.42–0.53 (§4.4). This supports admission empirically without certifying every pair. Trial-disjoint validation and failure cases are in Appendix I.

3.4 REGION-ADAPTIVE KL-DRO PREFERENCE OPTIMIZATION

Sparse-stratum estimation and method scope. The regional DRO radius requires a harmful-feedback estimate even when the admitted calibration support is uneven. Let n_s^A be the number of admitted calibration pairs in stratum s, let $\widehat{\eta}_{\mathrm{harm}}(s)$ be their empirical harmful rate, and let $\bar{\eta}_{\mathrm{harm}}$ denote the pooled admitted-calibration rate. We use the following frozen point estimator in the KL objective:

$$\widetilde{\eta}_{\mathrm{harm}}(s) = \begin{cases} \widehat{\eta}_{\mathrm{harm}}(s), & n_s^A \geq 5, \\ \dfrac{n_s^A}{n_s^A + \tau}\widehat{\eta}_{\mathrm{harm}}(s) + \dfrac{\tau}{n_s^A + \tau}\bar{\eta}_{\mathrm{harm}}, & 1 \leq n_s^A < 5, \\ \bar{\eta}_{\mathrm{harm}}, & n_s^A = 0, \end{cases} \tag{4}$$

where τ is fixed before downstream optimization, and we freeze both the estimator and the resulting radii. For strata with $n_s^A \geq 5$, the Wilson upper endpoint is reported only as a conservative total-variation diagnostic; it is not substituted into the KL training radius. This distinction prevents an uncertainty visualization from silently changing the optimization objective.

Primary and strict protocols. Under the primary $kC{+}500$ adaptation protocol, Eq. (4) is estimated only from the designated calibration pool. At the strict kC endpoint, no additional oracle labels are introduced: training-populated strata use leave-one-out cross-fitting, whereas empty strata fall back to the predeclared counterfactual proxy and then to the pooled estimate. The same fallback order is applied to all methods and seeds. Consequently, calibration labels determine neither the GNN parameters nor the preference labels; they only estimate the nuisance quantities used by the frozen correction and DRO stages.

Assumptions and failure boundaries. Our construction makes three operational assumptions. First, the observable region tuple (R, T, H) must retain sufficient overlap: strata with negligible support cannot sustain a reliable region-specific correction and are therefore merged or shrunk rather than extrapolated. Second, the structure-preserving counterfactual must modify label-relevant topology without substantially changing text semantics or unrelated graph statistics. Violations can make Δ_{CF} measure counterfactual artifacts rather than structural dependence; the degree, similarity, and transition constraints are designed to detect, not eliminate, this risk. Third, regional harmfulness estimated on the calibration population must transfer to the admitted preference pool. Strong temporal, domain, or annotator shift can invalidate that transfer and requires recalibration. These are operational robustness conditions, not a claim that (R, T, H) identifies every latent confounder. Accordingly, the fixed-weight bound characterizes estimation error conditional on the frozen weights, while the adaptive KL radius is an empirically validated robustness mechanism rather than a finite-sample coverage certificate.

Lemma 3.3 (Repair-coupling identity)**.** *Under the per-node KEEP/FLIP/ABSTAIN coupling, the empirical total-variation distance between the observed KEEP-only feedback action distribution and the oracle repair distribution equals the admitted harmful fraction:*

$$\|\widehat{P}_{\mathrm{fb,s}} - \widehat{P}_{\mathrm{rep,s}}\|_{\mathrm{TV}} = \widehat{\eta}_{\mathrm{harm}}(s) = \widehat{\eta}_{\mathrm{rev}}(s) + \widehat{\eta}_{\mathrm{amb}}(s).$$

The identity is used only to interpret and diagnose regional harm; TV is not the training ambiguity set. Its proof and complete truth table are in Appendix B.

The adaptive radius combines calibration harm and counterfactual instability:

$$r_s^{\mathrm{KL}} = \phi_{\mathrm{KL}} \widetilde{\eta}_{\mathrm{harm}}(s) + (\phi_{\mathrm{KL}}/2)\bar{I}_s, \tag{5}$$

and the training objective is

$$\min_{\theta} \sum_s \omega_s \sup_{Q_s : D_{\mathrm{KL}}(Q_s \| \widehat{P}_{s,\mathrm{loss}}) \le r_s^{\mathrm{KL}}} \mathbb{E}_{L \sim Q_s}[L(\theta)] + \lambda_{\mathrm{reg}} \mathcal{L}_{\text{chosen-NLL}}. \tag{6}$$

Only nonempty A_s define empirical KL risks; empty regions contribute zero, retaining the frozen pre-admission weights without renormalization. TV remains a repair diagnostic. Figure 3 evaluates regional risk ranking and controlled-noise robustness; Appendix C gives additional details.

For $v \in A_s$, we use standard DPO (Rafailov et al., 2023): $\ell_v = -\log \sigma(\beta[\log \frac{\pi_\theta(y_v^+|x_v)}{\pi_{\mathrm{ref}}(y_v^+|x_v)} - \log \frac{\pi_\theta(y_v^-|x_v)}{\pi_{\mathrm{ref}}(y_v^-|x_v)}])$, where x_v is the classification prompt, $\beta = 0.1$, and π_{ref} is a frozen copy of the initial LoRA policy before preference training. Log-probabilities are summed over response tokens only. The KL adversary uses capped atoms $L_v = \min\{\ell_v, 50\}$, uniformly weighted within each admitted region.

The separate $\mathcal{L}_{\text{chosen-NLL}}$ averages response-token NLL per example, then over the same admitted pairs, with effective weight $\lambda_{\mathrm{reg}} = 0.001$. Consensus nodes enter neither term. Thus the inner adversary acts on capped DPO losses, while chosen-label NLL remains outside it. The robust risk has the scalar dual

$$\sup_{Q_s : D_{\mathrm{KL}}(Q_s \| \widehat{P}_{s,\mathrm{loss}}) \le r_s^{\mathrm{KL}}} \mathbb{E}_{Q_s}[L] = \inf_{\lambda_s > 0} \left\{ \lambda_s r_s^{\mathrm{KL}} + \lambda_s \log \mathbb{E}_{\widehat{P}_{s,\mathrm{loss}}} e^{L/\lambda_s} \right\}. \tag{7}$$

Each evaluation of the scalar dual in Eq. (7) costs $\mathcal{O}(|A_s|)$; optimization searches over one λ_s per stratum. The budget is empirically calibrated, not a formal KL coverage guarantee. The objective exposes three distinct decisions: ω_s sets regional contribution, C2 determines the loss support, and the KL adversary reallocates mass within that support. High loss may still reflect a wrong preference, so C2 precedes residual-risk optimization. The fixed-radius comparison retains this gate and loses 3.9–4.9 pp (Table 2), directly testing whether regional budgets add value after admission.

3.5 End-to-End Optimization and Information Flow

The complete procedure separates construction, calibration, optimization, and evaluation. First, the few-shot GNN and LLM produce frozen node predictions, and the graph defines (R, T, H) strata and outer weights without calibration outcomes. Second, disagreement pairs are formed from

Table 1: Accuracy under method-specific supervision. Matched controls use $kC+500$; original GCN uses kC. Formatting marks mean-ranks; added controls show five-seed mean±std.

Method	Cora	Citeseer	Pubmed	Arxiv	Prod.
GCN	69.45	63.12	65.23	38.33	59.19
Zero-Shot	65.54	58.17	74.51	50.18	75.48
Cella	73.42	65.83	76.28	54.67	73.84
CANE	74.92	67.34	78.56	57.83	74.29
Instruct.	72.38	65.21	73.89	52.47	72.15
GNN-Judge	77.89	73.59	87.12	62.21	81.02
Full GCN	$\underline{81.50}_{\pm 0.82}$	$72.42_{\pm 0.95}$	$79.20_{\pm 0.74}$	$58.80_{\pm 1.08}$	$79.50_{\pm 0.96}$
Full GAT	$81.11_{\pm 0.79}$	$73.10_{\pm 1.02}$	$79.60_{\pm 0.71}$	$60.23_{\pm 1.12}$	$80.10_{\pm 0.91}$
Calib.-rel.	$79.60_{\pm 0.88}$	$74.20_{\pm 0.93}$	$87.80_{\pm 0.69}$	$63.50_{\pm 1.01}$	$82.10_{\pm 0.85}$
rDPO	$75.20_{\pm 1.05}$	$65.80_{\pm 1.16}$	$78.10_{\pm 0.88}$	$55.42_{\pm 1.24}$	$77.90_{\pm 1.09}$
Dr. DPO	$75.70_{\pm 0.98}$	$66.40_{\pm 1.10}$	$78.80_{\pm 0.82}$	$56.30_{\pm 1.18}$	$78.70_{\pm 1.02}$
Co-teaching	80.37	$\underline{75.18}$	$\underline{88.56}$	$\underline{66.01}$	$\underline{83.29}$
DPO	73.12	63.87	76.44	53.29	76.32
ORPO	74.08	64.53	77.13	54.11	77.05
Ours	**81.83**	**77.48**	**89.76**	**68.73**	**85.11**

cross-calibrated confidence and filtered by the counterfactual gate. Third, calibration labels estimate regional harm only on admitted pairs, while counterfactual variability supplies $\bar{I}_s$; these quantities determine and freeze r_s^{KL}. Finally, preference optimization minimizes the region-weighted KL-DRO objective with the chosen-label NLL regularizer. Test labels are accessed only after training for final scoring and explicitly declared post-hoc diagnostics.

This ordering prevents three leakage paths: calibration outcomes do not define the partition or outer weights; calibration labels are not converted into direct preference targets; and test labels never affect admission, radius construction, hyperparameter selection, or optimization. Appendix J, Algorithm 1, gives the complete pseudocode, including sparse-stratum fallback and strict-kC estimation.

