

CoFGraph: Interleaving Reciprocal Calibration with Confidence Feedback for LLM-Augmented Graph Learning

Feng Zhao, Tiewen Chen*, Xianggan Liu*, Ruilin Zhao, Yifan Song

Natural Language Processing and Knowledge Graph Lab,

School of Computer Science and Technology,

Huazhong University of Science and Technology, Wuhan, China

{zhaof, m202373953, liuxg, ruilinzha, songyf}@hust.edu.cn

Abstract

Graph learning requires jointly leveraging node semantics and graph structure, but traditional methods often suffer from label scarcity and shallow semantic utilization. Although LLMs provide strong semantic reasoning for graph learning, existing LLM-augmented pipelines face two major challenges: their unidirectional injection of semantics often propagates noise that conflicts with graph topology, and they incur substantial overhead due to LLM fine-tuning or real-time inference calls. We propose CoFGraph, a confidence-feedback framework that interleaves reciprocal calibration between semantic and structural views to enable reliable and efficient LLM-augmented graph learning. To avoid expensive LLM adaptation, we keep the LLM frozen and train a lightweight Interpreter as a semantic proxy, eliminating the need for LLM fine-tuning and continual test-time prompting. To further mitigate semantic noise and the lack of structural feedback, we introduce confidence-based dual-view learning with node-wise confidence estimation and dynamic role switching, allowing the more reliable view to correct the other. Experiments on four public benchmarks validate the effectiveness of CoFGraph, achieving consistent accuracy gains over strong baselines. Notably, CoFGraph demonstrates exceptional interference resistance under semantic noise tests, and maintains strong stability in low-resource regimes.

1 Introduction

Relational data are naturally modeled as graphs, motivating graph learning across many domains (Zhao et al., 2023). A core task is node classification, which must integrate node information with neighborhood structure (Yan et al., 2023). The challenge intensifies when nodes carry free-form natural-language descriptions, where semantics and topology must be modeled jointly. However, existing GNN pipelines still rely on costly

labeled supervision (Manchanda et al., 2023; Yang et al., 2016) and often encode text into fixed vectors without structural context (Yan et al., 2023), leading to semantic–structural mismatch (Zhou et al., 2025a; Zhu et al., 2025).

Recent LLMs offer strong semantic reasoning and generation capabilities, making them attractive tools for graph learning with language-described nodes (Vaswani et al., 2017; Brown et al., 2020). By providing low-supervision predictions and natural-language rationales, they mitigate label scarcity and supply richer semantics than traditional pipelines (Jin et al., 2024; He et al., 2024). Motivated by this, recent work explores integrating LLM-based semantic cues with the structural inductive bias of GNNs. A representative line of work treats the LLM as an interpreter, generating rationales to enrich node semantics for downstream GNN learning (He et al., 2023).

Despite their promise, most existing LLM-augmented pipelines remain unidirectional: an LLM is queried to produce pseudo-labels and natural-language rationales, which are then encoded and injected as features or supervision for a downstream GNN (He et al., 2024; Pan et al., 2024). Because LLMs are typically structure-agnostic, their text-driven outputs can easily contradict local neighborhood evidence. For example, in Figure 1, an paper titled "Analyzing Historical Documents using Neural Networks" may be misclassified as History due to the keyword "historical," ignoring its Computer Science citation neighborhood (Hu et al., 2024; Wang et al., 2024b). However, one-way designs cannot feed structural signals back to recalibrate semantic information, which can lead to severe error accumulation. Therefore, a bidirectional feedback process is required to resolve this issue. Furthermore, existing solutions often require LLM parameter updates or repeated inference calls, leading to substantial compute overhead and scalability constraints (Tang et al., 2024; Zhu et al.,

*Co-first author.

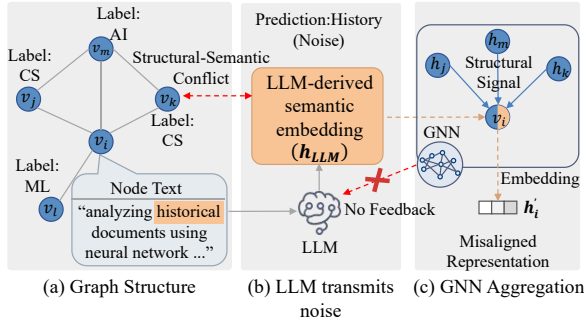


Figure 1: Semantic-noise propagation in unidirectional LLM-to-GNN pipelines. LLM outputs inferred from node text can conflict with neighborhood evidence; when used as supervision/features, the noise is further propagated by GNN aggregation, resulting in structural-semantic mismatch.

2024; Xue et al., 2023; Hu et al., 2024).

To overcome these limitations, we propose **CoF-Graph (Confidence-Feedback Graph Learning)**, a unified framework for that replaces static one-way injection with dynamic, confidence-driven feedback. CoFGraph consists of two modules: Confidence-Based Dual-View Learning and Semantic Proxy Training. First, we propose Confidence-Based Dual-View Learning, which addresses reliability under semantic noise and the missing structural feedback in prior unidirectional pipelines. We maintain a semantic view and a structural view during training, estimate their node-wise reliability, and dynamically assign the more reliable view as a temporary teacher to provide corrective supervision to the other. This bidirectional role-switching mitigates semantic noise via neighborhood consistency, while utilizing semantic evidence to compensate for sparse structural cues. Second, we propose Semantic Proxy Training to reduce LLM dependence. We keep the LLM frozen and train a lightweight Interpreter as a semantic proxy that distills the LLM’s semantic judgments (e.g., predictions and rationales) into a trainable model, and make the Interpreter isomorphic to the downstream Student for efficient interaction during dual-view learning. As a result, CoFGraph leverages LLM-level semantic knowledge without LLM fine-tuning or deployment-time LLM calls, reducing overhead while preserving the benefits of semantic augmentation.

The main contributions of this paper are summarized as follows:

- We propose a novel Confidence-Based Dual-View Learning mechanism that enables dynamic bidirectional calibration between semantic and structural views, effectively miti-

gating semantic noise and resolving structure-semantic mismatch.

- We propose Semantic Proxy Training to distill a frozen LLM’s reasoning into a lightweight Interpreter, achieving LLM-augmented graph learning without the overhead of fine-tuning or deployment-time inference calls.
- Extensive experiments on four public benchmarks demonstrate that CoFGraph exhibits exceptional robustness against semantic noise, consistently achieving superior performance and data efficiency over strong baselines.

