ICML
International Conference
On Machine Learning

When Do LLMs Help With Node Classification? A Comprehensive Analysis

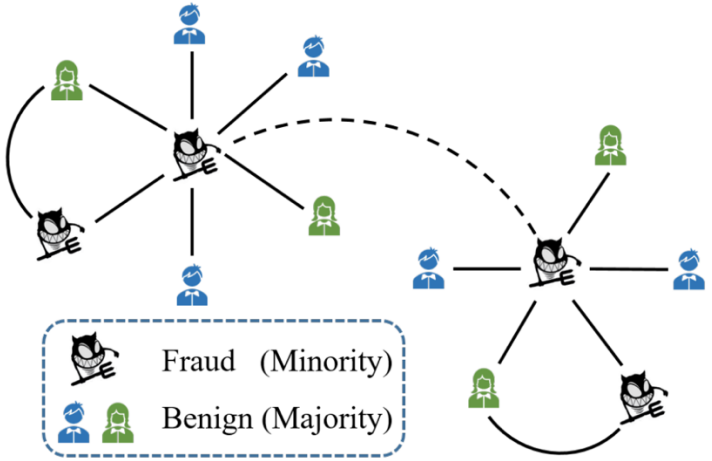
Xixi Wu[1] Yifei Shen[2] Fangzhou Ge[1] Caihua Shan[2]
Yizhu Jiao[3] Xiangguo Sun[1] Hong Cheng[1]

[1]The Chinese University of Hong Kong [2]Microsoft Research Asia
[3]University of Illinois Urbana-Champaign

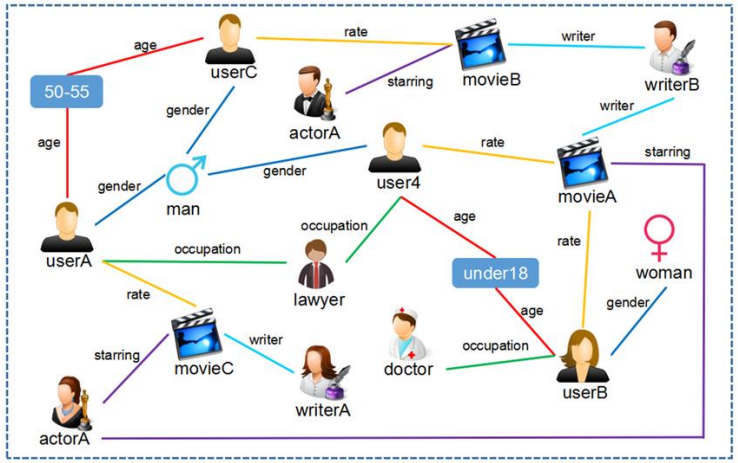
香港中文大學
The Chinese University of Hong Kong

Motivation

- Node classification is a fundamental task in graph analysis



Fraud detection



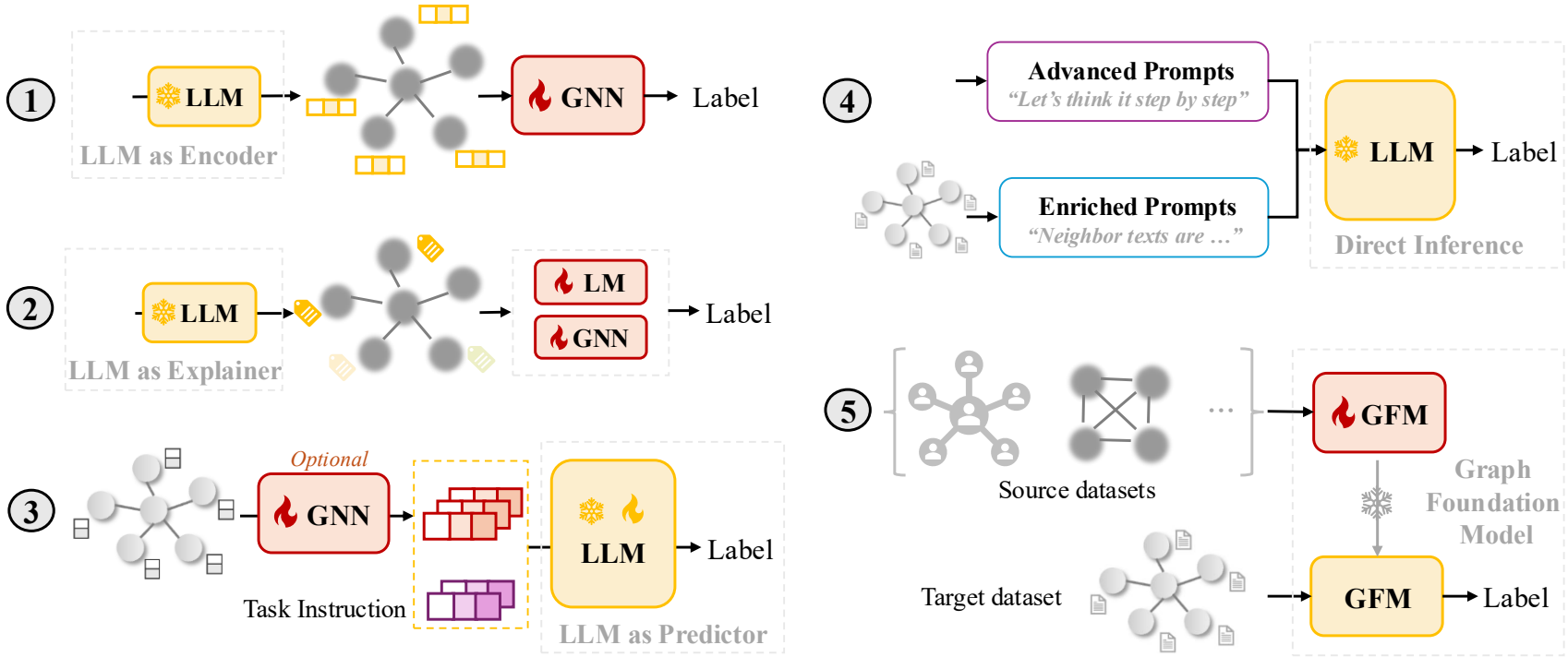
User profiling

Item tagging
...

Even a marginal improvement in classification accuracy could result in **substantial financial profits**

Motivation

- Leveraging LLMs for node classification has become popular



(a) Semi- / Supervised Settings

(b) Zero-shot Setting

Superior semantic comprehension *of LLMs overcome the limitations of shallow embeddings*

Motivation

- Designing principles for LLM-based node classification algorithms remain elusive
  - For each algorithm category, what is the most suitable setting?
  - Under what scenarios, LLMs can surpass traditional LMs like BERT?
  - ...

