School of Computing and Information Systems

August 3-7, 2025

KDD2025

# Quantizing Text-attributed Graphs for Semantic-Structural Integration

Jianyuan Bo[1], Hao Wu[2], Yuan Fang[1]
[1]Singapore Management University, Singapore
[2]Beijing Normal University
{jybo.2020, yfang}@smu.edu.sg, wuhao@bnu.edu.cn

Paper

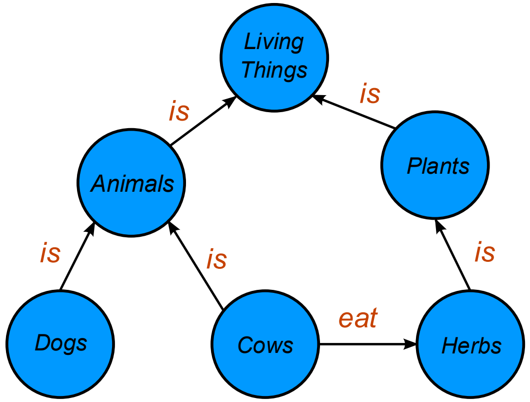
Code

# Overview

- Introduction
- Related Work
- Proposed Method
- Future Work

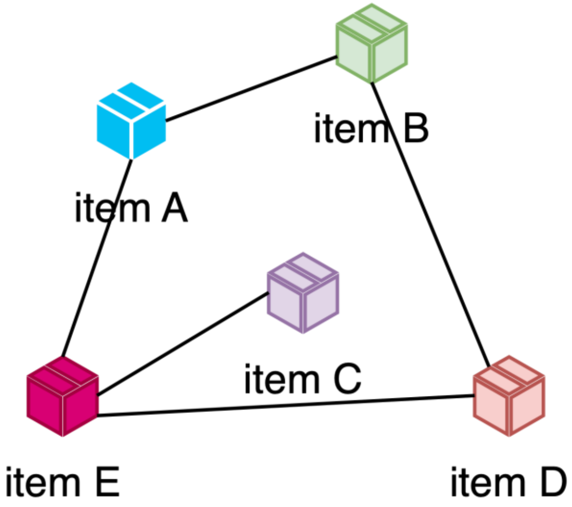
# Introduction

- Real-world graphs are **rich with textual information.**




https://commons.wikimedia.org/w/index.php?curid=37135596

**Knowledge Graph**

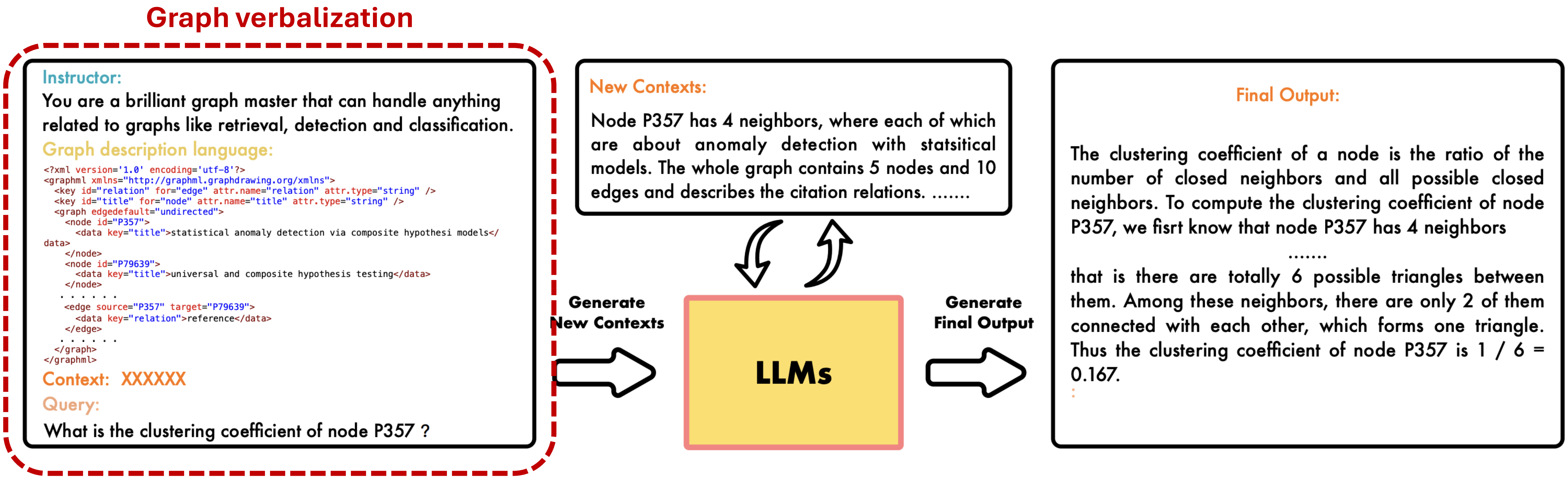



**Co-purchase Graph**

- **Graph + LLM** has great potential
  - Rich semantic understanding from LLMs
  - Few-shot and zero-shot transfer learning capabilities