2 Related Work

2.1 Graph Learning with GNNs

Graph learning studies how to encode relational data into representations that support downstream tasks such as node classification and link prediction. Early lines of work learn node embeddings from graph topology via random-walk objectives, including DeepWalk (Perozzi et al., 2014) and node2vec (Grover and Leskovec, 2016). More recently, Graph Neural Networks unify a broad family of methods under the message passing paradigm (Gilmer et al., 2017), where node representations are iteratively updated by aggregating information from local neighborhoods. Representative architectures include GCN (Kipf and Welling, 2017), GraphSAGE (Hamilton et al., 2017), and GAT (Velićković et al., 2018), which provide strong structural inductive biases and have become standard backbones for modern graph learning. For broader formulations and applications of GNNs, we refer to existing surveys (Wu et al., 2021).

2.2 LLMs for Graph Learning

Recent studies on integrating LLMs into graph learning can be broadly categorized into three paradigms: LLM-as-predictor, LLM-as-enhancer, and LLM-GNN alignment (Li et al., 2024a). The LLM-as-predictor line bypasses GNNs by linearizing graph structures into LLM-compatible text (e.g., prompts or node sequences) for direct inference (Ye et al., 2023; Perozzi et al., 2024; Chen et al., 2024a; Zhou et al., 2025b; Kong et al., 2025), yet such sequential formulations may still struggle to faithfully capture complex non-Euclidean dependencies (Qin et al., 2023). In contrast, LLM-as-enhancer leverages LLMs to provide richer semantic signals (e.g., pseudo labels, rationales, or

LLM-derived features) to augment graph models (He et al., 2024; Chen et al., 2024b; Pan et al., 2024; Yu et al., 2025; Sun et al., 2023; Chen et al., 2023; Duan et al., 2023). However, most enhancer pipelines remain static and unidirectional (LLM-to-GNN) and implicitly assume a globally superior teacher, whereas in content-rich graphs the more reliable expert view can vary across samples. Finally, LLM-GNN alignment seeks deeper coupling by aligning latent spaces, but typically requires fine-tuning the LLM and thus incurs substantial training costs that contradict our low-cost deployment objective (Wang et al., 2024a; Zhang et al., 2024; Tang et al., 2024; Li et al., 2024b). Therefore, our evaluation primarily focuses on the computationally efficient frozen-LLM paradigm.

3 Methods

3.1 Overview

We consider node classification on a graph $G = (V, E)$, where each node $v_i \in V$ is associated with a text description x_i (e.g., title/abstract/profile) and a label $y_i \in \{1, \dots, C\}$ is available for a subset of nodes. Given G and $\{x_i\}$, our goal is to predict labels for unlabeled nodes. To address the limitations of static, one-way LLM pipelines, we propose CoFGraph (Confidence-Feedback Graph Learning) to jointly model semantic and structural evidence and enable bidirectional calibration between them. CoFGraph is implemented through two modules: **Semantic Proxy Training**, which queries the frozen LLM to obtain pseudo hard labels, soft label distributions, rationales, and neighbor-importance assessments and distills these signals into the Interpreter; and **Confidence-Based Dual-View Learning**, which introduces an isomorphic Student to capture graph structure and performs confidence-aware bidirectional calibration, where the view that is more confident on a node temporarily acts as a teacher to calibrate the other via dynamic role switching. At inference, we further fuse the two views using uncertainty-aware weighting, so that predictions rely more on the view that is locally more reliable.

3.2 Semantic Proxy Training

Our objective is to construct a lightweight Interpreter, composed of LM and GNN, to serve as a faithful knowledge proxy for the frozen LLM. By constructing tailored prompts that provide the large language model with the textual attributes of

each central node and its local neighborhood, we obtain a comprehensive set of knowledge signals as the response. These semantic signals are designed to replace the original raw text, translating the LLM’s reasoning capability into actionable supervision and establishing a fine-grained, structure-aware foundation for the subsequent message passing stage. We detail these steps below.

Predictive Knowledge Extraction. We adopt a unified generation strategy to efficiently extract predictive supervision and explicit semantic reasoning from the LLM. By constructing an integrated prompt that contains the textual attributes of the central node and its local neighborhood, we instruct the LLM to output a pseudo-label y_n , a soft label distribution l_n , and a textual rationale r_n . The generation process is formulated as:

$$y_n, l_n, r_n = LLM(x_n, \{x_i\}_{v_i \in N(v_n)}; \text{prompt}_L) \quad (1)$$

where prompt_L denotes the unified prompt template. Crucially, r_n captures the underlying reasoning logic derived from the node’s structural and semantic context. To transfer such reasoning capability into the local model, we use r_n to construct informative initial node features, enabling the generated rationales to replace the original raw text as semantically enriched inputs aligned with the classification task.

Generative Neighborhood Evaluation. To further leverage topological context, we instruct the LLM to assess each neighbor relative to the central node and its predicted label, yielding for each neighbor $v_i \in N(v_n)$ (1) an importance score $s_{n,i}$ quantifying its contribution, and (2) an explanatory rationale $r_{n,i}$ that articulates supporting evidence. Formally, for a central node v_n and its neighborhood $\mathcal{N}(v_n)$, the extraction of these semantic signals is formulated as:

$$\begin{aligned} & \{s_{n,i}\}, \{r_{n,i}\}_{v_i \in \mathcal{N}(v_n)} \\ & = LLM(x_n, \{x_i\}_{v_i \in \mathcal{N}(v_n)}, y_n; \text{prompt}_E) \end{aligned} \quad (2)$$

where $s_{n,i}$ denotes the importance score of neighbor v_i , and $r_{n,i}$ denotes the generated natural-language rationale. We then incorporate these signals into the Interpreter’s graph propagation: the generated neighbor rationale $r_{n,i}$ is used to replace the original neighbor text for message construction, while $s_{n,i}$ serves as the aggregation weight. This design ensures that the graph convolution is driven by weighted, semantically distilled reasoning rather than noisy raw attributes.

