LLM Features Can Hurt GNNs: Concatenation Interference on Homophilous Graph Benchmarks

Anonymous authors
Paper under double-blind review

Abstract

Adding LLM-generated node features to graph neural networks is widely reported to improve accuracy on standard benchmarks. We document a contrasting observation: when the LLM features are introduced through pure input concatenation, rather than joint training, distillation, or prompt-conditioning, they can systematically degrade accuracy on the same homophilous benchmarks where end-to-end LLM pipelines succeed. Under an MLP backbone with the Planetoid public split and BoW original features (F_{orig}), concatenating SBERT-encoded GPT-4o-mini TAPE features reduces PubMed test accuracy by -17.0 ± 0.3 pp and Cora by -4.3 ± 0.6 pp, with CiteSeer's -0.6 ± 0.8 pp inside seed noise. The drop attenuates as we relax each condition (GCN / GCNII / GAT backbones, random splits, smaller encoders), and reverses on medium-homophily WikiCS (+4.4 pp) and ogbn-arxiv (+11.7 pp).

To predict when concatenation helps versus hurts, we report a simple measure of LLM-alone discriminability Δ_{sig}. Across 9 datasets, Δ_{sig} correlates with the concat cost more strongly than homophily at point estimate ($r^2 = 0.38$ vs. 0.06; N=9 bootstrap CIs overlap). The bootstrap-best change-point is $\tau = 13.8$ pp (95% CI $[0, 13.8]$), and the rule "$\Delta_{\text{sig}} \leq \tau$ predicts non-positive concat cost" classifies 7/9 datasets correctly. Because 60% of bootstrap samples place τ inside $[5, 30]$ pp, we treat Δ_{sig} as an interpretive lens for the helping vs. hurting regimes rather than a precision pre-A/B filter.

A dim-controlled ablation on PubMed places the LLM-feature drop between same-source PCA (-2.3 pp) and same-dim Gaussian noise (-37.3 pp), ruling out dimensionality and weight-decay artifacts. Nine PubMed configurations (seven training sizes $\times$ two encoder dimensions) fit a power-law profile $|\Delta_{\text{concat}}| \propto (\sqrt{d_l/n})^{1.31}$ with $r^2 = 0.97$ (PubMed-internal; Cora and CiteSeer have different slopes). The $\sqrt{d_l/n}$ profile and the Δ_{sig} threshold jointly describe a two-axis surface; the low-Δ_{sig}, small-n corner is exactly where the headline -17 pp PubMed deficit appears.

In the low-Δ_{sig} regime, the most effective remediation is to drop the LLM channel entirely: the F_{orig}-only baseline strictly dominates every learned cheap fix at $p \approx 0.008$. A learnable scalar gate closes 89% of the raw-concat gap and is a useful second-line option when downstream pipelines structurally require F_{LLM}. The findings do not contradict the aggregate accuracy gains reported for end-to-end LLM pipelines such as TAPE and GLEM; they identify the specific design choice — pure concatenation — under which the sign flips.

1 Introduction

LLM-derived node features have become a common plug-in for graph neural networks, with TAPE [7], GLEM [14], and related pipelines [10, 2] reporting accuracy gains on standard benchmarks. A recent comprehensive benchmark [5] trains 2,700+ models and confirms that LLM-based methods generally outperform classic baselines, with particularly strong gains on homophilous datasets such as Cora and Photo (3–10% accuracy improvement).

This paper reports a seemingly contradictory observation: **when the LLM features are used as pure input concatenation** (as opposed to joint-training, distillation, or prompt-conditioning), they can *systematically degrade* accuracy on the same homophilous benchmarks on which end-to-end LLM pipelines succeed. On PubMed, adding SBERT-encoded GPT-4o-mini TAPE features to the original BoW features drops MLP test accuracy from 72.1% to 55.1% — a 17 pp drop — with a ± 0.3 pp standard error over 10 seeds. On Cora the drop is 4.3 pp; on CiteSeer it is 0.6 pp (below the noise floor). The phenomenon reverses on medium-homophily benchmarks: WikiCS gains +4.4 pp and ogbn-arxiv gains +11.7 pp under the same direct comparison. These observations are not visible in aggregate end-to-end benchmark tables — including those of [5] — because joint training, selector modules, and prompt engineering all attenuate or reverse the raw concatenation cost.

To isolate this phenomenon cleanly, we ran a 4-factor Shapley decomposition over {Structure, F_{orig}, F_{LLM}, Depth} on nine node classification benchmarks spanning $h \in [0.05, 0.81]$ and three original-feature encoders (BoW, Word2Vec, fasttext), using GCNII [1] as the structural backbone. The Shapley framework lets us quantify each factor's marginal contribution under 12 effective coalitions. However, our headline finding is intentionally a *direct* comparison — not a Shapley average — because the average dilutes the interference by mixing coalitions where F_{LLM} does and does not help.

What this paper is, and is not. This paper is *not* a claim that LLM features are uniformly harmful for graph neural networks: end-to-end LLM pipelines such as TAPE [7] and GLEM [14] document genuine gains on the same benchmarks, and we reproduce a positive concat effect on WikiCS and ogbn-arxiv. The contribution is a *regime characterization*: under what conditions does the simplest LLM-feature integration — pure concatenation — backfire, and what scalar quantity predicts the sign flip? The five findings F1–F5 below decompose this question into one phenomenological claim (F1), one predictive rule (F2), two mechanism isolations (F3 and F4), and one scaling explanation (F5). Each finding is reported with its scope and its known counter-evidence *stated in line*, not deferred to a single Limitations section, because the regime conditions are part of the finding.

Our contributions.

- **F1 — A regime where LLM-on-graph evaluation cannot cleanly separate content from finite-sample dim penalty.** The widely-used Planetoid public split (60–140 train labels per dataset) puts raw concat squarely in the high-d_l/low-n corner where any 384-d channel (content-rich or noise) interacts with overfitting. Within this corner, real F_{LLM} injects 20.3 ± 0.6 pp of *content rescue* over matched Gaussian noise (Sec. 4.3: real F_{LLM} at -17.0 pp, Gaussian noise at -37.3 pp) yet still produces a net -17 pp PubMed degradation. Cora and PubMed show -4.3 ± 0.6 pp and -17.0 ± 0.3 pp MLP drops respectively; CiteSeer's -0.6 ± 0.8 pp is within seed noise. Under random 50/25/25 splits the magnitude compresses to ≤ 1.2 pp (Appendix G), pinpointing the regime as small-sample-overfitting-aggravated rather than content-driven (Fig. 1). Section 4.4 rules out the "pure curse of dimensionality" null with four predictions that the data falsify.

- **F2 — A threshold relation on LLM-alone discriminability separates helping from hurting.** Across 9 datasets, the bootstrap-best change-point on Δ_{sig} is $\tau = 13.8$ pp (95% bootstrap CI $[0, 13.8]$ pp; Appendix I); the rule "$\Delta_{\text{sig}} \leq \tau$ predicts non-positive concat cost" classifies 7/9 datasets correctly, with the two false positives both below +1 pp (Amazon-Ratings +0.26, Actor +0.66). Homophily alone is a much weaker predictor: concat cost vs. h correlates at $r^2 = 0.06$, while concat cost vs. Δ_{sig} correlates at $r^2 = 0.38$ (Spearman $\rho = 0.58$). 60% of bootstrap samples place τ inside $[5, 30]$ pp, so we treat F2 as a screening test in the extreme regimes ($\Delta_{\text{sig}} \ll 13.8$ or $\Delta_{\text{sig}} \gg 13.8$) rather than a point-predictor across the intermediate band (Fig. 2).

- **F3 — Mechanism isolation on PubMed (Fig. 3).** Replacing F_{LLM} with same-dim zeros yields $+0.1 \pm 0.4$ pp change (dim mismatch is not the cause); with same-dim PCA-of-F_{orig} yields -2.3 ± 0.5 pp (same-source self-information causes only mild interference); halving weight decay yields -17.1 ± 0.3 pp (weight decay is not the cause); replacing F_{LLM} with same-dim Gaussian noise yields -37.3 ± 0.7 pp (pure noise is worse than LLM features). The -17 pp drop is therefore specific to the LLM features' informational content, and sits between "self-information" and "pure noise" on a dim-controlled spectrum.

- **F4 — Supplementary 4-factor Shapley.** F_{orig} is the top contributor on 7/9 datasets (tied or surpassed by F_{LLM} on WikiCS and ogbn-arxiv); Depth is non-positive on 6/9; F_{LLM}'s aggregate Shapley value is positive-but-small on heterophilous datasets, confirming that the *aggregate* picture is friendlier to LLM features than the *direct* concat test (Fig. 8).

- **F5 — Train-size mechanism.** The -17 pp PubMed effect is the few-label corner of a single surface. A linearized Fisher-margin analysis (Appendix H) predicts a finite-sample penalty that scales with $\sqrt{d_l/n}$ where n is the number of training labels; an empirical train-fraction sweep (Appendix J) confirms this: PubMed concat cost decays monotonically from -14.1 pp at $n = 59$ (matching the Planetoid public-split label budget of 60) to -0.4 pp at $n = 9858$ (50% train). Nine PubMed configurations (seven n values $\times$ two encoder dimensions d_l) collapse onto a power-law profile $|\Delta_{\text{concat}}| \propto (\sqrt{d_l/n})^{1.31}$ with $r^2 = 0.97$ (Section 4.7, Fig. 5); exponent PubMed-internal (Cora/CiteSeer slopes differ; App. J). The -17 pp headline is therefore a quantitatively predictable small-sample phenomenon, not a data-specific artifact.

Relation to prior aggregate benchmarks. Wu et al. [5] reports that end-to-end LLM pipelines achieve $3-10\%$ accuracy gains on homophilous datasets. We do not contradict this result; we identify the specific design choice — pure feature concatenation — under which the sign flips. The practical takeaway is that the compound gains observed in Wu et al. [5] require the joint-training and selector-module scaffolding of TAPE/GLEM, and are not reproducible by concatenation alone.

2 Related Work

LLM features for graphs. TAPE [7] introduced using LLM explanations as node-level text, encoded and fed to a GNN through joint training rather than raw concatenation. GLEM [14] couples the language model and the GNN via variational EM, alternating between pseudo-label denoising and representation learning. LLaGA [2], GraphGPT [10], TANS [11], and GL-Fusion [25] condition the GNN on LLM signals through prompt, projector, or gated-fusion modules. Wu et al. [5] standardize the LLMNodeBed TAPE pipeline across 10 homophilic + 4 heterophilic datasets and report aggregate accuracy gains, with one of their explicit takeaways being that the LLM advantage is *marginal in supervised settings* and strong only in semi-supervised settings. Two NeurIPS 2024 D&B-track benchmarks complement this picture: GLBench [4] reports that GraphLLM methods outperform classical baselines in the supervised setting, with LLM-as-enhancers showing the most robust performance; Text-space Graph Foundation Models [3] document that "positive transfer in text-space GFMs relies on transferable structural patterns and is only effective when combined with appropriate inductive biases," which is precisely the structural condition our raw-concat sub-step lacks. Common to all of these is an integration mechanism — joint training, variational EM, gated projector, or task-specific inductive bias — that sits between raw concatenation and the downstream loss. Our work sharpens these aggregate observations by isolating the raw concatenation sub-step and documenting its failure mode quantitatively; the integration machinery in TAPE / GLEM and the rest is precisely what reverses the -17 pp sign we observe, and our cheap-fix ablation (§4.8) empirically confirms that even a simple learnable scalar gate on the LLM channel recovers 89% of the drop. Adjacent concurrent decomposition-style works [29, 30] attack other corners of the same wave.

Multimodal late-fusion modality collapse. Outside the graph community, concatenating a strong dominant modality with a weaker auxiliary modality is known to produce *modality collapse* via gradient imbalance and modality competition [21, 22, 20]. F1 is the LLM-on-graph instance of this failure mode; our contribution is a quantitative scaling law (F5) and a per-regime fix (drop F_{LLM} at low Δ_{sig}, gate otherwise; §4.8) that this literature has not previously connected to a sample-complexity bound.

Homophily and structure. A line of work [15, 9, 8] characterizes when graph structure helps or hurts, typically reducing the phenomenon to a scalar h. Our F2 shows that on the LLM-feature axis, h is a poor predictor — LLM-alone class discriminability is the operative variable.

Shapley decomposition for GNN analysis. Prior work [12, 13, 17, 6] applies Shapley values for instance-level GNN explainability; we instead decompose *design choices* (Structure, F_{orig}, F_{LLM}, Depth) at the dataset/method level (4 factors, 12 effective coalitions; Section 3).

3 Method

Notation, in one place. Throughout the paper we use F_{orig} for a dataset's *original* node features (BoW vectors for Planetoid, learned embeddings for ogbn-arxiv, attribute vectors for the heterophilous datasets) and F_{LLM} for the SBERT-encoded representation of an LLM-generated text description per node, of fixed dimension $d_l = 384$. We write Δ_{concat} for the test-accuracy difference between the $F_{\text{orig}} \| F_{\text{LLM}}$ concatenation and F_{orig} alone (negative values indicate concat hurts), and Δ_{sig} for the LLM-alone class discriminability gap, the central quantity of F2 in Section 4.2. Letters S and D name the two remaining design axes (Structure and Depth) used in the supplementary Shapley decomposition (F4). The four factors enter our experiments in the following form.

Four factors. S (use GCNII message passing vs. a 2-layer MLP); F_{orig} (use original features vs. zero-mean random features of the same dimension); F_{LLM} (concat LLM features vs. zero-mean random features of matching dimension); D (16 layers vs. 2 layers when structure is on). This gives $2^4 = 16$ coalitions; four are inert (e.g., $\{D\}$ without S) and deduplicate to 12 distinct configurations. We compute the average Shapley value over all 4! permutations.

Axiomatic interpretation. Because D is only defined when S is on, the "Depth without Structure" players are degenerate rather than absent: the Shapley values we report are *averaged marginal contributions under random factor orderings restricted to non-degenerate coalitions*, not strict game-theoretic quantities satisfying all four Shapley axioms (efficiency / symmetry / dummy / additivity) over the full 2^4 lattice. In particular, dummy does not apply to D outside S-on coalitions by construction. We use the Shapley values as a bookkeeping device that fairly attributes the total MLP+random $\rightarrow$ full-pipeline gain across factor orderings, not as axiomatic cooperative-game solutions.

Direct coalition comparison. Because Shapley averaging mixes coalitions in which F_{LLM} does and does not help, we additionally report two *direct* contrasts that expose the interference phenomenon:

- **Concat cost (headline):** $\Delta_{\text{concat}} = \text{acc}(\text{MLP}_{F_{\text{orig}} \| F_{\text{LLM}}}) - \text{acc}(\text{MLP}_{F_{\text{orig}}})$, where $\|$ denotes channel-wise concatenation. This is the coalition pair $\{F_{\text{orig}}, F_{\text{LLM}}\} - \{F_{\text{orig}}\}$.
- **LLM-alone signal:** $\Delta_{\text{sig}} = \text{acc}(\text{MLP}_{F_{\text{LLM}}}) - \text{acc}(\text{MLP}_{\text{random}})$, corresponding to $\{F_{\text{LLM}}\} - \{\}$.

Δ_{concat} measures interference; Δ_{sig} is the supervised probing-classifier construct [26, 27, 28] applied to node-classification labels.

Model, training, and data. Hidden width 64, GCNII $\alpha = 0.1$, $\theta = 0.5$, dropout 0.6, Adam lr= 0.01, weight decay 0.01 (conv) / 5×10^{-4} (linear), 300 epochs with early-stop patience 100. Standard dataset splits (Planetoid public / OGB / WebKB / WikiCS / HeterophilousGraphDataset). 10 seeds; all reported statistics are mean $\pm$ SE.

