NAACL 2025

Exploring the Potential of Large Language Models for Heterophilic Graphs

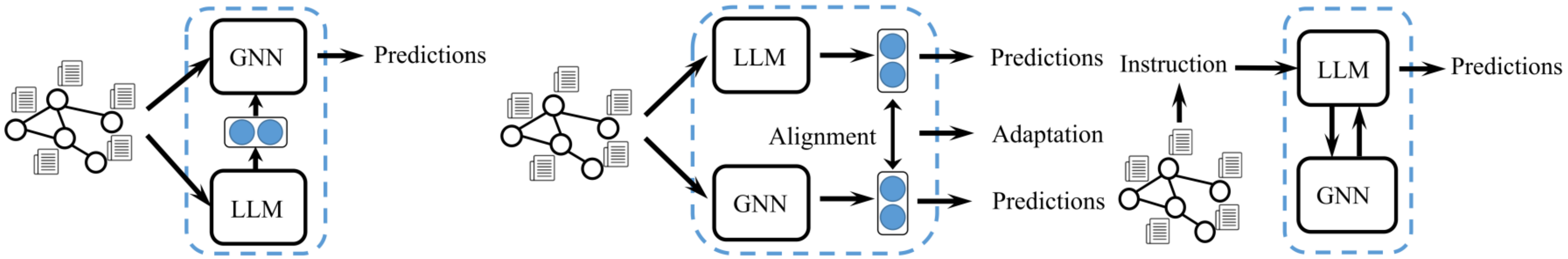
Yuxia Wu*
Singapore Management University
yieshah2017@gmail.com

Shujie Li*
Beijing University of Post and Telecommunication
shujieli@bupt.edu.cn

Yuan Fang
Singapore Management University
yfang@smu.edu.sg

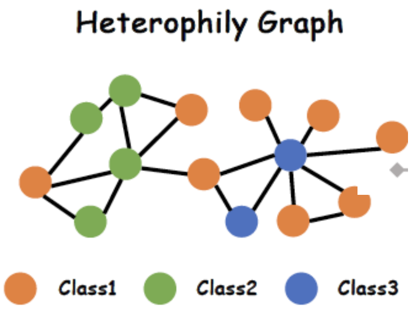
Chuan Shi
Beijing University of Post and Telecommunication
shichuan@bupt.edu.cn

Motivation: LLM for Graph



(a) GNN-centric methods. (b) Symmetric methods. (c) LLM-centric methods.

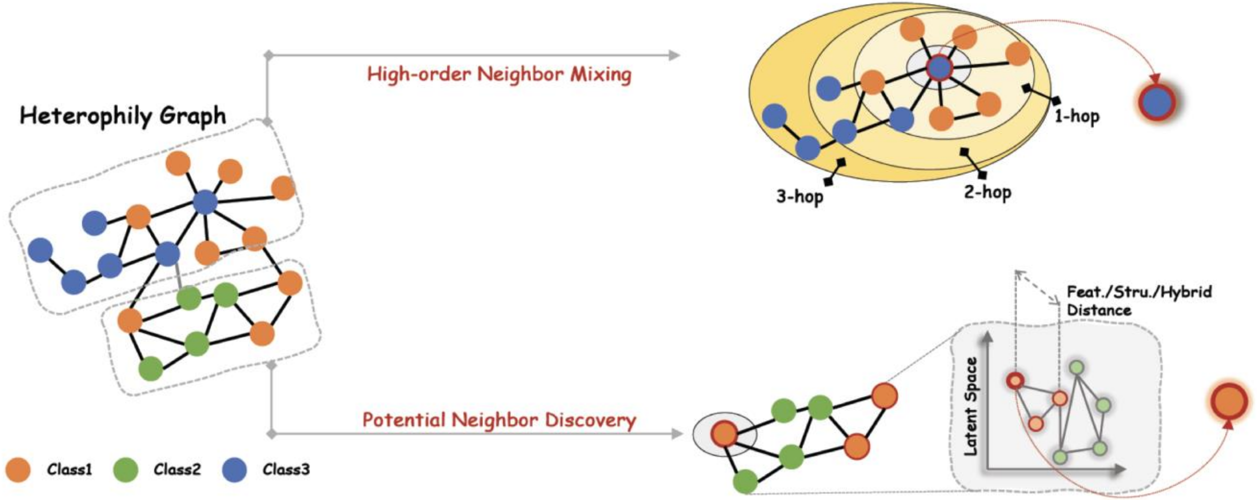
LLM for heterophilic graphs is largely unexplored.



