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

xxwu@se.cuhk.edu.hk https://wxxshirley.github.io/

香港中文大學 The Chinese University of Hong Kong

GitHub 知乎

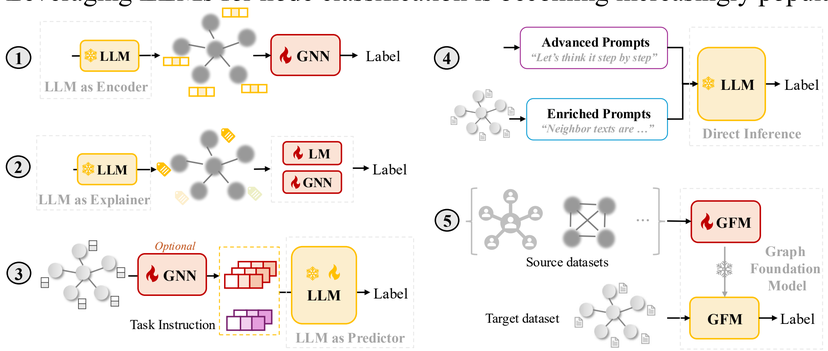
ICML International Conference On Machine Learning

Benchmark: LLMNodeBed

Node Classification is a fundamental task in graph analysis

- Fraud Detection
- User Profiling
- Item Tagging

Leveraging **LLMs** for node classification is becoming increasingly popular



(a) Semi- / Supervised Settings **(b) Zero-shot Setting**

Illustrations of LLM-based node classification algorithms in supervised and zero-shot settings

? Design principles for LLM-based methods remain elusive

LLM vs. GNN

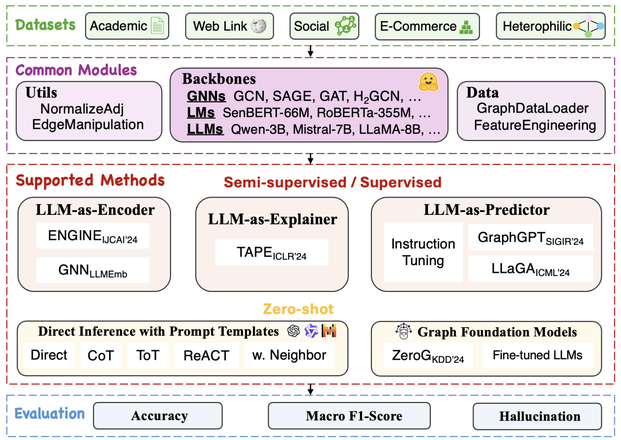
Is my graph suited for an LLM, or will a GNN suffice?

LLM vs. LM

In which scenarios can LLMs outperform LMs?

Practical Tips

What is the best recipe for each category?



A PyG-based testbed with **14 datasets**, **8 LLM-based algorithms**, **8 classic methods**, and **3 learning configurations**

Controlled Experiments: Findings & Tips

Semi-supervised & Supervised

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

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

Zero-shot

Type & LLM	Method	Cora (82.52) Acc	Cora (82.52) Macro-F1	WikiCS (68.67) Acc	WikiCS (68.67) Macro-F1	Instagram (63.35) Acc	Instagram (63.35) Macro-F1	Photo (78.50) Acc	Photo (78.50) Macro-F1	Avg. Acc	Avg. 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

Takeaways (6) **GFMs can outperform open-source LLMs** but still fall short of strong LLMs like GPT-4o. (7) **LLM direct inference** can be improved by appropriately **incorporating structural information**

LLM vs. LM as Encoder

Method	Encoder	Cornell	Texas	Wisconsin	Washington
Homophily Ratio (%)		11.55	6.69	16.27	17.07
MLP	SenBERT	50.59±3.14	56.67±2.15	71.98±1.59	63.26±2.89
	RoBERTa	59.08±2.57	67.47±1.29	73.87±1.62	65.43±3.44
	Qwen-3B	57.78±3.24	76.27±1.61	82.36±1.62	**75.11±1.92**
	Mistral-7B	**59.87±6.72**	**76.27±1.08**	**83.30±1.42**	74.24±0.88
GCN	SenBERT	46.80±2.13	54.93±0.68	58.30±2.56	52.61±1.35
	RoBERTa	47.06±2.19	55.20±2.78	54.91±3.40	54.89±1.50
	Qwen-3B	53.59±2.07	56.80±4.29	**63.02±2.16**	**64.56±4.06**
	Mistral-7B	**54.64±1.52**	**58.67±3.60**	62.08±2.61	61.52±3.61
SAGE	SenBERT	52.55±1.58	61.73±1.37	70.47±1.75	65.54±2.44
	RoBERTa	55.55±3.44	64.26±6.26	73.59±2.72	66.08±1.60
	Qwen-3B	**57.13±2.29**	**78.53±1.76**	83.21±1.39	72.18±3.66
	Mistral-7B	56.86±1.37	76.53±2.40	**83.96±1.55**	**73.91±0.97**
H_2GCN	SenBERT	56.34±1.67	66.67±2.95	73.40±1.68	70.55±4.95
	RoBERTa	60.00±3.54	68.13±2.93	75.66±2.12	71.52±1.22
	Qwen-3B	**61.57±3.89**	**80.13±6.45**	**84.53±0.70**	**74.67±1.77**
	Mistral-7B	59.22±4.54	72.93±8.21	81.89±1.51	68.59±4.46

Takeaways (8) LLM-as-Encoder significantly outperforms LMs when graph structure is less informative about the labels, e.g., **heterophilic ones**

Takeaways

(1) Appropriately incorporating LLMs **consistently** improves the performance

(2) LLM-based methods provide greater improvements **in semi-supervised settings** than in supervised settings.

(3) **LLM-as-Explainer** methods are highly effective when **labels heavily depend on text**.

(4) LLM-as-Encoder methods balance computational cost and accuracy effectively.

(5) **LLM-as-Predictor** methods are more effective when labeled data is **abundant**.

More experiments and findings can be found in our paper!

Project Page