CITE: Benchmarking Heterogeneous Text-Attributed Graph Models

Chenghao Zhang[1,2], Qingqing Long[1,2], Ludi Wang[1,2], Wenjuan Cui[1,2], Jianjun Yu[1,2,*], Yi Du[1,2,3,*]

[1]Computer Network Information Center, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences
{chzhang,qqlong,wld,wenjuancui}@cnic.cn, yujj@cnic.ac.cn, duyi@cnic.cn
[*]Corresponding Authors

Abstract

Recent advances in large language models (LLMs) and text-aware graph learning have increased interest in reasoning over text-attributed graphs (TAGs). In many real-world settings, such graphs are inherently heterogeneous, with most existing benchmarks remaining largely homogeneous in structure. As a result, the lack of large-scale benchmarks for heterogeneous text-attributed graphs has hindered systematic evaluation and fair comparison of existing methods. In this work, we introduce CITE - **C**atalytic **I**nformation **T**extual **E**ntities Graph, the first and largest heterogeneous text-attributed citation graph benchmark for catalytic materials. CITE contains over 438K nodes and 1.2M edges spanning four node types and four relation types, with rich node-level textual information. We establish standardized evaluation protocols for node classification and link prediction, and conduct ablation studies to assess the impact of graph heterogeneity and textual attributes. Using CITE, we benchmark four classes of learning paradigms, including homogeneous graph models, heterogeneous graph models, LLM-centric models, and LLM+Graph models. By providing a large-scale heterogeneous text-attributed benchmark together with standardized evaluation protocols and comprehensive baselines, CITE enables systematic assessment across diverse modeling paradigms and offers new insights into text-aware and LLM-enhanced graph learning. The dataset[1], codebase and evaluation suite[2] are publicly available.

1 Introduction

Graphs are frequently employed in the modeling of relationships and structures of objects in the real-world, covering a wide range of scenarios, such as citation networks (Cohan et al., 2020; Ju et al., 2024a), social networks (Myers et al., 2014), and recommendation systems (Linmei et al., 2019; Long et al., 2026). Moreover, many nodes of real-world graphs are often associated with textual attributes, leading to TAGs (Fang et al., 2021, 2022; Yang et al., 2021). In the context of TAGs, nodes are conventionally employed to represent text entities, such as documents or sentences (Ju et al., 2024b). In addition, there are multiple types of nodes and links in the real world (Mao et al., 2025; Yan et al., 2024). For instance, in a citation network, papers are connected together via papers, authors, venues, and keywords. Compared to homogeneous graphs, heterogeneous graphs are capable of processing and containing more information in nodes and links, and consequently establish a new development of data mining (Wang et al., 2020a; Zhang et al., 2026).

The key to learning on TAGs is the effective integration of two key elements: the node attributes, which encompass the textual semantics, and the graph topology, which encompasses the structural connections. Existing methods for TAGs representation learning are typically divided into three categories. Firstly, methods based on pretrained language models (PLMs) are effective at capturing the semantic content of textual information associated with nodes, enabling the characterization of individual node properties (Luo et al., 2025; Yan et al., 2023). In PLM-based methods, the textual content from the target node is typically input into the model, transforming the task into a text classification or link prediction problem (Chen et al., 2024b; Yasunaga et al., 2022). However, they often overlook the critical topological knowledge of graph structures (Zhao et al., 2022). Despite recent efforts to integrate graph structure with PLMs, such as using GNN-generated structural embeddings or graph-aware tokenization, these approaches have demonstrated unstable performance across different scenarios (Huang et al., 2024; Tang et al., 2024; Zhang et al., 2024).

[1]https://huggingface.co/datasets/kg4sci/CITE/tree/main
[2]https://github.com/kg4sci/CITE-NEW

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 31426–31448
July 2-7, 2026

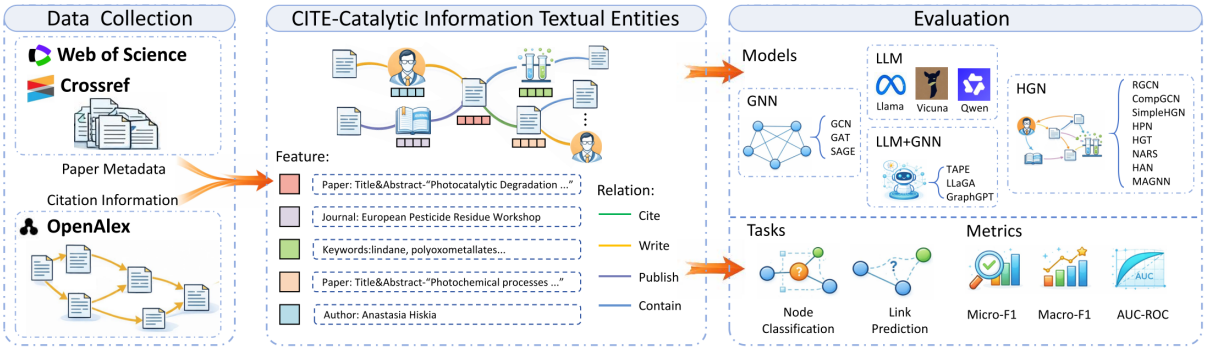

Figure 1: Overview of CITE structure and content.

Secondly, methods based on Graph Neural Networks (GNNs), can effectively capture structural relationships within a graph by leveraging the message-passing mechanism (Lin et al., 2024; Long et al., 2020). However, GNNs exhibit limited capacity when handling large-scale textual information. Some approaches attempt to address this by using LLMs to preprocess node text and feed embeddings into GNNs. However, GNNs generally treat node attributes as static, which hampers gradient backpropagation and limits end-to-end training (Wang et al., 2020b; Yan et al., 2023). Furthermore, most efforts focus on homogeneous graphs, with less exploration in heterogeneous settings.

Thirdly, in order to reap the benefits of both methods, recent research has illuminated the co-training paradigm, which can be called aligner methods (Jin et al., 2023; Zhao et al., 2022). The basic idea of the aligner methods is to enable GNNs and PLMs to facilitate the learning of topology and textual information, respectively. However, these methods can result in cumbersome models and severe scalability issues.

Current TAG representation learning frameworks primarily target homogeneous graphs. In contrast, real-world graphs often exhibit structural heterogeneity through diverse node and link types. For example, academic citation networks involve multiple entity types, such as papers, authors, journals, and keywords. In heterogeneous TAGs, nodes and edges have type-specific semantic properties, with entity features and relational patterns reflecting the complex structure and domain-specific knowledge of real-world systems.

Conventional homogeneous TAG representation learning methods struggle to capture type-specific attribute distributions and cross-type interactions, as they rely on uniform attribute spaces and single relation mechanisms. The lack of heterogeneous TAG datasets has hindered progress in this field, as existing datasets like OGB (Hu et al., 2020a), Cora (Sen et al., 2008), PubMed (White, 2020), CiteSeer (Giles et al., 1998) primarily focus on homogeneous graphs and fail to support heterogeneity or leverage text attributes. Thus, there is an urgent need for heterogeneous text-attributed graphs to advance TAG representation learning and support the development of next-generation graph learning architectures.

To address these gaps, we introduce CITE - **C**atalytic **I**nformation **T**extual **E**ntities, as illustrated in Figure 1. CITE is a large-scale heterogeneous TAG benchmark designed to support systematic evaluation of graph- and LLM-based models. We further conduct extensive experiments on CITE to establish reliable baselines and analyze model behavior under diverse settings. Our contributions are summarized as follows:

- We introduce CITE, a publicly available large-scale heterogeneous text-attributed citation graph benchmark that integrates rich textual information with multi-type entities and relations.
- We establish a comprehensive benchmarking suite on CITE that systematically evaluates a wide range of models, including homogeneous and heterogeneous graph models, LLM-centric models, and LLM+Graph models, across node classification and link prediction tasks.
- We provide extensive empirical analyses that reveal how structural heterogeneity, textual attributes, and label imbalance affect model performance, offering diagnostic insights into the capabilities and limitations of existing representation learning methods.

2 CITE Dataset

2.1 Overview of CITE

CITE builds a heterogeneous network across the domains of photocatalysis and electrocatalysis, based

Dataset	Nodes	Edges	Class	Modeling	Node types	Heterogeneity	Task	Raw-text	Multi-label
Cora (McCallum et al., 2000)	2,708	5,429	7	One-hop	1	✗	NC	✗	✗
CiteSeer (Giles et al., 1998)	3,312	4,732	6	One-hop	1	✗	NC	✗	✗
PubMed (White, 2020)	19,717	44,338	3	Tf-IDF	1	✗	NC	✗	✗
Ogbn-arXiv (Hu et al., 2020a)	169,343	1,166,243	40	Word2Vec	1	✗	NC	✓	✗
ACM (Wang et al., 2019)	8,860	-	56	BoW	2	✓	LP	✗	✗
CITE	**438,304**	**1,220,374**	**84/107**	**Bert**	**4**	✓	**NC, LP**	✓	✓

Table 1: Statistics of graphs. NC presents node classification, LP presents link prediction.

on peer-reviewed publications from 1922 to 2022. The network includes four types of nodes: Paper, Author, Journal, and Keyword, and their semantic relationships, such as citations, authorship, publication venues, and topics. With 438,304 nodes and 1,220,374 edges, CITE provides a comprehensive representation of the field's evolution, linking structural metadata, such as citation graphs, with rich textual attributes like abstracts and section content.

Compared to existing citation graph datasets, CITE excels in both heterogeneity and semantic integration. By unifying diverse node types as first-class entities, it supports nuanced analyses such as institutional collaboration dynamics and keyword-driven topic evolution, which are capabilities absent in homogeneous citation networks. CITE also embeds abstract content and keywords directly into graph nodes, enabling cross-modal queries that link methodological details to citation patterns, thus providing a robust foundation for longitudinal studies. As quantified in Table 1, CITE outperforms benchmarks in node diversity and attribute richness.

A key feature of CITE is its multi-label classification scheme, which allows each paper to be associated with more than one disciplinary label. The label distribution includes 84 subfields under the single-label setup and 107 subfields under the multi-label setup. For detailed statistics please refer to Appendix B.1.

CITE is publicly hosted on https://huggingface.co/datasets/kg4sci/CITE/tree/main, including raw JSON files metadata and processed structured files (JSON, CSV, and pt formats), ensuring transparent access and reproducibility.

2.2 Collection and Construction of CITE

CITE is a heterogeneous citation graph derived from peer-reviewed publications in photocatalysis and electrocatalysis, spanning from 1922 to 2022. It includes four types of nodes: Paper, Author, Journal, and Keyword, each associated with relationships such as Paper-Paper, Paper-Author, Paper-Journal, and Paper-Keyword. Paper nodes represent articles, with attributes such as title and abstract, and are classified according to SCI journal categories. CITE supports multi-label classification, allowing each paper to be associated with multiple disciplinary labels. Author nodes are identified by ORCID or institutional affiliation, while journal nodes are linked via ISSN. Keyword nodes represent the core research topics of each paper. Each entity is assigned a unique identifier to ensure data consistency. For more detailed metadata descriptions and the construction pipeline, please refer to Appendix B.2.

CITE is sourced from DOIs indexed in Web of Science, while citation relationships are derived from OpenAlex, ensuring the dataset's credibility for citation analysis and interdisciplinary research. The graph construction process, including extraction, cleaning, and normalization, involved multiple validation steps to ensure data quality. For further details on these processes, please refer to Appendix B.2.

