Semantic-Structural Integration in Text-Attributed Grahs

Yuan FANG
School of Computing and Information Systems
Singapore Management University

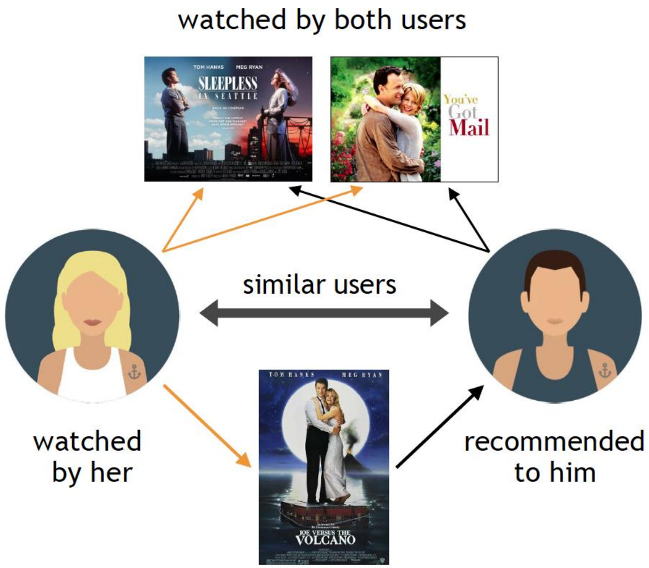
ISCSC 2025 @ Taiyuan, China
9 Nov 2025

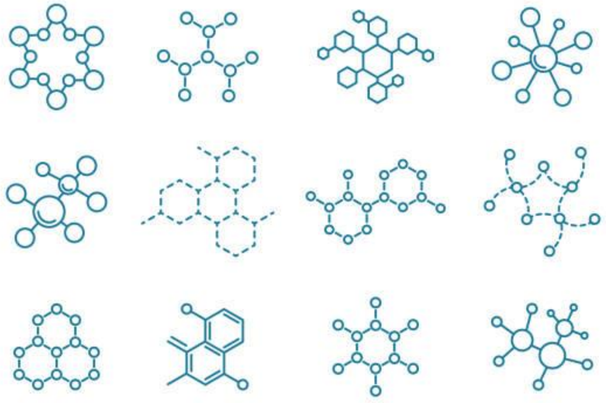
SMU SINGAPORE MANAGEMENT UNIVERSITY
School of Computing and Information Systems

Outline

- **Introduction: Graphs & text-attributed graphs**
- Jointly training graph and textual data
- Quantizing graphs into language tokens for LLMs
- Conclusions

Graph structures are prevalent





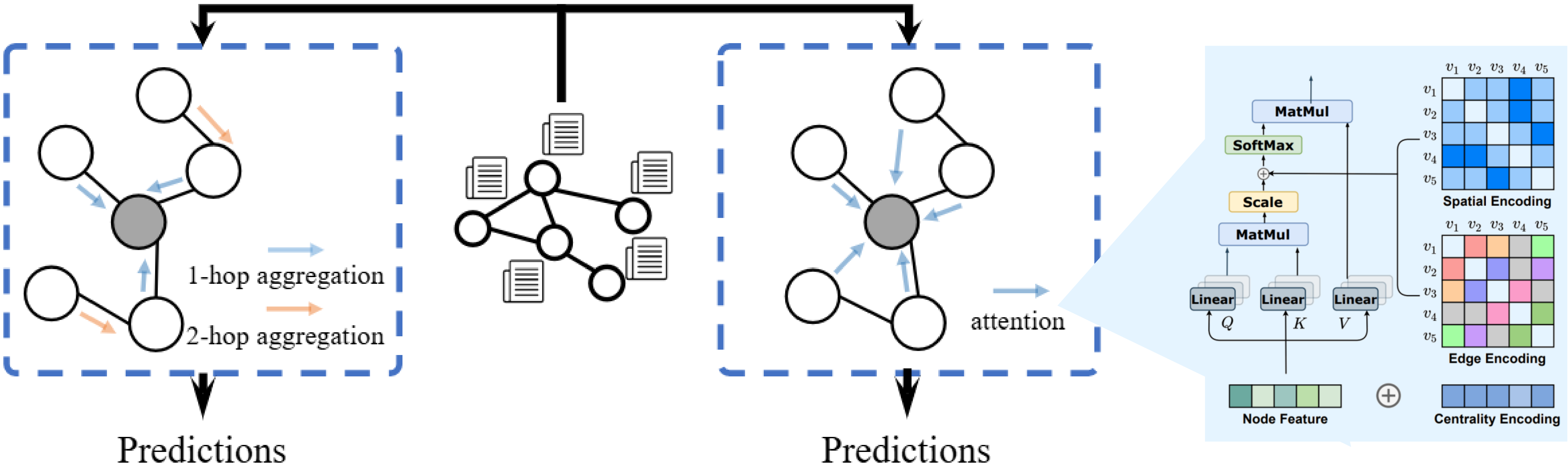
Social network

Recommendation System

Molecular graphs

Wu, *et al*. A comprehensive survey on graph neural networks. TNNLS 2021.

Learning from graph structures



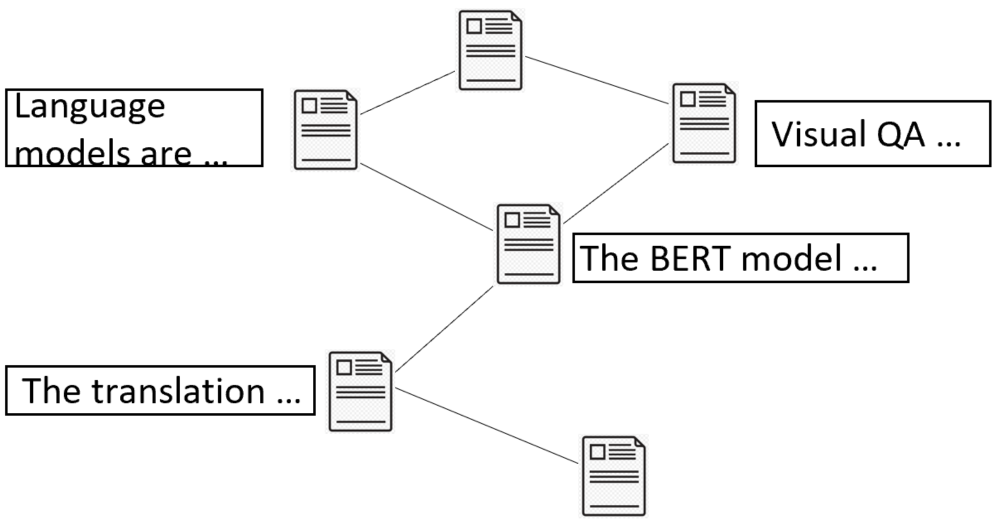
Message-passing graph neural networks (GNNs)

