JUDGEBLIND: CROSS-ENCODER FEEDBACK ALIASING IN GNN-TO-LLM PIPELINES

Anonymous authors
Paper under double-blind review

ABSTRACT

GNN-to-LLM pipelines expose a cross-encoder gap: a GNN judge approves graph-dependent feedback, while an LLM student learns from raw tokens through a separate encoder. We show that graph–text edits can preserve the judge's complete non-textual decision transcript—exactly for discrete fields and within fixed componentwise tolerances for continuous fields—while harming student learning. The problem is whether unchanged approval decisions imply unchanged training effects. JudgeBlind studies this boundary through a shared attack–diagnosis–defense mechanism. FC-CEFA constructs transcript-invariant paths with graph-side curvature correction; MC-CMCT diagnoses correctability and surviving student exposure; CT-CAFT suppresses that exposure over continuation tubes under IT/PT coverage constraints. Rank and Taylor analysis characterize exposed directions and second-order closure, yielding a cubic relaxed-state judge residual. A conditional finite-round bound separates probe mismatch from student-update approximation error. On five TAG benchmarks and two victims over five seeds, FC-CEFA attains pre-cap gate-pass rates of 73.8–90.0%. Against the matched first-order control on Cora, it raises feasibility from 51.4% to 90.0% and budget-normalized invariant feedback harm from 0.24 to 0.53. MC-CMCT predicts corrector convergence with AUROC 0.87–0.92. CT-CAFT reduces adaptive harm by 52.2–60.0%; on Cora, it improves over the margin control by 16.7 percentage points while halving clean-accuracy loss from 1.4 to 0.7 points. These results distinguish approval stability from learning stability in heterogeneous feedback pipelines.

1 INTRODUCTION

Text-attributed graphs (TAGs) encode both topological relationships and natural-language node attributes. A graph neural network (GNN) serves as a structural selector, certifying instruction-tuning and preference-tuning feedback for a large language model (LLM) student in pipelines such as GNN-as-Judge (Xu and Ding, 2026) and LLM–GNN Co-Teaching (Peng et al., 2026). These systems create a trust boundary: the GNN judge certifies supervision signals, but the downstream LLM student interprets approved text through a distinct tokenizer and encoder that need not share the GNN's representational assumptions.

Approval is only one step in a feedback pipeline: the selected text must train the student. Separate encoders can disagree on the effect of an edit even when every approval field stays fixed. JudgeBlind isolates this gap by measuring harm under the complete transcript constraint.

Existing attacks manipulate judge outputs, graph predictions, or downstream student gradients (Tong et al., 2025; Ding et al., 2026; Chaudhari et al., 2025; Ma et al., 2026b;a; Luo et al., 2026; Zhang et al., 2026b; Chen et al., 2026; Geiping et al., 2021; Duan et al., 2025; Cloud et al., 2026). Output-preserving geometry instead moves along model equivalence classes (Tron et al., 2024; Benfenati and Marta, 2021; Chen et al., 2024). Our setting jointly fixes every discrete field and bounds every continuous field of a graph-dependent judge transcript while targeting a separately encoded student's update. This fieldwise constraint distinguishes the problem from preserving aggregate teacher utility or selected outputs.

We pose a narrower but structurally harder question:

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.

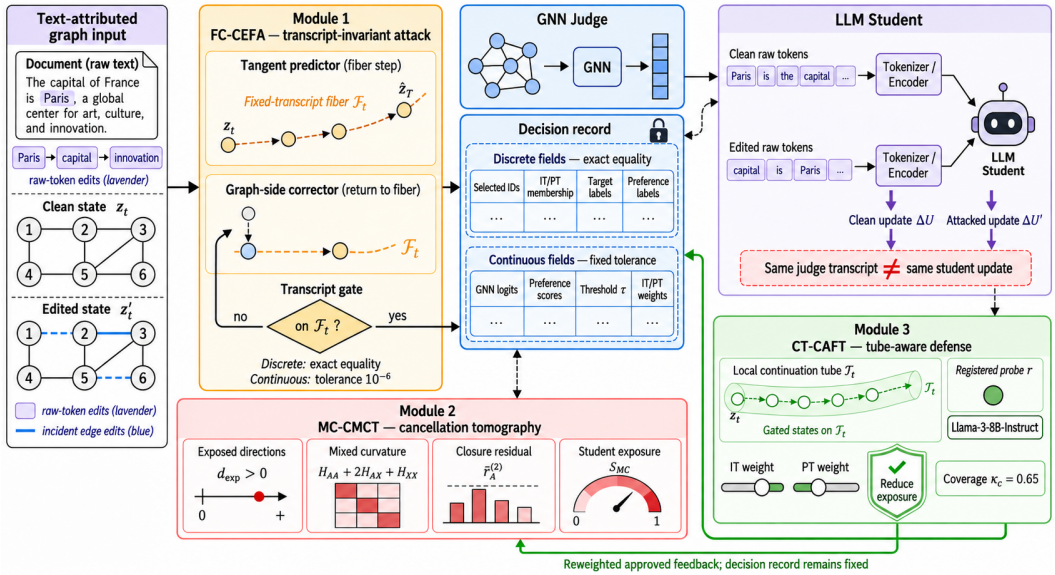


Figure 1: **Schematic JudgeBlind attack–diagnosis–defense loop.** FC-CEFA modifies raw node tokens and graph edges while requiring exact equality of $R_t^{\text{disc}} = (\mathcal{V}_{\text{selected},t}, \mathcal{V}_{\text{agreed},t}, \mathcal{V}'_{\text{disagreed},t}, \hat{\mathbf{y}}_t^G, \hat{\mathbf{y}}_t^L)$ and componentwise tolerance of $\mathbf{c}_t = (\boldsymbol{\ell}_t^G, \mathbf{S}_{\text{pref},t}, \tau_t, \mathbf{s}_t^{\text{IT}}, \mathbf{s}_t^{\text{PT}})$. MC-CMCT diagnoses mixed-curvature correctability and surviving student exposure; CT-CAFT suppresses the same path family over a local continuation tube and a registered student probe (Llama-3-8B-Instruct).

Core question. *Can an attacker preserve every discrete field of a GNN judge's decision transcript and keep every continuous field within tolerance fixed before test-set evaluation, while changing the LLM student's training gradient through raw-token edits interpreted by a separate encoder?*

Figure 1 links three operations on transcript-preserving directions: construction, diagnosis, and defense. The contribution connects complete approval constraints to downstream learning: graph edits cancel transcript curvature, while corrected student exposure guides diagnosis and feedback defense:

1. **FC-CEFA (Fiber-Continuation Cross-Encoder Feedback Aliasing).** Constructs harmful paths under a complete transcript gate; graph-side second-order correction improves feasibility beyond matched first-order constraints.
2. **MC-CMCT (Mixed-Curvature Cross-Modal Cancellation Tomography).** Diagnoses text, graph, and mixed-curvature blindness through an exposed-subspace certificate, closure tests, and corrected student exposure.
3. **CT-CAFT (Curvature-Tube Cancellation-Aware Feedback Training).** Reweights IT/PT feedback over continuation tubes with per-class coverage minima and provides an attack-specific structured defense with an explicit probe-gap term.

2 Related Work and Novelty Boundary

GNN-to-LLM Feedback. GNN-as-Judge (Xu and Ding, 2026), LLM–GNN Co-Teaching (Peng et al., 2026), and tool-mediated agents (Wang and Vemuri, 2026) study how graph judgments produce or exchange feedback, but not adversarial equivalence between a GNN decision record and a separately encoded student's update.

Judge Attacks and Downstream Propagation. Existing work studies output manipulation and rubric attacks (Tong et al., 2025; Ding et al., 2026), propagation and curator bias (Zhang et al., 2026a; Chaudhari et al., 2025), or bias directions in judge states (Xu et al., 2026a;b). FC-CEFA instead fixes the entire judge decision record while changing the downstream update.

Equivalence, Fiber, and Curvature Methods. Output-preserving manifolds and equivalence classes (Tron et al., 2024; Benfenati and Marta, 2021; Chen et al., 2024), null-space training (Hu et al., 2024), and curvature-based robustness (Singla and Feizi, 2020; Entesari et al., 2024; Bose et al., 2025; Sosnin et al., 2024) provide related geometric tools. Distributionally robust objectives (Namkoong and Duchi, 2016; Sagawa et al., 2020) are structurally related to CT-CAFT, but neither family imposes the complete transcript boundary or its probe-gap decomposition.

Graph and Clean-Label Poisoning. GRAPHTEXTACK (Ma et al., 2026a), BadGraph (Chen et al., 2026), and related TAG studies (Ma et al., 2026b; Luo et al., 2026) disrupt predictions, while clean-label transfer methods (Geiping et al., 2021; Draganov et al., 2026) steer gradients. Antidistillation Sampling (Savani et al., 2025) preserves teacher-side utility while harming student learning through teacher-controlled generation and likelihood constraints; our attacker-controlled graph–text edits instead preserve every discrete field exactly and every continuous field within fixed tolerances. VARMAT (Zhang et al., 2025) regularizes modality vulnerability; CleanCLIP (Bansal et al., 2023) weakens poisoned cross-modal associations. CT-CAFT instead targets transcript-invariant student exposure with explicit IT/PT coverage.

These boundaries motivate a threat model in which the attacker is powerful, but the judge-facing evidence is fixed field by field.

3 Threat Model and Problem Formulation

3.1 Threat Model

We fix a white-box attacker. Before graph processing at round t, the attacker may:

- Make at most B_X token-level substitutions, insertions, or deletions in total across the target nodes, subject to BERTScore ≥ 0.95 and perplexity ratio $\in [0.8, 1.2]$;
- Toggle at most B_A unordered edges incident to target nodes. We write $\|\Delta_A\|_{0,E}$ for the number of unordered additions/deletions, counting each undirected edge exactly once. Homophily maintenance requires $|\Delta h| \leq 0.05$, degree distribution KS-test $p > 0.1$, and no new disconnected components;
- Access the GNN model, judge encoder, selector parameters, current student probe (Llama-3-8B-Instruct), and feedback weights;
- Not modify ground-truth labels, train/validation splits, selector source code, or the total budget.

A gray-box extension replaces unknown tokenizers or checkpoints with a registered probe r (Llama-3-8B-Instruct); the probe gap δ_{probe} from Theorem 5.3 captures the approximation error when the attacker's probe differs from the true student encoder. The probe gap is zero when the registered probe and the true student are the same model (white-box). We do not evaluate black-box transfer. All baselines share the identical budget and observability thresholds.