3 Experiment

CITE is a heterogeneous textual attribute graph designed for scientific literature analysis in catalytic materials, integrating multiple node types and rich text attributes. Its domain-specific characteristics and compound heterogeneity-textual properties introduce unique challenges in modeling semantic interaction patterns, multi-granularity integration, and architectural adaptability across learning paradigms. To systematically assess the value of CITE, we design our experiments around three core research questions: **Q1**: What is the impact of heterogeneous text-attributed graph architectures on model generalization across diverse learning paradigms? **Q2**: Does CITE present unique challenges compared to existing citation graph benchmarks that current methods cannot address? **Q3**: Are multi-relational structures and textual semantic richness essential for robust graph representation learning?

To address these questions, we employ four

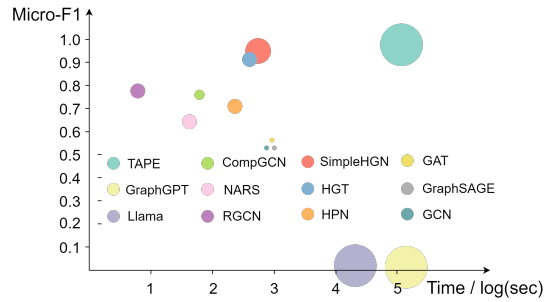

Figure 2: Performance, efficiency, and model scales.

learning paradigms: (1) Homogeneous graph models, (2) Heterogeneous graph models, (3) LLM-centric models, (4) LLM+Graph models. All experiments follow a 60%/20%/20% split for training, validation, test, and all graph models use BERT-base tokenization with 128-token truncation. In addition to the standard random split, we evaluate two time-aware protocols to assess potential temporal information leakage in citation networks, the full results are reported in Appendix D.2.

We focus on two key graph learning tasks: node classification, aiming to predict discipline-specific SCI taxonomy labels for scientific papers, and link prediction, which evaluates the ability to infer missing relational links based on heterogeneous structural and semantic cues. For comparability with prior TAG benchmarks and broad baseline coverage, the main text reports results on the single-label version of CITE. In addition, since CITE also supports multi-label categorization, we further construct a fine-grained multi-label node classification benchmark and report the full setup and results in Appendix D.5. Node classification performance is assessed using Macro-F1 and Micro-F1 in the single-label setting, while the multi-label setting is evaluated with standard multi-label F1 metrics. Link prediction performance is evaluated with AUC-ROC.

Because journal information is correlated with paper labels in this benchmark, we additionally report a stricter journal-masked evaluation and diagnostic analyses in Appendix D.4.

3.1 Baseline Model Comparison *(Q1)*

3.1.1 Experiment Setup

To systematically assess how heterogeneous text-attributed graph architectures impact model generalization across learning paradigms, we evaluate the four learning paradigms mentioned above: (1) Homogeneous graph models: GCN (Kipf and Welling, 2016), GAT (Velickovic et al., 2017), GraphSAGE (Hamilton et al., 2017); (2) Heterogeneous graph models: RGCN (Schlichtkrull et al., 2018), CompGCN (Vashishth et al., 2019), SimpleHGN (Lv et al., 2021), HPN (Ji et al., 2021), HGT (Hu et al., 2020b), NARS (Yu et al., 2020), HAN (Wang et al., 2019), MAGNN (Fu et al., 2020); (3) LLM-centric models: Llama-2-7B (Touvron et al., 2023), Vicuna-7B-v1.5 (Chiang et al., 2023), Llama-3.1-8B (Grattafiori et al., 2024), Qwen2.5-7B (Yang et al., 2025). (4) LLM+Graph models: TAPE (He et al., 2023) (LLM enhances GNN in prediction), LLaGA-HO (Chen et al., 2024a), and GraphGPT (Tang et al., 2024) (graph construction enhances LLM in prediction). Details about the models and their hyperparameters are provided in Appendix E.1.

To ensure fairness and isolate the effect of heterogeneity, architecture-specific adjustments are applied. For homogeneous models, we treat all heterogeneous nodes as a single type. Training, validation, and testing tasks are performed only on the paper nodes. For LLM-based models, we design structured prompt templates to inject heterogeneous metadata. The complete prompt templates for LLM-centric and LLM+Graph models are provided in Appendix C.1. All models adopt identical encoding models and data split rates. All paradigms are evaluated on node classification task, link prediction task are conducted only for heterogeneous graph models that natively support that, the target link is paper-paper. Figure 2 illustrates model performance relative to efficiency, with the x-axis representing the logarithm of execution time (in seconds) and the y-axis representing F1-scores. All experiments are performed on an NVIDIA A100 GPU. Homogeneous models have approximately 8M parameters; heterogeneous models range from 14M to 120M and LLM-based models have 7B parameters. The complete quantitative results are presented in Table 2, where the top three scores for each metric are highlighted in red, yellow, and blue. Sensitivity analysis is provided in Appendix D.1.

3.1.2 Discussion

(1) Heterogeneous models benefit from flexible, type-aware representations: Experimental results show that heterogeneous graph models with explicit relation-type modeling generally outperform homogeneous models. This is because homogeneous models fail to capture type-specific seman-

Category	Model Name	Micro-F1	Macro-F1	AUC-ROC
Homogeneous Graph Model	GCN	0.5057	0.0239	-
	GAT	0.5115	0.0457	-
	GraphSAGE	0.5117	0.0474	-
Heterogeneous Graph Model	RGCN	0.8626	0.1442	0.6990
	CompGCN	0.8173	0.2184	0.7153
	SimpleHGN	0.9907	0.5052	0.7392
	HPN	0.7863	0.1387	0.4978
	HGT	0.9791	0.4183	0.7288
	NARS	0.7273	0.0660	-
	HAN	0.7515	0.1706	0.6826
	MAGNN	0.5542	0.0268	0.5542
LLM-centric Model	Llama 2	0.0623	0.0037	-
	Vicuna	0.0077	0.0002	-
	Llama 3.1	0.0872	0.0061	-
	Qwen 2.5	0.0981	0.0094	-
LLM+Graph Model	GraphGPT	0.0214	0.0039	-
	TAPE	0.9936	0.5655	-
	LLaGA	0.3516	0.0304	-

Table 2: Performance across learning paradigms on CITE, where "-" denotes the model does not support the task.

tics. The limitation is amplified in CITE's highly multi-label distribution, where isotropic aggregation and the inability to distinguish heterogeneous nodes fail to leverage high-signal node types like journals.

Among heterogeneous models, dynamic graph models such as SimpleHGN and HGT consistently outperform those based on fixed meta-paths like MAGNN and HAN. The latter rely on predefined paths such as paper-author links, which carry low informational density. These edges, despite their abundance, carry low informational density and contribute to over-smoothing rather than semantic enrichment. This trend holds across tasks, highlighting that flexibility in capturing and transferring relational semantics is crucial for robust generalization.

(2) GNN-based LLM-graph integration improves robustness, but design choices matter: LLM+Graph models exhibit duality: TAPE uses a two-stage framework, first using LLM to predict pseudo-labels from paper text, then injecting these labels into a graph model for structural refinement. This decoupling mitigates the stochastic nature of LLM outputs, allowing TAPE to achieve robust performance. Conversely, GraphGPT explores an end-to-end paradigm where graph embeddings encoded by graph models serve as input to LLM for text-based classification. The performance was found to be unsatisfactory. We provide a detailed analysis of these challenges in Subsection 3.1.3.

To further understand the performance gap between these two LLM+Graph paradigms, we additionally conduct an ablation on TAPE by removing its graph propagation module. In this variant, the graph model prediction is made directly from the LLM-derived outputs. As shown in Table 3, removing graph propagation causes a substantial performance drop,. This result suggests that TAPE's advantage does not simply come from combining LLMs with graph information at the input level; rather, it mainly arises from explicit structural propagation on the true graph topology. Therefore, designs that use LLMs to provide semantic supervision and use graph models for topology-aware refinement appear to be more effective than end-to-end prompting schemes that ask LLMs to directly reason over graph-encoded inputs.

Method	Uses LLM	Uses graph propagation	Micro-F1	Macro-F1
GraphGPT	✓	×	0.0214	0.0039
TAPE w/o graph propagation	✓	×	0.0202	0.0027
TAPE	✓	✓	0.9936	0.5655

Table 3: Comparison of GraphGPT, TAPE w/o graph propagation, and TAPE.

(3) Output inconsistency remains a challenge for LLM-based models in multi-class settings: Both LLM-centric and LLM+Graph models face challenges in highly multi-class scenarios, with outputs showing non-deterministic behavior. TAPE's decoupled design helps mitigate this issue by leveraging stable graph model outputs for label prediction. Even the best heterogeneous graph and LLM+Graph models still struggle with minority class representation, as indicated by the average Macro-F1 score of 0.5.

3.1.3 Analysis of LLM-Based Models

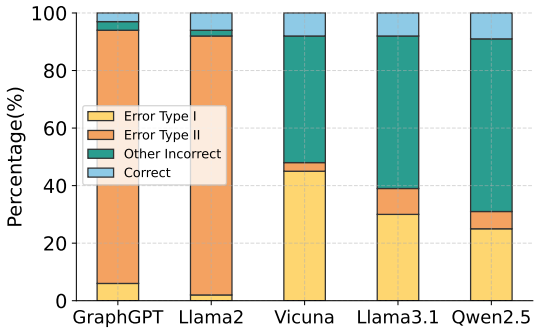

Figure 3: Response statistics of GraphGPT, Llama 2, Vicuna, Llama 3.1 and Qwen 2.5.

To better understand the degraded performance of GraphGPT and LLM-centric models on CITE, we conduct an in-depth analysis of their response behaviors and error patterns. We categorize fail-

ure cases into out-of list (Error Type I) and question repetition (Error Type II). Models' detailed responses are provided in Appendix C.2, and Figure 3 summarizes the corresponding statistics.

To explore the causes of LLM performance issues, we conducted a small-scale prompt-tuning experiment on a 200-node subgraph of CITE, testing four methods:

- Multi-turn dialogue: Iteratively narrowing the label set by asking follow-up questions.
- Candidate-set reduction: Providing the LLM with only the 10 most probable labels instead of the full 84-label list.
- Forced-choice prompting: We convert the task into a strict multi-class classification setting in which the model must output exactly one label set (ID or name) without any additional text.
- Likelihood ranking: We cast labeling as a likelihood-based multiple-choice problem over a fixed label set. The "ranking" is computed externally from token probabilities.

See detailed settings and results in Appendix D.6.

The results in Figure 4 indicate that prompt tuning and exploring LLM architectures improve performance, though they do not fully address the core challenges, including:

- Limited structural grounding and label alignment: LLM-based models lack mechanisms to capture relational dependencies in graph data, making label alignment difficult.
- Domain-specific limitations: Specialized terminology in the catalysis domain (e.g., Z-scheme vs. S-scheme photocatalysts) poses challenges for general-purpose LLMs without fine-tuning or domain adaptation.

While current LLMs have limitations, improvements in scaling model parameters, adding structured signals, and fine-tuning on domain-specific data could enhance performance.