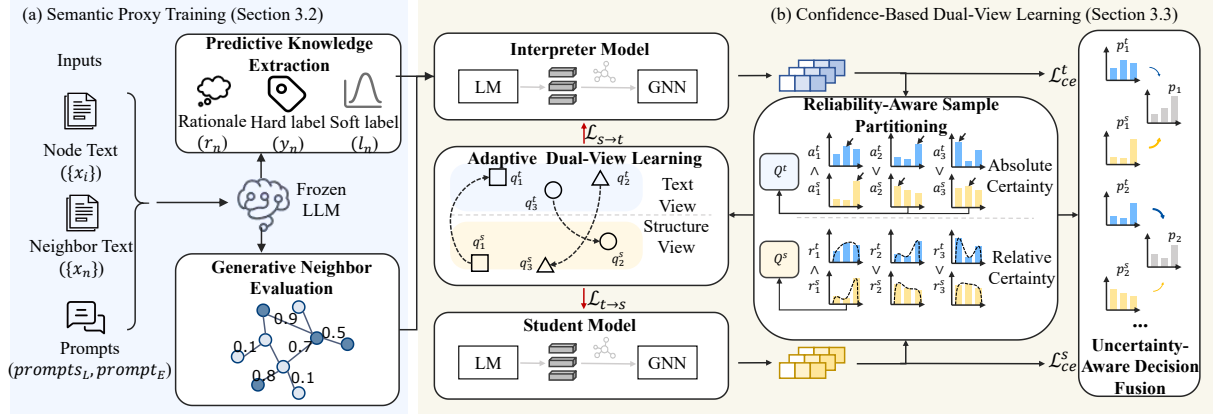


Figure 2: CoFGraph enables confidence-guided bidirectional collaboration between semantic and structural views. (a) A frozen LLM provides rationale-augmented pseudo supervision to train an Interpreter that approximates LLM semantics at low cost. (b) Based on absolute and relative certainty, samples are partitioned into reliability groups, where the more reliable view teaches the other through adaptive dual-view learning; predictions are further combined by uncertainty-aware decision fusion.

Training the Interpreter Model. Leveraging the synthesized rationales and supervisory signals distilled from the LLM, we optimize the Interpreter model. After initializing node features with the generated rationales and propagating them through the importance-weighted graph convolution layers, we obtain the final structure-aware node representations. The model then produces a predicted probability distribution $\hat{t}_n^T$ via an MLP projection head. To ensure comprehensive knowledge transfer, we employ a multi-task learning objective that enforces alignment with both the discrete pseudo-labels and the fine-grained soft distributions provided by the LLM. The total loss function is formulated as:

$$\mathcal{L}^T = \sum_{v_n \in \mathcal{V}_{train}} (\mathcal{L}_{label}^T(\hat{t}_n^T, y_n) + \lambda_1 \mathcal{L}_{logits}^T(\hat{t}_n^T, l_n)) \quad (3)$$

In this objective, $\mathcal{L}_{label}^T$ denotes the Cross-Entropy loss that fits the hard pseudo-labels y_n , while $\mathcal{L}_{logits}^T$ represents the Mean Squared Error loss designed to distill the rich latent semantic information encapsulated in the soft logits l_n . The hyperparameter λ_1 balances the contributions of these two supervision terms.

3.3 Confidence-Based Dual-View Learning

We view textual semantics (Interpreter) and structural topology (Student) as two complementary perspectives. To enable dynamic mutual calibration, this module operates in three steps: reliability-aware sample partitioning, adaptive dual-view distillation, and uncertainty-aware decision fusion.

Reliability-Aware Sample Partitioning (RSP).

Dual-view learning requires deciding which view (text or structure) is more reliable for each instance. We estimate view-specific reliability from the posterior distribution. For the i -th sample in view $v \in t, s$, we compute the probability of it belonging to class k by measuring its distance to the class prototype t_k^v . Using the Euclidean distance metric ψ , the posterior distribution is derived as:

$$\mathcal{P}(\hat{y}_i = k | x_i^v) = \frac{\exp(-\psi(q_i^v, t_k^v))}{\sum_{k'=1}^N \exp(-\psi(q_i^v, t_{k'}^v))} \quad (4)$$

where q_i^v denotes the query representation and t_k^v denotes the prototype of the k -th class.

We then quantify reliability using two complementary theoretically grounded metrics: absolute certainty a_i^v (derived from prototype-distance posteriors) and relative certainty r_i^v , defined as:

$$\begin{aligned} a_i^v &= \max_k \mathcal{P}(\hat{y}_i = k | x_i^v), \\ r_i^v &= \sum_{k=1}^N \mathcal{P}(\hat{y}_i = k | x_i^v) \log \mathcal{P}(\hat{y}_i = k | x_i^v) \end{aligned} \quad (5)$$

To ensure stable teacher selection, we assign a sample to the Text-Dominated set Q^t or Structure-Dominated set Q^s only when one view exceeds the other on both metrics:

$$\begin{aligned} Q^t &= \{(x_i^t, x_i^s) | (x_i^t, x_i^s) \in Q, a_i^t > a_i^s, r_i^t > r_i^s\}, \\ Q^s &= \{(x_i^t, x_i^s) | (x_i^t, x_i^s) \in Q, a_i^s > a_i^t, r_i^s > r_i^t\} \end{aligned} \quad (6)$$

Through this partitioning, we ensure that dual-view learning is guided only by the highly confident teacher view for each specific sample.

Adaptive Dual-View Learning (ADL). After partitioning samples into Q^t and Q^s , we perform sample-wise role switching: since no single view is globally reliable across the graph, the locally more reliable view serves as the teacher to calibrate the weaker student, overcoming the limitations of simple static fusion. We enforce cross-view consistency using KL divergence, but apply it directionally according to the reliability groups. Specifically, for Q^t we distill from text to structure, while for Q^s we distill from structure to text. Furthermore, to amplify the contribution of highly certain samples during distillation, we incorporate the absolute certainty score (a_i^t or a_i^s) as a weighting coefficient. The directional transfer losses are formulated as:

$$\begin{aligned}\mathcal{L}_{t \rightarrow s}(\theta^s) &= \frac{1}{\sum_{(x_i^t, x_i^s) \in Q^t} a_i^t} \sum_{(x_i^t, x_i^s) \in Q^t} a_i^t \mathcal{D}_{KL}(p_i^t, p_i^s) \\ \mathcal{L}_{s \rightarrow t}(\theta^t) &= \frac{1}{\sum_{(x_i^t, x_i^s) \in Q^s} a_i^s} \sum_{(x_i^t, x_i^s) \in Q^s} a_i^s \mathcal{D}_{KL}(p_i^s, p_i^t)\end{aligned}\quad (7)$$

where $\mathcal{L}_{t \rightarrow s}$ represents the distillation loss where the Interpreter calibrates the Student, and $\mathcal{L}_{s \rightarrow t}$ denotes the reverse. Here, p_i^s and p_i^t refer to the posterior distributions of the structure and text views, respectively.