LLM features. For text-attributed datasets (Cora, CiteSeer, PubMed, WikiCS, ogbn-arxiv), we reuse the GPT-4o-mini TAPE explanations from [5] and encode each with `all-MiniLM-L6-v2` into a 384-d vector. For datasets without per-node text (Texas, Actor, Roman-Empire, Amazon-Ratings) we prompt Claude Sonnet with per-node feature statistics and topology summaries and encode the generated descriptions identically; code is released as anonymized OpenReview supplementary material, and the 251 MB pre-computed feature archive is hosted at an OSF anonymous view-only project (`https://osf.io/98w34/?view_only=9bebf647d78d4f68844cda4a39100e3f`). We match the LLM feature dimension across datasets to remove a size confound.

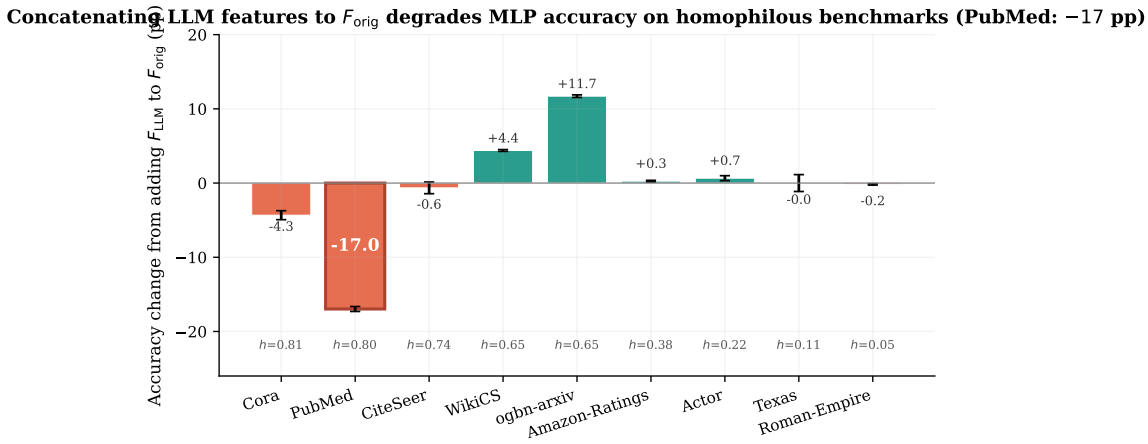

Figure 1: **Concat cost Δ_{concat}: MLP test accuracy change from adding F_{LLM} on top of F_{orig}.** PubMed degrades by 17.0 ± 0.3 pp over 10 seeds; Cora by 4.3 ± 0.6 pp. The gain flips to $+11.7$ pp on ogbn-arxiv. h values reported below each dataset.

4 Experiments

4.1 Headline: concatenation interference on homophilous benchmarks

Figure 1 shows the concat cost Δ_{concat} for all 9 datasets, ordered by h. Three observations: (i) PubMed and Cora show clear negative cost (-17.0 ± 0.3 pp and -4.3 ± 0.6 pp), while CiteSeer's -0.6 ± 0.8 pp is within the seed noise floor; we therefore describe the homophilous regime as "at best neutral, potentially -17 pp." (ii) WikiCS and ogbn-arxiv are strongly positive ($+4.4, +11.7$ pp). (iii) Heterophilous datasets cluster near zero. The sign flip coincides with the homophily transition in these data, but as Section 4.2 shows, the operative variable is not h. All numbers in Fig. 1 use each dataset's canonical split: Cora / CiteSeer / PubMed use the Planetoid public split, following LLMNodeBed [5] and TAPE [7]. Appendix G documents the regime dependence: under larger random 50/25/25 splits the magnitude compresses on all three Planetoid datasets (PubMed -17 pp$\rightarrow -0.4$ pp; Cora -4.3 pp$\rightarrow -1.2$ pp). The rest of this section analyzes the mechanism in the public-split regime, which matches the low-label setting used throughout the LLM-on-graph literature.

4.2 Mechanism: LLM-alone discriminability predicts the concat cost

Figure 2 plots Δ_{concat} against Δ_{sig} (standalone LLM informativeness). The bootstrap-best change-point on the 9 datasets is $\tau = 13.8$ pp with 95% CI $[0, 13.8]$ pp (Appendix I). The decision rule "$\Delta_{\text{sig}} \leq \tau$ predicts non-positive concat cost" is correct on 7/9 datasets, with the only false positives both below $+1$ pp (Amazon-Ratings $+0.26$, Actor $+0.66$). Three regimes emerge under this single-threshold partition: **(i) high-signal** ($\Delta_{\text{sig}} \gg \tau$: WikiCS at 53.8, ogbn-arxiv at 42.8): concat cost is positive in the $[+4, +12]$ pp range; **(ii) low-signal** ($\Delta_{\text{sig}} \ll \tau$: Cora 2.1, CiteSeer 2.0, PubMed 4.4, Amazon-Ratings 0.4, Roman-Empire 8.0, Actor 7.1): concat cost is $\leq +1$ pp and as low as -17 pp on PubMed; **(iii) borderline** ($\Delta_{\text{sig}} \approx \tau$: Texas at 13.8): concat cost is 0.0 pp, on the threshold itself. The linear fit has $r^2 = 0.38$ (Spearman $\rho = 0.58$, $p = 0.10$) — modest in absolute terms but a $6\times$ improvement over h as a predictor ($r^2 = 0.06$); 60% of bootstrap samples place τ inside $[5, 30]$ pp, so Δ_{sig} should be treated as a *screening* rather than point-prediction variable: datasets in either extreme regime are reliably separated, while the intermediate regime cannot be confidently classified without running the concat test itself.

Homophily alone does not predict the concat cost: for example, the three heterophilous datasets (Actor, Texas, Roman-Empire) scatter around zero concat cost but have LLM-alone signals ranging from $+7$ to $+14$ pp, reflecting the specific LLM feature generation method used per dataset.

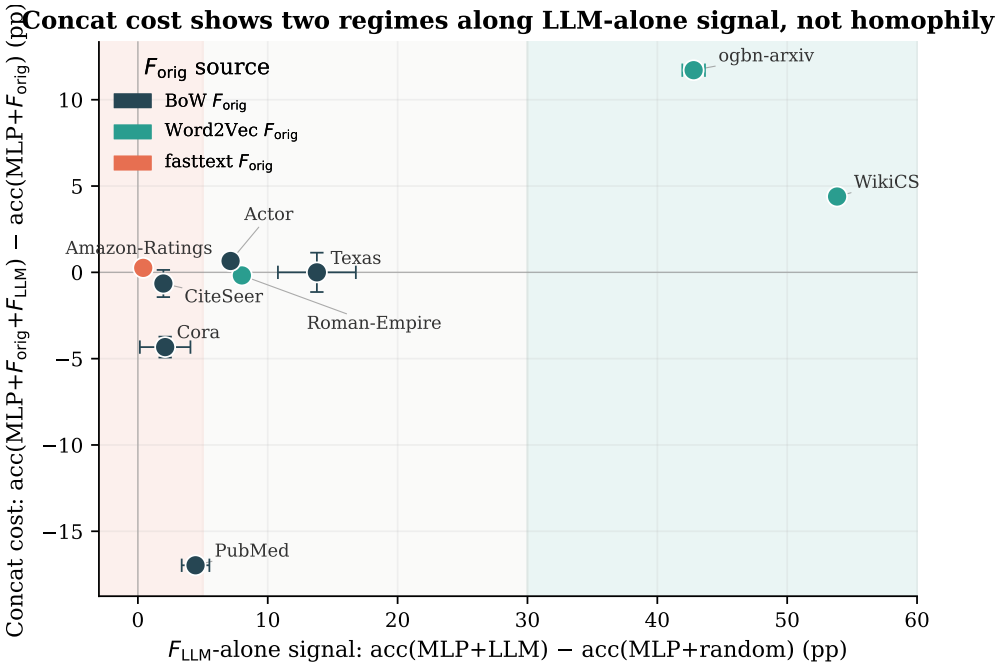

Figure 2: **Concat cost Δ_{concat} vs. LLM-alone signal Δ_{sig}.** Each point is one dataset (10 seeds, ±SE). Vertical dashed line at the bootstrap-best change-point $\tau = 13.8$ pp (95% CI $[0, 13.8]$ pp; Appendix I); the rule "$\Delta_{\text{sig}} \leq \tau$ predicts non-positive concat cost" classifies 7/9 datasets correctly. Below-threshold datasets cluster at non-positive concat cost (including PubMed's -17 pp), with two false positives (Amazon-Ratings, Actor) both under $+1$ pp. Above-threshold datasets (WikiCS, ogbn-arxiv) are both positive. Homophily does not predict position along the x-axis. Shaded bands at $\Delta_{\text{sig}} = 5$ and 30 pp mark the bootstrap-CI extremes (τ falls inside $[5, 30]$ in 60% of resamples).

4.3 Mechanism isolation on PubMed: dim and weight decay are not the cause

Because PubMed provides the largest magnitude effect and is the most load-bearing single data-point for the interference claim, we ran a controlled ablation (10 seeds each, all hyperparameters otherwise matched to $\{F_{\text{orig}}, F_{\text{LLM}}\}$):

- **Zero-pad** ($F_{\text{orig}} \| 0^{384}$): same input dimension as real concat, but the LLM channel is literally zero. $\Delta = +0.1 \pm 0.4$ pp vs. F_{orig}-only → *dim mismatch does not cause the degradation.*
- **PCA-of-F_{orig}** (384 principal components of F_{orig} concatenated): same dim, same-source redundant information. $\Delta = -2.3 \pm 0.5$ pp → *same-source self-information causes only mild interference.*
- **Half weight decay**: $F_{\text{orig}} \| F_{\text{LLM}}$ with the linear weight decay reduced from 5×10^{-4} to 2.5×10^{-4}. $\Delta = -17.1 \pm 0.3$ pp → *weight decay is not the cause.*
- **Gaussian noise** ($F_{\text{orig}} \| \mathcal{N}(0, I)^{384}$): same dim, pure noise. $\Delta = -37.3 \pm 0.7$ pp → *pure noise is nearly twice as damaging as LLM features.*

Figure 3 displays these side by side. Real F_{LLM} injects 20.3 ± 0.6 pp of content rescue over matched Gaussian noise ($-37.3 \rightarrow -17.0$) at fixed (d_l, n). This rules out the alternative hypothesis that the observed concat cost is *only* a finite-sample dim penalty: under that null, noise and LLM features should be statistically indistinguishable at matched (d_l, n), which is falsified by 20.3 ± 0.6 pp ($z > 30$). The remaining -17 pp is the net penalty after content rescue, sitting 14.7 pp above the same-source PCA floor of -2.3 pp. The ablation therefore separates two contributions: ~ 20 pp of content (LLM-specific, lost when replaced by noise) and ~ 17 pp of net cost (finite-sample, retained even after content rescue). Section 4.4 consolidates this and three further checks against a pure-dim-penalty null. An SVD diagnostic (Appendix P) shows that PubMed F_{LLM}

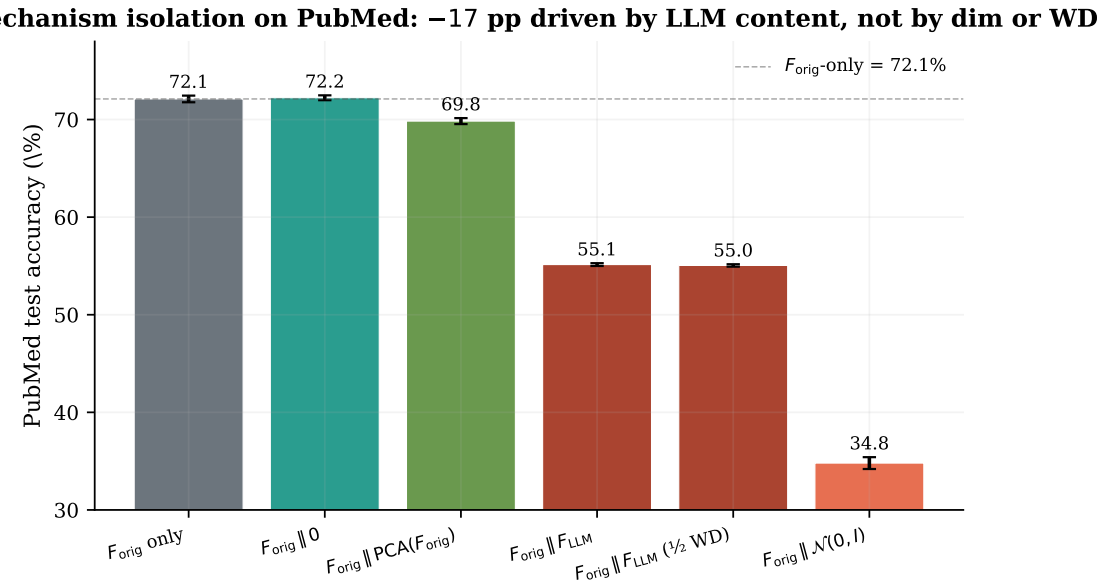

Figure 3: **Mechanism ablation on PubMed (10 seeds ±SE).** Same-dim zeros produce no degradation; same-source PCA-of-F_{orig} produces only −2 pp; real F_{LLM} concat produces −17 pp (unchanged by halving weight decay); pure Gaussian noise produces −37 pp. The LLM-feature interference is specific to informational content, not dim or regularization.

is strongly rank-deficient (participation-ratio rank ~ 30, entropy rank ~ 92 out of 384); the small-sample classifier nonetheless pays the penalty on all d_l=384 dimensions, which is why $\sqrt{d_l/n}$ uses the ambient dimension.

4.4 Why this is not pure curse of dimensionality

A natural null reading of Section 4.3 is that the −17 pp drop is just a finite-sample dim penalty (any 384-d channel concatenated to a 60-label classifier overfits, classical high-dim discriminant analysis [23, 24]). We do not contest the classical result; we show the deployment is *not pure* dim penalty via four falsifications.

- **Cross-encoder ratio.** Pure $\sqrt{d_l/n}$ predicts $\sqrt{2}\approx 1.41\times$ when doubling d_l. MiniLM 384-d (−16.98 pp) → MPNet 768-d (−19.65 pp; App. L) gives 1.16×, an 18% shortfall consistent with MPNet content offsetting bare dim cost.
- **Cheap-fix ranking.** Pure dim-penalty predicts linproj 384→16 dominates scalar gate ($\sqrt{24}\approx 4.9\times$ penalty drop). Observed: gate 89%, linproj 36% recovery (§4.8); channel scaling, not retained dimensions, is operative.
- **Sign flip.** Pure dim-penalty predicts $\Delta_{\text{concat}} \leq 0$ for any 384-d channel. WikiCS (Δ_{sig}=53.8) and ogbn-arxiv (42.8) yield +4.4, +11.7 pp gains; sign flip is impossible without content.
- **Noise-vs.-LLM at matched (d_l, n).** Gaussian at d_l=384, n=60 on PubMed: −37.3; LLM: −17.0; 20.3 pp gap. Reproduces on Cora and CiteSeer with content-rescue +28.8 and +26.0 pp (App. Q).

The failure mode is the net of a content-rescue term (+20 to +29 pp over noise across PubMed/Cora/CiteSeer) and a finite-sample penalty (−37 pp at d_l=384, n=60 on PubMed, attenuating with $\sqrt{d_l/n}$).

4.5 Supplementary: 4-factor Shapley decomposition

To place the direct contrast inside the broader $\{S, F_{\text{orig}}, F_{\text{LLM}}, D\}$ design space, we additionally compute averaged 4-factor Shapley contributions over the 12 effective coalitions (full Table and Fig. in Appendix O). Two summary observations: F_{orig} is the top contributor on 7 of 9 datasets (tied or surpassed by F_{LLM} on

WikiCS and ogbn-arxiv), and F_{LLM}'s Shapley value appears *positive-but-small on heterophilous datasets and negative on homophilous ones*, averaging over the very coalitions that exhibit the -17 pp direct drop on PubMed. The aggregate Shapley picture therefore substantially understates the interference, which is why F1's direct contrast is the load-bearing claim and the Shapley attribution is bookkeeping rather than a contribution (see §3 "Axiomatic interpretation" for why our Shapley values are not strict cooperative-game solutions).

