Toward Trustworthy LLM–GNN Fusion: A Fusion-Aware Evaluation and Reporting Framework

Zhifei Hu [1] **Alexandra I. Cristea** [1]

Abstract

Hybrid LLM–GNN systems are often evaluated primarily through task accuracy, which can obscure whether improvements arise from genuine cross-modal interaction or from a dominant single modality. We propose a fusion-aware evaluation framework that attributes gains relative to unimodal baselines and organises evaluation into three dimensions: *effectiveness*, *stability*, and *responsibility*. Through controlled case studies on Cora, Citeseer, WikiCS, and PubMed, we use lightweight baselines to demonstrate the evaluation protocol, rather than to claim state-of-the-art performance. The results show that fusion does not consistently improve in-domain accuracy, while benefits are more visible in cross-dataset transferability and selected robustness settings. *Responsibility* metrics further expose trade-offs, including fairness shifts and limited explanation-consistency gains. These findings illustrate that accuracy-only evaluation is insufficient for hybrid graph–language systems, and that fusion should be analysed as a multi-dimensional trade-off rather than a uniformly beneficial integration.

1. Introduction

Recent studies increasingly combine Large Language Models (LLMs) with Graph Neural Networks (GNNs) to bring together language understanding and structural reasoning. LLMs excel at processing unstructured text, while GNNs are effective at modelling relational structure, motivating a growing number of hybrid architectures that integrate these complementary capabilities. Such systems have been applied to graph question answering (Yao et al., 2024), recommendation (Zhou et al., 2021), knowledge reasoning (Lewis et al., 2020), and scientific discovery (Gilmer et al., 2017), consistently demonstrating performance gains over single-modality models. As LLM-GNN systems become more common across tasks and domains, the need for systematic and comparable evaluation becomes increasingly important. However, existing evaluation practices remain fragmented and lack a shared assessment framework.

This fragmentation largely stems from the separation of evaluation traditions in graph learning and language modelling. Graph learning benchmarks typically emphasise task accuracy under fixed data splits, sometimes supplemented with efficiency or scalability analysis (e.g., OGB (Hu et al., 2020)). In contrast, LLM evaluations focus on linguistic correctness, reasoning ability, or robustness to prompt variation (Liang et al.). When these paradigms are applied to hybrid systems, evaluation is often reduced to task-level accuracy. Yet prior work has shown that accuracy alone can give an incomplete or even misleading view of model behaviour (Ribeiro et al., 2020). As a result, it is often unclear whether reported improvements arise from effective fusion or are dominated by a single component, and what trade-offs in computation, stability, or reproducibility this integration entails. Previous studies further suggest that gains in multi-component models may arise from shortcuts or benchmark artifacts rather than meaningful interaction between components (Geirhos et al., 2020; Dehghani et al., 2021).

We argue that these issues call for an evaluation approach that goes beyond aggregate accuracy and explicitly examines the role of fusion and its trade-offs across system dimensions. This perspective aligns with prior work advocating system-level and reproducibility-aware evaluation in machine learning (Henderson et al., 2018; Pineau et al., 2021; Dodge et al., 2019). Accordingly, we propose a **fusion-aware evaluation framework for LLM-GNN systems**. Our study is guided by three core research questions for LLM-GNN fusion: ***RQ1****: To what extent do performance gains arise from genuine fusion rather than dominant single modalities?*; ***RQ2****: What trade-offs are introduced by fusion in terms of performance, cost, and stability?*; and ***RQ3****: How does fusion affect robustness and bias?*. Our

[1]Durham University, Durham DH1 3LE, United Kingdom. Correspondence to: Zhifei Hu <zhifei.hu@durham.ac.uk>, Alexandra I. Cristea <alexandra.i.cristea@durham.ac.uk>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea.

contributions are threefold:

- We identify a fundamental limitation of current evaluation practices for LLM–GNN systems, showing that accuracy-centric metrics often misattribute fusion gains and fail to capture cross-modal interactions.
- We propose a **unified fusion-aware attribution and evaluation framework** that systematically isolates genuine fusion effects relative to strong unimodal baselines across effectiveness, stability, and responsibility dimensions.
- Our results suggest that fusion tends to improve *transferability* and *robustness*, while providing limited gains in in-domain *performance* and introducing trade-offs, such as *fairness* degradation and limited *explainability* gains.

2. Related Work

This section reviews evaluation practices in graph learning and large language models, and discusses their limitations when applied to hybrid LLM-GNN systems.

2.1. Methodologies and Taxonomies in Graph and Language Model Evaluation

Standard benchmarks such as OGB (Hu et al., 2020) and large-scale studies of GNN architectures (Dwivedi et al., 2023) have improved comparability in graph representation learning. However, these efforts mainly focus on task accuracy under fixed data splits, with limited attention to efficiency, stability, or reproducibility. Several works have shown that benchmark design choices and reporting practices can strongly influence empirical conclusions (Henderson et al., 2018; Dehghani et al., 2021). In parallel, evaluation of LLMs has developed along different directions, emphasising language quality, reasoning ability, robustness to prompt variation, and human-aligned criteria, such as helpfulness and safety (Liang et al.). Recent studies have similarly argued that accuracy-centred evaluation does not fully capture model behaviour, motivating broader system-level assessment (Pineau et al., 2021). Despite their progress, these paradigms remain largely separate: graph benchmarks abstract away semantic reasoning and prompt variability, while LLM evaluations typically ignore graph structure and relational bias. As a result, neither paradigm alone is sufficient for evaluating systems that jointly rely on structural and semantic reasoning.