3.2 Full Judge-Decision Record

To define unchanged approval, write the GNN-as-Judge feedback at round t as:

$$\mathcal{D}_t^{\text{IT}} = \{(x_i, \hat{y}_i^G, s_{t,i}^{\text{IT}}) : i \in \mathcal{V}_{\text{agreed},t}\}, \tag{1}$$

$$\mathcal{D}_t^{\text{PT}} = \{(x_i, \hat{y}_i^G, \hat{y}_i^L, s_{t,i}^{\text{PT}}) : i \in \mathcal{V}'_{\text{disagreed},t}\}. \tag{2}$$

The **discrete decision record** at round t consists of the GNN selector outputs expressible as integers or identifiers:

$$R_t^{\text{disc}} = (\mathcal{V}_{\text{selected},t}, \mathcal{V}_{\text{agreed},t}, \mathcal{V}'_{\text{disagreed},t}, \hat{\mathbf{y}}_t^G, \hat{\mathbf{y}}_t^L). \tag{3}$$

The **continuous decision record** collects the remaining real-valued judge quantities:

$$\mathbf{c}_t = (\boldsymbol{\ell}_t^G, \mathbf{S}_{\text{pref},t}, \tau_t, \mathbf{s}_t^{\text{IT}}, \mathbf{s}_t^{\text{PT}}). \tag{4}$$

Throughout, $\mathbf{s}_t = [\mathbf{s}_t^{\mathrm{IT}}; \mathbf{s}_t^{\mathrm{PT}}]$ denotes the judge-produced reference weights included in the verified non-textual transcript; the shared tolerance ϵ_w is applied separately to both blocks. After transcript verification, CT-CAFT introduces separate defensive training weights $\boldsymbol{\alpha}_t = [\boldsymbol{\alpha}_t^{\mathrm{IT}}; \boldsymbol{\alpha}_t^{\mathrm{PT}}]$ and regularizes them toward $\mathbf{s}_t$. The transcript gate is evaluated on $\mathbf{s}_t$; CT-CAFT optimizes $\boldsymbol{\alpha}_t$. The full judge-facing observable is $o_t = (R_t^{\mathrm{disc}}, \mathbf{c}_t)$. o_t contains every judge-produced non-textual decision field used for selection, labeling, preference construction, thresholding, and weighting. The raw text x_i is an input payload passed to both encoders, not a judge-produced decision field; it is excluded from o_t. JudgeBlind claims invariance of the complete non-textual decision transcript, not byte-identical feedback tuples. On the fixed discrete cell defined by $R_t^{\mathrm{disc}} = R_t^{\mathrm{disc},0}$, define the **continuous transcript-residual map**:

$$\mathbf{F}_t(z) = \mathbf{c}_t(z) - \mathbf{c}_t(z_t^0). \tag{5}$$

3.3 Exact and Tolerant Invariance

Definition 3.1 (Transcript Invariance)**.** *Let z denote the graph–text state and $R_t^{disc,0}, \mathbf{c}_t^0$ the clean records. The **transcript-invariant fiber** at round t is:*

$$\mathcal{F}_t^{\boldsymbol{\epsilon}}(o_t^0) = \{z : R_t^{disc}(z) = R_t^{disc,0},\ |F_{t,j}(z)| \le \epsilon_j\ \forall j\}, \tag{6}$$

where $\boldsymbol{\epsilon} = (\epsilon_\ell, \epsilon_{pref}, \epsilon_\tau, \epsilon_w)$ are componentwise tolerances fixed before test-set evaluation. For all experiments, $\epsilon_\ell = \epsilon_{\mathrm{pref}} = \epsilon_\tau = \epsilon_w = 10^{-6}$. These tolerances bound numerical drift of continuous transcript fields only; by themselves they do not certify unchanged labels, rankings, memberships, or selected IDs. Exact equality of all selected IDs, IT/PT memberships, target labels, and preference labels is enforced independently by R_t^{disc}. The IT and PT judge-produced reference weights satisfy $\mathbf{s}_t^{\mathrm{IT}}, \mathbf{s}_t^{\mathrm{PT}} \in [0,1]$ elementwise and are not unit-sum normalized; the two channels share ϵ_w but are checked independently.

Together, exact discrete equality and componentwise continuous tolerances define the gate at every retained step and round. We now measure harm among edits passing this gate.

3.4 Adversary Budget and Invariant Feedback Harm

Let ϕ_T denote the student parameters after T rounds of GNN-judge-guided fine-tuning; here $T = T_{\mathrm{GAJ}} = 1$ for the GNN-as-Judge victim and $T = T_{\mathrm{CT}} = 20$ for the LLM–GNN Co-Teaching victim. Let $\mathcal{D}_{\mathrm{adv}}$ be an attack-development split drawn only from validation nodes, disjoint from the final target evaluation set $\mathcal{V}_{\mathrm{tar}} \subseteq \mathcal{D}_{\mathrm{test}}$. The attack objective is

$$\mathcal{L}_{\mathrm{adv}}(\phi_T) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathrm{adv}}}\big[\ell(\phi_T(x), y)\big]. \tag{7}$$

The adversary maximizes $\mathcal{L}_{\mathrm{adv}}$ using only $\mathcal{D}_{\mathrm{adv}}$. After all attack paths, hyperparameters, and stopping decisions are fixed, Harm and IFH are evaluated once on the untouched $\mathcal{V}_{\mathrm{tar}}$; no label from $\mathcal{V}_{\mathrm{tar}}$ is used for attack construction or model selection. The adversary solves $\max_{\Delta_A,\Delta_X} \mathcal{L}_{\mathrm{adv}}(\phi_T)$ subject to:

- Edit budget: $\|\Delta_X\|_{\mathrm{tok}} \le B_X$, $\|\Delta_A\|_{0,E} \le B_A$, where $\|\Delta_X\|_{\mathrm{tok}}$ counts Levenshtein token edits and $\|\Delta_A\|_{0,E}$ counts unordered edge additions/deletions;
- Semantic: BERTScore$(x_i, x_i') \ge 0.95$, perplexity ratio $\in [0.8, 1.2]$;
- Graph: $|\Delta h| \le 0.05$, degree KS $p > 0.1$, no new components, per-node incident cap;
- Transcript invariance: every retained step and every round passes $\mathcal{F}_t^{\boldsymbol{\epsilon}}$, with $\epsilon_\ell = \epsilon_{\mathrm{pref}} = \epsilon_\tau = \epsilon_w = 10^{-6}$ fixed before test-set evaluation.

We use $B_X = 20$ token edits and $B_A = 10$ unordered edge toggles. The budget-normalized cost is $C_{\mathrm{norm}}(\Delta_A, \Delta_X) = \|\Delta_A\|_{0,E}/B_A + \lambda_X \|\Delta_X\|_{\mathrm{tok}}/B_X$, with $\lambda_X = 0.5$ calibrated on clean validation.

Stealth is evaluated post hoc through a prespecified blind audit over crossed item–annotator judgments (Appendix A.6.2); it is supporting evidence, not a feasibility constraint.

The **invariant feedback harm** is defined only for transcript-invariant attacks. Let $\mathcal{V}_{\mathrm{tar}}$ be the fixed target-node evaluation set shared by all methods and seeds. We define $\mathcal{L}_{\mathrm{down}}(\phi) =$

$|\mathcal{V}_{\text{tar}}|^{-1}\sum_{i\in\mathcal{V}_{\text{tar}}}\text{CE}(f_\phi(x_i),y_i)$ after the final feedback round T. Harm is the unnormalized difference $\mathcal{L}_{\text{down}}(\phi'_T)-\mathcal{L}_{\text{down}}(\phi_T)$, and IFH divides it by the budget-normalized cost $C_{\text{norm}}(\Delta_A,\Delta_X)+\varepsilon$:

$$\text{IFH}=\frac{\mathcal{L}_{\text{down}}(\phi'_T)-\mathcal{L}_{\text{down}}(\phi_T)}{\|\Delta_A\|_{0,E}/B_A+\lambda_X\|\Delta_X\|_{\text{tok}}/B_X+\varepsilon}, \tag{8}$$

reported alongside unnormalized harm, edge edit count, and token edit count. Attacks that violate any field of R_t^{disc} or any component tolerance of $\mathbf{c}_t$ are considered visible and receive no IFH value.

4 METHOD: JUDGEBLIND

4.1 FC-CEFA: TRANSCRIPT-INVARIANT ATTACK CONSTRUCTION

FC-CEFA seeks harmful directions that leave the complete transcript unchanged. At a relaxed state, let $\mathbf{C}_t=D_z\mathbf{F}_t$. The predictor projects the attack-objective gradient g_t^{adv} onto the judge null space, $u_t=(I-\mathbf{C}_t^\dagger\mathbf{C}_t)g_t^{\text{adv}}$, then normalizes nonzero u_t. Graph-side second-order correction cancels local quadratic drift when the range condition in Theorem 5.2 holds. Newton refinement, sparse-edge projection, and tokenizer-aware realization produce discrete candidates. Each candidate must independently pass the complete transcript gate: relaxed closure alone does not guarantee discrete feasibility. Appendix A.3.1 gives implementation details.

4.2 MC-CMCT: MIXED-CURVATURE CANCELLATION TOMOGRAPHY

MC-CMCT separates visibility from harm: Proposition 5.1 detects student-visible directions, while the attack objective selects harmful candidates. The closure residual tests whether graph edits can cancel quadratic transcript drift. For the one-round student-update map $\Delta U_t^{(r)}$, write $\mathbf{A}_t^{(r)}=D_z\Delta U_t^{(r)}$ and $\mathcal{B}_t^{(r)}=D_z^2\Delta U_t^{(r)}$. With graph correction $d_A^{(2)\star}$ from Theorem 5.2 and graph-coordinate insertion $\mathbf{E}_A$, corrected exposure is

$$\mathcal{E}_{AX,t,\rho}^{(r)}=\left\|\rho\mathbf{A}_t^{(r)}v+\rho^2\left(\mathbf{A}_t^{(r)}\mathbf{E}_Ad_A^{(2)\star}+\tfrac{1}{2}\mathcal{B}_t^{(r)}[v,v]\right)\right\|_2. \tag{9}$$

Arguments $(z,\boldsymbol{\alpha}_t)$ are suppressed. This second-order approximation measures surviving student exposure and supplies CT-CAFT's defense target. Graph-assisted surplus is $\mathcal{S}_{\text{MC}}=[\mathcal{E}_{AX}-\mathcal{E}_X]_+$, where $\mathcal{E}_X$ uses $v_A=0$ without graph correction at the same state and weights. With both graph displacement and correction disabled, $\mu_{AX}=\mathcal{S}_{\text{MC}}=0$. Exact closure requires $\bar{r}_A^{(2)}=0$; numerical screening uses $\bar{r}_A^{(2)}\le\tau_r$ calibrated on validation. Appendix A.3.2 expands the diagnostics.

4.3 CT-CAFT: CURVATURE-TUBE CANCELLATION-AWARE DEFENSE

CT-CAFT jointly reweights the IT/PT channels around the transcript-preserved reference weights $\mathbf{s}_t$. Over at most K_{cont} gate-passing continuation states, $\mathfrak{E}_{t,\rho}^{(r)}$, $\mathfrak{S}_{t,\rho}^{(r)}$, and $\delta_{\text{probe},t}$ denote the worst corrected exposure, graph-assisted surplus, and registered-to-true-probe mismatch, respectively. CT-CAFT solves:

$$\begin{aligned}\min_{\phi,\boldsymbol{\alpha}_t^{\text{IT}},\boldsymbol{\alpha}_t^{\text{PT}}}\quad&\mathcal{L}_{\text{sup}}(\phi;\mathcal{D}_L)+\sum_{i\in\mathcal{V}_{\text{agreed}}}\alpha_{t,i}^{\text{IT}}\mathcal{L}_{\text{IT}}(\phi;x_i,\hat{y}_i^G)\\&+\lambda_{\text{PT}}\sum_{i\in\mathcal{V}'_{\text{disagreed}}}\alpha_{t,i}^{\text{PT}}\mathcal{L}_{\text{PT}}(\phi;x_i,\hat{y}_i^G,\hat{y}_i^L)\\&+\lambda_w\|\boldsymbol{\alpha}_t-\mathbf{s}_t\|_1+\lambda_T\mathfrak{E}_{t,\rho}^{(r)}(\boldsymbol{\alpha}_t)+\lambda_S\mathfrak{S}_{t,\rho}^{(r)}(\boldsymbol{\alpha}_t)\end{aligned} \tag{10}$$

subject to $\mathfrak{E}_{t,\rho}^{(r)}(\boldsymbol{\alpha}_t)\le\eta_t$, and for $q\in\{\text{IT},\text{PT}\}$ and every dataset class c, feasibility first requires $|\mathcal{C}_{t,c}^q|>0$—an empty class–channel cell makes that round and seed infeasible rather than vacuously satisfying coverage—and conditional on non-emptiness enforces $\sum_{i\in\mathcal{C}_{t,c}^q}\alpha_{t,i}^q\ge\kappa_c|\mathcal{C}_{t,c}^q|$, with $0\le\alpha_{t,i}\le1$. Complete suprema, optimizer details, and implementation are in Appendix A.3.3.