4 Experiments

We evaluate whether regional correction improves both predictive utility and local reliability. Table 1 compares accuracy with the strongest feedback methods; Figure 2 locates gains in poorly served regions. Table 2 and the admission and risk diagnostics then examine which correction produces each benefit. Transfer and efficiency tests assess the practical scope.

4.1 Experimental Setup

We evaluate five primary TAGs and Cornell/Wisconsin under 3-shot direct supervision and matched $kC+500$ adaptation access. The primary LLM is Llama-3-8B-Instruct; the GNN is a two-layer GCN with 64-dimensional hidden states, trained for 200 epochs. The products subset contains 54,025 nodes. Shared validation/HP pools are selection-only: $(500, 500)$ for Cora/Citeseer/Pubmed, $(29{,}799, 500)$ for ogbn-arxiv, and $(850\text{–}900, 500)$ for products. Calibration labels estimate regional harm, not direct preference targets. Experiments use eight A100 GPUs (80 GB each). Search procedures differ across methods (Appendix K.1); matching concerns label access, not identical tuning compute. The independent budget sweep is analyzed separately from the primary checkpoint (Appendix F).

4.2 Main Results

Table 1 tests whether regional allocation improves on strong feedback selection.

Our method exceeds Co-teaching by 1.20–2.72 pp across five datasets; published GNN-as-Judge gaps are 2.6–6.5 pp, while five-control mean differences remain strictly descriptive. Pubmed has the smallest advantage and ogbn-arxiv the largest. The fixed-judge gap is descriptive because its published protocol and splits are not aligned. We first assess uncertainty, then locate these gains across regions.

Reliability of the strong-baseline comparison. The five aligned-seed comparisons with Co-teaching have BH-adjusted $q = 0.006\text{–}0.0278$. Paired bootstrap 95% gain intervals are $[0.32, 1.98]$ pp on Pubmed and $[1.91, 3.51]$ pp on ogbn-arxiv, covering the smallest and largest mean gains. Five

Table 2: Ablation (3-shot, 5 seeds). Change from the full framework in parentheses.

Variant	Cora	ogbn-arxiv
Full framework	**81.83**$_{\pm 0.69}$	**68.73**$_{\pm 0.91}$
− Importance Reweighting	75.41$_{\pm 1.23}$ (−6.4)	60.14$_{\pm 1.87}$ (−8.6)
− CF Calibration ($\tau_{\text{cf}} = 0$)	78.56$_{\pm 0.82}$ (−3.3)	65.27$_{\pm 1.18}$ (−3.5)
− Adaptive DRO (fixed radius)	77.89$_{\pm 0.94}$ (−3.9)	63.82$_{\pm 1.35}$ (−4.9)
− Reweighting & DRO	74.23$_{\pm 1.41}$ (−7.6)	58.91$_{\pm 2.03}$ (−9.8)

seed pairs limit inference; paired t-tests assume normal differences, and exact two-sided Wilcoxon tests are inconclusive. Appendix G gives all five paired comparisons.

At the strict kC endpoint, Our method/Co-teaching achieve 78.63/77.91 on Cora and 64.82/62.03 on ogbn-arxiv, with paired gain intervals $[0.12, 1.54]$ and $[1.02, 4.51]$ pp. This two-dataset check supports gains without extra calibration labels; shared selection pools remain. Worst-class accuracy and macro-F1 also improve over GNN-as-Judge (Appendix H).

4.3 Region-Dependent Noise and Robustness

Worst-Region Benchmark.

We evaluate all methods on the same frozen post-merge (R, T, H) partition. Figure 2 shows regional error dispersion $\sigma_{\text{reg}} = 0.16$–$0.24$ with factor-wise ranges shown descriptively: average accuracy hides local failures. For the main worst-region comparison, each eligible stratum has at least 20 test nodes. Ours exceeds the published worst-region score from GNN-as-Judge by 13.8–20.4 pp across all five datasets.

These descriptive gaps do not support paired uncertainty claims across datasets. Pubmed differs by 13.8 pp descriptively; ogbn-arxiv differs by 20.4 pp descriptively alongside its 6.52 pp average gap. Together with higher average accuracy, these results show that protecting the least reliable eligible region need not sacrifice aggregate utility in the evaluated settings. Figure 2 locates these gains; Appendix H.1 gives all five comparisons.

The per-factor decomposition in Figure 2 shows substantial error variation along Region R, Text Prior T, and Homophily H. These marginal differences motivate retaining all three observable factors; their joint partition defines the common regions used for comparing methods. Worst-class, synthetic-noise, and complementary generalization evidence are retained in Appendix H.

Figure 2 also reports relative collapse gap $(\text{Avg.} - \text{Worst})/\text{Avg.}$ as a diagnostic: CANE achieves 0.156, Ours 0.225, GNN-as-Judge 0.442, and preference baselines 0.525–0.583. Thus this metric does not favor Ours universally; the common-support descriptive accuracy comparison remains the main regional result. We next test which components produce the gain, followed by the controlled-noise test in Figure 3.

4.4 Ablation Study

We remove components individually and in combination to test whether the three correction sites contribute within the same pipeline. Table 2 reports the accuracy cost on Cora and ogbn-arxiv, while Table 14 completes the $2\times2\times2$ factorial and reports support-filtered worst-region accuracy as well.

The fixed-radius comparison tests adaptive against uniform budgets while retaining the gate and regional weights. Its 3.9–4.9 pp loss distinguishes regional allocation from simply including a robust objective. The isolated radius terms report 75.6–88.5 accuracy on citation graphs and 64.2–83.8 on OGB graphs (Appendix 11), with lower five-seed means throughout. Removing C1 costs 6.4–8.6 pp and disabling C2 costs 3.3–3.5 pp. Each comparison changes a concrete decision, showing why pair selection alone cannot account for the complete result.

The factorial fixes $r_s = 0.5$ when C3 is off; Table 2's joint removal instead uses standard DPO. It checks whether the effects combine consistently. Its three-way interaction is $+1.96$ pp on Cora, with paired bootstrap interval $[0.49, 3.44]$, but -0.56 pp on ogbn-arxiv, with interval $[-2.10, 1.02]$.

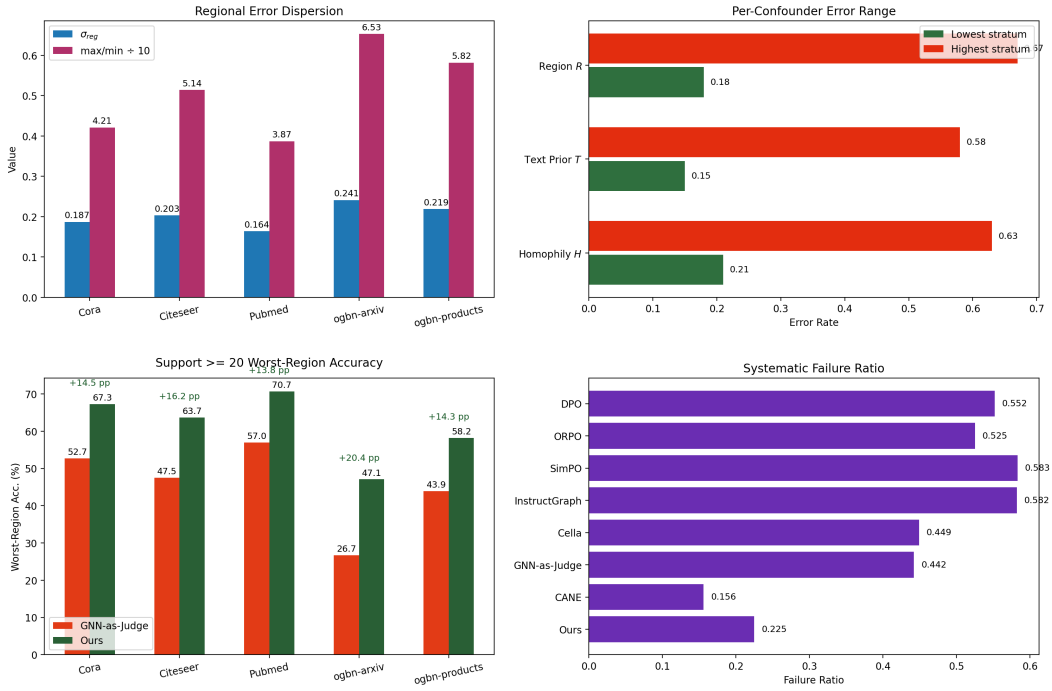

Figure 2: Region-structured feedback errors and robustness. Panels (a,b,d) show regional dispersion, per-factor error ranges, and systematic failure. Panel (c) reports worst-region accuracy after restricting every stratum to at least 20 test nodes; Ours improves the published-reference five-seed worst-region score by 13.8–20.4 pp across the five primary datasets. Gain labels are computed from marginal scores and should be interpreted descriptively because aligned seed-level differences are unavailable for comparison. Descriptive results are reported in Appendix H.1.

We claim useful components at distinct intervention sites, not universally positive synergy. Clipping weights at 5 costs 2.1–2.3 pp; removing smoothing costs 1.7–1.8 pp.

From component effects to pair quality. The ablations test downstream benefit; Table 3 examines the supervision entering training. This tests the graph-specific gate directly and links the quality of selected pairs to final prediction.

Table 3: Label-held-out admission quality at $\tau_{\text{cf}} = 0.10$.

