On the Intersection of Language and Graph Models

Chuxu Zhang

Associate Professor of Computer Science

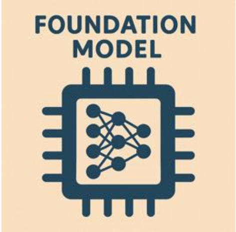
University of Connecticut (UConn)

ICDM ISIR-eCom Workshop, Nov. 2025

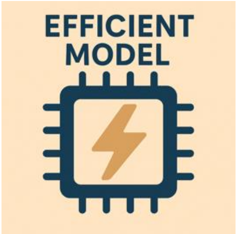
Research Overview: AI, ML, DS and Their Societal Applications

AI, Machine Learning, Data Science

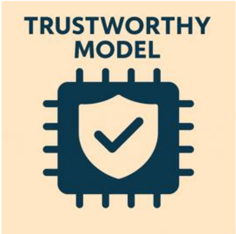
graph/network data (graph machine learning)
text/language data (large language models)



Graph Neural Networks, Large Language Models, Combination of GNN and LLM
HetGNN (KDD'19, 1900+ citations), GFM (NeurIPS'24), TANS (NAACL'25), GIT (ICML'25), GPM (ICML'25)

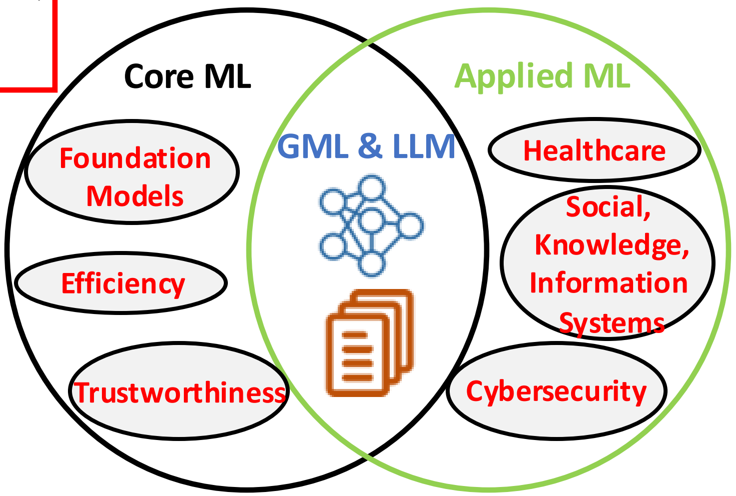


Data Efficiency: Self-supervised Learning, Few-shot Learning, Data Distillation
Model Efficiency: Prompt Learning, Model Distillation/Pruning
FSRL (AAAI'20, 300+ citations), TENT (KDD'22), ParetoGNN (ICLR'23), NOSMOG (ICLR'23), IAGPL (TMLR'25), MASS (ICML'25)



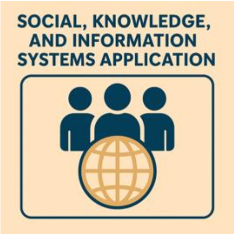
Robust Learning, Interpretable Learning
GAME (ICLR'23), G-FAME (WWW'23), CFExplainer (KDD'23), Dragon (ICLR'24), LIME (KDD'24)

This talk



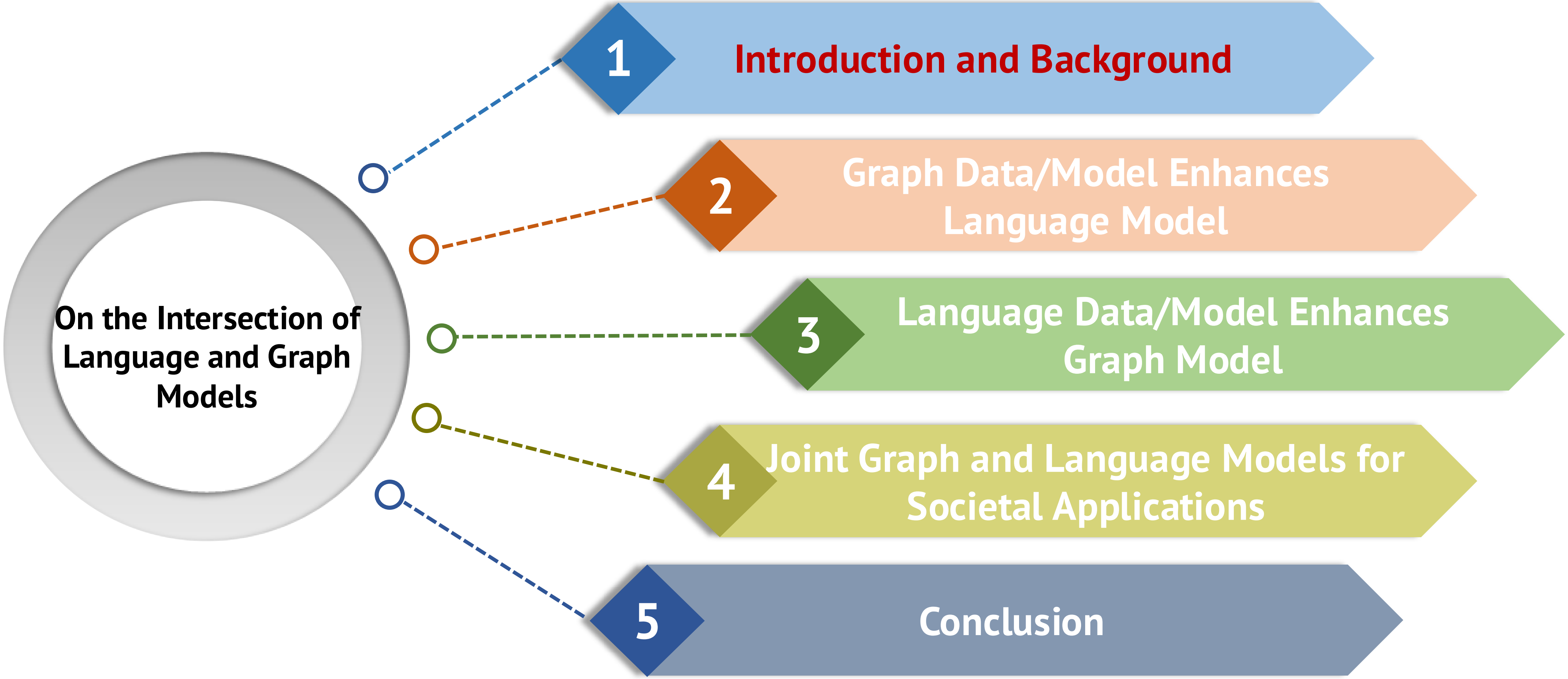
Healthcare: Combat the Drug Crisis:
RxNet (CIKM'21, Best Paper Award)
DHGNN (KDD'22, Best Paper Candidate)
Diet-ODIN (KDD'24)
Food & Nutrition, Drug Discovery:
MOPI-HFRS (KDD'25), MGNN (WWW'21), MOF-DDI (CIKM'23)

Anomaly/Malware Detection:
MSCRED (AAAI'19, 1100+ citations), MiST (WWW'19, Best Paper Candidate)
Rep2Vec (KDD'22), MetaHGNN (IJCAI'21)
Malicious Activity Detection:
MetaHG (NeurIPS'21), GraphBERT (ICDM'22)
LLM-HetGDT (ACL'25)

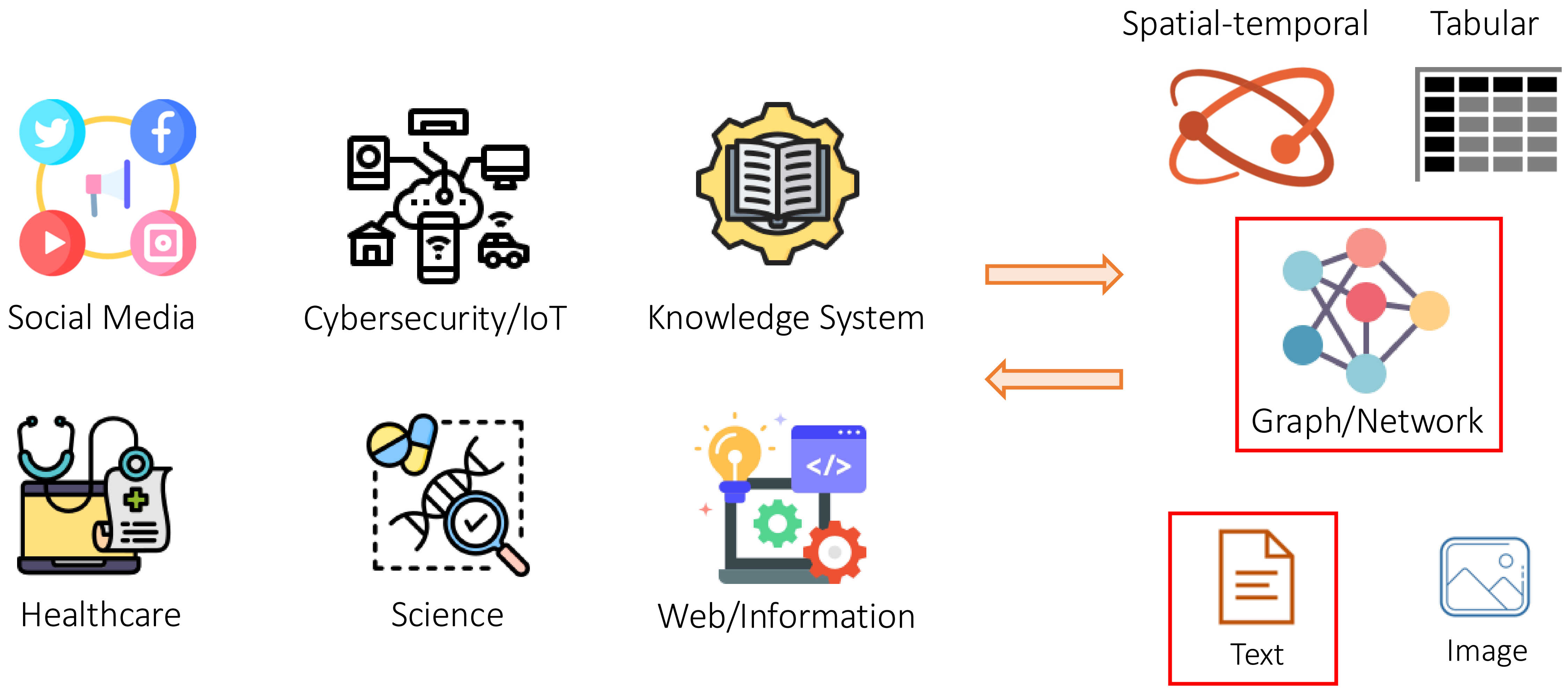


Knowledge Reasoning and QA:
FIRE (EMNLP'20), Grape (EMNLP'22)
SGCL (CIKM'22), NGQA (ACL'25)
Recommender Systems:
SHT (WWW'23), MMSSL (KDD'22), RecipeRec (IJCAI'22)

Outline


On the Intersection of Language and Graph Models
1
Introduction and Background
2
Graph Data/Model Enhances Language Model
3
Language Data/Model Enhances Graph Model
4
Joint Graph and Language Models for Societal Applications
5
Conclusion

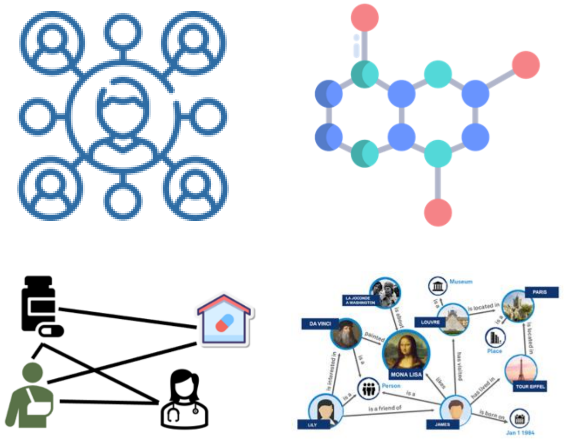
Introduction: Various Data in Real-world Applications



