Accelerator for LLM-Enhanced GNN with Product Quantization and Unified Indexing

Jiaming Xu*[1,2], Jinhao Li*[1], Jun Liu[1], Hao Zhou[1], Guohao Dai[1,2]

Equal contribution*, Shanghai Jiao Tong University[1], Infinigence-AI[2],

Correspondence to:

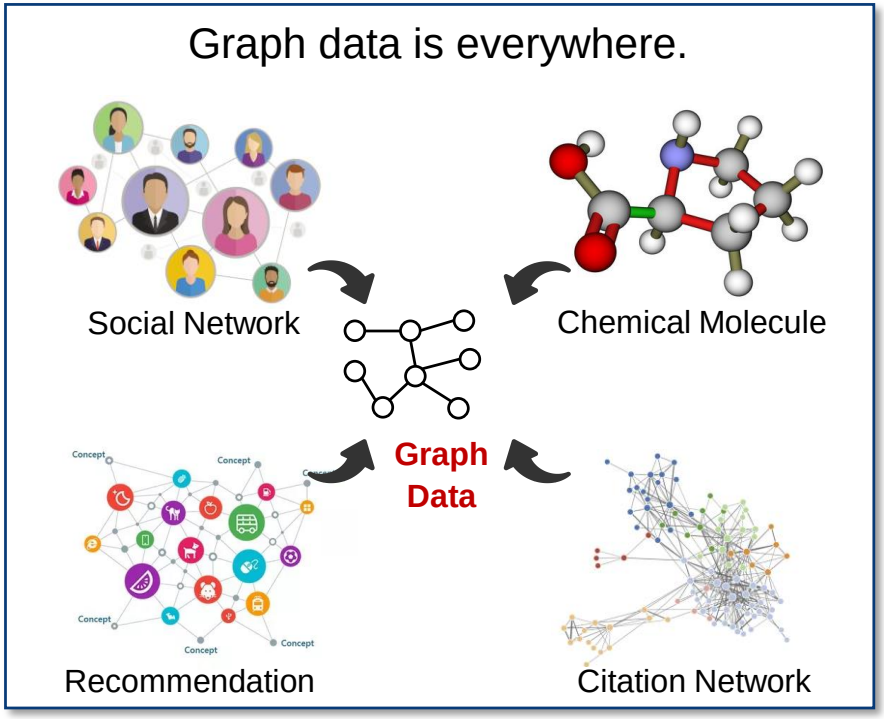
Guohao Dai <daiguohao@sjtu.edu.cn>

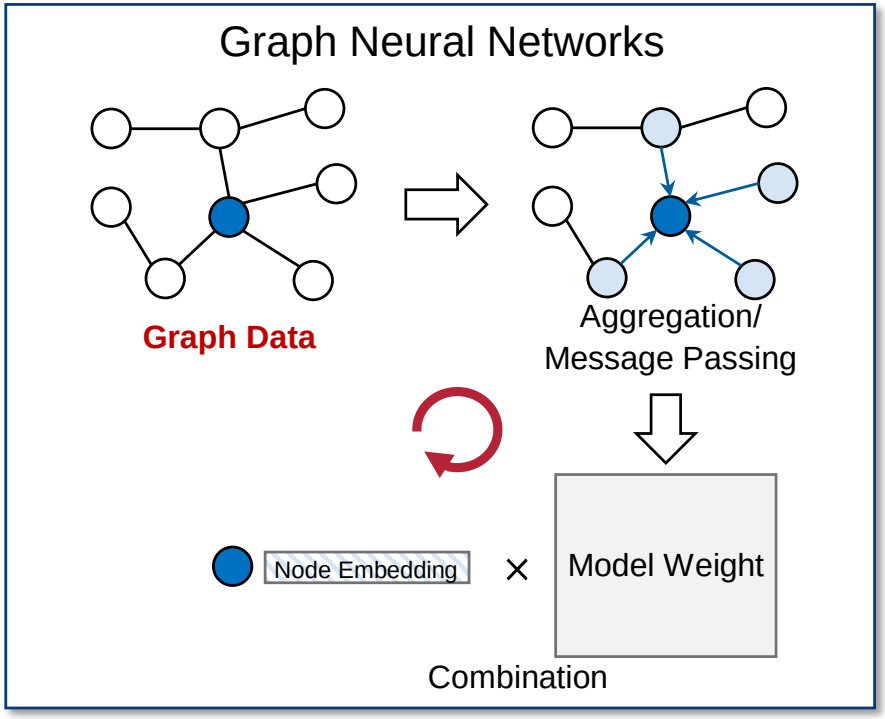
Outline

- Background and Challenge
- Techniques
  - Overview
  - PQ-based MatMul
  - Unified Architecture
  - Extensible GFM-ISA Design
- Experiment Results

Graph Neural Network for Graph Data

Graph neural networks (GNNs) are designed for graph data problem.

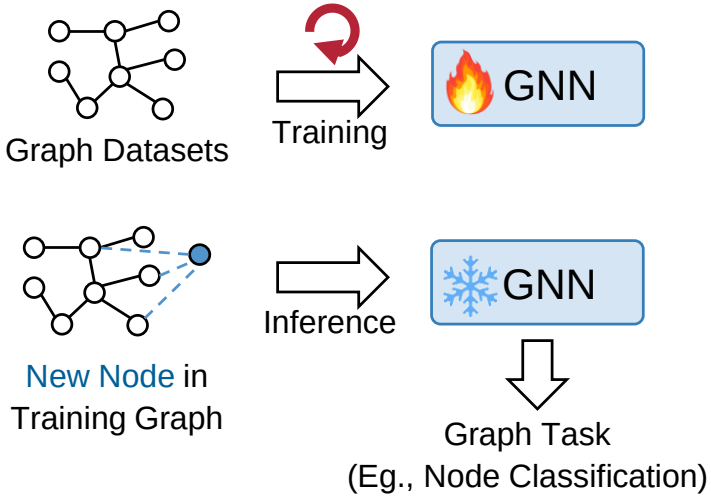




Problem of Traditional GNN

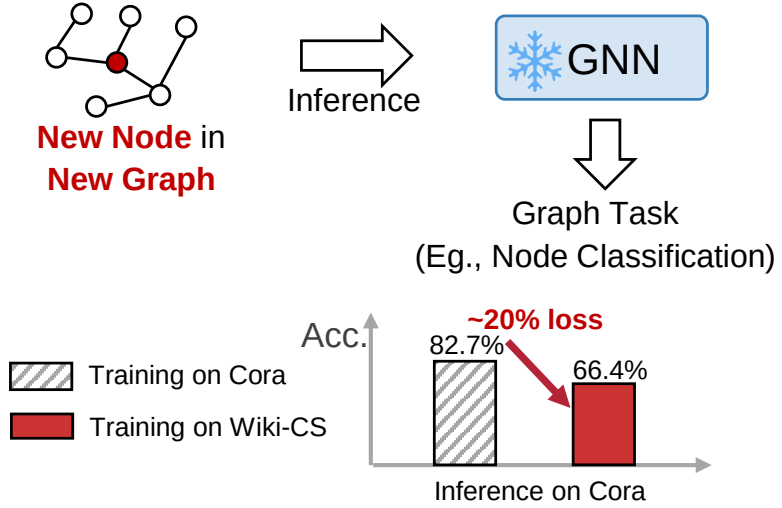
Vulnerability of traditional GNN on unseen graphs.

GNN Training and Inference



High Accuracy !

Poor Generalization on Unseen Graphs[1]



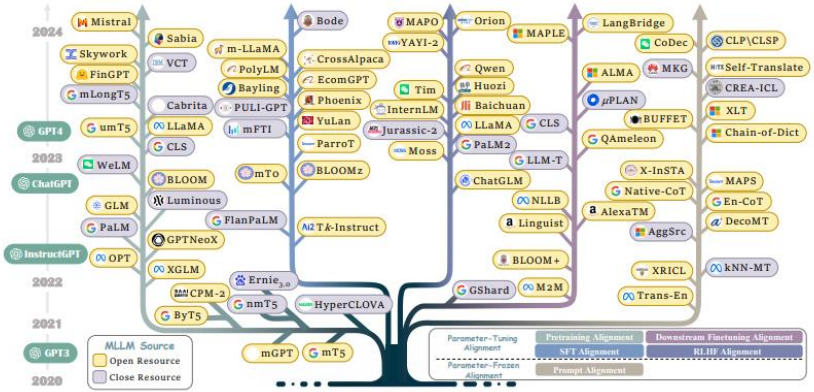
Poor Generalization !