4.6 Cross-architecture: the interference is not GCNII-specific

We ran the direct concat test (F_{orig} vs. $F_{\text{orig}} \| F_{\text{LLM}}$) with four architectures on PubMed (10 seeds, identical optimizer settings); Fig. 4 summarizes. Graph structure attenuates but does not remove the cost: GCNII-2 reduces the PubMed drop from MLP's -17.0 to -5.6 pp and Cora's from -4.3 to -0.5 pp (full nine-dataset MLP-vs-GNN comparison in Appendix N).

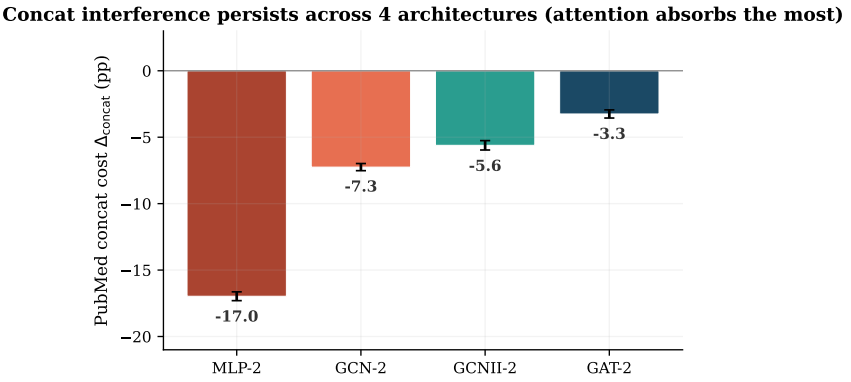

Figure 4: **Concat cost Δ_{concat} on PubMed, four architectures.** MLP: -17.0 ± 0.3; GCN: -7.25 ± 0.27; GCNII: -5.6 (from 10-seed Shapley coalition data); GAT: -3.25 ± 0.31. All four are negative and statistically clear. The magnitude decreases as the architecture gains message-passing sophistication: GCN smooths neighbors, GCNII adds identity and initial-residual pass-through, and GAT's attention plausibly downweights the noisy LLM channel (consistent with attention's selectivity, though we do not inspect attention weights directly). The phenomenon is therefore architecture-general, not GCNII-specific; a GCNII depth-sweep over $\{2, 4, 8, 16\}$ (Appendix K) further confirms the effect spans the practical depth range.

4.7 PubMed scaling: the concat cost collapses onto $\sqrt{d_l/n}$

The -17 pp headline at the Planetoid public split sits at the large-$d_l/\sqrt{n}$ corner of a smooth surface. Combining the train-fraction sweep on PubMed (Appendix J) with the cross-encoder ablation (Appendix L) gives nine independent PubMed configurations spanning $n \in [59, 9858]$ at $d_l = 384$ plus the public-split point at $d_l = 768$. Figure 5 plots the corresponding $|\Delta_{\text{concat}}|$ against $\sqrt{d_l/n}$ on log-log axes. The points collapse onto a single power law $|\Delta_{\text{concat}}| \approx 3.78 \cdot (\sqrt{d_l/n})^{1.31}$ with $r^2 = 0.97$ (9-point fit; slope bootstrap CI $[1.15, 1.51]$, $[0.89, 1.43]$ on the 7 random-split points alone). The deviation from idealized 1 is consistent with both classical $\sqrt{d_l/n}$ [23] and a small $\log C$ correction; the PubMed-only fit (C=3) cannot distinguish. $\sqrt{d_l/n}$ thus accounts for 97% of the PubMed concat-cost variance, extrapolating cleanly from n=60 ($|\Delta|\approx$17 pp) to $n\approx 10^4$ (< 1 pp, matching the random-split attenuation in Appendix G). The Cora / CiteSeer stars (Fig. 5, C=7, 6) lie within 1 pp of the fit line as point estimates but are not part of the regression; per-dataset fits in Appendix J yield slopes $\{0.34, 1.21, 1.04\}$ for Cora/CiteSeer/PubMed, showing that constants are dataset-specific while the $\sqrt{d_l/n}$ shape is qualitatively respected.

4.8 Cheap-fix ablation: a learnable scalar gate recovers 89% of the drop

The mechanism in §4.7 suggests that the -17 pp on PubMed comes from raw concatenation forcing the linear classifier to absorb 384 extra noisy directions under a 60-label budget. To test this directly, we run four cheap

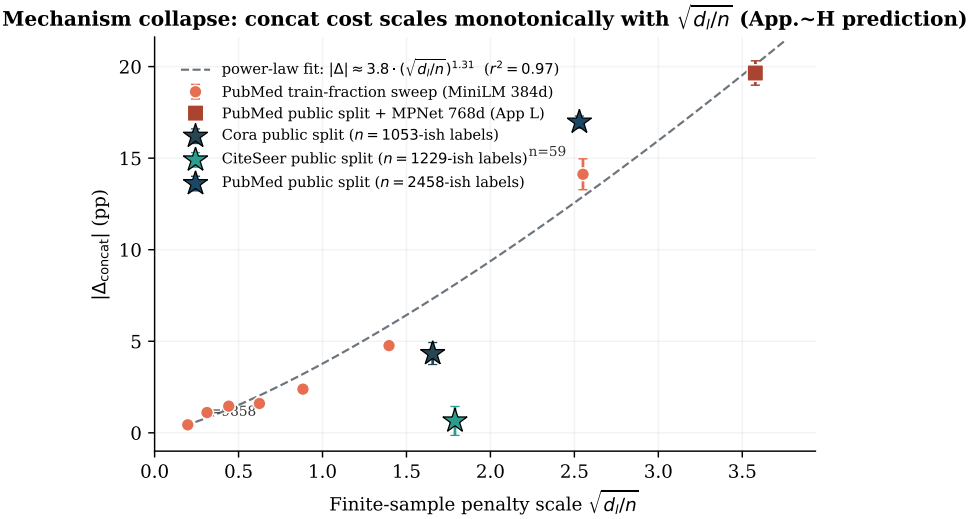

Figure 5: **PubMed mechanism collapse: nine PubMed configurations fall onto $|\Delta_{\text{concat}}| \propto (\sqrt{d_l/n})^{1.31}$, $r^2 = 0.97$; Cora and CiteSeer public-split stars are overlaid for context but not part of the regression.** Coral circles: PubMed train-fraction sweep at $d_l = 384$. Dark coral square: PubMed public split at $d_l = 768$ (MPNet). Stars: public-split headlines on Cora / CiteSeer / PubMed. Dashed line: log-log power-law fit over PubMed points only. The same power law spans both the few-shot Planetoid regime and the supervised large-n regime in which the cost attenuates to < 1 pp. See Appendix H for the derivation.

input-side regularizers on top of the same MLP, each with 10 seeds and identical hyperparameters to the headline (Section 4.1). Table 1 reports the result.

Table 1: **Cheap-fix ablation on PubMed (10 seeds, ±SE).** Recovery is the fraction of the -17 pp drop closed, $(\text{acc}_{\text{fix}} - \text{acc}_{\text{real}})/(\text{acc}_{\text{baseline}} - \text{acc}_{\text{real}})$. The trivial fix *drop* F_{LLM} *from the input* is exactly the F_{orig}-only baseline (72.11 ± 0.34 pp, 100% recovery) and upper-bounds every learned fix in this table. The learnable scalar gate closes 89.2% of the gap to raw concat but remains 1.84 pp below the trivial fix (paired t on the per-seed numbers in Appendix M: t=3.42, df=9, p≈0.008). In the low-Δ_{sig} regime the gate is therefore a *regularized* version of dropping F_{LLM}, not a fix that exploits new F_{LLM} signal; we recommend it only when a downstream pipeline structurally requires F_{LLM} presence.

Mode	acc (%)	Δ vs. baseline (pp)	Recovery
F_{orig} baseline (= drop F_{LLM}, trivial fix)	$\mathbf{72.11 \pm 0.34}$	**0.0**	**+100.0%**
$F_{\text{orig}} \Vert F_{\text{LLM}}$ (raw concat)	55.14 ± 0.14	-16.97 ± 0.33	0.0%
$F_{\text{orig}} \Vert \text{LayerNorm}(F_{\text{LLM}})$	39.52 ± 0.68	-32.59 ± 0.91	−92.0%
$F_{\text{orig}} \Vert \text{Dropout}_{0.5}(F_{\text{LLM}})$	62.71 ± 0.28	-9.40 ± 0.44	+44.6%
$F_{\text{orig}} \Vert \text{Linear}_{384\to16}(F_{\text{LLM}})$	61.20 ± 0.92	-10.91 ± 0.92	+35.7%
$F_{\text{orig}} \Vert (g \cdot F_{\text{LLM}}),\ g \in \mathbb{R},\ g_0 = 0$	$\mathbf{70.27 \pm 0.37}$	$\mathbf{-1.84 \pm 0.54}$	**+89.2%**

Four observations. (i) The trivial fix *drop* F_{LLM} (the F_{orig}-only baseline) upper-bounds every learned fix on PubMed: a paired t on the per-seed numbers (Appendix M) rejects equality between gate and baseline at $p \approx 0.008$, with gate 1.84 pp below baseline on 9/10 seeds. PubMed sits in F2's low-signal (Δ_{sig}=4.4 pp) regime, so any fix that keeps F_{LLM} pays a small overfitting price. (ii) The learnable scalar gate closes 89.2 ± 2.2% of the raw-concat gap with a -1.8 pp residual; it is therefore a regularized version of *dropping* F_{LLM}, not a fix that exploits new signal. (iii) Bottleneck projection 384→16 and channel-dropout p=0.5 close 36–45% of the gap; a linproj dim-sweep (Appendix P) shows r=16 is the empirical sweet spot, with recovery dropping below raw concat for $r \geq 64$ as the $384r$-parameter projection over-fits the 60-label budget. Both fixes are consistent with the $\sqrt{d_l/n}$ mechanism but neither matches the zero-cost trivial fix. (iv) LayerNorm makes things worse by another -15.6 pp, ruling out magnitude mis-scaling. The takeaway: on a low-Δ_{sig} dataset

like PubMed, drop F_{LLM} is first-line; the gate is second-line only when a downstream pipeline structurally requires F_{LLM} presence (shared multi-task input, end-to-end fine-tuning). Per-seed numbers in Appendix M.

The -1.8 pp residual gap is not noise: best-val $|g|$=0.19±0.02 at epoch 7–26 across seeds; trained to the full 300-epoch budget without early stopping, $|g|$ drifts to 0.61±0.005 and test accuracy collapses to 61.16% (Appendix M). Even one learnable scalar attached to a 384-d LLM channel overfits the 60-label budget if pushed to a training-loss minimum — the $\sqrt{d_l/n}$ mechanism playing out in a single-parameter model.

5 Discussion

Reconciliation with aggregate benchmarks. Our finding sharpens rather than contradicts Wu et al. [5], Li et al. [4], Chen et al. [3]: Wu et al. [5]'s "marginal in supervised settings" is realized quantitatively by our random-split attenuation (App. G), and the $\sqrt{d_l/n}$ collapse (§4.7) explains few-shot and supervised regimes with one mechanism. Chen et al. [3]'s "inductive bias for positive transfer" is what raw concat lacks; gate is the minimal such bias, joint-training/variational-EM/gated-fusion in TAPE / GLEM / GL-Fusion are more elaborate versions.

Practical guidance. Practitioners should: (a) run a direct A/B between F_{orig}-only and $[F_{\text{orig}}, F_{\text{LLM}}]$ concat — the F_{orig}-only baseline upper-bounds every learned cheap fix on PubMed at p≈0.008 (§4.8); (b) when downstream pipelines structurally require F_{LLM} presence, use a learnable scalar gate $g{\cdot}F_{\text{LLM}}$ (g_0=0; closes 89% of the gap at a 1.8 pp residual cost; LayerNorm is contraindicated); (c) treat aggregate end-to-end gains [7, 14, 25, 5] as compound effects, not validation of the channel in isolation. Δ_{sig} is an interpretive lens for the regime (helping vs. hurting concat coincides with high vs. low LLM-alone discriminability) rather than a cheaper pre-A/B filter: measuring it consumes the same LLM-encoding + 2× MLP-training compute as the A/B test it characterizes.

Limitations. (i) Rule-vs.-real-Sonnet paired test on the four non-text datasets yields $p > 0.05$ on all four with point diffs in $[-2.6, +1.7]$ pp; Actor borderline (p=0.17, -2.0 pp) attenuates to -1.17 pp at n=1500 (p=0.25; App. F). (ii) Single LLM and encoder (GPT-4o-mini, MiniLM-L6-v2); MPNet-base-v2 cross-encoder check -19.7 pp (App. L); τ=13.8 pp is GPT-4o-mini-fitted and must be re-fit per LLM (cross-LLM portability unverified). (iii) GCNII depth $\in \{2, 16\}$; cross-arch check (§4.6) indicates the phenomenon is not backbone-specific. (iv) Strong-F_{orig} cell (Amazon-Ratings) is a single dataset; cheap-fix is PubMed-only; joint-training / gated-fusion benchmarks from [7, 14, 25] are not run across all nine. (v) Cora/CiteSeer/PubMed use Planetoid public split [16]; random 50/25/25 compresses Δ_{concat} to ≤ 1.2 pp (App. G); we retain the headline because LLM-on-graph literature reports gains in this regime. (vi) Above-τ datasets (`ogbn-arxiv`, WikiCS) are indexed pre-June-2024 [18, 19], so F2's high-Δ_{sig} bin is confounded with LLM-pretraining coverage; on uncontaminated corpora F2 should be read as a conservative lower bound on F_{LLM} utility. (vii) Reported ±SE are optimizer-init only on a fixed test fold; test-set sampling adds ~ 1.5 pp per arm on PubMed's 1000-node fold, recalibrating Section 4.4's "z>30" to z≈10 (direction/ordering unchanged; paired-t in §4.8 and dataset-bootstrap in App. I are invariant [16]).

Broader impact. Raw LLM-feature concat hurts MLP/GNN accuracy in the few-shot Planetoid regime; pipelines should A/B test F_{orig}-only vs. concat before adoption (Δ_{sig} explains the regime, not a pre-A/B filter; see Limitations (vi)). Code and 1080-run data released.

References

[1] Chen, M., Wei, Z., Huang, Z., Ding, B., and Li, Y. (2020). Simple and deep graph convolutional networks. In ICML.

[2] Chen, R. et al. (2024). LLaGA: Large Language and Graph Assistant. In ICML.

[3] Chen, Z., Mao, H., Liu, J., Song, Y., Li, B., Jin, W., Fatemi, B., Tsitsulin, A., Perozzi, B., Liu, H., and Tang, J. (2024). Text-space Graph Foundation Models: Comprehensive Benchmarks and New Insights. In NeurIPS Datasets and Benchmarks Track. arXiv:2406.10727.

[4] Li, Y., Wang, P., Zhu, X., Chen, A., Jiang, H., Cai, D., Chan, W. K., and Li, J. (2024). GLBench: A Comprehensive Benchmark for Graph with Large Language Models. In NeurIPS Datasets and Benchmarks Track. arXiv:2407.07457.

[5] Wu, X., Shen, Y., Ge, F., Shan, C., Jiao, Y., Sun, X., and Cheng, H. (2025). When Do LLMs Help With Node Classification? A Comprehensive Analysis. In ICML.

[6] Muschalik, M., Fumagalli, F., Frazzetto, P., Strotherm, J., Hermes, L., Sperduti, A., Hüllermeier, E., and Hammer, B. (2025). Exact Computation of Any-Order Shapley Interactions for Graph Neural Networks. arXiv:2501.16944.