2.2. Evaluation of Hybrid LLM-GNN Systems

Recent work has proposed various hybrid architectures that integrate LLMs with graphs, including LLM-assisted graph construction (Yao et al., 2024), graph-grounded reasoning pipelines (Luo et al., 2024), and GNN-enhanced language models (Lewis et al., 2020). Although these approaches often report improvements over single-modality baselines, evaluation is still largely task-focused, relying mainly on accuracy or simple ablation studies. Prior research on multi-component and multimodal systems has shown that such evaluation makes it difficult to identify the true contribution of integration, since observed gains may be driven by dominant modules, shortcut learning, or dataset-specific artifacts (Geirhos et al., 2020). Consequently, existing evaluations offer limited insight into efficiency, robustness, reproducibility, and other system-level trade-offs introduced by fusion. In contrast, our work focuses on evaluation methodology itself, proposing a unified, fusion-aware framework to systematically assess both the benefits and the costs of integrating LLMs and GNNs across multiple dimensions.

3. Fusion-Aware Evaluation Framework

We propose a fusion-aware evaluation framework for LLM–GNN systems that extends beyond task accuracy by jointly characterising effectiveness, stability, and responsibility. As shown in Fig. 1, evaluation is organised into three hierarchical levels with increasing requirements on fusion design: **Effectiveness**, quantified by **Structure–Semantic Performance** (**SSP**) and **Cross-Modal Transferability** (**CMT**); **Stability**, captured by **Multi-Channel Robustness** (**MCR**) and **Fusion Reproducibility** (**FR**); and **Responsibility**, instantiated by **Graph–Text Consistency Explainability** (**GTCE**), **Cross-Modal Fairness** (**CMF**), and the **Fusion Overhead Ratio** (**FOR**). Across all levels, fusion effects are isolated via comparison with single-modality baselines using a unified fusion-aware attribution scheme and summarised through multi-dimensional trade-off profiles, with metric definitions and instantiations provided in Table 1.

3.1. Task Effectiveness and Fusion Gains (Level I)

A core challenge in evaluating LLM–GNN systems is *avoiding false attribution*: improvements observed in a hybrid pipeline may stem from the LLM or the GNN alone, rather than their interaction. To isolate the genuine effect of fusion, we evaluate the hybrid system against two single-modality baselines: *LLM-only* ($\mathcal{M}_{\text{LLM}}$) and *GNN-only* ($\mathcal{M}_{\text{GNN}}$).

To ensure a consistent interpretation across heterogeneous evaluation criteria, we define a universal **fusion-aware gain** $\Delta\mathcal{M}$. For metrics where *higher is better* (e.g., accuracy or alignment scores), the gain is defined as:

$$\Delta\mathcal{M}^{+} = \mathcal{M}(\text{LLM+GNN}) - \max\{\mathcal{M}_{\text{LLM}}, \mathcal{M}_{\text{GNN}}\}. \quad (1)$$

Conversely, for metrics where *lower is better* (e.g., variance,

Table 1. Operationalisation of the fusion-aware evaluation metrics. Fusion-aware gains $\Delta\mathcal{M}$ are computed uniformly using Eq. 1 or Eq. 2 depending on metric orientation.

Abbrev.	Full Name	Definition / Instantiation	Level
SSP	Structure–Semantic Performance	In-domain task accuracy measuring joint structural–textual effectiveness	I
CMT	Cross-Modal Transferability	Cross-dataset accuracy under distribution shift	I
MCR	Multi-Channel Robustness	Performance degradation under joint perturbations: $\text{Acc}_{\text{clean}} - \text{Acc}_{\text{perturbed}}$	II
FR	Fusion Reproducibility	Run-to-run variance across random seeds	II
GTCE	Graph–Text Consistency Explainability	Alignment between graph-side attribution patterns and textual explanations	III
CMF	Cross-Modal Fairness	Bias gap across sensitive groups in fused predictions	III
FOR	Fusion Overhead Ratio	Relative inference or memory overhead introduced by fusion (Eq. 5)	III
$\mathbf{p}$	Fusion Trade-off Profile	Vector aggregation of normalised fusion-aware gains for Pareto comparison	N/A