J. Liu, et al. "Graph Foundation Models: Concepts, Opportunities and Challenges." TPAMI'25

Existing Methods: Non-local neighbor extension

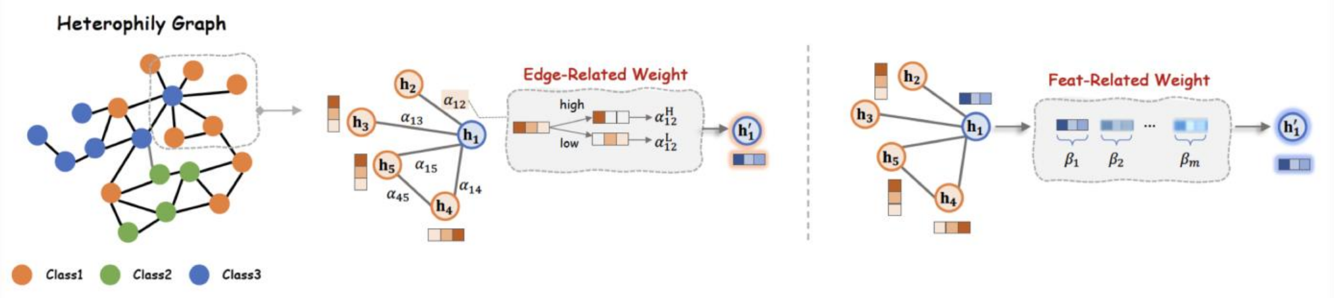
- High-order Neighbor Mixing: Mix latent information from neighbors at various distances
- Potential Neighbor Discovery: Identify suitable potential neighbors



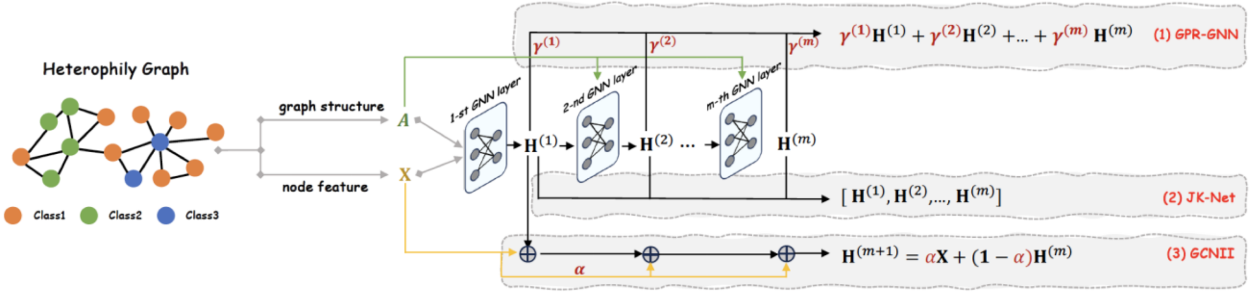
X. Zheng, *et al*. "Graph Neural Networks for Graphs with Heterophily: A Survey." ArXiv'24

Existing Methods: Architectural Refinement

- Identifiable Message Aggregation: Learn adaptive edge-aware weights for homophilic and heterophilic edges



- Inter-Layer Combination: Shallow layers: local. Deeper layers: global.