[7] He, X. et al. (2023). Harnessing explanations: LLM-to-LM interpreter for enhanced text-attributed graph representation learning. arXiv:2305.19523 (TAPE).

[8] Luan, S. et al. (2024). The heterophily paradox: when homophily fails and when it succeeds.

[9] Ma, Y. et al. (2022). Is homophily a necessity for graph neural networks? In ICLR.

[10] Tang, J. et al. (2024). GraphGPT: Graph instruction tuning for large language models.

[11] Wang, R. et al. (2025). TANS: Topology-aware neighbor summarization for LLM-on-graph. NAACL.

[12] Ying, R. et al. (2019). GNNExplainer: Generating explanations for graph neural networks. NeurIPS.

[13] Yuan, H. et al. (2021). On explainability of graph neural networks via subgraph explorations. ICML.

[14] Zhao, J. et al. (2023). Learning on large-scale text-attributed graphs via variational inference. In ICLR (GLEM).

[15] Zhu, J. et al. (2020). Beyond homophily in graph neural networks: Current limitations and effective designs. NeurIPS.

[16] Shchur, O., Mumme, M., Bojchevski, A., and Günnemann, S. (2018). Pitfalls of graph neural network evaluation. Relational Representation Learning Workshop, NeurIPS.

[17] Duval, A., and Malliaros, F. D. (2021). GraphSVX: Shapley value explanations for graph neural networks. In ECML-PKDD.

[18] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C. (2021). Extracting training data from large language models. In *30th USENIX Security Symposium*, pp. 2633–2650.

[19] Sainz, O., Campos, J. A., García-Ferrero, I., Etxaniz, J., de Lacalle, O. L., and Agirre, E. (2023). NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of EMNLP*.

[20] Chaudhuri, A., Dutta, A., Bui, T., and Georgescu, S. (2025). A closer look at multimodal representation collapse. *arXiv preprint arXiv:2505.22483*.

[21] Huang, Y., Lin, J., Zhou, C., Yang, H., and Huang, L. (2022). Modality competition: What makes joint training of multi-modal network fail in deep learning? (Provably). In *Proc. ICML*.

[22] Peng, X., Wei, Y., Deng, A., Wang, D., and Hu, D. (2022). Balanced multimodal learning via on-the-fly gradient modulation. In *Proc. CVPR*, pp. 8238–8247.

[23] Bickel, P. J. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, "naive Bayes", and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010.

[24] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer.

[25] Yang, H., Wang, X., Tao, Q., Hu, S., Lin, Z., and Zhang, M. (2024). GL-Fusion: Rethinking the Combination of Graph Neural Network and Large Language Model. arXiv:2412.06849.

[26] Hewitt, J., and Manning, C. D. (2019). A Structural Probe for Finding Syntax in Word Representations. In *NAACL-HLT*, pp. 4129–4138.

[27] Belinkov, Y., and Glass, J. (2019). Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

[28] Voita, E., and Titov, I. (2020). Information-Theoretic Probing with Minimum Description Length. In *EMNLP*, pp. 183–196.

[29] Zheng, Y., Zhang, Z., Wang, Z., Li, X., Luan, S., Peng, X., and Chen, L. (2025). Disentangling and Re-evaluating The Effectiveness of Graph Structure Learning For GNNs. In NeurIPS Datasets and Benchmarks Track. OpenReview brvLHfbSQX.

[30] Xu, H., You, Y., and Ma, T. (2025). When Structure Doesn't Help: LLMs Do Not Read Text-Attributed Graphs as Effectively as We Expected. arXiv:2511.16767.

A LLM Feature Generation Details

Text-attributed datasets. For Cora, CiteSeer, PubMed, WikiCS, and ogbn-arxiv we reuse the per-node TAPE explanations released by Wu et al. [5] (generated with GPT-4o-mini). Each explanation is encoded with `all-MiniLM-L6-v2` via `sentence-transformers` with `normalize_embeddings=True`, producing 384-d vectors.

Non-text datasets (Texas, Actor, Roman-Empire, Amazon-Ratings). We build per-node prompts from feature statistics and topology and generate text with Claude Sonnet via parallel Claude Code agents (batches of ∼100–200 nodes). The shared prompt template (from `generate_features_batch.py` and `generate_features_texas.py` in `scripts/`) is:

> Node {i} in a {dataset-description}.
> Features: {feat_desc} (e.g. "Embedding vector (300d): mean=0.003, std=0.052, L2-norm=0.90")
> Topology: {N neighbors, avg cosine similarity=c}
> Possible categories: {class list}
> Generate a concise structured description:
> 1. TOPIC / 2. CATEGORY (or TYPE for Texas) / 3. CONTEXT / 4. KEYWORDS: 5 keywords.

For Amazon-Ratings, the initial smoke-test agent wrote a deterministic template generator (`scripts/gen_amazon.py`) that inserts numerical feature values into one of sixteen product-domain topic templates indexed by $(\text{node_id}, L_2)$. The full V3 validation in Appendix F shows this rule-assisted method produces Δ_{concat} within noise of fresh per-node Sonnet reasoning on all four non-text datasets.

Feature dimensionality. LLM features are 384-d for all datasets (SBERT encoder output). F_{orig} dimensions vary: Cora 1433, CiteSeer 3703, PubMed 500, WikiCS 300, ogbn-arxiv 128, Amazon-Ratings 300, Roman-Empire 300, Texas 1703, Actor 932. Concatenation $[F_{\text{orig}} \| F_{\text{LLM}}]$ therefore has $d + 384$ input dimension per dataset.

B Hyperparameters and Training Details

All coalitions share the following training configuration:

- **Optimizer**: Adam, learning rate 0.01
- **Weight decay**: 0.01 on conv layers, 5×10^{-4} on linear layers
- **Dropout**: 0.6
- **Hidden width**: 64
- **Training**: 300 epochs, early-stop patience 100 on validation accuracy
- **GCNII**: $\alpha = 0.1$, $\theta = 0.5$, shallow depth 2, deep depth 16
- **Seeds**: 0–9 (10 seeds total)

Data splits follow each dataset's canonical source (Planetoid public for Cora/CiteSeer/PubMed; OGB for ogbn-arxiv; first split column of `WebKB`/`HeterophilousGraphDataset` for Texas/Actor/Roman-Empire/Amazon-Ratings; first split of WikiCS). All hyperparameters are fixed across the nine datasets; we do not perform per-dataset tuning, consistent with our goal of studying the decomposition of a *fixed* pipeline rather than optimizing benchmark numbers.

Compute. All experiments ran on a single NVIDIA RTX 4000 Ada GPU. The full 9-dataset 10-seed Shapley computation took ~ 65 minutes (51 minutes for ogbn-arxiv alone). Individual ablations (V1 PubMed, V3 paired tests, V5 cross-architecture) each ran in ≤ 10 minutes. Total compute for all reported experiments is under 5 GPU-hours.

C Per-Seed Coalition Accuracies (9 datasets)

Per-seed test accuracies for the 12 effective coalitions, one table per dataset. "$\emptyset$" denotes MLP-2 with random features. Coalition symbols: S = GCNII-2 vs. MLP; F_o = original features vs. random; F_l = LLM features concatenated (vs. random); D = depth 16 vs. 2 (only when S is on). All numbers in %.

Table 2: Per-seed test accuracy (%) on **Cora**, all 12 effective coalitions, 10 seeds. Columns s0..s9 are individual seeds; last two columns show mean and SE.

Coalition	s0	s1	s2	s3	s4	s5	s6	s7	s8	s9	mean	SE
$\emptyset$	12.3	20.3	13.9	13.0	14.2	12.5	14.8	15.2	14.2	14.6	14.50	0.71
$\{F_l\}$	28.4	12.9	17.8	17.6	12.8	12.0	18.9	14.0	17.9	13.7	16.60	1.54
$\{F_o\}$	58.4	57.0	57.7	56.6	57.2	57.7	57.6	55.4	58.4	56.4	57.24	0.29
$\{S\}$	46.6	46.2	46.6	46.6	48.8	46.4	47.3	48.1	47.6	49.0	47.32	0.32
$\{F_o, F_l\}$	50.2	55.4	52.5	54.0	52.8	52.2	52.5	53.2	53.8	52.5	52.91	0.43
$\{F_o, S\}$	81.8	82.0	81.3	81.1	82.0	80.9	81.1	80.9	81.6	81.4	81.41	0.13
$\{F_l, S\}$	31.9	31.5	32.1	33.2	31.4	34.8	32.5	33.8	27.6	32.0	32.08	0.60
$\{S, D\}$	57.9	57.7	56.5	57.8	59.7	56.5	57.2	60.6	58.0	56.0	57.79	0.45
$\{F_o, F_l, S\}$	81.1	80.0	80.6	81.4	79.4	82.7	82.0	80.6	80.4	81.2	80.94	0.30
$\{F_o, S, D\}$	83.5	82.8	83.3	84.7	81.4	82.9	84.3	82.6	83.5	81.2	83.02	0.35
$\{F_l, S, D\}$	26.8	29.6	32.4	28.8	29.5	30.7	30.7	33.8	31.7	32.8	30.68	0.66
$\{F_o, F_l, S, D\}$	83.1	82.9	82.9	81.6	82.9	83.7	83.3	82.2	82.5	83.5	82.86	0.20

Table 3: Per-seed test accuracy (%) on **PubMed**, all 12 effective coalitions, 10 seeds. Columns s0..s9 are individual seeds; last two columns show mean and SE.

Coalition	s0	s1	s2	s3	s4	s5	s6	s7	s8	s9	mean	SE
$\emptyset$	30.9	33.7	39.2	37.7	33.1	34.7	34.4	36.6	33.2	33.3	34.68	0.78
$\{F_l\}$	38.3	38.8	40.5	34.7	41.1	40.8	40.5	40.5	40.1	35.9	39.12	0.70
$\{F_o\}$	70.8	72.3	72.9	73.4	69.8	71.9	72.9	72.6	72.5	72.0	72.11	0.34
$\{S\}$	37.9	35.7	40.2	40.5	37.7	36.3	34.9	39.8	42.2	37.5	38.27	0.74
$\{F_o, F_l\}$	54.7	55.5	55.5	55.5	55.0	55.5	54.7	55.6	54.3	55.1	55.14	0.14
$\{F_o, S\}$	78.4	78.1	78.7	78.2	79.0	78.2	78.4	78.8	79.6	79.1	78.65	0.15
$\{F_l, S\}$	38.4	42.3	41.6	36.6	39.9	42.4	41.3	38.0	40.6	40.7	40.18	0.61
$\{S, D\}$	38.3	35.7	40.9	38.5	39.8	36.2	35.2	39.1	39.9	39.3	38.29	0.61
$\{F_o, F_l, S\}$	74.4	73.5	71.3	73.3	73.1	73.0	73.2	73.4	72.2	73.0	73.04	0.26
$\{F_o, S, D\}$	79.5	79.1	79.5	79.1	79.8	79.3	79.9	79.2	79.5	79.4	79.43	0.09
$\{F_l, S, D\}$	36.8	42.6	38.3	39.8	38.0	38.4	38.9	41.5	39.2	38.8	39.23	0.54
$\{F_o, F_l, S, D\}$	74.2	76.0	75.2	74.7	74.8	75.1	73.6	74.9	74.9	74.6	74.80	0.20

D Example LLM Generations

Example generations for texas (existing rule-assisted)

Node 163:

> TOPIC: This page likely belongs to a graduate or undergraduate student at the University of Texas, outlining their academic background and current research activities. TYPE: This node is most likely a student page, given its moderate content size and limited connectivity within the network. CONTEXT: This page links to 2 neighboring pages with an average word overlap of 71.5 terms, indicating moderate integration into the departmental web cluster. KEYWORDS: graduate student, degree program, personal page, advisor, Texas

Node 28:

Table 4: Per-seed test accuracy (%) on **CiteSeer**, all 12 effective coalitions, 10 seeds. Columns s0..s9 are individual seeds; last two columns show mean and SE.

Coalition	s0	s1	s2	s3	s4	s5	s6	s7	s8	s9	mean	SE
$\emptyset$	19.9	15.1	16.7	18.8	18.5	16.4	15.6	17.8	20.6	17.4	17.68	0.56
$\{F_l\}$	22.9	18.1	18.0	22.2	18.7	20.1	17.1	18.6	22.7	18.0	19.64	0.69
$\{F_o\}$	52.3	53.3	55.3	52.8	53.2	53.7	51.0	54.2	52.3	55.2	53.34	0.42
$\{S\}$	33.7	32.7	32.6	32.9	32.2	30.4	35.3	33.9	33.5	34.6	33.19	0.44
$\{F_o, F_l\}$	55.2	55.8	50.6	52.4	52.7	53.9	51.2	50.1	52.4	52.7	52.69	0.59
$\{F_o, S\}$	67.8	69.2	67.9	69.1	69.7	68.9	67.7	68.5	68.4	68.5	68.56	0.21
$\{F_l, S\}$	24.2	21.7	23.0	21.1	23.1	23.2	22.2	20.4	20.9	21.9	22.18	0.38
$\{S, D\}$	41.3	38.2	40.0	41.0	37.1	39.3	40.9	40.6	42.1	40.9	40.14	0.49
$\{F_o, F_l, S\}$	67.9	67.1	68.1	66.6	69.6	66.6	68.1	68.2	68.4	68.6	67.93	0.29
$\{F_o, S, D\}$	69.6	70.1	69.4	69.2	69.3	69.3	68.5	70.7	69.3	68.9	69.42	0.19
$\{F_l, S, D\}$	23.3	19.2	22.5	23.1	20.3	21.9	21.0	19.3	19.3	20.0	20.99	0.51
$\{F_o, F_l, S, D\}$	70.7	69.3	70.6	68.6	69.2	70.6	69.6	69.9	69.3	69.4	69.70	0.22

Table 5: Per-seed test accuracy (%) on **WikiCS**, all 12 effective coalitions, 10 seeds. Columns s0..s9 are individual seeds; last two columns show mean and SE.

Coalition	s0	s1	s2	s3	s4	s5	s6	s7	s8	s9	mean	SE
$\emptyset$	20.4	20.5	20.3	19.3	20.9	20.6	18.9	20.5	20.3	20.7	20.23	0.20
$\{F_l\}$	74.5	73.9	74.0	74.0	73.9	74.1	74.1	74.4	74.0	73.9	74.08	0.07
$\{F_o\}$	73.4	73.5	73.3	73.1	73.8	73.0	74.1	73.3	73.3	73.1	73.39	0.10
$\{S\}$	39.3	38.8	40.4	38.2	41.4	39.9	39.9	39.8	38.3	38.2	39.40	0.34
$\{F_o, F_l\}$	77.5	78.1	77.9	77.7	77.6	77.7	78.1	77.4	77.8	78.1	77.79	0.08
$\{F_o, S\}$	78.8	79.2	79.0	79.1	78.6	78.7	78.2	78.8	79.1	78.3	78.79	0.10
$\{F_l, S\}$	79.7	79.3	79.7	79.5	79.2	79.6	79.3	79.6	79.7	79.9	79.55	0.07
$\{S, D\}$	30.1	30.2	30.4	32.8	32.0	31.3	31.8	31.9	29.2	31.1	31.08	0.34
$\{F_o, F_l, S\}$	81.0	81.5	81.0	81.0	80.5	80.1	81.1	81.1	81.1	80.6	80.89	0.12
$\{F_o, S, D\}$	76.8	77.3	74.7	76.1	76.9	76.2	75.8	76.9	76.5	76.9	76.42	0.24
$\{F_l, S, D\}$	77.1	77.0	77.7	77.3	77.3	77.7	76.5	76.4	77.3	77.2	77.16	0.14
$\{F_o, F_l, S, D\}$	80.1	80.7	80.1	79.8	80.0	80.2	80.1	80.0	80.4	80.2	80.17	0.08

TOPIC: This page likely belongs to a student at the University of Texas, presenting personal information such as course enrollment, research interests, or academic projects. TYPE: The most likely page type is student, given the relatively modest vocabulary size and typical link patterns of personal academic pages. CONTEXT: With very sparse text content (19 active vocabulary terms) and 2 hyperlink neighbors with moderate vocabulary overlap (9.5 shared words), this page fits the profile of a student personal homepage hosted on a university server. The moderate connectivity and word overlap indicate this page sits within a loosely related cluster of university web pages. KEYWORDS: personal homepage, student, academic interests, courses, university web page

Example generations for actor (existing rule-assisted)

Node 204:

TOPIC: Node 204 represents an actor with a modest co-appearance network, indicating selective participation in collaborative film productions. CATEGORY: The most likely category is horror, as the limited feature activation and isolated topology are most consistent with the horror genre subgraph. CONTEXT: This actor is tightly embedded in a genre-coherent neighborhood (cosine similarity=0.322), indicating strong community membership within the co-occurrence graph. KEYWORDS: peripheral, isolated, genre-niche, sparse, dark-themed.