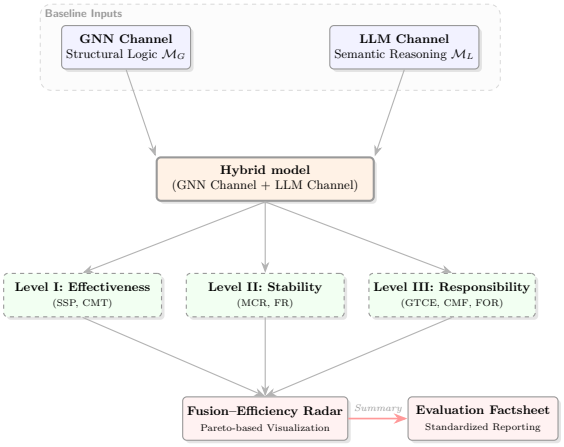

Figure 1. Fusion-aware evaluation framework for LLM–GNN systems. The evaluation factsheet records the experimental context required to interpret the metric profile, including evaluated components, datasets, perturbation settings, explanation mechanisms, sensitive attributes where applicable, and computation budget.

bias gap, or performance degradation), we define:

$$\Delta\mathcal{M}^{-} = \min\{\mathcal{M}_{\text{LLM}}, \mathcal{M}_{\text{GNN}}\} - \mathcal{M}(\text{LLM+GNN}). \quad (2)$$

Under this unified formulation, a positive value ($\Delta\mathcal{M} > 0$) indicates that the hybrid model outperforms both unimodal baselines, meaning that the hybrid system provides benefits beyond either modality alone.

Following the above fusion-aware attribution scheme, Level-I evaluation focuses on two core metrics: **Structure–Semantic Performance (SSP)** and **Cross-Modal Transferability (CMT)**.

Structure–Semantic Performance (SSP) SSP measures in-domain task performance, reflecting the model's ability to integrate graph topology with textual semantics. Following standard node classification evaluation practice (Hu et al., 2020), we report both the absolute in-domain accuracy, (Acc_{in}) and the corresponding fusion-aware gain ΔSSP. The latter is computed using Eq. (1), isolating the effectiveness attributable specifically to the fusion mechanism rather than to the strength of an individual encoder.

Cross-Modal Transferability (CMT) CMT measures cross-dataset generalisation under domain shifts, where graph structure or language style may vary. It is instantiated as cross-dataset accuracy ($\text{Acc}_{\text{cross}}$), and the corresponding gain ΔCMT is computed using Eq. (1) to determine whether fusion improves adaptability to unseen distributions compared with unimodal baselines.

3.2. Fusion Stability and Reliability Gains (Level II)

Level-II evaluation examines the *stability* of hybrid systems under input perturbations and stochastic optimisation, focusing on robustness across modalities and reproducibility.

Multi-Channel Robustness (MCR) We quantify robustness using performance degradation in accuracy Acc from clean to perturbed inputs. To avoid over-penalising a model due to a single failure mode, we adopt a weighted formulation:

$$\text{MCR} = \sum_{p\in\mathcal{P}} w_p(\text{Acc}_{\text{clean}} - \text{Acc}_p), \quad (3)$$

where $\mathcal{P}$ denotes the set of perturbations and w_p is the weight assigned to perturbation p. For simplicity, we assume uniform weights, i.e., $w_p = \frac{1}{|\mathcal{P}|}$. This formulation captures overall stability across perturbation types. To evaluate fusion gains without bias toward a single modality, we compare against the average unimodal baseline:

$$\Delta\text{MCR}^{\text{uni-mean}}_{\text{avg}} = \frac{\text{MCR}^{\text{GNN}}_{\text{avg}} + \text{MCR}^{\text{LLM}}_{\text{avg}}}{2} - \text{MCR}^{\text{Hybrid}}_{\text{avg}}, \quad (4)$$

where positive values indicate improved robustness due to fusion.

Fusion Reproducibility (FR) FR measures the stability of training outcomes under stochastic optimisation. It is

instantiated as the variance of model performance across multiple runs with different random seeds. The corresponding gain ΔFR is computed using Eq. (2), where positive values indicate reduced variance and improved reproducibility compared to unimodal baselines.

3.3. Fusion Responsibility and Deployment Efficiency (Level III)

Level-III evaluation examines whether fusion mechanisms are suitable for responsible and practical deployment. As summarised in Table 1, this level focuses on interpretability, fairness, and computational efficiency, ensuring that fusion gains are not achieved at hidden deployment costs.

Graph–Text Consistency Explainability (GTCE) GTCE measures the consistency between graph-based attention mechanisms and textual explanations, reflecting cross-modal interpretability. Following Eq. (1), the fusion-aware gain ΔGTCE quantifies whether fusion improves explanation alignment relative to unimodal baselines. In our case study, GTCE is instantiated as the alignment between graph-side attribution patterns and token-level importance scores derived from the language encoder. The framework itself is agnostic to the specific explanation method, allowing alternative graph or language explanation mechanisms to be substituted.

Cross-Modal Fairness (CMF) CMF evaluates the impact of fusion on predictive fairness and is instantiated as the bias gap $\mathcal{B}$ across sensitive groups, where lower values indicate fairer outcomes. The corresponding gain ΔCMF is computed using Eq. (2), capturing changes in bias attributable to fusion.