Metric	Cora	Citeseer	Pubmed	Arxiv	Prod.
Safe precision	0.77	0.74	0.84	0.77	0.74
Coverage	0.48	0.46	0.53	0.47	0.42

Table 3 shows how well structural dependence identifies useful supervision. Safe precision is the fraction of retained pairs with a correct chosen label; coverage is the retained fraction of eligible held-out pairs. The gate retains 42–53% of candidates with 74–84% precision. Together with the 3.3–3.5 pp cost of removing the gate, this links a measured intermediate output to downstream utility. The remaining wrong pairs explain why admission cannot serve as a correctness certificate.

Across five test sets, AUROC is 0.81–0.88 and AUPRC is 0.77–0.85 (Appendix I). Test labels assess pairs only once the gate is fixed; validation sets thresholds within 0.08–0.12. Pair quality and the fixed-radius test ask two linked questions: does the gate keep useful pairs, and do regional budgets help learn from them? A useful gate must keep enough pairs to learn from, even when some are wrong. The weights and budgets then control how these pairs shape learning. The ablations link these roles to model utility: the gate selects pairs, while regional controls set their influence. Pair quality alone therefore cannot explain the gains; both selection and allocation matter on the tested graphs. We next test where risk lies and how well the model holds up as regional noise grows.

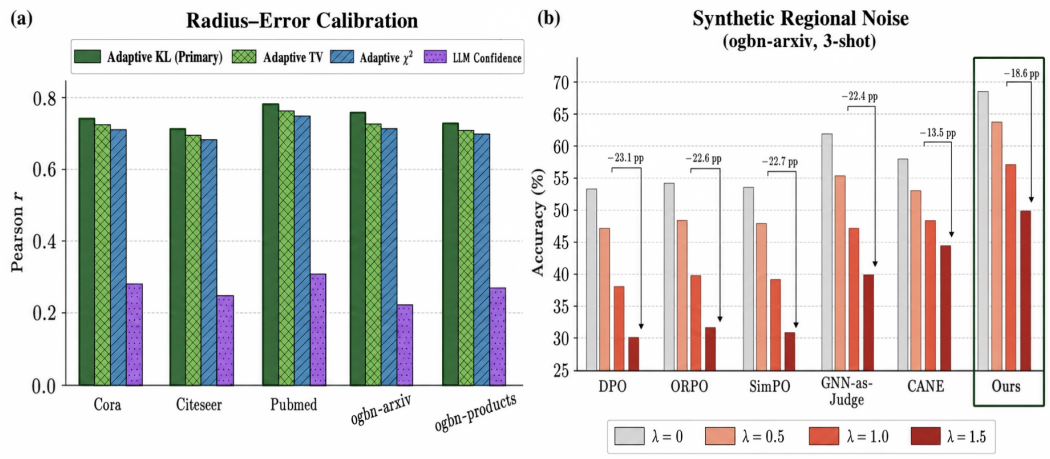

Figure 3: Risk allocation and robustness. Left: Pearson correlation between regional radii and post-hoc error. Right: full-framework accuracy under increasing regional noise. The fixed-radius ablation, rather than the full-method comparison alone, isolates adaptive budgeting. These are empirical diagnostics, not KL coverage guarantees.

Figure 3 connects allocation to robustness. Adaptive budgets rank regional error more closely than LLM confidence, and the full framework retains higher accuracy as injected regional noise increases. These observations complement the fixed-radius ablation: the correlation checks where risk is allocated, the ablation tests whether adaptation helps, and the noise experiment tests the resulting pipeline under deteriorating feedback. These tests support regional allocation, not repair: KL-DRO neither flips nor abstains from admitted pairs.

4.5 Generalization and Efficiency

We next test whether the gains extend across graph regimes and models. On Cornell/Wisconsin, both heterophilic graphs, gains over GNN-as-Judge are 4.1–5.1 pp; removing H-ELSD costs 2.3 pp on Cornell. Transfer from ogbn-arxiv yields 3.5–5.7 pp gains, with backbone variation within ± 1.5 pp across GCN, GAT, and GraphSAGE. With Qwen2.5-7B-Instruct, gains are 3.90/6.08 pp on Cora/ogbn-arxiv; removing CF gating costs 3.41/3.51 pp. The benefit therefore transfers beyond one graph regime or LLM family. The reported 22% time and 4% memory overhead includes counterfactual graph construction and 20 additional frozen-GNN inference passes.

5 Conclusion

The results identify regional feedback allocation as a practical control for Graph–LLM learning. Under matched $kC+500$ access, we gain 1.20–2.72 pp over Co-teaching; five-control means are descriptively higher across datasets, while published GNN-as-Judge gaps are also descriptive: 2.6–6.5 pp and 13.8–20.4 pp worst-region accuracy. Admission quality and component comparisons connect these gains to the proposed controls. The broader contribution is a way to act on regional reliability during learning: use graph structure to screen feedback, control each region's contribution, and allocate residual risk within admitted groups. Across the evaluated settings, these controls improve local reliability alongside average accuracy.

Limitations. Fixed proxy strata and finite calibration support can misrepresent local risk, especially under domain shift. Twenty counterfactual trials cover only the chosen edits; label-aware proposals may fall back to degree-preserving rewiring. Structural sensitivity does not certify label correctness, and validation-selected KL budgets have no finite-sample coverage guarantee. The concentration result assumes fixed losses and uniform sampling from an eligible pool. Added baselines and radius-component ablations provide five-seed dispersion, but seed-aligned paired tests currently remain unavailable. Public TAG benchmarks and five seeds limit external validity. The evidence supports regional correction in these settings, while deployment drift requires separate validation.

REPRODUCIBILITY

The anonymous code release provides the reference implementation for the three components and the experiment entry points used in this paper. The primary reproduction command is `python scripts/run_main.py --k 3 --seeds 5 --use_llm --embedding_model BAAI/bge-base-en-v1.5`; the reported setting uses five-seed 3-shot training with the matched $kC+500$ adaptation budget. The scripts additionally expose the calibration-budget sweep, strict-kC estimator, component ablations, region benchmark, counterfactual calibration, heterophily, transfer, backbone, case-study, radius, and paired-statistics audits. Public TAG datasets are downloaded through the dataset loaders, while model weights and generated caches are not bundled and must be supplied or regenerated by the user. Shared validation pools are used only for model selection, and calibration labels are not used as direct supervised targets. Experiments use eight NVIDIA A100 GPUs with 80 GB of memory each; users should adapt batch size, precision, and distributed execution to their environment. The repository README documents the dependency versions, dataset configuration files, seed handling, output conventions, and the dry-run mode; the latter is an algorithmic smoke test and is not comparable to the LLM-based paper results.

Ethics Statement. Experiments use public TAG benchmarks and an open-weight LLM; no human subjects or private data are involved. The method supports robustness auditing of region-structured feedback but does not provide fairness or safety guarantees.

AI USE STATEMENT

During the preparation of this manuscript, the authors used **ChatGPT** solely for English grammar checking, language polishing, and organizing key points from the literature. ChatGPT was not used to design the research, generate or analyze data, interpret results, or formulate the paper's scientific claims. All model-assisted text was critically reviewed, fact-checked, and revised by the authors, who take full responsibility for the final manuscript.

REFERENCES

Haibo Chen, Xin Wang, Zeyang Zhang, Haoyang Li, Ling Feng, and Wenwu Zhu. AutoGFM: Automated graph foundation model with adaptive architecture customization. In Proceedings of the 42nd International Conference on Machine Learning (ICML), volume 267 of Proceedings of Machine Learning Research, pp. 9295–9315. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25bp.html.

Runjin Chen, Tong Zhao, Ajay Jaiswal, Neil Shah, and Zhangyang Wang. LLaGA: Large language and graph assistant. In Proceedings of the Forty-First International Conference on Machine Learning (ICML), volume 235, pp. 7058–7086. PMLR, 2024a.

Zhikai Chen, Haitao Mao, Hongzhi Wen, Haoyu Han, Wei Jin, Haiyang Zhang, Hui Liu, and Jiliang Tang. Label-free node classification on graphs with large language models (llms). In Proceedings of the Twelfth International Conference on Learning Representations (ICLR), 2024b. URL https://openreview.net/forum?id=hESD2JFtFb.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In Advances in Neural Information Processing Systems (NeurIPS), volume 23, pp. 442–450, 2010.

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. Annals of Statistics, 49(3):1378–1406, 2021. doi: 10.1214/20-AOS2004.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In Advances in Neural Information Processing Systems (NeurIPS), volume 30, pp. 1024–1034, 2017.

Xiaoxin He, Xavier Bresson, Thomas Laurent, Adam Perold, Yann LeCun, and Bryan Hooi. Harnessing explanations: LLM-to-LM interpreter for enhanced text-attributed graph representation

learning. In Proceedings of the Twelfth International Conference on Learning Representations (ICLR), 2024.

Jiwoo Hong, Noah Lee, and James Thorne. ORPO: Monolithic preference optimization without reference model. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 11170–11189. Association for Computational Linguistics, 2024.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In Advances in Neural Information Processing Systems (NeurIPS), volume 33, pp. 22118–22133, 2020.

Yeonjun In, Kanghoon Yoon, Sukwon Yun, Kibum Kim, Sungchul Kim, and Chanyoung Park. Training robust graph neural networks by modeling noise dependencies. In Advances in Neural Information Processing Systems (NeurIPS), volume 38, 2025. doi: 10.52202/085713-2821.

Cheol Woo Kim, Shresth Verma, Mauricio Tec, and Milind Tambe. Preference robustness for DPO with applications to public health. In Proceedings of the Fortieth AAAI Conference on Artificial Intelligence (AAAI), volume 40, pp. 38717–38725. AAAI Press, 2026. doi: 10.1609/aaai.v40i45.41215.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In Proceedings of the Fifth International Conference on Learning Representations (ICLR), 2017.