Graph transformers

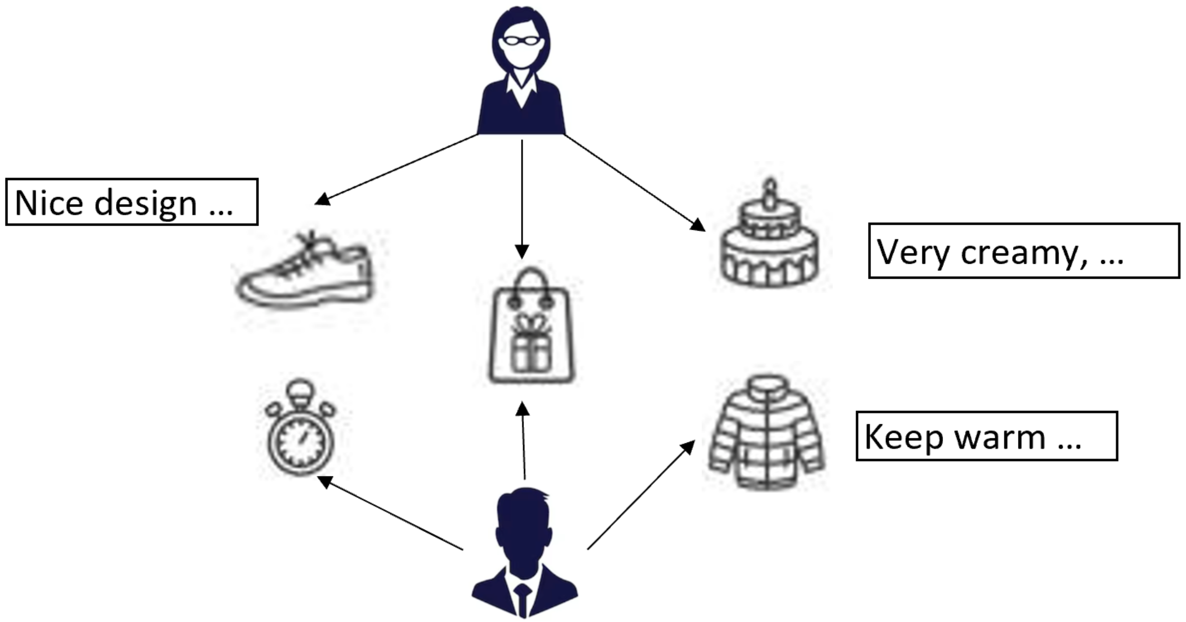
Ying, *et al*. Do Transformers Really Perform Bad for Graph Representation? NeurIPS 2021.

Liu, *et al.* Graph foundation models: Concepts, Opportunities and Challenges. TPAMI 2025.

Semantics on graphs: Text-Attributed Graphs



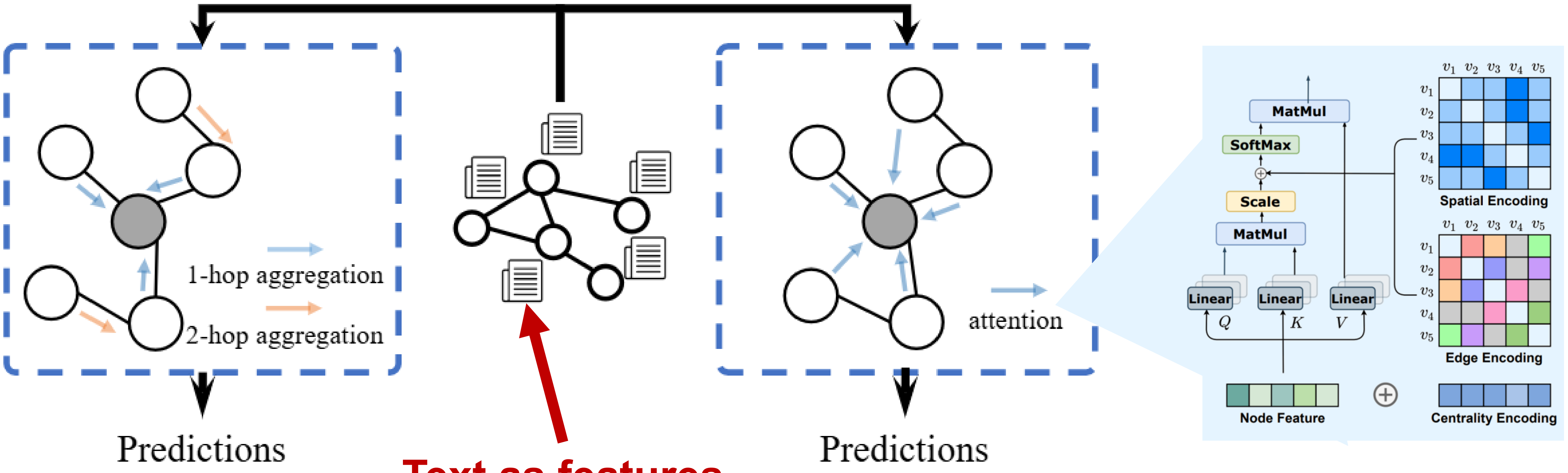
Citation graph for online articles



E-commerce item review graph

Can we **integrate** graph structures and textual semantics within one model?

Can GNNs or graph transformers utilize textual attributes? Yes, but ineffective



Text as features (embeddings, BOW)

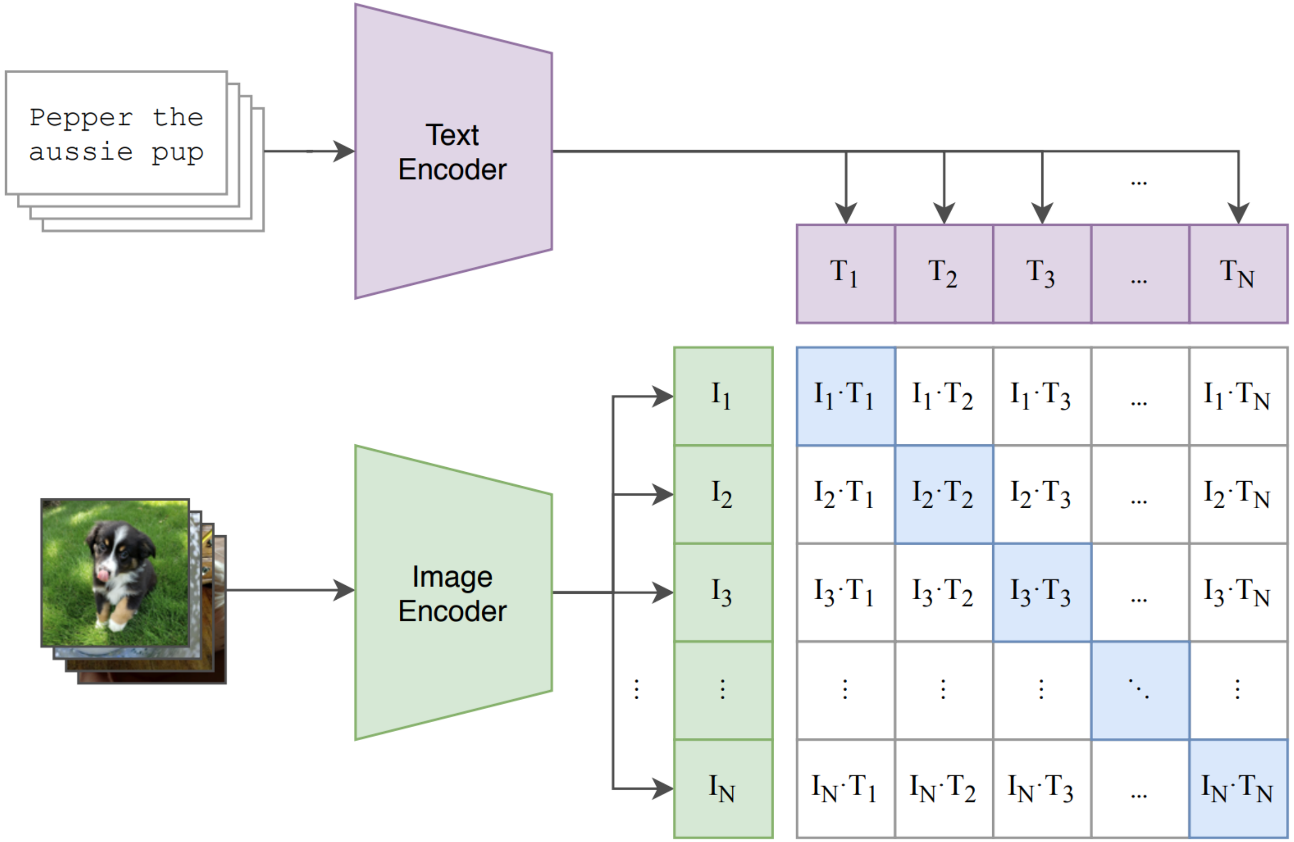
Ying, *et al*. Do Transformers Really Perform Bad for Graph Representation? NeurIPS 2021.
Liu, *et al.* Graph foundation models: Concepts, Opportunities and Challenges. TPAMI 2025.