Fusion Overhead Ratio (FOR) FOR quantifies the relative computational overhead introduced by fusion with respect to the GNN backbone. Let $\mathrm{Cost}(\cdot)$ denote a resource metric such as inference latency or memory consumption; FOR is defined as:

$$\mathrm{FOR} = \frac{\mathrm{Cost(LLM{+}GNN)}}{\mathrm{Cost(GNN)}} - 1. \tag{5}$$

Lower FOR values indicate more efficient fusion designs.

To summarise multi-dimensional fusion effects, we aggregate the fusion-aware gains into a unified **Fusion Trade-off Profile** $\mathbf{p}$:

$$\mathbf{p} = \begin{pmatrix} \Delta\mathrm{SSP},\ \Delta\mathrm{MCR},\ \Delta\mathrm{FR}, \\ \Delta\mathrm{CMT},\ \Delta\mathrm{CMF},\ \Delta\mathrm{GTCE},\ -\mathrm{FOR} \end{pmatrix} \tag{6}$$

By construction, all components of $\mathbf{p}$ follow a higher-is-better convention, enabling direct comparison of different fusion designs via Pareto dominance. In the case studies, we use $\mathbf{p}$ as the object summarised by the radar profiles, rather than as a single scalar score. This avoids collapsing heterogeneous criteria into one number and makes explicit where fusion improves, degrades, or trades off behaviour across effectiveness, stability, and responsibility.

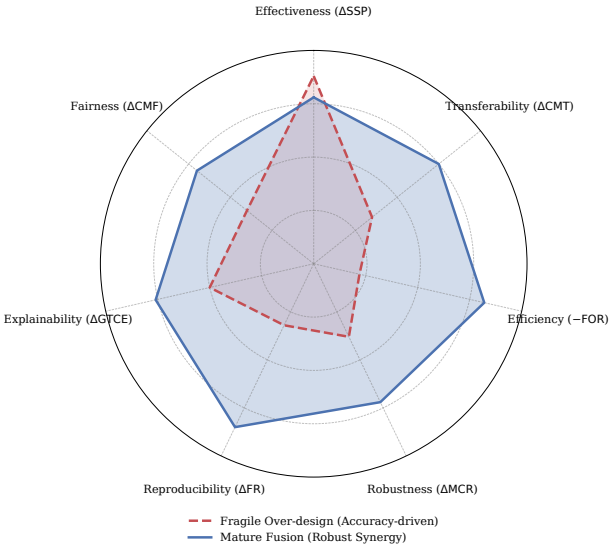

Figure 2. Conceptual Fusion–Efficiency Radar visualising the normalised trade-off profile $\mathbf{p}$.

3.4. Fusion–Efficiency Radar

To visualise multi-dimensional trade-offs, we introduce the **Fusion–Efficiency Radar** (Fig. 2). Here, $\mathbf{p}$ denotes the fusion trade-off profile defined in Eq. (6), aggregating the fusion-aware gains introduced in Levels I–III. Each component $p_i \in \mathbf{p}$ is normalised to the $[0, 1]$ range within a comparison set $\mathcal{S}$ as:

$$\widehat{p}_i = \frac{p_i - \min_{\mathcal{S}} p_i}{\max_{\mathcal{S}} p_i - \min_{\mathcal{S}} p_i + \epsilon}. \tag{7}$$

Since all dimensions of $\mathbf{p}$ follow a higher-is-better convention, the radar plot enables compact visual comparison across metrics. As the final component of the framework (Fig. 1), the radar summarises seven dimensions: ΔSSP, ΔCMT, ΔMCR, ΔFR, ΔGTCE, ΔCMF, and $-$FOR. Because *min–max* normalisation can visually amplify small differences and gives equal visual weight to heterogeneous metrics, the radar plot should be read as a diagnostic summary of the profile, rather than as a scalar ranking. Raw metric values are therefore reported in the corresponding tables.

4. Case Studies of Fusion-Aware Evaluation

This section presents case studies on four benchmarks (Cora, Citeseer, WikiCS, PubMed), to illustrate how the proposed

framework changes the evaluation of Graph–LLM fusion systems. Rather than aiming for benchmark completeness or state-of-the-art performance, we use controlled comparisons, to show how fusion-aware evaluation reveals deployment-relevant properties that are overlooked by accuracy-only reporting. Thus, the experiments are intended to demonstrate evaluation methodology, rather than to claim universal effectiveness of a specific fusion architecture. In addition, we adopt a lightweight language encoder (TinyBERT) to isolate fusion effects under controlled settings, rather than to study the scaling behaviour of large language models.