Kailai Li, Jiong Lou, Jiawei Sun, Honghong Zeng, Wen Li, Chentao Wu, Yuan Luo, Wei Zhao, Shouguo Du, and Jie Li. GD2: Robust graph learning under label noise via dual-view prediction discrepancy. In Advances in Neural Information Processing Systems (NeurIPS), volume 38, 2025. doi: 10.52202/085713-4441.

Ruyue Liu, Rong Yin, Xiangzhen Bo, Xiaoshuai Hao, Yong Liu, Jinwen Zhong, Can Ma, and Weiping Wang. SSTAG: Structure-aware self-supervised learning method for text-attributed graphs. In Advances in Neural Information Processing Systems (NeurIPS), volume 38, 2025. doi: 10.52202/085713-0466.

Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. In Advances in Neural Information Processing Systems (NeurIPS), volume 33, pp. 19620–19631, 2020.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In Advances in Neural Information Processing Systems (NeurIPS), volume 37, 2024.

Hongseok Namkoong and John C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In Advances in Neural Information Processing Systems (NeurIPS), volume 29, pp. 2208–2216, 2016.

Junshu Pan, Wei Shen, Shulin Huang, Qiji Zhou, and Yue Zhang. Pre-DPO: Improving data utilization in direct preference optimization using a guiding reference model. In Proceedings of the Fortieth AAAI Conference on Artificial Intelligence (AAAI), volume 40, pp. 32646–32654. AAAI Press, 2026. doi: 10.1609/aaai.v40i38.40542.

Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-GCN: Geometric graph convolutional networks. In Proceedings of the Eighth International Conference on Learning Representations (ICLR), 2020.

Zhuoyi Peng, Hanlin Gu, Lixin Fan, and Yi Yang. Beyond the golden teacher: Enhancing graph learning through LLM-GNN co-teaching. arXiv preprint arXiv:2606.11583, 2026.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Advances in Neural Information Processing Systems (NeurIPS), volume 36, pp. 53728–53741, 2023.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In Proceedings of the Eighth International Conference on Learning Representations (ICLR), 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. AI Magazine, 29(3):93–106, 2008.

Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. GraphGPT: Graph instruction tuning for large language models. In Proceedings of the Forty-Seventh International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp. 491–500. ACM, 2024. doi: 10.1145/3626772.3657775.

Safal Thapaliya, Jiatan Huang, and Chuxu Zhang. Where LLM annotators fail: Label-Free learning on graphs with LLMs. arXiv preprint arXiv:2605.27913, 2026. URL https://arxiv.org/abs/2605.27913.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In Proceedings of the Sixth International Conference on Learning Representations (ICLR), 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

Haoyu Peter Wang, Shikun Liu, Rongzhe Wei, and Pan Li. Generalization principles for inference over text-attributed graphs with large language models. In Proceedings of the 42nd International Conference on Machine Learning (ICML), volume 267 of Proceedings of Machine Learning Research, pp. 63755–63776. PMLR, 2025. URL https://proceedings.mlr.press/v267/wang25bq.html.

Jianing Wang, Junda Wu, Yupeng Hou, Yao Liu, Ming Gao, and Julian McAuley. InstructGraph: Boosting large language models via graph-centric instruction tuning and preference alignment. In Findings of the Association for Computational Linguistics: ACL 2024, pp. 13485–13510. Association for Computational Linguistics, 2024.

Ziqing Wang and Kaize Ding. HopRank: Self-supervised LLM preference-tuning on graphs for few-shot node classification. arXiv preprint arXiv:2604.17271, 2026.

Xixi Wu, Yifei Shen, Fangzhou Ge, Caihua Shan, Yizhu Jiao, Xiangguo Sun, and Hong Cheng. When do LLMs help with node classification? A comprehensive analysis. In Proceedings of the 42nd International Conference on Machine Learning (ICML), volume 267 of Proceedings of Machine Learning Research, pp. 67623–67649. PMLR, 2025.

Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. Provably robust DPO: Aligning language models with noisy feedback. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of Proceedings of Machine Learning Research. PMLR, 2024.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jiawei Chen, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. Towards robust alignment of language models: Distributionally robustifying direct preference optimization. In *International Conference on Learning Representations (ICLR)*, 2025. https://arxiv.org/abs/2407.07880.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In Proceedings of the Seventh International Conference on Learning Representations (ICLR), 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

Ruiyao Xu and Kaize Ding. GNN-as-Judge: Unleashing the Power of LLMs for Graph Few-shot Semi-supervised Learning with GNN Feedback. In Proceedings of the Fourteenth International Conference on Learning Representations (ICLR), 2026. URL https://iclr.cc/virtual/2026/poster/10007476.

Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In Proceedings of the Thirty-Third International Conference on Machine Learning (ICML), volume 48, pp. 40–48. PMLR, 2016.

Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. GNNExplainer: Generating explanations for graph neural networks. In Advances in Neural Information Processing Systems (NeurIPS), volume 32, pp. 9240–9251, 2019.

Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. In Advances in Neural Information Processing Systems (NeurIPS), volume 33, pp. 7793–7804, 2020.

Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. Adversarial attacks on neural networks for graph data. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 2847–2856. ACM, 2018. doi: 10.1145/3219819.3220078.

A Proof of Theorem 3.1 (Concentration of Stabilized Weighted Risk)

The result establishes concentration around the stabilized target $\mathcal{R}_\gamma$. Identification of the balanced or population risk requires additional overlap and sufficiency assumptions and is outside its scope; unobserved factors remain a source of bias.

Proof. **Conditioning convention.** Freeze the eligible pool U, proxy partition, predictor, and stabilized weights before sampling calibration nodes. The weights use all-node counts independent of calibration outcomes, while calibration labels enter only the loss/error statistic; re-estimating weights from those outcomes is outside the claim. Let $\mathcal{S}$ denote the post-merge strata (strata with < 3 nodes are merged by the same weighted-L_1 key distance). The stabilized target and estimator are

$$\tilde{w}_\gamma(s) = \frac{P_{\text{bal}}(s)}{\gamma_s P_{\text{obs}}(s) + (1-\gamma_s) P_{\text{bal}}(s)}, \mathcal{R}_\gamma = \mathbb{E}_{P_U}[\tilde{w}_\gamma(S)\ell], \qquad \hat{\mathcal{R}}_\gamma = \frac{1}{n}\sum_{v=1}^{n} \tilde{w}_\gamma(s_v)\ell_v. \tag{8}$$

Because $\tilde{P}_{\text{obs}}(s) \geq (1-\gamma_s)P_{\text{bal}}(s)$, $\tilde{w}_\gamma(s) \leq W_\gamma = 1/(1-\gamma_s) = 10$ at the selected $\gamma_s = 0.9$ (with $W_{\text{clip}} = \infty$). Smoothing trades exact balanced identification for reduced variance; the theorem concerns concentration around $\mathcal{R}_\gamma$.

Write $Z_v = \tilde{w}_\gamma(s_v)\ell_v \in [0, W_\gamma L_{\max}]$ for each fixed node in U. Uniform sampling without replacement gives $\mathbb{E}[\hat{\mathcal{R}}_\gamma] = |U|^{-1}\sum_{v\in U} Z_v = \mathcal{R}_\gamma$. Its exponential moment is bounded by that of independent uniform draws with replacement. Applying the bounded-variable moment bound and optimizing the Chernoff parameter yields

$$P(|\hat{\mathcal{R}}_\gamma - \mathcal{R}_\gamma| \geq \varepsilon) \leq 2\exp\left(-\frac{2n\varepsilon^2}{W_\gamma^2 L_{\max}^2}\right) \tag{9}$$

and hence

$$|\hat{\mathcal{R}}_\gamma - \mathcal{R}_\gamma| \leq W_\gamma L_{\max}\sqrt{\frac{\log(2/\delta)}{2n}} \tag{10}$$

with probability at least $1-\delta$. This is a fixed-predictor bound on the eligible-pool target; it does not identify whole-graph risk or cover subsequent model selection. The weights remain standard importance-weighting tools (Cortes et al., 2010). □

B Proof of Lemma 3.3 (Repair Coupling Identity)

The Repair Coupling Identity links observed Graph–LLM feedback to the harmful-rate decomposition on a unified action space.

Proof. **Case partition.** For each disagreement node v ($c_L \neq c_G$), let $C_v = \{c_L, c_G\}$, choose $\hat{d}_v \in \{L, G\}$ by calibrated confidence, and set $a_{\text{fb}}(v) = \texttt{KEEP}$ with $y^+ = c_{\hat{d}_v}$ and $y^- = c_{\{L,G\}\setminus\{\hat{d}_v\}}$.

The oracle action $a^*(v) \in \{\texttt{KEEP}, \texttt{FLIP}, \texttt{ABSTAIN}\}$ is

$$a^*(v) = \begin{cases} \texttt{KEEP}, & y_v^+ = y_v^* \\ \texttt{FLIP}, & y_v^+ \neq y_v^* \wedge y_v^- = y_v^* \\ \texttt{ABSTAIN}, & y_v^+ \neq y_v^* \wedge y_v^- \neq y_v^* \end{cases} \tag{11}$$

When $c_L = c_G$, no preference pair is generated, so consensus nodes are excluded from A_s. On the common action space $\mathcal{A} = \{\texttt{KEEP}, \texttt{FLIP}, \texttt{ABSTAIN}\}$, the natural nodewise coupling therefore gives

$$\|\hat{P}_{\text{fb},s} - \hat{P}_{\text{rep},s}\|_{\text{TV}} = \frac{1}{n_s^A} \sum_{v \in A_s} \mathbf{1}[a_{\text{fb}}(v) \neq a^*(v)] = \hat{\eta}_{\text{harm}}(s) = \hat{\eta}_{\text{rev}}(s) + \hat{\eta}_{\text{amb}}(s). \tag{12}$$

Here $\hat{\eta}_{\text{harm}}$ counts reversed and ambiguous harmful feedback. The Wilson upper endpoint u_s^{harm} is diagnostic only; for $n_s^A < 5$ we use shrinkage, and the KL training radius uses the empirical/shrunk point estimate rather than the Wilson endpoint. The structural instability $\bar{I}_s$ enters the diagnostic radius with the fixed $\beta^{\text{TV}} = 1$. □

C Correction Levels and KL-DRO Details

We distinguish regional allocation from pair selection before giving the KL objective.