Outline

- Introduction: Graphs & text-attributed graphs
- **Jointly training graph and textual data**
- Quantizing graphs into language tokens for LLMs
- Conclusions

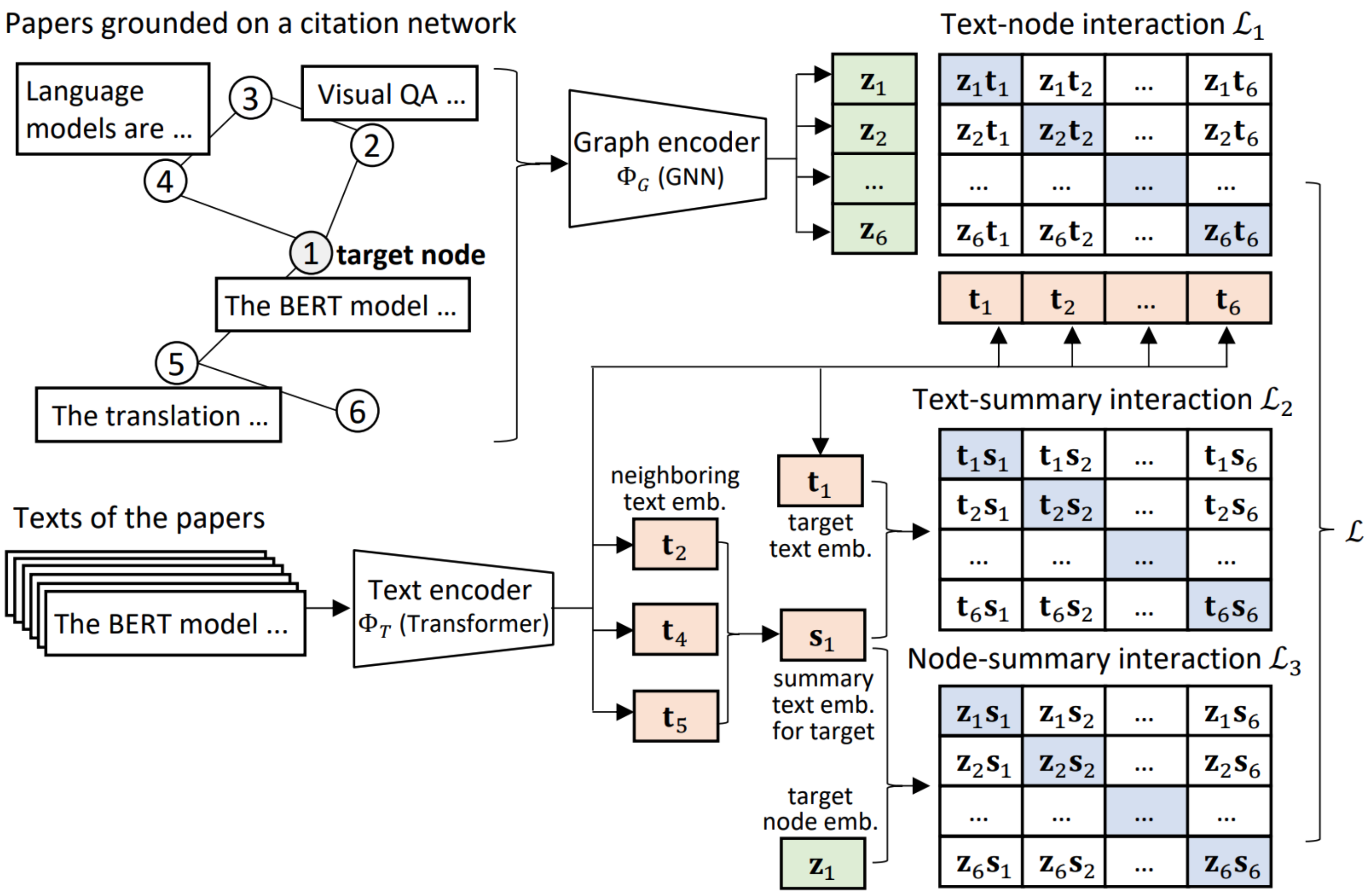
How are language-image models trained?

CLIP: Contrastive Language-Image Pre-training



Radford *et al*. Learning Transferable Visual Models From Natural Language Supervision. OpenAI 2021.

Graph-grounded **pre-training** and prompting (G2P2)



Learns a dual-modal embedding space by jointly pre-training a **text encoder** and **graph encoder**

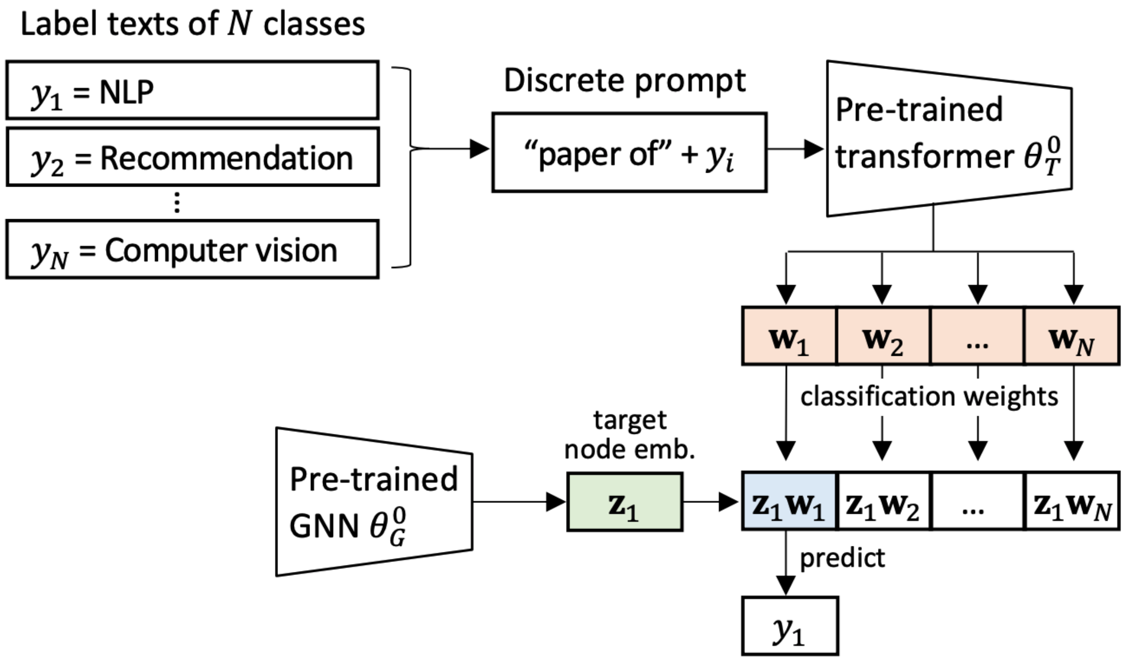
Exploits **three contrastive losses**

- $\mathcal{L}_1$: Text-node contrast
- $\mathcal{L}_2$: Text-summary contrast
- $\mathcal{L}_3$: Node-summary contrast