The overall training objective combines task-specific supervision with this mutual calibration. The total loss $\mathcal{L}$ is a weighted sum of the independent cross-entropy losses ($\mathcal{L}_{ce}^s, \mathcal{L}_{ce}^t$) and the bidirectional distillation losses:

$$\mathcal{L} = \mathcal{L}_{ce}^s + \mathcal{L}_{ce}^t + \lambda_2(\mathcal{L}_{t \rightarrow s}(\theta^s) + \mathcal{L}_{s \rightarrow t}(\theta^t)) \quad (8)$$

where λ_2 serves as a hyperparameter to balance the regularization strength of the dual-view learning term.

Uncertainty-Aware Decision Fusion (UDF). During the inference phase, naive aggregation (e.g., averaging) of predictions from different views often yields suboptimal results, particularly when one modality is compromised by inherent noise or sparsity. To address this, we introduce an adaptive fusion mechanism that dynamically calibrates the final decision based on model certainty. The core principle is to assign greater decision weight to the more confident modality. We leverage the absolute certainty scores (a_i^v) derived in the previous section as dynamic importance weights. Specifically, the final probability of assigning a query sample to class k is computed by evaluating the

confidence-weighted proximity to the class prototype. This process effectively re-normalizes the certainty-weighted similarities across all candidate classes. By modulating the influence of each view according to its certainty, our approach ensures that the final prediction is primarily governed by the expert view for each specific sample.

4 Experiments

4.1 Experimental Setting

Benchmarks. We evaluate on four public benchmarks: Cora (McCallum et al., 2000), PubMed (Sen et al., 2008), ogbn-products (Hu et al., 2020), and arxiv_2023 (He et al., 2024). The latter contains post-2023 papers to mitigate potential LLM data contamination. We default to a 60/20/20 train/validation/test split. To ensure fair comparison, we construct a unified text field (e.g., title plus abstract) for each node as the standardized semantic input for all models, with graph topology provided by citation or co-purchase links.

Baselines. We compare three categories: (i) text-centric (fine-tuned DeBERTa (He et al., 2021) and a direct LLM baseline); (ii) GNNs (MLP, GCN (Kipf and Welling, 2017), GraphSAGE (Hamilton et al., 2017), RevGAT (Li et al., 2021)) utilizing either shallow or DeBERTa embeddings; and (iii) frozen-LLM augmented graph learning (LLM4NG (Yu et al., 2025), LLM-GNN (Chen et al., 2024b), and TAPE (He et al., 2024)). All LLM-based methods share the same backend and prompting protocol, with LLM access strictly confined to the training split to ensure comparable computational budgets.

Implementation. Implemented via PyTorch Geometric (Fey and Lenssen, 2019) and HuggingFace, experiments are conducted on two NVIDIA A100 GPUs. We report the mean accuracy over four runs. We employ DeBERTa as the textual backbone and 4-layer GNNs, optimized via two-stage training (Interpreter pre-training followed by joint Student co-training).

4.2 Main Results

We compare our framework with strong baselines, and the overall results are reported in Section 3.3. Across all four datasets, our method achieves the highest accuracy, demonstrating consistent effectiveness across diverse domains and label spaces. We summarize two main findings below.

	Method	Cora	PubMed	ogbn-products	arxiv_2023
GNN(Shallow)	MLP	0.6458 $\pm$ 0.0207	0.8615 $\pm$ 0.0033	0.5390 $\pm$ 0.0024	0.6300 $\pm$ 0.0071
	GCN	0.8658 $\pm$ 0.0171	0.8510 $\pm$ 0.0107	0.7040 $\pm$ 0.0053	0.6510 $\pm$ 0.0129
	SAGE	0.8741 $\pm$ 0.0189	0.8665 $\pm$ 0.0088	0.6900 $\pm$ 0.0028	0.6560 $\pm$ 0.0049
	RevGAT	0.8713 $\pm$ 0.0133	0.8440 $\pm$ 0.0052	0.6955 $\pm$ 0.0019	0.6945 $\pm$ 0.0037
GNN(PLM)	MLP	0.7795 $\pm$ 0.0231	0.9320 $\pm$ 0.0042	0.6500 $\pm$ 0.0229	0.7300 $\pm$ 0.0014
	GCN	0.8718 $\pm$ 0.0179	0.9325 $\pm$ 0.0034	0.7500 $\pm$ 0.0042	0.7320 $\pm$ 0.0015
	SAGE	0.8612 $\pm$ 0.0136	0.9330 $\pm$ 0.0039	0.7450 $\pm$ 0.0034	0.7420 $\pm$ 0.0018
	RevGAT	0.8589 $\pm$ 0.0168	0.9335 $\pm$ 0.0021	0.7480 $\pm$ 0.0038	0.7570 $\pm$ 0.0020
LM	DeBERTa	0.7920 $\pm$ 0.0134	0.9510 $\pm$ 0.0042	0.7310 $\pm$ 0.0028	0.7380 $\pm$ 0.0008
LLM	Llama	0.7350 $\pm$ 0.0151	0.9315 $\pm$ 0.0026	0.7410 $\pm$ 0.0013	0.7355 $\pm$ 0.0019
LLM4NG(Yu et al., 2025)	GCN	0.8153 $\pm$ 0.0142	0.9117 $\pm$ 0.0063	0.7438 $\pm$ 0.0047	0.7364 $\pm$ 0.0031
LLM-GNN(Chen et al., 2024b)	GCN	0.8442 $\pm$ 0.0116	0.9261 $\pm$ 0.0034	0.7659 $\pm$ 0.0052	0.7527 $\pm$ 0.0028
TAPE(He et al., 2024)	MLP	0.8160 $\pm$ 0.0103	0.9495 $\pm$ 0.0041	0.7850 $\pm$ 0.0088	0.7980 $\pm$ 0.0013
	GCN	0.8727 $\pm$ 0.0092	0.9370 $\pm$ 0.0026	0.7980 $\pm$ 0.0055	0.7810 $\pm$ 0.0059
	SAGE	0.8759 $\pm$ 0.0055	0.9510 $\pm$ 0.0038	0.8120 $\pm$ 0.0048	0.8005 $\pm$ 0.0028
	RevGAT	0.8782 $\pm$ 0.0121	0.9505 $\pm$ 0.0032	0.8215 $\pm$ 0.0038	0.8060 $\pm$ 0.0026
CoFGraph(Proposed)	MLP	0.8410 $\pm$ 0.0090	0.9630 $\pm$ 0.0031	0.8120 $\pm$ 0.0068	0.8265 $\pm$ 0.0031
	GCN	0.8995 $\pm$ 0.0076	0.9520 $\pm$ 0.0018	0.8270 $\pm$ 0.0036	0.8105 $\pm$ 0.0032
	SAGE	0.9041 $\pm$ 0.0048	0.9660 $\pm$ 0.0028	0.8435 $\pm$ 0.0038	0.8320 $\pm$ 0.0021
	RevGAT	0.9065 $\pm$ 0.0088	0.9645 $\pm$ 0.0027	0.8525 $\pm$ 0.0031	0.8370 $\pm$ 0.0019