Proposition C.1 (Non-interchangeable correction levels)**.** *For the frozen target $r_s^* = \phi_{\text{KL}}\widetilde{\eta}_{\text{harm}}(s) + (\phi_{\text{KL}}/2)\bar{I}_s$, two regions with unequal targets cannot both be matched by one fixed radius. This establishes allocation non-equivalence, not superior accuracy or KL coverage. The fixed-radius and factorial ablations evaluate performance consequences.*

The proposition concerns the distinct levels at which fixed/adaptive radii and outer/inner risk operators act. The TV quantity below is retained only as an in-sample repair-containment diagnostic.

Proof. **Non-interchangeable radii.** Let $r_s^* = g(\widetilde{\eta}_{\text{harm}}(s), \bar{I}_s)$ denote the frozen monotone target budget used by the adaptive construction, with $g(a, b) = \phi_{\text{KL}}a + (\phi_{\text{KL}}/2)b$. If $r_{s_1}^* < r_{s_2}^*$, no single fixed radius can match both targets: a value below $r_{s_2}^*$ under-allocates relative to the calibrated target for s_2, whereas a value above $r_{s_1}^*$ over-allocates relative to the target for s_1. This is an operator-allocation statement, not a KL coverage guarantee.

Non-interchangeable operators. The outer weights appear only as $\sum_s \omega_s R_s$ and change how much each stratum contributes. The inner KL adversary acts only on the loss atoms within a fixed s and cannot change the outer stratum prevalence. Therefore, neither operator can implement the other.

Diagnostic corollary (not the proposition's coverage claim). From Lemma 3.3, on the calibration-admitted set:

$$\|\widehat{P}_{\text{fb},s} - \widehat{P}_{\text{rep},s}\|_{\text{TV}} = \widehat{\eta}_{\text{harm}}(s) \leq u_s^{\text{harm}} \tag{13}$$

The adaptive radius is

$$r_s^{\text{TV}} = \alpha^{\text{TV}} \cdot u_s^{\text{harm}} + \beta^{\text{TV}} \cdot \bar{I}_s \tag{14}$$

With $\alpha^{\text{TV}} \geq 1$ and $\beta^{\text{TV}} \geq 0$, this contains every eligible calibration stratum by construction; it is an in-sample repair-containment diagnostic rather than a finite-sample KL coverage guarantee.

KL-DRO training. KL-DRO uses the same capped atoms $L_v = \min\{\ell_v, 50\}$ as the main objective, uniformly weighted by $1/|\mathcal{A}_s|$ in each nonempty admitted stratum:

$$R_s^{\text{KL}}(\theta) = \inf_{\lambda_s > 0} \left[\lambda_s r_s^{\text{KL}} + \lambda_s \log \frac{1}{|\mathcal{A}_s|} \sum_{v \in \mathcal{A}_s} e^{L_v/\lambda_s} \right]. \tag{15}$$

The outer aggregation uses stabilized stratum masses $\omega_s = a_s / \sum_{s'} a_{s'}$ with $a_s = \widehat{P}_{\text{obs}}(s)\tilde{w}_\gamma(s)$:

$$\min_\theta \sum_{s \in \mathcal{S}} \omega_s \, R_s^{\text{KL}}(\theta) + \lambda_{\text{reg}} \mathcal{L}_{\text{NLL}}. \tag{16}$$

The scalar KL dual is solved once per stratum (not per node), and the outer mass is $\omega_s \propto P_{\text{obs}}(s) w_\gamma(s)$. The selected $\gamma = 0.9$ gives a near-balanced correction with the smoothing bound $W_\gamma = 10$. □

Table 4: Stratum support, estimator usage, and partition sensitivity. Panel A reports post-merge strata, the number with at least five calibration nodes, and the fractions served by the global, shrinkage, and empirical estimators at the default setting. Panel B reports accuracy / support-≥ 20 worst-region accuracy for the default and boundary partition settings.

Dataset	Post-merge	$n_s^{\text{cal}} \geq 5$	Global	Shrinkage	Empirical
Cora	40	26	0.08	0.24	**0.69**
Citeseer	40	31	0.10	0.25	**0.65**
Pubmed	40	19	0.07	0.27	**0.66**
ogbn-arxiv	828	15	0.08	0.41	**0.50**
ogbn-products‡	760	19	0.09	0.38	**0.53**

Panel B: Default and boundary settings.

Dataset	thr.3/B=5	thr.20/B=5	thr.3/B=3	thr.3/B=8
Cora	**81.83 / 67.3**	81.32 / 65.4	81.62 / 66.1	81.65 / 66.7
Citeseer	**77.48 / 63.7**	76.84 / 62.2	77.18 / 62.8	77.31 / 63.2
Pubmed	**89.76 / 70.7**	89.22 / 68.8	89.42 / 69.8	89.57 / 70.1
ogbn-arxiv	**68.73 / 47.1**	67.74 / 44.8	68.21 / 45.8	68.48 / 46.4
ogbn-products‡	**85.11 / 58.2**	84.35 / 56.19	84.76 / 56.9	84.93 / 57.5

D Dataset and Strata Statistics

Table 4 summarizes the stratum support, estimator usage, and partition sensitivity. Cora, Citeseer, and Pubmed are citation TAGs (Sen et al., 2008; Yang et al., 2016); ogbn-arxiv and ogbn-products (Hu et al., 2020) test scale, while Cornell and Wisconsin (Pei et al., 2020) are retained in the heterophily stress test. The same frozen (R, T, H) partition and $W_\gamma = 10$ stabilization are used throughout.

Pair-weighted fractions mask stratum sparsity: 590 of 828 ogbn-arxiv strata use the global harm mean, leaving structural instability as their only radius-varying term. Wilson remains diagnostic only; the default boundary setting performs best empirically.

Strata shrinkage estimator (for $\hat{\eta}$ estimation with small admitted-pair counts):

$$\widetilde{\eta}_{\text{harm}}(s) = \frac{n_s^A}{n_s^A + \tau}\,\widehat{\eta}_{\text{harm}}(s) + \frac{\tau}{n_s^A + \tau}\,\overline{\eta}_{\text{harm,global}}, \qquad \tau = 5. \tag{17}$$

Here $\overline{\eta}_{\text{harm, global}}$ is computed over all admitted calibration pairs; when $n_s^A = 0$, the estimator equals this global mean. For $n_s^A \geq 5$, the KL estimator is the empirical point estimate and the Wilson upper endpoint is computed separately for the TV diagnostic; for $n_s^A \in [1, 4]$, shrinkage is applied (Eq. 17).

One-sided Wilson upper endpoint. For $n_s^A \geq 5$, let $\hat{\eta} = x/n$ and $z_{0.95} = 1.645$:

$$u_s^{\text{harm}} = \frac{\hat{\eta} + z_{0.95}^2/(2n) + z_{0.95}\sqrt{\hat{\eta}(1-\hat{\eta})/n + z_{0.95}^2/(4n^2)}}{1 + z_{0.95}^2/n}.$$

This is a pointwise descriptive endpoint under the stated independent-binomial working model; $\widetilde{\eta}_{\text{harm}}$, not u_s^{harm}, enters the KL training radius. On ogbn-arxiv, $\widetilde{\eta}_{\text{harm}}(s)$ rank ordering across strata correlates at $r > 0.72$ with validation-set rankings, preserving relative noise ordering for DRO radius assignment.

D.1 Candidate-Direction Space and TV/KL Diagnostics

The following candidate-direction propositions operate on the binary space $\{\text{correct}, \text{harmful}\}$. Lemma 3.3 separately operates on the three-action repair space $\{\texttt{KEEP}, \texttt{FLIP}, \texttt{ABSTAIN}\}$. For each node v, the candidate set $C_v = \{c_L, c_G\}$ is fixed before error measurement, where $c_L =$ LLM top-1 and $c_G =$ GNN top-1. The **actual chosen identity** $\hat{d}_v \in \{L, G\}$ is determined by higher calibrated confidence after temperature scaling (rather than always choosing the LLM). The candidate-direction space therefore has exactly two categories.

Proposition D.1 (TV on Candidate-Set Direction with Actual Orientation)**.** *On the binary direction simplex, let $\hat{P}^*_{s,\text{cal}}$ be the all-correct oracle distribution, so $P^*_{\text{harmful}} = 0$, and let $\hat{P}_{s,\text{fb}}$ be the*

observed direction distribution whose harmful mass is $\hat{\eta}_{\mathrm{harm}}(s)$. For Bernoulli distributions on $\{\mathrm{correct}, \mathrm{harmful}\}$,

$$\|p - q\|_{\mathrm{TV}} = \tfrac{1}{2}(|p_c - q_c| + |p_h - q_h|) = |p_h - q_h|.$$

Therefore $\|\hat{P}^*_{s,\mathrm{cal}} - \hat{P}_{s,\mathrm{fb}}\|_{\mathrm{TV}} = \hat{\eta}_{\mathrm{harm}}(s)$, with diameter 1. This candidate-direction TV is distinct from the three-action repair TV used in Lemma 3.3; under deterministic KEEP coupling, they share the same harmful mass but are defined on different sample spaces.

The proof of Lemma 3.3 gives the chosen orientation and repair cases. We now turn from this diagnostic space to the loss distribution used for training.