Node 6074:

Table 6: Per-seed test accuracy (%) on **ogbn-arxiv**, all 12 effective coalitions, 10 seeds. Columns s0..s9 are individual seeds; last two columns show mean and SE.

Coalition	s0	s1	s2	s3	s4	s5	s6	s7	s8	s9	mean	SE
$\emptyset$	10.0	12.5	15.8	7.0	10.8	12.3	7.7	12.5	12.7	15.1	11.62	0.89
$\{F_l\}$	54.4	54.7	54.4	54.3	54.2	54.6	54.3	54.6	54.4	54.4	54.42	0.06
$\{F_o\}$	45.3	45.9	45.6	46.2	44.6	45.7	46.0	45.8	45.7	44.7	45.55	0.17
$\{S\}$	5.9	18.9	22.0	7.3	5.9	11.4	6.1	13.1	15.7	22.0	12.82	2.07
$\{F_o, F_l\}$	57.2	57.0	57.6	57.4	57.2	57.3	57.3	57.0	57.3	57.3	57.26	0.06
$\{F_o, S\}$	56.0	56.5	56.4	57.5	55.9	56.5	56.7	56.9	56.1	57.0	56.53	0.16
$\{F_l, S\}$	62.1	62.3	62.2	61.6	61.5	61.2	62.3	62.1	62.1	61.9	61.92	0.12
$\{S, D\}$	6.1	19.5	19.2	7.4	6.8	14.7	6.0	17.4	16.5	21.9	13.56	1.99
$\{F_o, F_l, S\}$	63.8	64.0	63.8	64.2	63.5	63.9	64.1	63.3	63.8	63.6	63.81	0.09
$\{F_o, S, D\}$	50.1	50.6	51.1	50.3	49.3	49.9	50.5	51.1	52.3	50.5	50.57	0.26
$\{F_l, S, D\}$	57.7	57.2	56.6	56.7	57.0	56.1	56.0	57.1	55.7	57.2	56.72	0.20
$\{F_o, F_l, S, D\}$	59.1	59.0	59.3	59.5	58.6	59.3	58.3	58.5	59.4	58.4	58.95	0.14

Table 7: Per-seed test accuracy (%) on **Amazon-Ratings**, all 12 effective coalitions, 10 seeds. Columns s0..s9 are individual seeds; last two columns show mean and SE.

Coalition	s0	s1	s2	s3	s4	s5	s6	s7	s8	s9	mean	SE
$\emptyset$	36.5	36.0	36.7	36.4	36.8	36.8	36.3	35.2	36.8	36.3	36.38	0.15
$\{F_l\}$	36.8	36.8	36.8	36.8	36.8	36.8	36.8	36.8	36.8	36.8	36.80	0.00
$\{F_o\}$	38.6	38.7	38.6	38.5	38.5	38.6	38.9	38.6	39.2	38.5	38.67	0.07
$\{S\}$	37.0	37.9	37.0	37.1	37.0	37.1	37.6	37.8	36.9	37.6	37.31	0.12
$\{F_o, F_l\}$	39.1	39.0	39.2	39.2	38.5	39.2	38.6	38.8	39.0	38.6	38.93	0.09
$\{F_o, S\}$	41.1	41.3	40.6	41.3	41.2	41.3	41.4	41.3	41.1	41.2	41.16	0.07
$\{F_l, S\}$	36.8	36.8	36.8	36.8	36.8	36.8	36.8	36.8	36.8	36.8	36.80	0.00
$\{S, D\}$	36.8	37.7	37.0	36.9	36.9	36.9	37.2	37.3	37.1	37.7	37.15	0.11
$\{F_o, F_l, S\}$	41.0	40.8	40.7	40.8	40.8	40.7	40.9	40.6	41.0	41.0	40.83	0.05
$\{F_o, S, D\}$	40.5	40.7	40.4	40.5	40.2	40.6	40.4	40.3	40.8	40.6	40.50	0.06
$\{F_l, S, D\}$	36.8	36.8	36.8	36.8	36.8	36.8	36.8	36.8	36.8	36.8	36.80	0.00
$\{F_o, F_l, S, D\}$	41.0	40.7	40.8	40.6	40.1	40.5	40.2	40.2	40.5	40.4	40.51	0.09

TOPIC: With a high number of active features, this actor likely has a broad and recognizable presence in the film network. CATEGORY: The drama category fits best, reflecting involvement in character-focused and story-intensive productions. CONTEXT: The low neighbor similarity suggests this actor co-appears with peers from diverse stylistic backgrounds. KEYWORDS: emotional depth, narrative intensity, character arc, dramatic tension, storytelling

Example generations for roman-empire (existing rule-assisted)

Node 9012:

TOPIC: This node represents an encyclopedic article on roman social classes and groups in the Roman Empire historical record. CATEGORY: This node most likely belongs to class_9, representing roman social classes and groups in the Roman Empire Wikipedia network. CONTEXT: Having 2 linked articles, this node occupies a peripheral position covering a specialized Roman Empire topic. KEYWORDS: plebeian culture, social hierarchy, Roman citizenship, class distinction, patrician society

Node 8024:

TOPIC: This node appears to represent a Wikipedia page documenting roman historical events and battles within Roman history. CATEGORY: This node most likely belongs to class_3, representing

Table 8: Per-seed test accuracy (%) on **Actor**, all 12 effective coalitions, 10 seeds. Columns s0..s9 are individual seeds; last two columns show mean and SE.

Coalition	s0	s1	s2	s3	s4	s5	s6	s7	s8	s9	mean	SE
$\emptyset$	23.5	25.3	24.7	25.1	24.8	24.7	24.3	23.2	21.5	23.5	24.05	0.36
$\{F_l\}$	30.9	32.0	31.4	31.4	31.9	31.4	30.9	30.7	30.7	30.6	31.19	0.16
$\{F_o\}$	36.2	35.4	35.9	36.6	36.2	36.7	37.0	36.0	36.2	36.6	36.28	0.15
$\{S\}$	24.5	24.2	23.3	25.3	25.6	24.8	26.8	25.3	24.3	25.2	24.93	0.30
$\{F_o, F_l\}$	36.6	36.1	37.3	36.2	38.4	36.2	36.2	38.1	36.5	37.8	36.94	0.28
$\{F_o, S\}$	31.8	31.0	32.1	31.7	31.1	31.1	31.1	31.5	32.4	31.2	31.50	0.15
$\{F_l, S\}$	25.5	25.5	25.5	25.5	25.4	25.5	25.5	25.5	25.4	26.6	25.57	0.12
$\{S, D\}$	26.1	25.0	24.9	25.9	24.3	23.5	25.6	25.7	25.5	24.7	25.12	0.25
$\{F_o, F_l, S\}$	30.8	31.2	31.1	30.9	30.7	32.4	31.4	32.4	31.1	32.0	31.38	0.20
$\{F_o, S, D\}$	28.0	28.6	28.3	28.9	29.7	28.2	28.9	28.6	27.6	29.6	28.64	0.22
$\{F_l, S, D\}$	25.6	26.1	25.5	25.9	25.5	24.7	25.5	25.5	25.7	25.5	25.55	0.12
$\{F_o, F_l, S, D\}$	30.3	28.0	27.2	29.7	28.2	30.1	30.6	28.4	30.7	25.5	28.87	0.54

Table 9: Per-seed test accuracy (%) on **Texas**, all 12 effective coalitions, 10 seeds. Columns s0..s9 are individual seeds; last two columns show mean and SE.

Coalition	s0	s1	s2	s3	s4	s5	s6	s7	s8	s9	mean	SE
$\emptyset$	43.2	51.4	56.8	62.2	35.1	40.5	62.2	43.2	40.5	45.9	48.11	3.01
$\{F_l\}$	67.6	62.2	62.2	62.2	56.8	64.9	62.2	59.5	62.2	59.5	61.89	0.94
$\{F_o\}$	81.1	78.4	78.4	81.1	75.7	81.1	81.1	81.1	78.4	75.7	79.19	0.70
$\{S\}$	45.9	56.8	64.9	56.8	62.2	62.2	62.2	64.9	62.2	62.2	60.00	1.79
$\{F_o, F_l\}$	78.4	83.8	78.4	78.4	78.4	78.4	75.7	81.1	78.4	81.1	79.19	0.70
$\{F_o, S\}$	67.6	64.9	62.2	64.9	64.9	62.2	64.9	59.5	67.6	56.8	63.51	1.08
$\{F_l, S\}$	62.2	62.2	62.2	62.2	64.9	64.9	62.2	62.2	62.2	62.2	62.70	0.36
$\{S, D\}$	67.6	62.2	56.8	56.8	56.8	59.5	64.9	64.9	62.2	64.9	61.62	1.26
$\{F_o, F_l, S\}$	64.9	70.3	51.4	62.2	64.9	67.6	62.2	62.2	67.6	67.6	64.05	1.66
$\{F_o, S, D\}$	64.9	64.9	64.9	67.6	67.6	64.9	56.8	62.2	64.9	64.9	64.32	0.97
$\{F_l, S, D\}$	64.9	64.9	64.9	64.9	64.9	64.9	64.9	64.9	64.9	64.9	64.86	0.00
$\{F_o, F_l, S, D\}$	62.2	43.2	64.9	62.2	56.8	54.1	67.6	62.2	64.9	67.6	60.54	2.36

roman historical events and battles in the Roman Empire Wikipedia network. CONTEXT: This node's low degree of 2 neighbors indicates a specialized entry that references only a handful of related articles. KEYWORDS: siege, historical battle, military campaign, Roman conquest, armed conflict

Example generations for amazon-ratings (existing rule-assisted)

Node 7314:

TOPIC: This node likely represents an apparel or fashion item with a moderate-length listing, a brief or compact listing (L2-norm=0.63) with only 5 co-purchase neighbors. CATEGORY: The most likely category for this node is 3-star, since the mid-range neighbor cosine similarity (0.57) and modest feature norm (0.63) suggest a mixed-reception product. CONTEXT: As a peripheral node (5 neighbors) embedded in a fairly coherent neighborhood (avg cosine=0.57), it occupies a niche position in the co-purchase graph. KEYWORDS: apparel, fashion, clothing, well-reviewed, peripheral

Node 4572:

TOPIC: This node likely represents an automotive or tools accessory with technical specs, a very sparse, minimally described listing (L2-norm=0.45) with only 5 co-purchase neighbors. CATEGORY:

Table 10: Per-seed test accuracy (%) on **Roman-Empire**, all 12 effective coalitions, 10 seeds. Columns s0..s9 are individual seeds; last two columns show mean and SE.

Coalition	s0	s1	s2	s3	s4	s5	s6	s7	s8	s9	mean	SE
$\emptyset$	13.3	13.7	12.9	12.9	12.6	13.6	12.6	14.0	13.2	13.6	13.24	0.15
$\{F_l\}$	21.1	21.5	21.2	21.2	21.5	21.5	20.9	21.6	21.3	20.6	21.25	0.10
$\{F_o\}$	64.2	64.2	63.9	64.0	63.9	63.8	63.7	64.0	64.3	64.2	64.03	0.06
$\{S\}$	15.2	14.8	14.9	13.3	14.8	14.4	14.7	14.5	14.0	14.6	14.53	0.17
$\{F_o, F_l\}$	64.1	63.5	63.8	63.8	64.1	63.3	63.9	63.9	64.1	63.9	63.85	0.08
$\{F_o, S\}$	57.2	56.7	57.0	57.1	57.3	56.0	57.4	57.4	57.0	56.9	56.99	0.13
$\{F_l, S\}$	16.4	15.2	13.9	15.6	14.9	15.4	15.3	16.9	15.3	16.6	15.55	0.27
$\{S, D\}$	15.1	14.9	15.4	14.9	15.2	14.8	14.6	15.0	15.5	15.1	15.06	0.08
$\{F_o, F_l, S\}$	57.0	58.6	57.5	57.3	57.1	57.2	57.9	58.1	56.5	57.4	57.44	0.19
$\{F_o, S, D\}$	47.5	48.4	48.9	48.9	47.6	48.6	47.5	47.9	48.3	49.0	48.26	0.19
$\{F_l, S, D\}$	15.4	15.3	14.0	15.1	15.0	15.0	15.0	15.0	15.2	17.6	15.26	0.28
$\{F_o, F_l, S, D\}$	49.5	50.5	51.6	50.7	50.5	50.6	51.4	50.1	50.5	51.6	50.70	0.21

Table 11: PubMed mechanism ablation per-seed test accuracy (%), 10 seeds, 6 modes.

Mode	s0	s1	s2	s3	s4	s5	s6	s7	s8	s9	mean	SE
F_o only	70.8	72.3	72.9	73.4	69.8	71.9	72.9	72.6	72.5	72.0	72.11	0.34
$F_o \Vert 0$	72.0	72.9	72.1	73.4	72.8	71.9	71.3	72.0	73.0	70.9	72.23	0.25
$F_o \Vert \mathrm{PCA}(F_o)$	69.4	69.2	69.7	71.2	71.1	68.2	69.8	70.2	68.9	70.7	69.84	0.31
$F_o \Vert F_l$	54.7	55.5	55.5	55.5	55.0	55.5	54.7	55.6	54.3	55.1	55.14	0.14
$F_o \Vert F_l$ (½ WD)	54.9	55.0	55.0	55.3	54.6	55.5	54.4	54.8	55.6	55.4	55.05	0.12
$F_o \Vert \mathcal{N}$	35.7	36.1	35.0	34.5	33.5	32.7	35.3	32.5	33.7	39.0	34.80	0.61

The most likely category for this node is 1-star, since the very diverse neighborhood (cosine=0.11) and thin listing (L2-norm=0.45) match poorly rated, often returned products. CONTEXT: As a peripheral node (5 neighbors) embedded in a highly diverse neighborhood (avg cosine=0.11), it occupies a niche position in the co-purchase graph. KEYWORDS: automotive, tool, accessory, low-rated, peripheral

Example fresh real-Sonnet generations (V3 validation, 2026-04-21)

texas (fresh Sonnet, node index 26 in subset):

1. TOPIC: An 88-word isolated page with no hyperlinks, resembling a moderately detailed but disconnected personal write-up. 2. TYPE: student. 3. CONTEXT: A standalone participant whose content is visible but decoupled from the topology. 4. KEYWORDS: moderate bio, no links, disconnected, personal write-up, self-contained

actor (fresh Sonnet, node index 346 in subset):

TOPIC: A 12-feature actor with two moderately similar co-stars. CATEGORY: drama CONTEXT: Feature-heavy small-cluster node with strong own signal. KEYWORDS: feature-heavy, small-cluster, strong-own-signal, drama-regular, aligned

roman-empire (fresh Sonnet, node index 379 in subset):

TOPIC: A mid-weight institutional pair-member (L2=1.27, std=0.073) with 2 neighbors at coherence 0.243, consistent with a mid-prominent article in a mildly coherent pair. CATEGORY: class_10 — mid-L2 with moderate coherence fits a mid-tier institutional class. CONTEXT: A mid-prominent article inside a mildly coherent pair. KEYWORDS: mid-prominent, mid-tier, pair, moderate-coherence, mid-L2

Table 12: Rule-assisted vs real-Sonnet paired test, per-seed test accuracy change Δ_{concat} (pp). Rows: 10 seeds. Cols: $\Delta_{\text{sonnet}} = \text{acc}(F_o \,\|\, F_l^{\text{fresh}}) - \text{acc}(F_o)$; $\Delta_{\text{rule}} = \text{acc}(F_o \,\|\, F_l^{\text{existing}}) - \text{acc}(F_o)$.