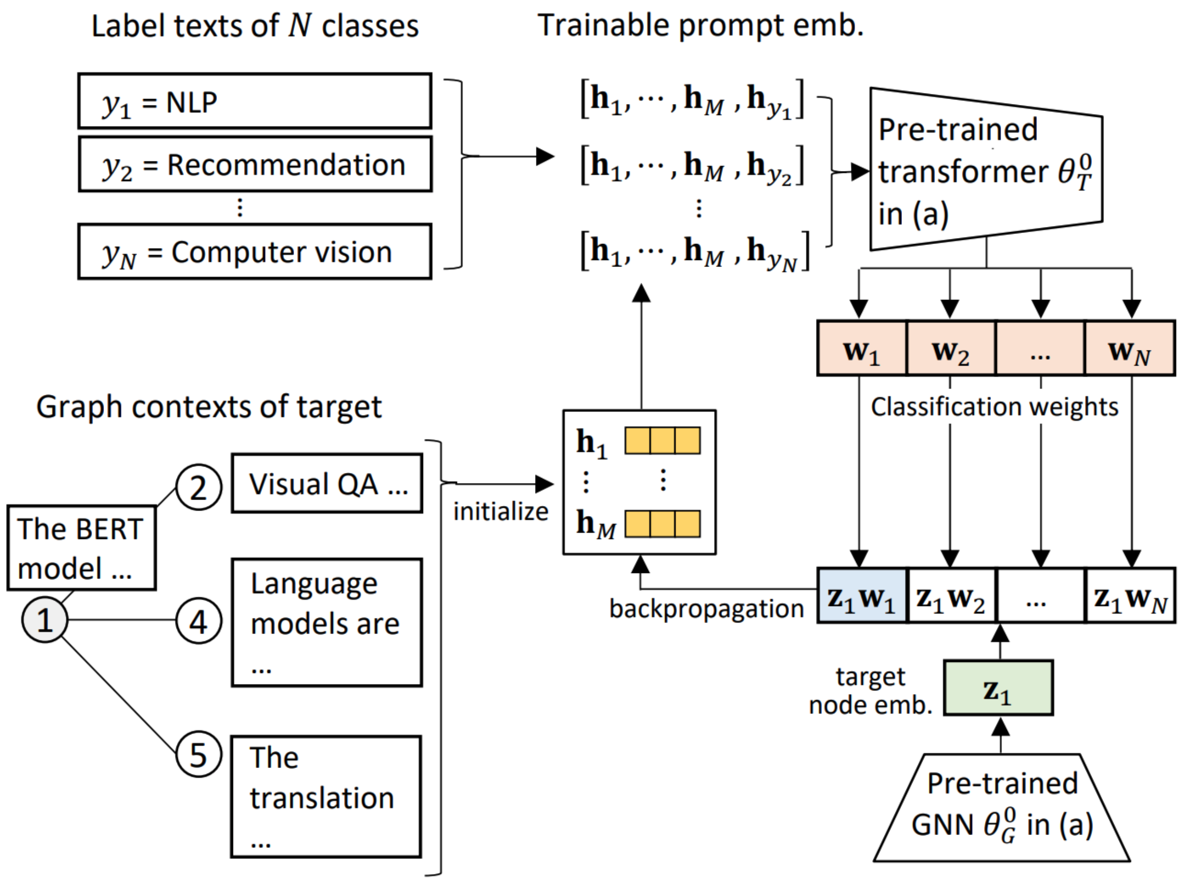
Wen *et al*. Augmenting Low-Resource Text Classification with Graph-Grounded Pre-training and Prompting. SIGIR 2023.

Graph-grounded pre-training and **prompting** (G2P2)

Zero-shot node classification with discrete prompts

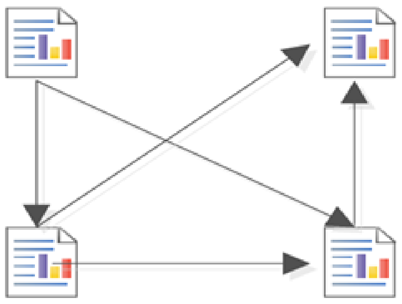


Few-shot node classification with continuous prompt tuning

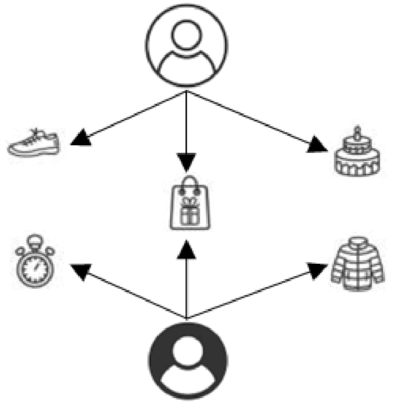


Datasets to evaluate G2P2

Dataset	Cora	Art	Industrial	M.I.
# Documents	25,120	1,615,902	1,260,053	905,453
# Links	182,280	4,898,218	3,101,670	2,692,734
# Avg. doc length	141.26	54.23	52.15	84.66
# Avg. node deg	7.26	3.03	2.46	2.97
# Classes	70	3,347	2,462	1,191



Cora is a collection of research papers with citation links



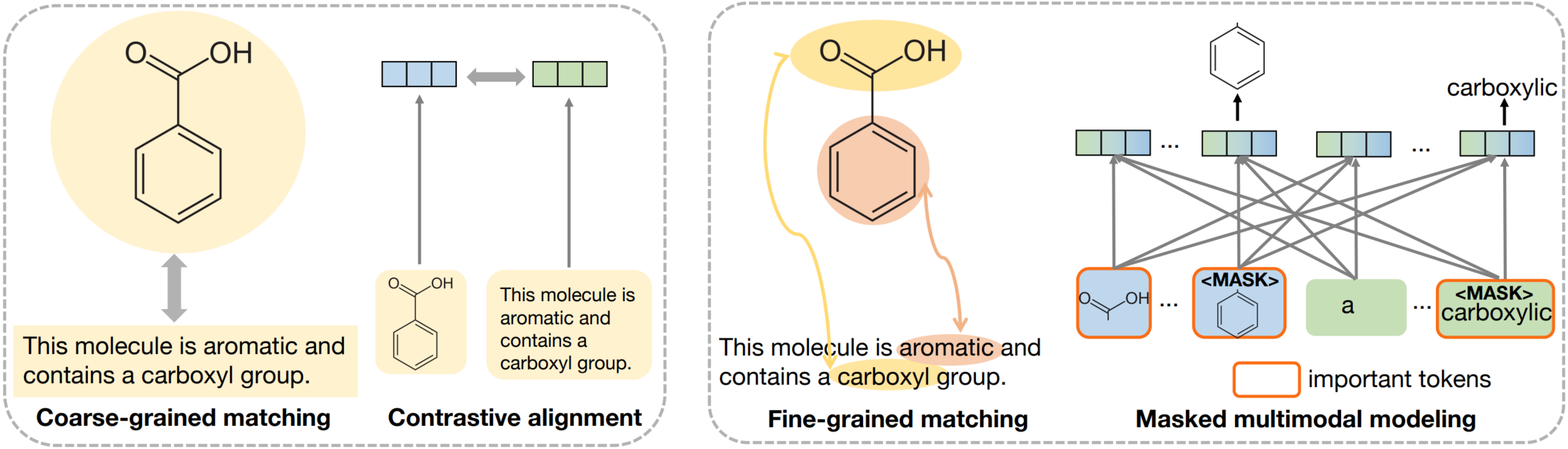
Art, **Industrial** and **Music Instruments** (M.I.) are three Amazon review datasets