Proposition D.2 (KL-DRO over Per-Node Loss Distribution)**.** *For the KL-DRO problem over the per-node empirical loss distribution $\{\mathcal{L}_{pref}(v) : v \in \text{stratum } s\}$ with n_s^A nodes, the worst-case distribution reweights nodes via the scalar dual:* $\sup_q \mathbb{E}_q[\mathcal{L}] = \inf_{\lambda>0}[\lambda r_s^{KL} + \lambda \log \frac{1}{n_s^A} \sum_{i=1}^{n_s^A} \exp(\mathcal{L}_i/\lambda)]$. *The KL radius $r_s^{KL} = \phi_{KL} \cdot \widetilde{\eta}_{harm}(s) + (\phi_{KL}/2) \cdot \bar{I}_s$ uses the validation-selected ϕ_{KL} and the empirical/shrunk point estimate from Eq. (4). The TV diagnostic coefficients α^{TV}, β^{TV} are fixed by construction ($\alpha^{TV} = \beta^{TV} = 1$).*

TV containment diagnostic and KL training radius. The per-stratum TV radius $r_s^{\mathrm{TV}} = \alpha^{\mathrm{TV}} u_s^{\mathrm{harm}} + \beta^{\mathrm{TV}} \bar{I}_s$ is a repair-containment diagnostic only; training uses the KL radius $r_s^{\mathrm{KL}} = \phi_{\mathrm{KL}} \widetilde{\eta}_{\mathrm{harm}}(s) + (\phi_{\mathrm{KL}}/2)\bar{I}_s$ with the frozen empirical/shrunk point estimate. The outer risk uses stabilized masses $\omega_s = a_s / \sum_{s'} a_{s'}$, while the KL nominal distribution is uniform over admitted node-level losses.

The TV radius is constructed on the calibration-admitted set and never enters training. The current support-filtered fixed/adaptive comparison is reported in Table 14; the diagnostic and training objectives remain distinct.

E Experimental Setup Details

Label usage summary and protocol. The main experimental setup defines label roles, shared access, and the matched $kC+500$ adaptation budget. Co-teaching selects temperature and operating points on the shared pools; the five-stage pipeline freezes predictors, constructs strata, estimates harm and gates pairs, optimizes KL-DRO, and evaluates held-out risk.

Statistics. Published GNN-Judge comparisons use five-seed marginal means descriptively; because its seeds and splits remain unaligned, paired tests, effect sizes, and multiplicity-adjusted significance claims are omitted from the paper's tables and formal conclusions.

Table 5 reports marginal accuracy differences for the primary comparison (Ours vs. GNN-as-Judge, 3-shot accuracy). Because baseline splits and seeds are unaligned, these differences remain descriptive only.

Dataset	$(\Delta\mathrm{Acc}\ (\mathrm{pp}), \mathrm{source})$
Cora	$(+3.93, \text{published})$
Citeseer	$(+3.89, \text{published})$
Pubmed	$(+2.64, \text{published})$
ogbn-arxiv	$(+6.52, \text{published})$
ogbn-products	$(+4.09, \text{published})$

Table 5: Descriptive accuracy differences for the primary Ours–GNN-as-Judge comparison. Each tuple gives $(\Delta\mathrm{Acc}\ (\mathrm{pp}), \mathrm{source})$; published baseline means have unaligned splits and seeds, so paired inference is omitted.

All five differences remain descriptive because seed alignment and split equivalence are unavailable.

F Strict k-Shot Protocol and Calibration-Budget Curve

To make the label budget fully explicit and to test whether gains depend on access to 500 calibration labels, we report two protocols.

Table 6: Independent calibration-budget sweep at 3-shot under the matched protocol. "Added" is the number of calibration labels added to the fixed kC training budget; "Adaptation total" is the train-plus-calibration count $kC + \text{Added}$, excluding the shared validation pools. Values are mean±std over five independent sweep seeds; CI is the paired percentile-bootstrap 95% interval for Δ, computed from the same-seed Ours-minus-GNN-Judge differences within this sweep. The Added$= 500$ row is an independent-sweep value and is not the primary Table 1 checkpoint.

Dataset	Added	Adaptation total	Ours	GNN-Judge	Δ (pp)	CI (pp)	n
Cora	0	21	78.63±0.89	77.26±0.53	+1.37	[0.85,1.88]	5
Cora	10	31	79.03±0.89	76.87±0.71	+2.16	[1.38,2.97]	5
Cora	25	46	80.19±0.53	76.63±0.57	+3.55	[3.19,3.94]	5
Cora	50	71	79.59±0.70	77.05±0.65	+2.53	[1.73,3.39]	5
Cora	100	121	81.18±1.21	77.06±0.85	+4.12	[2.99,5.25]	5
Cora	250	271	81.98±0.54	77.98±0.46	+4.00	[3.67,4.28]	5
Cora	500	521	82.01±0.66	78.27±0.63	+3.73	[3.62,3.84]	5
ogbn-arxiv	0	120	64.82±1.15	60.26±1.19	+4.56	[3.06,6.07]	5
ogbn-arxiv	10	130	64.83±1.85	60.15±1.17	+4.69	[3.79,6.14]	5
ogbn-arxiv	25	145	65.81±0.61	62.17±1.23	+3.64	[2.73,4.54]	5
ogbn-arxiv	50	170	67.98±0.75	61.53±1.71	+6.45	[5.09,7.32]	5
ogbn-arxiv	100	220	68.75±1.12	63.94±0.82	+4.82	[4.04,5.60]	5
ogbn-arxiv	250	370	68.62±1.29	63.78±1.28	+4.83	[3.51,6.11]	5
ogbn-arxiv	500	620	71.35±0.83	63.85±1.25	+7.50	[5.95,9.05]	5

F.1 Strict k-shot (kC labels only)

Under the strict protocol, adaptation uses exactly kC labels and no separate 500-label calibration set. Shared $\mathcal{V}_{\text{val}}$ and $\mathcal{V}_{\text{hpval}}$ remain selection-only, so this is not strict total labeled access. Training-populated strata use leave-one-out estimates; empty strata use the label-free CF-proxy/shrinkage fallback. All comparison baselines receive the same strict access, isolating the correction levels from added calibration labels.

F.2 Calibration-budget curve

The independent calibration-budget sweep adds $\{0, 10, 25, 50, 100, 250, 500\}$ labels to the fixed kC training budget. Table 6 reports all seven points on Cora and ogbn-arxiv under identical adaptation counts and shared selection pools. Gains remain positive: $+1.37$–$+4.12$ pp on Cora and $+3.64$–$+7.50$ pp on ogbn-arxiv. At the strict endpoint, LOO-supported pairs account for 0.42 and 0.18, respectively; other points use calibration-supported empirical, shrinkage, or fallback estimates.

The seven points are independent budget-sweep runs; the primary Table 1 checkpoint at Added$= 500$ is shown separately below as a reference and is not used to define the sweep trend. The corresponding adaptation totals are Cora $21, 31, 46, 71, 121, 271, 521$ and ogbn-arxiv $120, 130, 145, 170, 220, 370, 620$. **Primary-checkpoint reference (not part of the independent sweep).** The primary Added$= 500$ checkpoint corresponds to Table 1. No paired intervals are reported because the published GNN-Judge means lack aligned seed records for this checkpoint; these descriptive differences are not pooled with the independent sweep.

Dataset	Added	Adaptation total	Ours	GNN-Judge	Δ (pp)	Inference
Cora	500	521	81.83±0.69	77.89±1.28	+3.94	Descriptive
ogbn-arxiv	500	620	68.73±0.91	62.21±1.45	+6.52	Descriptive

The sweep and primary checkpoint use independent splits, checkpoint selections, and seed alignments; they are not pooled. Sweep gains compare methods within matched runs, rather than reproducing the absolute values in Table 1. Their Added$= 500$ means differ by $+0.18$ pp on Cora and $+2.62$ pp on ogbn-arxiv, with Welch 95% intervals $[-0.80, 1.16]$ and $[1.35, 3.89]$ pp, reflecting split-and-selection variability as well as budget effects. Table 7 shows which estimators supply feedback without extra calibration labels.

Table 7: estimator coverage at the strict kC endpoint. Pair fractions are computed over admitted training preference pairs. LOO-supported means that the pair's stratum contains training labels and is estimated by leave-one-out cross-fitting; CF-proxy/global is used for training-empty strata. Added-label runs do not use these strict estimators: they use calibration-supported empirical estimates (with Wilson used only for the TV diagnostic), calibration shrinkage, or global fallback, whose fractions are exported by the code.

Dataset	Added	Adaptation total	LOO-supported	CF-proxy/global	Sum
Cora	0	21	0.42	0.58	1.00
ogbn-arxiv	0	120	0.18	0.82	1.00

The strict protocol uses both training-supported and proxy estimates; neither requires additional calibration labels.

G Full 3-Shot Results and Supplementary Robustness Evidence

Table 1 gives the complete comparison of mean accuracy; the supplementary robustness evidence below keeps the shared benchmark-validation access explicit. The five-dataset accuracy dispersions below use population standard deviations ($\mathrm{ddof} = 0$); paired-difference standard deviations use $\mathrm{ddof} = 1$ for $t = \bar{\Delta}\sqrt{5}/s_\Delta$ and $d_z = \bar{\Delta}/s_\Delta$. Across five seeds, Ours has standard deviations 0.49–0.91 pp, GNN-as-Judge 0.64–1.45 pp, and Co-teaching 0.69–1.14 pp on the five primary datasets. The ogbn-products Co-teaching mean is 83.29±1.10, with seed accuracies 81.8, 82.7, 83.0, 84.1, and 84.9. Descriptive GNN-Judge gaps and paired Co-teaching effects are reported in their respective comparison tables.