**Text-Attributed Graphs (TAGs)**

# Related Work

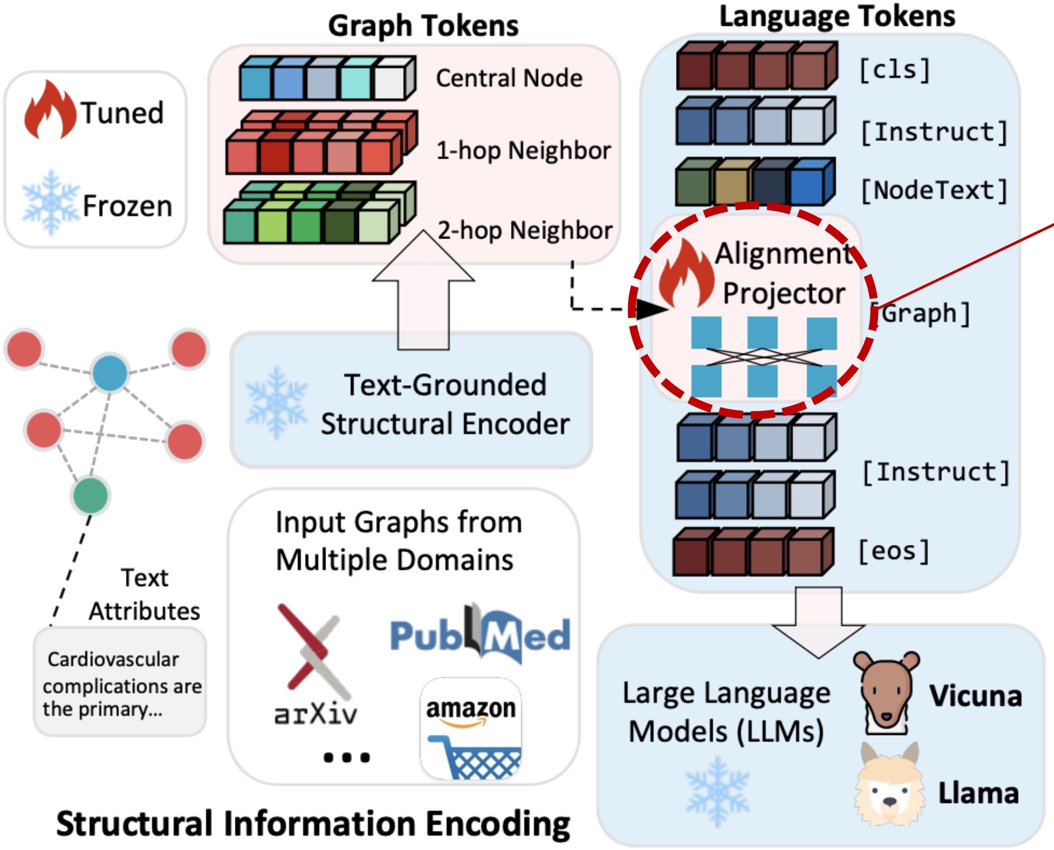



**GPT4Graph (Guo et al, 2024)**

Guo, J., Du, L., Liu, H., Zhou, M., He, X., & Han, S. (2023). Gpt4graph: Can large language models understand graph structured data? an empirical evaluation and benchmarking. *arXiv preprint arXiv:2305.15066*.