Empirical performance of G2P2

		Cora		Art		Industrial		M.I.	
		Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
End-to-end GNN	GCN	41.15±2.41	34.50±2.23	22.47±1.78	15.45±1.14	21.08±0.45	15.23±0.29	22.54±0.82	16.26±0.72
	SAGE_{sup}	41.42±2.90	35.14±2.14	22.60±0.56	16.01±0.28	20.74±0.91	15.31±0.37	22.14±0.80	16.69±0.62
	TextGCN	59.78±1.88	55.85±1.50	43.47±1.02	32.20±1.30	53.60±0.70	45.97±0.49	46.26±0.91	38.75±0.78
Pre-trained GNN	GPT-GNN	76.72±2.02	72.23±1.17	65.15±1.37	52.79±0.83	62.13±0.65	54.47±0.67	67.97±2.49	59.89±2.51
	DGI	78.42±1.39	74.58±1.24	65.41±0.86	53.57±0.75	52.29±0.66	45.26±0.51	68.06±0.73	60.64±0.61
	$\text{SAGE}_{\text{self}}$	77.59±1.71	73.47±1.53	76.13±0.94	65.25±0.31	71.87±0.61	65.09±0.47	77.70±0.48	70.87±0.59
Pre-trained Transformers	BERT	37.86±5.31	32.78±5.01	46.39±1.05	37.07± 0.68	54.00±0.20	47.57±0.50	50.14±0.68	42.96±1.02
	BERT*	27.22±1.22	23.34±1.11	45.31±0.96	36.28±0.71	49.60±0.27	43.36±0.27	40.19±0.74	33.69±0.72
	RoBERTa	62.10±2.77	57.21±2.51	72.95±1.75	62.25±1.33	76.35±0.65	70.49±0.59	70.67±0.87	63.50±1.11
	RoBERTa*	67.42±4.35	62.72±3.02	74.47±1.00	63.35±1.09	77.08±1.02	71.44±0.87	74.61±1.08	67.78±0.95
Prompt tuning	P-Tuning v2	71.00±2.03	66.76±1.95	76.86±0.59	66.89±1.14	79.65±0.38	74.33±0.37	72.08±0.51	65.44±0.63
	G2P2-p	79.16±1.23	74.99±1.35	79.59±0.31	68.26±0.43	80.86±0.40	74.44±0.29	81.26±0.36	74.82±0.45
	G2P2	**80.08***±1.33	**75.91***±1.39	**81.03***±0.43	**69.86***±0.67	**82.46***±0.29	**76.36***±0.25	**82.77***±0.32	**76.48***±0.52
	(improv.)	(+2.12%)	(+1.78%)	(+5.43%)	(+4.44%)	(+3.53%)	(+2.7%)	(+6.53%)	(+7.92%)

G2P2 outperforms the best baseline (at that time) by around 3–7%.

Fine-grained graph-text integration

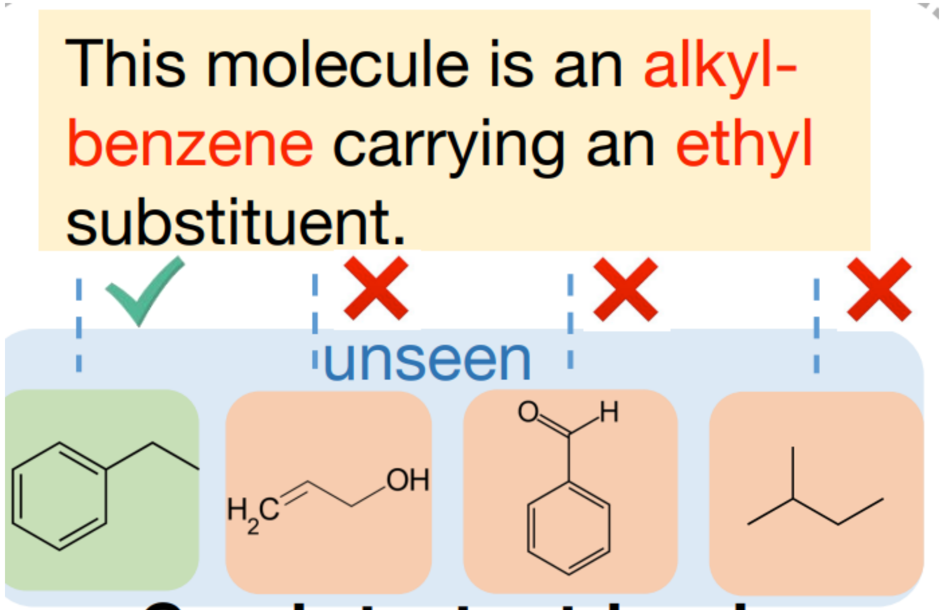


(a) Molecule-level learning

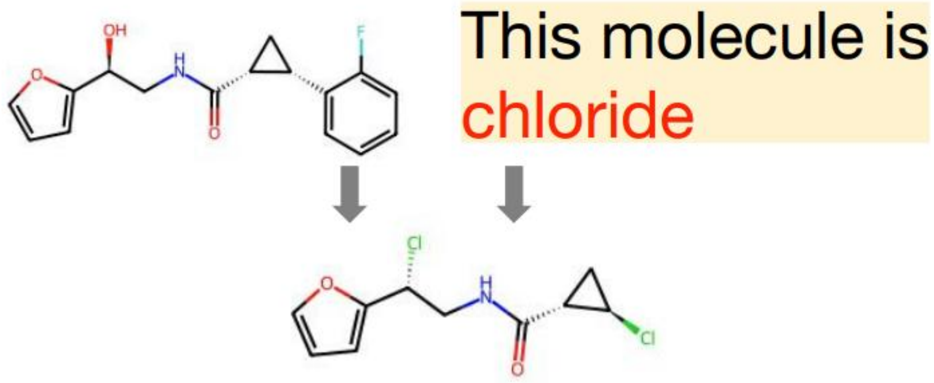
(b) Motif-level learning

Li *et al.* Advancing Molecular Graph-Text Pre-training via Fine-grained Alignment. KDD 2025.

Why is fine-grained alignment important?