Table 1: Main results of node classification. Comparison of test accuracy (mean $\pm$ std) against baselines on four benchmark datasets. The best performance is highlighted in **bold**.

Consistent gains across datasets. Our approach delivers the best performance on Cora, Pubmed, ogbn-products, and arxiv_2023, indicating that the proposed confidence-based adaptive dual-view learning generalizes beyond a single type of graph or textual content. Compared with the strongest competing methods, our framework yields stable gains, including improvements of about 2.8% on Cora, 3.1% on ogbn-products, and 3.1% on arxiv_2023. The advantage on arxiv_2023 is particularly informative because the dataset contains only recent papers, which reduces the likelihood that an LLM can directly rely on memorized test instances. These results collectively suggest that the performance gains are primarily driven by effective structure–semantics collaboration rather than simple dataset-specific shortcuts.

LLM outputs help, but feedback is crucial. Incorporating LLM outputs like rationales, pseudo-labels, and synthetic nodes consistently yields better performance than GNN baselines constrained to PLM embeddings, particularly on arxiv_2023. Such results indicate that LLM-derived knowledge delivers vital complementary evidence outside of static representations. Nevertheless, it is fundamentally suboptimal to merely inject LLM outputs as fixed features or synthetic nodes in a unidirectional manner. By consistently outperforming recent one-

way LLM-to-GNN transfer techniques, our method demonstrates the critical need for a framework capable of adaptively calibrating semantic information against topological structures via dynamic bidirectional feedback. To intuitively illustrate how structural feedback effectively suppresses semantic noise generated by the LLM, a qualitative case study is provided in Appendix A.

4.3 Efficiency and Reliability Analysis

We study performance under limited supervision and severe semantic noise, and quantify the overhead introduced by LLM supervision to demonstrate our train-for-inference paradigm.

Performance under Data Scarcity. We evaluate data efficiency by varying the labeled training ratio from 0.1% to 60%, as shown in Figure 3. A key observation is that our framework exhibits a pronounced advantage in extremely low-resource settings. When labels are scarce, the model cannot rely solely on supervised signals, and performance becomes highly sensitive to the inductive biases of the learning procedure. In this regime, our dual-view learning mechanism is particularly effective: the semantic view provides informative guidance through LLM-derived rationales, while the structural view enforces neighborhood consistency and prevents the model from over-committing to noisy

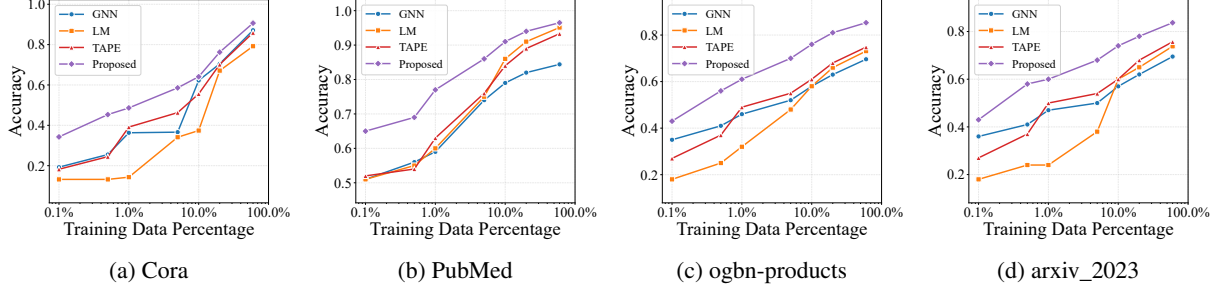


Figure 3: Data efficiency analysis. Test accuracy comparison across four datasets under varying proportions of training data, ranging from 0.1% to 60%.

Methods	#tokens	Training time	Test time	Avg.Acc
LM		118s	17s	0.7920
GNN(PLM)	0.7M	34s	16s	0.8428
TAPE		165s	31s+O(LLM)	0.8670
Proposed	1.7M	326s	21s	0.8878

Table 2: Comparison of token consumption, training time, and test-time between our proposed framework and other baselines.

semantic cues. Importantly, the performance gap does not collapse as more labeled data become available. Our method maintains a stable lead across the full spectrum of training ratios, indicating that the proposed collaboration between semantic and structural evidence remains beneficial beyond the few-shot setting.

Resource Trade-off Analysis. Section 4.3 details token usage and wall-clock time. Although CoFGraph consumes more tokens during training to construct neighborhood-aware semantic views, this upfront investment yields superior classification performance. Moreover, it completely eliminates test-time LLM calls. Unlike TAPE, which suffers from an $O(LLM)$ inference bottleneck due to real-time LLM invocations, CoFGraph achieves a fixed inference latency. This dual advantage of enhanced accuracy and massive deployment efficiency fully justifies the initial training overhead.

Robustness against Semantic Noise. To explicitly verify our mechanism’s ability to filter unreliable semantic noise, we conduct a high-intensity stress test by randomly perturbing 30% of the LLM pseudo-labels and corresponding rationales in the training set. As shown in Section 4.3, under such severe semantic noise, the strong baseline TAPE suffers a drastic 8.4% accuracy drop (from 0.8727 to 0.7996) on Cora. In contrast, CoFGraph exhibits only a marginal 3.7% decline. This empirical evidence strongly corroborates that our confidence-

Method	Original Acc	Corrupted Acc	Drop
TAPE	0.8727	0.7996	↓ 8.4%
CoFGraph	0.8995	0.8664	↓ 3.7%

Table 3: Robustness stress test on Cora with 30% injected semantic noise.

based dynamic role switching effectively detects reliability gaps and uses the structural view to override noisy semantic propagation.