X. Zheng, *et al*. "Graph Neural Networks for Graphs with Heterophily: A Survey." ArXiv'24

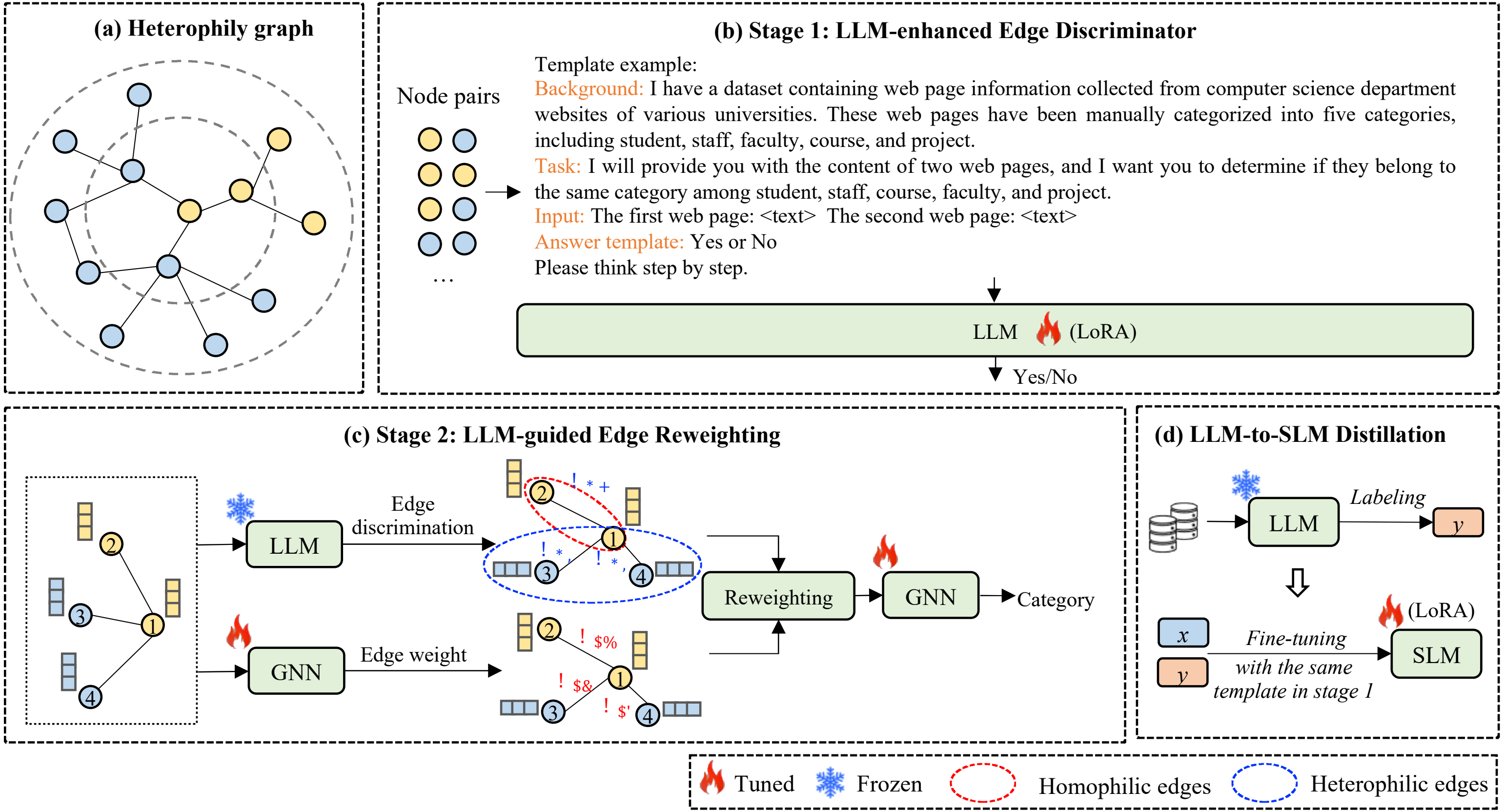
Motivation

- Limitation of Current works:
  - Heterophily-specific GNNs: Overlook the rich textual content associate with the nodes (bag-of-words, shallow embedding)
  - LLM for graphs: No current works for heterophilic graph
- Research Questions:

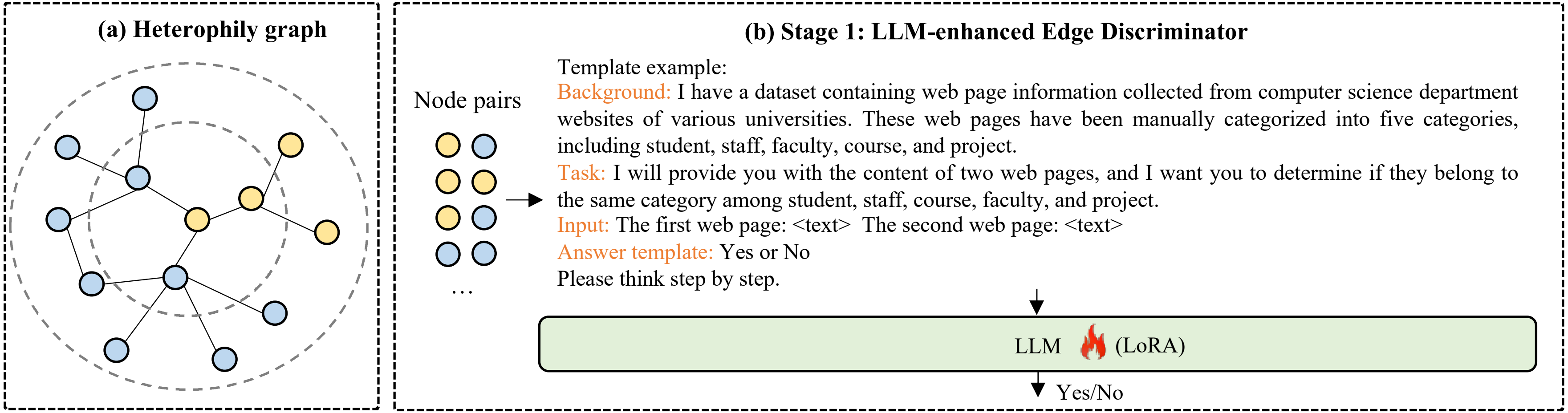
Can LLMs be effectively adapted to characterize heterophilic contexts?

Can LLMs effectively guide the fine-grained integration of heterophilic contexts into graph models?

Proposed Method: LLM for Heterophilic Graphs (LLM4HeG)