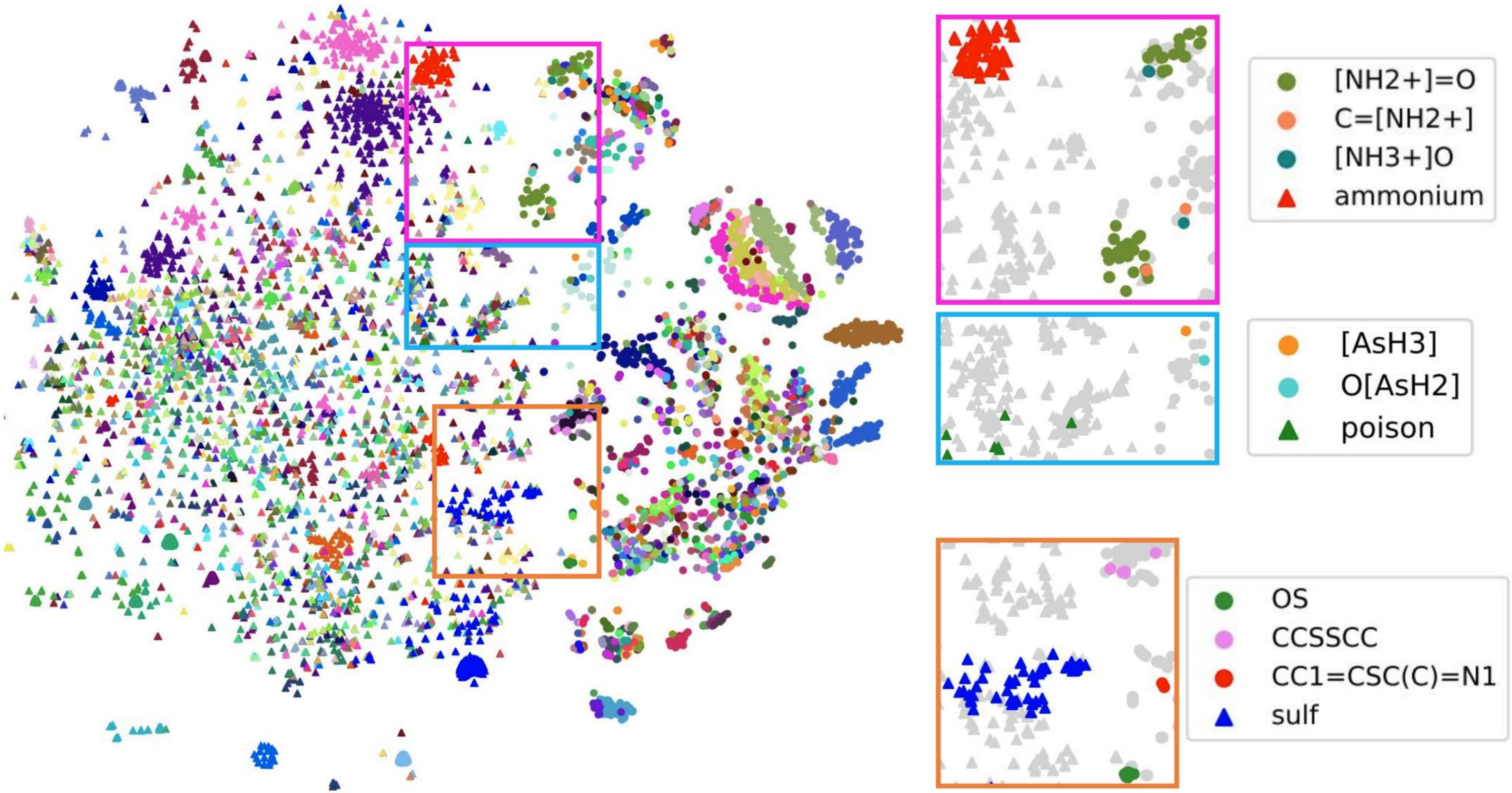


Graph-text retrieval



Text-based molecule editing

Visualization of learned word/motif embeddings

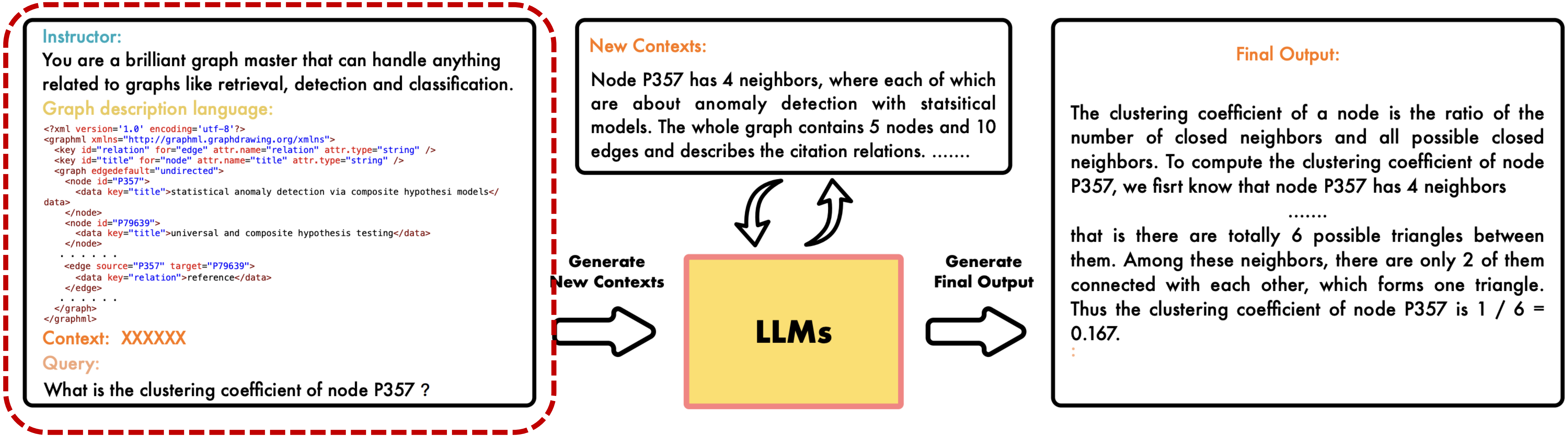


Outline

- Introduction: Graphs & text-attributed graphs
- Jointly training graph and textual data
- **Quantizing graphs into language tokens for LLMs**
- Conclusions

Integrating graph data in the era of LLMs

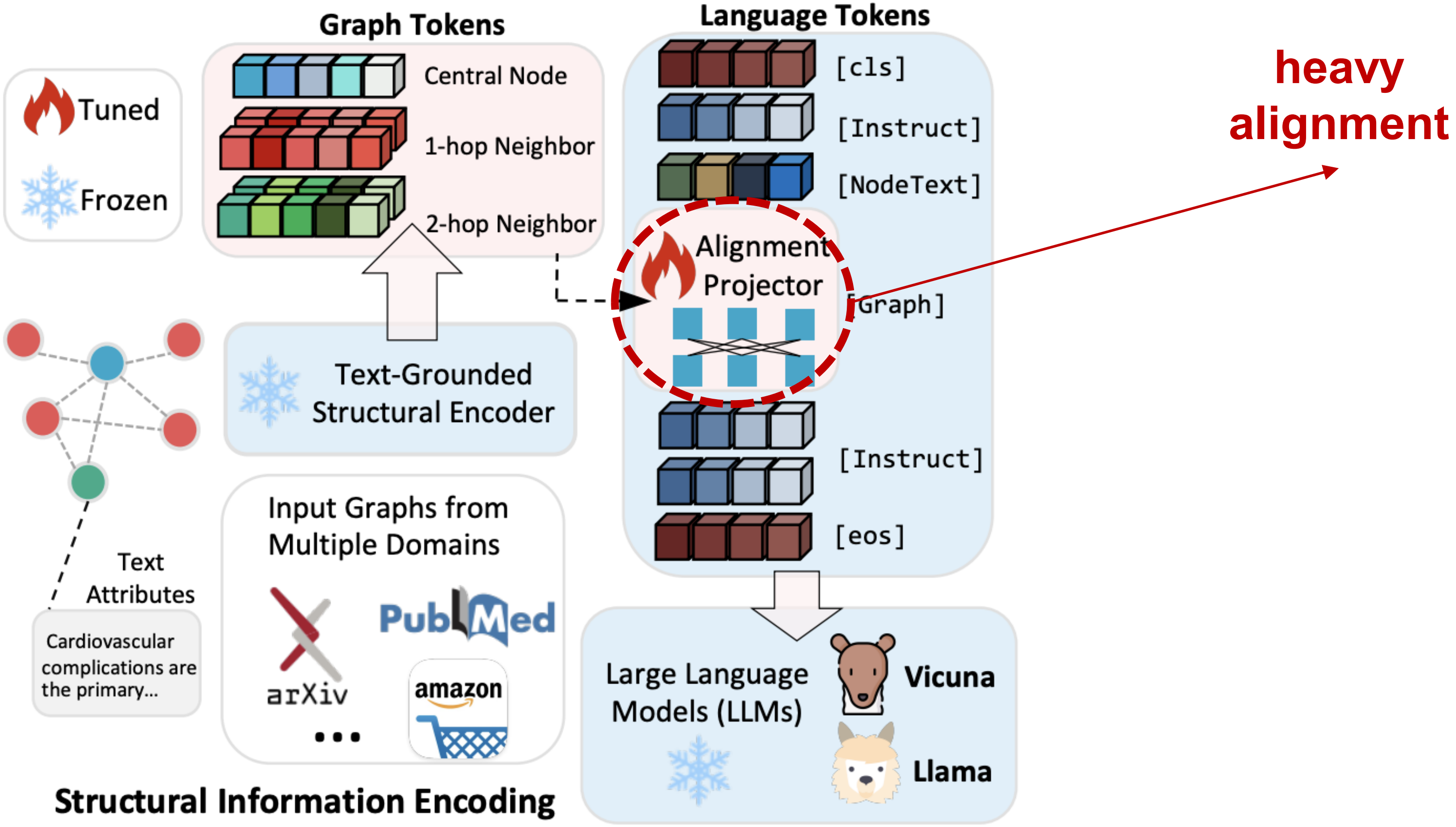
Graph verbalization



GPT4Graph

Guo *et al.* GPT4Graph: Can large language models understand graph structured data? an empirical evaluation and benchmarking. arXiv preprint 2023.