Related Work

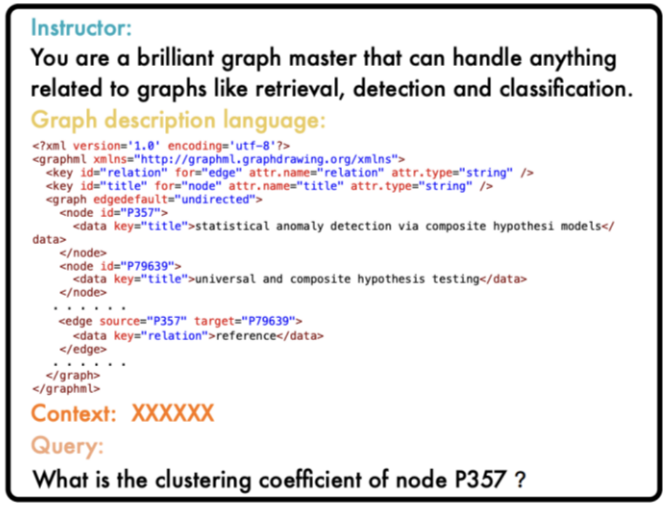
# Related Work




heavy alignment

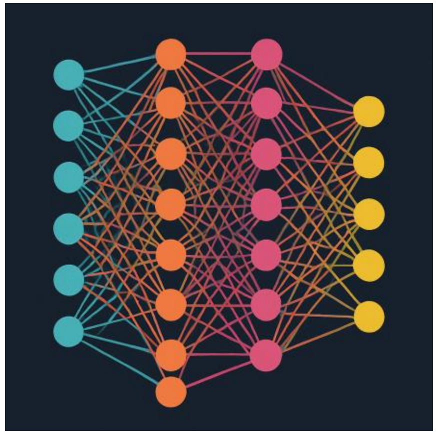
GraphGPT (Tang et al, 2024)