To examine whether the output instability observed on CITE is dataset-specific or reflects a broader limitation of LLM-based TAG classification, we further evaluate four instruction-tuned LLMs on ACM and arXiv under three output protocols: original free-form generation, forced-choice prompting, and likelihood ranking. A clear label-space effect emerges: predictions are much more stable on ACM (3 classes) than on arXiv (40 classes), while CITE (84 classes) remains the most challenging setting. Across datasets, stricter

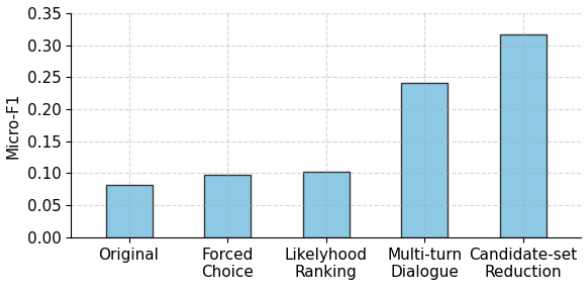

Figure 4: Performance of Llama 3.1 on prompt tuning.

protocols consistently reduce invalid outputs and improve performance. However, the gains become limited as the label space grows and remains highly fine-grained, and they do not close the gap between LLM-only methods and graph-based or hybrid approaches on CITE. These results suggest that the weakness of LLM-only models on CITE is not merely a prompt artifact, but is closely tied to label alignment difficulty in large multi-class TAG settings. Full results are provided in Appendix D.6.

1st Finding

On large-scale heterogeneous TAGs, models with flexible and relation-aware representations achieve stronger performance, while LLM-based methods remain sensitive to label alignment and output consistency.

3.2 Comparison with Existing Datasets *(Q2)*

3.2.1 Experiment Setup

To address Q2, we compare model performance on specialized benchmarks against CITE. We select four benchmark categories: (1) homogeneous non-text-attributed graphs (arXiv, Cora, PubMed) (Li et al., 2024); (2) homogeneous text-attributed graphs (Ogbn-arXiv (Hu et al., 2020a), Cora (Chen et al., 2024b), PubMed (Chen et al., 2024b)); (3) heterogeneous non-text-attributed graphs (ACM, DBLP) (Han et al., 2022); and (4) CITE, our heterogeneous text-attributed graph dataset. Details about the datasets are provided in Appendix E.2.

For node classification task, we include a subset of heterogeneous graph models together with all other model categories. For link prediction task, we select several representative heterogeneous models, and perform link prediction over all edge types available in each dataset. For dataset, we paired different kind of models with their "native" benchmarks: homogeneous graph models on non-text

version of arXiv, Cora, and PubMed, heterogeneous graph models on ACM and DBLP, and LLM-centric and LLM+Graph models on text version of Ogbn-arXiv, Cora, and PubMed. All models share identical hyperparameters, text encoders, and data split rates from Q1. Node classification results are reported in Table 4, while link prediction results are shown in Table 5, where 'p' stands for paper, 'j' stands for journal, and 'k' stands for keywords.

Model Category	Model	Dataset	Micro-F1	Macro-F1
Homogeneous Graph Model	GCN	Cora(non-text)	0.8125	0.8237
		PubMed(non-text)	0.8611	0.8587
		arXiv(non-text)	0.8621	0.8505
		CITE	**0.5057**	**0.0239**
	GAT	Cora(non-text)	0.8603	0.8488
		PubMed(non-text)	0.8640	0.8493
		arXiv(non-text)	0.7179	0.5496
		CITE	**0.5115**	**0.0457**
	GraphSAGE	Cora(non-text)	0.8125	0.8237
		PubMed(non-text)	0.8667	0.8586
		arXiv(non-text)	0.7185	0.5496
		CITE	**0.5117**	**0.0474**
Heterogeneous Graph Model	HGT	DBLP	0.8419	0.8334
		ACM	0.8803	0.8762
		CITE	**0.9791**	**0.4183**
	SimpleHGN	DBLP	0.9303	0.9245
		ACM	0.9063	0.9046
		CITE	**0.9907**	**0.5052**
	RGCN	DBLP	0.6942	0.5845
		ACM	0.7809	0.7567
		CITE	**0.8626**	**0.1442**
	HAN	DBLP	0.9033	0.8926
		ACM	0.9033	0.8926
		CITE	**0.7515**	**0.1706**
	MAGNN	DBLP	0.9107	0.8992
		ACM	0.6357	0.4986
		CITE	**0.5542**	**0.0268**
LLM-centric	Llama	arXiv(text)	0.0045	0.0015
		Cora(text)	0.2962	0.0936
		PubMed(text)	0.8402	0.8184
		CITE	**0.0623**	**0.0037**
LLM+Graph Model	TAPE	arXiv(text)	0.7275	0.5123
		Cora(text)	0.7573	0.7401
		PubMed(text)	0.8022	0.7948
		CITE	**0.9936**	**0.5655**
	GraphGPT	arXiv(text)	0.6258	0.2622
		Cora(text)	0.1256	0.0819
		PubMed(text)	0.7011	0.6491
		CITE	**0.0214**	**0.0039**

Table 4: Node classification results across datasets.

3.2.2 Discussion

(1) Long-tail distribution remains a major challenge: In the node classification task, we observe a significant gap between Micro-F1 and Macro-F1 scores across all models, highlighting the impact of class imbalance in CITE. This issue is notable for homogeneous graph models and LLM-based models. We compiled a frequency-bucket chart, which groups label counts into High (> 1000), Medium (101-1000), Low (11-100), and Rare ($\leq$ 10). As shown in Figure 6, across all models, bucket-wise accuracy declines with label frequency. This pattern confirms that the long-tail distribution in

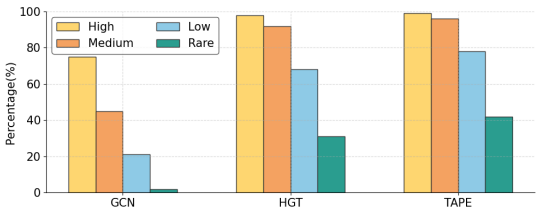

Figure 5: Bucket-wise label frequency and model accuracy

Model	Dataset	Edge Type				
		p-p	p-j	j-p	p-k	k-p
RGCN	ACM	0.7925	0.9636	0.3307	-	-
	DBLP	-	0.5000	0.2114	-	-
	CITE	**0.6990**	**0.9343**	**0.3870**	**0.6693**	**0.8065**
CompGCN	ACM	0.8645	0.9882	0.6366	-	-
	DBLP	-	0.9777	0.705	-	-
	CITE	**0.7153**	**0.9991**	**0.6845**	**0.8802**	**0.9203**
HAN	ACM	0.6902	0.9811	0.6782	-	-
	DBLP	-	0.7275	0.1923	-	-
	CITE	**0.6826**	**0.8709**	**0.5098**	**0.6969**	**0.8707**
MAGNN	ACM	0.4945	0.5187	0.4994	-	-
	DBLP	-	0.7187	0.4848	-	-
	CITE	**0.5542**	**0.7683**	**0.8171**	**0.7357**	**0.8107**
SimpleHGN	ACM	0.9083	0.9933	0.5761	-	-
	DBLP	-	0.9793	0.4546	-	-
	CITE	**0.7392**	**0.9998**	**0.7550**	**0.9049**	**0.9474**
HGT	ACM	0.8967	0.9821	0.6112	-	-
	DBLP	-	0.9515	0.3003	-	-
	CITE	**0.7288**	**0.9995**	**0.5937**	**0.8635**	**0.9097**

Table 5: Link prediction results (AUC-ROC) across datasets.

CITE remains a major challenge for current methods. This limitation is critical in realworld applications like new material discovery and classification, where minority classes are often crucial. To verify that this gap is partially mitigable without changing model architectures, we evaluate two standard long-tail interventions (class-balanced sampling and balanced softmax) and a class-balanced subset; both consistently improve Macro-F1 while preserving the relative model ranking (see Appendix D.3 for details). Moreover, existing work has leveraged active learning to prioritize such high-value yet underrepresented samples, enabling more targeted exploration of promising material candidates for accelerated discovery (Ma et al., 2025).

(2) Joint modeling of heterogeneity and text remains difficult: Current models struggle to integrate multi-relational structures with textual semantics. Homogeneous models fail to capture cross-type node semantics, while LLM-based models, even though pretraining, still have significant potential for learning from graph topology. Although TAPE achieved high Micro-F1 on CITE through

LLM-enhanced GNN, the model rely on homogenization of heterogeneous data, which distorts relational semantics. While some heterogeneous graph models achieve good results, they cannot fully leverage unstructured text as textual complexity increases. These observations highlight the need for architectures that natively unify textual and structural signals without sacrificing relational granularity.

(3) CITE exposes both challenging and informative relations: CITE's complex, heterogeneous structure poses a challenge for most models, particularly in paper-paper link prediction. However, some relation types, like paper-journal and paper-keywords, consistently yield higher AUC scores compared to ACM or DBLP across all tested models. This is due to clear one-to-one mappings, dense high-degree nodes, and tight semantic coupling. CITE's structure not only presents challenges but also provides high-quality relational signals that help reveal model strengths in leveraging both graph structure and textual semantics.

2nd Finding

CITE's structural heterogeneity, rich textual attributes, and long-tailed label distribution jointly expose limitations of existing representation learning methods.

3.3 Ablation Study on Heterogeneous Properties *(Q3)*

3.3.1 Experiment Setup

To address Q3, we design three ablation dimensions: (1) removal of structural heterogeneity, (2) reduction of textual features, (3) elimination of both. For heterogeneity removal, we remove specific node types (author, keywords, journal) and their relationships with the paper node. For text attribute ablation, we remove abstract content from paper nodes, retaining only title. In terms of full homogenization, we combine structural and textual ablations, resulting in a homogeneous graph with little text. For model selection, in node classification task, we selected GCN, GAT, GraphSAGE as homogeneous graph models, HGT, SimpleHGN, RGCN, MAGNN as heterogeneous graph models. For the LLM-centric model, we selected Llama, and for the LLM+Graph model, we selected TAPE. In the link prediction task, we selected all models supporting the task. Hyperparameters, text encoders, and data split ratios are consistent with Q1. Experiment results are shown in Table 6 - 7, with the best results in **bold** and the worst in underlined.

3.3.2 Discussion

(1) Structural information plays a dominant role in GNN-based models: Results indicate that structural information plays a dominant role in GNN-based representation learning on CITE. In homogeneous graph models, the performance of models upon removing heterogeneous information increased slightly and it increased more upon removing textual attributes reveals their inability of handling textual information. However the models achieve the optimal value when removing the text features and retaining the heterogeneous features, this suggests homogeneous graph models can benefit from controlled heterogeneous semantic fusion, but fail to disentangle structural and textual signals effectively. In terms of the performance of the heterogeneous graph models, the modeling performance degrade regardless of which heterogeneous node is eliminated, and this is especially noticeable on the journal node. This suggests that low-frequency (0.58%) but high-connectivity (average degree=50) nodes can form latent topic clusters.

(2) Textual semantics substantially influence LLM-based graph representations: Llama's performance increase slightly with text attributes and increase more under structural ablation. When making the graph full Homogenization, the results lie between both of them, indicating model performance indifference to graph topology. Also we can see the more heterogeneous nodes are integrated, the more performance degrades, likely because discontinuous entity mentions disrupt coherent context. Unlike GNNs, LLMs may treat each node as an isolated text snippet under prompt-based structural injection, thus failing to reliably leverage relational dependencies. Notably, the LLM-centric model achieves its best performance when structural/heterogeneous context is truncated, suggesting that textual semantics dominate its predictions and that explicit topology provides limited benefit. This observation is consistent with recent analyses on LLMs over text-attributed graphs (Guan et al., 2025; Wu et al., 2025; Xu et al., 2025), which show that transformer attention often struggles to capture inter-node relations from graph-structured inputs and that structural encodings can yield marginal or negative gains in node classification when labels are largely determined by semantics.