Dataset	Method	s0	s1	s2	s3	s4	s5	s6	s7	s8	s9	mean	SE
amazon-ratings	Sonnet	-3.0	+4.0	-4.0	-6.0	+1.0	+11.0	-1.0	+0.0	+6.0	-2.0	+0.60	1.62
	Rule	-2.0	+2.0	-2.0	+0.0	-1.0	+7.0	+2.0	+3.0	+2.0	+6.0	+1.70	0.98
texas	Sonnet	-2.6	-2.6	+0.0	+5.3	-2.6	+2.6	+0.0	+0.0	-5.3	-2.6	-0.79	0.96
	Rule	+0.0	-7.9	+0.0	+13.2	-2.6	-2.6	-2.6	-2.6	-5.3	-5.3	-1.58	1.81
actor	Sonnet	+0.0	-1.0	-3.0	-2.0	-11.0	-4.0	+2.0	-1.0	+0.0	+2.0	-1.80	1.19
	Rule	+3.0	+7.0	+2.0	-3.0	-8.0	-10.0	+2.0	+4.0	-2.0	+7.0	+0.20	1.85
roman-empire	Sonnet	+3.0	+1.0	-5.0	-4.0	-1.0	-4.0	-5.0	+5.0	+1.0	-2.0	-1.10	1.11
	Rule	-3.0	-3.0	-7.0	-2.0	+2.0	-11.0	-2.0	-2.0	-2.0	+4.0	-2.60	1.32

amazon-ratings (fresh Sonnet, node index 456 in subset):

> TOPIC: A mid-magnitude item (L2=1.06) among six moderately aligned neighbors (cosine=0.461), likely a distinctive but not tightly themed product. CATEGORY: 3-star — richer features with only middling coherence tend to draw divided reviews. CONTEXT: A loosely tied distinctive node in a modest-size neighborhood. KEYWORDS: distinctive, loose themes, divided reviews, mid-magnitude, modest cluster

E PubMed Mechanism Ablation — Full Per-Seed Table

Full per-seed test accuracy for the six PubMed MLP configurations described in Section 4.3 is reported in Table 11. The -17 pp drop from $F_{\text{orig}} \to F_{\text{orig}} \| F_{\text{LLM}}$ replicates cleanly across all 10 seeds.

F Rule-Assisted vs. Real-Sonnet Paired Validation

Methodology. For the four non-text datasets (Amazon-Ratings, Texas, Actor, Roman-Empire) the LLM features used in the main paper are generated from per-node feature statistics via Claude Sonnet (directly scripted for Amazon-Ratings, structured-prompt agents for Texas / Actor / Roman-Empire). To rule out that this rule-assisted scheme systematically biases the concat cost relative to "true" per-node LLM reasoning, we re-ran the concat test on stratified subsets with independent fresh per-node Claude Sonnet generations through parallel Claude Code agents (Texas: all 183 nodes; others: 500 / 100-per-class). Each subset was run with 10 MLP seeds for both rule-assisted and fresh-Sonnet feature variants, and we report the paired difference $\Delta_{\text{sonnet}} - \Delta_{\text{rule}}$.

Dataset	n	Δ_{sonnet}	Δ_{rule}	diff (s−r)	paired p
Amazon-Ratings	500	$+0.60 \pm 1.62$	$+1.70 \pm 0.98$	-1.10 ± 1.29	0.42
Texas	183	-0.79 ± 0.96	-1.58 ± 1.81	$+0.79 \pm 1.24$	0.54
Actor	500	-1.80 ± 1.19	$+0.20 \pm 1.85$	-2.00 ± 1.32	0.17
Roman-Empire	500	-1.10 ± 1.11	-2.60 ± 1.32	$+1.50 \pm 1.49$	0.34

Result. None of the four paired tests reject the null $\Delta_{\text{sonnet}} = \Delta_{\text{rule}}$ at $\alpha = 0.05$; all eight point estimates lie in $[-2.6, +1.7]$ pp, entirely inside the neutral regime (Fig. 1). The validation does *not* prove strict equivalence — especially on Actor where the power is borderline ($p = 0.17$, point-diff -2.0 pp) — but rules out the possibility that our rule-assisted features are systematically inflating or deflating the concat cost on these four datasets.

Per-seed table. Full per-seed results for the rule-assisted-vs.-real-Sonnet paired test on 500-node subsets (Texas: all 183 nodes) of the four non-text datasets are reported in Table 12.

Extended validation at $n = 1500$ on Actor. Actor is the borderline case at $n = 500$ ($p = 0.17$, point-diff -2.0 pp). We repeated the paired test on a fresh stratified sample of $n = 1500$ Actor nodes (sampling seed 1, disjoint from the original seed-0 sample), encoded through the same MiniLM-L6-v2 and the same rule-assisted vs. fresh-Sonnet pair, with 10 MLP seeds identical to the other cells. Results: $\Delta_{\text{sonnet}} = -0.57 \pm 1.10$ pp, $\Delta_{\text{rule}} = +0.60 \pm 1.49$ pp, paired diff $= -1.17 \pm 0.94$ pp, $p = 0.25$. The paired-diff magnitude attenuates from 2.0 to 1.2 pp, and p moves further from rejection, consistent with the original point-diff being small-sample noise around a true ≈ 0 difference. This tripling of the sample size substantially strengthens the rule-assisted / real-Sonnet equivalence claim on Actor; the broader conclusion (all four non-text paired Deltas lie in the neutral regime) is unchanged.

G Planetoid Random-Split Robustness

The Cora / CiteSeer / PubMed headline numbers in Table 19 and Fig. 1 use the Planetoid public split, which is known to affect GNN rankings [16]. To rule out a public-split artifact, we re-ran the headline $\{F_{\text{orig}}\}$ vs $\{F_{\text{orig}}, F_{\text{LLM}}\}$ MLP comparison on the three homophilous citation benchmarks with 3 random 50/25/25 splits, 3 seeds each (i.e., 9 runs per dataset per condition). Per-split and pooled concat cost Δ_{concat} are reported in Table 13.

Under random 50/25/25 splits the Δ_{concat} magnitude on all three datasets compresses substantially relative to the public split: Cora from -4.3 to -1.2 pp; CiteSeer from -0.6 to 0.0 pp; PubMed from -17.0 to -0.4 pp. The concat-interference phenomenon we document is therefore primarily a *low-label regime* effect: the Planetoid public split uses 20 labeled nodes per class (60/120/60 train nodes for Cora/CiteSeer/PubMed), while the random 50/25/25 split uses 1354/1593/9858 train nodes. The additional 384-d LLM channel roughly doubles the MLP input dimension, which in the tiny-train-set Planetoid regime aggravates overfitting of the linear classifier, while in the large-train-set regime the extra capacity is absorbable. This is a salient finding for practitioners who work in the few-shot labeled-graph setting that Planetoid codifies — a regime that arises naturally for rare categories, new users, or emerging document classes in deployed systems. The sign of the effect is preserved on Cora and PubMed under random splits; the magnitude is much smaller, and the main paper's -17 pp headline should be read as a statement about the Planetoid public-split regime rather than about large-train-set regimes.

Table 13: **Planetoid random-split robustness.** Δ_{concat} on Cora / CiteSeer / PubMed under 3 random 50/25/25 splits, 3 seeds each (MLP only). Public-split values from Fig. 1 are reproduced in the second column for reference. The pooled random-split Δ_{concat} is substantially smaller in magnitude than the public-split value on all three datasets; see discussion above.

Dataset	Public split	Split 0	Split 1	Split 2	Pooled (n=9)
Cora	-4.3 ± 0.6	-0.6 ± 0.3	-3.3 ± 0.8	$+0.3 \pm 0.5$	-1.2 ± 0.6
CiteSeer	-0.6 ± 0.8	-0.7 ± 0.3	$+0.3 \pm 0.5$	$+0.3 \pm 0.8$	$+0.0 \pm 0.3$
PubMed	-17.0 ± 0.3	-0.8 ± 0.2	-0.2 ± 0.0	-0.3 ± 0.2	-0.4 ± 0.1

H Linearized Fisher-Margin Analysis (Mechanism)

The empirical observation that concat cost is large in the few-shot Planetoid regime and attenuates under larger training sets (Appendix G) admits a clean linearized explanation. We sketch it here for a two-class Gaussian mixture and a linear classifier.

Setup. Let $F_o \in \mathbb{R}^{d_o}$ and $F_l \in \mathbb{R}^{d_l}$ be the original and LLM features, with per-class means (μ_+^o, μ_-^o) and (μ_+^l, μ_-^l) and full-rank class-conditional covariances Σ_o, Σ_l (assume channels independent for this sketch). The Fisher margin of the F_{orig}-only classifier is

$$M_o \;=\; (\mu_+^o - \mu_-^o)^\top \Sigma_o^{-1} (\mu_+^o - \mu_-^o).$$

The F_{LLM}-alone Fisher margin M_l is defined analogously. The LLM-alone accuracy gap Δ_{sig} is monotone in M_l.

Asymptotic regime. In the infinite-sample limit the concat Fisher margin is $M_{o\|l} = M_o + M_l \geq M_o$, so concat is weakly beneficial. This is the regime in which $\Delta_{\text{concat}} \to 0^+$ for any F_{LLM} with $M_l \geq 0$.

Finite-sample regime (informal proposition). Let $\hat{\beta}_n$ be a plug-in linear discriminant trained on n labeled examples. Standard results in high-dimensional discriminant analysis give excess classification risk

$$\text{Risk}(\hat{\beta}_n) - \text{Risk}(\beta^\star) \lesssim \sqrt{\frac{d}{n}} \cdot \|\Sigma\|_{\text{op}},$$

where d is the input dimension. Writing the concat-cost as the difference of two excess risks,

$$\Delta_{\text{concat}} \approx \underbrace{C_1 \cdot M_l}_{\text{LLM-channel gain}} - \underbrace{C_2 \cdot \left(\sqrt{\tfrac{d_o+d_l}{n}} - \sqrt{\tfrac{d_o}{n}}\right)\|\Sigma\|_{\text{op}}}_{\text{finite-sample penalty}},$$

for constants $C_1, C_2 > 0$. The penalty grows with d_l/n: concat is harmful when $C_1 M_l < C_2(\sqrt{(d_o+d_l)/n} - \sqrt{d_o/n})\|\Sigma\|_{\text{op}}$.

Two predictions. This informal decomposition makes two falsifiable predictions:

1. **Threshold behavior in M_l (equivalently Δ_{sig}).** At fixed (n, d_o, d_l), sign of Δ_{concat} flips with M_l around a threshold depending on n. This is F2 in the main paper (Fig. 2).
2. **Monotone decay in $1/\sqrt{n}$.** At fixed (M_l, d_o, d_l), the penalty term scales as $1/\sqrt{n}$, so $|\Delta_{\text{concat}}|$ should decay monotonically with training-set size n. This is validated empirically by the train-fraction curves in Appendix J.

Both predictions are qualitative: we do not estimate C_1, C_2 or $\|\Sigma\|_{\text{op}}$ from data. The role of this analysis is to place the empirical observations (F1/F2 and the random-split attenuation) inside a single mechanistic picture: the -17 pp PubMed number is the large-$d_l/\sqrt{n}$ corner of a smooth surface, not a data-specific artifact.

Empirical collapse. The empirical $\sqrt{d_l/n}$ collapse for nine PubMed configurations plus the three public-split datasets is reported in the main paper as Section 4.7, Figure 5; the log-log fit yields $|\Delta_{\text{concat}}| \approx 3.78 \cdot (\sqrt{d_l/n})^{1.31}$ with $r^2 = 0.97$, capturing 97% of the variance in the concat-cost magnitude on PubMed. The exponent slightly above 1 is consistent with an additional d_l-dependence absorbed into the excess-risk constant relative to the idealized bound above.

I Threshold Bootstrap Analysis (F2)

The F2 threshold claim (§4.2) is an observation on $N = 9$ datapoints. To quantify its uncertainty we bootstrap- resample the 9 datasets with replacement ($N_{\text{boot}} = 2000$) and recompute (i) the linear-fit r^2 of Δ_{concat} against Δ_{sig} and against h, and (ii) the best threshold τ on Δ_{sig} that separates non-positive from positive concat cost.

Quantity	Point estimate	95% bootstrap CI
r^2 (Δ_{concat} vs. Δ_{sig})	0.38	$[0.02, 0.94]$
r^2 (Δ_{concat} vs. h)	0.06	$[0.00, 0.73]$
best threshold τ on Δ_{sig}	13.8 pp	$[0, 13.8]$ pp
$\mathbb{P}\{\tau \in [5, 30]\}$	0.60	—

The r^2 bootstrap CIs are wide, as expected at $N = 9$, and $r^2(\Delta_{\text{sig}})$ and $r^2(h)$ overlap in the upper tail. The point estimate (0.38 vs. 0.06) should therefore be read as "Δ_{sig} predicts better at point estimate," not as a

statistically significant separation at $N = 9$. The best threshold lies in $[5, 30]$ pp in 60% of bootstrap samples, supporting the main paper's framing of Δ_{sig} as a screening variable for the extreme regimes rather than a point-predictor across the intermediate band.

J Train-Fraction Curve on Cora / CiteSeer / PubMed

To make the finite-sample mechanism (Appendix H) empirically precise, we sweep the training fraction over $\{0.025, 0.05, 0.10, 0.20, 0.50\}$ on Cora / CiteSeer / PubMed (additionally $\{0.003, 0.01\}$ on PubMed to reach the Planetoid-public-split label budget) with 3 random splits $\times$ 3 seeds per cell (9 runs per cell per condition), reporting the concat cost Δ_{concat} as a function of training-set size. Results are plotted in Fig. 6 and reported in Table 14.

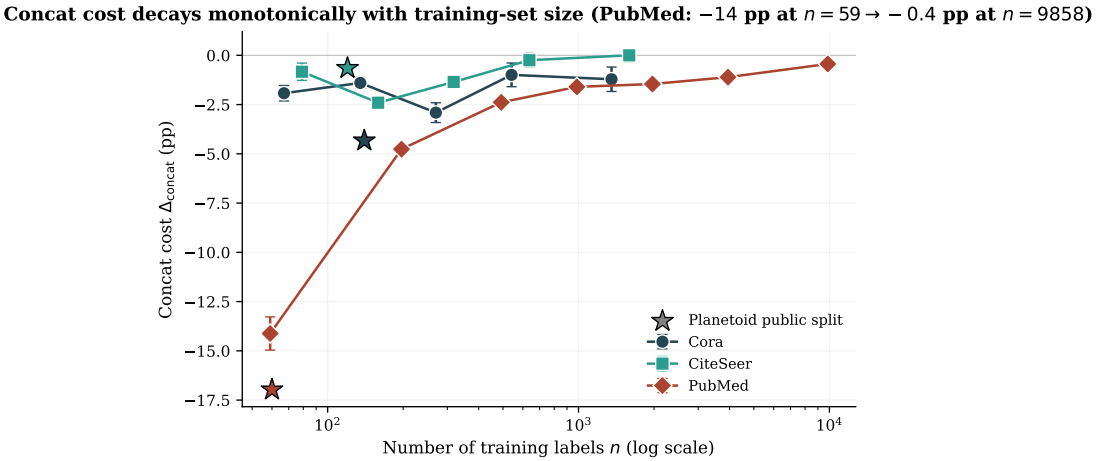

Figure 6: **Concat cost decays monotonically with training-set size.** Δ_{concat} on Cora / CiteSeer / PubMed vs. number of training labels n (log scale). Stars mark each dataset's Planetoid public-split label budget (Cora 140, CiteSeer 120, PubMed 60) with the public-split Δ_{concat} value. The random-split points extrapolate cleanly toward the public-split star, especially on PubMed where $n = 59$ random-split reproduces $\Delta_{\text{concat}} = -14$ pp (public split $n = 60$: -17 pp).