Introduction: Graph (Network) and Text (Language) Data

Graph/Network

Node and Edge

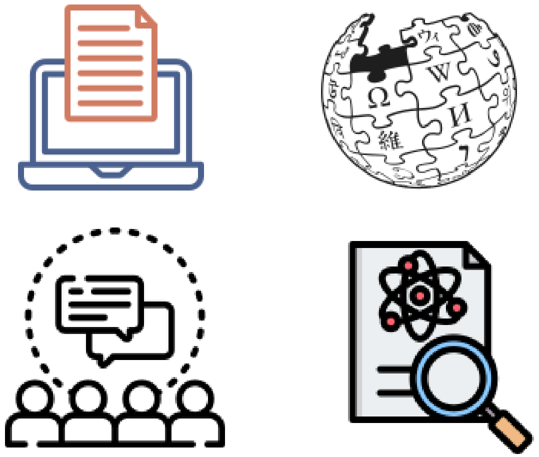


Structured Data
Encode Relational Information

Complement

Text/Language

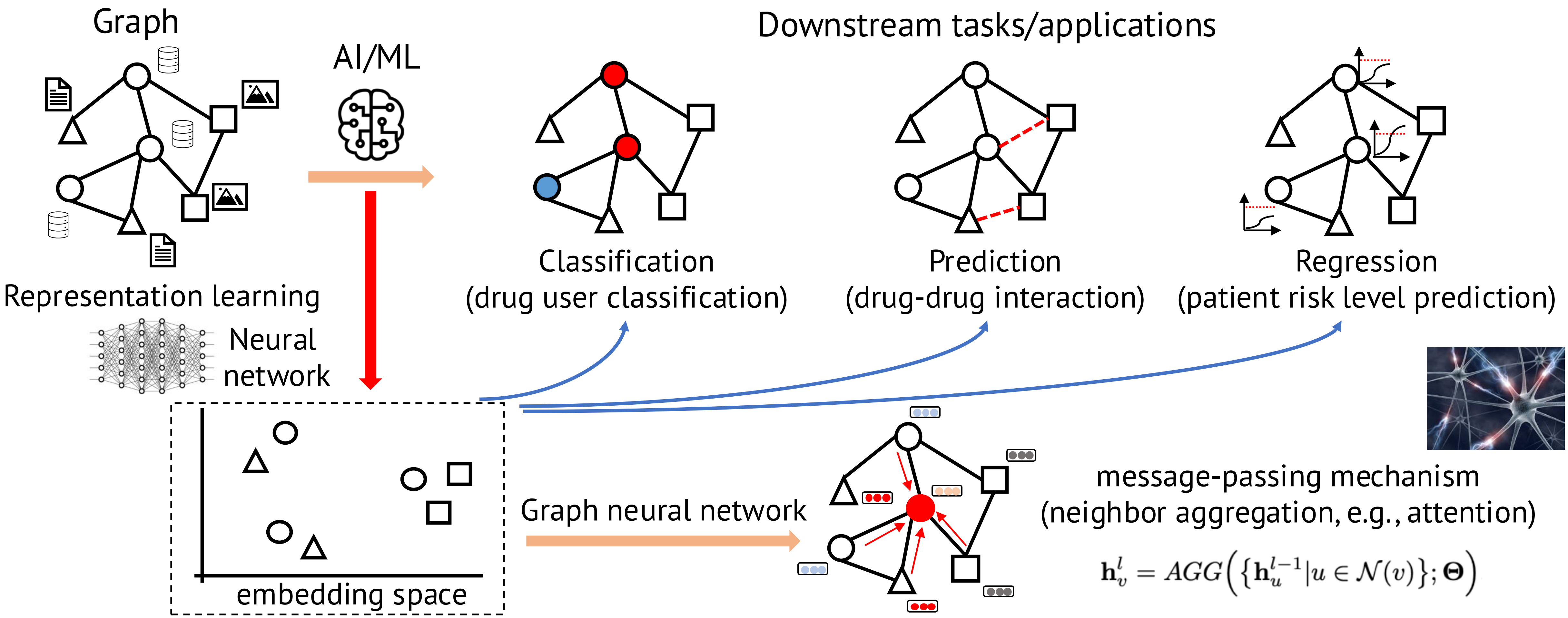
Token/Word



Unstructured Data
Encode Semantic Information

Introduction: Graph Neural Networks (GNNs)

- **GNNs learn representations of nodes by iteratively transforming and aggregating/propagating the features from their neighborhoods**

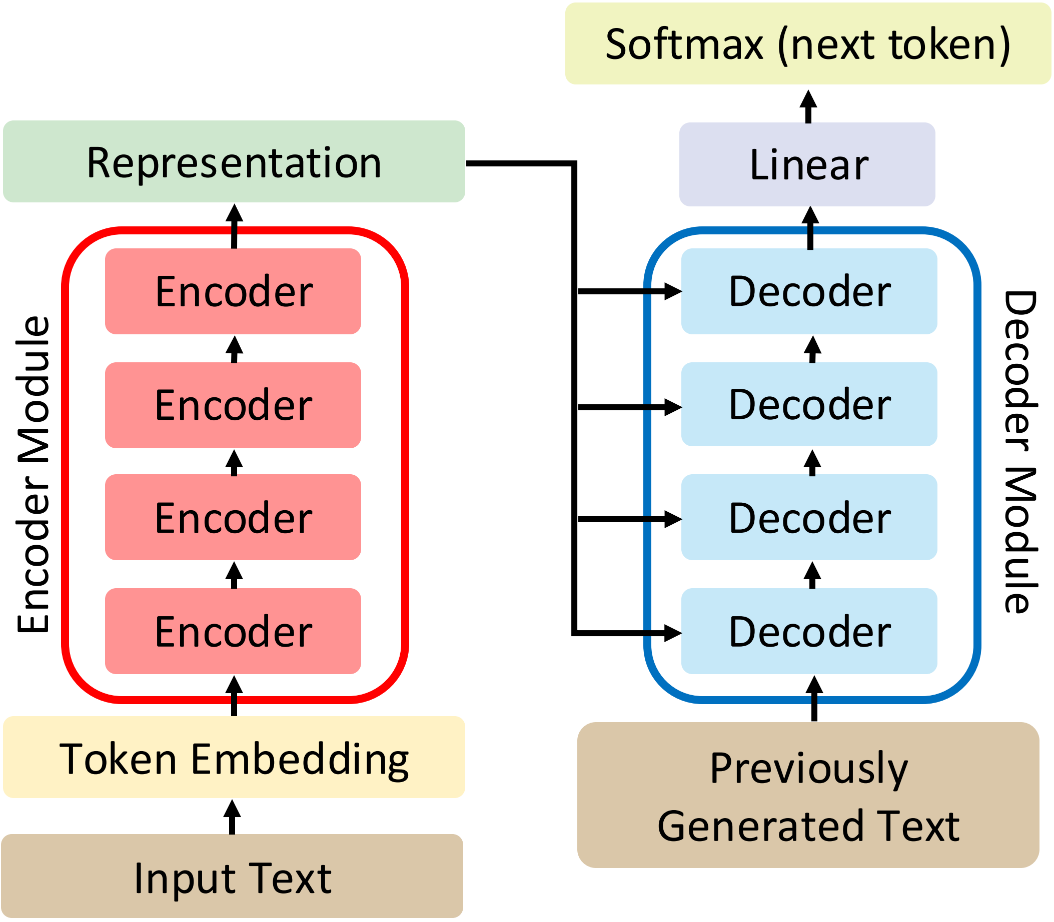


Introduction: Graph Neural Networks (GNNs)

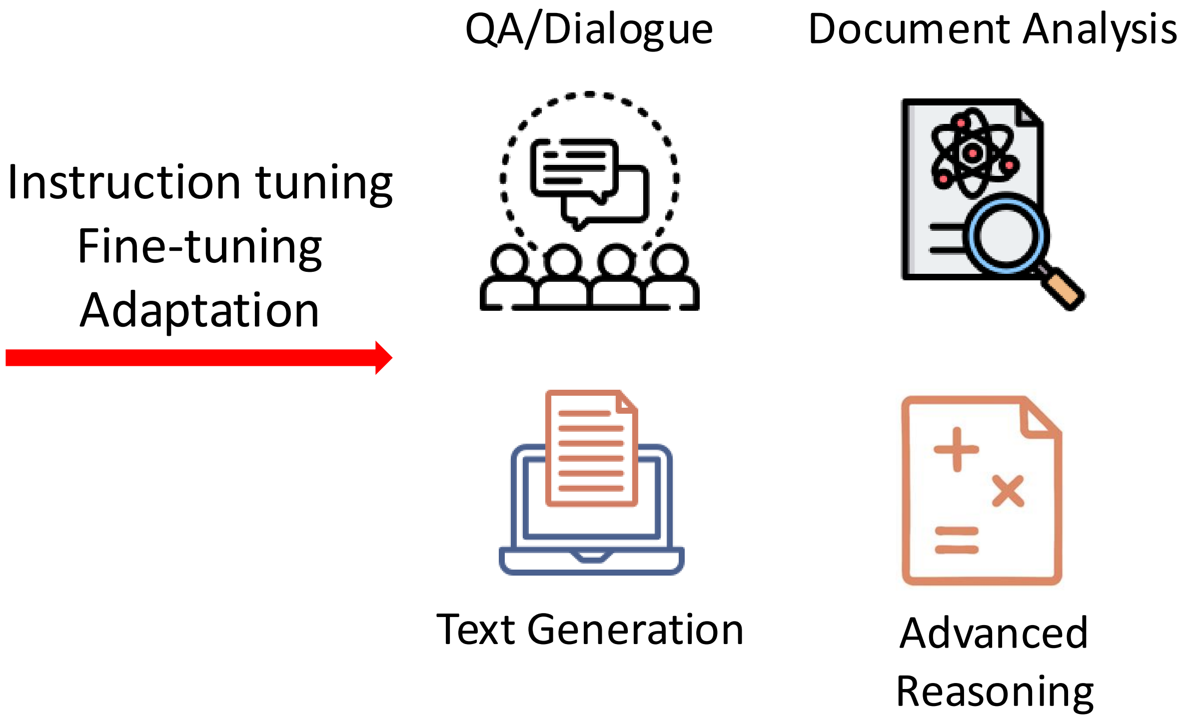
- **GCN (ICLR'16)**
  - Convolution Aggregation
- **GSAGE (NeurIPS'17)**
  - Pooling/Recurrent Aggregation
- **GAT (ICLR'18)**
  - Attention Aggregation
- **HetGNN (PhD work, KDD'19)**
  - The first GNN on Heterogeneous Graphs
  - 1900+ Citations, 2024 ICBS Frontiers of Science Award
- **Our recent works: GFT (NeurIPS'24), GIT (ICML'25), GPM (ICML'25), G^{2}PM (NeurIPS'25)**
  - Graph Foundation Models Across Datasets/Tasks/Domains

Introduction: Large Language Model (LLMs)