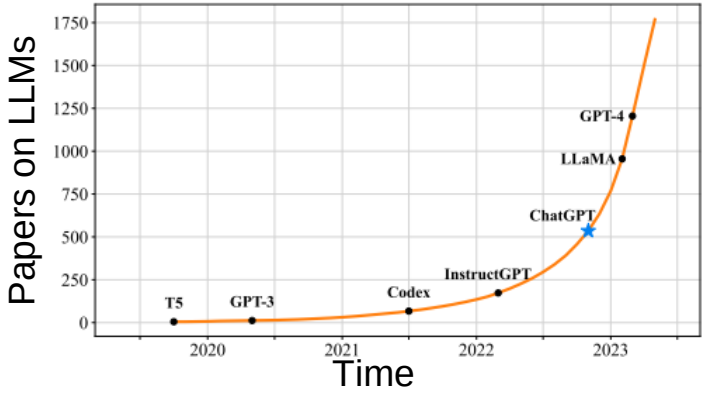
[1] Zheng X, Zhang M, Chen C, et al. Gnnevaluator: Evaluating gnn performance on unseen graphs without labels[J]. Advances in Neural Information Processing Systems, 2024, 36.

Era of Large Language Models (LLMs)

LLMs have attracted much attention.

Rapid Development of LLMs





One LLM per day on average[1,2]

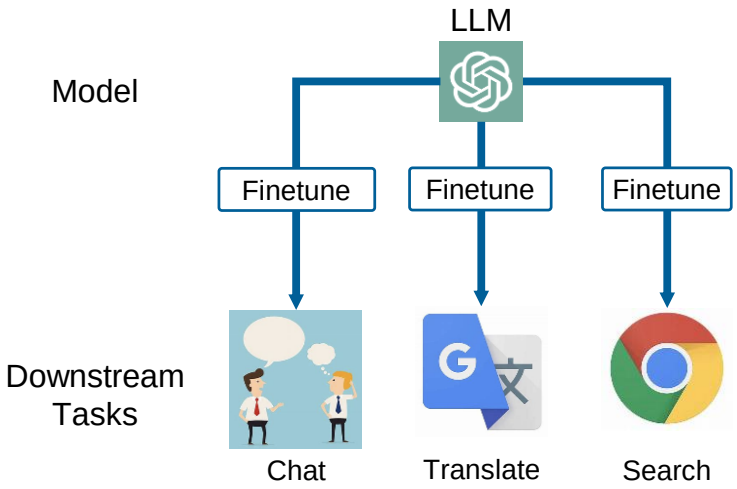
8.58 papers per day on average[3]

[1] Qin L, Chen Q, Zhou Y, et al. Multilingual large language model: A survey of resources, taxonomy and frontiers[J]. arXiv preprint arXiv:2404.04925, 2024.
[2] https://www.21jingji.com/article/20240516/herald/c7c069827f1f79b182604827b494511b.html
[3] Zhao W X, Zhou K, Li J, et al. A survey of large language models[J]. arXiv preprint arXiv:2303.18223, 2023.

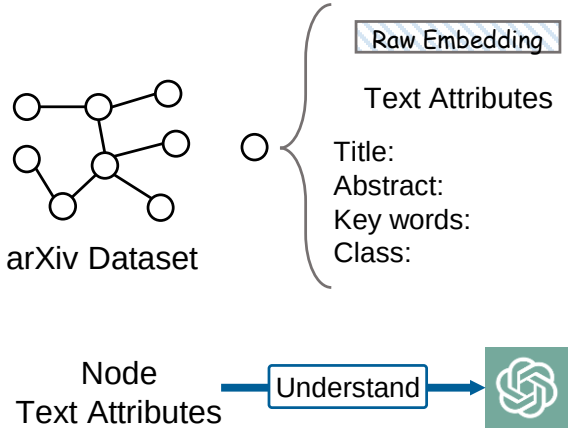
LLM Improves GNN

Utilize the strong generalization of LLM to improve GNN

Strong Generalization on Downstream Tasks



Utilization of LLM on Text-attributed graphs

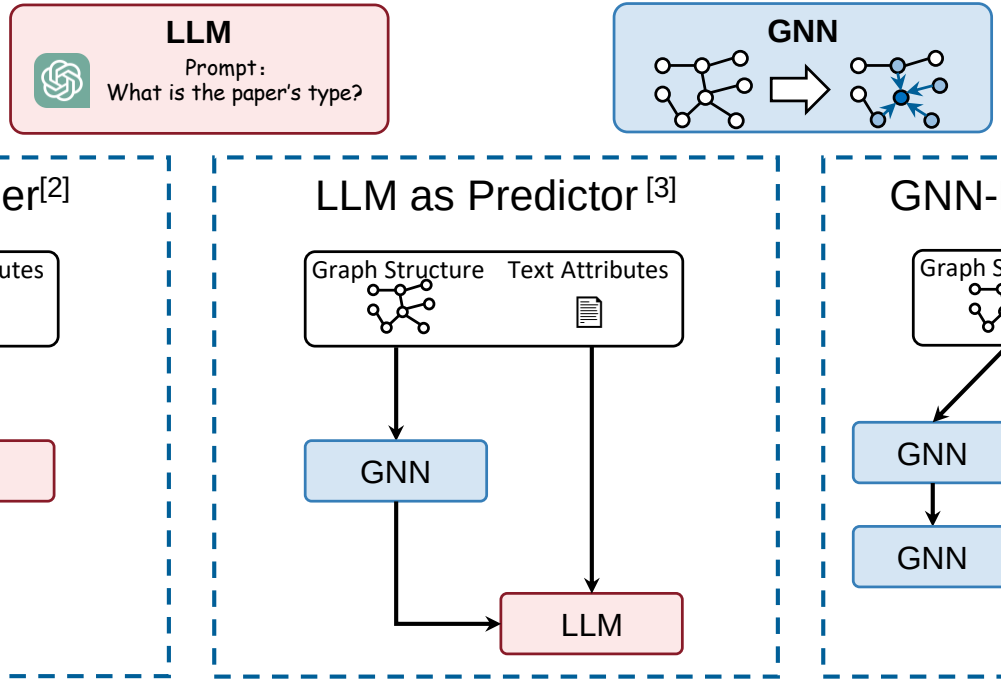


[1] Qin L, Chen Q, Zhou Y, et al. Multilingual large language model: A survey of resources, taxonomy and frontiers[J]. arXiv preprint arXiv:2404.04925, 2024.

Related Works

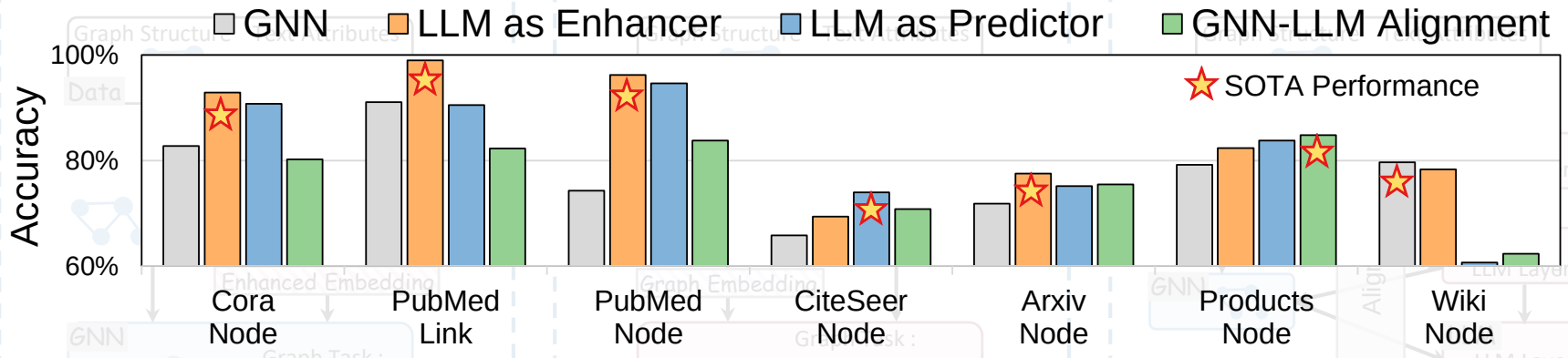
Combination Methods of LLM and GNN[1]