(3) LLM+Graph architecture is capable of

Model Type	Model Name	Original Graph	Remove Heterogeneous Nodes				Remove Text Attributes	Full Homogenization
			Keywords	Journal	Author	All		
Homogeneous Graph Models	**GCN**	0.5057/0.0239	0.5164/0.0294	0.5074/0.0243	0.5098/0.0261	0.5088/0.0186	**0.5295/0.0407**	0.5084/0.0186
	GAT	0.5115/0.0457	0.5262/0.0578	0.5139/0.0524	0.5237/0.0564	0.5121/0.0389	**0.5879/0.1146**	0.5123/0.0370
	GraphSAGE	0.5117/0.0474	0.5254/0.0572	0.5157/0.0499	0.5099/0.0261	0.5113/0.0363	**0.5873/0.1036**	0.5123/0.0378
Heterogeneous Graph Models	**SimpleHGN**	0.9907/0.5052	0.9904/0.4913	0.5471/0.0810	0.9881/0.4707	0.5027/0.0399	**0.9913/0.5224**	0.4995/0.0342
	HGT	**0.9791**/0.4183	0.9782/**0.4253**	0.5389/0.0613	0.9743/0.3915	0.5260/0.0646	0.9770/0.4094	0.5154/0.0554
	RGCN	0.8626/0.1442	0.8398/0.0936	0.5128/**0.1756**	0.8317/0.0940	0.4978/0.0127	**0.8637**/0.1231	0.4971/0.0127
	MAGNN	0.5542/0.0268	**0.5858/0.0439**	0.4713/0.0197	**0.5858**/0.0336	0.4166/0.0297	0.5759/0.0372	0.4265/0.0247
LLM-centric Model	**Llama**	0.0623/0.0037	0.0634/0.0038	0.0696/0.0041	0.0627/0.0038	**0.1316/0.0110**	0.0774/0.0046	0.0712/0.0041
LLM+Graph Model	**TAPE**	0.9936/0.5655	0.9937/0.5880	0.9912/0.5433	**0.9940/0.5904**	0.9926/0.6394	0.9907/0.5498	0.9908/0.5606

Table 6: Node classification results of ablation study (Micro-F1/Macro-F1).

Model Type	Model Name	Original Graph	Remove Heterogeneous Nodes				Remove Text Attributes	Full Homogenization
			Keywords	Journal	Author	All		
Heterogeneous Graph Models	**RGCN**	0.6990	0.6838	0.6978	0.6803	0.6988	0.6847	**0.6999**
	CompGCN	0.7153	**0.7517**	0.7158	0.7203	0.7414	0.7210	0.7415
	SimpleHGN	0.7392	**0.7422**	0.7383	0.7401	0.7159	0.7403	0.7084
	HPN	0.4978	0.5007	0.5007	0.5007	0.5005	0.4980	**0.5042**
	HGT	**0.7288**	0.7286	0.7230	0.7233	0.7199	0.7279	0.7281
	HAN	0.6826	0.6843	0.6902	0.6903	0.6895	**0.6921**	0.6901
	MAGNN	0.5542	0.4970	0.5038	0.5044	0.4962	**0.5066**	0.4976

Table 7: Link prediction results of ablation study (AUC-ROC).

comprehending both crucial graph structures and textual semantics: TAPE demonstrate robustness but reveal dual sensitivity patterns. When abstracts were removed, performance dropped, indicating that LLMs rely heavily on semantic richness. Removing author nodes slightly improved performance by eliminating low-information entities, thus increasing the textual signal-to-noise ratio. Structural sensitivity was shown when removing journal nodes caused a decline in Macro-F1, while removing keywords or authors improved performance. These results highlight the complementary roles of structural cues and textual semantics, underscoring the need for future work to fully exploit their synergistic potential.

(4) Link prediction is primarily driven by structural heterogeneity: Metapath methods (MAGNN, HAN) occasionally lead after aggressive ablations, suggesting that noise removal can outweigh expressivity losses for them. Attention models (SimpleHGN, HGT) dominate under most settings, they show minimal drops when ablating individual node types, highlighting their robust relation-specific attention. MAGNN is most sensitive, suffering consistent AUC drops, reflecting brittleness of fixed meta-path schemes when heterogeneity or text is reduced. The removal of abstracts had only a marginal impact, confirming that link prediction performance is primarily driven by structure rather than textual features. The largest performance drops come from removing all heterogeneous nodes, underscoring that some heterorelations are indispensable. Yet, selective removal can actually denoise and slightly boost certain models, revealing that distribution-sensitive sampling could further improve learning.

3rd Finding

High-degree heterogeneous nodes, rich textual semantics, and selective integration of heterogeneous components play a key role in graph representation learning on TAGs.

4 Conclusion

In this paper, we introduce CITE, a citation graph benchmark for catalytic materials, designed to address limitations of existing datasets, such as small scale, limited node and relationship types, and lack of textual information. We have curated over a century of research on photocatalysis and electrocatalysis, organizing it into citation graphs with four node types and four edge types. We then evaluate CITE across various learning paradigms. Our goal is to explore the challenges posed by heterogeneous and textual attributes to existing models and identify future research directions. Finally, we organize our code to allow researchers to easily replicate our experiments and evaluate new models using CITE.

Limitations

First, this work focuses on heterogeneous text-attributed citation graphs in the domain of catalytic materials, and the findings may not directly generalize to graphs from other domains or with substantially different relational structures. Second, our evaluation considers node classification and link prediction tasks, and does not cover other settings such as graph-level reasoning, or interactive retrieval scenarios. Third, the evaluation of LLM-based and LLM+Graph models depends on specific prompting strategies and integration designs, and alternative formulations may lead to different performance characteristics. Finally, from a broader impact perspective, this work contributes a unified knowledge base for catalytic materials that may support material discovery and systematic knowledge organization; however, heavy reliance on data-driven representations without sufficient domain context may risk oversimplifying complex scientific reasoning processes.

Ethical Considerations

This work introduces a benchmark dataset constructed from publicly available scientific publications and citation metadata, and is intended solely for research and academic purposes. The dataset is derived from scholarly publications and citation relations rather than user-generated personal data. The released benchmark may include author names as part of public bibliographic metadata, but does not use private contact information such as email addresses, and affiliations are not used in the benchmark construction or experiments. We retain only bibliographic and relational information necessary for benchmark construction, and offensive content is not a primary risk scenario in this domain-specific scientific corpus. While CITE enables systematic evaluation of graph and LLM-based models, the benchmark should not be used as the sole basis for scientific decision-making or material discovery without expert validation.

The long-tailed label distribution present in CITE may amplify performance disparities across categories, which should be carefully considered when interpreting model results. Additionally, over-reliance on data-driven representations and automated evaluation may oversimplify complex scientific reasoning processes if applied without sufficient domain context. We encourage responsible use of the dataset and transparent reporting of model limitations when applying CITE to tasks such as automated literature analysis, citation-based inference, or material discovery.

Acknowledgements

This work was supported in part by the Natural Science Foundation of China under Grant No. T2322027, 62442204 and 92470204, in part by the Key Research Program of the Chinese Academy of Sciences under Grant No. RCJJ-145-24-20. Generative AI assistants were used for language polishing and code debugging. All technical decisions, experimental results, and final manuscript content were reviewed and verified by the authors.

References

Marvin Alberts, Oliver Schilter, Federico Zipoli, Nina Hartrampf, and Teodoro Laino. 2024. Unraveling molecular structure: A multimodal spectroscopic dataset for chemistry. *Advances in Neural Information Processing Systems*, 37:125780–125808.

John Arevalo, Ellen Su, Anne Carpenter, and Shantanu Singh. 2024. Motive: A drug-target interaction graph for inductive link prediction. *Advances in Neural Information Processing Systems*, 37:140320–140333.

Yuanchen Bei, Weizhi Chen, Hao Chen, Sheng Zhou, Carl Yang, Jiapei Fan, Longtao Huang, and Jiajun Bu. 2025. Correlation-aware graph convolutional networks for multi-label node classification. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 37–48.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Runjin Chen, Tong Zhao, Ajay Kumar Jaiswal, Neil Shah, and Zhangyang Wang. 2024a. Llaga: Large language and graph assistant. In *International Conference on Machine Learning*, pages 7809–7823. PMLR.

Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, Hongzhi Wen, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, and 1 others. 2024b. Exploring the potential of large language models (llms) in learning on graphs. *ACM SIGKDD Explorations Newsletter*, 25(2):42–61.

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, and 1 others. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282.

Zheng Fang, Qingqing Long, Guojie Song, and Kunqing Xie. 2021. Spatial-temporal graph ode networks for traffic flow forecasting. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 364–373.

Zheng Fang, Lingjun Xu, Guojie Song, Qingqing Long, and Yingxue Zhang. 2022. Polarized graph neural networks. In *Proceedings of the ACM web conference 2022*, pages 1404–1413.

Xinyu Fu, Jiani Zhang, Ziqiao Meng, and Irwin King. 2020. Magnn: Metapath aggregated graph neural network for heterogeneous graph embedding. In *Proceedings of the web conference 2020*, pages 2331–2341.

C Lee Giles, Kurt D Bollacker, and Steve Lawrence. 1998. Citeseer: An automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, pages 89–98.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. In *Neural Information Processing Systems*. Curran Associates.

Zhong Guan, Likang Wu, Hongke Zhao, Ming He, and Jianping Fan. 2025. Attention mechanisms perspective: Exploring llm processing of graph-structured data. In *International Conference on Machine Learning*, pages 20612–20639. PMLR.

Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, and 1 others. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688.

Anjali Gupta, Prashant Kumar, Aniket Mishra, Abhishek Singh, Surender Kumar, Muthusamy Chelliah, Abhijnan Chakraborty, and Sayan Ranu. 2025. Persona identification in e-commerce with scarce labels and in-context graph learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 778–789.

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Hui Han, Tianyu Zhao, Cheng Yang, Hongyi Zhang, Yaoqi Liu, Xiao Wang, and Chuan Shi. 2022. Openhgnn: An open source toolkit for heterogeneous graph neural network. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 3993–3997.

Xiaoxin He, Xavier Bresson, Thomas Laurent, Adam Perold, Yann LeCun, and Bryan Hooi. 2023. Harnessing explanations: Llm-to-lm interpreter for enhanced text-attributed graph representation learning. *arXiv preprint arXiv:2305.19523*.

Rodrigo Hormazabal, Changyoung Park, Soonyoung Lee, Sehui Han, Yeonsik Jo, Jaewan Lee, Ahra Jo, Seung Hwan Kim, Jaegul Choo, Moontae Lee, and 1 others. 2022. Cede: A collection of expert-curated datasets with atom-level entity annotations for optical chemical structure recognition. *Advances in Neural Information Processing Systems*, 35:27114–27126.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020a. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020b. Heterogeneous graph transformer. In *Proceedings of the web conference 2020*, pages 2704–2710.

Xuanwen Huang, Kaiqiao Han, Yang Yang, Dezheng Bao, Quanjin Tao, Ziwei Chai, and Qi Zhu. 2024. Can gnn be good adapter for llms? In *Proceedings of the ACM Web Conference 2024*, pages 893–904.

Houye Ji, Xiao Wang, Chuan Shi, Bai Wang, and Philip S Yu. 2021. Heterogeneous graph propagation network. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):521–532.

Bowen Jin, Wentao Zhang, Yu Zhang, Yu Meng, Xinyang Zhang, Qi Zhu, and Jiawei Han. 2023. Patton: Language model pretraining on text-rich networks. *arXiv preprint arXiv:2305.12268*.