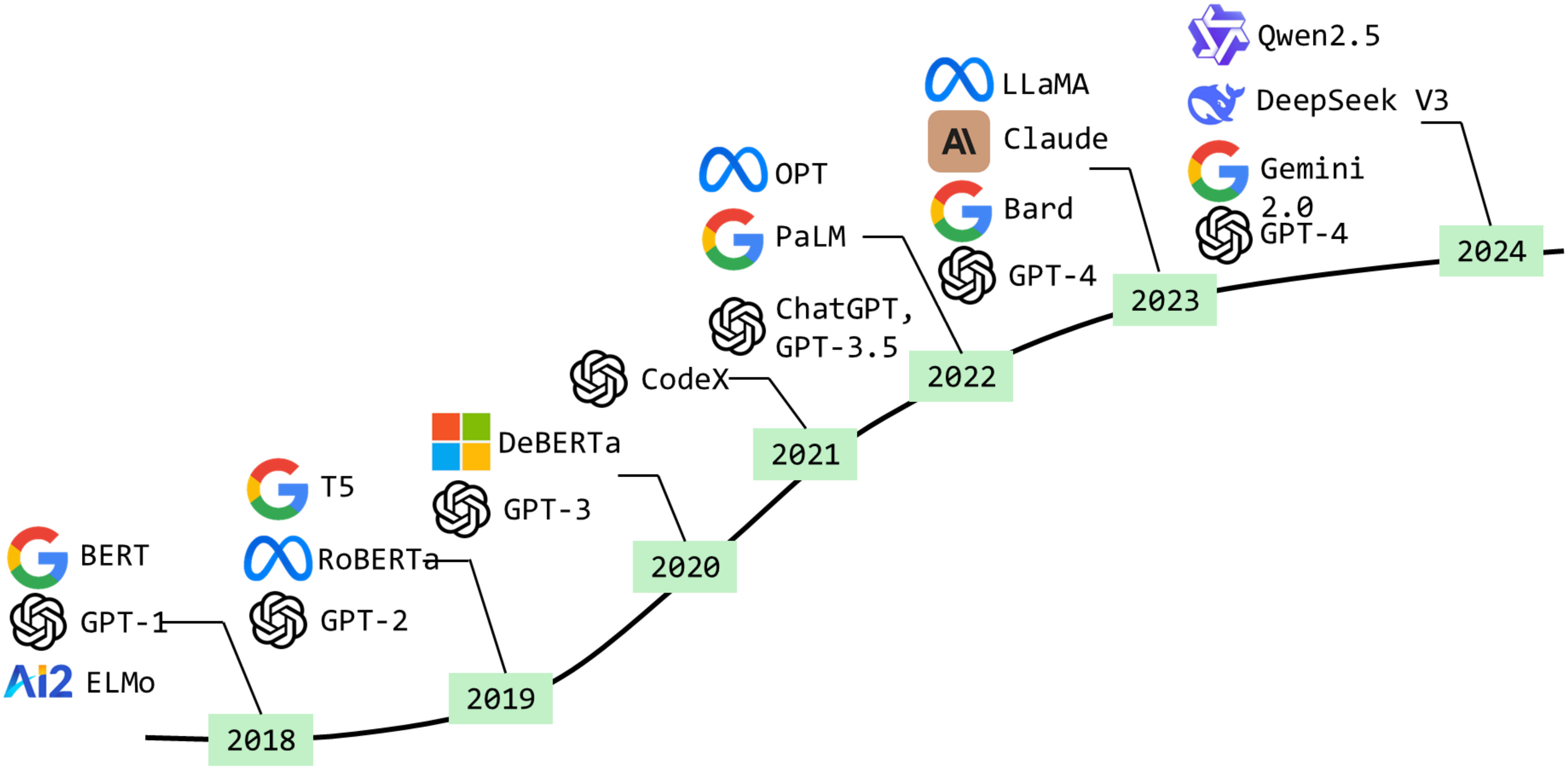
- **Pretrained Architecture**



Transformer



Introduction: Large Language Model (LLMs)

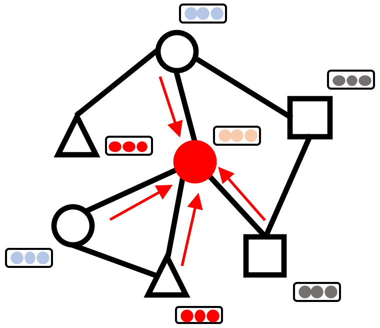


Introduction: Graph Model and Language Model

Graph/Network

Structured Data

Encode Relational Information



GNN

Complement

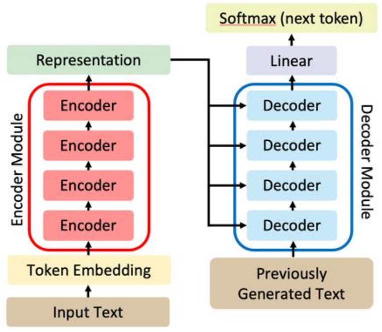
Enhance

Complement

Text/Language

Unstructured Data

Encode Semantic Information



LLM

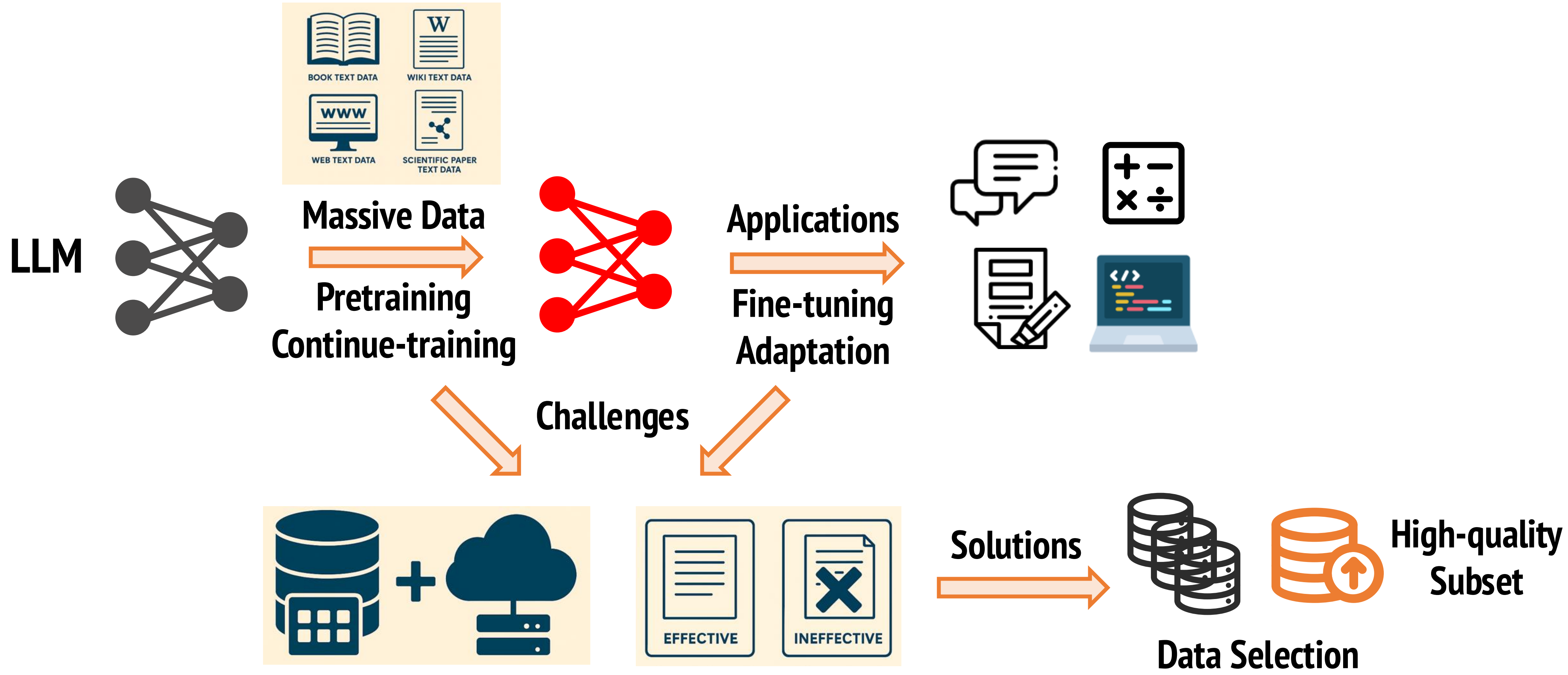
Outline

On the Intersection of Language and Graph Models

1. Introduction and Background
2. Graph Data/Model Enhances Language Model
3. Language Data/Model Enhances Graph Model
4. Joint Graph and Language Models for Societal Applications
5. Conclusion

Graph Data/Model Enhances Language Model: MASS

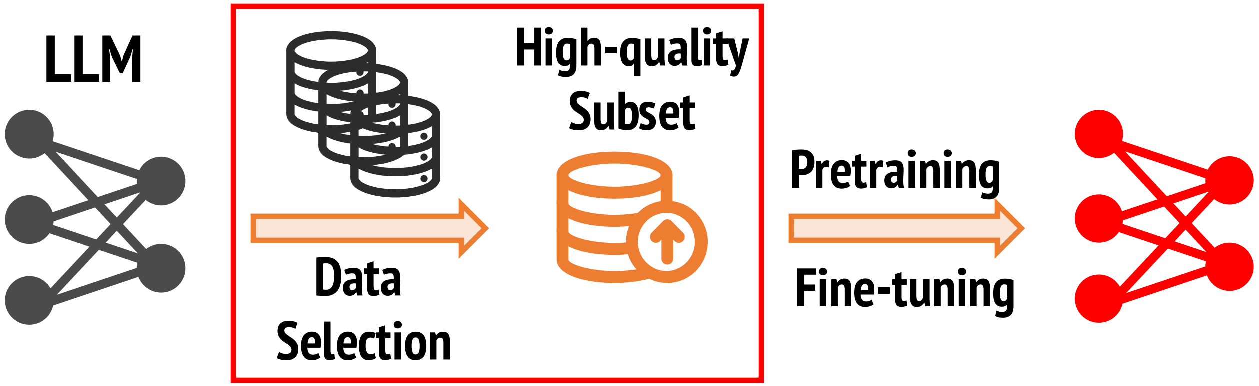
- **Data Selection for Pretraining LLMs: Improve Training Efficiency and Effectiveness**



MASS: MAthematical Data Selection via Skill Graphs for Pretraining Large Language Models, ICML'25

Graph Data/Model Enhances Language Model: MASS

- Data Selection for Pretraining LLMs: Improve Training Efficiency and Effectiveness



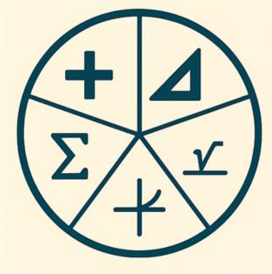
Skill Extraction Example

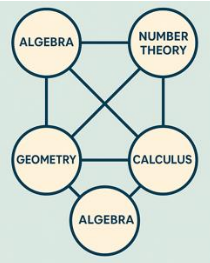
CONTENT: Find the greatest value of x such that $\frac{x^2-x-90}{x-9} = \frac{2}{x+7}$. The expression when simplified by factoring the numerator transforms into: $\frac{(x-9)(x+10)}{x-9} = \frac{2}{x+7}$. Canceling the $(x-9)$ factor on both sides, provided $x \neq 9$, we get: $x+10 = \frac{2}{x+7}$. Multiplying both sides by $(x+7)$ to eliminate the fraction yields: $(x+10)(x+7) = 2$. Expanding and rearranging this equation results in: $x^2+17x+70 = 2 \implies x^2+17x+68 = 0$. Factoring the quadratic gives: $(x+4)(x+13) = 0$. The solutions to this equation are $x=-4$ and $x=-13$. The greatest of these solutions is $\boxed{-4}$.

SKILLS: Equation solving, Factoring polynomials, Fraction manipulation, Quadratic equations, Root identification, Expression simplification, Algebraic transformation, Polynomial division, Inequality consideration, Solution verification

Prior methods: e.g., RHO-1 [NeurIPS'24], AutoDS [ICLR'24]
Limitation: concentrate on general domains while neglecting the underlying knowledge and their interrelations of specific domain data, such as mathematical skills for advanced reasoning capabilities.

Assumption: a data point reflecting (1) more important math skills or (2) more compositional information of math skills should receive a higher quality score.





