GLBench: A Comprehensive Benchmark for Graph with Large Language Models

Yuhan Li, Peisong Wang, Xiao Zhu, Aochuan Chen, Haiyun Jiang, Deng Cai, Victor Wai Kin Chan, Jia Li

HKUST(GZ), HKUST, THU, Tencent AI Lab

香港科技大學(廣州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

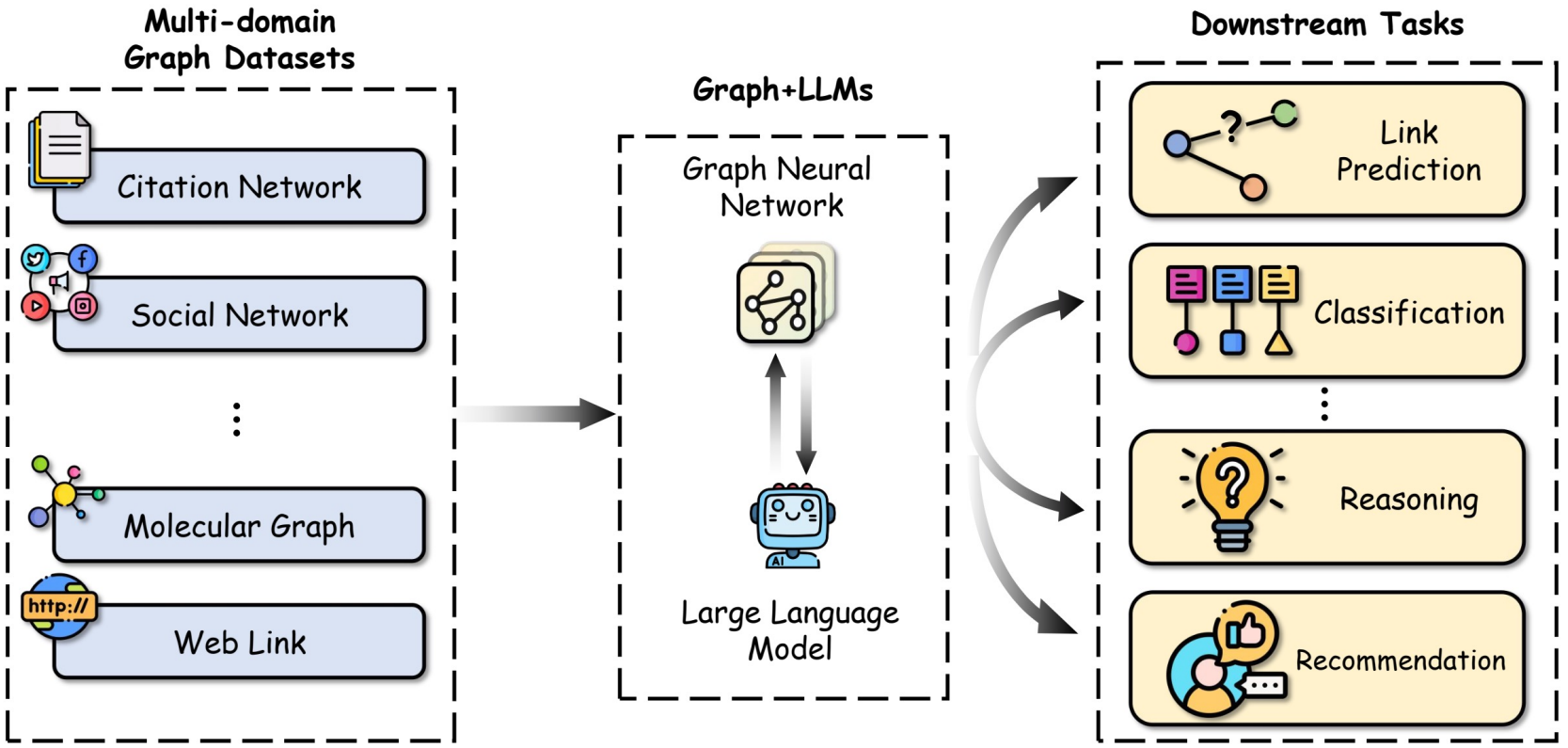
香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