Unified Benchmark
Evaluate all methods using consistent dataloaders, learning paradigms, backbones, and implementation codebases

Controlled Experiments
Consider comprehensive variables including learning paradigms, language model type & size, graph characteristics, etc

Benchmark - LLMNodeBed

- Contains **14** datasets with varying scales, domain, and homophily
- Integrates **8** LLM-based algorithms, **8** classic methods
- Supports **3** different learning paradigms: semi-supervised, supervised, and zero-shot

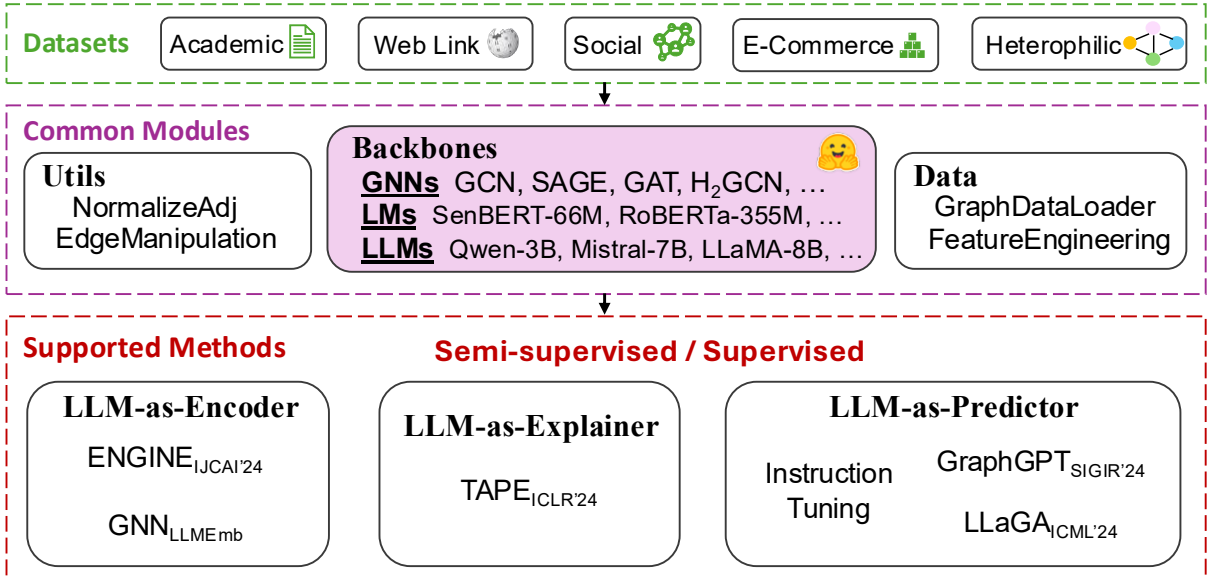


Table 1: **Statistics of supported datasets in LLMNodeBed.**

Statistics	Academic				Web Link	Social		E-Commerce			Heterophilic			
	Cora	Citeseer	Pubmed	arXiv	WikiCS	Instagram	Reddit	Books	Photo	Computer	Cornell	Texas	Wisconsin	Washington
# Classes	7	6	3	40	10	2	2	12	12	10	5	5	5	5
# Nodes	2,708	3,186	19,717	169,343	11,701	11,339	33,434	41,551	48,362	87,229	191	187	265	229
# Edges	5,429	4,277	44,338	1,166,243	215,863	144,010	198,448	358,574	500,928	721,081	292	310	510	394
Avg. # Token	183.4	210.0	446.5	239.8	629.9	56.2	197.3	337.0	201.5	123.1	594.6	453.2	639.1	469.0
Homophily (%)	82.52	72.93	79.24	63.53	68.67	63.35	55.52	78.05	78.50	85.28	11.55	6.69	16.27	17.07

Experiments & Findings

- Semi-supervised & Supervised

Comparing Classic and LLM-based, ***introducing LLMs to exploit textual information is useful***