5 THEORETICAL ANALYSIS

Let $\mathbf{C}_t = D_z\mathbf{F}_t$, $\mathbf{C}_{A,t}$ be its graph block, $\mathcal{H}_t = D_z^2\mathbf{F}_t$, and $\mathbf{A}_t^{(r)} = D_z\Delta U_t^{(r)}$ the student-update Jacobian; $\mathbf{E}_A$ inserts graph coordinates. Standard rank and Taylor identities yield an operational test: can graph edits cancel transcript curvature while retaining student exposure? Proof sketches follow; Appendix A.5 gives details.

Proposition 5.1 (Exposed Subspace Rank Certificate)**.** *A judge-blind, student-visible direction for probe r at state z_t exists iff $d_{exp}^{(r)} = \operatorname{rank}[\mathbf{C}_t; \mathbf{A}_t^{(r)}] - \operatorname{rank}(\mathbf{C}_t) > 0$. Furthermore, $d_{exp}^{(r)} = \dim \operatorname{Null}(\mathbf{C}_t) - \dim(\operatorname{Null}(\mathbf{C}_t) \cap \operatorname{Null}(\mathbf{A}_t^{(r)})) = \operatorname{rank}(\mathbf{A}_t^{(r)}|_{\operatorname{Null}(\mathbf{C}_t)})$.*

Proof sketch. By rank-nullity on the restriction of $\mathbf{A}_t^{(r)}$ to $\operatorname{Null}(\mathbf{C}_t)$, and the stacked-matrix dimension formula. Full derivation in Appendix A.5.

Theorem 5.2 (Structural Second-Order Closure and Cubic Residual)**.** *Assume $\mathbf{F}_t(z_t) = \mathbf{0}$, $\mathbf{C}_t v = \mathbf{0}$, $\mathcal{H}_t[v, v] \in \operatorname{Range}(\mathbf{C}_{A,t})$, and that $\mathbf{C}_{A,t}$ has constant rank in a neighborhood of z_t. Assume $D^3\mathbf{F}_t$ is locally Lipschitz ($\mathbf{F}_t \in C^4$ is sufficient). Let $d_A^{(2)\star} = -\frac{1}{2}\mathbf{C}_{A,t}^{\dagger}\mathcal{H}_t[v, v]$ and $z(\rho) = z_t + \rho v + \rho^2\mathbf{E}_A d_A^{(2)\star}$. The statement concerns the relaxed local state; the derivatives and remainder constants are assumed uniform on a compact neighborhood of the corrected path. Then:*

$$\mathbf{F}_t(z(\rho)) = \rho^3\Big(D^2\mathbf{F}_t[v, \mathbf{E}_A d_A^{(2)\star}] + \frac{1}{6}D^3\mathbf{F}_t[v, v, v]\Big) + O(\rho^4), \tag{11}$$

$$\|\mathbf{F}_t(z(\rho))\|_2 \le \frac{M_{corr,t}}{6}\rho^3 + O(\rho^4), \tag{12}$$

where $M_{corr,t} = 6\|D^2\mathbf{F}_t[v, \mathbf{E}_A d_A^{(2)\star}]\|_2 + \|D^3\mathbf{F}_t[v, v, v]\|_2$. The exact condition $\bar{r}_A^{(2)}(v) = 0 \iff \mathcal{H}_t[v, v] \in \operatorname{Range}(\mathbf{C}_{A,t})$ is necessary and sufficient for cancellation of the quadratic coefficient by a graph-side correction; without it the generic residual is $O(\rho^2)$.

Proof sketch. Taylor expand $\mathbf{F}_t(z(\rho))$; the $O(\rho)$ term vanishes by $\mathbf{C}_t v = \mathbf{0}$, the $O(\rho^2)$ term vanishes by the range condition and $d = d_A^{(2)\star}$, leaving the $O(\rho^3)$ term as stated. Full derivation in Appendix A.5.

Theorem 5.3 (CT-CAFT Conditional Finite-Round Deviation Bound)**.** *Let e_t be the clean–attacked student-state deviation and a_t the clean-update state Lipschitz constant. For an evaluated tube pair, let $h_t^{(r^\star)}$ be the second-order update vector whose norm defines $\mathcal{E}_{AX,t,\rho_t}^{(r^\star)}$. Let q_t be the realized attacked-minus-clean one-round update at the same student state. Assume independently that $\|q_t - h_t^{(r^\star)}\|_2 \le \xi_t$ uniformly over compared states; this includes Taylor, realization, and tube-base discrepancies. The transcript gate alone does not bound ξ_t. Under CT-CAFT with $\mathfrak{E}_{t,\rho_t}^{(r)}(\boldsymbol{\alpha}_t) \le \eta_t$ and the shared-tube probe gap in eq.* (23)*,*

$$\|e_{t+1}\|_2 \le a_t\|e_t\|_2 + b_t, \tag{13}$$

where $b_t = \eta_t + \delta_{\mathrm{probe},t}(\boldsymbol{\alpha}_t) + \xi_t$. Consequently, $\|e_T\|_2 \le (\prod_{j=0}^{T-1} a_j)\|e_0\|_2 + \sum_{k=0}^{T-1}(\prod_{j=k+1}^{T-1} a_j)b_k$.

Proof sketch. The triangle inequality gives $\|q_t\|_2 \le \|h_t^{(r^\star)}\|_2 + \xi_t \le \eta_t + \delta_{\mathrm{probe},t} + \xi_t$. The clean-update Lipschitz condition gives the recurrence, and induction gives the product-sum bound. Theorem 5.2 controls relaxed judge residuals only; a numerical student-deviation certificate additionally requires an independent bound on ξ_t. Appendix A.5 provides the decomposition.

6 EXPERIMENTS

6.1 SETUP

Datasets and Victims. Following GNN-as-Judge (Xu and Ding, 2026), we evaluate Cora, CiteSeer, PubMed, ogbn-arxiv, and ogbn-products under 5-shot supervision, with a shot-count sensitivity sweep (3 and 10) and six fixed conditions: clean, heterophily, three OpenRTAG degradations (Dai

et al., 2026), and a combined text–edge extension. Pre-cap averages target/path rates within each seed before five-seed mean±SD reporting; stress/shot summaries macro-average the ten dataset–victim pairs within seeds. The primary GNN-as-Judge victim combines a 2-layer GCN with LoRA-tuned Llama-3-8B-Instruct; for the LLM–GNN Co-Teaching victim (Peng et al., 2026), we preserve the published bidirectional schedule ($T_{\mathrm{CT}} = 20$ rounds), with the adversary modifying only the approved GNN-to-LLM payload before each LLM update and the LLM-to-GNN channel left unmodified. The complete GNN-to-LLM decision transcript is checked in every round; Harm/IFH is evaluated on the final-round LLM state. GNN-as-Judge uses $T_{\mathrm{GAJ}} = 1$ feedback round. Five seeds (0–4) are used. Complete data splits, architecture, subgraph construction, hyperparameters, bootstrap, baseline mapping, and validation-only tuning are in Appendices A.1–A.4.

Attacks and Defenses. We evaluate under the white-box primary threat model with $B_X = 20$ token edits, $B_A = 10$ unordered edge toggles, $\lambda_X = 0.5$, and the full transcript gate with $\epsilon_\ell = \epsilon_{\text{pref}} = \epsilon_\tau = \epsilon_w = 10^{-6}$. Attacks include nine native-objective baselines, three FC-CEFA variants, and two transcript-aware constrained baseline implementations. Defenses include similarity filtering, margin regularization, controls, and CT-CAFT. All methods share identical budgets, observability thresholds, and pretrained checkpoints. Appendix Tables 6 and 7 specify the matching protocol.

6.2 Native-Baseline Compliance and Invariant Harm (RQ1)

Table 1 separates two questions: whether native attacks preserve the transcript, and whether FC-CEFA improves over methods explicitly constrained to do so. The latter comparison uses matched Penalty/Lagrangian and First-order controls with the same budgets and gate. Cross-setting results then test whether compliant harm persists beyond Cora; Table 2 isolates the correction mechanism.

Table 1: **Main compliance and cross-setting attack results.** Left: Cora/GNN-as-Judge; right: ten dataset–victim settings. Pre-cap averages target/path rates within each seed; 14 proposals per path-round. Left-panel Pre-cap and Harm/IFH report five-seed mean±SD; right-panel cells report means. IFH uses budget-normalized cost; drift is the maximum. *P* denotes Pass. Appendix Table 6 gives the native-objective audit.

(a) Cora compliance

Method	Pre-cap	Harm / IFH	Drift / Gate
Native objectives (9)	–	.044–.127 / –	$\geq 10^{-4}$ / 0/9
Penalty/Lagrangian	.287±.051	.093±.013 / .17±.03	8.9×10^{-7} / P
First-order null	.514±.067	.117±.016 / .24±.04	9.6×10^{-7} / P
FC-CEFA text-origin	.800±.079	.142±.018 / .31±.02	10^{-15} / P
FC-CEFA coupled	.886±.064	.189±.024 / .44±.06	10^{-15} / P
FC-CEFA adaptive	**.900±.038**	**.221±.019 / .53±.03**	10^{-15} **/ P**

(b) Ret.% / IFH

Dataset	GAJ	CoT
Cora	90.0/.53	87.2/.50
CiteSeer	86.8/.51	84.5/.47
PubMed	85.1/.48	82.3/.44
ogbn-arxiv	80.7/.43	78.2/.40
ogbn-products	77.5/.39	73.8/.36

Finding (RQ1). The matched controls isolate gains under identical constraints. Adaptive FC-CEFA raises IFH from 0.24 to 0.53 (2.21 times) and unnormalized Harm from 0.117 to 0.221 against the First-order control. Pre-cap feasibility rises by 38.6 percentage points. Results indicate more harmful compliant edits, with gate-passing attacks in all ten settings. Native-objective failures identify interface differences; they do not establish inferiority.

Table 2 asks whether compliant harm comes from the continuation pipeline alone or from its graph-side closure mechanisms.

Table 2: **FC-CEFA mechanism ablation at Cora/GNN-as-Judge.** Every retained path passes the complete transcript gate in all five seeds; Pre-cap and IFH are five-seed mean±SD, and Drift is the worst retained-step value. The tangent+gate row is an internal ablation of adaptive FC-CEFA, not the matched First-order null-space control in Table 1.

Variant	Drift	Pre-cap	IFH
FC-CEFA: tangent + gate only	10^{-8}	$.55 \pm .03$	$.41 \pm .03$
w/o graph-side Newton	10^{-10}	$.84 \pm .02$	$.49 \pm .02$
w/o full $H_{AA} + 2H_{AX} + H_{XX}$	10^{-11}	$.78 \pm .02$	$.47 \pm .02$
Full adaptive	$\leq 10^{-15}$	$\mathbf{.900 \pm .038}$	$\mathbf{.53 \pm .03}$

Mechanism finding. At the unchanged gate, removing full mixed curvature lowers pre-cap feasibility by 12 percentage points; removing graph-side Newton lowers it by 6. Tangent+gate reaches 55%, versus 90% for the full method. Higher feasibility accompanies higher IFH: correction finds admissible edits with greater harm per normalized cost, beyond the local tangent approximation.

The ablations establish construction gains. We next test whether diagnostics predict correction outcomes, then examine whether suppressing the measured exposure improves the defense trade-off.

6.3 MC-CMCT Diagnostic Prediction (RQ2)

Can local geometry predict which candidates merit correction? Figure 2 compares held-out convergence and exposure diagnostics; Appendix Tables 10 and 12 provide statistics.