Wei Ju, Zheng Fang, Yiyang Gu, Zequn Liu, Qingqing Long, Ziyue Qiao, Yifang Qin, Jianhao Shen, Fang Sun, Zhiping Xiao, and 1 others. 2024a. A comprehensive survey on deep graph representation learning. *Neural Networks*, 173:106207.

Wei Ju, Yifan Wang, Yifang Qin, Zhengyang Mao, Zhiping Xiao, Junyu Luo, Junwei Yang, Yiyang Gu, Dongjie Wang, Qingqing Long, and 1 others. 2024b. Towards graph contrastive learning: A survey and beyond. *arXiv preprint arXiv:2405.11868*.

Kuzma Khrabrov, Anton Ber, Artem Tsypin, Konstantin Ushenin, Egor Rumiantsev, Alexander Telepov, Dmitry Protasov, Ilya Shenbin, Anton Alekseev, Mikhail Shirokikh, and 1 others. 2024. ∇ 2 dft: A universal quantum chemistry dataset of drug-like molecules and a benchmark for neural network potentials. *SESAR Innovation Days*, 2023:82–83.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Yuhan Li, Peisong Wang, Xiao Zhu, Aochuan Chen, Haiyun Jiang, Deng Cai, Victor W Chan, and Jia Li. 2024. Glbench: A comprehensive benchmark for graph with large language models. *Advances in Neural Information Processing Systems*, 37:42349–42368.

Guojiao Lin, Meng Zhen, Dongjie Wang, Qingqing Long, Yuanchun Zhou, and Meng Xiao. 2024. Gume: Graphs and user modalities enhancement for long-tail multimodal recommendation. In *Proceedings of the 33rd ACM international conference on information and knowledge management*, pages 1400–1409.

Hu Linmei, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. 2019. Heterogeneous graph attention networks for semi-supervised short text classification. 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)*, pages 4821–4830.

Yunhui Liu, Qizhuo Xie, Jinwei Shi, Jiaxu Shen, and Tieke He. 2025. Multi-scale heterogeneous text-attributed graph datasets from diverse domains. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 757–760.

Qingqing Long, Haotian Chen, Chenyang Zhao, Xiaolei Du, Xuezhi Wang, Pengyao Wang, Chengzan Li, Yuanchun Zhou, and Hengshu Zhu. 2026. Sciencedb ai: An llm-driven agentic recommender system for large-scale scientific data sharing services. *arXiv preprint arXiv:2601.01118*.

Qingqing Long, Yilun Jin, Guojie Song, Yi Li, and Wei Lin. 2020. Graph structural-topic neural network. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1065–1073.

Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, and 1 others. 2025. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*.

Qingsong Lv, Ming Ding, Qiang Liu, Yuxiang Chen, Wenzheng Feng, Siming He, Chang Zhou, Jianguo Jiang, Yuxiao Dong, and Jie Tang. 2021. Are we really making much progress? revisiting, benchmarking and refining heterogeneous graph neural networks. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 1150–1160.

Yujie Ma, Yang Gao, Ludi Wang, Ming Chen, Wenjuan Cui, Bin Wang, and Yi Du. 2025. Accelerating materials discovery through active learning: Methods, challenges and opportunities. *The Innovation Informatics*, 1(1):100013–1.

Zhengyang Mao, Wei Ju, Siyu Yi, Yifan Wang, Zhiping Xiao, Qingqing Long, Nan Yin, Xinwang Liu, and Ming Zhang. 2025. Learning knowledge-diverse experts for long-tailed graph classification. *ACM Transactions on Knowledge Discovery from Data*, 19(2):1–24.

Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. 2000. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3:127–163.

Seth A Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. 2014. Information network or social network? the structure of the twitter follow graph. In *Proceedings of the 23rd international conference on world wide web*, pages 493–498.

Kaichen Ouyang, Shiyun Zhang, Song-Ling Liu, Jiachuan Tian, Yuanhao Li, Hua Tong, Hai-Yang Bai, Wei-Hua Wang, and Yuan-Chao Hu. 2025. Graph learning metallic glass discovery from wikipedia. *AI for Science*, 1(2):025004.

Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, and 1 others. 2020. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The semantic web: 15th international conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, proceedings 15*, pages 593–607. Springer.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. 2008. Collective classification in network data. *AI magazine*, 29(3):93–93.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246.

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*, pages 491–500.

Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 990–998.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. 2019. Composition-based multi-relational graph convolutional networks. *arXiv preprint arXiv:1911.03082*.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, and 1 others. 2017. Graph attention networks. *stat*, 1050(20):10–48550.

Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuan-Jing Huang. 2020a. Heterogeneous graph neural networks for extractive document summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6209–6219.

Duo Wang, Ao Xie, Shengcun Ma, Wei Tong, Shiqiang Zou, and Anubhav Jain. 2025. Computational catalysis and machine learning applications to water treatment technologies. *AI for Science*, 1(1):013001.

Junshan Wang, Ziyao Li, Qingqing Long, Weiyu Zhang, Guojie Song, and Chuan Shi. 2020b. Learning node representations from noisy graph structures. In *2020 IEEE international conference on data mining (ICDM)*, pages 1310–1315. IEEE.

Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. 2019. Heterogeneous graph attention network. In *The world wide web conference*, pages 2022–2032.

Jacob White. 2020. Pubmed 2.0. *Medical reference services quarterly*, 39(4):382–387.

Xixi Wu, Yifei Shen, Fangzhou Ge, Caihua Shan, Yizhu Jiao, Xiangguo Sun, and Hong Cheng. 2025. When do llms help with node classification? a comprehensive analysis. In *International Conference on Machine Learning*, pages 67623–67649. PMLR.

Haotian Xu, Yuning You, and Tengfei Ma. 2025. When structure doesn't help: Llms do not read text-attributed graphs as effectively as we expected. In *The Fourth Learning on Graphs Conference*.

Hao Yan, Chaozhuo Li, Ruosong Long, Chao Yan, Jianan Zhao, Wenwen Zhuang, Jun Yin, Peiyan Zhang, Weihao Han, Hao Sun, and 1 others. 2023. A comprehensive study on text-attributed graphs: Benchmarking and rethinking. *Advances in Neural Information Processing Systems*, 36:17238–17264.

Yuchen Yan, Peiyan Zhang, Zheng Fang, and Qingqing Long. 2024. Inductive graph alignment prompt: bridging the gap between graph pre-training and inductive fine-tuning from spectral perspective. In *Proceedings of the ACM Web Conference 2024*, pages 4328–4339.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Junhan Yang, Zheng Liu, Shitao Xiao, Chaozhuo Li, Defu Lian, Sanjay Agrawal, Amit Singh, Guangzhong Sun, and Xing Xie. 2021. Graphformers: Gnn-nested transformers for representation learning on textual graph. *Advances in Neural Information Processing Systems*, 34:28798–28810.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016.

Yanpeng Ye, Jie Ren, Shaozhou Wang, Yuwei Wan, Imran Razzak, Bram Hoex, Haofen Wang, Tong Xie, and Wenjie Zhang. 2024. Construction and application of materials knowledge graph in multidisciplinary materials science via large language model. *Advances in Neural Information Processing Systems*, 37:56878–56897.

Lingfan Yu, Jiajun Shen, Jinyang Li, and Adam Lerer. 2020. Scalable graph neural networks for heterogeneous graphs. *arXiv preprint arXiv:2011.09679*.

Haodong Zhang, Xinyue Wang, Tao Ren, Yifan Wang, Siyu Yi, Fanchun Meng, Zeyu Ma, Qingqing Long, and Wei Ju. 2026. Fairgc: Fostering individual and group fairness for deep graph clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 28194–28202.

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 Web Conference 2024*, pages 1003–1014.

Jianan Zhao, Meng Qu, Chaozhuo Li, Hao Yan, Qian Liu, Rui Li, Xing Xie, and Jian Tang. 2022. Learning on large-scale text-attributed graphs via variational inference. *arXiv preprint arXiv:2210.14709*.

Peng-Cheng Zhao, Hong-Ze Zhou, Qiong Wang, Zhen-Yi Wu, Hui Yu, and Jian-Yu Shi. 2025. A comprehensive survey of ai-based retrosynthesis planning: Datasets, models, and tools. *The Innovation Informatics*, pages 100026–1.

A Related Work

Studies on Citation Graphs. Several well-known citation graph datasets have been widely used in graph learning tasks. Mccallum et al. established Cora (McCallum et al., 2000), which is a scientific publication dataset of computer science. Cora contains over 50,000 papers which are categorized into seven categories. The most commonly used version is designed for node classification, which contains 2,708 scientific publications and 5,429 citation links. The PubMed dataset (White, 2020) is based on the PubMed database, designed for node classification. It contains 19,717 scientific publications related to diabetes, which are grouped into three categories. The cross-reference network of these publications contained 44,338 edges. The CiteSeer dataset (Giles et al., 1998) is designed for node classification, which contains 3,312 scientific publications, six categories, and 4,732 citation links.

In addition, other more complex or heterogeneous datasets have been introduced. The OGB dataset (Hu et al., 2020a) is designed for node classification, link prediction and graph property prediction. The dataset assembles graphs of various sizes and topics, e.g., proteins, arXiv papers, etc., ranging in size from 100k to 100M. Recently, Liu et al. reorganized existing citation benchmarks such as arXiv and DBLP into heterogeneous text-attributed graph formats. However, their work consolidates previously available data rather than introducing new domain-specific datasets with independently curated entity types and relation structures (Liu et al., 2025). The Aminer dataset (Tang et al., 2008) is extracted from DBLP, ACM, MAG, etc. The first version contains 629,814 papers and 632,752 citations. Each paper is stored as a JSON file and is associated with an abstract, author, year, etc. All of the above information is represented as strings and numbers, and there is no information about the paper's citation graphs. ACM dataset (Wang et al., 2019) is a heterogeneous graph that comprises 3,025 papers, 5,835 authors and 56 subjects. The Microsoft Academic Graph (MAG) (Sinha et al., 2015) is a heterogeneous graph containing scientific publication records, citation relationships, and information about authors, institutions, journals, conferences, and fields of study. The dataset contains over 150 million nodes and more than 2 billion edges.

Studies on Graphs in Chemistry. In the chemistry domain, several datasets provide valuable resources for tasks related to molecular structure and properties. With the aim of improving the performance of neural architectures for chemical image interpretation, Hormazabal et al. introduced a dataset that contains one million molecular images and over 700,000 curated bounding boxes for chemical entities (Hormazabal et al., 2022). To improve molecular property prediction and conformational optimization, the ∇^2DFT dataset is introduced, which contains 15,716,667 conformations, including energies, forces, 17 molecular properties, Hamiltonian and overlap matrices, and a wavefunction object (Khrabrov et al., 2024). In molecular structure prediction, Alberts et al. proposed a multimodal spectroscopic dataset to support foundation models for structure elucidation and spectra prediction (Alberts et al., 2024).

Recent advancements have also highlighted the role of LLMs in chemistry tasks. Guo et al. identify three key chemistry-related capabilities of LLM, including understanding, reasoning and explaining. They also established a benchmark containing eight chemistry tasks (Guo et al., 2023). Beyond molecular-level tasks, Zhao et al. provide a comprehensive survey of AI-based retrosynthesis planning, covering GNN and Transformer methods, large-scale reaction datasets, and the emerging role of LLMs in synthesis route prediction, while highlighting persistent data quality challenges such as long-tail bias and inconsistent annotations in chemical reaction databases (Zhao et al., 2025). Ma et al. has integrated graph neural networks with large language models to efficiently navigate vast chemical spaces for catalytic materials discovery under limited experimental budgets (Ma et al., 2025). Ouyang et al. proposed a graph-based recommendation framework for metallic glass discovery, where node features are encoded from Wikipedia using language models, demonstrating the potential of combining graph neural networks with language-derived representations for data-driven materials

Table 8: Statistics of nodes and edges in CITE.