清華大學
Tsinghua University

Tencent AI Lab

GraphLLM

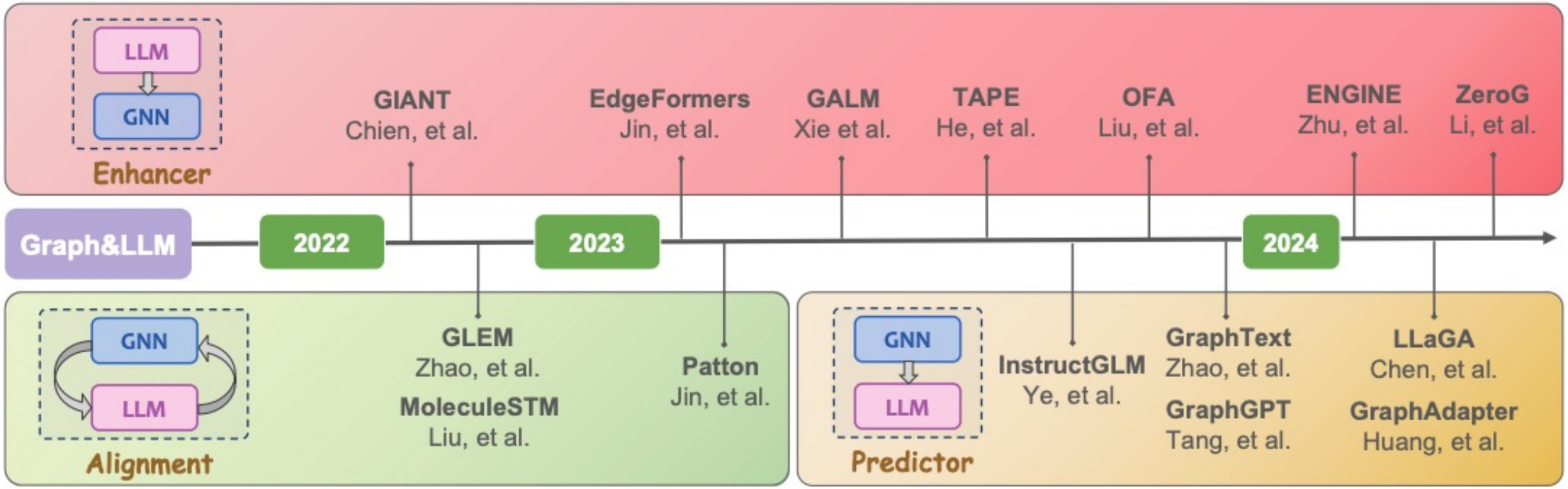
- The integration of GNNs and LLMs (GraphLLM) across a myriad of domains



Li Y, Li Z, Wang P, et al. A Survey of Graph Meets Large Language Model: Progress and Future Directions. IJCAI 2024.

Timeline of GraphLLM

- Existing methods can be divided into three categories based on the role palyed by LLMs.



Benchmarking GraphLLM

Role	Method	Predictor	GNN	PLM/LLM	Techniques Used		Learning Scenarios		Venue	Code
					Fine-tune	Prompt	Supervised	Zero-shot		
Enhancer	GIANT [9]	GNN	GraphSAGE, etc.	BERT	✗	✗	✓	✗	ICLR'22	Link
	TAPE [13]	GNN	RevGAT	ChatGPT	✗	✓	✓	✗	ICLR'24	Link
	OFA [26]	GNN	R-GCN	Sentence-BERT	✗	✓	✓	✓	ICLR'24	Link
	ENGINE [54]	GNN	GraphSAGE	LLaMA-2	✓	✓	✓	✗	IJCAI'24	Link
	ZeroG [25]	GNN	SGC	Sentence-BERT	✓	✓	✗	✓	SIGKDD'24	Link
Predictor	InstructGLM [50]	LLM	-	FLAN-T5/LLaMA-v1	✓	✓	✓	✗	EACL'24	Link
	GraphText [53]	LLM	-	ChatGPT/GPT-4	✓	✓	✓	✗	Arxiv	Link
	GraphAdapter [17]	LLM	GraphSAGE	LLaMA-2	✓	✓	✓	✗	WWW'24	Link
	GraphGPT [40]	LLM	GT	Vicuna	✓	✓	✓	✓	SIGIR'24	Link
	LLaGA [6]	LLM	-	Vicuna/LLaMA-2	✓	✓	✓	✗	ICML'24	Link
Aligner	GLEM [52]	GNN/LLM	GraphSAGE, etc.	RoBERTa	✓	✗	✓	✗	ICLR'23	Link
	PATTON [21]	LLM	GT	BERT/SciBERT	✓	✗	✓	✗	ACL'23	Link

Motivation

- 1. The use of different datasets, data processing approaches, and data splitting strategies in previous GraphLLM works.
- 2. The lack of benchmarks for zero-shot graph learning has led to limited exploration in this area.
- 3. Each method's computation and memory costs often overlooked.