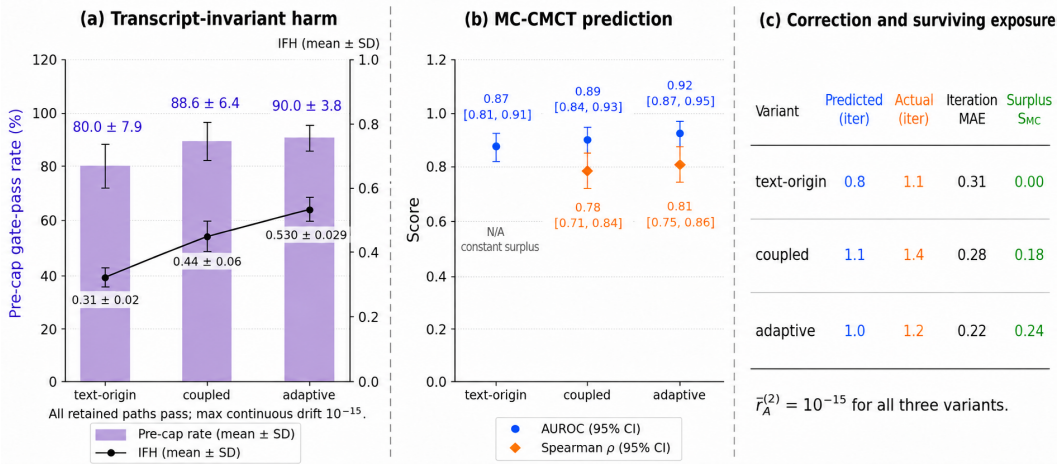


Figure 2: **FC-CEFA invariant harm and MC-CMCT diagnostics.** All retained FC-CEFA paths pass the complete transcript gate. Panel (a) reports five-seed mean±SD pre-cap gate-pass rate (averaged within seeds) and budget-normalized IFH; panel (b) reports direction-group bootstrap 95% CIs for corrector AUROC and Spearman ρ (N/A for reported constant surplus); panel (c) reports predicted versus actual Newton iterations, iteration MAE, and surviving exposure $\mathcal{S}_{\mathrm{MC}}$.

Finding (RQ2). MC-CMCT predicts convergence within 20 Newton iterations (AUROC 0.87–0.92; iteration MAE 0.22–0.31). Surplus correlates with direction-level IFH for coupled and adaptive variants ($\rho = 0.78/0.81$); the reported text-only surplus is constant, so its correlation is undefined. Convergence estimates address correction feasibility; surplus assesses the exposure remaining after that correction. These scores help prioritize correction candidates; Table 2 tests the construction mechanism. Experiments establish diagnostic value, without measuring runtime savings.

6.4 CT-CAFT Adaptive Defense (RQ3)

Does suppressing exposed directions preserve useful feedback? Figure 3 compares harm reduction against clean-accuracy loss and tests both defense components; Appendix A.10 reports class-wise coverage across ten settings.

Finding (RQ3). On Cora, CT-CAFT improves harm reduction over the margin control from 43.3% to 60.0% (paired gain 16.7 pp; descriptive five-seed bootstrap CI $[14.0, 19.4]$ pp), while clean-accuracy loss decreases from 1.4 to 0.7 pp. Removing surplus and tube reduces harm reduction by 14.9 and 7.6 pp, respectively. Across ten independently trained settings, reduction remains 52.2–60.0%, accuracy changes remain -1.9 to -0.7 pp, and every class in both feedback channels meets $\kappa_c = 0.65$ on every seed. Lower harm accompanies retained feedback and smaller clean-accuracy loss within the evaluated adaptive attack family.

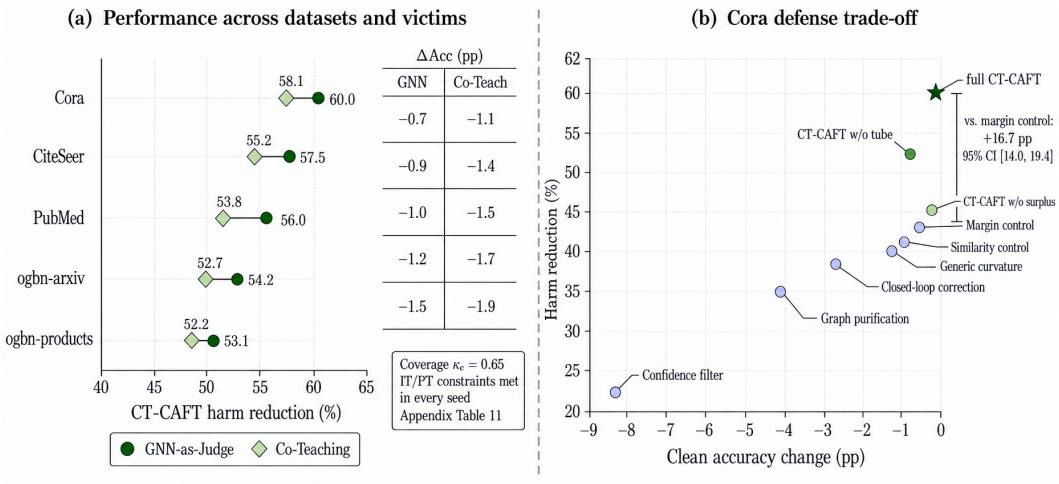


Figure 3: **CT-CAFT performance and harm–utility trade-off.** Margin and similarity controls denote internal implementations; Appendix Table 11 reports IT/PT coverage across ten settings.

Cross-RQ interpretation. JudgeBlind adds a downstream learning check to graph-based feedback approval. The contribution is to test approved feedback at the student update, using one transcript-preserving family for attack, diagnosis, and defense. RQ1 measures harmful edits under matched constraints; RQ2 tests whether local geometry predicts correction and exposure; RQ3 measures harm reduction together with clean utility and class-wise coverage. This yields a concrete evaluation criterion: approved feedback must be tested for downstream harm under the same decision constraints and budgets. The experiments support this claim for the evaluated graph–text setting, tube, and probe; they do not establish universal defense superiority.

7 Limitations

Our theory is local: the cubic result applies to relaxed states in the evaluated tube, while discrete token/edge realizability is established only by the empirical gate; rank or selector-margin violations are handled by backtracking. Probe gaps are estimated in student-update norm, Hessian-vector products contain finite-difference noise, and the gray-box study re-optimizes paths rather than proving frozen-path transfer. Finally, five-seed utility intervals are descriptive and are not equivalence intervals.

8 Conclusion

We introduced JudgeBlind, a closed-loop attack–diagnosis–defense framework for the cross-encoder feedback gap in GNN-to-LLM pipelines. Raw-token edits can alter the LLM student's update while preserving the GNN judge's complete non-textual transcript, with exact discrete equality and fixed continuous tolerances. FC-CEFA constructs the paths, MC-CMCT diagnoses closure and exposure, and CT-CAFT provides an attack-specific structured defense under per-class IT/PT coverage constraints. Across five TAG benchmarks and two victims, FC-CEFA attains a five-seed mean pre-cap gate-pass rate of 73.8–90.0%, MC-CMCT predicts corrector convergence with AUROC 0.87–0.92, and CT-CAFT yields 52.2–60.0% harm reduction with clean-accuracy changes from -1.9 to -0.7 pp. For feedback pipelines, the lesson is simple: evaluate text by its effect on student learning as well as its approval record. The attack, diagnostic, and defense make this distinction testable; the finite-round bound remains conditional on the evaluated tube and approximation error.

REPRODUCIBILITY

The supplement provides theorem-level sanity checks, unit tests, full-scale entry points, locked dependencies, configurations, and fail-closed scripts for Tables 1–3 and appendix summaries. Full-scale reproduction requires the stated TAG data and pretrained checkpoints. Figures 2–3 are rendered from reported summaries. The Figure 2 plotting script is supplied; the Figure 3 plotting script and seed-level metric logs are not included. The released package also contains the 5x2 matrix runner, the node-ID asset, the formula-to-code mapping, and the MIT License. Every full-scale entry point fails closed when required data, checkpoints, or configuration are missing. Complete details are in Appendix A.4.

ETHICS

The visibility audit involved three consenting adult annotators compensated at or above the local hourly rate, used only public benchmark data, and collected no sensitive personal information. The authors release only benchmark-scoped evaluation code.

AI USE

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

Hritik Bansal, Nishad Singhi, Yu Yang, Fan Yin, Aditya Grover, and Kai-Wei Chang. CleanCLIP: Mitigating data poisoning attacks in multimodal contrastive learning. arXiv preprint arXiv:2303.03323, 2023.

Alessandro Benfenati and Alessio Marta. A singular Riemannian geometry approach to deep neural networks II: Reconstruction of 1-D equivalence classes. arXiv preprint arXiv:2112.10583, 2021.

Avinandan Bose, Laurent Lessard, Maryam Fazel, and Krishnamurthy Dj Dvijotham. Keeping up with dynamic attackers: Certifying robustness to adaptive online data poisoning. In Proceedings of the 28th International Conference on Artificial Intelligence and Statistics, volume 258 of PMLR, pp. 4438–4446, 2025.

Harsh Chaudhari, Jamie Hayes, Matthew Jagielski, Ilia Shumailov, Milad Nasr, and Alina Oprea. Cascading adversarial bias from injection to distillation in language models. arXiv preprint arXiv:2505.24842, 2025.

Lu Chen, Shaofeng Li, Benhao Huang, Fan Yang, Zheng Li, Jie Li, and Yuan Luo. Contrasting adversarial perturbations: The space of harmless perturbations. arXiv preprint arXiv:2402.02095, 2024.

Zihui Chen, Yuling Wang, Pengfei Jiao, Kai Wu, Xiao Wang, Xiang Ao, and Dalin Zhang. Can LLMs fool graph learning? exploring universal adversarial attacks on text-attributed graphs. arXiv preprint arXiv:2603.21155, 2026.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. Subliminal learning: Language models transmit behavioral traits via hidden signals in data. arXiv preprint arXiv:2507.14805, 2025.

Yuze Dai, Zhihan Zhang, Yan Zhao, Ruoyu Wu, Xunkai Li, Zekai Chen, Qiangqiang Dai, Hongchao Qin, and Ronghua Li. OpenRTAG: A comprehensive benchmark for robust text-attributed graph learning under data quality degradation. arXiv preprint arXiv:2607.19108, 2026.

Ruomeng Ding, Yifei Pang, He Sun, Yizhong Wang, Zhiwei Steven Wu, and Zhun Deng. Rubrics as an attack surface: Stealthy preference drift in LLM judges. arXiv preprint arXiv:2602.13576, 2026.

Andrew Draganov, Tolga H. Dur, Anandmayi Bhongade, and Mary Phuong. Phantom transfer: Data poisoning can survive data-level defences. arXiv preprint arXiv:2602.04899, 2026.

Kaiwen Duan, Hongwei Yao, Yufei Chen, Ziyun Li, Tong Qiao, Zhan Qin, and Cong Wang. BadReward: Clean-label poisoning of reward models in text-to-image RLHF. arXiv preprint arXiv:2506.03234, 2025.

Taha Entesari, Sina Sharifi, and Mahyar Fazlyab. Compositional curvature bounds for deep neural networks. In Proceedings of the 41st International Conference on Machine Learning, volume 235 of PMLR, pp. 12527–12546, 2024.

Jonas Geiping, Liam Fowl, W. Ronny Huang, Wojciech Czaja, Gavin Taylor, Michael Moeller, and Tom Goldstein. Witches' brew: Industrial scale data poisoning via gradient matching. In The Ninth International Conference on Learning Representations (ICLR 2021), 2021.

Hanyi Hu, Qiao Han, Kui Chen, and Yao Yang. NPAT null-space projected adversarial training towards zero deterioration. arXiv preprint arXiv:2409.11754, 2024.

Qi Luo, Minghui Xu, Dongxiao Yu, and Xiuzhen Cheng. Graph-aware stealthy poison-text backdoors for text-attributed graphs. arXiv preprint arXiv:2603.20339, 2026.

Jiaji Ma, Puja Trivedi, and Danai Koutra. GRAPHTEXTACK: A realistic black-box node injection attack on LLM-enhanced GNNs. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2026), 2026a.

Yuhang Ma, Jie Wang, and Zheng Yan. Are LLM-enhanced graph neural networks robust against poisoning attacks? In 2026 IEEE Symposium on Security and Privacy, 2026b.