Node/Edge Type	Numbers	Percentage	Source Coverage
Paper	127,690	29.13%	-
Paper-Paper	335,118	27.46%	96.5%
Author	221,097	50.44%	-
Paper-Author	506,328	41.49%	94.7%
Keywords	86,964	19.84%	-
Paper-Keywords	253,142	20.74%	39.7%
Journal	2,553	0.58%	-
Paper-Journal	125,786	10.31 %	98.5%

design (Ouyang et al., 2025). In the electrocatalysis domain, Wang et al. discussed how computational modeling and machine learning can accelerate the search for new catalyst materials, highlighting the integration of machine-learned force fields with experimental platforms for rapid catalyst identification (Wang et al., 2025). In the domain of chemical materials, the MKG dataset has been developed, that organizes chemical entities such as "formulas", "names", and "acronyms" into a structured triad containing 162,605 nodes and 731,772 edges (Ye et al., 2024). MOTIVε integrates morphological data and graph-based modeling by compiling Cell Painting features for 11,000 genes and 3,600 compounds, enhancing the drug-target interaction performance (Arevalo et al., 2024).

B CITE Construction and Statistics

B.1 Overview

As shown in Table 8, CITE demonstrates balanced distributions across nodes, edges, and semantic labels, addressing common weaknesses of domain-specific graphs by preventing the dominance of a single entity type and ensuring robust representation of heterogeneous interactions. For example, despite the relatively small proportion of journal nodes, paper-journal edges cover 98% of paper nodes, while most edge types exhibit similarly high coverage, indicating that CITE forms a densely interconnected and well-structured network. Author nodes account for approximately half of all nodes but only 41% of edges, reflecting domain-specific authorship dynamics without disproportionate amplification.

Notably, journals and keywords together represent just 20.42% of nodes but contribute to 31.5% of all edges. This structural characteristic suggests two critical insights: (1) Research topics exhibit natural concentration around photocatalysis and electrocatalysis domains as manifested through keyword co-occurrence patterns, and (2) Journal entities, despite constituting merely 0.58% of nodes, account for 10.5% of edges (paper-journal affiliations), indicating concentrated paper dissemination where a limited subset of journals structurally anchor the majority of publications. Moreover, CITE exhibits strong advantages in data integrity, as summarized in Table 9, which reports the null-value rates of key fields, underscoring the overall quality and consistency of the dataset.

Table 9: Missing rates for key metadata.

Metadata	Missing Rate(%)	Metadata	Missing Rate(%)
id	0	doi	0
title	0	year	0.001
authors	0	publish_date	0.005
volume	2.077	page_start	1.458
page_end	3.773	print_issn	2.488
abstract	3.463	journal	1.594
cited_times	0	reference_counts	0
citations	0	references	0
organization	0	authors_split	0.001

B.2 Construction of CITE

Data Sources and Domain Scope: CITE is derived from peer-reviewed publications in the interdisciplinary fields of photocatalysis and electrocatalysis, which include leading journals and conferences in catalysis, such as *Applied Catalysis B: Environmental*, *ACS Catalysis*, and *Electrochimica Acta*. The temporal scope covers the period from 1922 to 2022, capturing the evolution of the field from foundational studies on TiO2-based photocatalysts to modern advances in atomic-scale electrocatalyst design. All papers in CITE are sourced from DOIs indexed in Web of Science, ensuring that each entry corresponds to peer-reviewed SCI journals. This ensures the reliability and credibility of the data, making it a robust foundation for downstream tasks such as citation analysis, trend discovery, and interdisciplinary research.

Data Extraction and Integration: After consulting with domain experts, we first identified scientific publications using the keywords “photocataly*” and “electrocataly*” in Web of Science, retrieving over 150,000 DOIs. Metadata for each paper was then collected from the CrossRef database, including bibliographic attributes such as titles, authors, journal and conference names, and publication years. The resulting metadata was organized into structured JSON files, each corresponding to a unique DOI and containing fields such as title, publication year, first author, and complete author list. This process yielded complementary components necessary for heterogeneous graph construction. Multiple rounds of validation and consistency checks were conducted throughout the process.

Data Cleaning and Entity Normalization: We performed a comprehensive quality control to address missing or duplicate entries. Records lacking all associated metadata were removed, and deduplication was performed using DOI identifiers. For journal nodes, an ISSN-based normalization was applied, reducing journal entries from

Table 10: Metadata and detailed information.

Field	Description	Example
id	Unique paper ID	j_000033981...
doi	Paper DOI number	10.1016/j.jhazmat...
title	Paper title	Fabrication of...
year	Year of publication	2011
authors	Ordered author list	Shouqiang Wei...
publish_date	Publication date	2011/10/1
volume	Volume set	194
page_start	Start page of paper	243
page_end	End page of paper	249
print_issn	ISSN number	0920-5861
abstract	Abstract text	Spherical core...
journal	Official journal name	Catalysis Today
cited_times	Citation count	18
reference_times	Reference count	50
organization	Authors' affiliation	{'org_name':'Seri...'
citations	List of citing papers	{'id': '23...','title':...
references	List of cited papers	{'id': '79...','title':...
authors_split	Author list in order	{'person_id': '872..'
fields	Journal research areas	CHEMISTRY;

3,590 to 2,553 unique titles. For author nodes, we adopted ORCID as the primary disambiguation key. When unavailable, author identity was inferred through email and institutional affiliation, resulting in 217,977 validated author records covering 98.6% of all data. For keyword nodes, only apparent noise like stopwords, articles, or numeric tokens was removed to preserve technical terminology, resulting in 86,964 records. After cleaning, the dataset contained 127,690 valid paper records.

Structural and Semantic Validation: We conducted multiple validation procedures to ensure structural and semantic reliability. For node validation, we computed the entropy of paper-journal and paper-author edge distributions, comparing them to the theoretical uniform entropy. The observed entropies (author-16.90 bits, journal-8.16 bits) correspond to 95.2% and 72.1% of their respective maxima, indicating well-balanced node distributions. For edge validation, OpenAlex provides officially curated citation data, ensuring high reliability of citation edges. Regarding graph sparsity, node degrees were calculated as follows: paper = 9.5, author = 2.29, keyword = 2.91, and journal = 49. The resulting average node degree of 5.57 demonstrates that the graph remains dense and well-connected after cleaning. Finally, for semantic consistency, we removed paper nodes lacking journal or abstract information, ensuring alignment between structural and textual dimensions and preserving the dataset's integrity for downstream analysis.

CITE consists of four core entity types: Paper, Author, Journal, and Keywords, which collectively correspond to four types of relationships: Paper-Paper, Paper-Author, Paper-Journal, and Paper-Keywords.

Paper nodes represent article content, with title and abstract attributes extracted from metadata. Labels are assigned based on the SCI classification of the journal, or 'uncategorized' if not indexed. Because certain journals are indexed under multiple SCI subject categories, CITE supports a multi-label classification setting, where each paper may be associated with more than one label. Author nodes represent article authors, identified by ORCID, or by institutional affiliation if ORCID is unavailable and the names match. Journal nodes represent publishing journals or conferences, with their ISSNs extracted from metadata for official title conversion. Keyword nodes reflect the core research topics and are extracted from paper metadata. Table 10 provides a detailed description of the parsed metadata fields for each entity type, including identifiers, textual attributes, and cross-entity associations used in graph construction.

Edges capture both explicit relationships (citations, authorship) and implicit semantic connections (topical similarity). Citation edges (Paper-Paper) are based on metadata references from OpenAlex. Authorship edges (Paper-Author) are derived from author lists, with each author uniquely identified. Publication venue edges (Paper-Journal) connect papers to journals via ISSN. Topical relevance edges (Paper-Keywords) are created from paper keywords, with data cleaning applied. The overall pipeline for constructing the heterogeneous graph from raw metadata is illustrated below.

B.3 Preprocess Workflow

This section details the preprocessing procedure for constructing CITE from structured bibliographic metadata. Algorithm 1 outlines the sequential steps of data parsing, entity normalization, relation extraction, and graph assembly.

C Prompt Design and Analysis of LLM Responses

C.1 Input Prompts

GraphGPT prompts.
Stage1:
Given a sequence of graph tokens <graph> that constitute a subgraph...
1.Title: {title}, Abstract: {abstract},

Algorithm 1 Preprocess Workflow

Input: Json file F of metadata, including P_F, A_F, J_F, K_F, I_F, which denotes set of Paper (Title & Abstract), Author (Name, Orcid), Journal (Journal Name), Keywords, ISSN
Output: CITE-$G(V, E, T)$, where V denotes node set, E denotes edge set, T denotes text set
Assign Unique Identifier p_i, a_i for $p_F \in P_F$, $a_F \in A_F$
Construct Paper set $V_P \in V$, Author set $V_A \in V$
for item i in F **do**
    Construct CSV column C_i
end for
for item in Journal column $j_C \in J_C$ **do**
    Standardize j_C with i_C, construct journal set $V_J \in V$
    Assign Unique Identifier j_i
end for
for item in Keywords column $k_C \in K_C$ **do**
    Construct keywords set $V_K \in V$
    Assign Unique Identifier k_i
end for
for row i in CSV **do**
    if j_i **then**
        Construct paper-journal edge set $E_J \in E$, text set $T_J \in T$
        Assign label L_i
    else if k_i **then**
        Construct paper-keywords edge set $E_K \in E$, text set $T_K \in T$
    else if a_i **then**
        Construct paper-author edge set $E_A \in E$, text set $T_A \in T$
    else if p_i(citation/reference) **then**
        Construct paper-paper edge set $E_P \in E$, text set $T_P \in T$
    end if
end for

Journal: `{journal}`, Authors: `{authors}`, Keywords: `{key_words}`. 2.Title:`{title}`...
Question: Please reorder the list of papers according to the order of graph tokens (i.e., complete the matching of graph tokens and papers).
Stage2:
Given a citation graph: `<graph>` where the 0th node is the target paper... Title: `{title}`, Abstract: `{abstract}`, Journal: `{journal}`, Authors: `{authors}`, Keywords: `{key_words}`.
Question: Which SCI category does this paper belong to? Here is a list of SCI categories: 1.ENGINEERING...

LLMs&TAPE prompts.
Question:
Title: `{title}`, Abstract: `{abstract}`, Journal: `{journal}`, Authors: `{authors}`, Keywords: `{key_words}`.
Which of the following categories does this paper belong to: 1.ENGINEERING...

C.2 Examples of LLM Responses

Examples of GraphGPT responses.
Correct Response: Based on the given citation, the paper belongs to the category "CHEMISTRY"
Incorrect Response: Based on the given citation, the paper belongs to the category "MATERIAL SCIENCE"
Error Type I-Out of list: Based on the given citation, the paper is likely to belong to the category "Nanoparticles of CdS and Indole"(outside the label list)
Error Type II-Question repetition: Based on the given citation, we obtain the following results: 1.ENGINEERING 2.MATERIALS SCIENCE 3.PHYSICS..(repeat the question without answering)

Examples of LLMs responses.
Correct Response: This paper belongs to the category of "CHEMISTRY"
Incorrect Response: This paper belongs to the category of "MATERIALS SCIENCE".
Error Type I-Out of list: This paper belongs to the category of SCIENCERESEARCH (outside the label list)

Error Type II-Question repetition: Title:`{title}`, Abstract: ...(repeat the question without answering)

C.3 Prompt-Tuning Methods

Multi-turn Dialogue

Round 1
I would like to determine which SCI category a photocatalysis-related paper belongs to. Here is the basic information of the paper: Title: `{title}`,Abstract: `{abstract}`. Based on this information, could you summarize the key research themes and technical focus of this paper?
Round 2
Here is some additional information about the paper: Journal: `{journal}`, Authors: `{authors}`, Keywords: `{keywords}`. Given this, could you further refine your judgment on which general scientific disciplines this paper most likely falls under? Please suggest 3 to 5 possible fields.
Round 3
Now, please choose the most appropriate SCI categories from the following list of 84 options, and explain your reason: 1. Chemistry 2. Materials Science...