Graph Task :
Eg., Node Classification



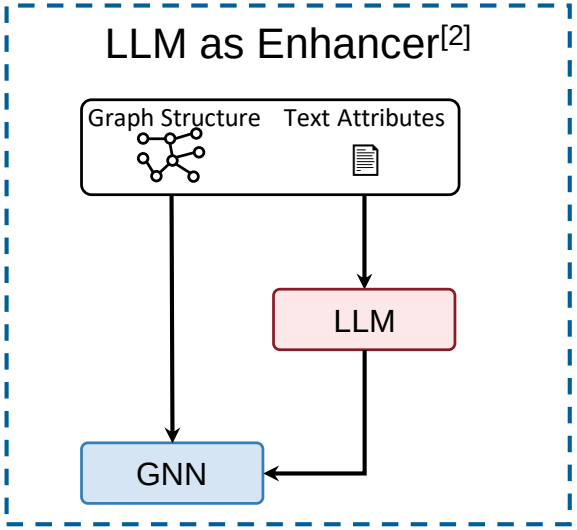
[1] Li Y, Li Z, Wang P, et al. A survey of graph meets large language model: Progress and future directions[J]. arXiv preprint arXiv:2311.12399, 2023.
[2] Liu H, Feng J, Kong L, et al. One for all: Towards training one graph model for all classification tasks[J]. arXiv preprint arXiv:2310.00149, 2023.
[3] Tang J, Yang Y, Wei W, et al. Graphgpt: Graph instruction tuning for large language models[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024: 491-500.
[4] Zhao J, Qu M, Li C, et al. Learning on large-scale text-attributed graphs via variational inference[J]. arXiv preprint arXiv:2210.14709, 2022.

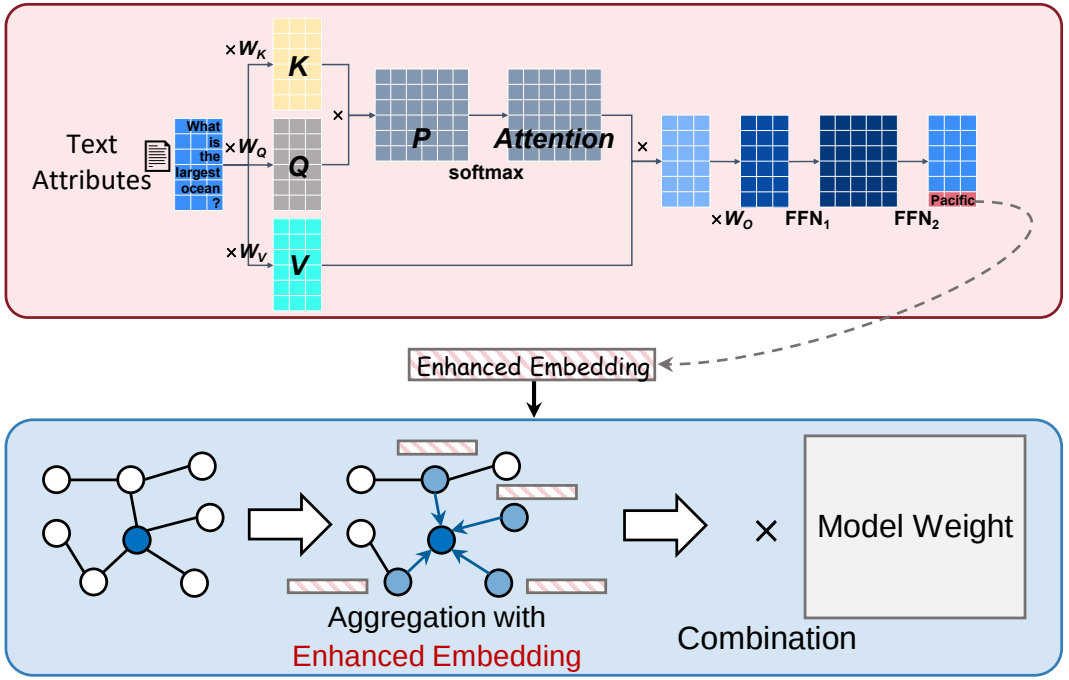
Related Works



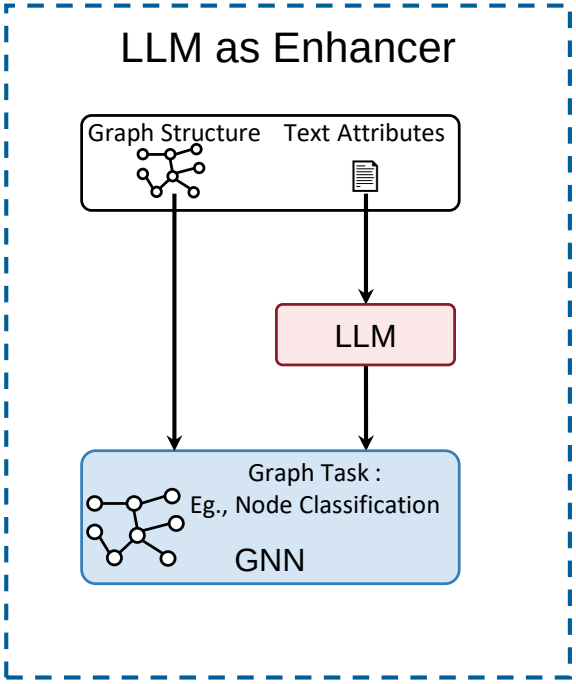
***LLM as Enhancer* achieves *SOTA performance* on most graph datasets.**