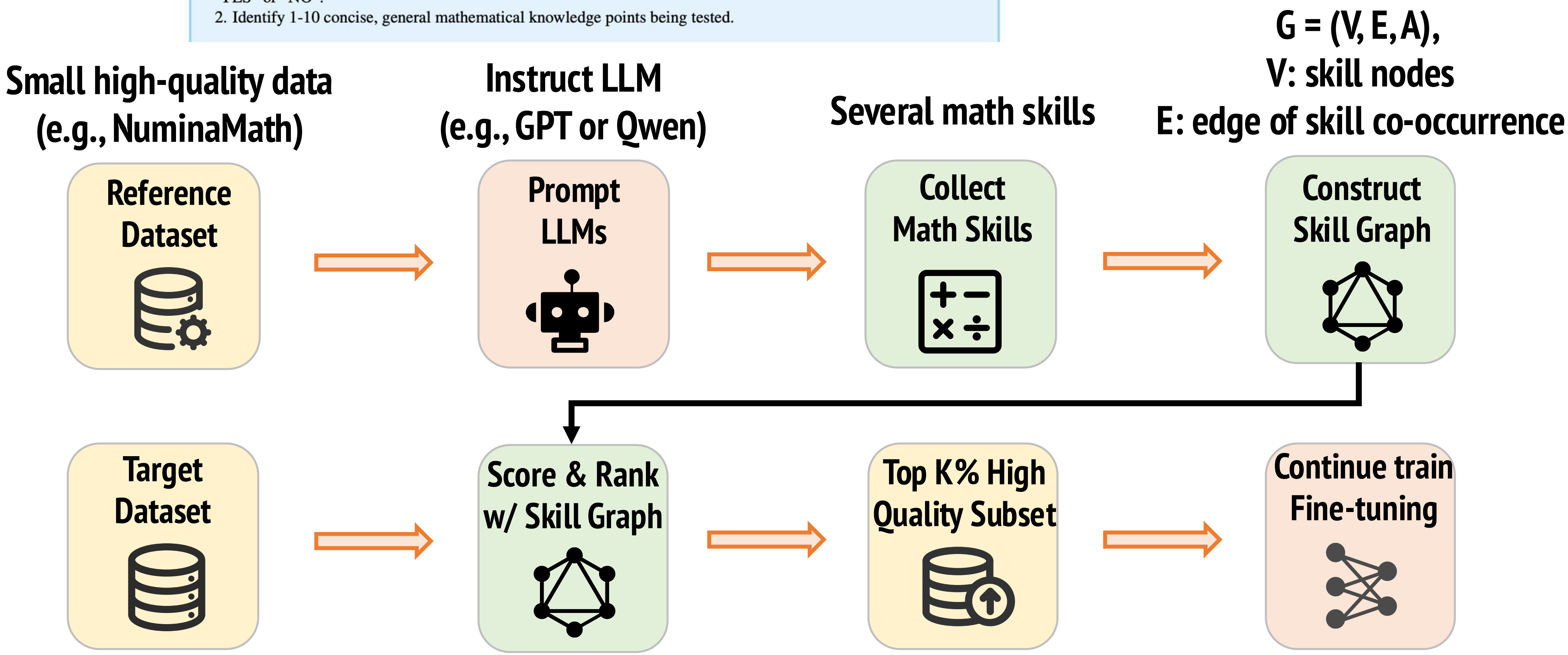
MASS: MAthematical Data Selection via Skill Graphs for Pretraining Large Language Models, ICML'25

Graph Data/Model Enhances Language Model: MASS

Prompt Template to Extract Skills

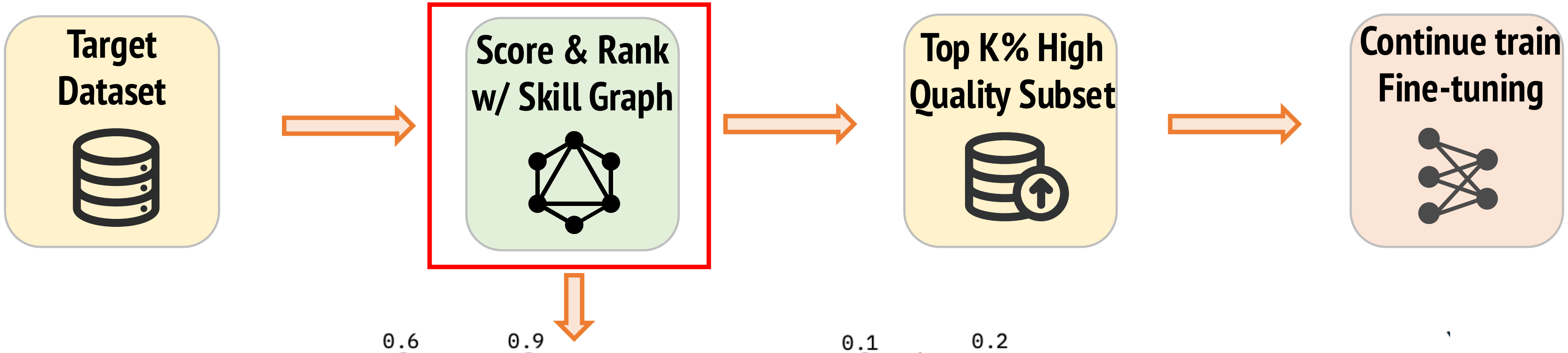
Please assume the role of a math teacher and analyze the provided question with the following steps:

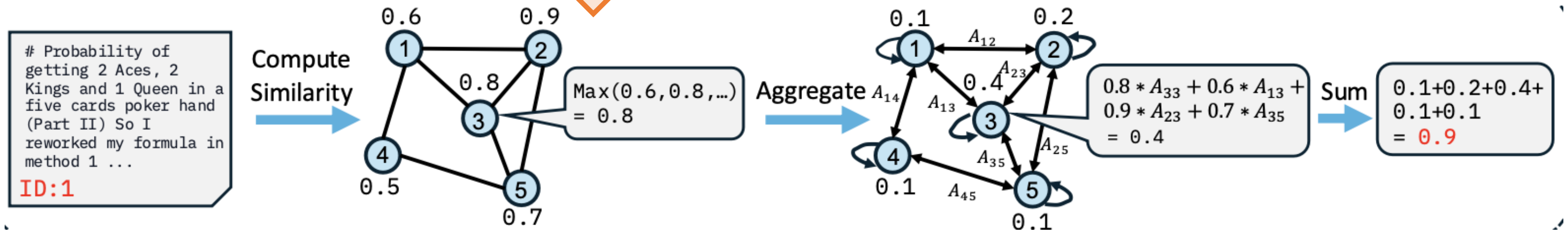
1. Determine if the text involves mathematical knowledge, reasoning, or problem-solving skills. Respond with "YES" or "NO".
2. Identify 1-10 concise, general mathematical knowledge points being tested.



MASS: MAthematical Data Selection via Skill Graphs for Pretraining Large Language Models, ICML'25

Graph Data/Model Enhances Language Model: MASS





$$\mathbf{A}_{i,i} = \sigma(v_i^{cnt}, T) = \frac{\exp\left(\frac{v_i^{cnt}}{T}\right)}{\sum_{j=1}^{|V|} \exp\left(\frac{v_j^{cnt}}{T}\right)}$$

$$\mathbf{A}_{i,j} = \sigma(e_{ij}^{cnt}, T) = \frac{\exp\left(\frac{e_{ij}^{cnt}}{T}\right)}{\sum_{(e_{ij} \in E)} \exp\left(\frac{e_{ij}^{cnt}}{T}\right)}$$

$$\mathrm{sim}(x_i, v_j) = \max_{k \in v_j^{ids}} \cos(\mathrm{Emb}(x_i), \mathrm{Emb}(d_k)).$$

$$\mathrm{score}(x) = \sum_{j=1}^{|V|} \mathrm{sim}_{agg}(x, v_j)$$

$$= \sum_{j=1}^{|V|} \mathbf{A}_{j,j} \mathrm{sim}(x, v_j) + \sum_{j=1}^{|V|} \sum_{v_k \in \mathcal{N}(v_j)} \mathbf{A}_{j,k} \mathrm{sim}(x, v_k)$$

Why it works

(1) more important math skills

(2) more compositional information on math skills

MASS: MAthematical Data Selection via Skill Graphs for Pretraining Large Language Models, ICML'25

Graph Data/Model Enhances Language Model: MASS

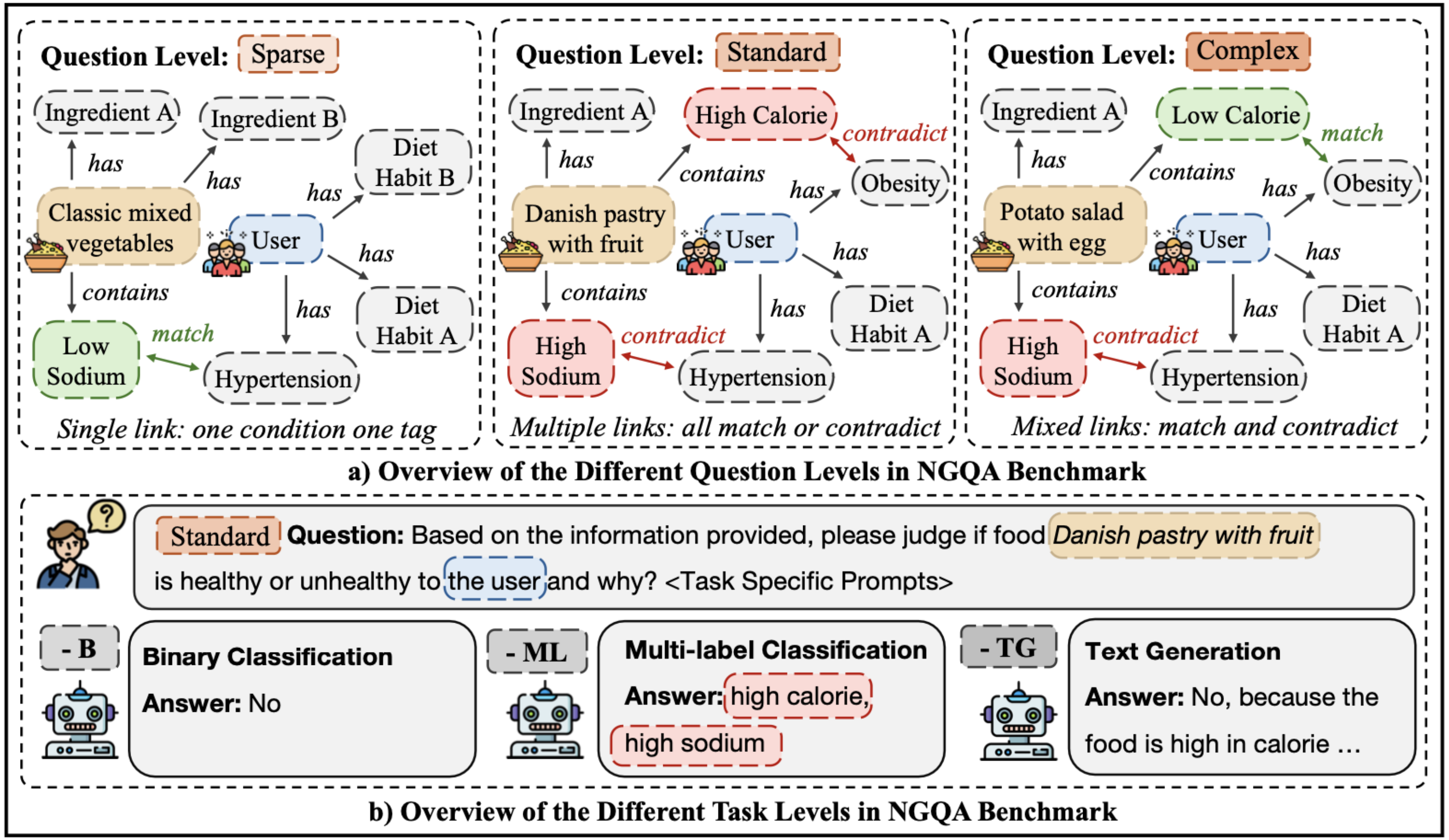
Table 2. The main experimental results. `TinyLlama-1.1B` and `Mistral-7B` are continuously pretrained using both the original and selected subsets of OpenWebMath, OpenWebMath-pro, and Jiuzhang3.0. The **bolded** entries indicate the best results within each setting. * indicates that results are from ProX (Zhou et al., 2024a)