4.1. Datasets, Baselines, and Experimental Setup

We conduct experiments on four widely used homogeneous graph benchmarks: **Cora** (McCallum et al., 2000), **Citeseer** (Giles et al., 1998), **PubMed** (Sen et al., 2008), and **WikiCS** (Mernyei & Cangea, 2020), focusing on the node classification task. These datasets are selected to vary in graph scale, semantic domain (computer science, biomedical, Wikipedia), graph construction mechanisms (citation vs. hyperlink graphs), and feature characteristics. As baseline models, we employ a Graph Convolutional Network (**GCN**) (Kipf & Welling, 2017) to capture graph structural information, a Graph Attention Network (**GAT**) (Veličković et al., 2017) to model adaptive neighbour importance via attention mechanisms, and a Sentence-BERT (**SBERT**) (Reimers & Gurevych, 2019) encoder to model textual semantics without graph propagation. These models are chosen as established and representative *graph-only* and *text-only* architectures, to enable controlled comparison and analysis of fusion effects. Finally, as our goal is to evaluate the proposed framework, rather than compare fusion strategies, we use a single representative fusion model, to demonstrate how the framework captures multi-dimensional trade-offs. This hybrid model integrates TinyBERT-based language representations with GCN-based graph propagation through a lightweight fusion module, without introducing additional architectural complexity. All models are trained under identical data splits and optimisation settings. Experiments are conducted on an NVIDIA GeForce RTX 4070 Ti GPU using PyTorch 2.5.1 and CUDA 12.4, and results are averaged over five runs with different random seeds.

4.2. Level-I Effectiveness as a Case Study of Fusion Utility(RQ1)

The metrics reported in Table 2 follow the Level I definitions introduced in section 3, *Structure–Semantic Performance* (SSP) is instantiated by in-domain node classification accuracy $\mathrm{Acc}_{\mathrm{in}}$; *Cross-Modal Transferability* (CMT) is instantiated by cross-dataset accuracy $\mathrm{Acc}_{\mathrm{cross}}$ under a fixed-source setting (trained on Cora and evaluated on target datasets without retraining); and fusion effectiveness is measured by the fusion-aware gain ΔSSP computed according to Eq. (1) relative to the stronger unimodal baseline. Results (see Table 2) show that ΔSSP is positive on Cora, Citeseer, and PubMed, while a slight negative value is observed on WikiCS, indicating that fusion does not consistently improve in-domain accuracy across datasets. In contrast, $\mathrm{Acc}_{\mathrm{cross}}$ improves substantially across target datasets, suggesting that fusion provides more consistent gains in cross-dataset transferability than in in-domain performance. This indicates that Level I effectiveness cannot be reliably assessed using accuracy alone and is better reflected in transferability.

Table 2. Level-I effectiveness evaluation. SSP is instantiated by in-domain accuracy; CMT denotes cross-dataset accuracy under a fixed-source transfer setting (Cora → target), and is not applicable to Cora itself. Fusion-aware gain ΔSSP is computed as per Eq. 1.

Dataset	Model	SSP ↑	CMT ↑	ΔSSP
Cora	GNN (GCN)	0.8472±0.0141	N/A	–
	GNN (GAT)	0.8517±0.0183	N/A	–
	Language-only (SBERT)	0.6993±0.0114	N/A	–
	Hybrid (GCN+TinyBERT)	0.8583±0.0151	N/A	+0.0111
Citeseer	GNN (GCN)	0.7254±0.0153	0.1793±0.0254	–
	GNN (GAT)	0.7185±0.0100	0.1753±0.0242	–
	Language-only (SBERT)	0.7630±0.0142	0.1693±0.0139	–
	Hybrid (GCN+TinyBERT)	0.7743±0.0190	0.7592±0.0055	+0.0113
WikiCS	GNN (GCN)	0.8171±0.0050	0.1174±0.0118	–
	GNN (GAT)	0.8157±0.0080	0.1751±0.0200	–
	Language-only (SBERT)	0.6208±0.0130	0.1016±0.0259	–
	Hybrid (GCN+TinyBERT)	0.8116±0.0123	0.8086±0.0109	−0.0055
PubMed	GNN (GCN)	0.8099±0.0028	0.1861±0.0107	–
	GNN (GAT)	0.8240±0.0074	0.2087±0.0120	–
	Language-only (SBERT)	0.8280±0.0045	0.0979±0.0769	–
	Hybrid (GCN+TinyBERT)	0.8567±0.0037	0.8565±0.0039	+0.0287

4.3. Level-II Stability as a Case Study of Modality Complementarity (RQ3)

We analyse modality-specific robustness under isolated perturbations (Table 3), where each perturbation is applied at a fixed 50% rate. Table 3 reports retained accuracies under perturbation, i.e., $\mathrm{Acc}_{\mathrm{perturbed}}$, rather than performance degradation. The results show that GNN-based models are more sensitive to structural perturbations (node/edge drop), while language-only models are more affected by textual perturbations, whereas the hybrid model maintains consistently higher retained accuracy across both channels, suggesting improved robustness across perturbation types. To quantify robustness in a unified manner, Table 4 reports the average-based Multi-Channel Robustness ($\mathrm{MCR}_{\mathrm{avg}}$) defined in Eq. (3) as the mean performance degradation ($\mathrm{Acc}_{\mathrm{clean}} - \mathrm{Acc}_{\mathrm{perturbed}}$) across perturbations. The corresponding fusion gain $\Delta\mathrm{MCR}_{\mathrm{avg}}^{\text{uni-mean}}$ is computed using Eq. (4), which compares the hybrid model against the average unimodal baseline. The results show that the hybrid model achieves lower or comparable $\mathrm{MCR}_{\mathrm{avg}}$ values across most datasets, with mostly positive $\Delta\mathrm{MCR}_{\mathrm{avg}}^{\text{uni-mean}}$, indicating that fusion improves robustness primarily by compen-