Benchmarking GraphLLM | GLBench

Comparison with existing benchmarks

Benchmark	#Datasets (Node-level)	#Domains	Text	#Models (GraphLLM)	Model Type	Supervision Scenario
Sen et al. [38]	2 (2)	1	✗	8 (0)	Classical	Supervised
Shchur et al. [39]	8 (8)	2	✗	8 (0)	GNN	Supervised
OGB [15]	14 (5)	3	✗	20 (0)	GNN	Supervised
CS-TAG [47]	8 (6)	2	✓	16 (2)	GNN, PLM, Enhancer	Supervised
`GLBench`	7 (7)	3	✓	18 (12)	GNN, PLM, GraphLLM	Supervised and Zero-shot

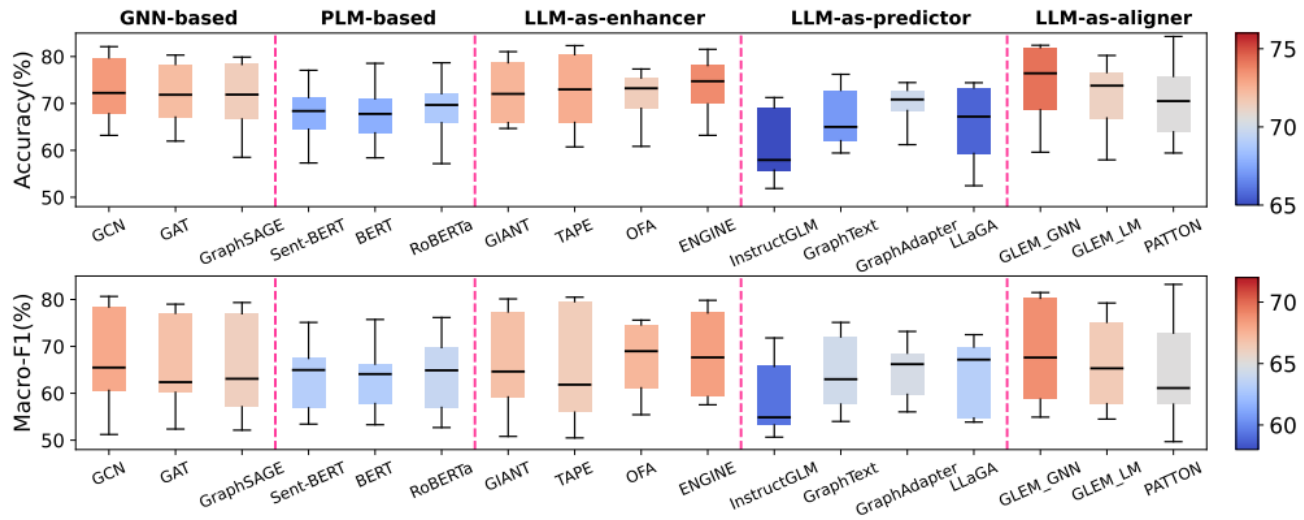
Datasets

Dataset	# Nodes	# Edges	Avg. # Deg	Avg. # Tok	# Classes	# Train	Node Text	Domain
Cora	2,708	5,429	4.01	186.53	7	5.17%	Paper content	Citation
Citeseer	3,186	4,277	2.68	213.16	6	3.77%	Paper content	Citation
Pubmed	19,717	44,338	4.50	468.56	3	0.30%	Paper content	Citation
Ogbn-arxiv	169,343	1,166,243	13.77	243.19	40	53.70%	Paper content	Citation
WikiCS	11,701	216,123	36.94	642.04	10	4.96%	Entity description	Web link
Reddit	33,434	198,448	11.87	203.84	2	10.00%	User's post	Social
Instagram	11,339	144,010	25.40	59.25	2	10.00%	User's profile	Social

Benchmarking GraphLLM | GLBench

- **Supervised Scenario**
  - Effectiveness
  - LLM-as-predictor
  - LLM-as-enhancer
  - LLM-as-aligner
  - Scaling law