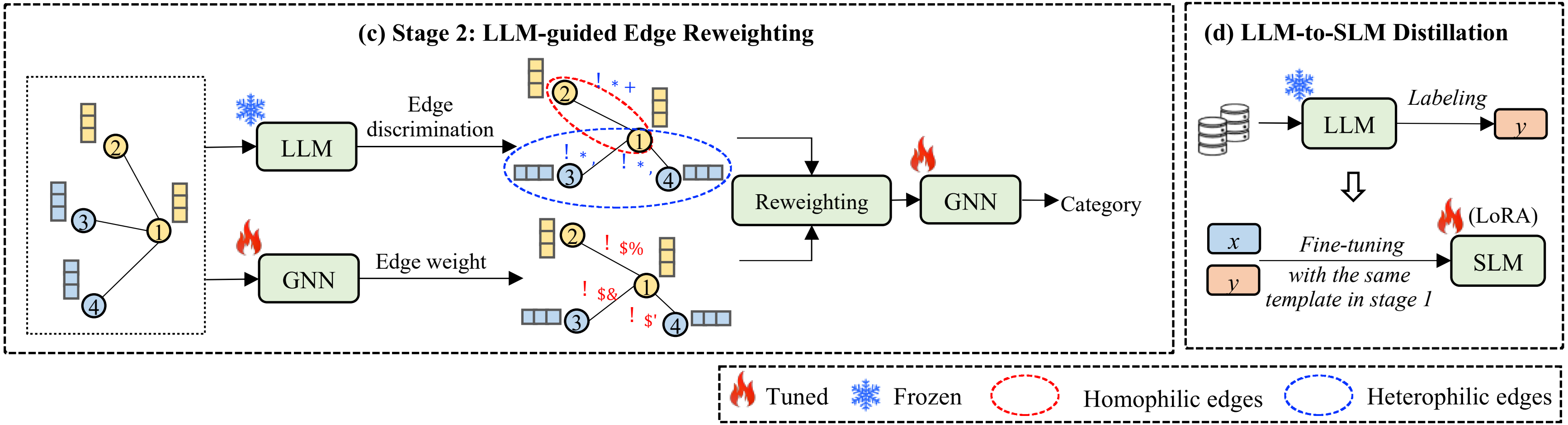


LLM4HeG: LLM-enhanced Edge Discriminator



- Construct the ground truth labels from the training set.
- Design a language template to describe the task of heterophilic edge discrimination.
- Parameter-efficient fine-tuning LLM: LoRA

LLM4HeG: LLM-guided Edge Reweighting



Edge weight from LLM:

$$w_{uv}^{\text{LLM}} = \begin{cases} \tanh(w_{\text{Ho}}) & \text{if } O_{\text{LLM}}(u,v) = \textit{Yes}, \\ \tanh(w_{\text{He}}) & \text{if } O_{\text{LLM}}(u,v) = \textit{No}, \end{cases}$$

Learnable parameter for homophilic edges and heterophilic edges

Reweighting:

$$w_{uv} = \frac{1}{2}\left(w_{uv}^{\text{LLM}} + w_{uv}^{\text{G}}\right).$$

Various GNN models for heterophilic graph

FAGCN: $w_{uv}^{\text{G}} = \tanh\left(\mathbf{g}^\top [\mathbf{h}_u \parallel \mathbf{h}_v]\right),$

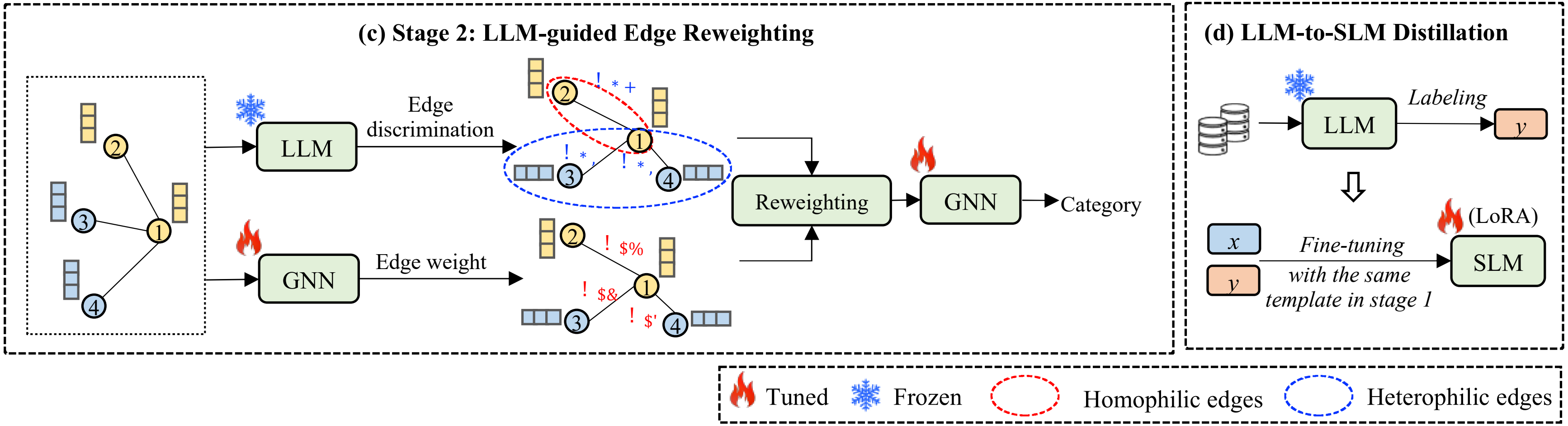
GNN prediction:

$$\mathbf{h}_v^{(0)} = \sigma(\texttt{LLM}(x_v)\mathbf{W}_e),$$

$$\mathbf{h}_v^{(l)} = \epsilon \mathbf{h}_v^{(0)} + \sum_{u \in \mathcal{N}_i(v)} \frac{w_{uv}}{\sqrt{d_u d_v}} \mathbf{h}_u^{(l-1)},$$

$$\mathbf{h}_{\text{out}} = \mathbf{W}_o \mathbf{h}_v^{(L)},$$

LLM4HeG: LLM-to-SLM Distillation



- Teacher model : fine-tuned LLM in Stage 1
- Expanded label set:
  - Pseudo-labels for additional node pairs + ground-truth labels
  - Fine-tune small language model (SLM)
- Inference: SLM

Experiments: Datasets

We collect publicly available raw text directly from the original data providers.