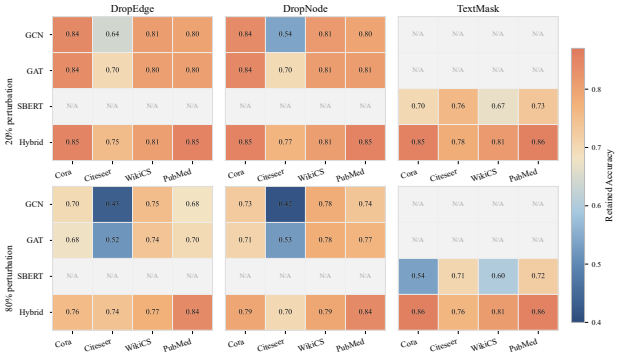

Figure 3. Robustness heatmaps under 20% (top) and 80% (bottom) perturbations. Each cell shows retained accuracy under DropEdge, DropNode, and TextMask across four datasets.

Table 3. Retained accuracy under isolated perturbations at a fixed perturbation rate. Values in this table are perturbed accuracies (not degradations). Higher values indicate that more task performance is preserved under the corresponding perturbation channel.

Dataset	Model	Node drop(50%)	Edge drop(50%)	Text mask(50%)
Cora	GNN(GCN)	0.7993 ± 0.0136	0.8074 ± 0.0216	–
	GNN(GAT)	0.8085 ± 0.0142	0.7974±0.0049	–
	Language-only(SBERT)	–	–	0.6539 ± 0.0146
	Hybrid (GCN+TinyBERT)	0.8354 ± 0.0260	0.8074 ± 0.0288	0.8579 ± 0.0179
Citeseer	GNN(GCN)	0.4897 ± 0.1547	0.5154 ± 0.0930	–
	GNN(GAT)	0.6426 ± 0.0182	0.6429 ± 0.0191	–
	Language-only(SBERT)	–	–	0.7517 ± 0.0178
	Hybrid (GCN+TinyBERT)	0.7326 ± 0.0142	0.7652 ± 0.0153	0.7762 ± 0.0165
WikiCS	GNN(GCN)	0.8028 ± 0.0062	0.7919± 0.0092	–
	GNN(GAT)	0.7802 ± 0.0087	0.7991 ± 0.0059	–
	Language-only(SBERT)	–	–	0.6475 ± 0.0095
	Hybrid (GCN+TinyBERT)	0.8030 ± 0.0124	0.7942 ± 0.0054	0.8109 ± 0.0134
PubMed	GNN(GCN)	0.7749±0.0082	0.7651 ± 0.0135	–
	GNN(GAT)	0.7952±0.0026	0.7684±0.0098	–
	Language-only(SBERT)	–	–	0.7190 ± 0.0090
	Hybrid (GCN+TinyBERT)	0.8463±0.0050	0.8445±0.0021	0.8576±0.0029

sating for modality-specific sensitivities rather than relying on a single perturbation setting. Negative MCR values may arise when perturbed inputs yield higher accuracy than clean inputs, indicating instability or noise sensitivity in the baseline model. This trend is further supported by Fig. 3, which visualises retained accuracy under varying perturbation intensities (20% and 80%); the hybrid model consistently maintains higher performance across datasets and perturbation types, especially under stronger perturbations, reinforcing its robustness advantage.

4.4. Level-III Responsibility as a Case Study of Deployment Diagnostics (RQ2)

Fusion responsibility corresponds to Level III of the proposed hierarchy and evaluates trade-offs among efficiency, explainability, fairness, and stability. Based on Table 5, Acc_{in} is used to compute fusion-aware gains via Eq. (1) and Eq. (2), while inference latency is used to compute the Fusion Overhead Ratio (FOR) in Eq. (5). As shown in Table 6, ΔGTCE and ΔCMF quantify explainability and fairness,

Table 4. Level-II stability evaluation using average robustness under 50% perturbations. MCR_{avg} is defined as the average performance degradation across perturbations. $\Delta\text{MCR}_{\text{avg}}^{\text{uni-mean}}$ measures robustness gain over the average unimodal baseline (GNN + LLM).

Dataset	Model	$\text{Acc}_{\text{clean}}$	Avg. $\text{Acc}_{\text{perturbed}}$	$\text{MCR}_{\text{avg}}\downarrow$	$\Delta\text{MCR}_{\text{avg}}^{\text{uni-mean}}\uparrow$
Cora	GNN (GCN)	0.8472	0.8034	0.0438	–
	GNN (GAT)	0.8517	0.8029	0.0488	–
	Language-only (SBERT)	0.6993	0.6539	0.0454	–
	Hybrid (GCN+TinyBERT)	0.8583	0.8336	0.0247	+0.0212
Citeseer	GNN (GCN)	0.7254	0.5026	0.2228	–
	GNN (GAT)	0.7185	0.6428	0.0757	–
	Language-only (SBERT)	0.7630	0.7517	0.0113	–
	Hybrid (GCN+TinyBERT)	0.7743	0.7580	0.0163	+0.0640
WikiCS	GNN (GCN)	0.8171	0.7974	0.0197	–
	GNN (GAT)	0.8157	0.7897	0.0260	–
	Language-only (SBERT)	0.6208	0.6475	-0.0267	–
	Hybrid (GCN+TinyBERT)	0.8116	0.8027	0.0089	-0.0108
PubMed	GNN (GCN)	0.8099	0.7700	0.0399	–
	GNN (GAT)	0.8240	0.7818	0.0422	–
	Language-only (SBERT)	0.8280	0.7190	0.1090	–
	Hybrid (GCN+TinyBERT)	0.8567	0.8495	0.0072	+0.0678