4.4 Validation and Stability of Confidence Metrics

To evaluate the theoretical soundness of our prototype-distance and entropy-based confidence measures, we examine the empirical properties of the partitioned node sets. Figure 4 quantitatively details the cross-view accuracy, mean absolute certainty, and localized entropy across the Text-Dominated set Q^t and Structure-Dominated set Q^s on the Cora dataset.

The empirical results strongly demonstrate that our metrics reflect genuine localized reliability and successfully identify local experts. First, the sample partitioning is highly interpretable regarding task performance. As shown in Figure 4a, nodes assigned to Q^t consistently exhibit a clear accuracy advantage in the text view over the structural view, whereas nodes in Q^s display the exact opposite pattern, confirming that the structural view effectively corrects ambiguous text semantics in this subset. Second, this behavioral gap is firmly supported by the underlying prediction distributions. As illustrated in Figure 4b and Figure 4c, the dominant view within each partition possesses significantly higher absolute certainty a and lower localized entropy H compared to the weaker view. This sharp contrast verifies that our information-theoretic and metric-learning criteria successfully capture sample-specific certainty, enabling robust bidirectional calibration by allowing the more con-

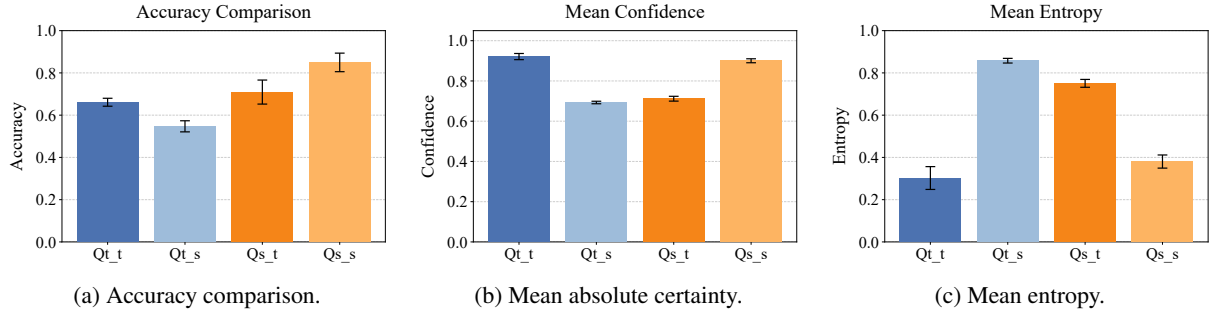


Figure 4: Empirical validation of the confidence metrics on the Cora dataset. The three bar charts demonstrate that the partitioning mechanism accurately and intuitively distinguishes between text-dominated and structure-dominated nodes based on localized certainty, task performance, and informational sharpness.

Table 4: Ablation study of Interpreter features. Evaluation of the impact of different enhanced features on the performance of the interpreter model.

Method	MLP	GCN	SAGE	RevGAT
w/o labels	0.7568	0.8382	0.8435	0.8455
w/o evidence	0.7609	0.8421	0.8476	0.8494
w/o evaluation	0.7585	0.8403	0.8455	0.8477
Interpreter	0.7725	0.8545	0.8602	0.8621

Table 5: Ablation study on dual-view learning mechanisms. Validation of the effectiveness of key components, including reliability metrics and the adaptive dual-view learning loop.

Method	MLP	GCN	SAGE	RevGAT
w/o AC	0.7554	0.8347	0.8390	0.8412
w/o RC	0.7953	0.8788	0.8833	0.8856
w/o ADL	0.7603	0.8401	0.8444	0.8467
w/o UDF	0.7676	0.8482	0.8526	0.8548
Proposed	0.8140	0.8995	0.9041	0.9065

fidient view to guide the learning process.

4.5 Ablation Study

To quantify the contribution of each component and to verify that the observed gains are not attributable to a single design choice, we conduct a two-stage ablation analysis. We first ablate the inputs used to construct the semantic view (Interpreter), and then ablate the key mechanisms in the proposed dual-view learning loop across different structural backbones.

Analysis of Interpreter Components. Table 4 reports ablations on Cora isolating the effect of different Interpreter inputs. The results demonstrate consistent performance degradation across all structural backbones when any single component is removed. This confirms the Interpreter inherently benefits from complementary information sources,

justifying our design of combining multiple semantic features to construct a robust semantic proxy.

Analysis of Dual-View Learning. We further evaluate the proposed dual-view interaction mechanisms in Table 5. Removing either the absolute or relative certainty criterion triggers a noticeable degradation in test accuracy, confirming multi-metric reliability estimation is critical for stable teacher selection. Furthermore, substantial performance drops when disabling the adaptive dual-view learning loop or uncertainty-aware decision fusion underscore that dynamic bidirectional interaction is indispensable. CoFGraph consistently achieves peak accuracy across all backbones, validating the synergy of these modules in mitigating structure-semantics mismatch.

5 Conclusion

This work advances confidence-aware graph representation learning via CoFGraph, which distills LLM semantic knowledge into graph models while mitigating semantic noise and structure-semantics misalignment. Unlike standard unidirectional transfer, CoFGraph introduces a dual-view feedback loop where semantic and structural views reciprocally calibrate predictions and rationales. By introducing a lightweight Interpreter, we bridge the gap between frozen LLMs and GNNs without the high cost of LLM fine-tuning or test-time LLM queries. Extensive experiments demonstrate the effectiveness of CoFGraph, achieving consistent accuracy improvements over strong baselines and yielding more reliable representations in low-resource and severe semantic noise settings.

Limitations

Even with strong empirical results, our framework has certain limitations. First, feature quality ultimately depends on the reasoning strength of the frozen LLM; in specialized domains, generic LLMs may produce weak rationales, which can cap performance despite our noise-filtering mechanism. Second, while we avoid LLM fine-tuning and test-time querying, training-time LLM annotation introduces non-trivial preprocessing overhead, potentially limiting scalability to very large graphs.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2023YFF0905503, National Natural Science Foundation of China under Grants No.62472188.