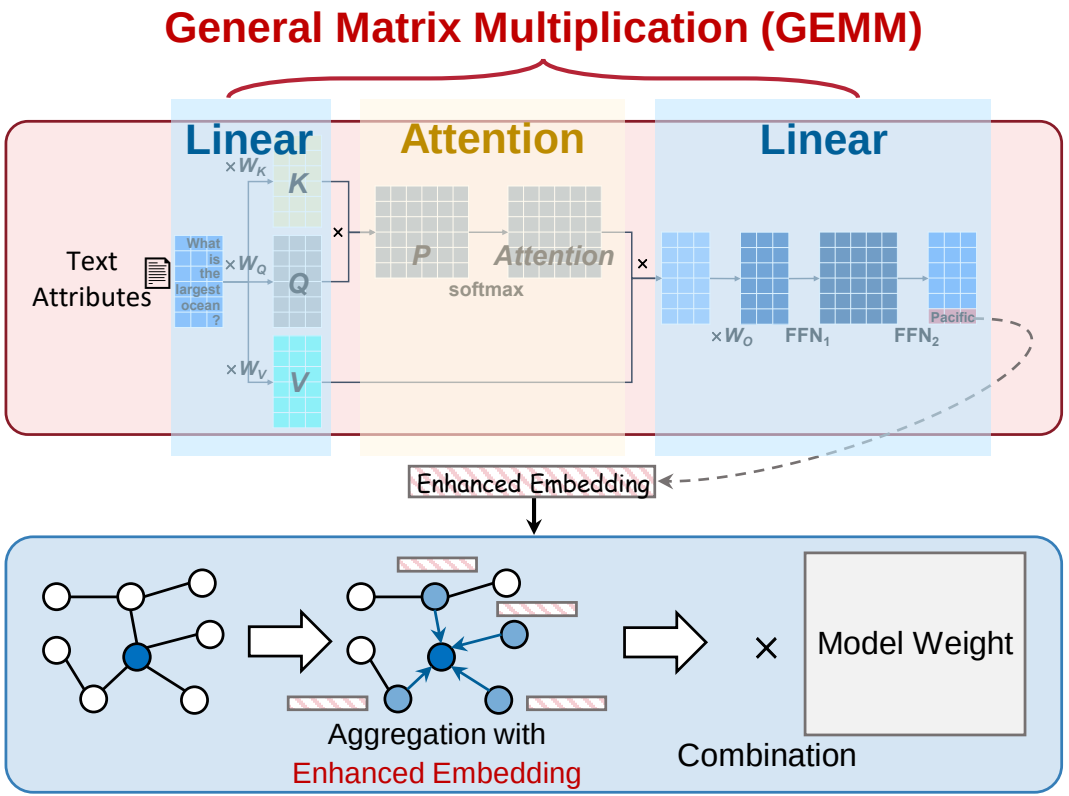
LLM-Enhanced GNN Dataflow



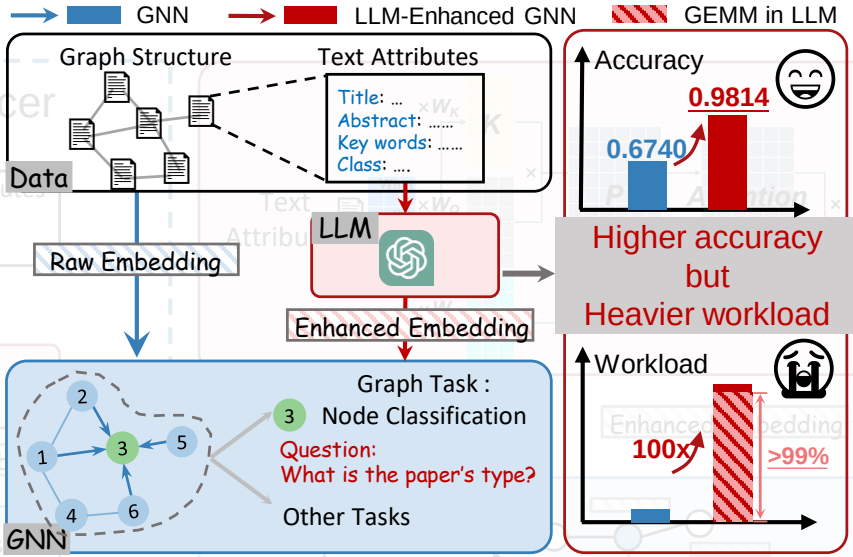


LLM-Enhanced GNN Dataflow





Challenge



LLMs bring ~100× more workload than GNNs with GEMMs accounting for more than 99%.

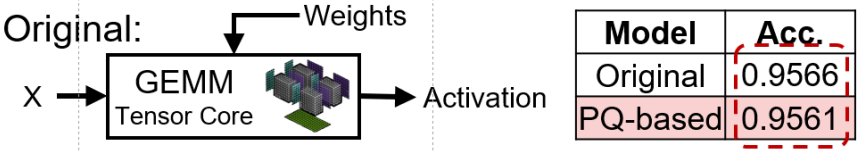
Outline

- Background and Challenge
- Techniques
  - Overview
  - PQ-based MatMul
  - Unified Architecture
  - Extensible GFM-ISA Design
- Experiment Results

Techniques Overview

Algorithm Level: PQ-based MatMul

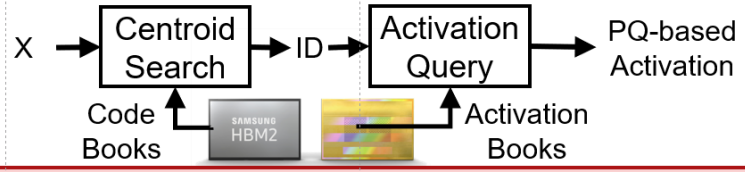
Original:



Model	Acc.
Original	0.9566
PQ-based	0.9561

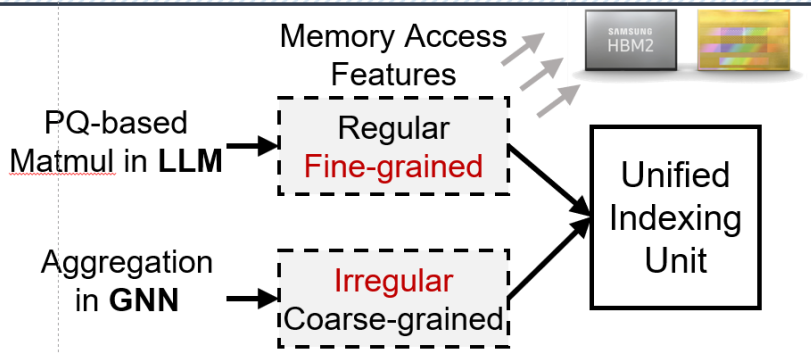
Loss. < 0.05%

Ours: PQ-based MatMul



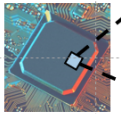
>70% computation reduction

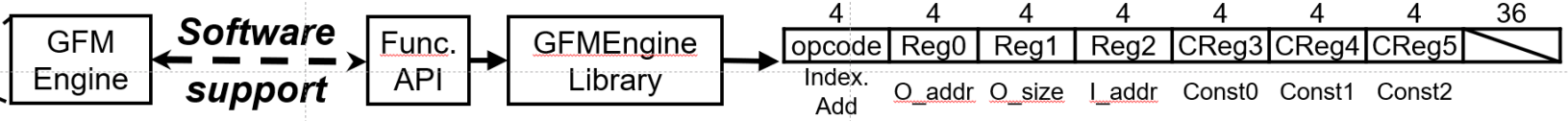
Hardware Level: Unified Indexing Unit



~30% memory access reduction

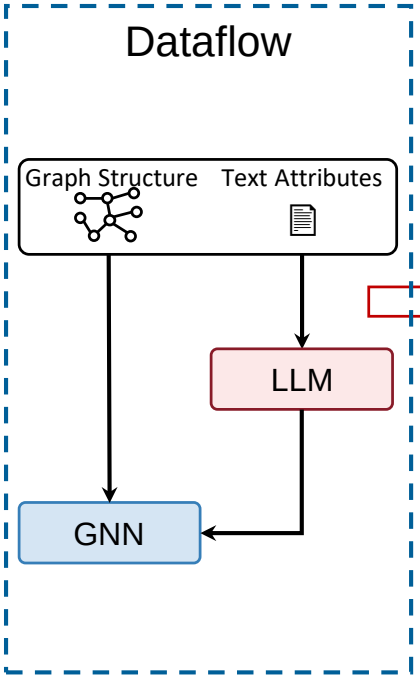
Compilation Level: An Extensible ISA Design

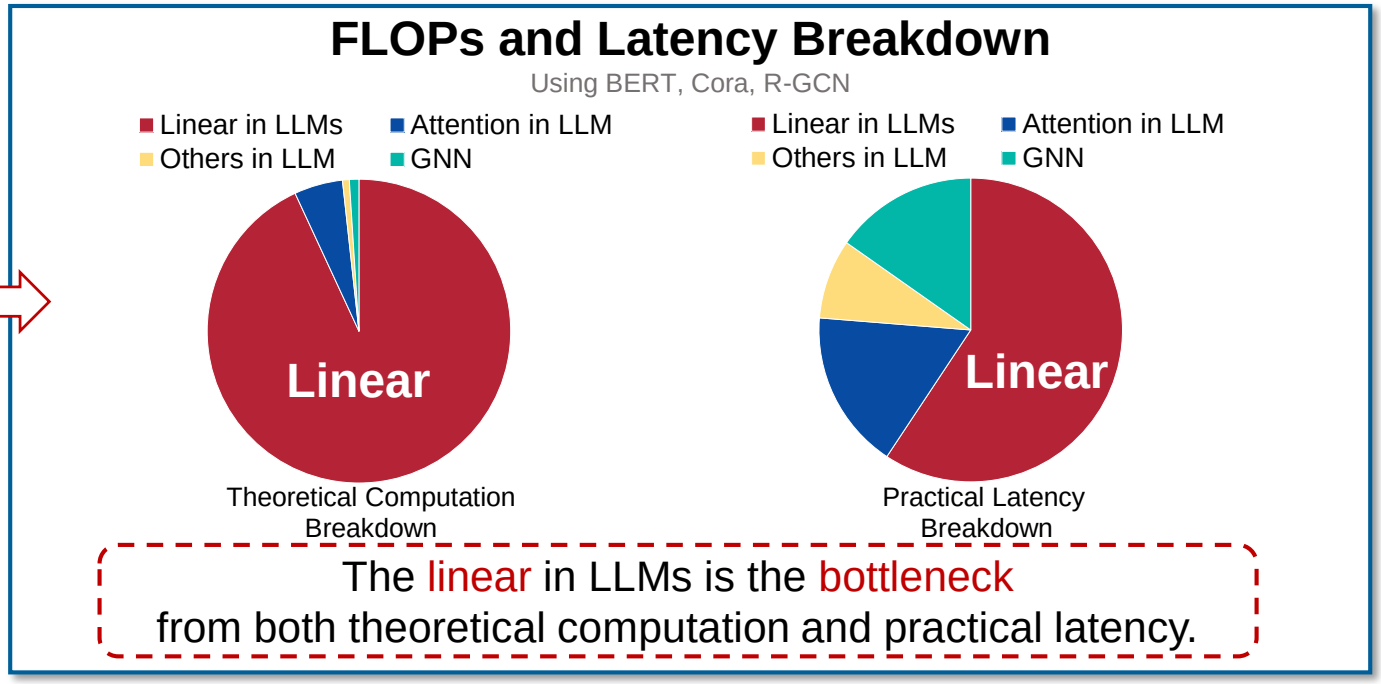




Technique 1: PQ-based MatMul

Motivation The **intensive GEMMs in LLM** need to be alleviated while ensuring the accuracy for end-to-end acceleration.