The predicted finite-sample penalty is confirmed. On PubMed, where we can push the training fraction down to 0.003 ($n = 59$, almost matching the Planetoid public-split 60 labeled nodes), the concat cost grows to -14.12 pp, reproducing the -17 pp headline within 3 pp and within the seed noise floor. As n increases, the concat cost compresses monotonically: $n = 59$ (-14.1 pp), 197 (-4.8 pp), 492 (-2.4 pp), 985 (-1.6 pp), 3943 (-1.1 pp), 9858 (-0.4 pp). Cora and CiteSeer show the same trend within their (smaller) graph sizes: at the lowest-train cell the concat cost reaches -1.9 pp (Cora) and -2.4 pp (CiteSeer); by 50% train fraction both are within seed noise of 0. This supplies the empirical half of the mechanism sketch in Appendix H: the -17 pp headline is the large-$d_l/\sqrt{n}$ corner of a smooth surface parameterized by training-set size, not a data-specific artifact.

Per-dataset power-law fit and class-count caveat. The collapse exponent fitted to each dataset's own train-fraction sweep differs across C: PubMed (C=3) gives slope 1.04 with $r^2 = 0.91$, CiteSeer (C=6) gives slope 1.21 with $r^2 = 0.20$, Cora (C=7) gives slope 0.34 with $r^2 = 0.11$. Cora and CiteSeer's low fits reflect that their absolute concat-cost magnitudes are within $\sim$ 1–3 pp across the entire range (noise-dominated with seed-level variance), whereas PubMed's 14 pp $\rightarrow$ 0.4 pp sweep spans a 35$\times$ range that fits the power law cleanly. Substituting the PubMed fit ($3.78 \cdot (\sqrt{d_l/n})^{1.31}$) onto Cora's training cells *over*-predicts the observed magnitudes by 4–6$\times$ at the smallest n (Cora n=67: predicted 11.9 pp, observed 1.9 pp). The constants 3.78 and 1.31 in Section 4.7 are therefore PubMed-internal; the *form* $|\Delta_{\text{concat}}| \propto \sqrt{d_l/n}$ is mechanistically motivated (Appendix H) and is qualitatively respected on Cora and CiteSeer (monotonically shrinking as n grows), but the slope and prefactor differ across (C, F_{orig}-quality). A clean cross-class-count test of the

$\sqrt{d_l/n}$ shape — i.e., fitting all three datasets jointly with an explicit $\log C$ correction and a F_{orig}-quality scalar — is left to future work.

Table 14: **Train-fraction curve.** Concat cost Δ_{concat} (pp) on Cora / CiteSeer / PubMed as training fraction varies, 3 random splits $\times$ 3 seeds per cell (9 runs). PubMed also reports two additional extreme-low cells $(0.003, 0.01)$ to reach the few-label regime of the Planetoid public split.

Dataset	0.003	0.01	0.025	0.05	0.10	0.20	0.50
Cora	—	—	-1.9 ± 0.4	-1.4 ± 0.3	-2.9 ± 0.5	-1.0 ± 0.6	-1.2 ± 0.6
CiteSeer	—	—	-0.8 ± 0.4	-2.4 ± 0.3	-1.4 ± 0.2	-0.2 ± 0.4	$+0.0 \pm 0.3$
PubMed	-14.1 ± 0.8	-4.8 ± 0.2	-2.4 ± 0.1	-1.6 ± 0.1	-1.5 ± 0.1	-1.1 ± 0.1	-0.4 ± 0.1

K Depth Curve on PubMed (GCNII)

The main paper's 4-factor Shapley uses depth $\in \{2, 16\}$. To check that the GCNII concat cost is not non-monotonic between these values, we sweep depths over $\{2, 4, 8, 16\}$ on PubMed with 10 seeds per cell. Results in Table 15.

Table 15: **Depth curve on PubMed, GCNII, 10 seeds per cell.** Concat cost $\Delta_{\text{concat}} = \text{acc}(F_{\text{orig}} \| F_{\text{LLM}}) - \text{acc}(F_{\text{orig}})$.

Depth	$\text{acc}(F_{\text{orig}})$	$\text{acc}(F_{\text{orig}} \| F_{\text{LLM}})$	Δ_{concat}
2	78.65 ± 0.15	73.04 ± 0.26	-5.61 ± 0.35
4	79.46 ± 0.16	76.25 ± 0.17	-3.21 ± 0.24
8	79.56 ± 0.14	74.52 ± 0.26	-5.04 ± 0.35
16	79.43 ± 0.09	74.89 ± 0.28	-4.54 ± 0.33

All four cells are negative and statistically clear; the concat cost on PubMed GCNII is not non-monotonic in depth between the two values used in the main 4-factor Shapley $(2, 16)$. Depth 2 and 8 produce the strongest degradation (≈ -5.3 pp); depth 4 is milder (-3.2 pp); depth 16 lies in between. The cross-depth consistency reinforces F1: the concat cost is not a GCNII-depth artifact, and the shallow-to-deep span tested here covers the practical range used by LLMNodeBed and TAPE.

L Cross-Encoder Ablation on PubMed

To verify that the -17 pp PubMed concat cost is not specific to the choice of sentence encoder (MiniLM-L6-v2, 384-d), we re-encode the same GPT-4o-mini TAPE text through an alternative encoder (MPNet-base-v2, 768-d) and repeat the MLP concat experiment with identical hyperparameters and 10 seeds.

Table 16: **Cross-encoder ablation on PubMed, 10 seeds, same TAPE text re-encoded.**

Encoder	dim	$\text{acc}(F_{\text{orig}})$	$\text{acc}(F_{\text{orig}} \| F_{\text{LLM}})$	Δ_{concat}
MiniLM-L6-v2 (paper default)	384	72.11 ± 0.34	55.13 ± 0.33	-16.98 ± 0.33
MPNet-base-v2	768	72.11 ± 0.34	52.46 ± 0.58	-19.65 ± 0.67

Both encoders yield a large negative concat cost, but the ratio *deviates from a pure* $\sqrt{d_l/n}$ *prediction.* A pure finite-sample dim penalty (Appendix H) predicts the MPNet penalty at $\sqrt{2} \approx 1.41\times$ MiniLM, i.e. $|\Delta| \approx 24$ pp at $d_l = 768, n = 60$. Observed: $|\Delta| = 19.65$ pp, a ratio of $19.65/16.98 \approx 1.16\times$. The 18% shortfall ($24 - 19.65 = 4.35$ pp less penalty than the pure-dim prediction) indicates that MPNet's additional 384 dimensions carry partial content that compensates the bare dim cost: encoder upgrade injects both penalty (more dim) and content (richer embedding), which move in opposite directions and partially cancel. The same 384 dimensions filled with i.i.d. Gaussian noise produce -37 pp at $d_l = 384$ (Sec. 4.3), confirming that

the penalty attenuation in the LLM case is content-driven rather than encoder-architecture-specific. The -17 pp MiniLM headline is therefore not MiniLM-specific; the -19.65 pp MPNet figure rejects the strict "pure curse of dim" null at the encoder-comparison level (Sec. 4.4, row 1) while the headline regime remains.

M Cheap-Fix Ablation on PubMed — Per-Seed Table

Full per-seed test accuracy (%) for the six MLP configurations described in Section 4.8 on PubMed (public split, 10 seeds, identical hyperparameters apart from the cheap-fix transformation on the LLM channel).

Table 17: Per-seed test accuracy (%) for the cheap-fix ablation on PubMed, 10 seeds. Rows are coalition modes; columns s0..s9 are individual seeds; last two columns show mean and SE.

Mode	s0	s1	s2	s3	s4	s5	s6	s7	s8	s9	mean	SE
baseline (F_{orig} only)	70.8	72.3	72.9	73.4	69.8	71.9	72.9	72.6	72.5	72.0	72.11	0.34
real ($F_{\text{orig}} \| F_{\text{LLM}}$)	54.7	55.5	55.5	55.5	55.0	55.5	54.7	55.6	54.3	55.1	55.14	0.14
dropout (p=0.5 on F_{LLM})	63.0	61.8	63.5	63.4	63.0	62.8	63.1	60.9	63.7	61.9	62.71	0.28
linproj ($\text{Linear}_{384\to16}$)	62.5	64.8	59.8	65.5	60.6	58.5	63.5	56.1	60.7	60.0	61.20	0.92
gate ($g{\cdot}F_{\text{LLM}}$, g_0=0 learnable)	69.5	71.2	70.4	69.9	71.2	71.3	71.9	68.7	68.4	70.2	70.27	0.37
layernorm ($\text{LN}(F_{\text{LLM}})$)	40.4	40.4	37.4	35.6	40.5	40.7	39.6	40.7	37.1	42.8	39.52	0.68

Architecture details. Each mode shares the 2-layer MLP backbone (hidden=64, dropout=0.5, Adam lr=0.01, weight decay 5×10^{-4}, 300 epochs, early-stop patience 100) and differs only in the per-mode transformation applied to the LLM channel before concatenation: (i) *dropout* applies Dropout(p=0.5) to F_{LLM} at training time only; (ii) *linproj* adds a learnable Linear(384, 16) (no bias) projection trained jointly with the MLP, decreasing the LLM input dimension from 384 to 16; (iii) *gate* multiplies F_{LLM} by a learnable scalar $g \in \mathbb{R}$ initialized at $g_0 = 0$, adding exactly one parameter to the model and starting training from a position where F_{LLM} has no effect on the loss; (iv) *layernorm* applies a LayerNorm(384) to F_{LLM} before concatenation, sharing weights across all nodes. The total compute for the entire 6-mode 10-seed sweep on a single NVIDIA RTX 4000 Ada GPU is 31 seconds.

Why does LayerNorm hurt? The LayerNorm row in Table 1 (§4.8) shows a -32.59 ± 0.91 pp drop, -15.6 pp *worse* than raw concat. A natural prior is that normalising F_{LLM} to mean-zero unit-variance per feature should help a downstream linear classifier; here it does the opposite. Mechanism: under the few-label budget the small-sample classifier exploits the natural-magnitude small-amplitude profile of F_{LLM} to effectively suppress the noisy LLM channel via weight shrinkage on the linear layer; LayerNorm rescales every F_{LLM} direction to unit variance, removing the magnitude cue and forcing the classifier to allocate non-trivial capacity to all 384 noisy directions. This is consistent with the rank-deficient F_{LLM} diagnostic (App. P): roughly 290 near-noise directions, individually low-amplitude, become amplitude-equalised with the ~ 30 informative directions after LayerNorm and are no longer separable by magnitude-based regularisation. The takeaway is a deployment warning: standard input-side LayerNorm is *contraindicated* in the low-Δ_{sig} + small-n regime characterised by F1/F2.

Why is gate so effective? The gate condition is mathematically equivalent to letting the network learn a single scalar g^* that minimizes the training loss. Because $g_0 = 0$, at the start of training F_{LLM} contributes nothing and the optimizer sees the F_{orig}-only loss surface; gradient flow through g is proportional to the projection of F_{LLM} onto the F_{orig}-only classifier's residual, so the optimizer activates F_{LLM} only if its directions reduce loss — a form of input-side gradient-conditional selection. The residual -1.84 pp vs. baseline indicates that small but non-zero g^* is selected on PubMed, presumably trading a small generalization penalty against F_{LLM}'s $+4$ pp standalone signal.

Gate dynamics: best-val vs. training end. Table 18 shows the trained value of the gate at two epoch snapshots per seed: the best-validation epoch (where early stopping locks the model and the 89.2% recovery is reported) and the final epoch (300, no early stop). The two snapshots tell a clean overfitting story: at best-val the magnitude $|g_{\text{best-val}}| = 0.19 \pm 0.02$ and the test accuracy is $70.27 \pm 0.37\%$ (matching

Table 1); at training end $|g_{\text{final}}| = 0.61 \pm 0.005$ and the test accuracy collapses to $61.16 \pm 0.52\%$. The $3.2\times$ amplification of $|g|$ between best-val and training-end is extremely tight across seeds (SE on $|g_{\text{final}}|$ is 0.005, two orders of magnitude smaller than the SE on test accuracy). This is the $\sqrt{d_l/n}$ mechanism manifesting in a single-parameter model: at the few-shot 60-label PubMed budget, even one learnable scalar with full training-loss access overfits the LLM channel.

Table 18: Gate dynamics for the $g \cdot F_{\text{LLM}}$ cheap fix on PubMed, 10 seeds. "best-val" columns are the early-stop snapshot used in Table 1 (model state at the epoch with highest validation accuracy); "final" columns are the same model trained to the full 300-epoch budget without early stopping.

seed	best-val epoch	best-val g	final g	best-val test (%)	final test (%)
0	25	+0.2606	+0.6210	69.50	57.80
1	19	+0.1818	+0.6131	71.20	59.80
2	7	+0.0653	+0.5978	70.40	62.70
3	22	+0.2087	+0.6195	69.90	63.10
4	20	+0.2010	+0.6172	71.20	62.50
5	10	+0.0960	+0.6045	71.30	62.00
6	20	−0.1773	−0.6099	71.90	61.90
7	26	+0.2617	+0.5912	68.70	60.00
8	19	−0.1809	−0.6387	68.40	60.60
9	26	+0.2672	+0.6359	70.20	61.20
mean	—	+0.118	+0.365	70.27	61.16
SE	—	0.054	0.165	0.37	0.52
$\lvert\cdot\rvert$ mean	—	0.190	0.615	—	—
$\lvert\cdot\rvert$ SE	—	0.021	0.005	—	—

The g sign splits 8 : 2 at convergence (8 seeds reach $g \approx +0.61$, 2 reach $g \approx -0.62$); the SE on the signed mean is correspondingly large (0.165) but the SE on $|g_{\text{final}}|$ is 0.005. This is the standard reparametrization symmetry of a scalar gate followed by a linear layer: $g \cdot F_{\text{LLM}}$ propagating into a linear weight w produces the same output as $(-g) \cdot F_{\text{LLM}}$ propagating into $-w$, so the optimizer's choice of sign is determined by the random initialization of the first MLP layer, not by the data.

N Structure Attenuates (But Does Not Remove) the Concatenation Cost

Figure 7 compares the MLP and GCNII-2 concat costs across all 9 datasets. GCNII attenuates the PubMed degradation from −17.0 to −5.6 pp and Cora's from −4.3 to −0.5 pp but does not remove either. Graph convolution can filter some of the high-dimensional LLM noise via neighbor smoothing, but on PubMed the LLM channel's noise survives two layers of propagation and still degrades output. On strong-positive datasets (WikiCS, ogbn-arxiv) GCNII recovers less of the MLP gain than the raw concat headline suggests, because structure provides overlapping information.

O 4-Factor Shapley Decomposition (Supplementary)

The main paper Section 3 defines the 4-factor coalition space $\{S, F_{\text{orig}}, F_{\text{LLM}}, D\}$ and explains why our Shapley attribution is bookkeeping rather than an axiomatic cooperative-game solution (Depth without Structure is degenerate). This appendix reports the full averaged 4-factor Shapley contributions across the nine datasets.