Forced-choice Prompting

You are a careful classifier. I would like you to determine which category a paper belongs to. You must follow the user's output format exactly. Do not add any explanation, punctuation, quotes, or extra words.
Task: Assign exactly ONE label to the paper based on the information below.
Paper: Title: {title}, Journal: {journal}, Authors: {authors}. Label set (choose exactly one; output must match one label verbatim): 1 = Engineering (for CITE) / 1 = cs.LG (for arXiv) / 1 = Database (for ACM)...
Output format rules (strict):
- Output exactly one label from the label set, verbatim. - Output nothing else (no reasoning, no extra text, no quotes, no ID numbers unless they are part of the label). - The entire response must be a single line containing only the label.
Answer:

For each input x, we build a shared prompt prefix $P(x)$ (paper metadata + label list + "Answer:"). For each candidate label y_i with tokenization $(t_1, \ldots, t_L)$, we compute a length-normalized teacher-forced log-likelihood:

$$\text{score}(y_i \mid x) = \frac{1}{L} \sum_{j=1}^{L} \log p(t_j \mid P(x), t_{<j}).$$

The prediction is then

$$\hat{y} = \arg\max_i \text{score}(y_i \mid x).$$

See prompts below:

Likelihood Ranking

You are a careful classifier. I would like you to determine which category a paper belongs to. You must follow the user's output format exactly.
Task: Assign exactly ONE label to the paper based on the information below.
Paper: Title: {title}, Journal: {journal}, Authors: {authors}. Label set (choose exactly one; output must match one label verbatim): 1 = Engineering (for CITE) / 1 = cs.LG (for arXiv) / 1 = Database (for ACM)...
Output format rules (strict):
- Output exactly one label from the label set, verbatim. - Output nothing else.
Answer:

D Supplementary Experiment

D.1 Sensitivity Analysis

We conducted five independent experiments on the baseline models to obtain the error bar results currently presented. In each round, we computed the mean and standard deviation (± 2σ). All experiments were performed on an NVIDIA A100 GPU, the top three scores for each metric are highlighted in red, yellow, and blue.

D.2 Temporal Split Experiment

The standard random 60/20/20 split used in the main experiments follows common practice in prior node-classification benchmarks on citation graphs (e.g., Cora/CiteSeer/PubMed-style protocols). However, citation networks exhibit temporality, as future studies are based on past studies, which may introduce temporal information leakage. Since CITE includes publication-time metadata (e.g., year/publish_date) and spans a long temporal range (1922-2022), temporal evaluation is

Table 11: Results and uncertainties of baseline model comparison (mean ± 2σ).

Category	Model Name	Micro-F1	Macro-F1	AUC-ROC
Homogeneous Graph Model	GCN	0.5057±0.0001	0.0239±0.0001	-
	GAT	0.5115±0.0043	0.0457±0.0051	-
	GraphSAGE	0.5117±0.0007	0.0474±0.0031	-
Heterogeneous Graph Model	RGCN	0.8626±0.0018	0.1442±0.0215	0.6990±0.0134
	CompGCN	0.8173±0.0324	0.2184±0.0150	0.7153±0.0217
	SimpleHGN	0.9907±0.0005	0.5052±0.0172	0.7392±0.0039
	HPN	0.7863±0.0328	0.1387±0.0141	0.4978±0.0221
	HGT	0.9791±0.0019	0.4183±0.0089	0.7288±0.0033
	NARS	0.7273±0.0130	0.0660±0.0014	-
	HAN	0.7515±0.0023	0.1706±0.0015	0.6826±0.0022
	MAGNN	0.5542±0.0020	0.0268±0.0016	0.5542±0.0019
LLM-centric Model	Llama 2	0.0623±0.0015	0.0037±0.0006	-
	Vicuna	0.0077±0.0001	0.0002±0.0001	-
	Llama 3.1	0.0872±0.0021	0.0061±0.0009	-
	Qwen 2.5	0.0981±0.0024	0.0094±0.0014	-
LLM+Graph Model	GraphGPT	0.0214±0.0013	0.0039±0.0009	-
	TAPE	0.9936±0.0030	0.5655±0.0111	-
	LLaGA	0.3516±0.0026	0.0304±0.0019	-

feasible. We conduct two complementary time-aware experiments.

Time-ordered split. We add a year-based time-ordered split for paper-node classification. Concretely, we sort paper nodes by publication year and choose two cutoff years (2017, 2020) so that the numbers of paper nodes in train/val/test are approximately 60%/20%/20% (with sampling within the boundary year, using fixed seeds). We then rerun representative methods, including GAT, HAN, HGT, and TAPE, while keeping the same experimental settings as in the main benchmark.

No-future-visibility protocol. To more strictly eliminate potential temporal information leakage, we further evaluate a "no-future-visibility" protocol. Specifically, we train models using only papers published up to the cutoff year 2017, and remove all papers published after 2017 together with their incident edges from the training graph. For other node types, we keep the nodes that are present through historical papers (i.e., connected to the $\leq$ 2017 subgraph), since these entities naturally span across years. For validation/testing, we add each future paper node as an unseen node with its textual attributes and then perform inference using the model trained on the past-only graph.

Table 12 reports the results. The time-ordered split is harder and leads to lower absolute scores due to temporal distribution shift. Importantly, the main trends remain consistent with Section 3.1.2: heterogeneous models still outperform the homogeneous baseline, and the GNN-based LLM+Graph approach remains best. The no-future-visibility protocol is stricter and further reduces performance across models, since future paper nodes and edges are completely removed during training and papers are introduced only at inference time. Nevertheless, the relative ordering is unchanged (HGT > HAN > GAT, and TAPE remains best), reinforcing that type-aware heterogeneous modeling and LLM+Graph integration remain effective under realistic temporal constraints.

Table 12: Node classification on CITE under temporal evaluation protocols.

Protocol	Model	Micro-F1	Macro-F1
Time-ordered	GAT	0.6124	0.0453
	HAN	0.7193	0.1468
	HGT	0.9431	0.3812
	TAPE	0.9736	0.5455
No-future-visibility	GAT	0.5891	0.0294
	HAN	0.6826	0.1313
	HGT	0.9389	0.2883
	TAPE	0.9355	0.5264

D.3 Long-Tail Analysis and Class-Balanced Evaluation

To investigate whether the large Micro–Macro F1 gap observed in Section 3.2.2 is mitigable and whether it affects our core conclusions, we conduct two complementary experiments on three representative models (GraphSAGE, HGT, and TAPE) without modifying their architectures or hyperparameters.

Long-tail interventions. We apply two standard strategies to the full 84-class setting: (1) class-balanced sampling (Chawla et al., 2002), which re-samples training instances with probability inversely proportional to class frequency at each epoch, and (2) balanced softmax (Ren et al., 2020), which adjusts logits by the log class prior $\log \pi_c$ estimated from the training set. Results are shown in Table 13. Both methods consistently improve Macro-F1 at a modest cost to Micro-F1, confirming that rare-class under-performance is a primary driver of the Micro-Macro gap.

Class-balanced subset. We further construct a reduced subset by removing classes with fewer than 50 samples and capping each remaining class at 200 samples (no oversampling), yielding 24 classes. Train/val/test splits follow the same 60/20/20 random protocol. Results are shown in Table 14. The Micro-Macro gap narrows substantially, further confirming that the gap is driven by the long tail.

Crucially, across both experiments the relative model ranking remains unchanged (TAPE > HGT > GraphSAGE), indicating that the core findings in Section 3.1.2: heterogeneous models outperform homogeneous ones, and GNN-based LLM+Graph

integration is most robust, are not artifacts of label imbalance.

Table 13: Long-tail interventions on the original 84-class split.

Model	Strategy	Micro-F1	Macro-F1
GraphSAGE	Original	0.5117	0.0474
GraphSAGE	Class-balanced	0.5321	0.0718
GraphSAGE	Balanced-Softmax	0.5256	0.0792
HGT	Original	0.9791	0.4183
HGT	Class-balanced	0.9568	0.4117
HGT	Balanced-Softmax	0.9453	0.4275
TAPE	Original	0.9936	0.5655
TAPE	Class-balanced	0.9886	0.5908
TAPE	Balanced-Softmax	0.9819	0.6079

Table 14: Class-balanced subset evaluation.

Model	Setting	Classes	Micro-F1	Macro-F1
GraphSAGE	Original (full)	84	0.5117	0.0474
HGT	Original (full)	84	0.9791	0.4183
TAPE	Original (full)	84	0.9936	0.5655
GraphSAGE	Balanced subset	24	0.5977	0.4626
HGT	Balanced subset	24	0.9680	0.6862
TAPE	Balanced subset	24	0.9929	0.7341

D.4 Journal-Masked Protocol and Diagnostic Baselines

Paper labels in our node classification benchmark are derived from SCI journal categories, making journal information strongly correlated with the prediction target. As suggested by the ablation results, removing journal nodes substantially reduces the performance of heterogeneous models. To make this issue explicit, we define two protocols. NC-Full uses the original heterogeneous graph with journal nodes and paper–journal edges retained. NC-NoJ removes all journal nodes and paper-journal edges; for LLM-based baselines, journal-related metadata is also excluded from prompts and model inputs.

To quantify the strength of venue identity alone, we introduce a Journal-ID baseline for NC-Full: each journal is assigned the majority SCI label of its training papers, with unseen journals mapped to the global majority class. This baseline is not applicable to NC-NoJ by construction.

Table 15 shows that Journal-ID already achieves strong performance under NC-Full, confirming that journal identity is a highly informative signal. Consistently, representative models perform substantially better in NC-Full than in NC-NoJ, indicating

Table 15: Performance comparison on NC-Full and NC-NoJ.

Model	NC-Full (Micro/Macro)	NC-NoJ (Micro/Macro)
Journal-only	0.9925/0.6063	–
SimpleHGN	0.9907/0.5052	0.5471/0.0810
HGT	0.9791/0.4183	0.5389/0.0613

that the full-graph setting is influenced by venue-related information. We therefore use NC-NoJ as a complementary diagnostic setting. This stricter protocol is also more realistic for applications such as classifying preprints or manuscripts with unknown venues.

D.5 Multi-Label Node Classification

Besides the 84-class single-label setting used in the main text, CITE also supports a finer-grained 107-label multi-label taxonomy. We report the single-label benchmark in the main tables to maintain a unified comparison across all four learning paradigms, since several LLM-centric and LLM+Graph baselines are naturally formulated as single-label prediction and cannot be directly extended to multi-label classification without substantial changes.

To complement the main benchmark, we additionally evaluate node classification on the 107-label version of CITE. We adapt representative baselines by replacing softmax cross-entropy with sigmoid activation and binary cross-entropy (BCE), without changing model architectures or hyperparameters. We further include two representative multi-label graph methods, Triper (Gupta et al., 2025) and CorGCN (Bei et al., 2025), where Triper formulates multi-label prediction as node-label link prediction and CorGCN explicitly models label correlations. We report Micro-F1 and Macro-F1 under standard sigmoid-thresholding.