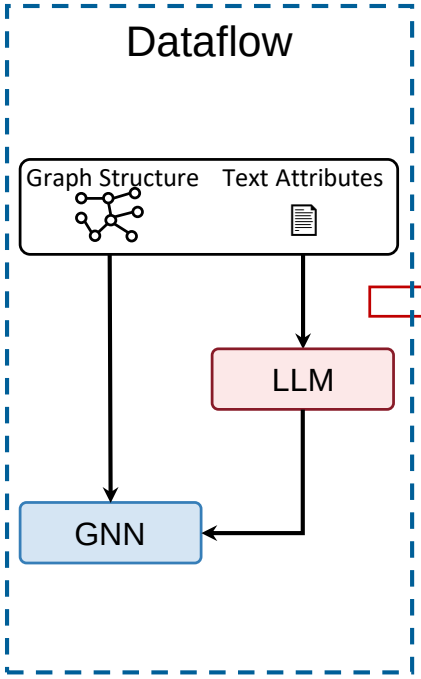


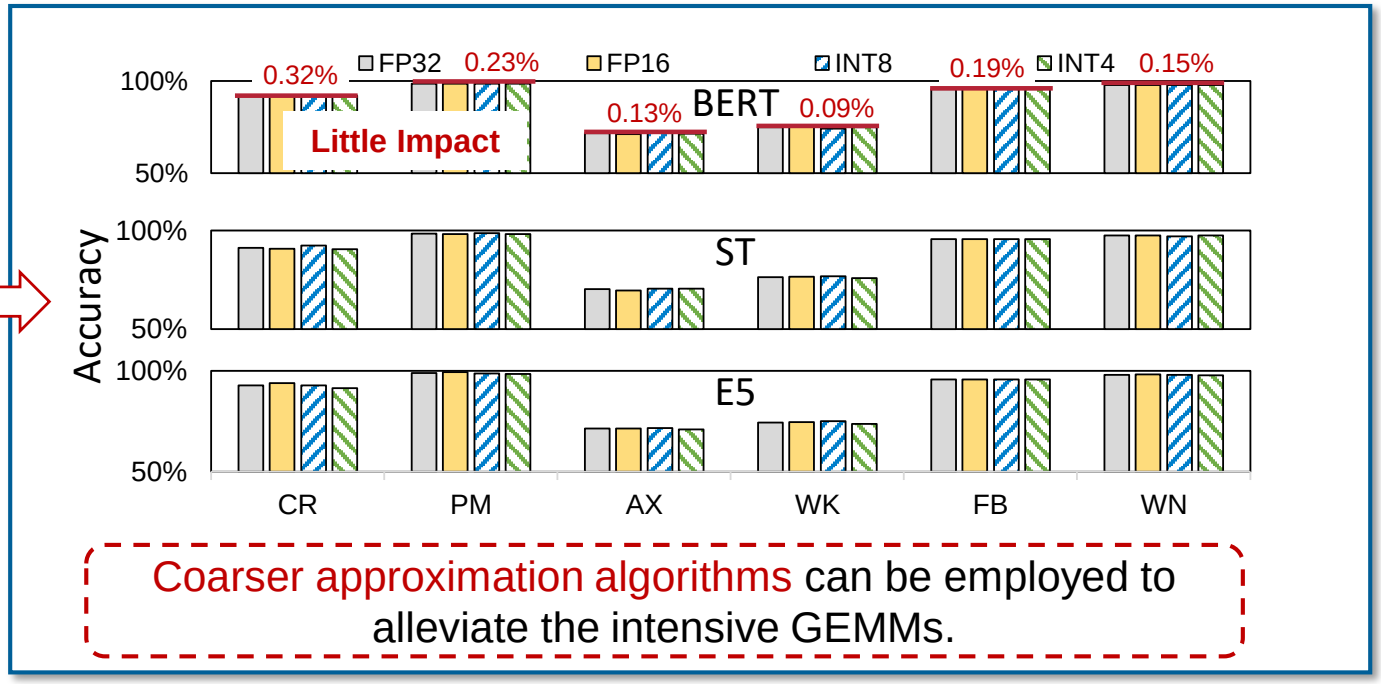


Technique 1: PQ-based MatMul

Insight

The computational precision of LLM has little impact on GNN training and inference.