Dataset	Method	Unique Tokens	Trained Tokens	GSM8K	MATH	SVAMP	ASDiv	MAWPS	TAB	MQA	MMLU STEM	SAT MATH	Avg.
`TinyLlama-1.1B`													
w/o continual pretraining				2.7	2.8	10.9	17.9	20.5	12.5	14.0	16.3	21.9	13.3
OpenWebMath	-	14.6B	14.6B	5.2	3.0	20.7	31.4	41.0	14.6	10.1	19.5	**37.5**	20.3
	RULE*	6.5B	15B	4.5	2.8	17.5	29.4	39.3	15.1	12.4	19.4	25.0	18.4
	RHO-1*	14.6B	9B	7.1	**5.0**	23.5	41.2	53.8	-	**18.0**	-	-	-
	ProX	5.1B	14.6B	8.6	3.0	23.8	40.2	51.6	19.6	14.9	**26.1**	25.0	23.6
	DSIR	4.9B	14.6B	5.5	2.6	24.1	37.8	54.3	16.9	12.1	25.4	22.3	22.1
	AutoDS	4.9B	14.6B	7.3	2.4	22.9	39.2	52.7	18.4	13.8	23.2	24.1	22.7
	MASS	4.9B	14.6B	**9.0**	4.4	**24.9**	**41.4**	**54.8**	**21.5**	13.9	20.3	25.0	**23.9**
OpenWebMath -pro	-	5.1B	14.6B	8.6	3.0	23.8	40.2	51.6	19.6	14.9	26.1	25.0	23.6
	DSIR	3B	14.6B	8.8	3.2	**24.1**	41.5	53.1	18.9	14.3	**27.6**	27.5	24.4
	AutoDS	3B	14.6B	9.1	4.5	22.4	40.8	54.3	23.2	13.1	26.5	28.0	24.7
	MASS	3B	14.6B	**10.2**	**5.8**	23.8	**42.3**	**57.9**	**25.3**	**15.3**	27.0	**34.4**	**26.9**
Jiuzhang3.0	-	3.4B	6.8B	22.3	19.0	46.4	60.1	73.2	29.6	19.1	**24.0**	34.4	36.4
	DSIR	2.4B	6.8B	24.5	21.3	48.2	63.9	74.4	28.8	19.2	22.1	33.6	37.3
	AutoDS	2.4B	6.8B	26.7	20.8	51.3	66.7	73.5	31.1	19.3	22.4	32.8	38.3
	MASS	2.4B	6.8B	**30.1**	**24.8**	**52.5**	**69.1**	**80.7**	**32.9**	**20.4**	22.7	34.4	**40.8**
`Mistral-7B`													
w/o continual pretraining				41.1	10.6	64.9	68.5	87.3	54.8	33.9	49.9	65.6	53.0
OpenWebMath	-	14.4B	9.6B	44.5	19.0	60.6	68.4	87.8	50.5	44.5	50.9	56.2	53.6
	MASS	4.8B	9.6B	**47.7**	**23.2**	**64.6**	**74.7**	**90.5**	**55.7**	**50.7**	**52.6**	**65.6**	**58.4**
OpenWebMath -pro	-	5.1B	5.1B	47.1	21.8	63.2	73.7	89.5	**58.2**	42.6	52.2	56.2	56.1
	MASS	3B	5.1B	**53.2**	**25.6**	**67.0**	**76.8**	**90.4**	57.6	**51.8**	**54.5**	**81.2**	**62.0**
Jiuzhang3.0	-	3.8B	3.8B	66.4	39.4	82.9	**85.9**	90.8	35.3	61.8	40.1	50.0	61.4
	MASS	2.7B	3.8B	**70.0**	**43.8**	**84.3**	85.7	**93.7**	**35.7**	**63.5**	**46.9**	**65.6**	**65.5**

MASS: MAthematical Data Selection via Skill Graphs for Pretraining Large Language Models, ICML'25

Graph Data/Model Enhances Language Model: More Study

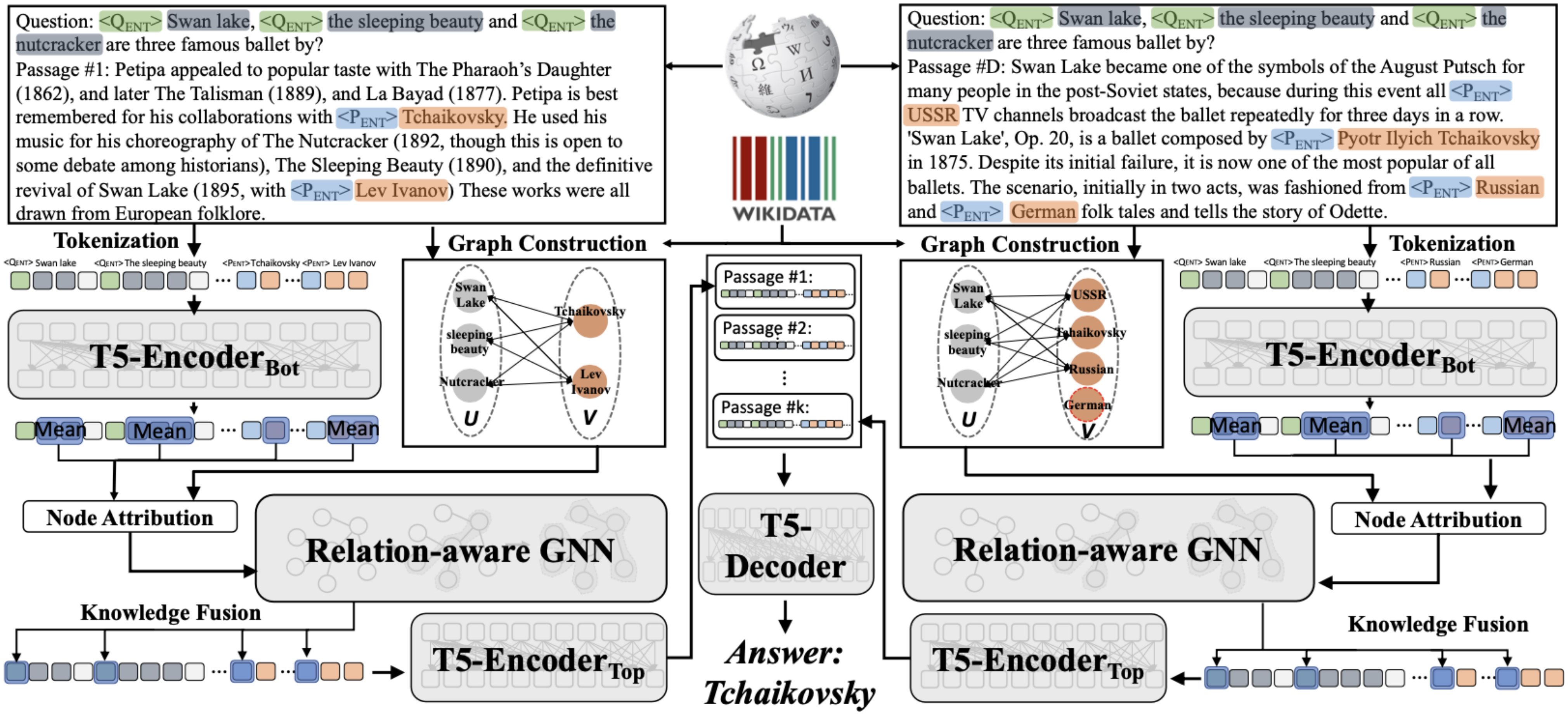
- **QA: Knowledge Graph as Retrieval Augmentation for LLMs**



NGQA: A Nutritional Graph Question Answering Benchmark for Personalized Health-aware Nutritional Reasoning, ACL 2025

Graph Data/Model Enhances Language Model: More Study

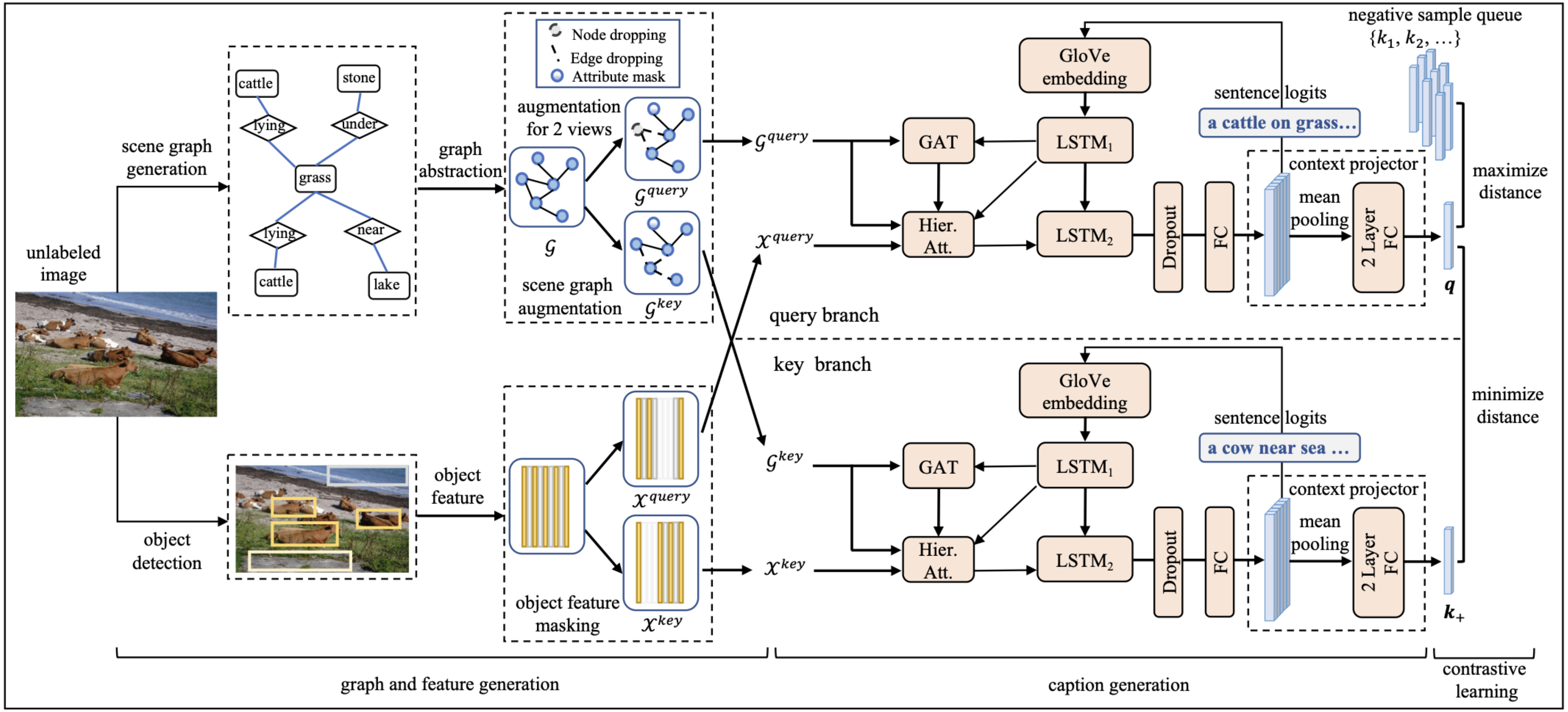
- **QA: Graph as Augmented Information for LLMs**



Knowledge Graph Enhanced Passage Reader for Open-domain Question Answering, EMNLP 2022