Semi-supervised		Cora	Citeseer	Pubmed	WikiCS	Instagram	Reddit	Books	Photo	Computer	Avg.
Classic	$\text{GCN}_{\text{ShallowEmb}}$	$82.30_{\pm 0.19}$	$70.55_{\pm 0.32}$	$78.94_{\pm 0.27}$	$79.86_{\pm 0.19}$	$63.50_{\pm 0.11}$	$61.44_{\pm 0.38}$	$68.79_{\pm 0.46}$	$69.25_{\pm 0.81}$	$71.44_{\pm 1.19}$	71.79
	$\text{SAGE}_{\text{ShallowEmb}}$	$82.27_{\pm 0.37}$	$69.56_{\pm 0.43}$	$77.88_{\pm 0.44}$	$79.67_{\pm 0.25}$	$63.57_{\pm 0.10}$	$56.65_{\pm 0.33}$	$72.01_{\pm 0.33}$	$78.50_{\pm 0.15}$	$81.43_{\pm 0.27}$	73.50
	$\text{GAT}_{\text{ShallowEmb}}$	$81.30_{\pm 0.67}$	$69.94_{\pm 0.74}$	$78.49_{\pm 0.70}$	$79.99_{\pm 0.65}$	$63.56_{\pm 0.04}$	$60.60_{\pm 1.17}$	$74.35_{\pm 0.35}$	$80.40_{\pm 0.45}$	$83.39_{\pm 0.22}$	74.67
	SenBERT-66M	$66.66_{\pm 1.42}$	$60.52_{\pm 1.62}$	$36.04_{\pm 2.92}$	$77.77_{\pm 0.75}$	$59.00_{\pm 1.17}$	$56.05_{\pm 0.41}$	$83.68_{\pm 0.19}$	$73.89_{\pm 0.31}$	$70.76_{\pm 0.15}$	64.93
	RoBERTa-355M	$72.24_{\pm 1.14}$	$66.68_{\pm 2.03}$	$42.32_{\pm 1.56}$	$76.81_{\pm 1.04}$	$63.52_{\pm 0.44}$	$59.27_{\pm 0.34}$	$84.62_{\pm 0.16}$	$74.79_{\pm 1.13}$	$72.31_{\pm 0.37}$	68.06
	GLEM	$81.30_{\pm 0.88}$	$68.80_{\pm 2.46}$	**$81.70_{\pm 1.07}$**	$76.43_{\pm 0.55}$	$60.25_{\pm 3.66}$	$55.13_{\pm 1.41}$	$83.28_{\pm 0.39}$	$76.93_{\pm 0.49}$	$80.46_{\pm 1.45}$	73.81
Encoder	$\text{GCN}_{\text{LLMEmb}}$	$83.33_{\pm 0.75}$	$71.39_{\pm 0.90}$	$78.71_{\pm 0.45}$	$80.94_{\pm 0.16}$	**$67.49_{\pm 0.43}$**	$68.65_{\pm 0.75}$	$83.03_{\pm 0.34}$	$84.84_{\pm 0.47}$	$88.22_{\pm 0.16}$	**78.51**
	ENGINE	**$84.22_{\pm 0.46}$**	**$72.14_{\pm 0.74}$**	$77.84_{\pm 0.27}$	$80.94_{\pm 0.19}$	$67.14_{\pm 0.46}$	$69.67_{\pm 0.16}$	$82.89_{\pm 0.14}$	$84.33_{\pm 0.57}$	$86.42_{\pm 0.23}$	78.40
Explainer	TAPE	$84.04_{\pm 0.24}$	$71.87_{\pm 0.35}$	$78.61_{\pm 1.23}$	**$81.94_{\pm 0.16}$**	$66.07_{\pm 0.10}$	$62.43_{\pm 0.47}$	**$84.92_{\pm 0.26}$**	**$86.46_{\pm 0.12}$**	**$89.52_{\pm 0.04}$**	78.43
Predictor	LLM_{IT}	$67.00_{\pm 0.16}$	$54.26_{\pm 0.22}$	**$80.99_{\pm 0.43}$**	$75.02_{\pm 0.16}$	$41.83_{\pm 0.47}$	$54.09_{\pm 1.02}$	$80.92_{\pm 1.38}$	$71.28_{\pm 1.81}$	$66.99_{\pm 2.02}$	65.76
	GraphGPT	$64.72_{\pm 1.50}$	$64.58_{\pm 1.55}$	$70.34_{\pm 2.27}$	$70.71_{\pm 0.37}$	$62.88_{\pm 2.14}$	$58.25_{\pm 0.37}$	$81.13_{\pm 1.52}$	$77.48_{\pm 0.78}$	$80.10_{\pm 0.76}$	70.02
	LLaGA	$78.94_{\pm 1.14}$	$62.61_{\pm 3.63}$	$65.91_{\pm 2.09}$	$76.47_{\pm 2.20}$	$65.84_{\pm 0.72}$	**$70.10_{\pm 0.38}$**	$83.47_{\pm 0.45}$	$84.44_{\pm 0.90}$	$87.82_{\pm 0.53}$	75.07

Supervised		Cora	Citeseer	Pubmed	arXiv	WikiCS	Instagram	Reddit	Books	Photo	Computer	Avg.
Classic	$\text{GCN}_{\text{ShallowEmb}}$	$87.41_{\pm 2.08}$	$75.74_{\pm 1.20}$	$89.01_{\pm 0.59}$	$71.39_{\pm 0.28}$	$83.67_{\pm 0.63}$	$63.94_{\pm 0.61}$	$65.07_{\pm 0.38}$	$76.94_{\pm 0.26}$	$73.34_{\pm 1.34}$	$77.16_{\pm 3.80}$	76.37
	$\text{SAGE}_{\text{ShallowEmb}}$	$87.44_{\pm 1.74}$	$74.96_{\pm 1.20}$	$90.47_{\pm 0.25}$	$71.21_{\pm 0.18}$	$84.86_{\pm 0.91}$	$64.14_{\pm 0.47}$	$61.52_{\pm 0.60}$	$79.40_{\pm 0.45}$	$84.59_{\pm 0.32}$	$87.77_{\pm 0.34}$	78.64
	$\text{GAT}_{\text{ShallowEmb}}$	$86.68_{\pm 1.12}$	$73.73_{\pm 0.94}$	$88.25_{\pm 0.47}$	$71.57_{\pm 0.25}$	$83.94_{\pm 0.61}$	$64.93_{\pm 0.75}$	$64.16_{\pm 1.05}$	$80.61_{\pm 0.49}$	$84.84_{\pm 0.69}$	$88.32_{\pm 0.24}$	78.70
	SenBERT-66M	$79.61_{\pm 1.40}$	$74.06_{\pm 1.26}$	$94.47_{\pm 0.33}$	$72.66_{\pm 0.24}$	$86.51_{\pm 0.86}$	$60.11_{\pm 0.93}$	$58.70_{\pm 0.54}$	$85.99_{\pm 0.58}$	$77.72_{\pm 0.35}$	$74.22_{\pm 0.21}$	76.40
	RoBERTa-355M	$83.17_{\pm 0.84}$	$75.90_{\pm 1.69}$	**$94.84_{\pm 0.06}$**	$74.12_{\pm 0.12}$	**$87.47_{\pm 0.83}$**	$63.75_{\pm 1.13}$	$60.61_{\pm 0.24}$	**$86.65_{\pm 0.38}$**	$79.45_{\pm 0.37}$	$75.76_{\pm 0.30}$	78.17
	GLEM	$86.81_{\pm 1.19}$	$73.24_{\pm 1.55}$	$93.98_{\pm 0.32}$	$73.55_{\pm 0.22}$	$79.81_{\pm 0.45}$	$67.39_{\pm 1.73}$	$53.11_{\pm 2.96}$	$83.98_{\pm 0.97}$	$78.16_{\pm 0.45}$	$81.63_{\pm 0.46}$	77.17
Encoder	$\text{GCN}_{\text{LLMEmb}}$	**$88.15_{\pm 1.79}$**	$76.45_{\pm 1.19}$	$88.38_{\pm 0.68}$	$74.39_{\pm 0.31}$	$84.78_{\pm 0.86}$	$68.27_{\pm 0.45}$	$70.65_{\pm 0.75}$	$84.23_{\pm 0.20}$	$86.07_{\pm 0.20}$	$89.52_{\pm 0.31}$	81.09
	ENGINE	$87.00_{\pm 1.60}$	$75.82_{\pm 1.52}$	$90.08_{\pm 0.16}$	$74.69_{\pm 0.36}$	$85.44_{\pm 0.53}$	$68.87_{\pm 0.25}$	**$71.21_{\pm 0.77}$**	$84.09_{\pm 0.09}$	$86.98_{\pm 0.06}$	$89.05_{\pm 0.13}$	81.32
Explainer	TAPE	$88.05_{\pm 1.76}$	$76.45_{\pm 1.60}$	$93.00_{\pm 0.13}$	$74.96_{\pm 0.14}$	$87.11_{\pm 0.66}$	$68.11_{\pm 0.54}$	$66.22_{\pm 0.83}$	$85.95_{\pm 0.59}$	**$87.72_{\pm 0.28}$**	**$90.46_{\pm 0.18}$**	**81.80**
Predictor	LLM_{IT}	$71.93_{\pm 1.47}$	$60.97_{\pm 3.97}$	$94.16_{\pm 0.19}$	**76.08**	$80.61_{\pm 0.47}$	$44.20_{\pm 3.06}$	$58.30_{\pm 0.48}$	$84.80_{\pm 0.13}$	$78.27_{\pm 0.54}$	$74.51_{\pm 0.53}$	72.38
	GraphGPT	$82.29_{\pm 0.26}$	$74.67_{\pm 1.15}$	$93.54_{\pm 0.22}$	$75.15_{\pm 0.14}$	$82.54_{\pm 0.23}$	$67.00_{\pm 1.22}$	$60.72_{\pm 1.47}$	$85.38_{\pm 0.72}$	$84.46_{\pm 0.36}$	$86.78_{\pm 1.14}$	79.25
	LLaGA	$87.55_{\pm 1.15}$	**$76.73_{\pm 1.70}$**	$90.28_{\pm 0.91}$	$74.49_{\pm 0.23}$	$84.03_{\pm 1.10}$	**$69.16_{\pm 0.72}$**	$71.06_{\pm 0.38}$	$85.56_{\pm 0.30}$	$87.62_{\pm 0.30}$	$90.41_{\pm 0.12}$	81.69