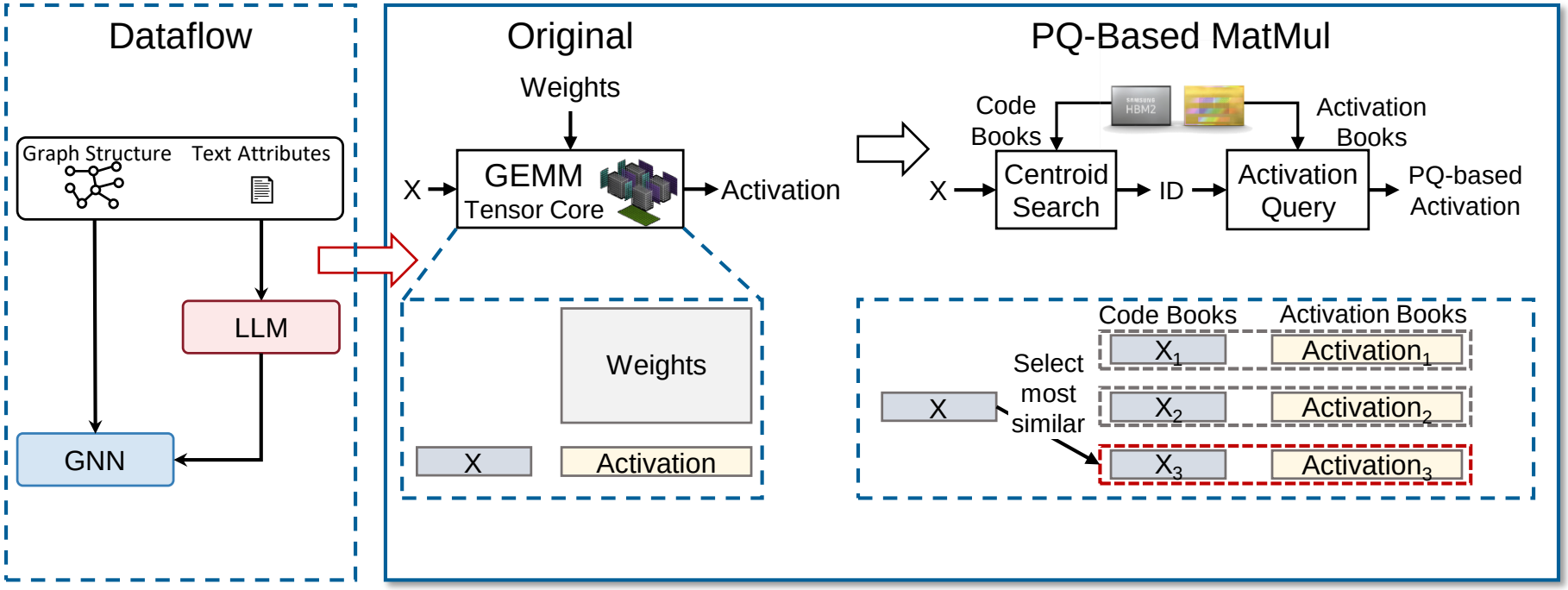




Technique 1: PQ-based MatMul

Approach

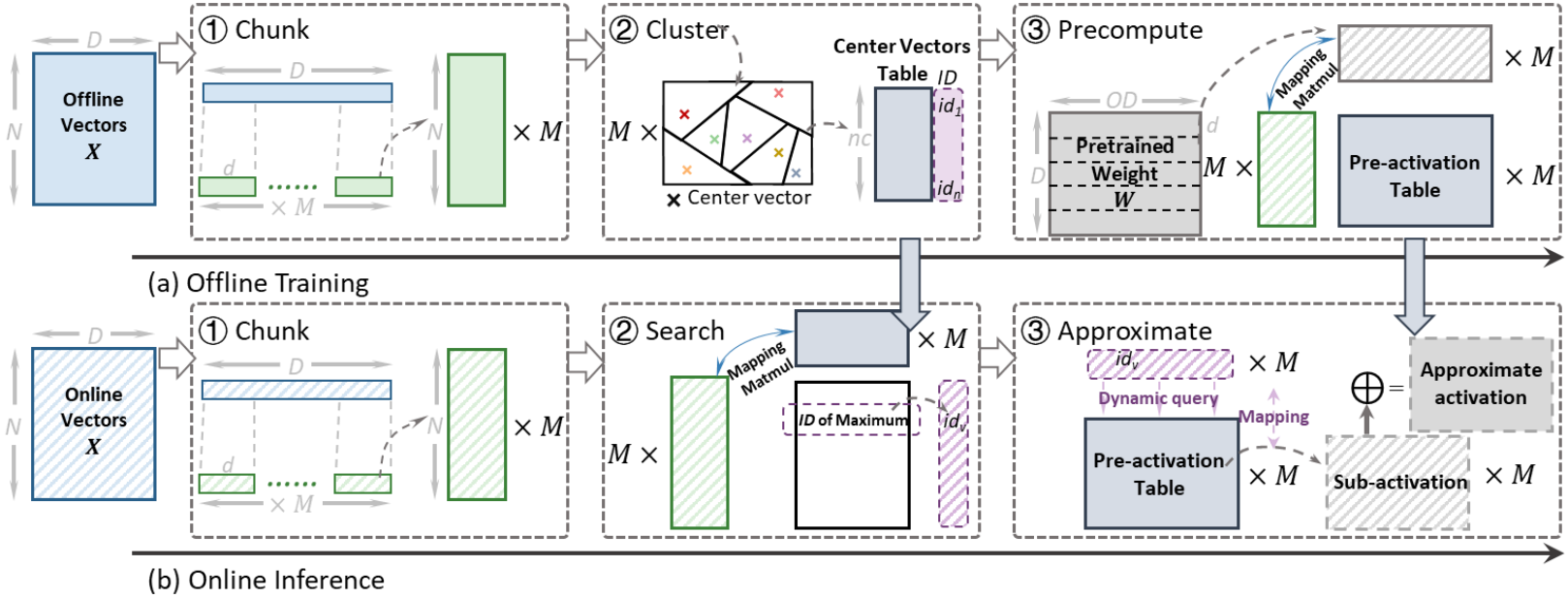
Product Quantization is used to approximate the linear operation in LLMs (PQ-based MatMul).



Technique 1: PQ-based MatMul

Approach

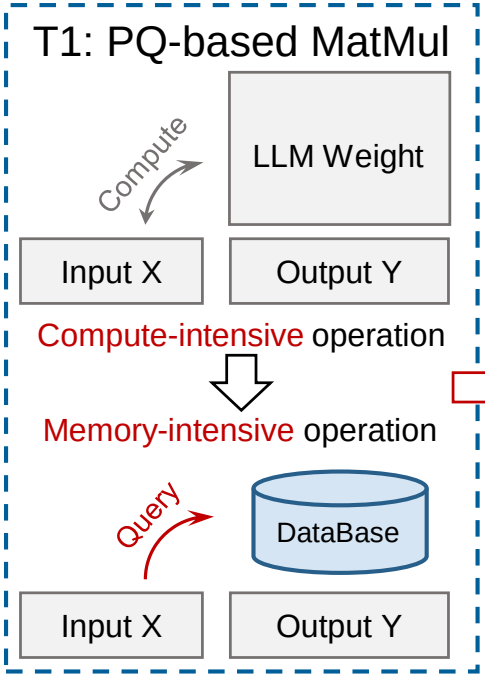
Product Quantization is used to approximate the linear operation in LLMs (PQ-based MatMul).

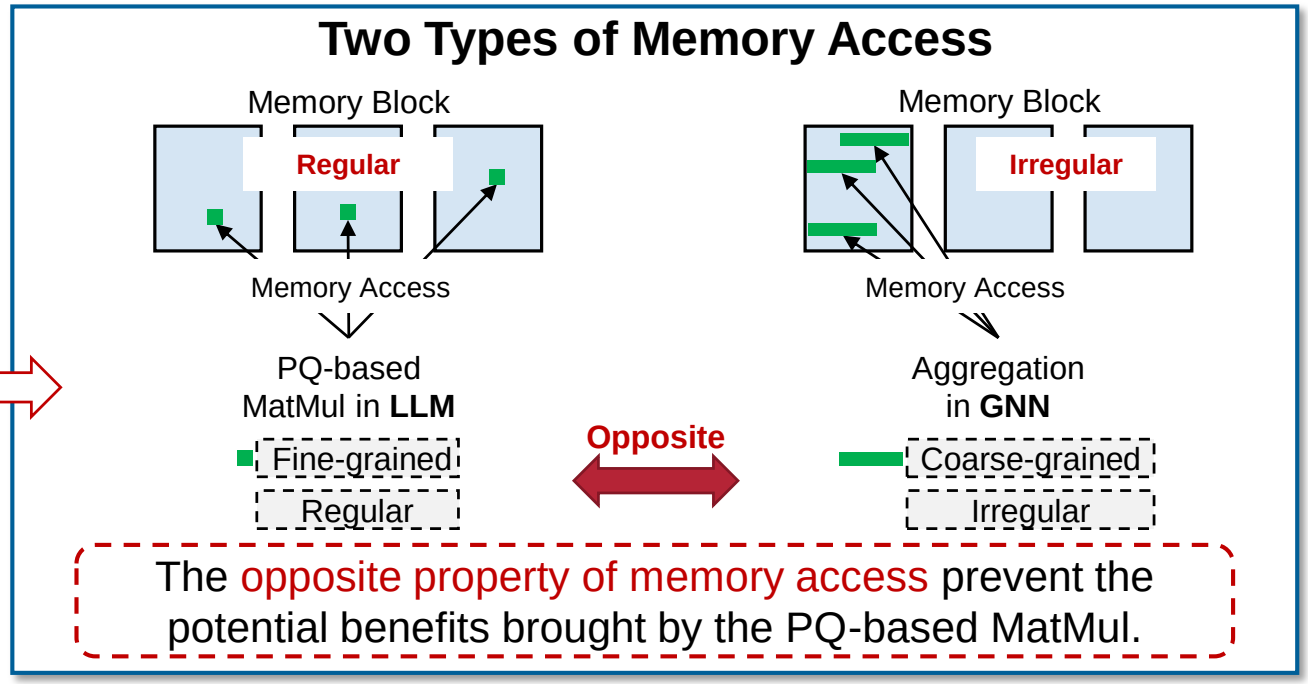


Reduce >70% overall computation workload

Technique 2: Unified Architecture

Motivation The **two types of memory access** in the LLM-enhanced GNN lower hardware utilization.