References

- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Runjin Chen, Tong Zhao, Ajay Kumar Jaiswal, Neil Shah, and Zhangyang Wang. 2024a. [LLaGA: Large language and graph assistant](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.
- Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, Hongzhi Wen, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, and Jiliang Tang. 2023. [Exploring the potential of large language models \(LLMs\) in learning on graphs](#). *SIGKDD Explor.*, 25(2):42–61.
- Zhikai Chen, Haitao Mao, Hongzhi Wen, Haoyu Han, Wei Jin, Haiyang Zhang, Hui Liu, and Jiliang Tang. 2024b. [Label-free node classification on graphs with large language models \(LLMs\)](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Keyu Duan, Qian Liu, Tat-Seng Chua, Shuicheng Yan, Wei Tsang Ooi, Qizhe Xie, and Junxian He. 2023. [SimTeG: A frustratingly simple approach improves textual graph learning](#). *CoRR*, abs/2308.02565.
- Matthias Fey and Jan Eric Lenssen. 2019. [Fast graph representation learning with pytorch geometric](#). *CoRR*, abs/1903.02428.
- Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. 2017. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*.
- Aditya Grover and Jure Leskovec. 2016. [node2vec: Scalable feature learning for networks](#). In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 855–864.
- William L. Hamilton, Zhitao Ying, and Jure Leskovec. 2017. [Inductive representation learning on large graphs](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 1024–1034.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [DeBERTa: Decoding-enhanced BERT with disentangled attention](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Xiaoxin He, Xavier Bresson, Thomas Laurent, and Bryan Hooi. 2023. [Explanations as features: LLM-based features for text-attributed graphs](#). *CoRR*, abs/2305.19523.
- Xiaoxin He, Xavier Bresson, Thomas Laurent, Adam Perold, Yann LeCun, and Bryan Hooi. 2024. [Harnessing explanations: LLM-to-LM interpreter for enhanced text-attributed graph representation learning](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. [Open graph benchmark: Datasets for machine learning on graphs](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Zhengyu Hu, Yichuan Li, Zhengyu Chen, Jingang Wang, Han Liu, Yumin Lee, and Kaize Ding. 2024. [Let’s ask GNN: Empowering large language model for graph in-context learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1396–1409, Miami, Florida, USA. Association for Computational Linguistics.
- Bowen Jin, Gang Liu, Chi Han, Meng Jiang, Heng Ji, and Jiawei Han. 2024. [Large language models on graphs: A comprehensive survey](#). *IEEE Trans. Knowl. Data Eng.*, 36(12):8622–8642.
- Thomas N. Kipf and Max Welling. 2017. [Semi-supervised classification with graph convolutional networks](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

- Lecheng Kong, Jiarui Feng, Hao Liu, Chengsong Huang, Jiaxin Huang, Yixin Chen, and Muhan Zhang. 2025. [GOFA: A generative one-for-all model for joint graph language modeling](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Guohao Li, Matthias Müller, Bernard Ghanem, and Vladlen Koltun. 2021. [Training graph neural networks with 1000 layers](#). In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 6437–6449. PMLR.
- Yuhan Li, Zhixun Li, Peisong Wang, Jia Li, Xiangguo Sun, Hong Cheng, and Jeffrey Xu Yu. 2024a. [A survey of graph meets large language model: Progress and future directions](#). In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 8123–8131. ijcai.org.
- Yuhan Li, Peisong Wang, Zhixun Li, Jeffrey Xu Yu, and Jia Li. 2024b. [ZeroG: Investigating cross-dataset zero-shot transferability in graphs](#). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 1725–1735. ACM.
- Sahil Manchanda, Shubham Gupta, Sayan Ranu, and Srikanth J. Bedathur. 2023. [Generative modeling of labeled graphs under data scarcity](#). In *Learning on Graphs Conference, 27-30 November 2023, Virtual Event*, volume 231 of *Proceedings of Machine Learning Research*, page 32. PMLR.
- Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. 2000. [Automating the construction of internet portals with machine learning](#). *Inf. Retr.*, 3(2):127–163.
- Bo Pan, Zheng Zhang, Yifei Zhang, Yuntong Hu, and Liang Zhao. 2024. [Distilling large language models for text-attributed graph learning](#). In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*, pages 1836–1845. ACM.
- Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. DeepWalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 701–710.
- Bryan Perozzi, Bahare Fatemi, Dustin Zelle, Anton Tsitsulin, Seyed Mehran Kazemi, Rami Al-Rfou, and Jonathan Halcrow. 2024. [Let your graph do the talking: Encoding structured data for LLMs](#). *CoRR*, abs/2402.05862.
- Yijian Qin, Xin Wang, Ziwei Zhang, and Wenwu Zhu. 2023. [Disentangled representation learning with large language models for text-attributed graphs](#). *CoRR*, abs/2310.18152.
- Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. 2008. [Collective classification in network data](#). *AI Mag.*, 29(3):93–106.
- Shengyin Sun, Yuxiang Ren, Chen Ma, and Xuechang Zhang. 2023. [Large language models as topological structure enhancers for text-attributed graphs](#). *CoRR*, abs/2311.14324.
- Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. 2024. [GraphGPT: Graph instruction tuning for large language models](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 491–500. ACM.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. [Graph attention networks](#). In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.
- Duo Wang, Yuan Zuo, Fengzhi Li, and Junjie Wu. 2024a. [LLMs as zero-shot graph learners: Alignment of GNN representations with LLM token embeddings](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Jianing Wang, Junda Wu, Yupeng Hou, Yao Liu, Ming Gao, and Julian McAuley. 2024b. [InstructGraph: Boosting large language models via graph-centric instruction tuning and preference alignment](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13492–13510, Bangkok, Thailand. Association for Computational Linguistics.
- Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. 2021. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24.
- Rui Xue, Xipeng Shen, Ruozhou Yu, and Xiaorui Liu. 2023. [Efficient end-to-end language model fine-tuning on graphs](#). *Preprint*, arXiv:2312.04737.
- Hao Yan, Chaozhuo Li, Ruosong Long, Chao Yan, Jianan Zhao, Wenwen Zhuang, Jun Yin, Peiyan Zhang, Weihao Han, Hao Sun, Weiwei Deng, Qi Zhang, Lichao Sun, Xing Xie, and Senzhang Wang. 2023. [A comprehensive study on text-attributed graphs: Benchmarking and rethinking](#). In

Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.

Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. 2016. [Revisiting semi-supervised learning with graph embeddings](#). In *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 40–48. JMLR.org.

Ruosong Ye, Caiqi Zhang, Runhui Wang, Shuyuan Xu, and Yongfeng Zhang. 2023. [Natural language is all a graph needs](#). *CoRR*, abs/2308.07134.

