When Structure Doesn't Help: LLMs Do Not Read Text-Attributed Graphs as Effectively as We Expected

Haotian Xu [1] Yuning You[2] Tengfei Ma [1]

[1]Stony Brook University [2]California Institute of Technology

Research Motivations & Questions

- **Motivation:** Despite recent progress in LLM–GNN hybrids, the true benefit of structural information remains uncertain.
- **Research Question:** Do LLMs truly need explicit structural encoding, or can semantic signals alone achieve strong graph reasoning?
- **Findings:** Structural encodings provide little to no performance gain and can even harm results when rich semantics exist.
- **Contribution:** This study systematically challenges the necessity of structural modeling and advocates a semantics-centered paradigm for LLM-based graph learning.

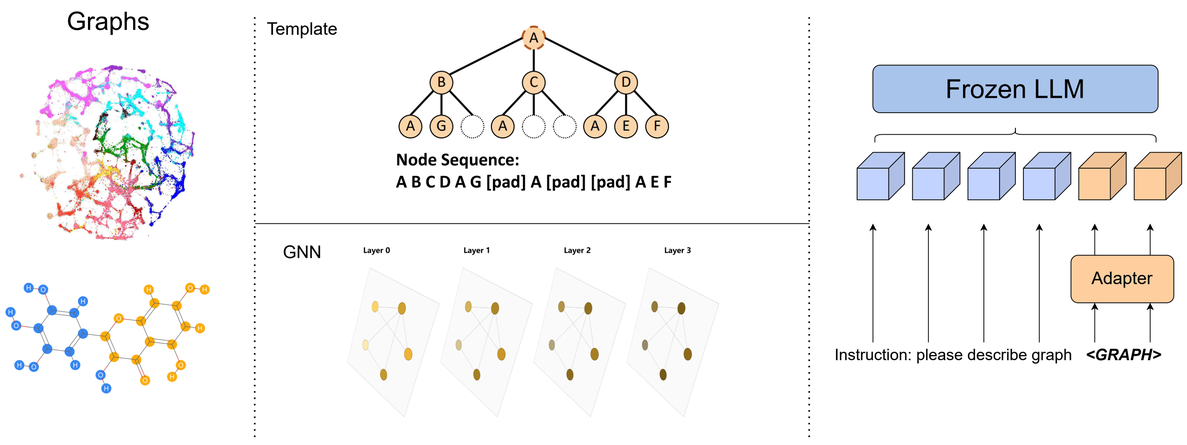


Figure 1. A common paradigm for aligning graph type data into LLMs.

Template-based Encodings

- **ND (Neighborhood Detail):** Uses a handcrafted k-hop B-tree subgraph and Laplacian positional encodings to encode structural context.
- **HN (Hop Neighbor):** Randomly samples a subset of k-hop neighbors to form an order-invariant sequence of node descriptions, removing explicit graph structure encodings.
- **CO (Center Only):** Provides only the central node description, entirely omitting neighbor information.

Table 1. ND underperforms HN and CO, especially on heterophilic graphs, showing that LLMs rely more on semantic cues than structural encodings, with isolated node semantics and unordered neighbor aggregation often sufficing.

Setting	Dataset	Node Classification			Link Prediction	
		ND	HN-1	CO	ND	HN-1
Homophilic	Cora	88.07% (0.74%)	**88.56%** (0.80%)	85.42% (1.78%)	85.56% (1.33%)	**87.27%** (1.56%)
	Citeseer	**80.31%** (0.81%)	80.20% (0.94%)	77.74% (0.31%)	86.73% (0.63%)	**88.79%** (0.84%)
	Pubmed	92.56% (0.71%)	94.80% (0.17%)	**94.84%** (0.04%)	88.25% (0.31%)	**90.98%** (0.38%)
Heterophilic	Shool	66.43% (3.69%)	82.02% (12.79%)	**91.13%** (1.66%)	**68.61%** (0.21%)	68.12% (1.51%)
	Roman Empire	48.56% (1.17%)	59.70% (2.42%)	**62.24%** (0.19%)	81.59% (0.50%)	**83.81%** (0.12%)
	Amazon Ratings	40.97% (0.56%)	**41.67%** (0.22%)	40.38% (1.14%)	80.26% (2.01%)	**84.51%** (0.53%)
Across Datasets		69.48%	74.49%	**75.29%**	81.83%	**83.91%**

GNN-based Encodings

Table 2. Replacing GNNs with simple MLPs in the GraphToken framework shows comparable performance, indicating that LLMs can rely on semantic representations alone without explicit structural modeling. **Best** results are bolded, second best are underlined.