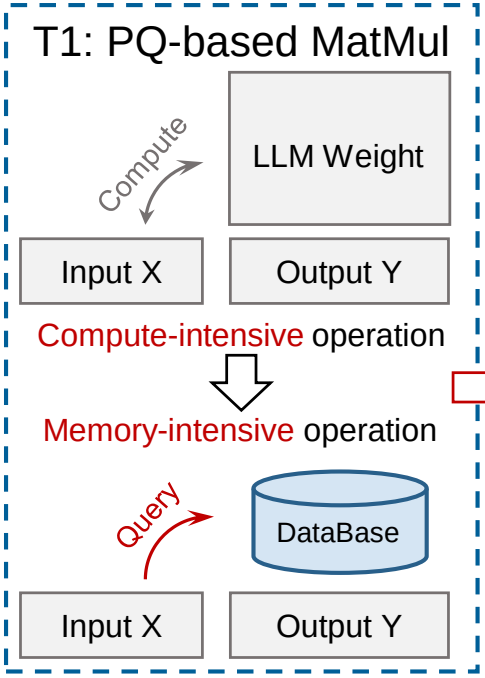


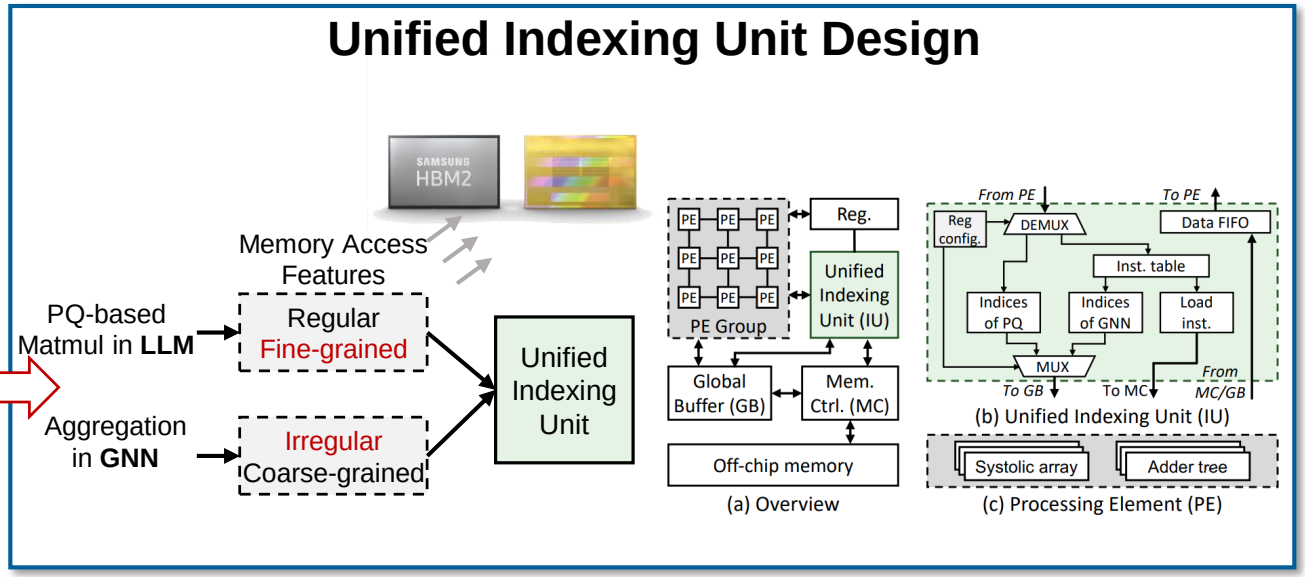


Technique 2: Unified Architecture

Approach

A hardware architecture with unified indexing unit is designed to support both types of computation.

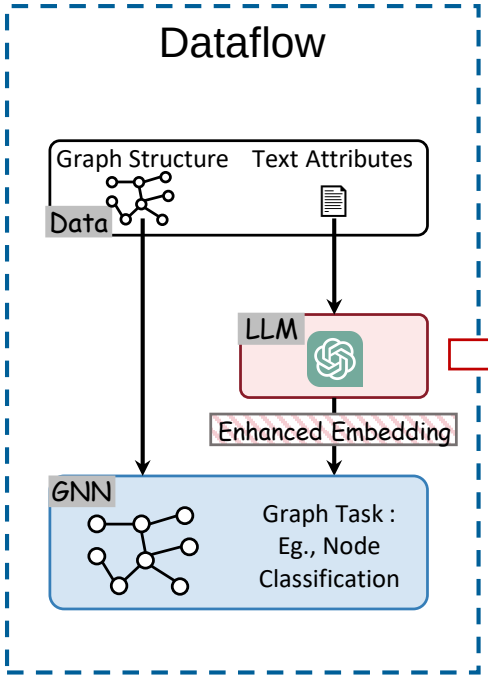


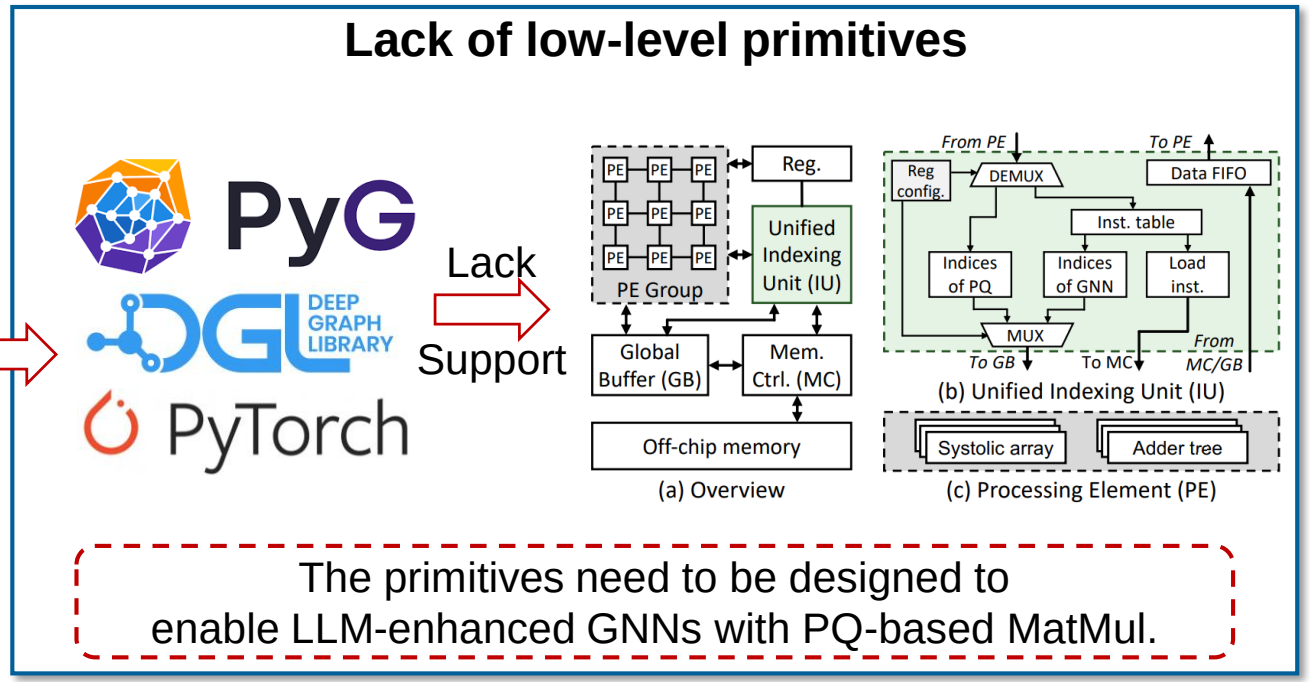


Reduce ~30% Memory Access

Technique 3: Extensible GFM-ISA Design

Motivation Existing software frameworks **lacking low-level primitives** for LLM-enhanced GNNs with PQ-based MatMul.