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

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.

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 The Eighth International Conference on Learning Representations (ICLR 2020), 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Yash Savani, Asher Trockman, Zhili Feng, Yixuan Even Xu, Avi Schwarzschild, Alexander Robey, Marc Finzi, and J. Zico Kolter. Antidistillation sampling. In Advances in Neural Information Processing Systems 38 (NeurIPS 2025), 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/aad5e2a483869d9ba3fab491686c3bf2-Paper-Conference.pdf.

Ilia Shumailov, Zakhar Shumaylov, Dmitry Kazhdan, Yiren Zhao, Nicolas Papernot, Murat A. Erdogdu, and Ross Anderson. Manipulating SGD with data ordering attacks. In Advances in Neural Information Processing Systems, volume 34, 2021.

Sahil Singla and Soheil Feizi. Second-order provable defenses against adversarial attacks. In Proceedings of the 37th International Conference on Machine Learning, volume 119 of PMLR, pp. 8981–8991, 2020.

Philip Sosnin, Mark N. Müller, Maximilian Baader, Calvin Tsay, and Matthew Wicker. Certified robustness to data poisoning in gradient-based training. arXiv preprint arXiv:2406.05670, 2024.

Terry Tong, Fei Wang, Zhe Zhao, and Muhao Chen. BadJudge: Backdoor vulnerabilities of LLM-as-a-Judge. In The Thirteenth International Conference on Learning Representations (ICLR 2025), 2025.

Eliot Tron, Nicolas Couellan, and Stéphane Puechmorel. Adversarial attacks on neural networks through canonical riemannian foliations. Machine Learning, 113:8655–8686, 2024. doi: 10.1007/s10994-024-06624-w. arXiv:2203.00922.

Zhongyuan Wang and Pratyusha Vemuri. When the tool decides: LLM agents defer blindly to graph neural network tools, and stronger backbones defer more. arXiv preprint arXiv:2606.14476, 2026.

Jian Xu, Delu Zeng, John Paisley, and Qibin Zhao. When can you debias an LLM judge? identifiability limits, a test, and designs for top-k ranking. arXiv preprint arXiv:2607.02104, 2026a.

Ruiyao Xu and Kaize Ding. GNN-as-Judge: Unleashing the power of LLMs for graph learning with GNN feedback. In The Fourteenth International Conference on Learning Representations (ICLR 2026), 2026. URL https://openreview.net/forum?id=nOlhDjNXKa.

Zixiang Xu, Sixian Li, Huaxing Liu, Xiang Wang, Shuai Li, Zirui Song, and Xiuying Chen. Inside the unfair judge: A mechanistic interpretability account of LLM-as-Judge bias. arXiv preprint arXiv:2607.11871, 2026b.

Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. arXiv preprint arXiv:2206.09391, 2022.

Junrui Zhang, Xinyu Zhao, Jie Peng, Chenjie Wang, Jianmin Ji, and Tianlong Chen. Vulnerabilityaware robust multimodal adversarial training. arXiv preprint arXiv:2511.18138, 2025.

Xing Zhang, Yanwei Cui, Guanghui Wang, Ziyuan Li, Wei Qiu, Bing Zhu, and Peiyang He. The blind curator: How a biased judge silently disables skill retirement in self-evolving agents. arXiv preprint arXiv:2607.07436, 2026a.

Yuxiang Zhang, Bin Ma, and Enyan Dai. Poisoning the inner prediction logic of graph neural networks for clean-label backdoor attacks. arXiv preprint arXiv:2603.05004, 2026b.

A Appendix

A.1 Dataset Statistics

Table 3 makes the graph sizes and supervision setting explicit before the robustness comparisons.

Table 3: Summary statistics of the five evaluation datasets. Raw-text preprocessing follows Xu and Ding (2026); ogbn-products reports the fixed 50,000-node induced subgraph used in all experiments. The degradation column lists the six fixed conditions; stress/shot summary results are macro-averaged over the ten dataset–victim pairs in Appendix A.1.1.

Dataset	Nodes	Edges	Classes	Shot (primary)	Degradation
Cora	2,708	5,429	7	5	clean, heterophily, 3 OpenRTAG, combined
CiteSeer	3,327	4,732	6	5	clean, heterophily, 3 OpenRTAG, combined
PubMed	19,717	44,338	3	5	clean, heterophily, 3 OpenRTAG, combined
ogbn-arxiv	169,343	1,166,243	40	5	clean, heterophily, 3 OpenRTAG, combined
ogbn-products	50,000	74,420	47	5	clean, heterophily, 3 OpenRTAG, combined

These graphs span different sizes and class counts. Clean and heterophily are not OpenRTAG atomic scenarios; the combined text–edge condition is our extension. The fixed ogbn-products induced subgraph is produced by stratified sampling with NumPy v1.24 and seed 42; its exact sorted node IDs are released as data/ogbn-products_seed42_node_ids.txt (50,000 unique, strictly increasing IDs in the legal range), with the file's SHA-256 recorded in the README and verified by the loader before use. Raw-text preprocessing follows Xu and Ding (2026).

A.1.1 Stress-Condition and Shot Sensitivity

Table 4 tests whether the clean-setting conclusions persist under data degradation.

Table 4: Stress-condition summary for full FC-CEFA adaptive and full CT-CAFT, macro-averaged over the ten dataset–victim pairs (Cora, CiteSeer, PubMed, ogbn-arxiv, ogbn-products $\times$ GNN-as-Judge, LLM–GNN Co-Teaching). For each seed, metrics are first computed for each dataset–victim pair and then macro-averaged over the ten pairs with equal weight; the table reports the five-seed mean$\pm$SD. Pre-cap gate-pass is the pre-cap gate-pass rate (target/path rates averaged within seeds, with 14 proposals per path-round); IFH is budget-normalized invariant feedback harm; CT-CAFT Harm Red. is relative to the undefended baseline; ΔAcc is the clean-accuracy percentage-point difference, CT-CAFT minus undefended. No test condition is used for tuning.

Condition	Pre-cap gate-pass rate	IFH	CT-CAFT Harm Red. (%)	ΔAcc (pp)
Clean (primary)	0.826 ± 0.008	0.45 ± 0.02	55.3 ± 1.8	-1.3 ± 0.4
Heterophily	0.812 ± 0.009	0.44 ± 0.02	53.9 ± 1.9	-1.5 ± 0.4
Structure sparsity	0.800 ± 0.009	0.43 ± 0.02	52.8 ± 2.0	-1.6 ± 0.4
Text noise	0.803 ± 0.009	0.42 ± 0.02	53.2 ± 1.9	-1.5 ± 0.4
Structure/edge noise	0.792 ± 0.010	0.41 ± 0.02	52.2 ± 2.1	-1.7 ± 0.5
Combined text–edge	0.775 ± 0.011	0.40 ± 0.02	51.2 ± 2.2	-1.9 ± 0.5

The combined condition gives the lowest IFH and defense reduction, while both effects persist across all listed conditions.

Table 5 then varies label availability while keeping the evaluation protocol fixed.

Table 5: Shot sensitivity of full FC-CEFA adaptive and full CT-CAFT, macro-averaged over the ten dataset–victim pairs. For each seed, metrics are first computed for each dataset–victim pair and then macro-averaged over the ten pairs with equal weight; the table reports the five-seed mean$\pm$SD. The 5-shot row is the primary operating point. No test condition is used for tuning.

Shots per class	Pre-cap gate-pass rate	IFH	CT-CAFT Harm Red. (%)	ΔAcc (pp)
3	0.800 ± 0.009	0.43 ± 0.02	52.8 ± 2.0	-1.6 ± 0.4
5	0.826 ± 0.008	0.45 ± 0.02	55.3 ± 1.8	-1.3 ± 0.4
10	0.840 ± 0.008	0.46 ± 0.02	56.2 ± 1.7	-1.1 ± 0.4

Across three to ten shots, IFH and defense reduction increase modestly, placing the primary five-shot setting between the two endpoints.

A.2 Baseline Adaptation and Matching Protocol

Table 6 separates native objectives from transcript-aware controls so that gate failures have a clear interpretation.

Table 6: **(a) Attack baselines: adaptation and matched-evaluation protocol.** All methods share identical pretrained checkpoints, observability thresholds, $B_X = 20$, and $B_A = 10$. Native objectives are audited without gate-aware optimization. The two transcript-aware constrained implementations use the same 14-proposal budget per target/path and round as FC-CEFA and are evaluated with the complete transcript gate.

Method	Native Objective	Allowed Edits	Budget Mapping	Tuning Split
Witches' Brew (Geiping et al., 2021)	gradient matching	token/edge adapter	B_X=20, B_A=10	validation only
BadReward (Duan et al., 2025)	clean-label poisoning	token adapter	B_X=20	validation only
Subliminal Learning (Cloud et al., 2026)	hidden trait transfer	token adapter	B_X=20	validation only
Phantom Transfer (Draganov et al., 2026)	gradient matching	token/edge adapter	B_X=20, B_A=10	validation only
Riemannian foliation (Tron et al., 2024)	output-preserving manifold	token adapter	B_X=20	validation only
Co-Attack adaptation (Zhang et al., 2022)	multi-surface gradient	token/edge adapter	B_X=20, B_A=10	validation only
Rubric-drift (Ding et al., 2026)	benchmark-compliant edit	token adapter	B_X=20	validation only
BadJudge (Tong et al., 2025)	judge backdoor	token adapter	B_X=20	validation only

Continued on next page

Table 6 (continued)

Method	Native Objective	Allowed Edits	Budget Mapping	Tuning Split
Data-ordering attack (Shumailov et al., 2021)	ordering perturbation	token adapter	B_X=20	validation only
Penalty/Lagrangian	harm + transcript/margin penalties	token/edge adapter	B_X=20, B_A=10	validation only
First-order null-space	$P_N g_{\mathrm{adv}}$ + gate	token/edge adapter	B_X=20, B_A=10	validation only
FC-CEFA text-origin	transcript-invariant	token adapter	B_X=20	validation only
FC-CEFA coupled	transcript-invariant	token/edge adapter	B_X=20, B_A=10	validation only
FC-CEFA adaptive	transcript-invariant	token/edge adapter	B_X=20, B_A=10	validation only

The constrained rows provide the matched construction comparison; native rows describe compliance under their documented attack interfaces.

Table 7 identifies the evaluated defense controls by their objectives.

Table 7: **(b) Defense controls: graph–text objectives.** Each row states the objective or filter documented for the evaluated control and its validation setting. The similarity and margin controls are not reproductions of CleanCLIP or VARMAT. Figure 3 labels these internal controls by their evaluated objectives. The margin control (43.3%) is the strongest evaluated control and supplies the RQ3 reference.

Method	Exact graph–text adaptation	Validation hyperparameter	Tuning	Status
Entropy filter	keep i if $\mathcal{H}_i^{\mathrm{mean}} = \mathrm{mean}_j H(p_{ij}) < \tau_H$; drop otherwise	$\tau_H = 0.55$	validation only	implementation control
Graph purification	drop edge (i, j) if text-sim. $s_{ij} < \tau_A$; no node drop	$\tau_A = 0.72$	validation only	implementation control
Closed-loop correction	$\alpha_i \propto 1/(1 + \lambda_{\mathrm{cl}} E_i)$ reweight per exposure E_i	$\lambda_{\mathrm{cl}} = 0.3$	validation only	implementation control
Generic curvature reg.	add $\lambda_H \,\mathrm{tr}(\mathcal{H}_{XX})$ to training loss	$\lambda_H = 0.5$	validation only	implementation control
Similarity control	graph view = encoded text emb. e_i; text view = raw token emb. t_i; filter by contrastive sim. $\sigma(e_i, t_i) < \tau_C$ (mapping image–text to graph–text)	$\tau_C = 0.5$	validation only	implementation control
Margin control	class-margin objective: $\mathcal{L} = \mathrm{CE} + \lambda_V \lVert\phi(x) - \mu\rVert_{\Sigma^{-1}}$ per class, margin m	$\lambda_V = 0.5, m = 1.0$	validation only	implementation control

These objectives define the scope of the defense comparison: the reported advantage is over the listed implementations.