Jianxiang Yu, Yuxiang Ren, Chenghua Gong, Jiaqi Tan, Xiang Li, and Xuecang Zhang. 2025. [Leveraging large language models for node generation in few-shot learning on text-attributed graphs](#). In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 13087–13095. AAAI Press.

Mengmei Zhang, Mingwei Sun, Peng Wang, Shen Fan, Yanhu Mo, Xiaoxiao Xu, Hong Liu, Cheng Yang, and Chuan Shi. 2024. [GraphTranslator: Aligning graph model to large language model for open-ended tasks](#). In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 1003–1014. ACM.

Jianan Zhao, Meng Qu, Chaozhuo Li, Hao Yan, Qian Liu, Rui Li, Xing Xie, and Jian Tang. 2023. [Learning on large-scale text-attributed graphs via variational inference](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Chuang Zhou, Jiahe Du, Huachi Zhou, Hao Chen, Feiran Huang, and Xiao Huang. 2025a. [Text-attributed graph learning with coupled augmentations](#). In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 10865–10876. Association for Computational Linguistics.

Huachi Zhou, Jiahe Du, Chuang Zhou, Chang Yang, Yilin Xiao, Yuxuan Xie, and Xiao Huang. 2025b. [Each graph is a new language: Graph learning with LLMs](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 17548–17559. Association for Computational Linguistics.

Yun Zhu, Haizhou Shi, Xiaotang Wang, Yongchao Liu, Yaoke Wang, Boci Peng, Chuntao Hong, and Siliang Tang. 2025. [GraphCLIP: Enhancing transferability in graph foundation models for text-attributed graphs](#). In *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025- 2 May 2025*, pages 2183–2197. ACM.

Yun Zhu, Yaoke Wang, Haizhou Shi, and Siliang Tang. 2024. [Efficient tuning and inference for large language models on textual graphs](#). In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 5734–5742. Main Track.

A Qualitative Analysis and Failure Modes

To explicitly answer when LLM-generated explanations help versus hurt, and to demonstrate how our Confidence-Based Dual-View Learning resolves these conflicts, we conduct a qualitative case study on the Cora dataset. Furthermore, we provide a transparent discussion of the framework’s failure modes.

A.1 Case Study: Bidirectional Reciprocal Calibration

Table 6 and the following examples illustrate how the fusion mechanism dynamically shifts its trust based on localized certainty (a_t, a_s) and entropy (H_t, H_s).

Case 1: Semantic Corrects Structure (Node 1827). This paper discusses “Efficient Algorithms for Inverting Evolution” using Maximum Likelihood Estimations. Its true label is *Theory*. However, its citation neighborhood is noisy, connecting to both *Neural Networks* and *Theory* nodes. Consequently, the structural view (Student) becomes confused, predicting *Neural Networks* with high uncertainty ($H_s = 1.3801$, $a_s = 0.4270$). In contrast, the semantic view (Interpreter) extracts clear rationale from the abstract’s algorithmic focus, confidently predicting *Theory* ($H_t = 0.0744$, $a_t = 0.9869$). Recognizing this reliability gap, CoFGraph assigns a dominant fusion weight ($w = 0.7868$) to the semantic view, demonstrating that when correct structural supervision is scarce, semantic information successfully provides richer and more accurate features to complement the missing topological guidance.

Case 2: Structure Corrects Semantic Hallucination (Node 1693). This paper is titled “Incremental Self-Improvement for Life-Time Multi-Agent Reinforcement Learning”. Its true label is *Reinforcement Learning* (RL). Here, the semantic view suffers from hallucination or text ambiguity, incorrectly predicting *Case Based* with extremely high entropy ($H_t = 1.4598$). Fortunately, the structural neighborhood provides overwhelming homophilous evidence for RL. The structural view predicts RL with near-perfect certainty

Table 6: Qualitative case studies on the Cora dataset. We compare the absolute certainty (a), entropy (H), and the dynamic semantic fusion weight (w) across different scenarios.

Case Type	Node	Rationale / Text Status	Semantic View (t)		Structural View (s)		Fusion	Final Outcome
			Pred (a_t)	Ent (H_t)	Pred (a_s)	Ent (H_s)	Weight (w)	
Semantic Corrects Structure	1827	Clear, highly informative algorithmic description	Theory (0.9869)	0.0744	Neural Net. (0.4270)	1.3801	0.7868	Semantic view trusted; semantic overrides structure
Structure Corrects Semantic	1693	Ambiguous reasoning; semantic hallucination	Case Based (0.3797)	1.4598	RL (0.9999)	0.0008	0.1886	Rationale suppressed; structural view overrides

($H_s = 0.0008$, $a_s = 0.9999$). The dynamic mechanism acutely detects the semantic hallucination, shifts the fusion weight heavily toward the structural view ($w = 0.1886$, meaning $1 - w = 0.8114$ for structure), and yields the correct final prediction.

In such cases of correlated systematic bias, reciprocal calibration cannot recover the ground truth, as neither view acts as a reliable expert.

A.2 Boundary Conditions and Limitations

While CoFGraph demonstrates strong robustness, we identify two extreme edge cases where the confidence estimates are misleading or where both views are confidently wrong.

Failure Mode 1: Misleading Semantic Confidence due to Missing Features (Node 537). In real-world graphs, node attributes can be incomplete. For Node 537 (True Label: *Genetic Algorithms*), the abstract is completely missing (“[NOT FOUND]”). Deprived of semantic context, the Interpreter falls back on spurious patterns and predicts *Case Based*. Critically, it does so with a misleadingly high certainty ($a_t = 0.9244$). Because the structural view is highly uncertain ($H_s = 1.7126$), the fusion mechanism is deceived by the false semantic confidence ($w = 0.7962$), leading to an incorrect final prediction. This highlights a vulnerability: our entropy-based metrics may not distinguish between genuine reasoning certainty and overconfident guessing caused by missing inputs.

Failure Mode 2: Correlated Systematic Bias (Node 153). Node 153 belongs to *Neural Networks*, but its text describes a robot learning sensori-motor associations with delayed rewards using Q-learning. These are quintessential *Reinforcement Learning* (RL) keywords. Consequently, the semantic view confidently predicts RL ($a_t = 0.5634$). Concurrently, because the paper heavily cites RL literature, the structural neighborhood is dominated by RL nodes, leading the Student model to also confidently predict RL ($a_s = 0.9998$). When the topological connections and the textual semantics share the same interdisciplinary ambiguity, both views become confidently wrong in the same direction.