# Motivation: Semantic-Structural Gap



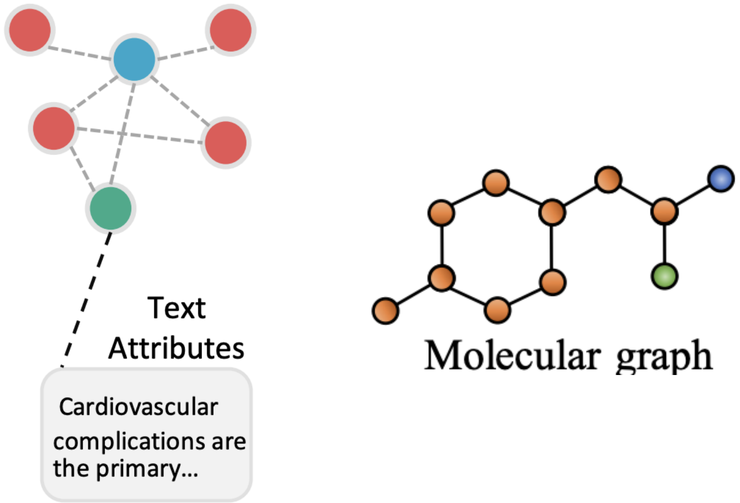

Graph Verbalization

**Structural information loss**



Projector-based Alignment

**High computational cost**



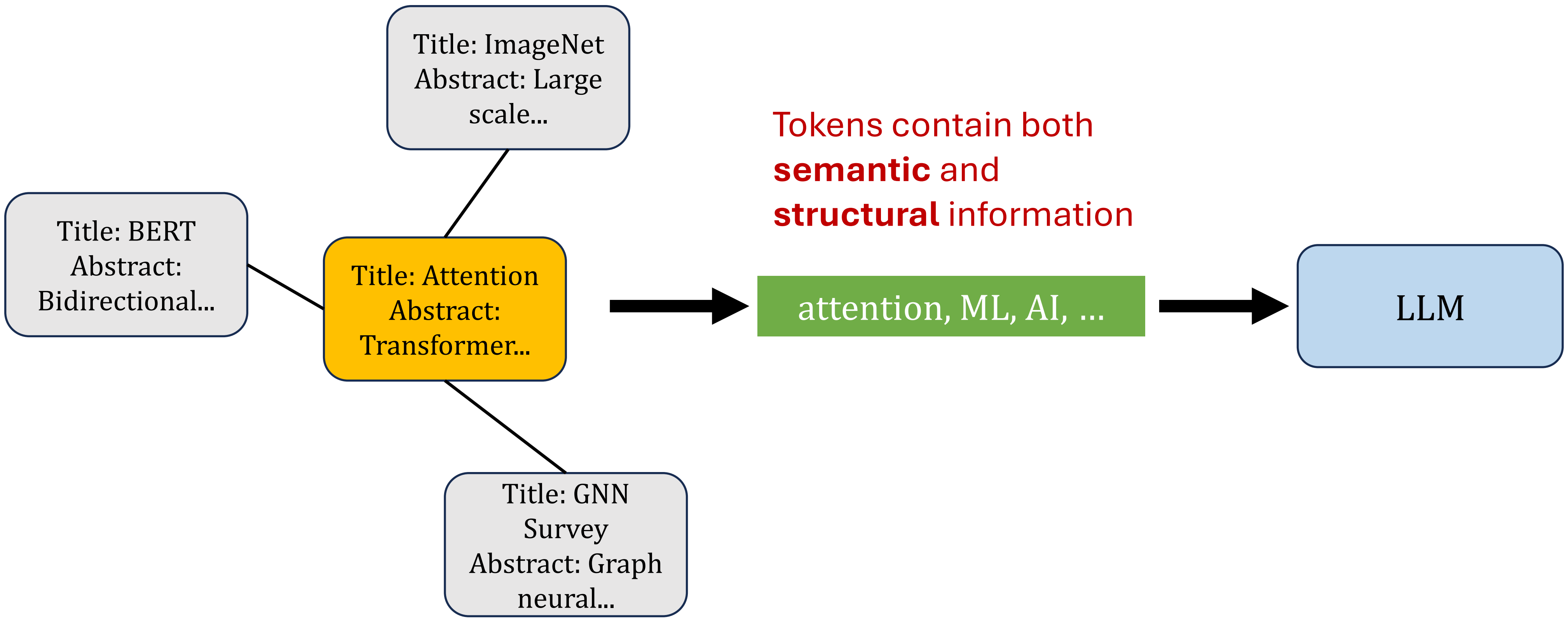

Transfer learning

**Poor generalization**

**Continuous vs. Discrete**

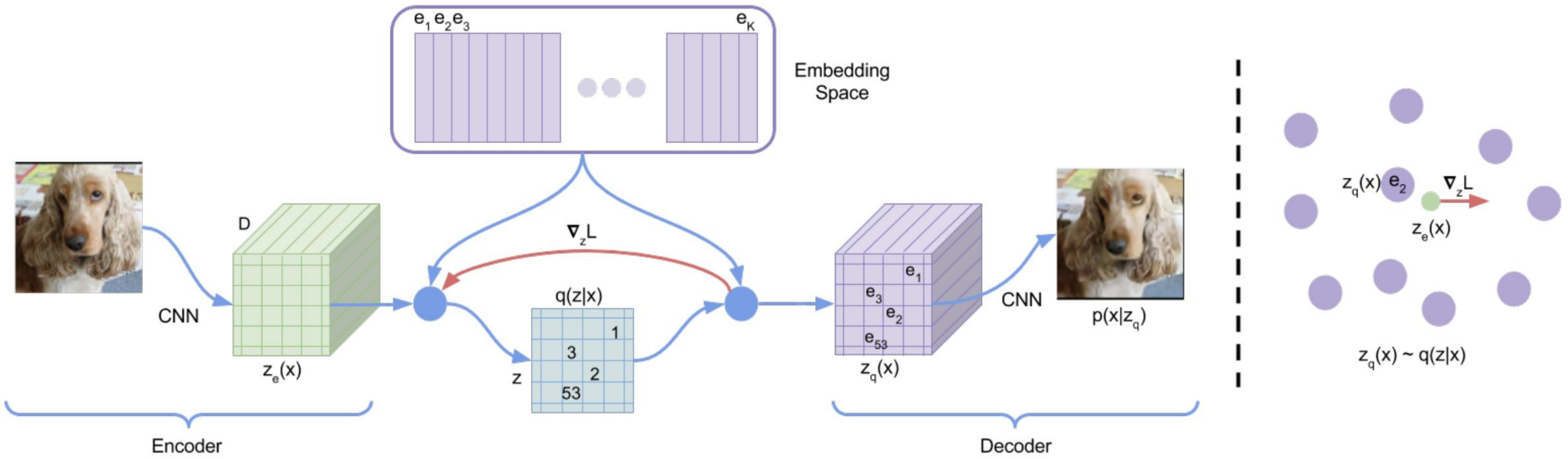
Graph embeddings ↔ LLM tokens

# Tokenization of Graph



SMU Singapore Management University | School of Computing and Information Systems