Experiments & Findings

- Semi-supervised & Supervised

Comparing Semi-supervised and Supervised,
LLMs can bring greater improvements in semi-supervised settings than supervised

Semi-supervised		Cora	Citeseer	Pubmed	WikiCS	Instagram	Reddit	Books	Photo	Computer	Avg.
Classic	$\text{GCN}_{\text{ShallowEmb}}$	$82.30_{\pm 0.19}$	$70.55_{\pm 0.32}$	$78.94_{\pm 0.27}$	$79.86_{\pm 0.19}$	$63.50_{\pm 0.11}$	$61.44_{\pm 0.38}$	$68.79_{\pm 0.46}$	$69.25_{\pm 0.81}$	$71.44_{\pm 1.19}$	71.79
	$\text{SAGE}_{\text{ShallowEmb}}$	$82.27_{\pm 0.37}$	$69.56_{\pm 0.43}$	$77.88_{\pm 0.44}$	$79.67_{\pm 0.25}$	$63.57_{\pm 0.10}$	$56.65_{\pm 0.33}$	$72.01_{\pm 0.33}$	$78.50_{\pm 0.15}$	$81.43_{\pm 0.27}$	73.50
	$\text{GAT}_{\text{ShallowEmb}}$	$81.30_{\pm 0.67}$	$69.94_{\pm 0.74}$	$78.49_{\pm 0.70}$	$79.99_{\pm 0.65}$	$63.56_{\pm 0.04}$	$60.60_{\pm 1.17}$	$74.35_{\pm 0.35}$	$80.40_{\pm 0.45}$	$83.39_{\pm 0.22}$	74.67
	SenBERT-66M	$66.66_{\pm 1.42}$	$60.52_{\pm 1.62}$	$36.04_{\pm 2.92}$	$77.77_{\pm 0.75}$	$59.00_{\pm 1.17}$	$56.05_{\pm 0.41}$	$83.68_{\pm 0.19}$	$73.89_{\pm 0.31}$	$70.76_{\pm 0.15}$	64.93
	RoBERTa-355M	$72.24_{\pm 1.14}$	$66.68_{\pm 2.03}$	$42.32_{\pm 1.56}$	$76.81_{\pm 1.04}$	$63.52_{\pm 0.44}$	$59.27_{\pm 0.34}$	$84.62_{\pm 0.16}$	$74.79_{\pm 1.13}$	$72.31_{\pm 0.37}$	68.06
	GLEM	$81.30_{\pm 0.88}$	$68.80_{\pm 2.46}$	**$81.70_{\pm 1.07}$**	$76.43_{\pm 0.55}$	$60.25_{\pm 3.66}$	$55.13_{\pm 1.41}$	$83.28_{\pm 0.39}$	$76.93_{\pm 0.49}$	$80.46_{\pm 1.45}$	73.81
Encoder	$\text{GCN}_{\text{LLMEmb}}$	$83.33_{\pm 0.75}$	$71.39_{\pm 0.90}$	$78.71_{\pm 0.45}$	$80.94_{\pm 0.16}$	**$67.49_{\pm 0.43}$**	$68.65_{\pm 0.75}$	$83.03_{\pm 0.34}$	$84.84_{\pm 0.47}$	$88.22_{\pm 0.16}$	**78.51**
	ENGINE	**$84.22_{\pm 0.46}$**	**$72.14_{\pm 0.74}$**	$77.84_{\pm 0.27}$	$80.94_{\pm 0.19}$	$67.14_{\pm 0.46}$	$69.67_{\pm 0.16}$	$82.89_{\pm 0.14}$	$84.33_{\pm 0.57}$	$86.42_{\pm 0.23}$	78.40
Explainer	TAPE	$84.04_{\pm 0.24}$	$71.87_{\pm 0.35}$	$78.61_{\pm 1.23}$	**$81.94_{\pm 0.16}$**	$66.07_{\pm 0.10}$	$62.43_{\pm 0.47}$	**$84.92_{\pm 0.26}$**	**$86.46_{\pm 0.12}$**	**$89.52_{\pm 0.04}$**	78.43
Predictor	LLM_{IT}	$67.00_{\pm 0.16}$	$54.26_{\pm 0.22}$	**$80.99_{\pm 0.43}$**	$75.02_{\pm 0.16}$	$41.83_{\pm 0.47}$	$54.09_{\pm 1.02}$	$80.92_{\pm 1.38}$	$71.28_{\pm 1.81}$	$66.99_{\pm 2.02}$	65.76
	GraphGPT	$64.72_{\pm 1.50}$	$64.58_{\pm 1.55}$	$70.34_{\pm 2.27}$	$70.71_{\pm 0.37}$	$62.88_{\pm 2.14}$	$58.25_{\pm 0.37}$	$81.13_{\pm 1.52}$	$77.48_{\pm 0.78}$	$80.10_{\pm 0.76}$	70.02
	LLaGA	$78.94_{\pm 1.14}$	$62.61_{\pm 3.63}$	$65.91_{\pm 2.09}$	$76.47_{\pm 2.20}$	$65.84_{\pm 0.72}$	**$70.10_{\pm 0.38}$**	$83.47_{\pm 0.45}$	$84.44_{\pm 0.90}$	$87.82_{\pm 0.53}$	75.07