Dataset	Classes	Nodes	Edges	$\mathcal{H}(G)$
Cornell	5	195	304	0.13
Texas	5	187	328	0.12
Wisconsin	5	265	530	0.20
Actor	5	4,416	12,172	0.56
Amazon	5	24,492	93,050	0.38

Table 1: Dataset statistics.

The level of homophily

1 : perfect homophily

0: total heterophily

Dataset	Cornell	Texas	Wisconsin	Actor	Amazon
Training	4,186	3,741	7,626	36,248	23,210
Distillation$\star$	916	991	1,299	1,781	11,422

$\star$: the number of additional samples for distillation .

Table 5: The number of node pairs in Stage 1 and distillation.

Experiment: Accuracy

Methods	Cornell	Texas	Wisconsin	Actor	Amazon
		Classic GNNs			
GCN	52.86±1.8	43.64±3.3	41.40±1.8	66.70±1.3	39.33±1.0
GraphSAGE	75.71±1.8	81.82±2.5	80.35±1.3	70.37±0.1	46.63±0.1
GAT	54.28±5.1	51.36±2.3	50.53±1.7	63.74±6.7	35.12±6.4
		Heterophily-specific GNNs			
H2GCN	69.76±3.0	79.09±3.5	80.18±1.9	70.73±0.9	47.09±0.3
FAGCN	76.43±3.1	84.55±4.8	83.16±1.4	75.58±0.5	49.83±0.6
JacobiConv	73.57±4.3	81.80±4.1	76.31±11.3	73.81±0.3	49.43±0.5
GBK-GNN	66.19±2.8	80.00±3.0	72.98±3.3	72.49±1.0	44.90±0.3
OGNN	71.91±1.8	85.00±2.3	79.30±2.1	72.08±2.4	47.79±1.6
SEGSL	66.67±4.1	85.00±2.0	79.30±1.8	72.73±0.8	47.38±0.2
DisamGCL	50.48±2.0	65.00±1.2	57.89±0.0	67.78±0.3	43.90±0.4
		LLM4HeG (fine-tuned LLM/SLMs and distilled SLMs)			
Vicuna 7B	**77.62**±2.9	**89.09**±3.3	86.14±2.1	**76.82**±0.5	51.53±0.4
Bloom 560M	75.48±2.1	80.00±4.0	86.49±1.9	76.16±0.6	51.52±0.5
Bloom 1B	75.71±1.4	83.86±2.8	83.86±1.7	74.99±0.5	**52.33**±0.6
7B-to-560M	75.00±4.0	88.18±2.2	**87.19**±2.5	75.78±0.2	51.51±0.4
7B-to-1B	77.38±2.7	88.18±4.0	86.14±1.5	75.37±0.9	51.58±0.4

Directly fine-tune

Distillation

Table 2: Accuracy for node classification of different methods. (Best results bolded; runners-up underlined.)

- Heterophily-specific GNNs generally outperform classic GNNs
- Our methods consistently achieve the best performance
- Fine-tuned LLM > Fine-tuned SLMs
- Fine-tuned LLM ~= Distilled SLMs

We use the initial node features derived from the Vicuna 7B model for all methods.

Experiment: Analysis of edge discrimination by LLM/SLMs

Model	Cornell	Texas	Wisconsin	Actor	Amazon	Average
Vicuna 7B	65.71	64.00	92.66	81.50	44.68	69.71
Bloom 560M	47.62	26.51	71.62	79.02	56.26	56.21
Bloom 1B	40.86	23.91	79.76	79.52	59.89	56.78
7B-to-560M	50.85	64.86	80.75	81.03	50.77	65.65
7B-to-1B	51.72	80.00	75.95	80.47	51.48	67.92

Table 3: F1 scores for edge discrimination of fine-tuned LLM/SLMs and distilled SLMs.