Table 8: paired Ours–Co-teaching comparison under the unified $kC+500$ protocol. Each row uses $n = 5$ aligned seeds and the same split realization for both methods. Δ is Ours minus Co-teaching in percentage points; CI is a paired percentile-bootstrap 95% interval. BH values are recomputed jointly from the five displayed raw p-values; confidence intervals and effect sizes retain the reported paired estimates.

Dataset	Δ mean	Δ std	95% CI	$t(4)$	Raw p	BH q	d_z
Cora	+1.45	0.61	[0.84,2.08]	5.34	0.0059	0.00875	2.39
Citeseer	+2.30	0.73	[1.49,3.05]	7.05	0.0021	0.00600	3.15
Pubmed	+1.20	0.79	[0.32,1.98]	3.37	0.0278	0.0278	1.51
ogbn-arxiv	+2.72	0.89	[1.91,3.51]	6.81	0.0024	0.00600	3.05
ogbn-products	+1.80	0.78	[0.9,2.7]	5.14	0.007	0.00875	2.3

Table 8 shows positive paired gains over Co-teaching on all five datasets, extending the comparison beyond the fixed GNN judge.

H Regional Robustness and Mechanism Evidence

The joint (R, T, H) partition captures graph-community, text-prior, and local-regularity structure while keeping the number of proxy factors fixed under the 500-label calibration budget. Table 9 tests robustness to injected noise, heterophily, transfer, and backbone changes; Table 10 then examines local reliability.

Under controlled noise injection, DPO/ORPO/SimPO decline by 22.6–23.1 pp (41.7–43.4% relative) from $\lambda = 0$ to $\lambda = 1.5$. Ours declines by 18.6 pp (27.0% relative).

Worst-class accuracy and macro-F1 follow the same ordering as overall and support-filtered worst-region accuracy.

On heterophilic graphs, gains are 4.1–5.1 pp. Low-homophily regions are identified as high-noise strata and receive larger robustness budgets. H-ELSD contributes 2.3 pp on Cornell.

Few-shot transfer gains (3.5–5.7 pp) hold in the evaluated source-target settings.

Few-shot transfer protocol. Target embeddings, predictions, and (R, T, H) profiles are recomputed, each target stratum is mapped to its nearest source-stratum centroid, and target labels do not re-estimate the transferred DRO parameters.

Performance is stable across GCN, GAT, and GraphSAGE (±1.5 pp). Importance reweighting operates on the feedback signal for any GNN architecture.

H.1 Worst-Region Robustness Under a Minimum Test-Support Threshold

To avoid conclusions driven by tiny groups, this analysis uses test support ≥ 20 and reports eligible-group count, worst-group accuracy, P10 accuracy, group-size-weighted $\mathrm{CVaR}_{0.1}$ loss, and worst-group support. Panel A reports marginal point estimates and support diagnostics; Panel B reports descriptive marginal gains without paired uncertainty for comparison. When raw per-node outcomes

Table 9: Regional robustness and mechanism evidence. Panel A reports the clean and highest-noise endpoints of the synthetic experiment; intermediate points preserve the same ordering. Panel B reports the retained heterophily and transfer comparisons. Panel C gives the observed backbone ranges.

Panel A: Synthetic region-dependent noise on ogbn-arxiv.

Method	$\lambda = 0$	$\lambda = 1.5$
DPO	53.29	30.18
ORPO	54.11	31.52
SimPO	53.67	30.94
GNN-as-Judge	62.21	39.85
CANE	57.83	44.37
Ours	**68.73**	**50.14**

Panel B: Heterophily and transfer.

	Heterophily		**Transfer from ogbn-arxiv**		
Method	**Cornell**	**Wisconsin**	**→ Cora**	**→ Citeseer**	**→ Pubmed**
GNN-as-Judge	$84.65_{\pm 1.23}$	$76.95_{\pm 2.03}$	$68.27_{\pm 0.91}$	$56.67_{\pm 0.89}$	$83.41_{\pm 0.76}$
Ours	$\mathbf{88.71}_{\pm 0.73}$	$\mathbf{82.04}_{\pm 0.68}$	$\mathbf{73.15}_{\pm 0.64}$	$\mathbf{62.38}_{\pm 0.71}$	$\mathbf{86.92}_{\pm 0.53}$

Panel C: Observed range across backbones.

Backbone sweep	**Cora**	**Citeseer**	**ogbn-arxiv**
GCN/GAT/GraphSAGE	81.67–82.41	77.48–78.12	67.25–68.73

are supplied, each worst-group bootstrap replicate resamples seed records, resamples nodes within every eligible stratum, and reselects the minimum eligible stratum; this includes worst-stratum selection uncertainty. Legacy grouped records use the explicitly labeled stratum-count sufficient-statistic fallback and are not described as raw-node intervals. P10 and CVaR use seed-level mean bootstrap intervals. We additionally report the descriptive Ours−GNN-Judge worst-region gap without paired seed inference here. The descriptive gain is computed from reported marginal means and does not imply aligned seeds, common splits, or paired statistical uncertainty. Across the five primary datasets, the support-filtered worst-region gain is 13.8–20.4 pp. The descriptive gaps remain positive across all five datasets, under the same eligibility rule for reported groups.

I Stratum-Radius Calibration and Sensitivity Analyses

Table 11 gives the macro and size-weighted Spearman diagnostics underlying the radius-calibration panel. The radius is frozen before post-hoc evaluation on $\mathcal{V}_{\text{test}}$; test labels never enter construction, selection, pair admission, or training. Macro Spearman is 0.66–0.68, size-weighted Spearman is 0.67–0.68, and the bootstrap 95% intervals span 0.59–0.73. Pearson $r = 0.71$–0.78 for adaptive KL versus 0.22–0.31 for LLM confidence is a complementary post-hoc diagnostic, not a training guarantee.

The positive rank correlations support the budget ordering without certifying individual predictions. Both single-component radii have lower five-seed means. Median admitted support is sparse on the OGB graphs, while positive conditional correlations show that the radius still tracks residual regional risk after structural instability is fixed.

Prompt and CF controls. On Cora (3-shot), role, exemplar, and output-format variations change accuracy by at most ± 0.8 pp for Ours and ± 1.4 pp for GNN-as-Judge. The held-out CF table below uses 10 construction and 10 validation trials from the same ELSD/H-ELSD family with disjoint seeds; its positive class is a safe pair, the gate retains scores at or above τ_{cf}, and coverage is evaluated on held-out pairs.

The label-held-out panel is the primary pair-validity evidence and supplies the compact precision–coverage summary in Table 3. The trial-disjoint panel tests structural sensitivity separately; neither diagnostic establishes correctness for every admitted pair.

Tables 13 and 14 test sensitivity to selection access and component combinations.

The method ordering persists across validation-pool sizes. The full factorial performs best on both datasets, while its individual interventions show which controls matter.

Table 10: Support-filtered worst-region evidence ($k = 3$). Panel A reports marginal point estimates and support diagnostics: accuracy, eligible groups, support, P10, and CVaR after restricting every stratum to at least 20 test nodes. Panel B reports descriptive Ours−GNN-Judge gaps from published marginal means without paired inference. Accuracy is in percentage points and CVaR is a 0–1 loss. Raw-node bootstrap uses replicate-level minimum re-selection; grouped legacy records use the labeled count-level fallback.

Panel A: Worst-region accuracy and support.

Dataset	Method	Eligible	Worst	Worst n	P10	CVaR$_{0.1}$
Cora	GNN-Judge	40	52.7	65	64.0	0.43
Cora	Ours	40	67.3	71	78.0	0.29
Citeseer	GNN-Judge	40	47.5	77	60.4	0.47
Citeseer	Ours	40	63.7	77	74.7	0.31
Pubmed	GNN-Judge	40	57.0	73	70.2	0.38
Pubmed	Ours	40	70.7	64	80.9	0.24
ogbn-arxiv	GNN-Judge	70	26.7	66	40.0	0.63
ogbn-arxiv	Ours	70	47.1	97	59.0	0.49
ogbn-products	GNN-Judge	70	43.9	77	56.2	0.51
ogbn-products	Ours	70	58.2	62	72.7	0.35

Detailed marginal intervals are exported in the source-labeled `worst_support.csv`.

Panel B: Descriptive worst-region gap.

Dataset	Gain	Baseline	Inference
Cora	14.5	Published	Descriptive
Citeseer	16.2	Published	Descriptive
Pubmed	13.8	Published	Descriptive
ogbn-arxiv	20.4	Published	Descriptive
ogbn-products	14.3	Published	Descriptive

The displayed gap subtracts published baseline means from reported Ours means; it is descriptive and does not represent a paired confidence interval.

Table 11: Radius calibration, component ablations, and support diagnostics. Spearman values are mean±std with bootstrap 95% CI; P lists Pearson r for adaptive KL/TV/χ^2/LLM confidence. Harm and Inst. use only the harmful-rate or structural-instability radius term; conditional r fixes structural instability.

Panel A: Radius–error correlation.

Dataset	Macro Spearman	Weighted Spearman	P tuple
Cora	0.66±.07 [.59,.71]	0.68±.05 [.64,.73]	.74/.72/.71/.28
Citeseer	0.67±.06 [.60,.72]	0.68±.05 [.62,.73]	.71/.69/.68/.25
Pubmed	0.67±.05 [.61,.72]	0.68±.05 [.63,.73]	.78/.76/.75/.31
ogbn-arxiv	0.68±.05 [.64,.73]	0.68±.06 [.64,.73]	.76/.73/.72/.22
ogbn-products	0.66±.06 [.59,.71]	0.67±.06 [.61,.73]	.73/.71/.70/.27

Panel B: Component accuracy (%) and support diagnostics.