Supervised		Cora	Citeseer	Pubmed	arXiv	WikiCS	Instagram	Reddit	Books	Photo	Computer	Avg.
Classic	$\text{GCN}_{\text{ShallowEmb}}$	$87.41_{\pm 2.08}$	$75.74_{\pm 1.20}$	$89.01_{\pm 0.59}$	$71.39_{\pm 0.28}$	$83.67_{\pm 0.63}$	$63.94_{\pm 0.61}$	$65.07_{\pm 0.38}$	$76.94_{\pm 0.26}$	$73.34_{\pm 1.34}$	$77.16_{\pm 3.80}$	76.37
	$\text{SAGE}_{\text{ShallowEmb}}$	$87.44_{\pm 1.74}$	$74.96_{\pm 1.20}$	$90.47_{\pm 0.25}$	$71.21_{\pm 0.18}$	$84.86_{\pm 0.91}$	$64.14_{\pm 0.47}$	$61.52_{\pm 0.60}$	$79.40_{\pm 0.45}$	$84.59_{\pm 0.32}$	$87.77_{\pm 0.34}$	78.64
	$\text{GAT}_{\text{ShallowEmb}}$	$86.68_{\pm 1.12}$	$73.73_{\pm 0.94}$	$88.25_{\pm 0.47}$	$71.57_{\pm 0.25}$	$83.94_{\pm 0.61}$	$64.93_{\pm 0.75}$	$64.16_{\pm 1.05}$	$80.61_{\pm 0.49}$	$84.84_{\pm 0.69}$	$88.32_{\pm 0.24}$	78.70
	SenBERT-66M	$79.61_{\pm 1.40}$	$74.06_{\pm 1.26}$	$94.47_{\pm 0.33}$	$72.66_{\pm 0.24}$	$86.51_{\pm 0.86}$	$60.11_{\pm 0.93}$	$58.70_{\pm 0.54}$	$85.99_{\pm 0.58}$	$77.72_{\pm 0.35}$	$74.22_{\pm 0.21}$	76.40
	RoBERTa-355M	$83.17_{\pm 0.84}$	$75.90_{\pm 1.69}$	**$94.84_{\pm 0.06}$**	$74.12_{\pm 0.12}$	**$87.47_{\pm 0.83}$**	$63.75_{\pm 1.13}$	$60.61_{\pm 0.24}$	**$86.65_{\pm 0.38}$**	$79.45_{\pm 0.37}$	$75.76_{\pm 0.30}$	78.17
	GLEM	$86.81_{\pm 1.19}$	$73.24_{\pm 1.55}$	$93.98_{\pm 0.32}$	$73.55_{\pm 0.22}$	$79.81_{\pm 0.45}$	$67.39_{\pm 1.73}$	$53.11_{\pm 2.96}$	$83.98_{\pm 0.97}$	$78.16_{\pm 0.45}$	$81.63_{\pm 0.46}$	77.17
Encoder	$\text{GCN}_{\text{LLMEmb}}$	**$88.15_{\pm 1.79}$**	$76.45_{\pm 1.19}$	$88.38_{\pm 0.68}$	$74.39_{\pm 0.31}$	$84.78_{\pm 0.86}$	$68.27_{\pm 0.45}$	$70.65_{\pm 0.75}$	$84.23_{\pm 0.20}$	$86.07_{\pm 0.20}$	$89.52_{\pm 0.31}$	81.09
	ENGINE	$87.00_{\pm 1.60}$	$75.82_{\pm 1.52}$	$90.08_{\pm 0.16}$	$74.69_{\pm 0.36}$	$85.44_{\pm 0.53}$	$68.87_{\pm 0.25}$	**$71.21_{\pm 0.77}$**	$84.09_{\pm 0.09}$	$86.98_{\pm 0.06}$	$89.05_{\pm 0.13}$	81.32
Explainer	TAPE	$88.05_{\pm 1.76}$	$76.45_{\pm 1.60}$	$93.00_{\pm 0.13}$	$74.96_{\pm 0.14}$	$87.11_{\pm 0.66}$	$68.11_{\pm 0.54}$	$66.22_{\pm 0.83}$	$85.95_{\pm 0.59}$	**$87.72_{\pm 0.28}$**	**$90.46_{\pm 0.18}$**	**81.80**
Predictor	LLM_{IT}	$71.93_{\pm 1.47}$	$60.97_{\pm 3.97}$	$94.16_{\pm 0.19}$	**76.08**	$80.61_{\pm 0.47}$	$44.20_{\pm 3.06}$	$58.30_{\pm 0.48}$	$84.80_{\pm 0.13}$	$78.27_{\pm 0.54}$	$74.51_{\pm 0.53}$	72.38
	GraphGPT	$82.29_{\pm 0.26}$	$74.67_{\pm 1.15}$	$93.54_{\pm 0.22}$	$75.15_{\pm 0.14}$	$82.54_{\pm 0.23}$	$67.00_{\pm 1.22}$	$60.72_{\pm 1.47}$	$85.38_{\pm 0.72}$	$84.46_{\pm 0.36}$	$86.78_{\pm 1.14}$	79.25
	LLaGA	$87.55_{\pm 1.15}$	**$76.73_{\pm 1.70}$**	$90.28_{\pm 0.91}$	$74.49_{\pm 0.23}$	$84.03_{\pm 1.10}$	**$69.16_{\pm 0.72}$**	$71.06_{\pm 0.38}$	$85.56_{\pm 0.30}$	$87.62_{\pm 0.30}$	$90.41_{\pm 0.12}$	81.69