Setting	Dataset	Node Classification			
		MLP	GCN	GAT	GIN
Homophilic	Cora	87.09% (0.66%)	87.64% (0.84%)	**88.25%** (0.53%)	83.03% (5.41%)
	Citeseer	79.39% (1.38%)	**80.20%** (0.13%)	79.74% (0.41%)	79.32% (1.11%)
	Pubmed	**94.76%** (0.10%)	92.24% (1.23%)	92.01% (0.24%)	91.40% (0.63%)
Heterophilic	Shool	**90.17%** (3.62%)	67.87% (3.24%)	64.75% (0.00%)	70.02% (2.19%)
	Roman Empire	**65.39%** (0.29%)	36.51% (18.06%)	36.97% (13.92%)	46.92% (22.37%)
	Amazon Ratings	**40.78%** (0.35%)	40.52% (0.51%)	40.71% (0.23%)	38.76% (0.18%)
Across Datasets		**76.26%**	67.50%	67.07%	68.24%

Setting	Dataset	Link Prediction			
		MLP	GCN	GAT	GIN
Homophilic	Cora	90.72% (0.85%)	90.51% (1.19%)	**91.05%** (0.93%)	87.86% (1.20%)
	Citeseer	87.67% (2.71%)	**89.32%** (0.53%)	88.53% (0.46%)	78.34% (1.99%)
	Pubmed	**89.14%** (0.19%)	89.11% (0.37%)	88.58% (0.38%)	87.54% (0.55%)
Heterophilic	Shool	59.40% (1.92%)	59.40% (3.26%)	**62.78%** (3.98%)	56.55% (1.25%)
	Roman Empire	51.60% (0.62%)	52.64% (0.68%)	51.00% (1.02%)	**53.63%** (0.24%)
	Amazon Ratings	**72.59%** (0.34%)	72.10% (1.04%)	66.24% (11.19%)	71.51% (0.19%)
Across Datasets		75.19%	**75.51%**	74.70%	72.57%

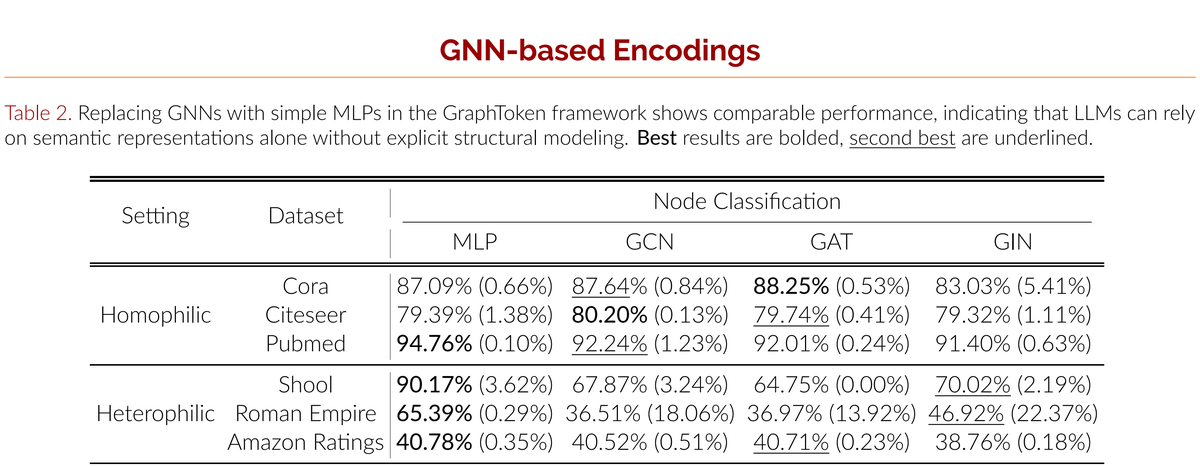


Figure 2. Increasing adapter depth in GraphToken degrades performance with GNNs but has little effect with MLPs, suggesting that rich semantic representations alone suffice for LLM-based graph reasoning, while explicit structural augmentation can be unnecessary or detrimental.