Graph Data/Model Enhances Language Model: More Study

- **Text Generation: Graph as Augmented Data for LLMs**



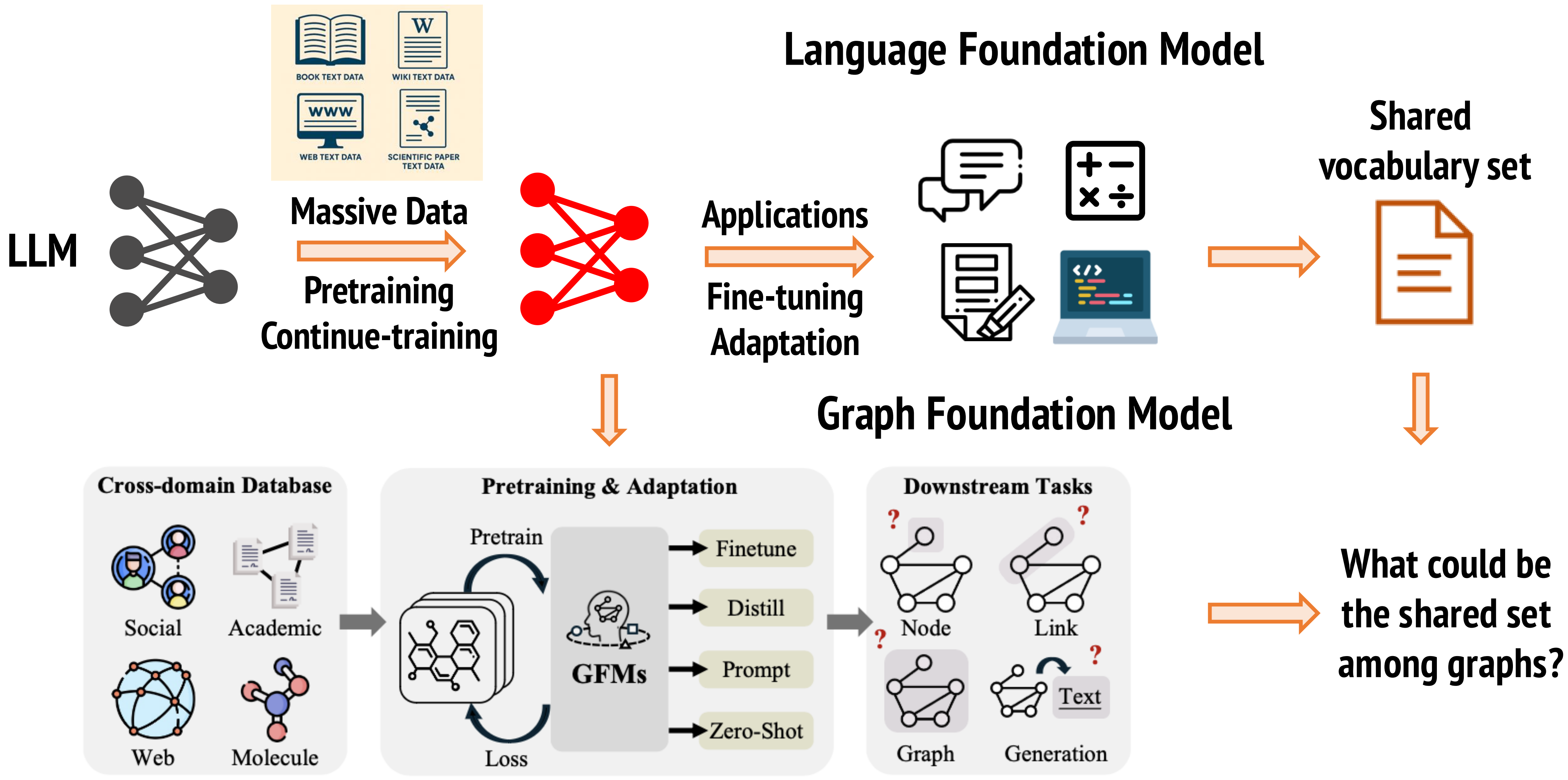
Look Twice as Much as You Say: Scene Graph Contrastive Learning for Self-Supervised Image Caption Generation, CIKM 2022

Outline

On the Intersection of Language and Graph Models

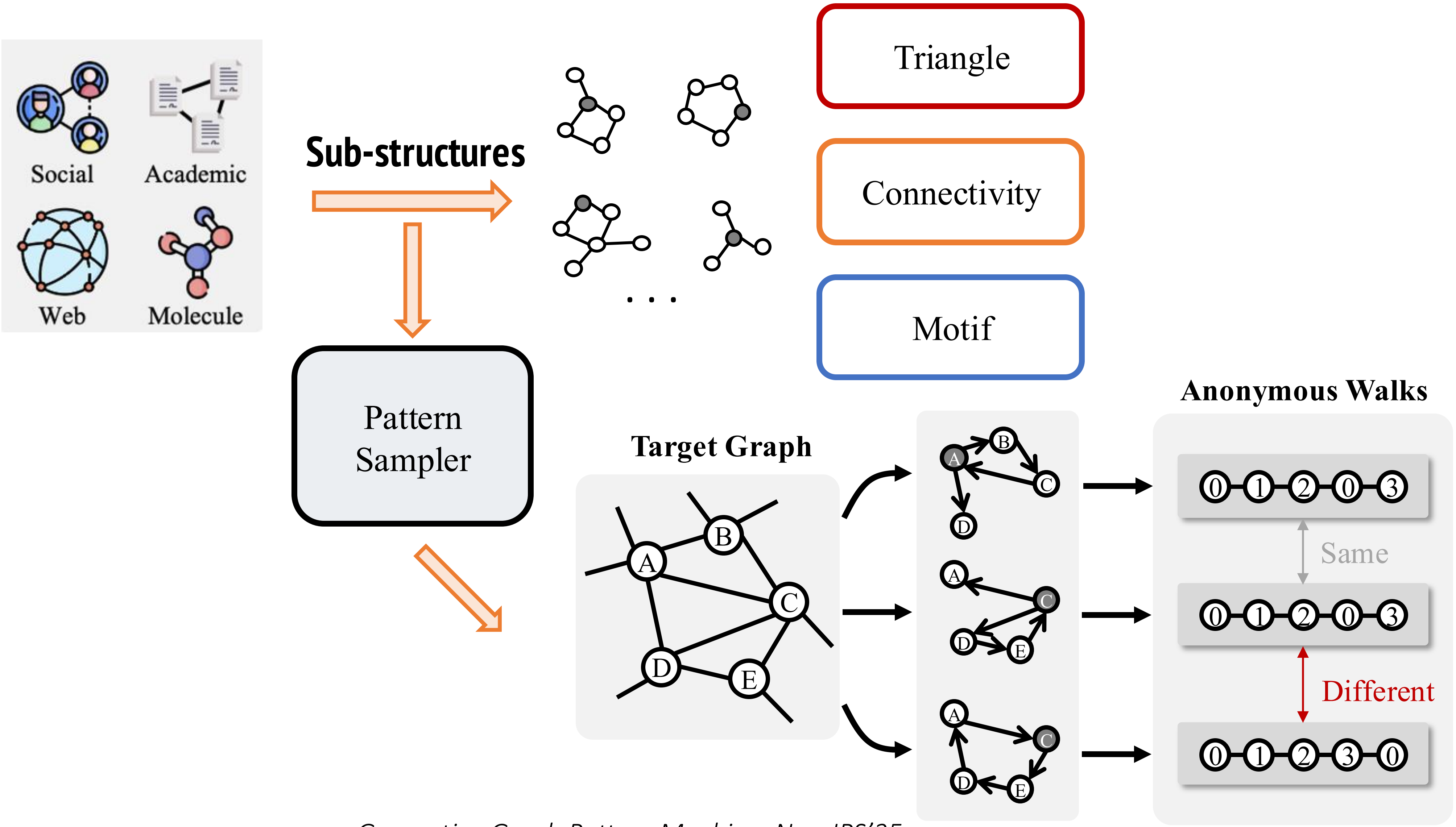
1. Introduction and Background
2. Graph Data/Model Enhances Language Model
3. Language Data/Model Enhances Graph Model
4. Joint Graph and Language Models for Societal Applications
5. Conclusion

Language Data/Model Enhances Graph Model: Foundation Model



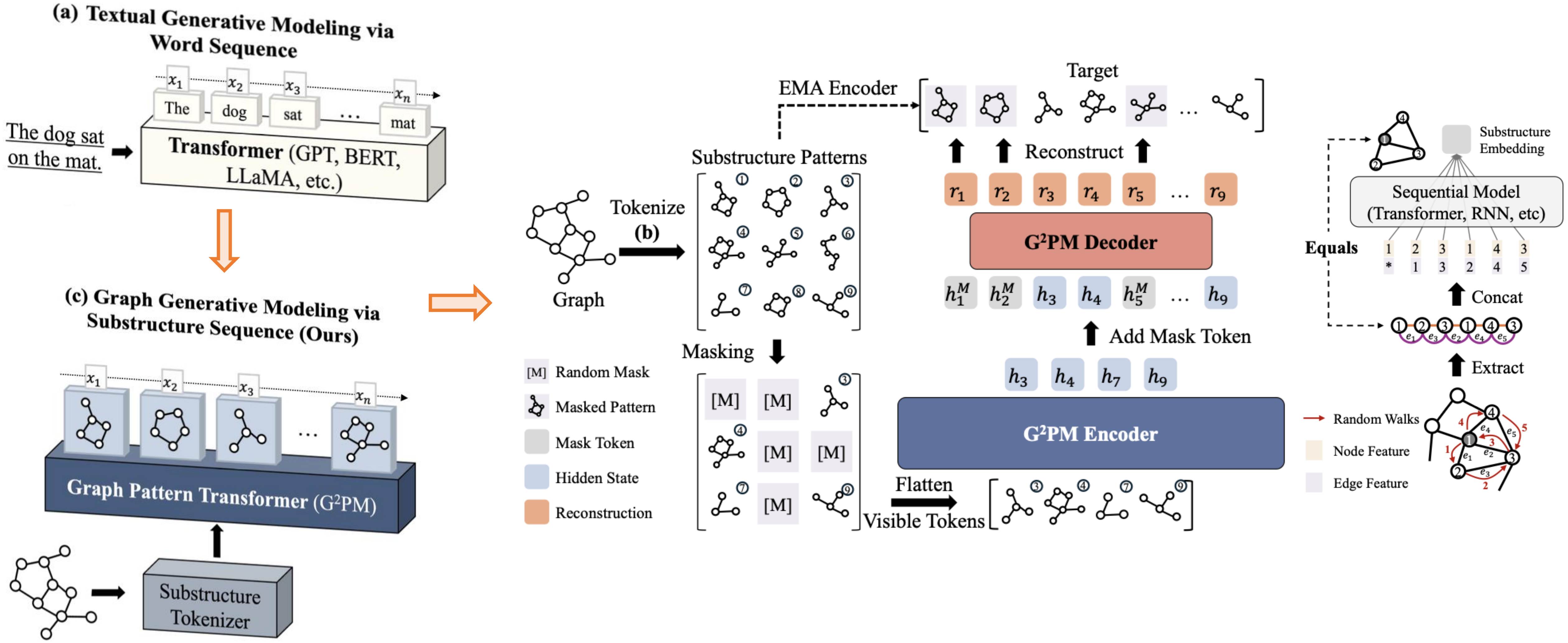
Generative Graph Pattern Machine, NeurIPS'25

Language Data/Model Enhances Graph Model: Graph Foundation Model



Generative Graph Pattern Machine, NeurIPS'25

Language Data/Model Enhances Graph Model: Graph Foundation Model



Generative Graph Pattern Machine, NeurIPS'25

Language Data/Model Enhances Graph Model: Graph Foundation Model

		Pubmed	Photo	Computers	WikiCS	Flickr	Arxiv	Products	A.R.
	# Nodes	19,717	7,650	13,752	11,701	89,250	169,343	2,449,029	-
	# Edges	88,648	238,162	491,722	431,206	899,756	2,315,598	123,718,024	-
Supervised	GAT [55]	83.1 ± 0.3	91.9 ± 0.5	87.9 ± 0.5	76.9 ± 0.8	50.7 ± 0.3	72.10 ± 0.13	79.45 ± 0.59	5.7
	GPM [62]	84.7 ± 0.1	92.7 ± 0.3	90.0 ± 0.4	80.2 ± 0.4	52.2 ± 0.2	72.89 ± 0.68	82.62 ± 0.39	1.3
Contrastive	GCA [78]	83.3 ± 0.5	92.4 ± 0.2	87.1 ± 0.2	77.4 ± 0.1	49.0 ± 0.1	71.23 ± 0.09	78.39 ± 0.03	6.9
	BGRL [48]	83.9 ± 0.3	92.5 ± 0.2	88.2 ± 0.2	77.5 ± 0.8	49.7 ± 0.2	70.51 ± 0.03	78.59 ± 0.02	5.7
	CCA-SSG [72]	81.8 ± 0.5	91.8 ± 0.6	88.6 ± 0.3	75.3 ± 0.8	47.5 ± 0.2	71.24 ± 0.20	75.27 ± 0.05	8.6
Generative	GraphMAE [21]	81.0 ± 0.5	92.0 ± 0.3	**89.2 ± 0.5**	77.1 ± 0.5	50.5 ± 0.1	71.75 ± 0.17	78.89 ± 0.01	6.0
	GraphMAE 2 [22]	81.3 ± 0.4	92.4 ± 0.2	88.3 ± 0.9	77.6 ± 0.4	50.4 ± 0.1	71.89 ± 0.03	79.33 ± 0.01	5.4
	S2GAE [47]	80.1 ± 0.5	91.4 ± 0.1	85.3 ± 0.1	75.3 ± 0.8	48.1 ± 0.8	67.77 ± 0.36	76.70 ± 0.03	10.3
	Bandana [76]	83.5 ± 0.5	91.4 ± 0.7	87.7 ± 0.2	77.3 ± 0.3	47.9 ± 0.6	71.09 ± 0.24	77.68 ± 0.05	8.1
	G^2PM w/o Pretrain	83.9 ± 0.2	92.8 ± 0.2	87.1 ± 0.3	78.5 ± 0.4	50.7 ± 0.1	69.64 ± 0.08	76.90 ± 0.16	6.0
	G^2PM	**84.3 ± 0.1**	**92.9 ± 0.2**	88.8 ± 0.3	**79.0 ± 0.4**	**51.0 ± 0.0**	**72.31 ± 0.07**	**80.56 ± 0.01**	2.0