Experiments & Findings

- Semi-supervised & Supervised

Semi-supervised		Cora	Citeseer	Pubmed	WikiCS	Instagram	Reddit	Books	Photo	Computer	Avg.
Classic	$\text{GCN}_{\text{ShallowEmb}}$	$82.30_{\pm 0.19}$	$70.55_{\pm 0.32}$	$78.94_{\pm 0.27}$	$79.86_{\pm 0.19}$	$63.50_{\pm 0.11}$	$61.44_{\pm 0.38}$	$68.79_{\pm 0.46}$	$69.25_{\pm 0.81}$	$71.44_{\pm 1.19}$	71.79
	$\text{SAGE}_{\text{ShallowEmb}}$	$82.27_{\pm 0.37}$	$69.56_{\pm 0.43}$	$77.88_{\pm 0.44}$	$79.67_{\pm 0.25}$	$63.57_{\pm 0.10}$	$56.65_{\pm 0.33}$	$72.01_{\pm 0.33}$	$78.50_{\pm 0.15}$	$81.43_{\pm 0.27}$	73.50
	$\text{GAT}_{\text{ShallowEmb}}$	$81.30_{\pm 0.67}$	$69.94_{\pm 0.74}$	$78.49_{\pm 0.70}$	$79.99_{\pm 0.65}$	$63.56_{\pm 0.04}$	$60.60_{\pm 1.17}$	$74.35_{\pm 0.35}$	$80.40_{\pm 0.45}$	$83.39_{\pm 0.22}$	74.67
	SenBERT-66M	$66.66_{\pm 1.42}$	$60.52_{\pm 1.62}$	$36.04_{\pm 2.92}$	$77.77_{\pm 0.75}$	$59.00_{\pm 1.17}$	$56.05_{\pm 0.41}$	$83.68_{\pm 0.19}$	$73.89_{\pm 0.31}$	$70.76_{\pm 0.15}$	64.93
	RoBERTa-355M	$72.24_{\pm 1.14}$	$66.68_{\pm 2.03}$	$42.32_{\pm 1.56}$	$76.81_{\pm 1.04}$	$63.52_{\pm 0.44}$	$59.27_{\pm 0.34}$	$84.62_{\pm 0.16}$	$74.79_{\pm 1.13}$	$72.31_{\pm 0.37}$	68.06
	GLEM	$81.30_{\pm 0.88}$	$68.80_{\pm 2.46}$	**$81.70_{\pm 1.07}$**	$76.43_{\pm 0.55}$	$60.25_{\pm 3.66}$	$55.13_{\pm 1.41}$	$83.28_{\pm 0.39}$	$76.93_{\pm 0.49}$	$80.46_{\pm 1.45}$	73.81
Encoder	$\text{GCN}_{\text{LLMEmb}}$	$83.33_{\pm 0.75}$	$71.39_{\pm 0.90}$	$78.71_{\pm 0.45}$	$80.94_{\pm 0.16}$	**$67.49_{\pm 0.43}$**	$68.65_{\pm 0.75}$	$83.03_{\pm 0.34}$	$84.84_{\pm 0.47}$	$88.22_{\pm 0.16}$	**78.51**
	ENGINE	**$84.22_{\pm 0.46}$**	**$72.14_{\pm 0.74}$**	$77.84_{\pm 0.27}$	$80.94_{\pm 0.19}$	$67.14_{\pm 0.46}$	$69.67_{\pm 0.16}$	$82.89_{\pm 0.14}$	$84.33_{\pm 0.57}$	$86.42_{\pm 0.23}$	78.40
Explainer	TAPE	$84.04_{\pm 0.24}$	$71.87_{\pm 0.35}$	$78.61_{\pm 1.23}$	**$81.94_{\pm 0.16}$**	$66.07_{\pm 0.10}$	$62.43_{\pm 0.47}$	**$84.92_{\pm 0.26}$**	**$86.46_{\pm 0.12}$**	**$89.52_{\pm 0.04}$**	78.43
Predictor	LLM_{IT}	$67.00_{\pm 0.16}$	$54.26_{\pm 0.22}$	**$80.99_{\pm 0.43}$**	$75.02_{\pm 0.16}$	$41.83_{\pm 0.47}$	$54.09_{\pm 1.02}$	$80.92_{\pm 1.38}$	$71.28_{\pm 1.81}$	$66.99_{\pm 2.02}$	65.76
	GraphGPT	$64.72_{\pm 1.50}$	$64.58_{\pm 1.55}$	$70.34_{\pm 2.27}$	$70.71_{\pm 0.37}$	$62.88_{\pm 2.14}$	$58.25_{\pm 0.37}$	$81.13_{\pm 1.52}$	$77.48_{\pm 0.78}$	$80.10_{\pm 0.76}$	70.02
	LLaGA	$78.94_{\pm 1.14}$	$62.61_{\pm 3.63}$	$65.91_{\pm 2.09}$	$76.47_{\pm 2.20}$	$65.84_{\pm 0.72}$	**$70.10_{\pm 0.38}$**	$83.47_{\pm 0.45}$	$84.44_{\pm 0.90}$	$87.82_{\pm 0.53}$	75.07