Findings on Text-Attributed Graphs Experiments

- **Template Encodings:** ND underperforms structure-free variants (HN, CO), indicating structural priors may hinder LLM reasoning.
- **Semantic Reliance:** LLMs perform well on TAGs using only node semantics, treating graph reasoning as a set-based task.
- **GNN Encodings:** Replacing GNNs with MLPs yields similar results, showing limited benefit from structural modeling.
- **Depth Effect:** Deeper GNN adapters reduce performance, suggesting overfitting and diminishing returns.
- **Conclusion:** Structural encodings add little value—LLMs reason effectively from semantic cues alone.

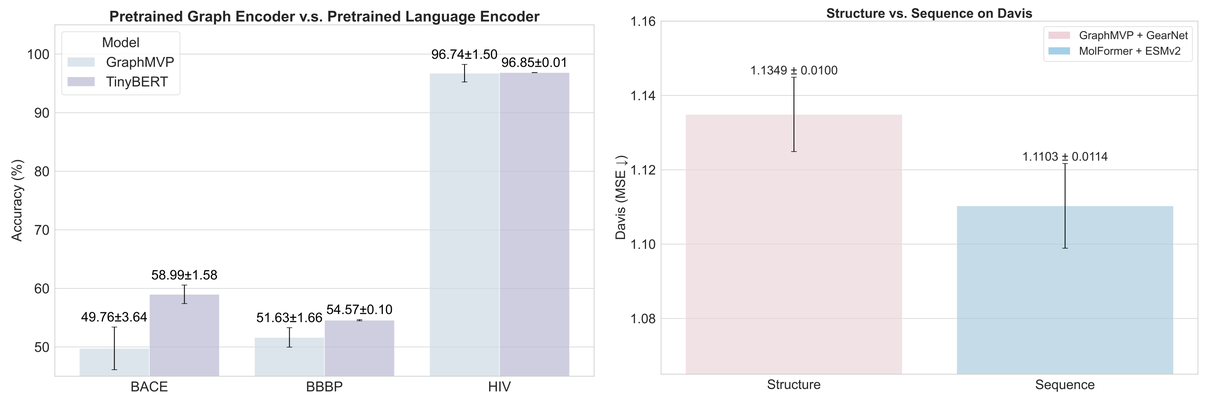
Natural Molecular Graphs

Table 3. LLMs can effectively perform molecular graph tasks using only rich node-level semantic embeddings, with explicit structural modeling providing little to no additional benefit.

Dataset	Molecular Property Prediction			
	MLP	GCN	GIN	GAT
BACE	**58.99%** (1.66%)	58.77% (9.13%)	**58.99%** (5.52%)	57.46% (3.62%)
BBBP	54.57% (1.38%)	57.84% (0.49%)	**60.29%** (0.49%)	51.96% (1.47%)
HIV	**96.85%** (0.01%)	96.81% (0.03%)	96.79% (0.00%)	96.82% (0.03%)

Pretrained Encoders and Geometric Deep Learning

Figure 3. *Left:* For molecular graphs with intrinsic structural priors, LLMs using semantic embeddings from pretrained language models outperform those using graph-specific encoders, highlighting that semantic content dominates over explicit structural information. *Right:* Even on structurally demanding datasets like Davis DTI, LLMs using sequence-based semantic encodings perform comparably to or better than structure-based encoders, suggesting that current graph benchmarks often overemphasize structural reasoning.



Insights for Experiments on Molecule and Protein Graphs

- **Semantic Dominance on Molecular Graphs:** Even simple MLPs without structural modeling match or outperform GNN-based adapters, showing LLMs can rely solely on semantic embeddings.
- **Pretrained Encoders:** Comparing GraphMVP (graph encoder) and TinyBERT (language encoder) reveals no consistent advantage for structural pretraining, even in chemistry.
- **Broader Implication on Geometric Deep Learning:** Structural information offers marginal gains, calling for benchmarks that better capture relational and semantic reasoning.
- **Key Insight:** High-quality semantic embeddings, not explicit topology, primarily determine LLM performance on graph reasoning tasks.