Model	Cora		Citeseer		Pubmed		Ogbn-arxiv		WikiCS		Reddit		Instagram	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
GCN [23]	**82.11**	**80.65**	69.84	65.49	79.10	79.19	72.24	51.22	80.35	77.63	63.19	62.49	65.75	58.75
GAT [43]	80.31	79.00	68.78	62.37	76.93	76.75	71.85	52.38	79.73	77.40	61.97	61.78	65.38	58.60
GraphSAGE [10]	79.88	79.35	68.23	63.10	76.79	76.91	71.88	52.14	79.87	77.05	58.51	58.41	65.12	55.85
Sent-BERT (22M) [36]	69.73	67.59	68.39	64.97	65.93	67.33	72.82	53.43	77.07	75.11	57.31	57.09	63.07	56.68
BERT (110M) [22]	69.71	67.53	67.77	64.10	63.69	64.93	72.29	53.30	78.55	75.74	58.41	58.33	63.75	57.30
RoBERTa (355M) [30]	69.68	67.33	68.19	64.90	71.25	72.19	72.94	52.70	78.67	76.16	57.17	57.10	63.57	56.87
GIANT [9]	81.04	80.13	65.82	62.31	76.89	76.05	72.04	50.81	80.48	78.67	64.67	64.64	66.01	56.11
TAPE [13]	80.95	79.79	66.06	61.84	79.87	79.30	72.99	51.43	82.33	80.49	60.73	60.50	65.85	50.49
OFA [26]	75.24	74.20	**73.04**	**68.98**	75.61	75.60	73.23	57.38	77.34	74.97	64.86	64.95	60.85	55.44
ENGINE [54]	81.54	79.82	72.15	67.65	74.74	75.21	75.01	57.55	81.19	79.08	63.20	59.34	**67.62**	**59.22**
InstructGLM [50]	69.10	65.74	51.87	50.65	71.26	71.81	39.09	24.65	45.73	42.70	55.78	53.24	57.94	54.87
GraphText [53]	76.21	74.51	59.43	56.43	74.64	75.11	49.47	24.76	67.35	64.55	61.86	61.46	62.64	54.00
GraphAdapter [17]	72.85	70.66	69.57	66.21	72.75	73.19	74.45	56.04	70.85	66.49	61.21	61.13	67.40	58.40
LLaGA [6]	74.42	72.50	55.73	54.83	52.46	68.82	72.78	53.86	73.88	70.90	**67.19**	**67.18**	62.94	54.62
GLEM_{GNN} [52]	**82.11**	80.00	71.16	67.62	81.72	81.48	**76.43**	**58.07**	**82.40**	**80.54**	59.60	59.41	66.10	54.92
GLEM_{LLM} [52]	73.79	72.00	68.78	65.32	79.18	79.25	74.03	58.01	80.23	78.30	57.97	57.56	65.00	54.50
PATTON [21]	70.50	67.97	63.60	61.12	**84.28**	**83.22**	70.74	49.69	80.81	77.72	59.43	57.85	64.27	57.48