All nine attack baselines are evaluated under their published or native objectives without re-optimization for the transcript gate. Budget mapping converts each method's native edit interface (token-only, edge-only, or joint) to the shared $B_X = 20$, $B_A = 10$, $\lambda_X = 0.5$ constraint. Tuning is restricted to validation splits only; no test-set information is used for adaptation. For Co-Attack, we preserve its collaborative two-surface objective and map the original image and text perturbation surfaces to unordered graph-edge toggles and token edits, respectively; this is an adaptation, not the original task interface. The transcript gate is applied uniformly to all methods for post-hoc evaluation; methods that fail the gate are reported as visible (Gate=Fail) and receive no IFH value.

GRAPHTEXTACK is excluded from the matched edit-budget table because node injection changes the candidate universe and violates the fixed-record interface. BadGraph is discussed only as an adjacent threat model because it is a universal black-box test-time attack on model predictions, not a transcript-constrained attack on the approved GNN-to-LLM training payload.

Relationship to DRO and min-max objectives. The CT-CAFT objective in Section 4.3 shares structural elements with distributionally robust optimization (Namkoong and Duchi, 2016; Sagawa et al., 2020): it minimizes a worst-case loss over a pre-constructed uncertainty set (the FC-CEFA continuation tube). The tube is not an arbitrary DRO ball—it is the specific transcript-invariant path family produced by the FC-CEFA attack, which gives the uncertainty set a structural interpretation tied to the judge-blind fiber geometry. The per-class coverage constraints further distinguish CT-CAFT from unconstrained DRO formulations. Retrieval-based defenses that rely on clean neighbor representations likewise do not reweight feedback under a transcript-invariant attack path; they are

not directly comparable to CT-CAFT, which optimizes weights over a judge-blind exposure tube rather than retrieving neighbors.

Victim protocol. For the LLM–GNN Co-Teaching victim, the published bidirectional schedule and both native update rules are preserved: the GNN-to-LLM feedback update and the LLM-to-GNN label update both run as in (Peng et al., 2026). GNN-as-Judge uses $T_{\mathrm{GAJ}} = 1$ feedback round. LLM–GNN Co-Teaching uses its native $T_{\mathrm{CT}} = 20$ co-teaching rounds: both peers update in every round, RPL-PO preference pairs are mined from consecutive odd/even rounds, the adversary modifies only the approved GNN-to-LLM payload before each LLM update, and the LLM-to-GNN channel receives no adversarial modification. The complete GNN-to-LLM decision transcript (all selected-node and membership fields, target/preference labels, and continuous fields) is checked immediately after selection in every one of the 20 rounds, and Harm/IFH is evaluated on the final round-20 LLM state. The 10 IT epochs are optimizer epochs within a GNN-as-Judge feedback round and are not co-teaching rounds.

METRIC DEFINITIONS

For the main results table, **Retention** is the *pre-cap gate-pass rate*. Each target/path has 14 proposals per feedback round; a proposal passes only if Newton convergence, structural, student-visibility, and complete transcript checks all pass. For seed s, average these rates over its evaluated target/path–round units: $p_s = M_s^{-1} \sum_{j=1}^{M_s} g_{sj}/14$, where g_{sj} counts passing proposals and M_s counts evaluated units. Tables report $\bar{p} = \frac{1}{5}\sum_{s=0}^{4} p_s$ and sample SD $[\frac{1}{4}\sum_{s=0}^{4}(p_s - \bar{p})^2]^{1/2}$. Thus, 14 is the per-unit budget, not the total seed-level denominator. The post-cap fraction per unit is at most $10/14$; Pre-cap counts passes before that cap. **Max Run** is the longest consecutive retained-step run within a feedback round, computed per seed and reported as mean±SD. **Max Drift** is $\max_{t,k,j} |F_{t,j}(z_{t,k})|$ over all retained steps and continuous fields, reported as the worst observed value across the five seeds. **Harm** is the unnormalized final-round target-set CE increase; **IFH** = Harm $/(C_{\mathrm{norm}} + \varepsilon)$ uses $C_{\mathrm{norm}} = \|\Delta_A\|_{0,E}/B_A + \lambda_X\|\Delta_X\|_{\mathrm{tok}}/B_X$ and is defined only for gate-passing attacks. Thus, IFH measures harm per budget-normalized edit cost, not per raw edit. Discrete fields require exact ID/integer equality; continuous fields are tolerant, with worst float64 drift $\leq 10^{-15}$ below the preregistered 10^{-6} componentwise tolerance.

For the defense tables, **Harm Reduction** $= 100 \times (\mathrm{Harm}_{\mathrm{undefended}} - \mathrm{Harm}_{\mathrm{defended}})/\mathrm{Harm}_{\mathrm{undefended}}$; **Min. Coverage** is the per-class minimum over the IT and PT channels, and every seed satisfies $\min_c \mathrm{Cov} \geq \kappa_c$ at $\kappa_c = 0.65$ up to 10^{-6} tolerance; **ΔAcc** is the clean-accuracy percentage-point difference, defended minus undefended. All internal-control thresholds (τ_H, τ_A, λ_{cl}, λ_H) are selected on validation only and frozen before test evaluation; the margin control supplies the RQ3 comparison; neither this control nor similarity filtering establishes performance against the original external methods.

A.3 COMPLETE METHOD DEFINITIONS AND ALGORITHMIC DETAILS

A.3.1 FC-CEFA PREDICTOR–CORRECTOR DETAILS

Let $\mathbf{F}_t$ be the continuous transcript-residual map of eq. (5); the judge Jacobian is $\mathbf{C}_{t,k} = D_z\mathbf{F}_t(z_{t,k})$ and the vector-valued Hessian is $\mathcal{H}_{t,k} = D_z^2\mathbf{F}_t(z_{t,k})$.

Tangent predictor. FC-CEFA computes $u_{t,k} = \widehat{\mathbf{P}}_{N,t,k}\, g_{t,k}^{\mathrm{adv}}$ with the nullspace projector $\widehat{\mathbf{P}}_{N,t,k} = \mathbf{I} - \widehat{\mathbf{C}}_{t,k}^{\dagger}\widehat{\mathbf{C}}_{t,k}$. If $\|u_{t,k}\|_2 \leq 10^{-12}$ (float64), this objective gradient supplies no usable tangent direction, so the step is rejected; otherwise $v_{t,k} = u_{t,k}/(\|u_{t,k}\|_2 + \varepsilon_v)$, and the tangent step $z_{t,k}^{\mathrm{pred}} = z_{t,k} + \rho_k v_{t,k}$ satisfies $\widehat{\mathbf{C}}_{t,k}v_{t,k} \approx 0$.

Blockwise second-order closure. Taylor expansion gives $\mathbf{F}_t(z_{t,k}^{\mathrm{pred}}) = \mathbf{F}_t(z_{t,k}) + \rho_k\mathbf{C}_{t,k}v_{t,k} + \frac{\rho_k^2}{2}\mathcal{H}_{t,k}[v_{t,k}, v_{t,k}] + O(\rho_k^3)$. At an exact fiber base and tangent, the first two terms vanish. FC-CEFA cancels quadratic drift with the graph-side closure:

$$d_A^{(2)} = d_{AA}^{(2)} + d_{AX}^{(2)} + d_{XX}^{(2)}, \quad \text{where} \tag{14}$$
$$d_{AA}^{(2)} = -\frac{1}{2}\mathbf{C}_{A,t}^{\dagger}\,\mathcal{H}_{AA,t}[v_A, v_A],$$

$$d_{AX}^{(2)} = -\mathbf{C}_{A,t}^{\dagger}\,\mathcal{H}_{AX,t}[v_A, v_X],$$
$$d_{XX}^{(2)} = -\frac{1}{2}\mathbf{C}_{A,t}^{\dagger}\,\mathcal{H}_{XX,t}[v_X, v_X].$$

Newton correction. Starting from $d_A^{(0)} = \rho_k^2 d_A^{(2)}$, sparse Newton steps refine:

$$d_A^{(\ell+1)} = d_A^{(\ell)} - \mathbf{C}_A(z_{t,k}^{\text{pred}} + \mathbf{E}_A d_A^{(\ell)})^{\dagger}\,\mathbf{F}_t(z_{t,k}^{\text{pred}} + \mathbf{E}_A d_A^{(\ell)}). \tag{15}$$

A proximal top-B_A projection converts the result into sparse edge edits; token candidates use tokenizer-aware beam search (beam width per dataset in Appendix).

A.3.2 MC-CMCT Exposure Definitions

Rank certificate. For judge Jacobian $\mathbf{C}_t$ and student Jacobian $\mathbf{A}_t^{(r)}$:

$$d_{\text{exp}}^{(r)} = \operatorname{rank}\begin{bmatrix}\mathbf{C}_t \\ \mathbf{A}_t^{(r)}\end{bmatrix} - \operatorname{rank}(\mathbf{C}_t) > 0 \tag{16}$$

iff a judge-blind, student-visible direction exists for probe r.

Curvature decomposition. Partition a unit tangent $v = [v_A; v_X]$:

$$\mathcal{H}_t[v, v] = \mathcal{H}_{AA,t}[v_A, v_A] + 2\mathcal{H}_{AX,t}[v_A, v_X] + \mathcal{H}_{XX,t}[v_X, v_X]. \tag{17}$$

The normalized interaction index is $\mu_{AX,t}(v) = \|2\mathcal{H}_{AX,t}[v_A, v_X]\|_2/(\|\mathcal{H}_{AA,t}[v_A, v_A]\|_2 + \|\mathcal{H}_{XX,t}[v_X, v_X]\|_2 + \varepsilon)$.

Joint corrected student exposure. Let $\Delta U_t^{(r)}(z; \boldsymbol{\alpha}_t)$ be the one-round student update under defensive weights $\boldsymbol{\alpha}_t = [\boldsymbol{\alpha}_t^{\text{IT}}; \boldsymbol{\alpha}_t^{\text{PT}}]$ at continuation state z, with local derivatives:

$$\mathbf{A}_t^{(r)}(z, \boldsymbol{\alpha}_t) = D_z \Delta U_t^{(r)}(z; \boldsymbol{\alpha}_t), \qquad \mathcal{B}_t^{(r)}(z, \boldsymbol{\alpha}_t) = D_z^2 \Delta U_t^{(r)}(z; \boldsymbol{\alpha}_t). \tag{18}$$

Let the local graph-side closure coefficient be $d_{A,t}^{(2)\star}(z, v) = -\frac{1}{2}\mathbf{C}_{A,t}(z)^{\dagger}\,\mathcal{H}_t(z)[v, v]$. For student probe r, radius ρ, and a direction passing the closure test:

$$\mathcal{E}_{AX,t,\rho}^{(r)}(z, v, \boldsymbol{\alpha}_t) = \left\|\rho \mathbf{A}_t^{(r)}(z, \boldsymbol{\alpha}_t)v + \rho^2\big(\mathbf{A}_t^{(r)}(z, \boldsymbol{\alpha}_t)\mathbf{E}_A d_{A,t}^{(2)\star}(z, v) + \tfrac{1}{2}\mathcal{B}_t^{(r)}(z, \boldsymbol{\alpha}_t)[v, v]\big)\right\|_2, \tag{19}$$

$$\mathcal{S}_{\text{MC},t}^{(r)}(z, v, \boldsymbol{\alpha}_t) = \max\{0, \mathcal{E}_{AX,t,\rho}^{(r)}(z, v, \boldsymbol{\alpha}_t) - \mathcal{E}_{X,t,\rho}^{(r)}(z, v, \boldsymbol{\alpha}_t)\}, \tag{20}$$

where $\mathcal{E}_{X,t,\rho}^{(r)}(z, v, \boldsymbol{\alpha}_t)$ is the text-only exposure with $v_A = 0$ and no graph correction at the same state z and weights $\boldsymbol{\alpha}_t$.