		HIV	PCBA	Sider	MUV	ClinTox	IMDB-B	REDDIT-M12K	A.R.
	# Graphs	41,127	437,929	1,427	93,087	1,478	1,000	11,929	-
	# Nodes	∼25.5	∼26.0	∼33.6	∼24.2	∼26.1	∼19.8	∼391.4	-
	# Edges	∼27.5	∼28.1	∼70.7	∼52.6	∼55.5	∼193.1	∼913.8	-
Supervised	GIN [67]	75.8 ± 0.8	70.3 ± 0.3	57.7 ± 0.8	74.4 ± 0.9	83.4 ± 0.6	73.3 ± 0.5	39.4 ± 1.4	6.3
	GPM [62]	77.0 ± 0.9	75.1 ± 0.3	59.0 ± 0.0	74.6 ± 1.4	82.4 ± 0.3	82.7 ± 0.5	43.1 ± 0.3	3.0
Contrastive	GraphCL [69]	75.5 ± 0.3	72.4 ± 2.1	57.3 ± 0.9	68.3 ± 2.6	82.9 ± 0.3	71.1 ± 0.4	37.9 ± 2.4	8.0
	JOAO [70]	76.8 ± 0.3	73.4 ± 1.5	58.5 ± 0.5	72.3 ± 1.0	82.2 ± 0.3	70.2 ± 3.1	39.9 ± 0.6	6.0
	MVGRL [18]	75.7 ± 0.7	70.4 ± 2.1	60.5 ± 0.6	71.5 ± 1.2	83.6 ± 0.2	74.2 ± 0.7	39.5 ± 1.8	5.7
	InfoGCL [66]	77.3 ± 0.6	74.6 ± 0.7	58.7 ± 0.7	73.4 ± 1.0	80.3 ± 0.7	75.1 ± 0.9	39.3 ± 0.5	5.4
Generative	GraphMAE [21]	77.8 ± 0.9	73.2 ± 1.4	60.6 ± 0.0	73.7 ± 0.8	84.8 ± 0.5	75.5 ± 0.7	37.6 ± 2.5	4.1
	S2GAE [47]	75.6 ± 0.8	72.9 ± 0.0	58.0 ± 0.9	71.6 ± 0.8	80.6 ± 0.4	75.8 ± 0.6	37.9 ± 1.8	7.0
	G^2PM w/o Pretrain	69.8 ± 0.1	68.4 ± 0.0	58.8 ± 0.3	66.3 ± 1.4	80.0 ± 1.8	80.0 ± 0.8	37.5 ± 0.3	8.3
	G^2PM	**78.7 ± 0.1**	**75.6 ± 0.1**	**61.2 ± 0.2**	**75.7 ± 0.4**	**86.6 ± 0.8**	**83.0 ± 0.8**	**41.8 ± 0.3**	1.0

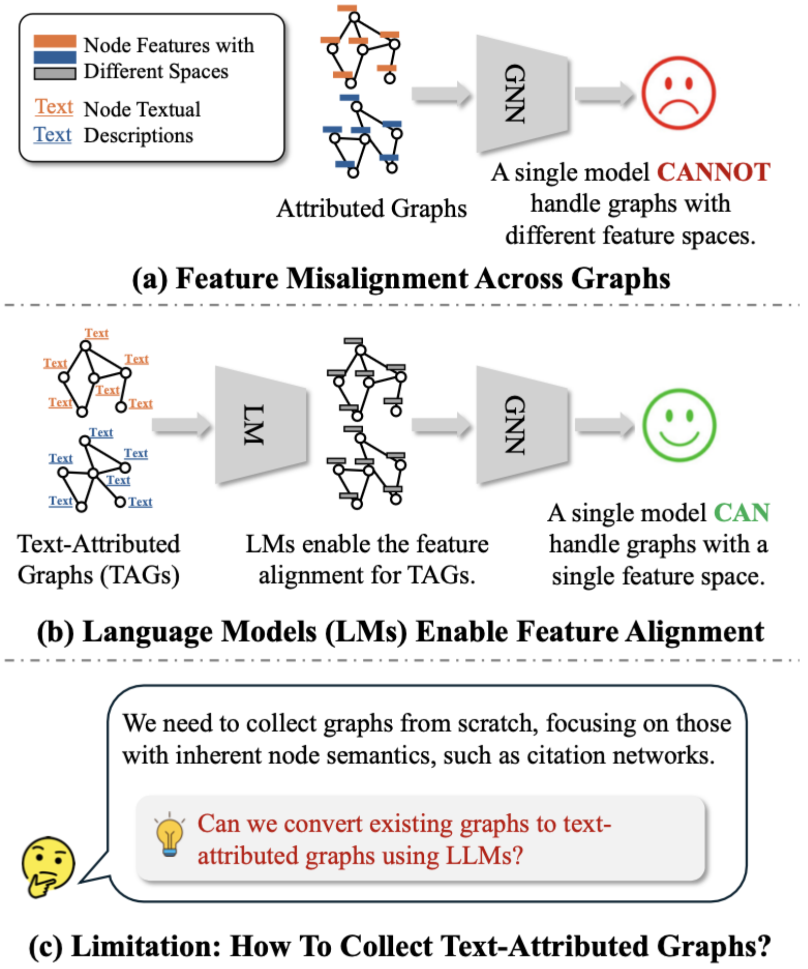
Source	**Arxiv**		**HIV**	
Target	**Products**	**HIV**	**Arxiv**	**PCBA**
GNN [55, 67]	78.3 (1.2 ↓)	70.1 (5.7 ↓)	71.1 (1.0 ↓)	71.9 (**1.6 ↑**)
GPM [62]	82.0 (0.6 ↓)	74.3 (2.7 ↓)	71.4 (1.5 ↓)	76.4 (**1.3 ↑**)
BGRL [48]	78.8 (**0.2 ↑**)	72.5 (3.8 ↓)	68.6 (1.9 ↓)	72.9 (0.6 ↓)
GraphMAE [21]	77.5 (1.4 ↓)	74.7 (3.1 ↓)	69.9 (1.9 ↓)	73.4 (**0.2 ↑**)
G^2PM	**81.3** (**0.7 ↑**)	**76.8** (1.9 ↓)	**72.6** (**0.3 ↑**)	**77.9** (**2.3 ↑**)

Pretrain	**Arxiv + FB15K237 + ChemBL**		
Downstream	**Arxiv (Academia)**	**FB15K237 (Knowledge Graph)**	**HIV (Molecule)**
BGRL [48]	70.8 ± 0.2	86.5 ± 0.3	68.5 ± 1.6
GraphMAE [21]	70.3 ± 0.3	87.8 ± 0.4	64.1 ± 0.5
OFA [32]	71.4 ± 0.3	84.7 ± 1.3	72.0 ± 1.6
GFT [60]	71.9 ± 0.1	**89.3 ± 0.2**	72.3 ± 2.0
G^2PM	**72.5 ± 0.1**	88.9 ± 0.5	**74.1 ± 1.3**

Generative Graph Pattern Machine, NeurIPS'25

Language Data/Model Enhances Graph Model

- **LLM as Text/Attribute Generation for GNNs**



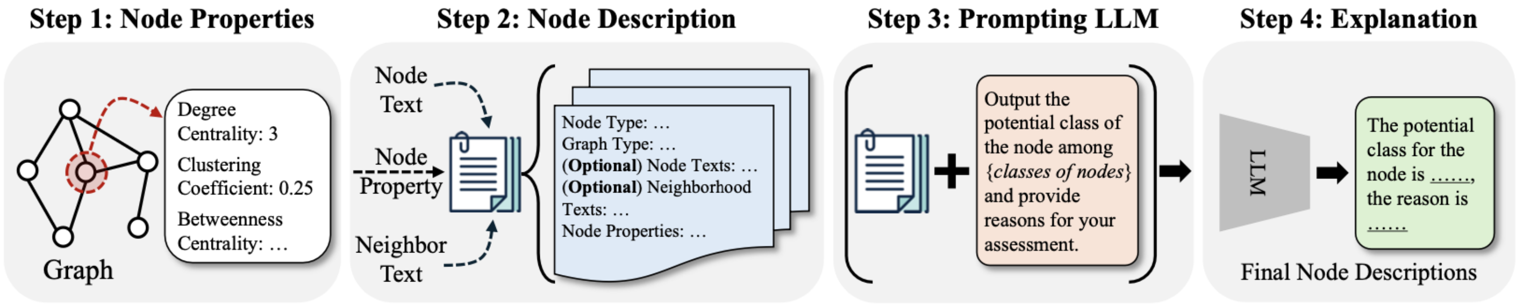


Figure 2: The framework of our topology-aware node description synthesis (TANS).

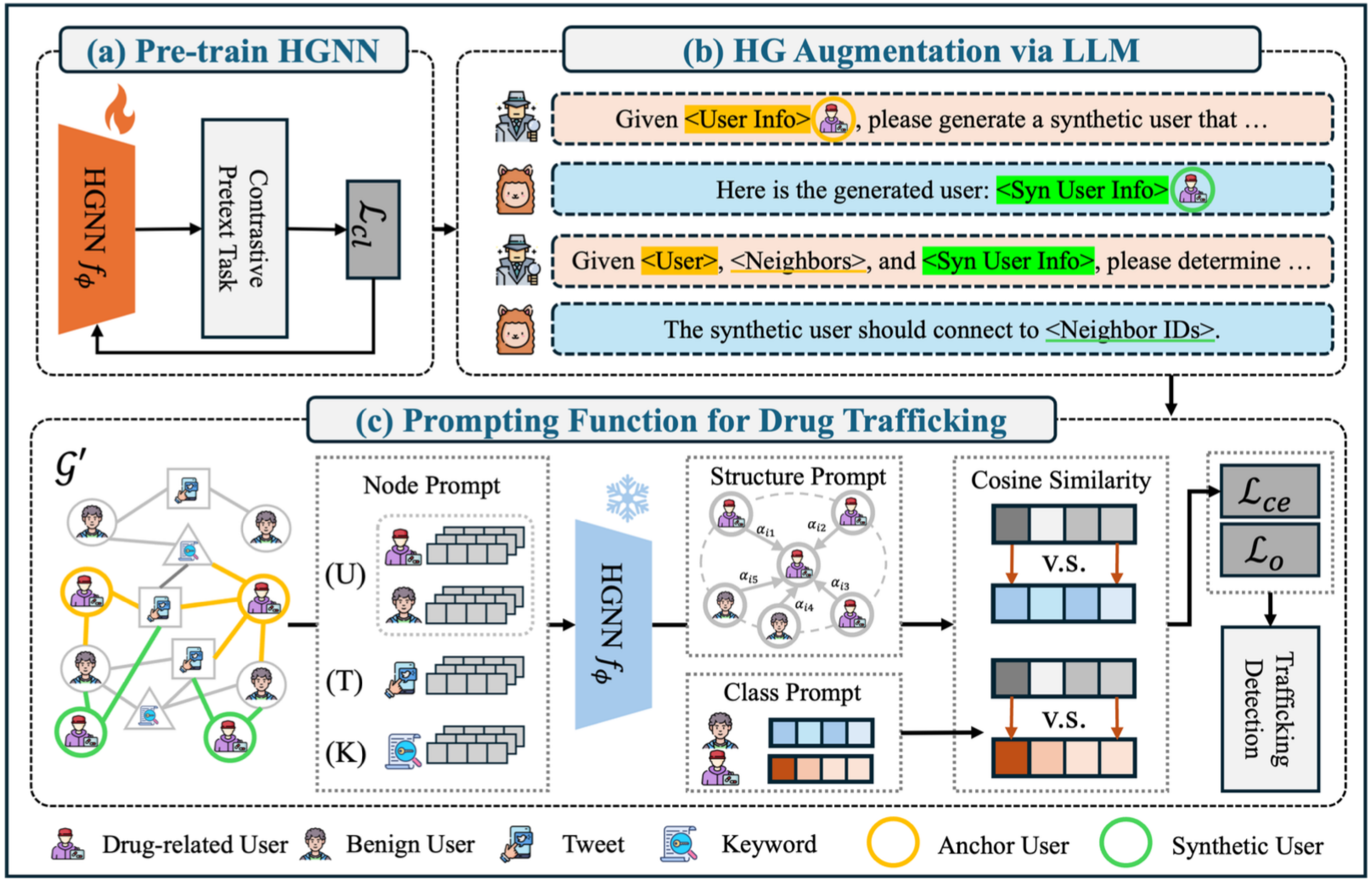
Step 2: Generate Basic Node Descriptions	
Prefix	`Given a node from a {Graph Type} graph, where the node type is {Node Type} with {Node Number} nodes, and the edge type is {Edge Type} with {Edge Number} edges.`
Node Text (Optional)	`The original node description is {Original Textual Descriptions}.`
Neighbor Text (Optional)	`The following are the textual information of {k} connected nodes. The descriptions are: {Textual Descriptions of Selected Neighborhoods}.`
Node Property	`The value of {Node Property} is {Value of The Given Property}, ranked as {Rank of The Node}% among {Node Number} nodes.`
Step 3: Prompting LLMs	
Suffix	`Output the potential {k} classes of the node and provide reasons for your assessment. The classes include {Classes of Nodes}. Your answer should be less than 200 words.`

Table 2: Prompt templates.