# Tokenization of Graph




**VQ-VAE (van den Oord et al., 2017)**

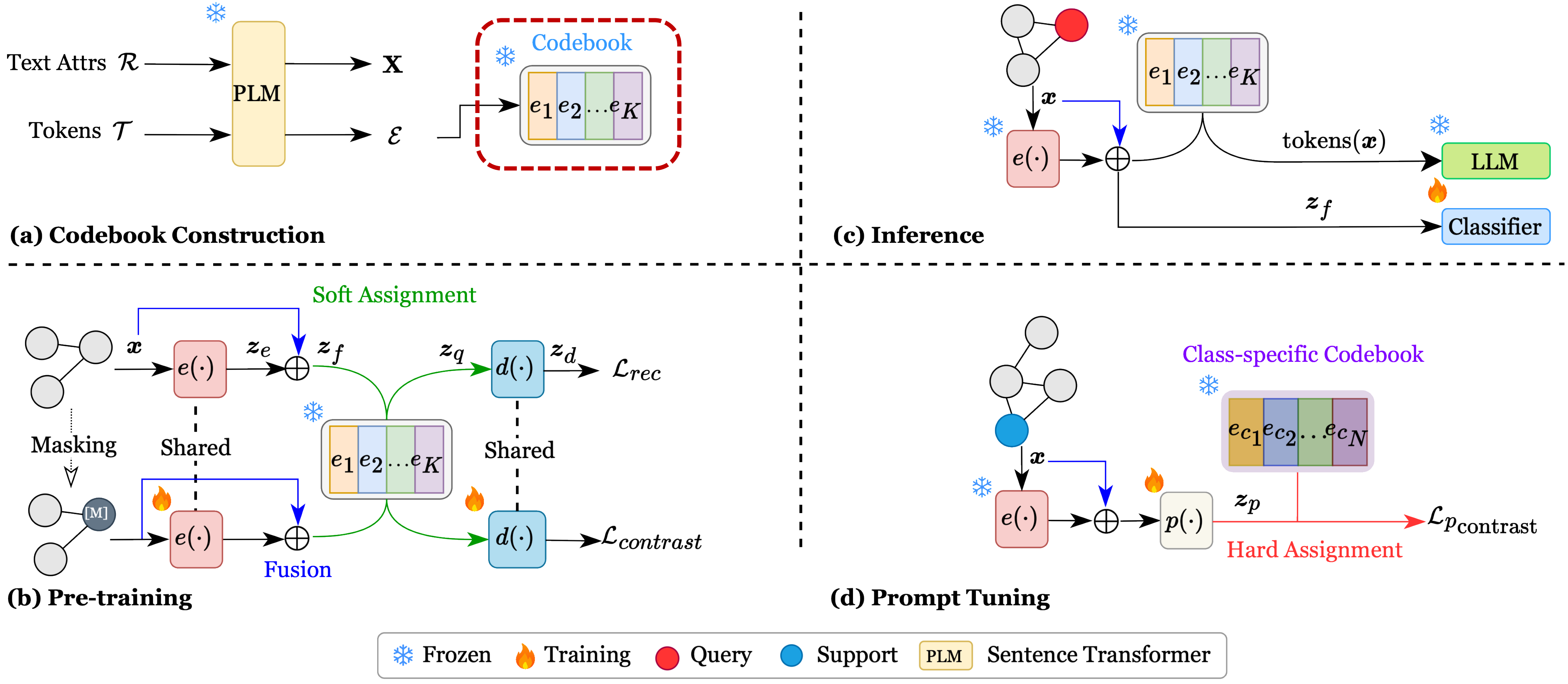
- **VQ-VAE**
  - Learn discrete codebook of semantic info
  - Map continuous features to discrete tokens
  - Enable generation and compression

- **Technical Challenges for Graphs**
  - No Natural Tokenization Structure
  - Hard Assignment Problems
  - Structure-Semantics Dilemma