Supervised		Cora	Citeseer	Pubmed	arXiv	WikiCS	Instagram	Reddit	Books	Photo	Computer	Avg.
Classic	$\text{GCN}_{\text{ShallowEmb}}$	$87.41_{\pm 2.08}$	$75.74_{\pm 1.20}$	$89.01_{\pm 0.59}$	$71.39_{\pm 0.28}$	$83.67_{\pm 0.63}$	$63.94_{\pm 0.61}$	$65.07_{\pm 0.38}$	$76.94_{\pm 0.26}$	$73.34_{\pm 1.34}$	$77.16_{\pm 3.80}$	76.37
	$\text{SAGE}_{\text{ShallowEmb}}$	$87.44_{\pm 1.74}$	$74.96_{\pm 1.20}$	$90.47_{\pm 0.25}$	$71.21_{\pm 0.18}$	$84.86_{\pm 0.91}$	$64.14_{\pm 0.47}$	$61.52_{\pm 0.60}$	$79.40_{\pm 0.45}$	$84.59_{\pm 0.32}$	$87.77_{\pm 0.34}$	78.64
	$\text{GAT}_{\text{ShallowEmb}}$	$86.68_{\pm 1.12}$	$73.73_{\pm 0.94}$	$88.25_{\pm 0.47}$	$71.57_{\pm 0.25}$	$83.94_{\pm 0.61}$	$64.93_{\pm 0.75}$	$64.16_{\pm 1.05}$	$80.61_{\pm 0.49}$	$84.84_{\pm 0.69}$	$88.32_{\pm 0.24}$	78.70
	SenBERT-66M	$79.61_{\pm 1.40}$	$74.06_{\pm 1.26}$	$94.47_{\pm 0.33}$	$72.66_{\pm 0.24}$	$86.51_{\pm 0.86}$	$60.11_{\pm 0.93}$	$58.70_{\pm 0.54}$	$85.99_{\pm 0.58}$	$77.72_{\pm 0.35}$	$74.22_{\pm 0.21}$	76.40
	RoBERTa-355M	$83.17_{\pm 0.84}$	$75.90_{\pm 1.69}$	**$94.84_{\pm 0.06}$**	$74.12_{\pm 0.12}$	**$87.47_{\pm 0.83}$**	$63.75_{\pm 1.13}$	$60.61_{\pm 0.24}$	**$86.65_{\pm 0.38}$**	$79.45_{\pm 0.37}$	$75.76_{\pm 0.30}$	78.17
	GLEM	$86.81_{\pm 1.19}$	$73.24_{\pm 1.55}$	$93.98_{\pm 0.32}$	$73.55_{\pm 0.22}$	$79.81_{\pm 0.45}$	$67.39_{\pm 1.73}$	$53.11_{\pm 2.96}$	$83.98_{\pm 0.97}$	$78.16_{\pm 0.45}$	$81.63_{\pm 0.46}$	77.17
Encoder	$\text{GCN}_{\text{LLMEmb}}$	**$88.15_{\pm 1.79}$**	$76.45_{\pm 1.19}$	$88.38_{\pm 0.68}$	$74.39_{\pm 0.31}$	$84.78_{\pm 0.86}$	$68.27_{\pm 0.45}$	$70.65_{\pm 0.75}$	$84.23_{\pm 0.20}$	$86.07_{\pm 0.20}$	$89.52_{\pm 0.31}$	81.09
	ENGINE	$87.00_{\pm 1.60}$	$75.82_{\pm 1.52}$	$90.08_{\pm 0.16}$	$74.69_{\pm 0.36}$	$85.44_{\pm 0.53}$	$68.87_{\pm 0.25}$	**$71.21_{\pm 0.77}$**	$84.09_{\pm 0.09}$	$86.98_{\pm 0.06}$	$89.05_{\pm 0.13}$	81.32
Explainer	TAPE	$88.05_{\pm 1.76}$	$76.45_{\pm 1.60}$	$93.00_{\pm 0.13}$	$74.96_{\pm 0.14}$	$87.11_{\pm 0.66}$	$68.11_{\pm 0.54}$	$66.22_{\pm 0.83}$	$85.95_{\pm 0.59}$	**$87.72_{\pm 0.28}$**	**$90.46_{\pm 0.18}$**	**81.80**
Predictor	LLM_{IT}	$71.93_{\pm 1.47}$	$60.97_{\pm 3.97}$	$94.16_{\pm 0.19}$	**76.08**	$80.61_{\pm 0.47}$	$44.20_{\pm 3.06}$	$58.30_{\pm 0.48}$	$84.80_{\pm 0.13}$	$78.27_{\pm 0.54}$	$74.51_{\pm 0.53}$	72.38
	GraphGPT	$82.29_{\pm 0.26}$	$74.67_{\pm 1.15}$	$93.54_{\pm 0.22}$	$75.15_{\pm 0.14}$	$82.54_{\pm 0.23}$	$67.00_{\pm 1.22}$	$60.72_{\pm 1.47}$	$85.38_{\pm 0.72}$	$84.46_{\pm 0.36}$	$86.78_{\pm 1.14}$	79.25
	LLaGA	$87.55_{\pm 1.15}$	**$76.73_{\pm 1.70}$**	$90.28_{\pm 0.91}$	$74.49_{\pm 0.23}$	$84.03_{\pm 1.10}$	**$69.16_{\pm 0.72}$**	$71.06_{\pm 0.38}$	$85.56_{\pm 0.30}$	$87.62_{\pm 0.30}$	$90.41_{\pm 0.12}$	81.69

Encoder
A robust choice

Explainer
Suitable for graphs with labels heavily depend on text

Predictor
Require rich supervision

Experiments & Findings

- Zero-shot

Type & LLM	Method	Cora (82.52)		WikiCS (68.67)		Instagram (63.35)		Photo (78.50)		Avg.	
		Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1
LLM GPT-4o	Direct	68.08	69.25	68.59	63.21	44.53	42.77	63.99	61.09	61.30	59.08
	CoT	68.89	69.86	70.75	**66.23**	**47.87**	**47.57**	61.61	60.62	62.28	61.07
	ToT	68.29	69.13	70.78	65.69	44.16	42.68	60.84	59.16	61.02	59.16
	ReAct	68.21	69.28	69.45	66.03	44.49	43.16	63.63	60.82	61.44	59.82
	w. Neighbor	70.30	71.44	69.69	64.51	42.42	39.79	69.93	68.55	63.09	61.07
	w. Summary	**71.40**	**72.13**	**70.90**	65.42	45.02	44.62	**72.63**	**70.84**	**64.99**	**63.25**
LLM LLaMA-8B	Direct	62.64	63.02	56.77	53.04	37.58	29.70	41.23	44.26	49.56	47.50
	CoT	62.04	62.61	58.88	56.00	42.00	39.06	44.22	47.13	51.78	51.20
	ToT	34.06	33.30	40.35	41.15	**45.33**	**45.27**	31.31	34.00	37.76	38.43
	ReAct	36.55	38.04	22.40	25.76	44.67	44.42	27.03	28.96	32.66	34.30
	w. Neighbor	64.55	64.41	59.43	54.16	36.98	28.32	45.49	50.44	51.61	49.33
	w. Summary	**64.69**	**64.62**	**62.69**	**56.40**	37.59	30.91	**48.11**	**52.20**	**53.27**	**51.03**
GFM	ZeroG	**62.55**	**57.56**	**62.71**	**57.87**	**50.71**	**50.43**	46.27	**51.52**	**55.56**	**54.35**
	LLM_{IT}	52.58	51.89	60.83	53.59	41.58	26.26	**49.23**	44.85	51.06	44.15
	LLaGA	18.82	8.49	8.20	8.29	47.93	47.70	39.18	4.71	28.53	17.30

GFMs can outperform open-source LLMs but still fall short of strong LLMs like GPT-4o