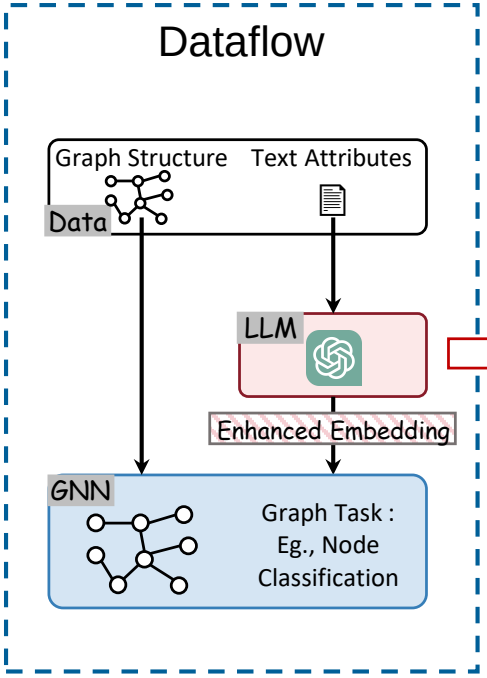


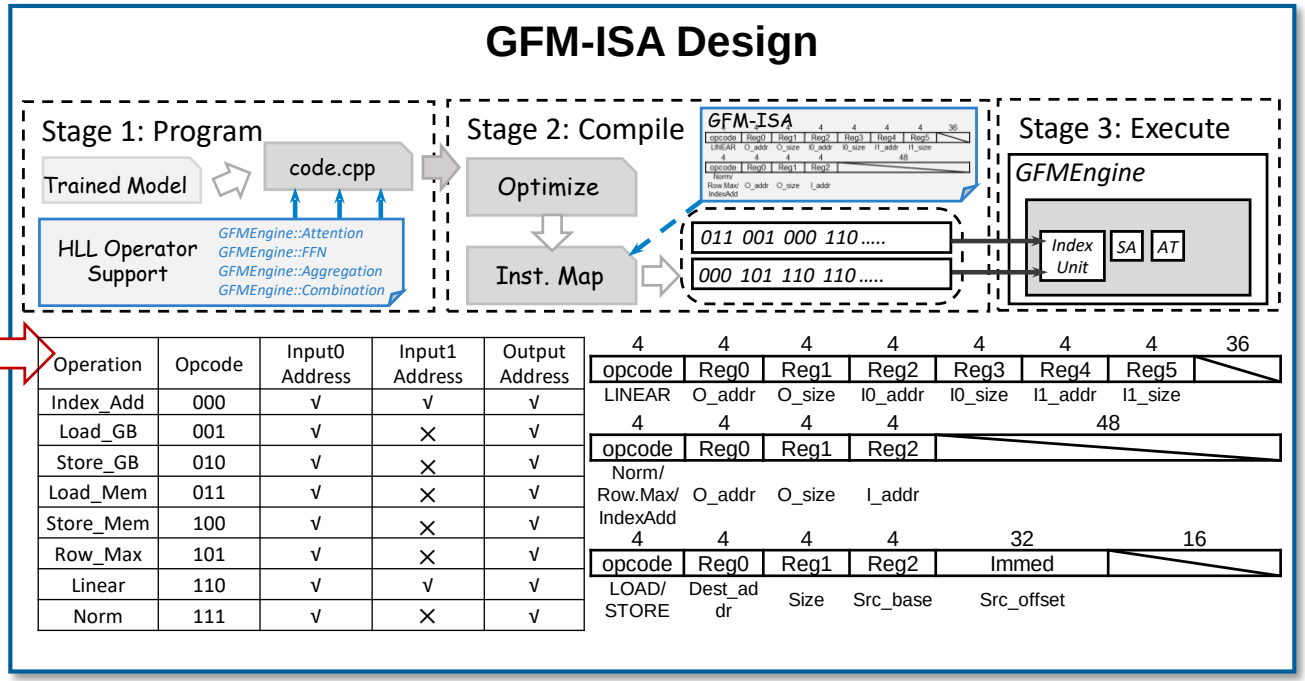


The primitives need to be designed to enable LLM-enhanced GNNs with PQ-based MatMul.

Technique 3: Extensible GFM-ISA Design

Approach **Extensible GFM-ISA design is used to offer low-level software primitives based on the unified architecture.**





Operation	Opcode	Input0 Address	Input1 Address	Output Address
Index_Add	000	√	√	√
Load_GB	001	√	×	√
Store_GB	010	√	×	√
Load_Mem	011	√	×	√
Store_Mem	100	√	×	√
Row_Max	101	√	×	√
Linear	110	√	√	√
Norm	111	√	×	√

4	4	4	4	4	4	4	36
opcode	Reg0	Reg1	Reg2	Reg3	Reg4	Reg5	
LINEAR	O_addr	O_size	I0_addr	I0_size	I1_addr	I1_size	

4	4	4	4	48
opcode	Reg0	Reg1	Reg2	
Norm/ Row.Max/ IndexAdd	O_addr	O_size	I_addr	

4	4	4	4	32	16
opcode	Reg0	Reg1	Reg2	Immed	
LOAD/ STORE	Dest_addr	Size	Src_base	Src_offset	

Outline

- Background and Challenge
- Techniques
  - Overview
  - PQ-based MatMul
  - Unified Architecture
  - Extensible GFM-ISA Design
- Experiment Results

Experiment Setup

- Tools for Evaluation: Ramulator 2.0 and TSMC 28nm process library
- Benchmark Models:
  - LLMs: BERT, Sentence Transformer, E5-large-v2, Llama2-7B/13B
  - GNN: R-GCN
- Baselines: NVIDIA A100 GPU, SGCN, MEGA, FACT
- Datasets

Table 1: Detailed Information of Datasets

Dataset	Task	Nodes	Avg. Edges
Cora (CR)	Node/Link	2,708	10,556
PubMed (PM)	Node/Link	19,717	44,338
ogbn-arxiv (AX)	Node	169,343	1,166,243
ogbn-products (PR)	Node	2,449,029	61,859,140
Wiki-CS (WK)	Node	11,701	216,123
FB15K237 (FB)	Link	14,541	310,116
WN18RR (WN)	Link	40,943	93,003

Table 2: System Configurations

		Freq.	Compute Unit	On-chip Memory	Bandwidth
GNN	**SGCN** [27]	1GHz	32x32 SA	512KB	256GB/s HBM2
GNN	**MEGA** [28]	1GHz	4x8x32 BSEs	392KB	256GB/s HBM1
LLM	**FACT** [18]	0.5GHz	16x32 SA	256KB	256GB/s HBM2
	GFMEngine	1GHz	16PEs (each with 4x16 SA 2x8 AT)	256KB	256GB/s HBM2

Accuracy Evaluation

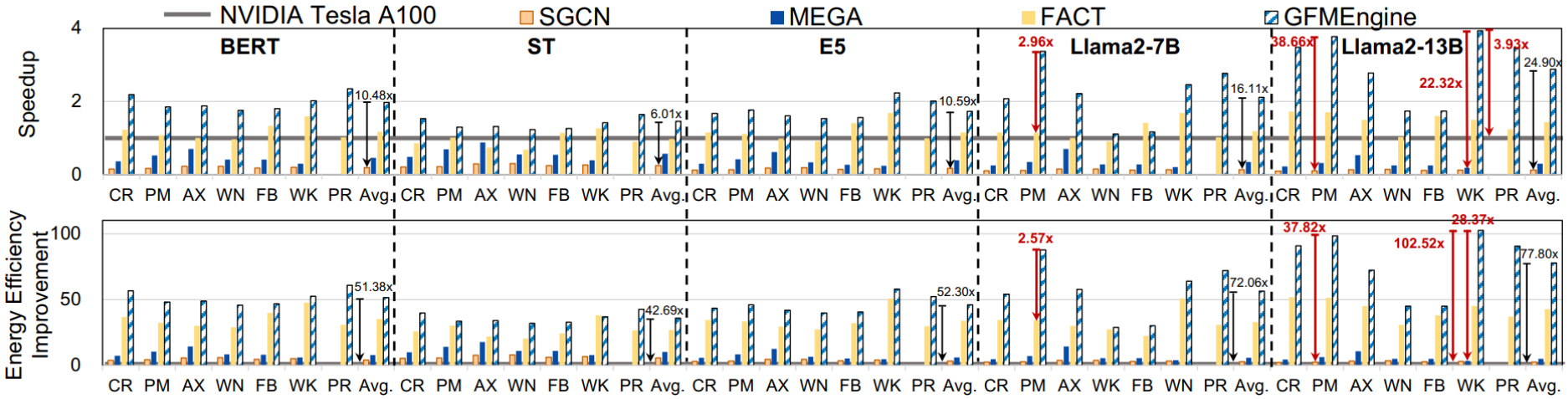
Table 3: Accuracy Loss

	BERT*	ST*	E5*	Llama2-7B*	Llama2-13B*
CR-Link	1.97	-0.81	0.92	0.51	-0.02
CR-Node	0.95	-3.64	-0.13	-0.86	0.58
PM-Link	0.41	0.09	0.19	0.16	0.47
PM-Node	2.39	0.54	1.40	1.37	0.36
AX	0.63	-0.92	0.02	-0.09	0.22
PR	1.19	0.28	0.11	-0.79	1.18
WK	2.62	2.41	0.61	0.31	0.62
FB	0.73	-0.06	0.6	0.33	1.2
WN	0.12	0.12	0.16	0.96	-1.69
Avg. Loss (↓)	1.22	-0.22	0.43	0.21	0.32

* represents the accuracy loss: the accuracy of original LLM-enhanced GNNs minus the accuracy of those models with PQ-based MatMul)

The accuracy loss of the LLM-enhanced GNNs with PQ-based MatMul can be negligible in most cases.

Hardware Evaluation



Compared with A100 GPU and other DSA, we achieve up to 3.93×, and 38.66× speedup and up to 102.52×, 37.82× energy efficiency.

Accelerator for LLM-Enhanced GNN with Product Quantization and Unified Indexing

Jiaming Xu, supervised by Prof. Guohao Dai
jiamingxu@sjtu.edu.cn, daiguohao@sjtu.edu.cn