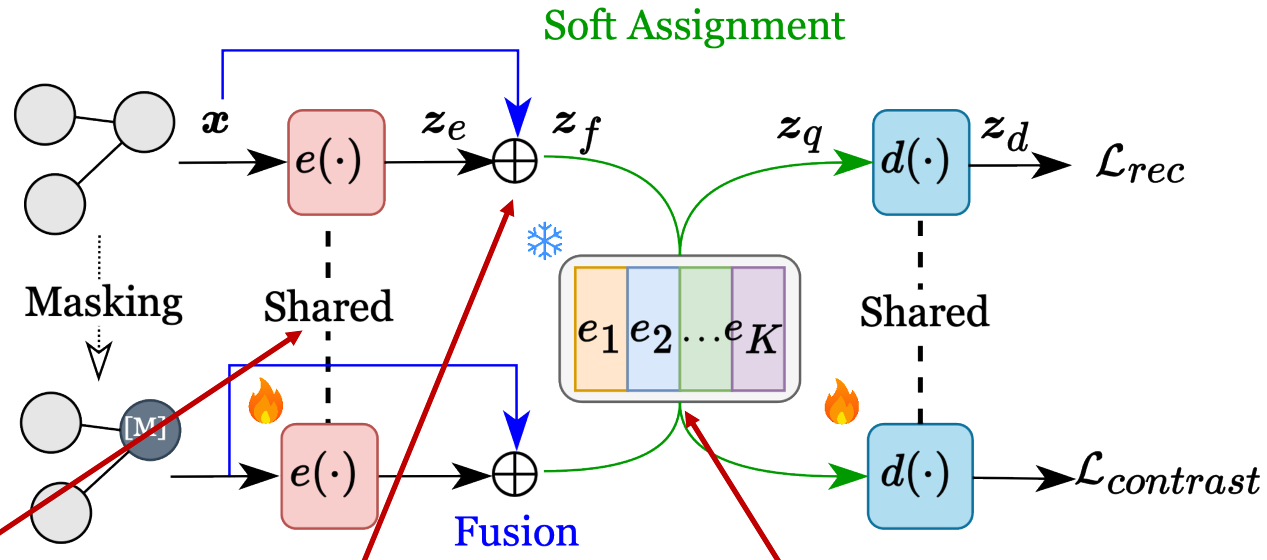


Van Den Oord, A., & Vinyals, O. (2017). Neural discrete representation learning. *Advances in neural information processing systems*, *30*.

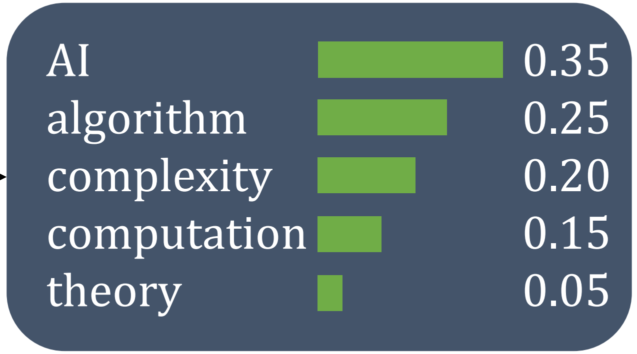
# Proposed Method



# Self-supervised Pre-training

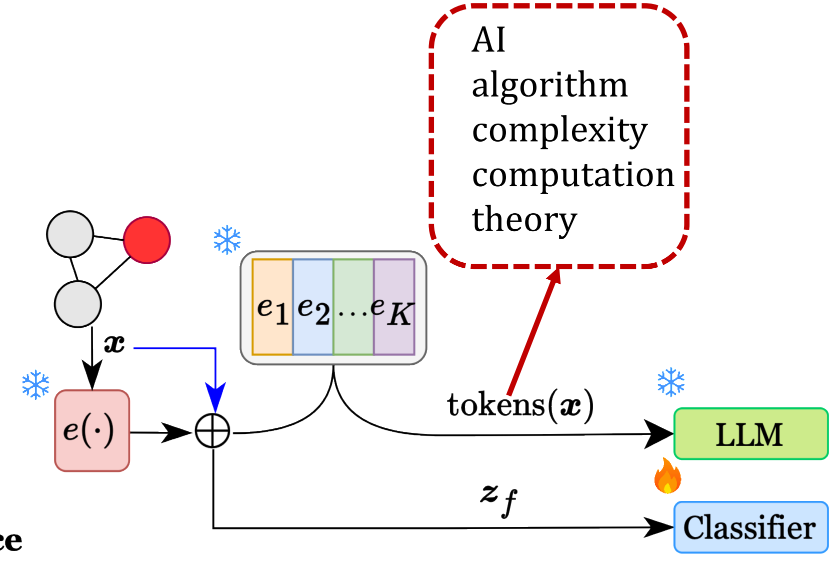



GNN Encoder (Text Features) → Semantic-Structural Fusion → Soft Assignment + KL Loss →



# Flexible Inference

- **With LLMs**
  - Extract **top-k tokens** from code distribution
  - Few-shot: Include **support examples** in prompt
  - Zero-shot: Direct LLM classification
- **Without LLMs**
  - **Linear probing** on frozen embeddings
  - Direct comparison with traditional methods
- **Graph Prompt Tuning**
  - Lightweight adaptation for **domain transfer**




(c) Inference

# Inference with LLM

**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:

Optional, remove for zero-shot inference

# Experiments

- Few-shot node classification
- Zero-shot node classification
- Ablation Studies
- Task Generalization
    - Link prediction
    - Edge classification
    - Subgraph classification