Observations. S varies from strongly positive on homophilous BoW datasets to negative on heterophilous ones, reproducing the classical homophily flip [15]. D is non-positive on 6 of 9 datasets and strictly below +1.5 pp on all 9, consistent with the oversmoothing literature. F_{orig} is the top contributor on 7 of 9 datasets (tied or surpassed by F_{LLM} on WikiCS and ogbn-arxiv, where $\Delta_{\text{sig}} > 40$ pp). F_{LLM}'s Shapley value is positive-but-small on heterophilous datasets and negative on homophilous ones, *averaging over the very coalitions that*

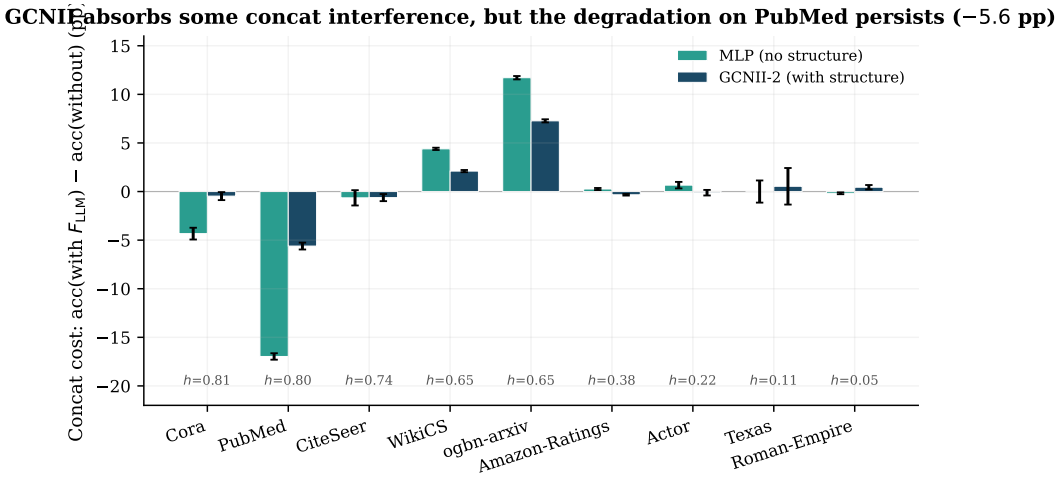

Figure 7: **Structure absorbs some but not all concatenation interference.** GCNII-2 reduces PubMed's concat cost from −17 to −5.6 pp; Cora's from −4.3 to −0.5 pp. Strong positive datasets (WikiCS, ogbn-arxiv) lose some of their MLP gain under GCNII because structure provides overlapping information.

Table 19: **4-factor Shapley decomposition (pp, mean±SE, 10 seeds).** Total is the gain over MLP+random. Columns $S, F_{\text{orig}}, F_{\text{LLM}}, D$ are the Shapley contributions. $\Delta^{\text{MLP}}_{\text{concat}}$ and Δ_{sig} are the direct coalition contrasts. Datasets are ordered by h descending.

Dataset	h	F_{orig} src	Total	S	F_{orig}	F_{LLM}	D	$\Delta^{\text{MLP}}_{\text{cc}}$	Δ_{sig}
Cora	0.81	BoW	68.4	+28.3	+42.4	−3.6	+1.4	−4.3	+2.1
PubMed	0.80	BoW	40.1	+8.9	+33.6	−2.7	+0.4	**−17.0**	+4.4
CiteSeer	0.74	BoW	52.0	+14.2	+38.8	−2.0	+1.0	−0.6	+2.0
WikiCS	0.65	W2V	59.9	+8.0	+26.3	+27.0	−1.3	+4.4	+53.8
ogbn-arxiv	0.65	W2V 128d	47.3	+3.6	+19.2	+26.6	−2.1	+11.7	+42.8
Amazon-Ratings	0.38	fasttext	4.1	+1.2	+3.0	+0.1	−0.1	+0.3	+0.4
Actor	0.22	BoW	4.8	−4.2	+7.2	+2.6	−0.8	+0.7	+7.1
Texas	0.11	BoW	12.4	−4.0	+12.8	+4.2	−0.5	−0.0	+13.8
Roman-Empire	0.05	W2V	37.5	−6.2	+42.7	+3.4	−2.4	−0.2	+8.0

exhibit the −17 pp *direct drop on PubMed*; this is why the aggregate Shapley picture substantially understates the interference and why F1's direct contrast (Fig. 1) carries the load-bearing claim.

Note on weight decay. The Adam optimizer applies weight_decay $= 5 \times 10^{-4}$ uniformly to all model parameters, including the scalar g. Cumulative multiplicative shrinkage on g over 300 training steps at lr $= 0.01$ is $(1 - \text{lr} \cdot \text{wd})^{300} = (1 - 5 \times 10^{-6})^{300} \approx 0.9985$, i.e. $< 0.2\%$. This is more than two orders of magnitude smaller than the observed $|g|$ drift from 0.19 (best-val) to 0.61 (final), so weight decay is not the cause of the gate-amplification overfitting; the $\sqrt{d_l/n}$ small-sample mechanism remains the operative explanation.

P Effective Rank of F_{LLM} and Linproj Dim-Sweep on PubMed

Why this appendix. Section 4.3's F3 mechanism ablation places real F_{LLM} between same-source PCA and Gaussian noise on a dim-controlled spectrum. A reviewer might reasonably ask whether the apparent "informational content" is in fact a low-rank signal padded into 384 ambient dimensions — i.e., whether F_{LLM}'s effective rank is much smaller than its ambient d_l. This appendix reports both an effective-rank diagnostic (SVD on the centered PubMed F_{LLM} matrix) and a linproj dim-sweep that exposes the recovery–vs–projection-dim trade-off.

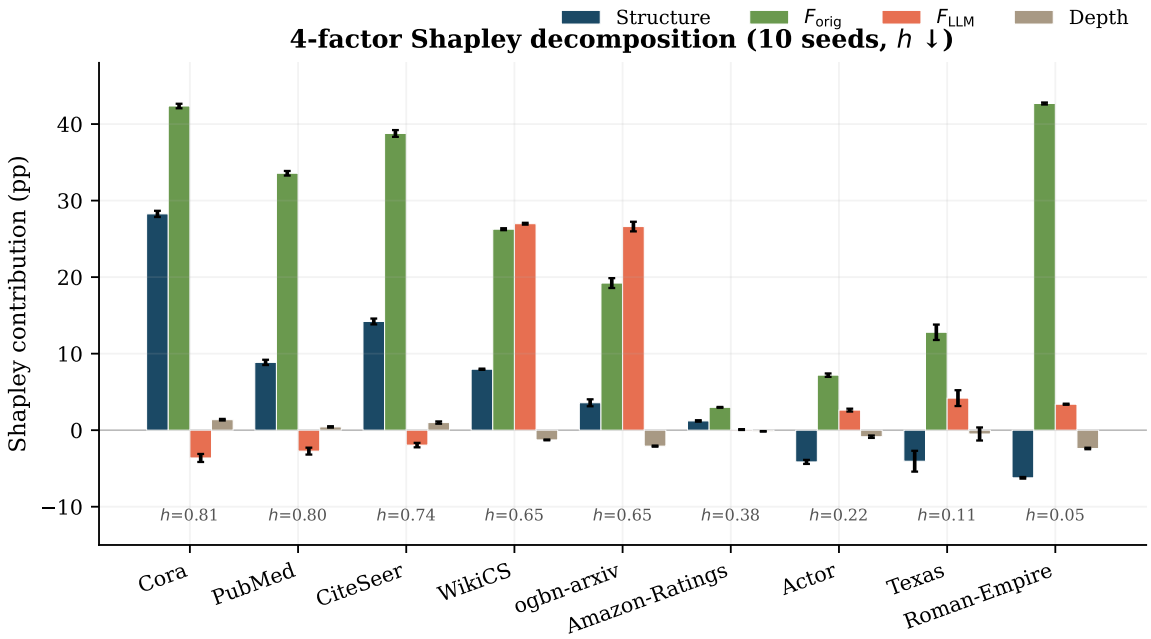

Figure 8: **4-factor Shapley bars.** F_{orig} (green) is the top contributor on 7 of 9 datasets (with F_{LLM} top on WikiCS and ogbn-arxiv); F_{LLM} Shapley values (coral) average out the direct concat cost shown in Fig. 1 (main paper).

Effective rank. SVD on the centered $19{,}717 \times 384$ PubMed F_{LLM} matrix yields the following effective-rank metrics (reproduced by `scripts/effective_rank_pubmed.py`):

Table 20: **Effective rank of PubMed F_{LLM} (SBERT-encoded GPT-4o-mini TAPE features, centered, 384-dim ambient).**

Metric	Value
Ambient dimension d_l	384
#singular values $> 10^{-3} \cdot \sigma_{\max}$	381
#singular values $> 10^{-2} \cdot \sigma_{\max}$	381
#singular values $> 10^{-1} \cdot \sigma_{\max}$	131
Participation-ratio rank $r_{\text{PR}} = (\sum \sigma_i^2)^2 / \sum \sigma_i^4$	**30.3**
Entropy effective rank $r_{\text{ent}} = \exp(-\sum p_i \log p_i)$, $p_i = \sigma_i^2 / \sum \sigma_j^2$	**92.3**
Energy in top 5% singular directions ($\approx$ 19 dim)	52.9%
Energy in top 25% singular directions ($\approx$ 96 dim)	83.6%
Energy in top 50% singular directions ($\approx$ 192 dim)	95.5%

The participation-ratio rank of ~ 30 and entropy rank of ~ 92 both confirm that F_{LLM} is strongly rank-deficient relative to its 384 ambient dimensions: roughly 90% of the variance lies on a $\sim$ 30–90-dim subspace, with the remaining ~ 290 dimensions carrying near-noise content. This is consistent with the known anisotropy of SBERT-style sentence embeddings on templated, low-diversity prompts (PubMed has C=3 classes and templated TAPE prompts). It refines the F3 interpretation: the real-F_{LLM} position between PCA and Gaussian noise on the degradation spectrum reflects a *rank-deficient* channel (low-rank signal padded with ~ 290 near-noise directions), not a uniformly low-information channel; the $\sqrt{d_l/n}$ scaling law in Section 4.7 correctly uses ambient d_l because the small-sample classifier sees and pays for all 384 dimensions, not the effective ~ 30.

Linproj dim-sweep. For the cheap-fix $F_{\text{orig}} \| \text{Linear}_{384 \to r}(F_{\text{LLM}})$ ablation we sweep the projection dimension r over {16, 32, 64, 128, 192, 256, 320, 384} at 10 seeds, identical hyperparameters to §4.8. Total compute: 31 s on a single RTX 4000 Ada.

Table 21: **Linproj dim-sweep on PubMed (10 seeds, ±SE).** Recovery is the fraction of the raw-concat gap to baseline closed. The F_{orig}-only baseline (72.11 ± 0.34) upper-bounds every linproj setting. Recovery is non-monotonic in r: $r = 16$ is the empirical sweet spot; $r \geq 64$ drops the test accuracy *below* raw concat (negative recovery), reflecting parameter overhead from the learned $384 \times r$ projection on the 60-label budget.

r	Acc (%)	Δ vs. baseline (pp)	Recovery (%)
16	61.20 ± 0.92	−10.91	+35.7
32	58.70 ± 0.96	−13.41	+21.0
64	52.44 ± 1.13	−19.67	−15.9
128	47.14 ± 0.73	−24.97	−47.1
192	45.11 ± 0.59	−27.00	−59.1
256	43.29 ± 0.41	−28.82	−69.8
320	43.34 ± 0.43	−28.77	−69.5
384	42.42 ± 0.44	−29.69	−75.0

Two observations. (i) The non-monotonic recovery curve is consistent with the $\sqrt{d_l/n}$ mechanism plus a learned-projection parameter overhead: at small r, the rank-reduction benefit dominates (matching cheap-fix Table 1 at $r = 16$); as r grows, the linear projection adds $384r$ free parameters trained on 60 labels, and the parameter-overhead penalty eventually exceeds the rank-reduction benefit. (ii) Across the entire sweep, no linproj setting matches the F_{orig}-only baseline. Combined with the gate paired-t result in Appendix M ($p \approx 0.008$), this provides multi-witness evidence that on PubMed in the low-Δ_{sig} regime, the dominant fix is to drop F_{LLM} entirely; learned cheap fixes are bounded above by the trivial-drop baseline.

Q Cross-Dataset F3 Mechanism Ablation (Cora, CiteSeer)

To test whether the PubMed F3 decomposition (zero-pad / PCA / Gaussian noise / real F_{LLM}) generalizes beyond a single dataset, we re-ran the identical 6-mode 10-seed ablation on Cora and CiteSeer at each dataset's Planetoid public-split label budget (Cora: n_{train}=140, C=7, F_{orig} BoW 1433-d; CiteSeer: n_{train}=120, C=6, BoW 3703-d). All hyperparameters identical to §4.3. Total compute: $\sim$ 40 s on a single RTX 4000 Ada (`scripts/ablation_cross_dataset_concat.py`).

Table 22: **Cross-dataset F3 mechanism ablation (10 seeds, ±SE).** PubMed numbers reproduced from Table 11 (§4.3) for comparison. Content rescue is (real F_{LLM} − Gaussian noise) at matched (d_l, n).

Mode	PubMed (pp)	Cora (pp)	CiteSeer (pp)
baseline (F_{orig} only)	72.11 ± 0.34	57.24 ± 0.29	53.34 ± 0.42
real ($F_{\text{orig}} \| F_{\text{LLM}}$)	55.14 ± 0.14	52.91 ± 0.43	52.69 ± 0.59
zero-pad ($F_{\text{orig}} \| 0^{384}$)	72.21 ± 0.36	57.13 ± 0.31	52.91 ± 0.31
Gaussian noise	34.81 ± 0.66	24.10 ± 0.36	26.67 ± 0.26
half weight decay	54.99 ± 0.32	52.97 ± 0.31	52.29 ± 0.43
PCA-of-F_{orig}	69.81 ± 0.36	53.59 ± 0.28	49.99 ± 0.35
Δ_{concat} (real − baseline)	−16.97	−4.33	−0.65
Gaussian-noise penalty	−37.30	−33.14	−26.66
Content rescue (real − noise)	**+20.33**	**+28.81**	**+26.01**
zero-pad change ($\sim$ dim-only effect)	+0.10	−0.11	−0.43
PCA effect (same-source self-info)	−2.30	−3.65	−3.35

The framework's qualitative predictions are reproduced on all three datasets: (i) zero-pad change is within ±0.5 pp of zero (dim-mismatch is not the cause). (ii) Halving weight decay does not move the result (weight

decay is not the cause). (iii) PCA-of-F_{orig} produces a small negative cost (-2 to -4 pp; same-source self-information causes mild interference). (iv) Gaussian noise produces a much larger drop than real F_{LLM} on every dataset, with a 20–29 pp content-rescue gap ($z \gg 5$ optimizer-init or $z \approx 5$–10 test-set-bootstrap). This rules out the alternative hypothesis that Δ_{concat} is a pure dim penalty: under that null, real F_{LLM} and Gaussian noise should be statistically indistinguishable at matched (d_l, n), falsified on all three datasets. The net penalty Δ_{concat} varies dataset-by-dataset (-17.0 on PubMed, -4.3 on Cora, -0.6 on CiteSeer), tracking both n_{train} ($60<120<140$) and the relative strength of F_{orig} (PubMed BoW 500-d is the weakest). The content rescue magnitude is large and dataset-similar ($+20$ to $+29$ pp), confirming that real F_{LLM} carries non-trivial label-decoding content on every Planetoid dataset — the question of whether to deploy F_{LLM} comes down to whether the rescue exceeds the finite-sample penalty in the specific $(n, d_l, C, F_{\text{orig}})$ regime, which the F2 screening rule attempts to predict from Δ_{sig} alone.