Diagnostic outputs. MC-CMCT reports μ_{AX}, $\bar{r}_A^{(2)}$, predicted/actual Newton iterations, corrector success, and $\mathcal{S}_{\text{MC}}$. Corrector AUROC uses the independently observed Newton-convergence label (convergence within 20 iterations) and score $-\bar{r}_A^{(2)}$. Iteration counts are predicted by a Huber regressor fitted only on disjoint calibration direction groups, using $-\log_{10}(\bar{r}_A^{(2)} + 10^{-10})$ as its scalar input and clipping predictions to $[0, 20]$; MAE is evaluated on held-out groups. For every held-out closure-passing direction whose final discrete realization passes the complete transcript gate, direction-level $\text{IFH}(v)$ is recomputed from that direction's own realized candidate via Eq. equation 8; no round-level or path-level IFH is copied across directions. For each seed, Spearman ρ is computed over held-out direction-level pairs $(\mathcal{S}_{\text{MC}}(v), \text{IFH}(v))$ with nonconstant surplus; the reported value is the five-seed mean. The reported text-origin surplus is zero; this follows structurally only when both graph displacement and correction are disabled. Its reported correlation and interval are N/A. Group-bootstrap confidence intervals resample complete continuation-path groups within a seed and recompute the five-seed mean, while calibration and evaluation groups remain disjoint.

A.3.3 CT-CAFT Tube Quantities

Let $\mathcal{T}_{t,\rho}$ contain at most K_{cont} gate-passing FC-CEFA continuation states of radius ρ. The worst tube exposure for the registered probe r is:

$$\mathfrak{E}_{t,\rho}^{(r)}(\boldsymbol{\alpha}_t) = \sup_{\substack{z \in \mathcal{T}_{t,\rho} \\ v \in \mathcal{V}_{AX,t}^{(2)}(z)}} \mathcal{E}_{AX,t,\rho}^{(r)}(z, v, \boldsymbol{\alpha}_t). \tag{21}$$

The worst tube surplus for the registered probe r is:

$$\mathfrak{S}_{t,\rho}^{(r)}(\boldsymbol{\alpha}_t) = \sup_{\substack{z\in\mathcal{T}_{t,\rho} \\ v\in\mathcal{V}_{AX,t}^{(2)}(z)}} \mathcal{S}_{\mathrm{MC},t}^{(r)}(z, v, \boldsymbol{\alpha}_t). \tag{22}$$

The probe approximation gap when the true student $r^\star$ differs from the registered probe r over the same tube and direction family:

$$\delta_{\text{probe},t}(\boldsymbol{\alpha}_t) = \sup_{\substack{z\in\mathcal{T}_{t,\rho} \\ v\in\mathcal{V}_{AX,t}^{(2)}(z)}} \left[\mathcal{E}_{AX,t,\rho}^{(r^\star)}(z, v, \boldsymbol{\alpha}_t) - \mathcal{E}_{AX,t,\rho}^{(r)}(z, v, \boldsymbol{\alpha}_t)\right]_+. \tag{23}$$

Coverage constraints (expanded). For every round t, channel $q \in \{\mathrm{IT}, \mathrm{PT}\}$, and dataset class c, feasibility first requires $|\mathcal{C}_{t,c}^q| > 0$; an empty class–channel cell makes that round and seed infeasible rather than vacuously satisfying coverage. Conditional on non-emptiness, CT-CAFT enforces per-class IT coverage $\sum_{i\in\mathcal{C}_{t,c}^{\mathrm{IT}}} \alpha_i^{\mathrm{IT}} \geq \kappa_c|\mathcal{C}_{t,c}^{\mathrm{IT}}|$ and per-class PT coverage $\sum_{i\in\mathcal{C}_{t,c}^{\mathrm{PT}}} \alpha_i^{\mathrm{PT}} \geq \kappa_c|\mathcal{C}_{t,c}^{\mathrm{PT}}|$ for every class c. All evaluated seeds contain every class in both clean approved channels, so the reported minima range over all dataset classes and all evaluated rounds, with no empty cells omitted. The transcript gate preserves the $(\hat{y}^G, \hat{y}^L)$ preference pair.

A.4 Implementation and Hyperparameters

GCN: 2 layers, 64-dim hidden, dropout $\in [0.3, 0.5, 0.7]$, batch normalization, lr = 10^{-2}, 200 epochs, patience = 100. LLM: Llama-3-8B-Instruct, LoRA (rank 8, $\alpha = 16$, dropout 0.1, batch size 8). IT: lr 5×10^{-6}, 10 epochs. CT-CAFT: lr 10^{-5}, 8 epochs, $\lambda_{\mathrm{PT}} = 1.0$, $\lambda_w = 0.1$, $\lambda_T = 1.0$, $\lambda_S = 0.5$, η_t calibrated on validation per dataset; optimization uses SLSQP for the toy verification and projected primal–dual with joint Lagrangian for full-scale experiments, where both coverage and exposure constraints are enforced simultaneously rather than through sequential independent projections. FC-CEFA uses a judge selector candidate pool $K_{\mathrm{sel}} = 1500$, selector threshold $\tau = 0.7$, a continuation horizon of $K_{\mathrm{cont}} = 10$ states retained after the cap, a proposal budget of 14 registered continuation proposals per target/path in each feedback round, tokenizer-aware realization beam width 5, and at most 20 Newton iterations per proposed state; $\epsilon_\ell = \epsilon_{\text{pref}} = \epsilon_\tau = \epsilon_w = 10^{-6}$ (logits in float64). GNN-as-Judge uses $T_{\mathrm{GAJ}} = 1$ feedback round, and LLM–GNN Co-Teaching uses its native $T_{\mathrm{CT}} = 20$ co-teaching rounds; in both victims the 10 IT epochs are optimizer epochs within a GNN-as-Judge feedback round and are not co-teaching rounds. The CT-CAFT tube contains at most $K_{\mathrm{cont}} = 10$ final gate-passing continuation states per round. The SVD-based nullspace projector reports rank and minimum non-zero singular value at each step; the graph-side Newton corrector uses backtracking line search; the transcript gate checks exact integer/ID equality for all five discrete fields (R_t^{disc}) and componentwise tolerance for all five continuous fields ($\boldsymbol{\ell}_t^G$, $\mathbf{S}_{\text{pref}}$, τ_t, $\mathbf{s}_t^{\mathrm{IT}}$, $\mathbf{s}_t^{\mathrm{PT}}$). All hyperparameters were fixed before test-set evaluation and held identical across seeds. Held-out probes: GPT-2 tokenizer, Llama-3-70B.

A.5 Complete Proofs

Proposition 5.1 (Rank Certificate). Let $\mathbf{C}_t \in \mathbb{R}^{m\times n}$ and $\mathbf{A}_t^{(r)} \in \mathbb{R}^{p\times n}$. Consider the restriction $\mathbf{A}_t^{(r)}|_{\mathrm{Null}(\mathbf{C}_t)}$. By the rank-nullity theorem applied to this restricted map, $\dim(\mathrm{Null}(\mathbf{C}_t)) - \dim(\mathrm{Null}(\mathbf{C}_t) \cap \mathrm{Null}(\mathbf{A}_t^{(r)})) = \mathrm{rank}(\mathbf{A}_t^{(r)}|_{\mathrm{Null}(\mathbf{C}_t)})$. This equals $\mathrm{rank}[\mathbf{C}_t; \mathbf{A}_t^{(r)}] - \mathrm{rank}(\mathbf{C}_t)$ by the dimension formula for stacked matrices. Thus $d_{\text{exp}}^{(r)} = \mathrm{rank}(\mathbf{A}_t^{(r)}|_{\mathrm{Null}(\mathbf{C}_t)})$ is the dimension of the exposed quotient $\mathrm{Null}(\mathbf{C}_t)/(\mathrm{Null}(\mathbf{C}_t) \cap \mathrm{Null}(\mathbf{A}_t^{(r)}))$. Non-existence iff $d_{\text{exp}}^{(r)} = 0 \Leftrightarrow \mathrm{Null}(\mathbf{C}_t) \subseteq \mathrm{Null}(\mathbf{A}_t^{(r)})$. The quotient geometry is illustrated in Figure 4(a).

Theorem 5.2 (Cubic Residual). Assume $\mathbf{F}_t(z_t) = \mathbf{0}$, $\mathbf{C}_t v = \mathbf{0}$, $\mathcal{H}_t[v, v] \in \mathrm{Range}(\mathbf{C}_{A,t})$, and $\mathbf{C}_{A,t}$ has constant rank in a neighborhood of z_t. Assume $D^3\mathbf{F}_t$ is locally Lipschitz ($\mathbf{F}_t \in C^4$ is sufficient). Let $d_A^{(2)\star} = -\frac{1}{2}\mathbf{C}_{A,t}^\dagger \mathcal{H}_t[v, v]$.

Consider the path $z(\rho) = z_t + \rho v + \rho^2 \mathbf{E}_A d_A^{(2)\star}$. Taylor expand $\mathbf{F}_t(z(\rho))$ to third order:

$$\mathbf{F}_t(z(\rho)) = \mathbf{F}_t(z_t) + D\mathbf{F}_t(z_t)[\rho v + \rho^2 \mathbf{E}_A d]$$

$$+\frac{1}{2}D^2\mathbf{F}_t(z_t)[\rho v+\rho^2\mathbf{E}_A d,\rho v+\rho^2\mathbf{E}_A d]$$
$$+\frac{1}{2}\int_0^1(1-s)^2D^3\mathbf{F}_t(z_t+s\Delta z)[\Delta z,\Delta z,\Delta z],ds.$$

Since $\mathbf{F}_t(z_t)=\mathbf{0}$ (clean state is on the fiber), $D\mathbf{F}_t[\rho v]=\rho\mathbf{C}_t v=\mathbf{0}$, and $D\mathbf{F}_t[\rho^2\mathbf{E}_A d]=\rho^2\mathbf{C}_{A,t}d$:

$$\mathbf{F}_t(z(\rho))=\rho^2\mathbf{C}_{A,t}d+\frac{\rho^2}{2}(\mathcal{H}_t[v,v]+O(\rho))+O(\rho^3)$$
$$=\rho^2(\mathbf{C}_{A,t}d+\frac{1}{2}\mathcal{H}_t[v,v])+O(\rho^3).$$

By the range condition, $\frac{1}{2}\mathcal{H}_t[v,v]=-\mathbf{C}_{A,t}d_A^{(2)\star}$, so the $O(\rho^2)$ term vanishes. Expanding the $O(\rho^3)$ term yields the cross-term $D^2\mathbf{F}_t[v,\mathbf{E}_A d_A^{(2)\star}]$ and the third derivative $D^3\mathbf{F}_t[v,v,v]/6$, giving the stated bound with $M_{\mathrm{corr},t}=6\|D^2\mathbf{F}_t[v,\mathbf{E}_A d_A^{(2)\star}]\|_2+\|D^3\mathbf{F}_t[v,v,v]\|_2$. Without structural correction, the $O(\rho^2)$ term is generically nonzero. Figure 4(b–c) visualizes the cancellation of the quadratic term.