Table 16: Multi-label node classification results on the 107-label version of CITE.

Model	Micro-F1	Macro-F1
GAT	0.2925	0.0055
SimpleHGN	0.3623	0.0062
HGT	0.2567	0.0255
HAN	0.3729	0.0063
Triper	0.1124	0.0070
CorGCN	0.3736	0.0063
TAPE	0.3939	0.0297

As shown in Table 16, the overall conclusions remain consistent with the main text: heterogeneous

models remain competitive under the multi-label protocol, while the Micro-F1/Macro-F1 gap becomes even larger, further highlighting the severe long-tail and label-imbalance challenge in CITE. Dedicated multi-label methods are competitive, but do not overturn the overall trend.

D.6 Cross-Dataset Evaluation of Output-Constrained LLM Protocols

To test whether the output instability observed on CITE is specific to this dataset or reflects a broader limitation of LLM-based TAG classification, we conduct an additional cross-dataset study on ACM, arXiv, and CITE. These datasets differ substantially in label-space size, with 3, 40, and 84 classes, respectively.

We evaluate the same four instruction-tuned LLMs used in our main analysis. For each dataset, we test three output protocols: (1) Original, which uses the standard free-form prompting setup; (2) Forced-choice and (3) Likelihood ranking. Compared with free-form decoding, the latter two protocols increasingly constrain the output space and reduce format-related failures. We report Micro-F1 and Macro-F1 for all settings. Table 17-19 summarizes the results.

Table 17: Performance of different output protocols on ACM.

LLM	Protocol	Micro-F1	Macro-F1
Llama 2	Original	0.6822	0.6587
	Forced-choice	0.7215	0.7102
	Ranking	0.7340	0.7119
Vicuna	Original	0.5778	0.5584
	Forced-choice	0.6230	0.6021
	Ranking	0.6417	0.6153
Llama3.1	Original	0.7422	0.7194
	Forced-choice	0.7810	0.7635
	Ranking	0.7941	0.7863
Qwen2.5	Original	0.7576	0.7298
	Forced-choice	0.8030	0.7857
	Ranking	0.8177	0.8010

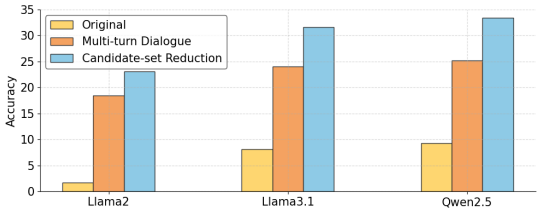

Figure 6: Performance of prompt-tuning methods across LLMs.

Table 18: Performance of different output protocols on arXiv.

LLM	Protocol	Micro-F1	Macro-F1
Llama 2	Original	0.0757	0.0052
	Forced-choice	0.1345	0.0215
	Ranking	0.1799	0.0503
Vicuna	Original	0.0531	0.0006
	Forced-choice	0.0933	0.0214
	Ranking	0.1332	0.0483
Llama 3.1	Original	0.1035	0.0292
	Forced-choice	0.1326	0.0344
	Ranking	0.1452	0.0393
Qwen 2.5	Original	0.1257	0.0215
	Forced-choice	0.1202	0.0247
	Ranking	0.1624	0.0493

Table 19: Performance of different output protocols on CITE.

LLM	Protocol	Micro-F1	Macro-F1
Llama 2	Original	0.0623	0.0037
	Forced-choice	0.1176	0.0095
	Ranking	0.1122	0.0089
Vicuna	Original	0.0077	0.0002
	Forced-choice	0.0319	0.0008
	Ranking	0.0423	0.0010
Llama 3.1	Original	0.0872	0.0061
	Forced-choice	0.0982	0.0079
	Ranking	0.1022	0.0082
Qwen 2.5	Original	0.0981	0.0094
	Forced-choice	0.1132	0.0119
	Ranking	0.1145	0.0126

These results support the interpretation in Section 3.1.3 that the weak performance of LLM-only methods on CITE is not merely caused by a particular prompt template. Rather, it reflects a broader difficulty of maintaining output consistency and label alignment in large multi-class TAG settings. At the same time, label-space size is not the only factor: compared with ACM and arXiv, CITE additionally involves stronger long-tail skew, more domain-specific terminology, and a greater reliance on heterogeneous relational signals. We therefore view this cross-dataset study as controlled supporting evidence that output-constrained prompting helps, but only partially mitigates the underlying challenge.

E Implementation Details

E.1 Models and Hyperparameters

- **GCN.** (Kipf and Welling, 2016) GCN is a classical model that works by per-

forming a linear approximation to spectral graph convolutions. Hyperparameters: hidden_channel=16, num_layers=2, dropout=0.3, learning rate=0.01.

- **GAT.** (Velickovic et al., 2017)GAT introduces the attention mechanism to capture the importance of neighboring nodes when aggregating information from the graph. Hyperparameters: hidden_channel=16, num_layers=2, dropout=0.3, learning rate=0.01.

- **GraphSAGE.** (Hamilton et al., 2017)GraphSAGE is a GNN model that focuses on inductive node classification, but can also be applied for transductive settings. Hyperparameters: hidden_channel=16, num_layers=2, dropout=0.3, learning rate=0.01.

- **RGCN.** (Schlichtkrull et al., 2018) RGCN is designed to model multi-relational knowledge graphs. It improve tasks such as link prediction and entity classification by aggregating information across different relation types. Hyperparameters: learning rate = 0.01, dropout = 0.2, hidden_dim = 64, num_layers = 3, batch_size = 128, max_epoch = 50.

- **CompGCN.** (Vashishth et al., 2019) CompGCN is a graph convolutional framework that learns embeddings for both nodes and relations in multi-relational graphs. It combines entity-relation composition from knowledge graph embedding techniques. Hyperparameters: learning rate = 0.01, hidden_dim = 32, out_dim = 32, num_layers = 2, batch_size = 128, max_epoch = 500.

- **SimpleHGN.** (Lv et al., 2021) SimpleHGN is a graph neural network designed for heterogeneous graphs. It simplifies the message-passing process between different node and relation types, aggregating features to model diverse relational structures. Hyperparameters:learning rate = 0.001, dropout = 0.2, hidden_dim = 256, num_layers = 3, num_heads = 8, batch_size = 2048, max_epoch = 500.

- **HPN.** (Ji et al., 2021) HPN is a heterogeneous graph neural network that reduce semantic confusion by weighting node semantics during aggregation and learning meta-paths. Hyperparameters: learning rate = 0.005, dropout = 0.6, hidden_dim = 64, out_dim = 16, num_layers = 2, max_epoch = 200.

- **HGT.** (Hu et al., 2020b) HGT is designed for large-scale heterogeneous graphs. It uses node and edge-type specific parameters with heterogeneous attention mechanisms. Hyperparameters: learning rate = 0.001, dropout = 0.4, hidden_dim = 64, out_dim = 16, num_layers = 2, num_heads = 8, batch_size = 5120, max_epoch = 500.

- **NARS.** (Yu et al., 2020) NARS is a method for heterogeneous graphs that trains classifiers on neighbor-averaged features from sampled relation subgraphs. It optimizes memory efficiency during training and inference while achieving state-of-the-art accuracy on multiple benchmarks. Hyperparameters: learning rate = 0.003, dropout = 0.7, hidden_dim = 64, out_dim = 16, num_heads = 8, num_hops = 2, max_epoch = 200.

- **HAN.** (Wang et al., 2019) HAN uses hierarchical attention, with node-level attention to assess node-meta-path importance and semantic-level attention to capture meta-path significance, generating node embeddings by aggregating features in a hierarchical manner. Hyperparameters: learning rate = 0.005, dropout = 0.6, hidden_dim = 128, num_heads = 8, out_dim = 16, max_epoch = 200.

- **MAGNN.** (Fu et al., 2020) MAGNN incorporates node content transformation, intra-metapath aggregation to include intermediate nodes, and inter-metapath aggregation to combine information from metapaths. Hyperparameters: learning rate = 0.005, dropout = 0.5, hidden_dim = 64, num_layers = 2, num_heads = 8, out_dim = 3, max_epoch = 10.

- **Llama-2-7b-chat-hf.** (Touvron et al., 2023) Released by Meta in July 2023, fine-tuned for chat-based applications with 7 billion parameters. The model can be accessed at meta-llama/Llama-2-7b-chat-hf. Hyperparameters: hidden_size=4096, num_attention_heads=32, max_position_embeddings=4096, num_hidden_layers=32, vocab_size=32000, initializer_range=0.02.

- **Vicuna-7b-v1.5-16k.** (Chiang et al., 2023) Developed by the LMSYS team based on

Meta's Llama 2, with 7 billion parameters and trained on a large conversational dataset from ShareGPT. The model is available at https://huggingface.co/lmsys/vicuna-7b-v1.5-16k. Hyperparameters: hidden_size=4096, num_attention_heads=32, num_hidden_layers=32, max_sequence_length=16384, vocab_size=32000, graph_hidden_size=128.

- **GraphGPT.** (Tang et al., 2024) GraphGPT is a framework that integrates LLMs with graph structural knowledge through instruction tuning. It includes a text-graph grounding module and a dual-stage tuning process with a lightweight graph-text alignment projector, allowing LLMs to understand complex graph structures. Hyperparameters: learning rate=2e-5, gnn_input, hid, output=128, context_length=128, embed_dim=128, transformer_heads=8, transformer_layers=12, transformer _width=512, vocab_size=49408, gnn_type=gt, num_nodes =127690, gt_layers=3, att_d_model=128, att_norm=true.

- **TAPE.** (He et al., 2023) TAPE leverages LLMs to extract textual features from TAGs by using zero-shot classification with explanations. These explanations are then interpreted into informative features for GNNs. In this work, we use Llama-2-7B as LLM and GraphSAGE as graph model. Hyperparameters: num_layers: [2, 3, 4], hidden_dim: [64, 128, 256], dropout: [0.3, 0.5, 0.6].

E.2 Datasets

- **arXiv(non-text)** (Hu et al., 2020a) is part of the Open Graph Benchmark (OGB), is a citation network constructed from arXiv papers. In this dataset, nodes represent academic papers, and edges denote citation relationships between them. Each node is enriched with textual features derived from the paper's title and abstract.

- **Cora(non-text)** (McCallum et al., 2000) is a widely used citation network where nodes represent scientific papers, and edges denote citation relationships. Each paper is categorized into one of several predefined topics, and each node has associated features derived from the paper's content.

- **PubMed(non-text)** (White, 2020) is a citation network derived from the PubMed repository of biomedical papers. Nodes represent papers, and edges indicate citation links. Each paper is categorized into one of several medical topics.

- **Ogbn-arXiv(text), Cora(text), PubMed(text)** sourced from Github repository provided in Chen et al (Chen et al., 2024b)

- **ACM** (Wang et al., 2019) ACM dataset is a set of extract papers published in KDD, SIGMOD, SIGCOMM, MobiCOMM, and VLDB. Wang et al divide the papers into three classes (Database, Wireless Communication, Data Mining). Then they construct a heterogeneous graph that comprises 3025 papers, 5835 authors and 56 subjects.

- **DBLP** (Tang et al., 2008) DBLP is a computer science bibliography website. Fu et al. adopt a subset of DBLP, containing 4057 au-thors, 14328 papers, 7723 terms, and 20 publication venues. The authors are divided into four research areas.