# Few-shot Node Classification

| Pre-train data | Method | LLM | Target data | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Cora | Cora Full | CiteSeer | PubMed | WikiCS | ogbn-arxiv | ogbn-products |
| Same as target | GCN | ✗ | 76.10±4.26 | 82.81±7.40 | 59.95±6.92 | 66.35±6.71 | 70.55±8.26 | 76.61±7.72 | 80.08±7.41 |
| | GAT | ✗ | 79.60±5.00 | 84.72±7.83 | 60.85±6.78 | 67.40±7.18 | 77.95±7.81 | 80.80±7.99 | 81.94±7.07 |
| No pre-train | Raw Text | ✓ | 63.40±9.07 | 71.66±7.84 | 62.10±5.35 | 85.00±5.48 | 77.15±6.92 | 54.74±9.21 | 87.58±5.48 |
| | Raw Feat + Quantization | ✓ | 54.85±6.28 | 73.74±8.10 | 56.40±5.48 | 48.30±7.78 | 73.40±8.19 | 63.75±9.98 | 65.84±9.96 |
| | Raw Feat + Linear Probing | ✗ | 70.25±7.22 | 81.29±7.47 | 63.00±6.72 | 68.30±6.28 | 78.05±7.47 | 83.05±7.40 | 77.53±7.16 |
| Cora Full | DGI | ✗ | 77.05±5.12 | 83.32±8.12 | 63.85±5.39 | 68.20±7.57 | 78.65±6.90 | 81.30±8.51 | 79.90±7.20 |
| | GraphMAE2 | ✗ | 77.70±6.92 | 84.74±7.42 | 65.25±5.84 | 66.35±6.09 | 80.95±4.96 | 80.04±8.15 | 73.93±7.57 |
| | GPPT | ✗ | 27.16±7.61 | 67.90±12.72 | 28.66±7.60 | 21.53±10.91 | 29.00±8.08 | 36.92±10.32 | 24.32±5.13 |
| | G2P2 | ✗ | 74.90±7.47 | 81.10±7.44 | 59.65±9.68 | 67.85±8.02 | 69.90±10.52 | 68.75±10.14 | 70.97±10.03 |
| | Prodigy | ✗ | 39.50±6.75 | 60.80±6.38 | 42.90±5.02 | 43.68±6.91 | 43.25±6.91 | 47.85±6.89 | 30.70±5.94 |
| | OFA | ✗ | 45.95±4.52 | 56.95±5.31 | 36.80±5.50 | 49.40±4.75 | 46.45±4.67 | 50.80±4.73 | 33.60±4.26 |
| | STAG | ✓ | 67.60±6.72 | 80.95±8.02 | 62.45±7.02 | 54.50±7.83 | 79.20±8.41 | 71.56±10.32 | 69.34±9.93 |
| | + Linear Probing | ✗ | 78.50±5.62 | 86.04±6.70 | **66.70±5.36** | **69.00±6.31** | **84.05±5.78** | 82.99±8.10 | 79.62±7.12 |
| | + Prompt Tuning | ✓ | 73.30±4.77 | 85.20±7.59 | 65.40±5.98 | 66.20±5.70 | 79.45±7.53 | 79.18±8.28 | 73.94±9.67 |
| | + Prompt Tuning* | ✗ | **78.65±5.93** | **86.66±7.67** | 65.80±7.03 | 68.25±6.80 | 83.55±5.94 | **83.57±8.30** | **80.48±6.86** |

# Zero-shot Node Classification

| Pre-train data | Method | LLM | Target data | | | |
|---|---|---|---|---|---|---|
| | | | Cora | Cora Full | WikiCS | ogbn-arxiv |
| No pre-train | Raw Feat + Q | ✓ | 47.10±5.98 | 60.33±10.88 | 70.40±8.88 | 25.48±5.54 |
| | Raw Feat + $\mathcal{C}$ | ✗ | 62.20±8.45 | 77.23±8.96 | 73.85±8.02 | 72.85±10.43 |
| Cora Full | G2P2 | ✗ | 60.45±7.58 | 64.29±11.56 | 50.25±8.43 | 19.66±6.38 |
| | OFA | ✗ | 20.30±2.93 | 23.85±3.58 | 21.45±3.99 | 17.60±3.74 |
| | STAG | ✓ | 48.05±6.15 | 62.63±11.70 | **76.25±8.48** | 26.01±7.52 |
| | STAG + $\mathcal{C}$ | ✗ | **66.55±7.48** | **82.90±9.52** | 75.15±7.81 | **74.23±9.35** |