Experiments & Findings

- LLM-as-Encoder vs. LM-as-Encoder

Method	Encoder	Semi-supervised				Supervised			
		Cornell	Texas	Wisconsin	Washington	Cornell	Texas	Wisconsin	Washington
Homophily Ratio (%)		11.55	6.69	16.27	17.07	11.55	6.69	16.27	17.07
MLP	SenBERT	50.59$_{\pm 3.14}$	56.67$_{\pm 2.15}$	71.98$_{\pm 1.59}$	63.26$_{\pm 2.89}$	66.15$_{\pm 1.92}$	76.32$_{\pm 3.72}$	81.51$_{\pm 7.00}$	70.44$_{\pm 8.65}$
	RoBERTa	59.08$_{\pm 2.57}$	67.47$_{\pm 1.29}$	73.87$_{\pm 1.62}$	65.43$_{\pm 3.44}$	66.67$_{\pm 8.88}$	74.21$_{\pm 6.09}$	80.00$_{\pm 9.88}$	76.96$_{\pm 7.48}$
	Qwen-3B	57.78$_{\pm 3.24}$	76.27$_{\pm 1.61}$	82.36$_{\pm 1.62}$	**75.11$_{\pm 1.92}$**	77.95$_{\pm 4.76}$	88.95$_{\pm 3.07}$	88.68$_{\pm 6.64}$	83.48$_{\pm 1.74}$
	Mistral-7B	**59.87$_{\pm 6.72}$**	**76.27$_{\pm 1.08}$**	**83.30$_{\pm 1.42}$**	74.24$_{\pm 0.88}$	**78.46$_{\pm 4.17}$**	**90.53$_{\pm 3.16}$**	**89.43$_{\pm 5.15}$**	**83.91$_{\pm 5.60}$**
GCN	SenBERT	46.80$_{\pm 2.13}$	54.93$_{\pm 0.68}$	58.30$_{\pm 2.56}$	52.61$_{\pm 1.35}$	50.77$_{\pm 10.18}$	59.47$_{\pm 5.16}$	61.13$_{\pm 8.65}$	61.30$_{\pm 1.62}$
	RoBERTa	47.06$_{\pm 2.19}$	55.20$_{\pm 2.78}$	54.91$_{\pm 3.40}$	54.89$_{\pm 1.50}$	51.79$_{\pm 7.68}$	58.42$_{\pm 7.33}$	59.24$_{\pm 8.82}$	61.31$_{\pm 5.39}$
	Qwen-3B	53.59$_{\pm 2.07}$	56.80$_{\pm 4.29}$	**63.02$_{\pm 2.16}$**	**64.56$_{\pm 4.06}$**	58.46$_{\pm 10.56}$	64.74$_{\pm 7.37}$	**65.28$_{\pm 6.82}$**	**67.83$_{\pm 3.74}$**
	Mistral-7B	**54.64$_{\pm 1.52}$**	**58.67$_{\pm 3.60}$**	62.08$_{\pm 2.61}$	61.52$_{\pm 3.61}$	**59.49$_{\pm 6.96}$**	**65.79$_{\pm 6.66}$**	64.90$_{\pm 5.67}$	66.96$_{\pm 4.84}$
SAGE	SenBERT	52.55$_{\pm 1.58}$	61.73$_{\pm 1.37}$	70.47$_{\pm 1.75}$	65.54$_{\pm 2.44}$	68.72$_{\pm 4.97}$	80.00$_{\pm 5.91}$	83.02$_{\pm 6.31}$	76.96$_{\pm 4.88}$
	RoBERTa	55.55$_{\pm 3.44}$	64.26$_{\pm 6.26}$	73.59$_{\pm 2.72}$	66.08$_{\pm 1.60}$	70.26$_{\pm 8.37}$	80.53$_{\pm 2.68}$	81.89$_{\pm 7.42}$	74.35$_{\pm 7.95}$
	Qwen-3B	**57.13$_{\pm 2.29}$**	**78.53$_{\pm 1.76}$**	83.21$_{\pm 1.39}$	72.18$_{\pm 3.66}$	74.87$_{\pm 2.99}$	89.47$_{\pm 1.67}$	**91.32$_{\pm 2.82}$**	**83.48$_{\pm 3.25}$**
	Mistral-7B	56.86$_{\pm 1.37}$	76.53$_{\pm 2.40}$	**83.96$_{\pm 1.55}$**	**73.91$_{\pm 0.97}$**	**77.44$_{\pm 2.99}$**	**91.05$_{\pm 2.69}$**	89.44$_{\pm 4.24}$	81.74$_{\pm 4.48}$
H_2GCN	SenBERT	56.34$_{\pm 1.67}$	66.67$_{\pm 2.95}$	73.40$_{\pm 1.68}$	70.55$_{\pm 4.95}$	73.85$_{\pm 7.14}$	84.21$_{\pm 4.40}$	86.42$_{\pm 6.01}$	77.83$_{\pm 7.20}$
	RoBERTa	60.00$_{\pm 3.54}$	68.13$_{\pm 2.93}$	75.66$_{\pm 2.12}$	71.52$_{\pm 1.22}$	74.87$_{\pm 7.68}$	83.16$_{\pm 6.14}$	84.53$_{\pm 9.04}$	79.13$_{\pm 5.43}$
	Qwen-3B	**61.57$_{\pm 3.89}$**	**80.13$_{\pm 6.45}$**	**84.53$_{\pm 0.70}$**	**74.67$_{\pm 1.77}$**	**76.41$_{\pm 2.99}$**	**92.11$_{\pm 2.88}$**	**89.81$_{\pm 3.29}$**	85.22$_{\pm 3.99}$
	Mistral-7B	59.22$_{\pm 4.54}$	72.93$_{\pm 8.21}$	81.89$_{\pm 1.51}$	68.59$_{\pm 4.46}$	75.89$_{\pm 3.84}$	89.47$_{\pm 3.72}$	89.43$_{\pm 5.42}$	**86.09$_{\pm 3.25}$**

LLM-as-Encoder significantly outperforms LMs in less informative graphs, e.g., ***heterophilic ones***

Contribution Summary

- A Testbed
  - LLMNodeBed, a PyG-based testbed including 14 datasets, 8 LLM-based algorithms, 8 classic algorithms, and 3 learning configurations
- Comprehensive Experiments
  - Training and evaluating over 2,700 models, we analyze how learning paradigm, homophily, language model type and size, and prompt design impact the performace
- Insights and Tips
  - Our work provides intuitive explanations, practical tips, and insights about the strengths and limitations of each algorithm category.

Resources

- Paper: https://arxiv.org/pdf/2502.00829
- Code: https://github.com/WxxShirley/LLMNodeBed
- Dataset: https://huggingface.co/datasets/xxwu/LLMNodeBed
- Chinese Blog: https://zhuanlan.zhihu.com/p/19135360567179679 76

Thank you for your attention!