Scaling Ineffectiveness & Semantic Content

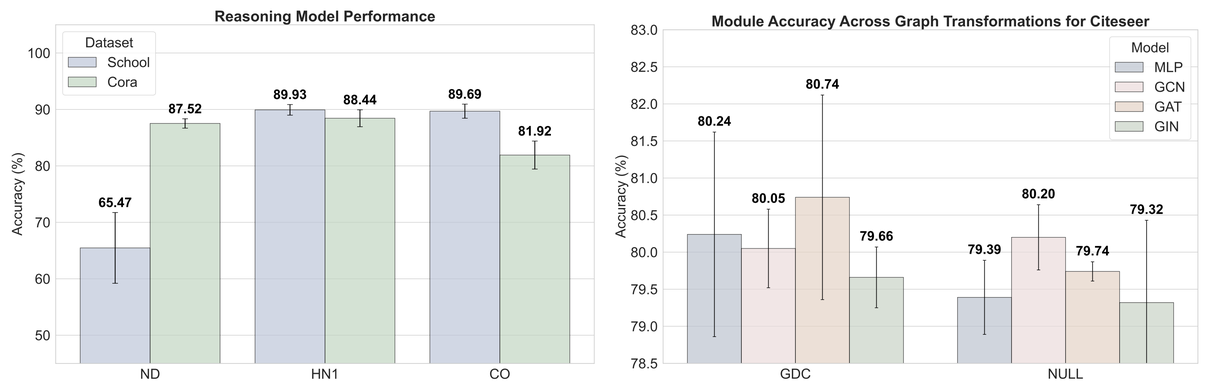
Table 4. Switching LLM backbones preserves our finding that structure may be unnecessary for LLMs processing graphs. Even with weak semantic content, LLMs still reveal the same pattern.

Model Architecture	Dataset	Node Classification		Link Prediction	
		ND	HN-1	ND	HN-1
Llama2-7B	Cora	87.76%(0.21%)	**88.01%**(0.56%)	85.48%(0.38%)	**87.04%**(0.75%)
	School	70.98%(0.83%)	**92.09%**(2.49%)	61.82%(2.88%)	**69.09%**(1.92%)
Llama2-13B	Cora	**87.58%**(0.59%)	87.45%(0.19%)	84.24%(0.89%)	**86.05%**(0.55%)
	School	69.30%(3.24%)	**89.45%**(3.40%)	61.21%(1.28%)	**67.15%**(1.52%)

Semantic Content	Dataset	Node Classification		Link Prediction	
		ND	HN-1	ND	HN-1
sparse	Cora	**83.96%**(2.74%)	82.17%(0.56%)	69.19%(1.15%)	**74.81%**(0.85%)
	School	56.95%(6.19%)	**73.62%**(7.21%)	63.63%(0.63%)	**65.09%**(3.93%)
full	Cora	83.39%(0.37%)	**84.81%**(0.46%)	70.81%(1.89%)	**75.84%**(0.74%)
	School	59.47%(3.97%)	**60.19%**(1.10%)	63.15%(5.91%)	**70.06%**(3.30%)

More Recent LLMs and Approaches to Better Leverage Structures

Figure 4. *Left*: Though reasoning model can perform structured decision-making, it does not rely on structure information. *Right*: Altering the node sequence via GDC can gain some enhancement at a time.



Large Text-Attributed Graphs

Table 5. Our findings hold consistently on larger text-attributed graphs, suggesting that structural information contributes only marginally to LLMs' graph inference.

Dataset	Node Classification		
	ND	HN-1	CO
Products	83.45% (0.39%)	**83.87%** (0.24%)	80.10% (0.27%)
ArXiv	**75.65%** (0.50%)	75.41% (0.21%)	74.46% (0.18%)

Ablation Summarization

- **Scaling Ineffective:** Increasing LLaMA model size from 7B to 13B does not improve structural sensitivity; structure-free templates often perform better.
- **Semantic Content:** Even with sparse node descriptions, structure-free templates match or exceed structure-aware ones, showing LLMs rely mainly on semantics.
- **Large Reasoning Models:** Even reasoning-oriented models (e.g., Nemotron-7B) gain little from explicit structural encodings like Laplacian embeddings.
- **Better Leveraging Structure:** Graph Diffusion Convolution (GDC) modestly improves performance by emphasizing long-range dependencies, suggesting minimal but useful structural cues.
- **Dataset Size:** Larger graph datasets (Products, ArXiv) show no significant performance gap between structure-aware and structure-free settings.