**True zero-shot with no labeled data from source domain**

# Pretrain Once, Apply All (Few-shot setting)

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

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

# Ablation Studies

| Method | LLM Inference | | | Linear Probing | | |
|---|---|---|---|---|---|---|
| | Cora Full | WikiCS | ogbn-arxiv | Cora Full | WikiCS | ogbn-arxiv |
| Full Model | **80.95**$\pm$**8.02** | **79.20**$\pm$**8.41** | **71.56**$\pm$**10.32** | **86.04**$\pm$**6.70** | **84.05**$\pm$**5.78** | **82.99**$\pm$**8.10** |
| $\neg$ Fusion | 37.49$\pm$6.66 | 29.10$\pm$9.65 | 29.49$\pm$8.01 | 46.73$\pm$6.66 | 35.05$\pm$8.61 | 34.12$\pm$6.79 |
| $\neg\mathcal{L}_{KL}$ | 69.74$\pm$11.04 | 57.95$\pm$11.48 | 59.79$\pm$8.32 | 81.83$\pm$8.33 | 75.05$\pm$7.21 | 76.18$\pm$9.21 |
| $\neg$ Soft | 37.07$\pm$8.01 | 31.55$\pm$8.98 | 28.21$\pm$7.63 | 67.77$\pm$9.89 | 61.35$\pm$8.81 | 50.90$\pm$9.53 |

- All components contribute to performance
- **Feature fusion** is most critical

SMU Singapore Management University | School of Computing and Information Systems

# Task Generalization

## Zero-shot Link Prediction

| Method | Cora | ogbn-products |
|---|---|---|
| LLaGA | 87.35 | 92.99 |
| STAG | 63.00 | 92.65 |
| STAG (non-LLM) | **93.20** | **96.85** |

## N-way 5-shot Edge Classification

| Method | WN18RR | FB15K237 |
|---|---|---|
| OFA | 34.35 | 19.55 |
| STAG | 41.75 | 56.60 |
| STAG + Linear Probing | **58.30** | **74.80** |

## 5-way 5-shot Subgraph Classification

| Method | Cora | Cora Full | Arxiv |
|---|---|---|---|
| Raw Feat + Quantization | 67.75 | 78.32 | 65.18 |
| STAG | **69.60** | **79.25** | **68.41** |

# Conclusion

- **Soft Tokenization for TAGs**
  - Conducts vector quantization on TAGs
- **LLM-Agnostic Framework**
  - seamlessly integrates with any LLM architecture, or without LLM
- **Superior Performance and Efficiency**
  - outperforms existing methods across diverse domains
- **Few-shot and Zero-shot Transfer Learning**
  - achieves true zero-shot capability without source domain labels
  - advancing towards Graph Foundation Models

# Our TPAMI Position Paper on Graph Foundation Models (GFMs)

## Graph Foundation Models: Concepts, Opportunities and Challenges

Jiawei Liu*, Cheng Yang*, Zhiyuan Lu, Junze Chen, Yibo Li,
Mengmei Zhang, Ting Bai, Yuan Fang, Lichao Sun, Philip S. Yu, and Chuan Shi

**Abstract**—Foundation models have emerged as critical components in a variety of artificial intelligence applications, and showcase significant success in natural language processing and several other domains. Meanwhile, the field of graph machine learning is witnessing a paradigm transition from shallow methods to more sophisticated deep learning approaches. The capabilities of foundation models in generalization and adaptation motivate graph machine learning researchers to discuss the potential of developing a new graph learning paradigm. This paradigm envisions models that are pre-trained on extensive graph data and can be adapted for various graph tasks. Despite this burgeoning interest, there is a noticeable lack of clear definitions and systematic analyses pertaining to this new domain. To this end, this article introduces the concept of Graph Foundation Models (GFMs), and offers an exhaustive explanation of their key characteristics and underlying technologies. We proceed to classify the existing work related to GFMs into three distinct categories, based on their dependence on graph neural networks and large language models. In addition to providing a thorough review of the current state of GFMs, this article also outlooks potential avenues for future research in this rapidly evolving domain.

**Index Terms**—Graph Foundation Models, Large Language Models

北京郵電大學
Beijing University of Posts and Telecommunications

GAMMA
—图数据挖掘与机器学习实验室—

SMU
SINGAPORE MANAGEMENT UNIVERSITY

LEHIGH
UNIVERSITY.

UIC
UNIVERSITY OF
ILLINOIS CHICAGO

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

# Thank you!

Paper

Code