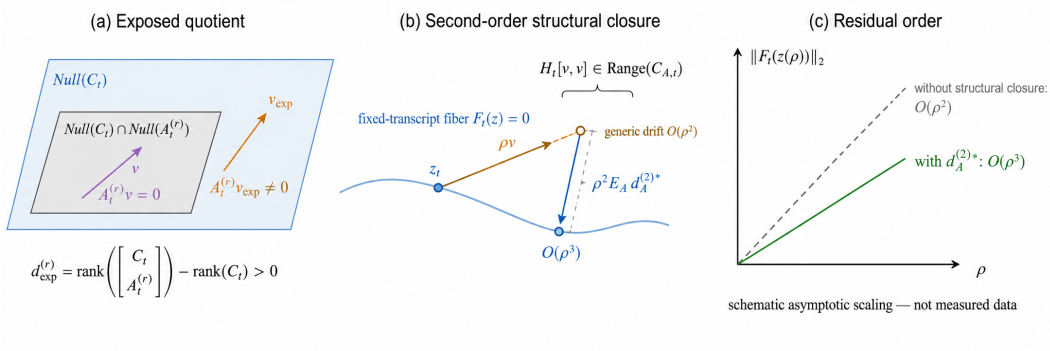


Figure 4: **Geometry behind MC-CMCT and FC-CEFA correction.** (a) The exposed quotient contains directions that the judge Jacobian cannot see but the student update can. (b) When the full graph–text curvature lies in $\mathrm{Range}(\mathbf{C}_{A,t})$, the minimum-norm graph-side correction cancels the quadratic transcript drift. (c) This changes the generic judge residual from $O(\rho^2)$ to $O(\rho^3)$; the lines are schematic asymptotic rates, not measured data.

Figure 4 explains local judge stability. The remaining question is how student-update discrepancies accumulate over rounds.

Theorem 5.3 (Conditional Finite-Round Bound). Write the clean and realized attacked update maps as U_t^0 and U_t^1, including their prescribed feedback weights. Assume U_t^0 is a_t-Lipschitz in student state on the compared trajectories. At ϕ_t', define

$$q_t=U_t^1(\phi_t')-U_t^0(\phi_t'),\qquad r_t=q_t-h_t^{(r^\star)},\qquad \|r_t\|_2\le\xi_t. \tag{24}$$

Here $h_t^{(r^\star)}$ is the vector inside the norm defining the true-student corrected exposure at the evaluated tube pair. The residual r_t includes the local expansion remainder, discrete realization error, and any displacement between the expansion base and the clean reference. The shared-tube probe-gap definition and the CT-CAFT exposure constraint imply

$$\|h_t^{(r^\star)}\|_2=\mathcal{E}_{AX,t,\rho_t}^{(r^\star)}\le\mathcal{E}_{AX,t,\rho_t}^{(r)}+\delta_{\mathrm{probe},t}\le\eta_t+\delta_{\mathrm{probe},t}. \tag{25}$$

Thus, by the triangle inequality,

$$\|e_{t+1}\|_2=\|U_t^1(\phi_t')-U_t^0(\phi_t)\|_2$$
$$\le\|q_t\|_2+\|U_t^0(\phi_t')-U_t^0(\phi_t)\|_2$$

$$\leq \eta_t + \delta_{\mathrm{probe},t} + \xi_t + a_t \|e_t\|_2. \tag{26}$$

Induction yields the stated product-sum bound. This is a conditional decomposition, not an empirically evaluated numerical certificate. In particular, a small judge-transcript residual cannot by itself bound r_t: student-visible directions in the judge null space are precisely the phenomenon studied here. Theorem 5.2 establishes a relaxed-state cubic judge residual and does not remove this independent student-update requirement.

A.6 Extended Experimental Results

A.6.1 Probe Sensitivity under Independent Re-optimization

Table 8 examines how changing the student probe affects independently optimized attacks.

Table 8: Probe sensitivity under independent re-optimization. Gate-pass rate and IFH use independently optimized paths. The empirical gap $\widehat{\delta}_{\mathrm{probe}}$ uses the shared tube/direction set $\mathcal{Q}_t$ in eq. (23), measured in student-update norm; the registered gap is zero by definition. This evaluates probe mismatch, not frozen-path transfer.

Probe	**Pre-cap gate-pass rate**	**Max Drift**	**IFH**	$\widehat{\delta}_{\mathrm{probe}}$
Registered (Llama-3-8B)	0.900±0.038	10^{-15}	0.530±0.029	0.000
GPT-2 tokenizer	0.721±0.062	10^{-12}	0.38±0.07	0.023±0.012
Llama-3-70B checkpoint	0.845±0.041	10^{-14}	0.49±0.05	0.011±0.009

The empirical probe gap $\widehat{\delta}_{\mathrm{probe}}$ is the maximum positive exposure difference over the shared evaluation set $\mathcal{Q}_t$, computed in the student-update norm via eq. (23). Registered row is 0.000 by definition; GPT-2 and Llama-3-70B values are the five-seed mean±std of the per-seed maximum positive difference. The largest observed IFH difference is 0.15 (registered Llama-3-8B versus the GPT-2 tokenizer probe); the difference versus the Llama-3-70B checkpoint is 0.04. These differences are reported only for interpretation and are not substituted into Theorem 5.3. Pre-cap gate-pass rates for held-out probes refer to independently re-optimized FC-CEFA paths under each surrogate student; the GTID (GNN judge) and proposed-step sets differ from the registered probe. GPT-2 tokenizer is a tokenizer-mismatch probe embedded within a Llama-3 student shell, not a standalone checkpoint replacement.

A.6.2 Semantic and Graph Visibility Audit

A blind audit with three annotators (Fleiss' $\kappa = 0.78$) evaluated 50 clean-vs-attacked pairs, 10 per dataset, producing 150 crossed item–annotator judgments. Discrimination accuracy was $78/150 = 0.520$. Since judgments are clustered by pair and annotator, we report no iid binomial confidence interval or p-value. BERTScore averaged 0.972 ± 0.018, perplexity ratio 1.04 ± 0.08, and the degree KS-test $p = 0.23 \pm 0.11$. These statistics support semantic and structural stealth but do not prove it.

A.7 Sensitivity Analysis

ϵ_ℓ was swept from 10^{-12} to 10^{-4} on Cora 5-shot validation, with $\epsilon_{\mathrm{pref}} = \epsilon_\tau = \epsilon_w = 10^{-6}$ fixed. The attack is re-optimized independently at each ϵ_ℓ, so the proposed-step sets differ and monotonicity is not guaranteed. We evaluate all thresholds on one fixed proposal set, for which the pre-cap gate-pass rate is non-decreasing in ϵ_ℓ. The $\epsilon_\ell = 10^{-6}$ fixed before test-set evaluation sits in a stable high-pre-cap-gate-pass plateau. λ_X (normalized token-cost weight) was swept from 0.1 to 2.0 on Cora; the IFH ranking across FC-CEFA variants is stable for $\lambda_X \in [0.3, 1.0]$. κ_c was varied from 0.50 to 0.80; CT-CAFT coverage holds up to $\kappa_c = 0.70$ with $\leq 1\%$ additional clean-utility cost.

A.8 Supplementary Defense Comparison

Table 9 separates the defense components and compares their harm–utility trade-offs.

Table 9: Complete CT-CAFT defense comparison at Cora/GNN-as-Judge (five-seed means). Baseline adaptations and metric definitions follow Appendix A.2. Full CT-CAFT exceeds the strongest evaluated implementation control, margin regularization, by 16.7 pp in harm reduction (descriptive five-seed paired-bootstrap CI: [14.0, 19.4] pp).

Defense	Harm Red. (%)	Min. Coverage	ΔAcc (pp)
Confidence filter (impl. control)	22.3	N/A	−8.3
Graph purification (impl. control)	34.9	N/A	−4.1
Closed-loop corr. (impl. control)	38.5	N/A	−3.2
Generic curvature (impl. control)	40.1	N/A	−1.9
Similarity control	40.9	N/A	−1.6
Margin control	43.3	N/A	−1.4
w/o surplus	45.1	0.65	−1.1
w/o tube	52.4	0.65	−1.6
full CT-CAFT	**60.0**	**0.65**	−0.7

Full CT-CAFT combines the largest harm reduction with the smallest clean-accuracy loss; each ablation weakens that trade-off.

A.9 Supplementary MC-CMCT Comparison

Table 10 tests whether the diagnostic remains informative across different construction methods.

Table 10: Complete MC-CMCT diagnostics including non-invariant baselines (five-seed means). For FC-CEFA, $\bar{r}_A^{(2)}$ is reported at retained closure-passing points; AUROC uses all held-out directions. Table 12 reports direction-group intervals.

Method	Variant	μ_{AX}	$\bar{r}_A^{(2)}$	Pred./Act. Iters	Corr. AUROC	$\mathcal{S}_{\text{MC}}$	Spearman ρ
FC-CEFA text-origin	text-origin	0.00	10^{-15}	0.8/1.1	0.87	0.00	N/A
FC-CEFA coupled	graph-corr.	0.77	10^{-15}	1.1/1.4	0.89	0.18	0.78
FC-CEFA adaptive	graph-corr.	0.82	10^{-15}	1.0/1.2	0.92	0.24	0.81
Witches' Brew	non-invar.	0.34	10^{-4}	–/fail	0.51	–	0.34
Riemannian	non-invar.	0.41	10^{-6}	–/fail	0.48	–	0.28
Co-Attack adaptation (Zhang et al., 2022)	non-invar.	0.28	10^{-5}	–/fail	0.43	–	0.19

FC-CEFA convergence AUROCs exceed those of the listed non-invariant methods; the zero text-origin surplus carries no rank correlation.

A.10 Full Dataset–Victim Results and Statistical Intervals

Table 11 checks whether pooled summaries conceal failures in individual dataset–victim settings.

Table 11: Full dataset–victim results (five-seed means). Every row passes the exact-discrete, 10^{-6}-continuous gate; Worst IT/PT is the minimum over seeds and classes, and all ten settings are feasible on 5/5 seeds.

Dataset	Victim	Gate	Ret. (%)	IFH	CT-CAFT Harm Red. (%)	Worst IT/PT	ΔAcc (pp)
Cora	GNN-as-Judge	Pass	90.0	0.53	60.0	0.6501/0.6520	−0.7
Cora	Co-Teaching	Pass	87.2	0.50	58.1	0.6500/0.6518	−1.1
CiteSeer	GNN-as-Judge	Pass	86.8	0.51	57.5	0.6500/0.6513	−0.9
CiteSeer	Co-Teaching	Pass	84.5	0.47	55.2	0.6500/0.6510	−1.4
PubMed	GNN-as-Judge	Pass	85.1	0.48	56.0	0.6500/0.6508	−1.0
PubMed	Co-Teaching	Pass	82.3	0.44	53.8	0.6500/0.6506	−1.5
ogbn-arxiv	GNN-as-Judge	Pass	80.7	0.43	54.2	0.6500/0.6504	−1.2
ogbn-arxiv	Co-Teaching	Pass	78.2	0.40	52.7	0.6500/0.6502	−1.7
ogbn-prod.	GNN-as-Judge	Pass	77.5	0.39	53.1	0.6500/0.6501	−1.5
ogbn-prod.	Co-Teaching	Pass	73.8	0.36	52.2	0.6500/0.6500	−1.9

All ten settings retain positive IFH and defense benefit, with the reported class-wise coverage meeting the prescribed minimum.

Table 12 complements these point estimates with uncertainty over held-out direction groups.

Table 12: Direction-group bootstrap statistics for MC-CMCT (five-seed means). AUROC and Spearman ρ use held-out direction groups; $\mathrm{IFH}(v)$ is recomputed from each direction's own gate-passing realization. Brackets are 95% group-bootstrap CIs; $N_{\mathrm{dir}}/N_{\mathrm{grp}}$ count directions/groups.

Method	N_{dir}	N_{grp}	**Corr. AUROC**	**Spearman ρ**	**Iters MAE**
FC-CEFA text-origin	840	42	0.87 [0.81, 0.91]	N/A	0.31
FC-CEFA coupled	1250	45	0.89 [0.84, 0.93]	0.78 [0.71, 0.84]	0.28
FC-CEFA adaptive	1380	48	0.92 [0.87, 0.95]	0.81 [0.75, 0.86]	0.22

The coupled and adaptive surplus correlations remain positive throughout their reported intervals, supporting their use as exposure diagnostics.

Together, Tables 9–12 connect the main findings to individual settings and diagnostic uncertainty.