- Fine-tuned LLM > Fine-tuned SLMs
- Fine-tuned LLM ~= Distilled SLMs

Experiment: Efficiency study

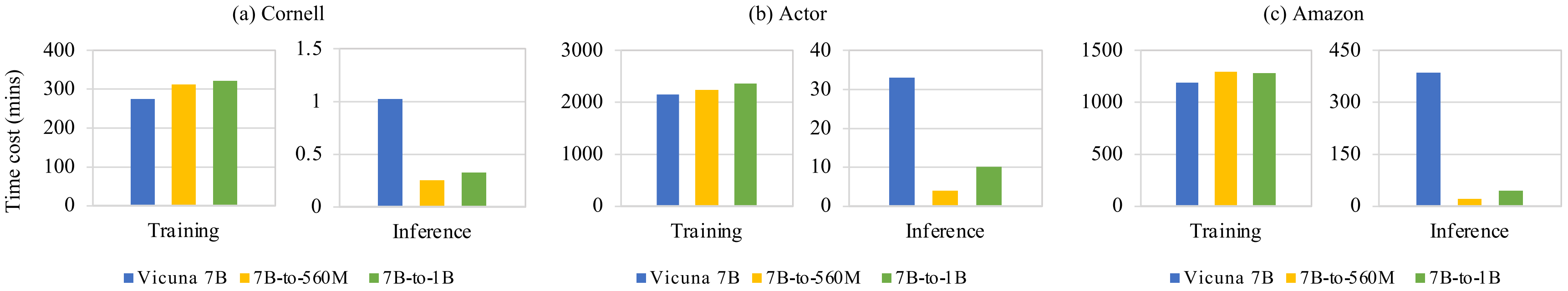


Figure 3: Analysis on the efficiency of the fine-tuned LLM and distilled SLMs.

- Training time:
  - LLM: fine-tune time in Stage 1
  - Distilled SLMs: fine-tuning LLM + generating the pseudo-labels + fine-tuning the SLM

- The inference time of SLMs are significantly lower than LLMs
- The distilled SLMs can be more easily deployed

Experiment: Plug-and-play with various backbones

	Cornell	Texas	Wisconsin	Actor	Amazon
GCN	52.86±1.8	43.64±3.3	41.40±1.8	66.70±1.3	39.33±1.0
+LLM4HeG	66.19±1.0	68.18±2.0	76.84±2.6	71.68±1.0	40.98±0.7
GAT	54.28±5.1	51.36±2.3	50.53±1.7	63.74±6.7	35.12±6.4
+LLM4HeG	58.57±4.9	58.18±2.3	57.54±6.1	70.78±0.7	36.01±5.8
H2GCN	69.76±3.0	79.09±3.5	80.18±1.9	70.73±0.9	47.09±0.3
+LLM4HeG	76.43±3.6	84.77±1.0	86.49±1.1	74.51±0.6	52.14±0.4
FAGCN	76.43±3.1	84.55±4.8	83.16±1.4	75.58±0.5	49.83±0.6
+LLM4HeG	77.62±2.9	89.09±3.3	86.14±2.1	76.82±0.5	51.53±0.4

Table 4: The accuracy for node classification of LLM4HeG with different backbones.

- Our method can be integrated with various GNN backbones.
- Our method enhances the performance of various backbones.

Summary:

- We explored the potential of LLMs to enhance the performance of GNNs for node classification on heterophilic graphs.
- We introduced a novel two-stage framework LLM4HeG, including an LLM-enhanced edge discriminator and an LLM-guided edge reweighting module.
- We implemented model distillation techniques to create smaller models that achieve much faster inference while maintaining competitive performance.

Thanks & QA

Our paper:
https://arxiv.org/pdf/2408.14134

Homepage: https://yuxiawu.github.io/