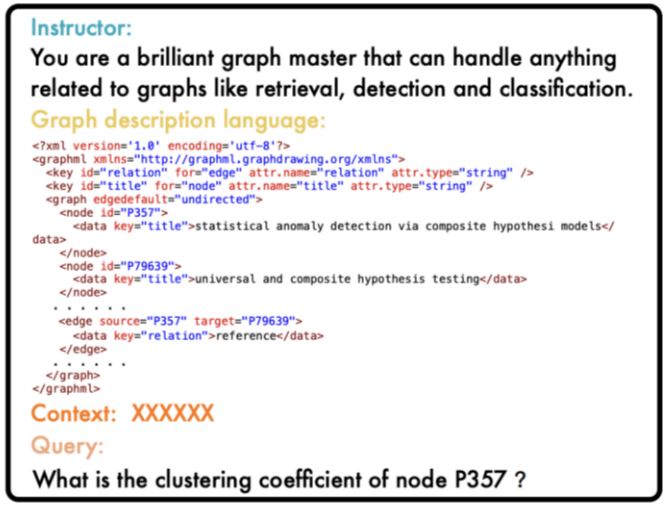
Integrating graph data in the era of LLMs



GraphGPT

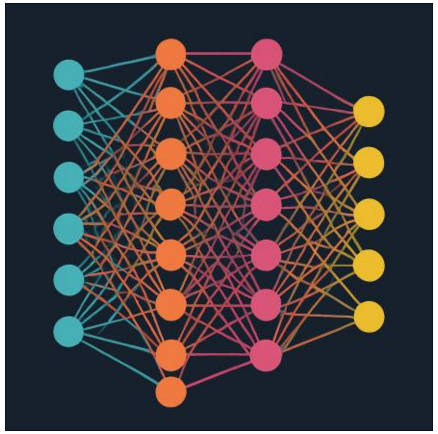
Tang *et al.* GraphGPT: Graph Instruction Tuning for Large Language Models. SIGIR 2024.

Structural-semantic gap



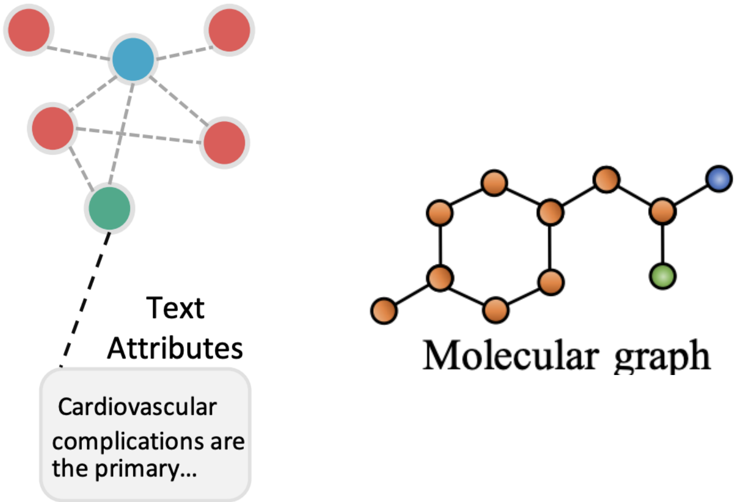
Graph Verbalization

Structural information loss



Projector-based Alignment

High computational cost



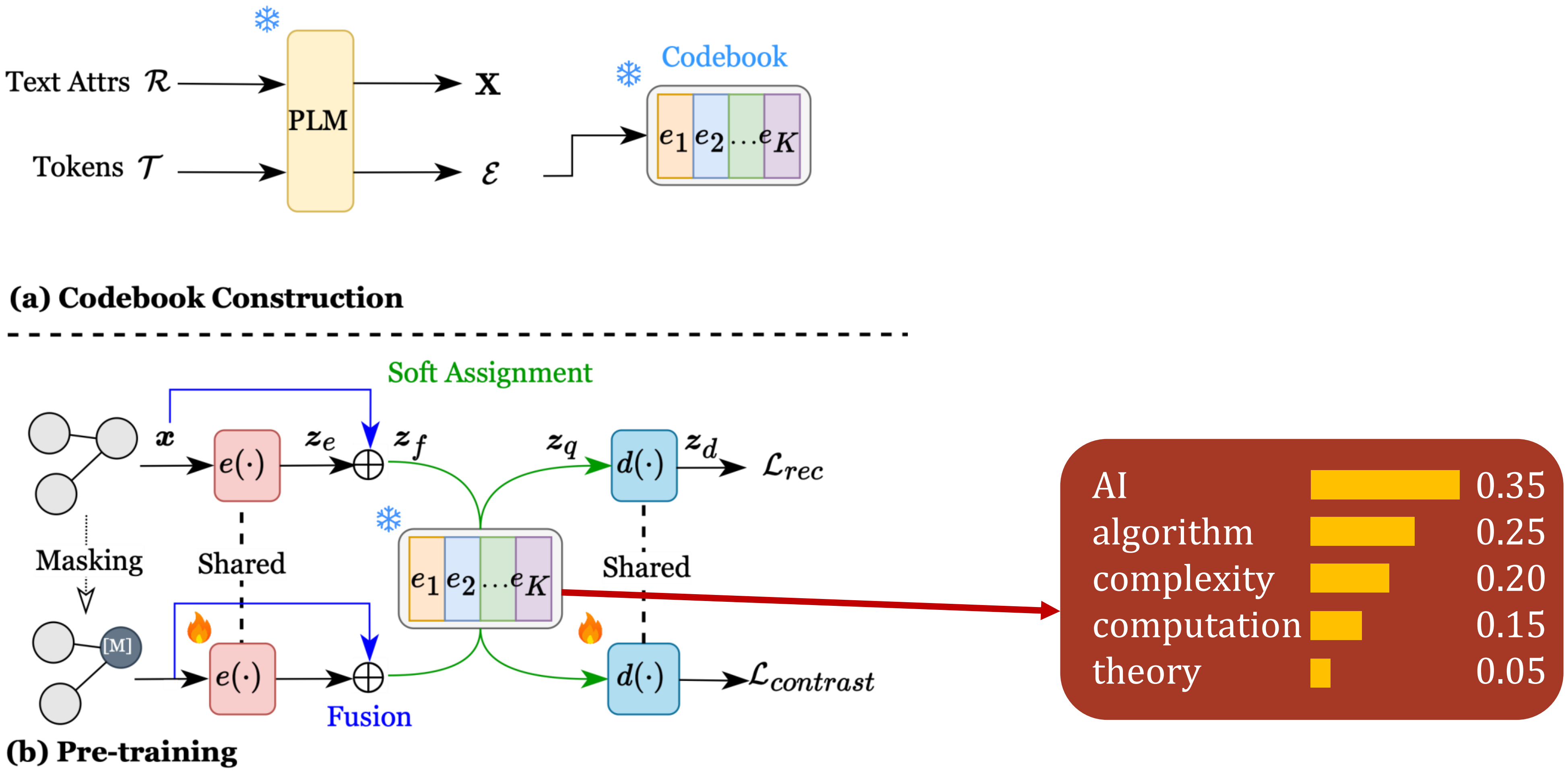
Transfer learning

Poor generalization

Continuous vs. Discrete

Graph embeddings ↔ LLM tokens

Soft Tokenization of Text-attributed Graphs (STAG)



Bo *et al*. Quantizing Text-attributed Graphs for Semantic-Structural Integration. KDD 2025.