Table 5. Raw outputs for Level-I and Level-III metrics. SSP (Acc_{in}) is identical to Table 2 and serves as the input for fusion-aware gains. Cost denotes average per-instance inference latency measured under identical hardware and batch settings.

Dataset	Model	SSP (Acc_{in}) ↑	Cost (Latency, s / instance) ↓
Cora	GNN (GCN)	0.8472 ± 0.0141	6.00e−06 ± 1.00e−06
	GNN (GAT)	0.8517 ± 0.0183	1.40e−05 ± 1.00e−06
	Language-only (SBERT)	0.6993 ± 0.0114	2.48e−03 ± 6.20e−05
	Hybrid (GCN+TinyBERT)	0.8583 ± 0.0151	1.21e−05 ± 3.25e−06
Citeseer	GNN (GCN)	0.7254 ± 0.0153	6.00e−06 ± 1.00e−06
	GNN (GAT)	0.7185 ± 0.0100	1.10e−05 ± 1.00e−06
	Language-only (SBERT)	0.7630 ± 0.0142	2.41e−03 ± 6.30e−05
	Hybrid (GCN+TinyBERT)	0.7743 ± 0.0190	6.81e−06 ± 1.35e−06
WikiCS	GNN (GCN)	0.8171 ± 0.0050	3.00e−06 ± 0
	GNN (GAT)	0.8157 ± 0.0080	1.80e−05 ± 1.00e−06
	Language-only (SBERT)	0.6208 ± 0.0130	3.99e−03 ± 2.98e−04
	Hybrid (GCN+TinyBERT)	0.8116 ± 0.0123	1.12e−05 ± 3.71e−06
PubMed	GNN (GCN)	0.8099 ± 0.0028	1.00e−06 ± 0
	GNN (GAT)	0.8240 ± 0.0074	5.00e−06 ± 0
	Language-only (SBERT)	0.8280 ± 0.0045	2.68e−03 ± 4.50e−05
	Hybrid (GCN+TinyBERT)	0.8567 ± 0.0037	3.96e−06 ± 1.00e−06

ΔFR captures reproducibility, and FOR measures efficiency. The hybrid model exhibits varying computational overhead across datasets. However, ΔGTCE remains near zero and ΔCMF is consistently negative, indicating that fusion provides limited explainability gains and may introduce fairness degradation, while ΔFR is also close to zero, indicating no substantial improvement in reproducibility. Negative values indicate that the hybrid model underperforms the strongest unimodal baseline under the corresponding fusion-aware definition (Eq. (1)–(2)). Consistent with the radar profiles in Fig. 4, fusion improves transferability (ΔCMT) but yields limited gains in in-domain effectiveness (ΔSSP), and exhibits mixed or negative trends in responsibility-related metrics. Overall, these results show that efficiency gains do not necessarily translate into improved responsibility, and that fusion introduces non-trivial trade-offs not captured by accuracy-centric evaluation.

Table 6. Level-II (Stability) and Level-III (Responsibility) evaluation. ΔGTCE and ΔCMF quantify explainability and fairness gains, ΔFR measures reproducibility gain, and −FOR represents the efficiency gain (inverted overhead). Results are reported only for the hybrid model (GCN+TinyBERT), as fusion-aware gains are undefined for non-fusion baselines.

Dataset	ΔGTCE ↑	ΔCMF ↑	ΔFR ↑	−FOR ↑
Cora	0.0044	-0.0371	−2.0e−05	0.9969
Citeseer	-0.0009	-0.1065	−2.6e−04	0.9983
WikiCS	-0.0079	-0.0174	−4.7e−05	-2.4585
PubMed	0.0045	-0.0111	−4.0e−06	-2.3706

Figure 4. Dataset-wise fusion-efficiency radar profiles. Each radar summarises the normalised fusion trade-off profile for one case study. For target datasets, the profile includes ΔSSP, ΔCMT, −FOR, ΔMCR, ΔFR, ΔGTCE, and ΔCMF. For Cora, which serves as the source dataset in the fixed-source transfer setting, ΔCMT is not applicable and is therefore omitted. Values are min-max normalised across datasets and softly rescaled for visualisation.

5. Discussion

Our results show that fusion effects vary substantially across metrics and datasets, indicating that hybrid LLM–GNN systems should be understood as trade-offs rather than uniformly beneficial integrations. We highlight several insights on when fusion helps, when it fails, and why these behaviours arise.