Dataset	Harm	Inst.	Radius median [IQR]	Pairs median [IQR]	Conditional r
Cora	$80.40_{\pm 0.81}$	$79.80_{\pm 0.86}$	.34 [.24,.46]	6 [3,11]	.61
Citeseer	$76.10_{\pm 0.95}$	$75.60_{\pm 0.99}$	.36 [.25,.48]	6 [3,10]	.58
Pubmed	$88.50_{\pm 0.72}$	$88.10_{\pm 0.76}$	.30 [.21,.42]	8 [4,14]	.65
ogbn-arxiv	$64.20_{\pm 1.14}$	$67.20_{\pm 1.08}$	.41 [.33,.52]	1 [1,2]	.31
ogbn-products	$82.90_{\pm 0.93}$	$83.80_{\pm 0.89}$	.39 [.30,.50]	1 [1,3]	.39

Co-teaching tuning fairness. Co-teaching uses a method-specific temperature/operating-point search on the shared validation pools; strict-kC disables these adaptations.

Table 12: Held-out counterfactual validation across five datasets. Panel A reports trial-disjoint structural validation using disjoint construction and validation trials from the same ELSD/H-ELSD family. Panel B reports label-held-out preference-pair quality on test nodes.

Panel A: Trial-disjoint structural validation.

Dataset	K_{con}	K_{val}	Spearman	95% CI
Cora	10	10	0.62	[0.47, 0.77]
Citeseer	10	10	0.59	[0.42, 0.76]
Pubmed	10	10	0.68	[0.55, 0.81]
ogbn-arxiv	10	10	0.63	[0.47, 0.79]
ogbn-products	10	10	0.61	[0.44, 0.78]

Panel B: Label-held-out preference-pair quality.

Dataset	AUROC	AUPRC	Safe precision	Safe recall	Coverage
Cora	0.84	0.81	0.77	0.66	0.48
Citeseer	0.81	0.77	0.74	0.60	0.46
Pubmed	0.88	0.85	0.84	0.72	0.53
ogbn-arxiv	0.83	0.79	0.77	0.64	0.47
ogbn-products	0.82	0.77	0.74	0.61	0.42

Metric definitions. Spearman evaluates the relation between construction-trial $\mathrm{CF}_{\mathrm{dep}}$ and validation-trial prediction shifts. AUROC and AUPRC use the continuous $\mathrm{CF}_{\mathrm{dep}}$ score. Safe precision, safe recall, and coverage use the fixed gate $\tau_{\mathrm{cf}} = 0.10$. The positive class is a safe pair whose chosen label is correct, and the rejected label is incorrect. Higher $\mathrm{CF}_{\mathrm{dep}}$ predicts a safe pair; the gate retains scores at or above τ_{cf} and filters lower scores.

Table 13: low-validation-budget sensitivity on Cora and ogbn-arxiv (3-shot, five seeds). The train/calibration budget, backbone, and search rule are fixed while $|\mathcal{V}_{\mathrm{val}}|/|\mathcal{V}_{\mathrm{hpval}}|$ is reduced. Values are accuracy mean±std in percentage points; stability is the modal fraction over $d = 12$ selected components, and ogbn-arxiv full/full denotes $29{,}799/500$.

Dataset	$V_{\mathrm{val}}/V_{\mathrm{hpval}}$	GNN	Co-teach	Ours	Stability
Cora	kC/kC	76.8±1.6	79.0±1.2	**80.8±1.0**	60%
Cora	100/100	77.4±1.4	79.7±1.1	**81.5±0.7**	74%
Cora	500/500	77.89±1.28	80.37±0.69	**81.83±0.69**	88%
ogbn-arxiv	kC/kC	60.7±1.8	63.9±1.5	**66.5±1.5**	60%
ogbn-arxiv	100/100	61.5±1.6	65.2±1.3	**67.8±1.2**	74%
ogbn-arxiv	500/500	62.21±1.45	66.01±1.14	**68.73±0.91**	88%
ogbn-arxiv	full/full	62.4±1.3	66.2±1.0	**68.9±0.9**	90%

Table 14: Complete factorial ablation under the primary protocol. Bits toggle weighting, gating, and adaptive radii; C3=0 retains KL-DRO at $r_s = 0.5$. Entries are accuracy / support-≥ 20 worst-region accuracy.

Bits	Cora	ogbn-arxiv
factorial_111 (all)	**81.83 / 67.3**	**68.73 / 47.1**
factorial_011 (no reweight)	75.41 / 52.1	60.14 / 34.4
factorial_101 (no CF gate)	78.56 / 61.4	65.27 / 42.9
factorial_110 (fixed radius)	77.89 / 60.1	63.82 / 40.5
factorial_001 (adaptive only)	70.84 / 45.7	52.61 / 30.1
factorial_010 (gate)	73.92 / 50.3	56.84 / 33.8
factorial_100 (weights)	76.18 / 53.8	60.91 / 35.9
factorial_000 (fixed r)	68.95 / 42.6	50.42 / 28.4

J Full Algorithm

Algorithm 1 gives a compact executable summary of the pipeline described in §3. **Algorithm 1** Region-Conditional Graph Feedback Reweighting

Input: TAG $\mathcal{G}$, $\mathcal{V}_{\text{train}}$, calibration set, LLM $\mathcal{M}$, GNN f_ϕ; ϕ_{KL}, λ_{reg}, τ_{cf}, K_R, B, K

Output: Fine-tuned $\mathcal{M}_\theta$

Five-stage pipeline: freeze f_ϕ predictions and LLM outputs; construct and merge (R, T, H) strata; generate shared whole-graph CF trials and retain disagreement pairs with $\mathrm{CF}_{\mathrm{dep}} \geq \tau_{\mathrm{cf}}$; estimate admitted regional harm by empirical/shrunk/global estimators; freeze the radius and optimize the outer-weighted KL-DRO objective. Strict-kC uses LOO/CF-proxy estimates, while the primary protocol uses calibration-supported estimates.

Radius and objective: $r_s^{\mathrm{KL}} = \phi_{\mathrm{KL}}\widetilde{\eta}_{\mathrm{harm}}(s) + (\phi_{\mathrm{KL}}/2)\bar{I}_s$ and $r_s^{\mathrm{TV}} = \alpha^{\mathrm{TV}} u_s^{\mathrm{harm}} + \beta^{\mathrm{TV}} \bar{I}_s$ (diagnostic only); $\min_\theta \sum_s \omega_s \sup_{q_s : D_{\mathrm{KL}}(q_s \| \widehat{P}_{s,\mathrm{loss}}) \leq r_s^{\mathrm{KL}}} \mathbb{E}_{q_s}[\mathcal{L}_{\mathrm{pref}}] + \lambda_{\mathrm{reg}} \mathcal{L}_{\mathrm{sft}}$, with $\omega_s \propto \widehat{P}_{\mathrm{obs}}(s)\tilde{w}_\gamma(s)$.

Return $\mathcal{M}_\theta$.

The algorithm freezes the correction before optimization, preserving the separation between calibration and training.

K IMPLEMENTATION DETAILS

Algorithm 1 uses minibatch KL risks with frozen outer weights and no within-batch renormalization. Absent regions supply no gradient; detached moving averages are bookkeeping constants. KL loss atoms are capped at 50.

Non-default implementation choices. We use $K = 20$ whole-graph degree-preserving double-edge-swap trials; each accepted swap preserves node degrees while changing neighbor identities, with rewiring strength 0.25. H-ELSD is triggered when $\bar{H} < 0.3$ and constrains the class-transition matrix by $\|\mathbf{T}^{cf} - \mathbf{T}^{orig}\|_1 \leq \tau_T$ with $\tau_T = 0.05$. CF scores are evaluated on the shared set of K graphs.

Pair and stratum details. On disagreement, the top-1 label with higher calibrated confidence is y^+; the other is y^-. We retain a pair iff $\mathrm{CF}_{\mathrm{dep}} \geq \tau_{\mathrm{cf}}$ and call it harmful iff $y^+ \neq y^*$, including reversed and ambiguous feedback; consensus nodes enter neither training loss. Confidence ties within 0.01 are discarded. Prediction scores and DPO both sum response-token log-probabilities; the auxiliary NLL instead uses a token mean followed by an example mean. Sparse strata use exact weighted-L_1 matching on (R, T, H) keys, with region differences normalized by B_R and bin differences by $1/B$; this is the same empirical merge used for the reported values.

K.1 TUNING AUDIT AND HYPERPARAMETER SEARCH

Tuning audit. Ours executes 48 candidate configurations; Co-teaching uses sequential rather than joint search: six temperatures per model are evaluated on calibration, followed by 20 threshold/fraction operating points on HP-validation with temperatures frozen. The audit records one final full-optimizer run per seed (five total) and these 2 temperature fits plus 20 operating-point evaluations. The other baselines use fixed or official configurations. The primary configuration uses Llama-3-8B-Instruct and a two-layer GCN (64 hidden dimensions, 200 epochs); Qwen2.5-7B-Instruct is the separate cross-family replication. The selected values are $K_R = 10/5/50$ for citation/Cornell–Wisconsin/OGB, $B = 5$, minimum stratum size $= 3$, $K = 20$, $\tau_{\mathrm{cf}} = 0.10$, $\gamma_s = 0.9$, $\phi_{\mathrm{KL}} = 0.5$, $W_{\mathrm{clip}} = \infty$, $\alpha^{\mathrm{TV}} = \beta^{\mathrm{TV}} = 1$, $\beta_{\mathrm{DPO}} = 0.1$, $\lambda_{\mathrm{reg}} = 0.001$ (`lambda_sft`$= 0.1$ times `sft_sequence_scale`$= 0.01$), LoRA rank $r = 8$ and scale $\alpha = 16$, learning rate 3e-5, 3 epochs, and batch size 8; candidate grids remain in the anonymous configuration files.