Soft Tokenization of Text-attributed Graphs (STAG)

✓ **With LLMs**

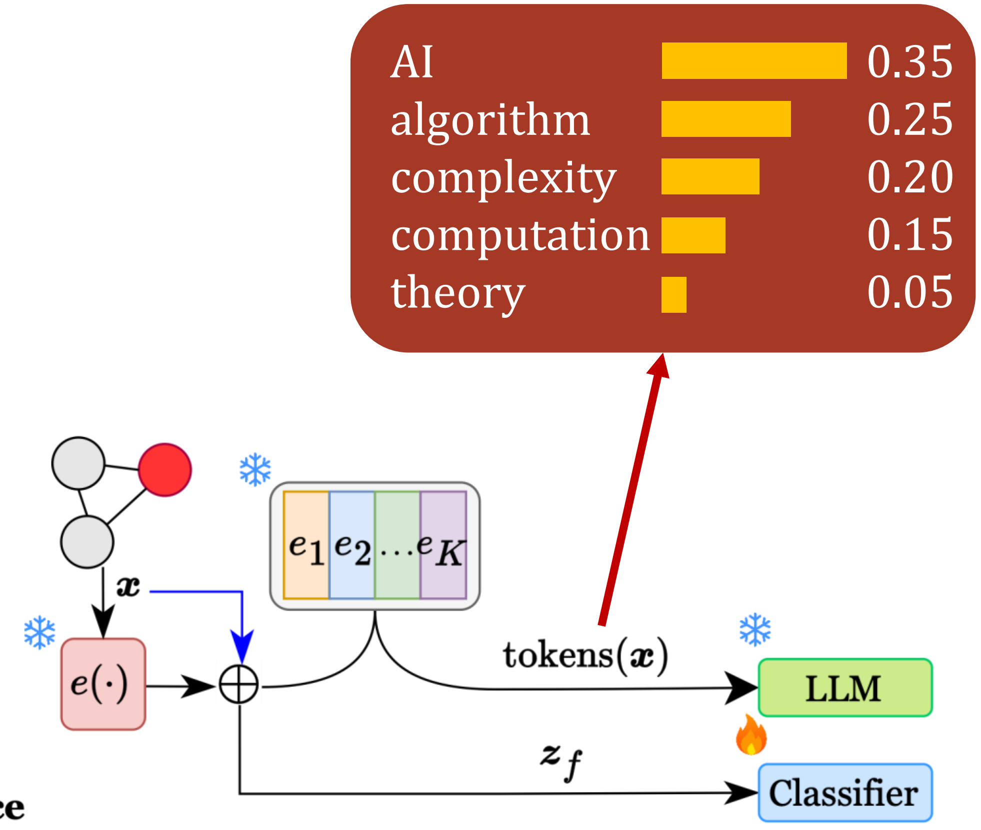
- Extract top-k tokens
- Few-shot: In-context learning
- Zero-shot: Direct LLM classification

✓ **Without LLMs**

- Linear probing on frozen embeddings

✓ **Prompt Tuning**

- Lightweight adaptation for domain transfer
- Supports both LLMs and without LLMs



(c) Inference

Inference with LLMs

System Prompt: You are a node classifier. Given a list of tokens representing a node's features, predict its class from the following options: [Research Paper, Dataset, Software].

Few-shot examples: Node tokens: [research, methodology, experiment] Class: Research Paper

Node tokens: [benchmark, statistics, collection] Class: Dataset

Node tokens: [implementation, code, library] Class: Software

Test Node: Node tokens: [algorithm, computation, optimization] Predict the class:

Pre-train once, apply all

LLM	Cora Full	WikiCS	ogbn-arxiv	CiteSeer
LLaMA2-7B	$76.66_{\pm 7.79}$	$79.00_{\pm 7.96}$	$65.33_{\pm 10.46}$	$54.35_{\pm 9.54}$
+ PT	$81.05_{\pm 7.77}$	$79.90_{\pm 7.69}$	$77.42_{\pm 10.48}$	$58.45_{\pm 8.61}$
LLaMA2-13B	$77.62_{\pm 8.67}$	$79.80_{\pm 7.30}$	$69.38_{\pm 8.83}$	$54.60_{\pm 8.79}$
+ PT	$81.95_{\pm 7.06}$	$80.45_{\pm 7.66}$	$77.75_{\pm 9.01}$	$57.30_{\pm 9.20}$
Vicuna-7B	$74.12_{\pm 6.47}$	$80.30_{\pm 7.02}$	$64.84_{\pm 9.38}$	$49.25_{\pm 6.72}$
+ PT	$80.77_{\pm 6.75}$	$80.10_{\pm 7.39}$	$76.95_{\pm 9.43}$	$52.25_{\pm 8.23}$
Vicuna-13B	$77.76_{\pm 8.58}$	$79.35_{\pm 7.98}$	$66.03_{\pm 9.34}$	$52.25_{\pm 6.39}$
+ PT	$81.38_{\pm 7.65}$	$79.25_{\pm 7.50}$	$75.65_{\pm 9.59}$	$53.00_{\pm 8.16}$
LLaMA3-8B	$79.22_{\pm 8.45}$	$78.40_{\pm 8.05}$	$70.37_{\pm 8.95}$	$61.25_{\pm 7.14}$
+ PT	$82.88_{\pm 8.09}$	$78.35_{\pm 7.61}$	$76.71_{\pm 10.20}$	$64.20_{\pm 7.39}$
GPT-4o-mini	$79.25_{\pm 8.42}$	$81.05_{\pm 6.80}$	$71.32_{\pm 9.13}$	$61.90_{\pm 7.22}$
+ PT	$83.04_{\pm 7.84}$	$\mathbf{81.90}_{\pm \mathbf{6.16}}$	$77.51_{\pm 9.58}$	$65.90_{\pm 7.04}$
GPT-4o	$\mathbf{81.40}_{\pm \mathbf{7.41}}$	$\mathbf{81.45}_{\pm \mathbf{7.10}}$	$\mathbf{72.75}_{\pm \mathbf{8.83}}$	$\mathbf{62.95}_{\pm \mathbf{6.61}}$
+ PT	$\mathbf{83.28}_{\pm \mathbf{7.06}}$	$81.60_{\pm 7.19}$	$\mathbf{78.85}_{\pm \mathbf{9.74}}$	$\mathbf{65.90}_{\pm \mathbf{7.03}}$

- **Larger** models perform better
- **Newer** architectures show advantages
- **Prompt tuning** provides consistent gains

Outline

- Introduction: Graphs & text-attributed graphs
- Jointly training graph and textual data
- Quantizing graphs into language tokens for LLMs
- **Conclusions**

Conclusions

- Text-attributed graphs contain rich semantics
- Graph structures and semantics can be jointly pre-trained
- Quantizing graphs is promising for integration with LLMs

Acknowledgement

Zhihao Wen, Yuan Fang. Augmenting Low-Resource Text Classification with Graph-Grounded Pre-training and Prompting. *SIGIR* 2023.

Yibo Li, Yuan Fang, Mengmei Zhang, Chuan Shi. Advancing Molecular Graph-Text Pre-training via Fine-grained Alignment. *KDD* 2025.

Jianyuan Bo, Hao Wu, Yuan Fang. Quantizing Text-attributed Graphs for Semantic-Structural Integration. *KDD* 2025.

Jiawei Liu, Cheng Yang, Zhiyuan Lu, Junze Chen, Yibo Li, Mengmei Zhang, Ting Bai, Yuan Fang, Lichao Sun, Philip S. Yu, Chuan Shi. Graph Foundation Models: Concepts, Opportunities and Challenges. *TPAMI* 2025.

The research is made possible with support from the School of Computing and Information Systems, Singapore Management University. Full publications, codes and data are available at http://www.yfang.site/.