- **Fusion gains are driven by distribution shift.** Fusion yields limited gains in in-domain performance (ΔSSP) but substantial improvements in cross-dataset transferability (ΔCMT), suggesting that combining structure-aware and semantic representations is most effective under distribution shift rather than aligned settings(e.g., on WikiCS and PubMed).

- **Observed robustness gains suggest complementary failure modes.** Although fusion reduces average degradation (MCR_{avg}), Table 3 shows no consistent per-perturbation improvement, indicating that robustness emerges from cross-modal error compensation rather than uniform resilience across structural and textual perturbations.

- **The evaluated fusion model shows increased fairness gaps.** Consistently negative ΔCMF values suggest that the evaluated hybrid model can worsen fairness gaps, likely because modality biases are aggregated without explicit calibration, particularly on WikiCS and PubMed.

- **Fusion behaviour is inherently multi-dimensional.** The radar profiles (Fig. 4) show that improvements in transferability and robustness can coexist with neutral or negative outcomes in fairness and explainability, confirming that accuracy alone is insufficient for evaluation. For instance, Citeseer shows stronger robustness but weaker transferability, whereas WikiCS shows the opposite.

Overall, these findings suggest that LLM–GNN fusion should be understood as a context-dependent trade-off mechanism rather than a uniformly beneficial integration. Its gains primarily emerge under distribution shift through complementary modality interactions, while in aligned settings the additional modality provides limited benefit and may introduce side effects such as fairness degradation. This indicates that improvements in hybrid systems are not intrinsic to fusion itself, but arise from how modality-specific biases and inductive strengths interact. The proposed framework makes these interactions explicit, providing a diagnostic lens for identifying when fusion yields genuine synergy and when it merely redistributes errors across modalities.

6. Conclusion

This work proposes a *fusion-aware evaluation framework for LLM–GNN systems* that moves beyond accuracy-centric assessment by jointly analysing *effectiveness, stability, and responsibility*. Our results show that fusion is better understood as a trade-off rather than a consistently beneficial integration: while it enhances transferability and robustness in our controlled case studies, its impact on in-domain accuracy is limited and it may introduce negative effects such as fairness degradation and minimal explainability gains. These findings highlight the importance of multi-dimensional evaluation for understanding hybrid systems and provide a structured framework for more transparent analysis. The framework is model-agnostic and can be directly extended to larger LLMs. While we include several datasets, our evaluation covers only a subset of available

benchmarks; future work should explore more architectures and broader benchmarks to improve generality.

Acknowledgments

This research was supported by the China Scholarship Council (No. 202408610044).

Impact Statement

This work proposes an *evaluation and reporting framework for hybrid Graph Neural Network (GNN) and Large Language Model (LLM) systems*, extending assessment beyond accuracy to *robustness*, *transferability*, and *responsibility*-oriented metrics. It provides a structured basis for researchers and practitioners to analyse system behaviour and compare methods more consistently. The framework's novelty is methodological, working well with existing models and datasets. Potential risks are indirect, as biases in underlying data may persist, and evaluation choices may influence research directions. Overall, this work emphasises that robust and responsible evaluation should be treated as a core component of Graph-LLM systems.

References

Dehghani, M., Tay, Y., Gritsenko, A. A., Zhao, Z., Houlsby, N., Diaz, F., Metzler, D., and Vinyals, O. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021.

Dodge, J., Gururangan, S., Card, D., Schwartz, R., and Smith, N. A. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2185–2194, 2019.

Dwivedi, V. P., Joshi, C. K., Luu, A. T., Laurent, T., Bengio, Y., and Bresson, X. Benchmarking graph neural networks. *Journal of Machine Learning Research*, 24 (43):1–48, 2023.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Giles, C. L., Bollacker, K. D., and Lawrence, S. Citeseer: An automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, pp. 89–98, 1998.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. Pmlr, 2017.

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., and Meger, D. Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., and Leskovec, J. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Luo, L., Li, Y., Haffari, G., and Pan, S. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *ICLR 2024: The Twelfth International Conference on Learning Representations*. ICLR, 2024.

McCallum, A. K., Nigam, K., Rennie, J., and Seymore, K. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163, 2000.

Mernyei, P. and Cangea, C. Wiki-cs: A wikipedia-based benchmark for graph neural networks. *arXiv preprint arXiv:2007.02901*, 2020.

Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Larochelle, H. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of machine learning research*, 22(164):1–20, 2021.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., and Eliassi-Rad, T. Collective classification in network data. *AI magazine*, 29(3):93–93, 2008.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Yao, Y., Wang, X., Zhang, Z., Qin, Y., Zhang, Z., Chu, X., Yang, Y., Zhu, W., and Mei, H. Exploring the potential of large language models in graph generation. *arXiv preprint arXiv:2403.14358*, 2024.

Zhou, K., Wang, X., Zhou, Y., Shang, C., Cheng, Y., Zhao, W. X., Li, Y., and Wen, J.-R. Crslab: An open-source toolkit for building conversational recommender system. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pp. 185–193, 2021.