Benchmarking GraphLLM | GLBench

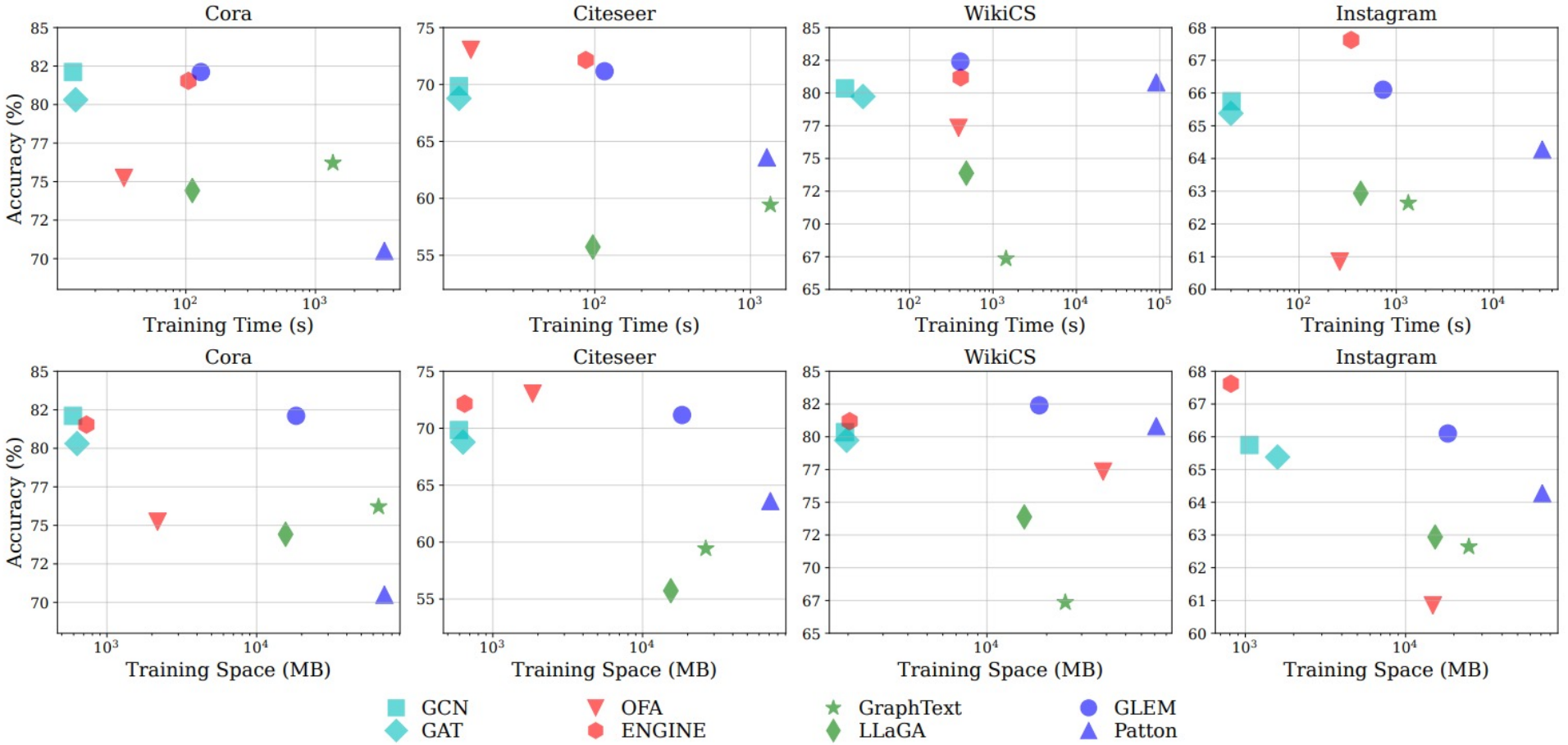
Zero-shot Scenario

- LLMs
- Semantics/Structures?
- Even a simple baseline can outperform existing GraphLLM methods.

Category	Model	$\mathcal{A}$	$\mathcal{S}$	Cora		Citeseer		Pubmed		WikiCS		Instagram	
				Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Graph SSL	DGI [44]	✓	✗	17.50	12.44	21.67	13.53	44.88	38.72	9.03	6.13	**63.64**	50.13
	GraphMAE [14]	✓	✗	27.08	23.66	15.24	14.44	22.03	15.65	10.74	6.69	53.56	52.18
LLMs	LLaMA3 (70B) [42]	✗	✓	67.99	68.05	51.44	49.98	77.00	64.18	**73.64**	**72.62**	38.23	36.41
	GPT-3.5-turbo [35]	✗	✓	65.67	63.22	50.58	49.34	75.99	69.90	68.75	66.56	49.39	49.67
	GPT-4o [1]	✗	✓	**68.62**	**68.49**	53.55	52.42	77.96	71.79	71.52	70.06	42.02	40.96
	DeepSeek-chat [3]	✗	✓	65.62	65.77	50.35	48.32	**79.23**	74.30	70.77	69.91	40.58	39.27
Training-free	**Emb w/ NA**	✓	✓	63.59	58.23	51.75	49.51	74.66	73.15	52.30	48.40	45.52	45.14
Enhancer	OFA [26]	✓	✓	23.11	23.30	32.45	28.67	46.60	35.04	34.27	33.72	53.63	51.10
	ZEROG [25]	✓	✓	62.52	57.53	**58.92**	**54.58**	79.08	**77.94**	60.46	57.24	56.13	**52.50**
Predictor	GraphGPT [40]	✓	✓	24.90	7.98	13.95	13.89	39.85	20.07	38.02	29.46	43.94	43.49

Benchmarking GraphLLM | GLBench

- **Efficiency**



Paper

Code

The End, Thanks!

GLBench: A Comprehensive Benchmark for Graph with Large Language Models

Yuhan Li, Peisong Wang, Xiao Zhu, Aochuan Chen, Haiyun Jiang, Deng Cai, Victor Wai Kin Chan, Jia Li