Can LLMs Convert Graphs to Text-Attributed Graphs? NAACL 2025

Language Data/Model Enhances Graph Model

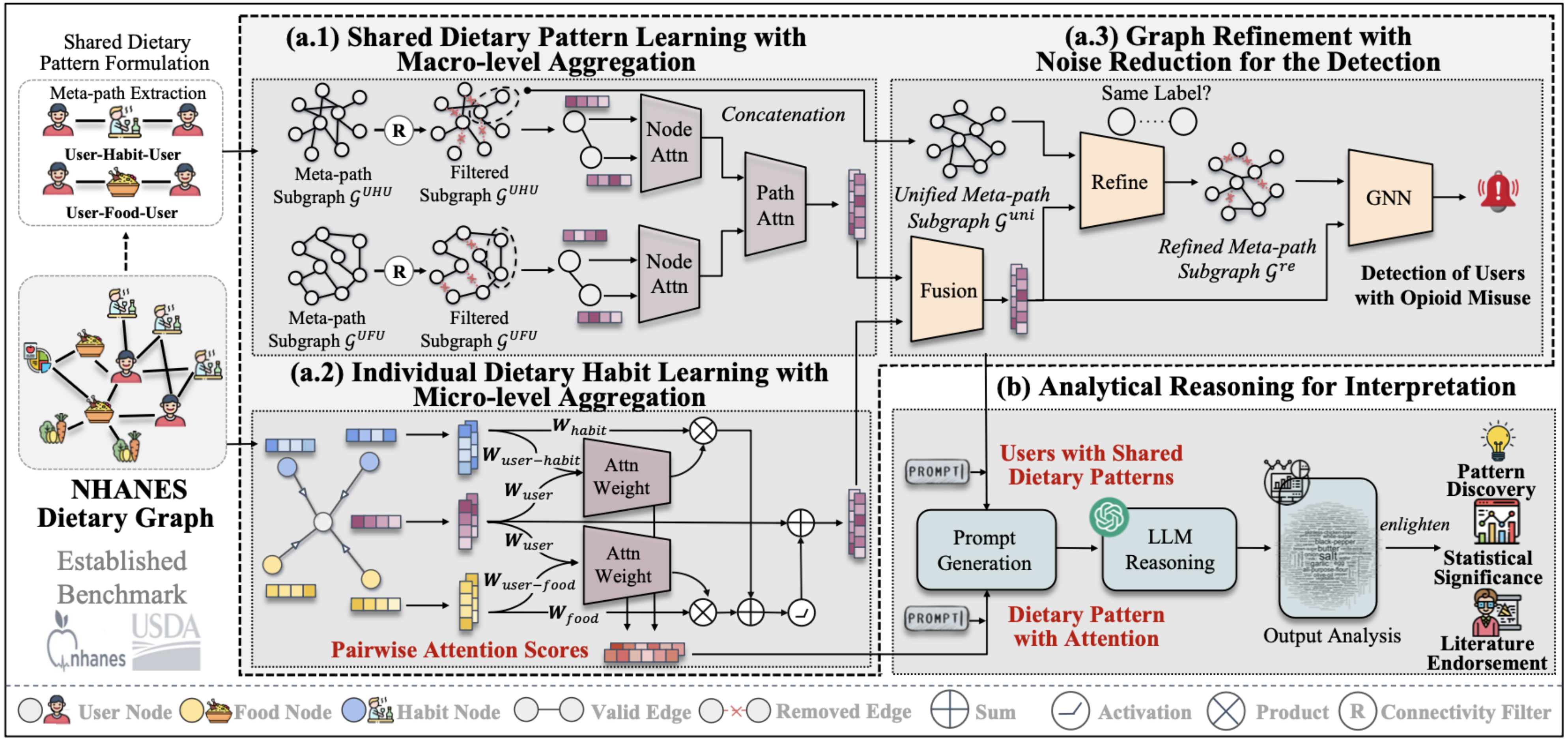
- **LLM as Data Augmentation for GNNs**



LLM-Empowered Class Imbalanced Graph Prompt Learning for Online Drug Trafficking Detection, ACL 2025

Language Data/Model Enhances Graph Model

- **Healthcare: LLM as Interpretator for GNN Output**



Diet-ODIN: A Novel Framework for Opioid Misuse Detection with Interpretable Dietary Patterns, KDD 2024

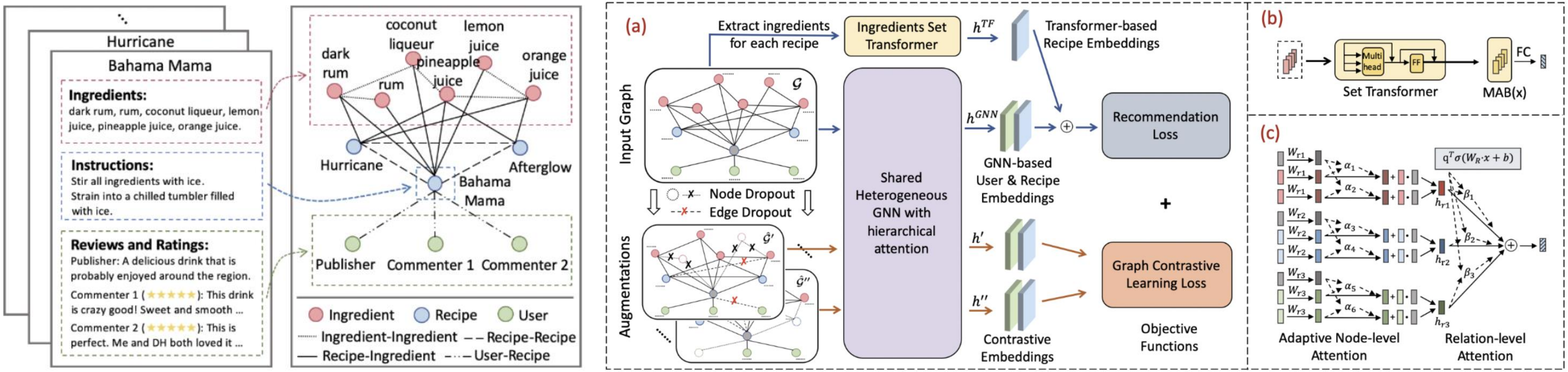
Outline

On the Intersection of Language and Graph Models

1. Introduction and Background
2. Graph Data/Model Enhances Language Model
3. Language Data/Model Enhances Graph Model
4. **Joint Graph and Language Models for Societal Applications**
5. Conclusion

Joint Graph and Language Models for Various Applications

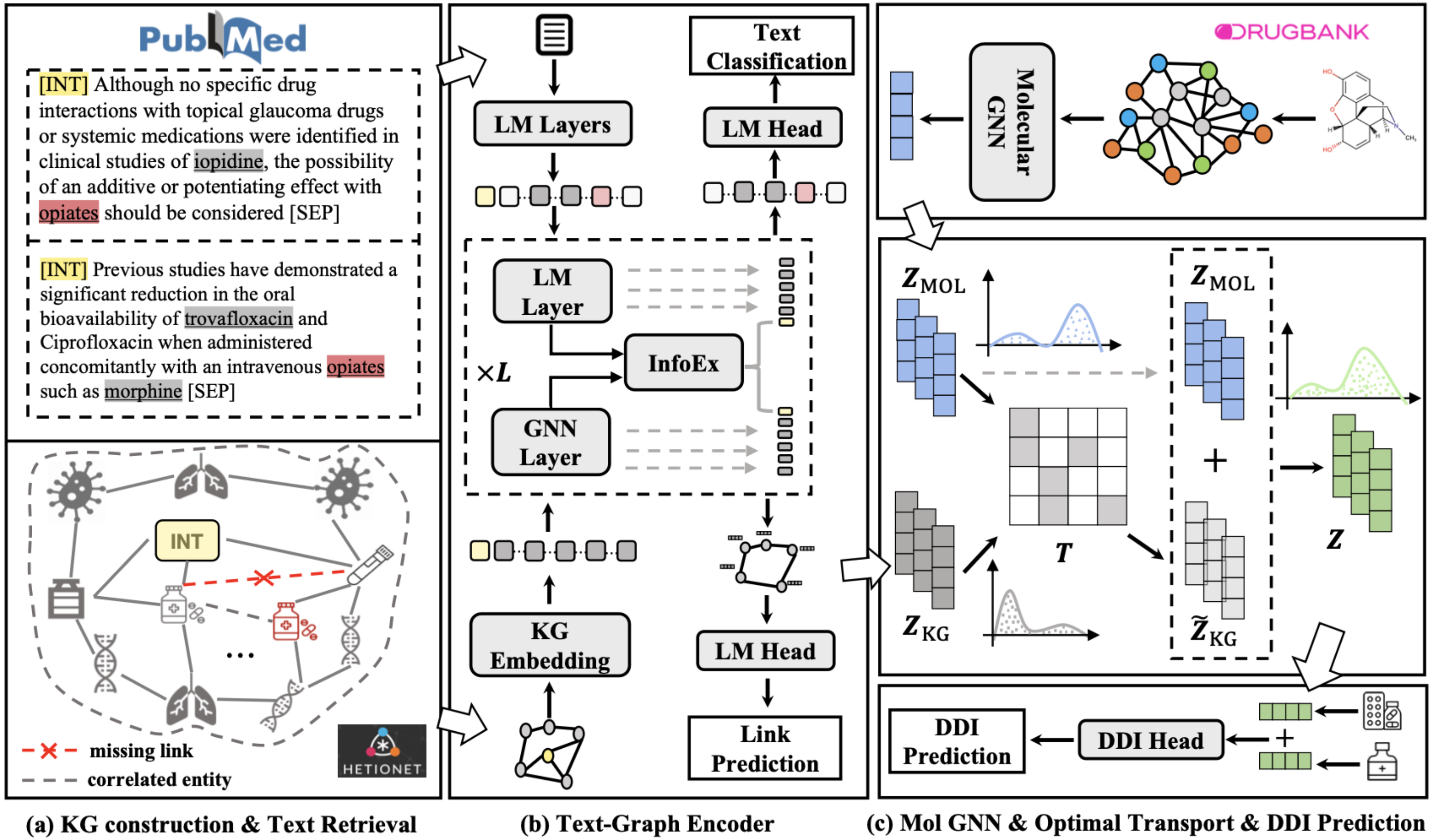
- **Recommender System: GNN and LLM as Multi-modal Data Encoder**



RecipeRec: A Heterogeneous Graph Learning Model for Recipe Recommendation, IJCAI 2022

Joint Graph and Language Models for Various Applications

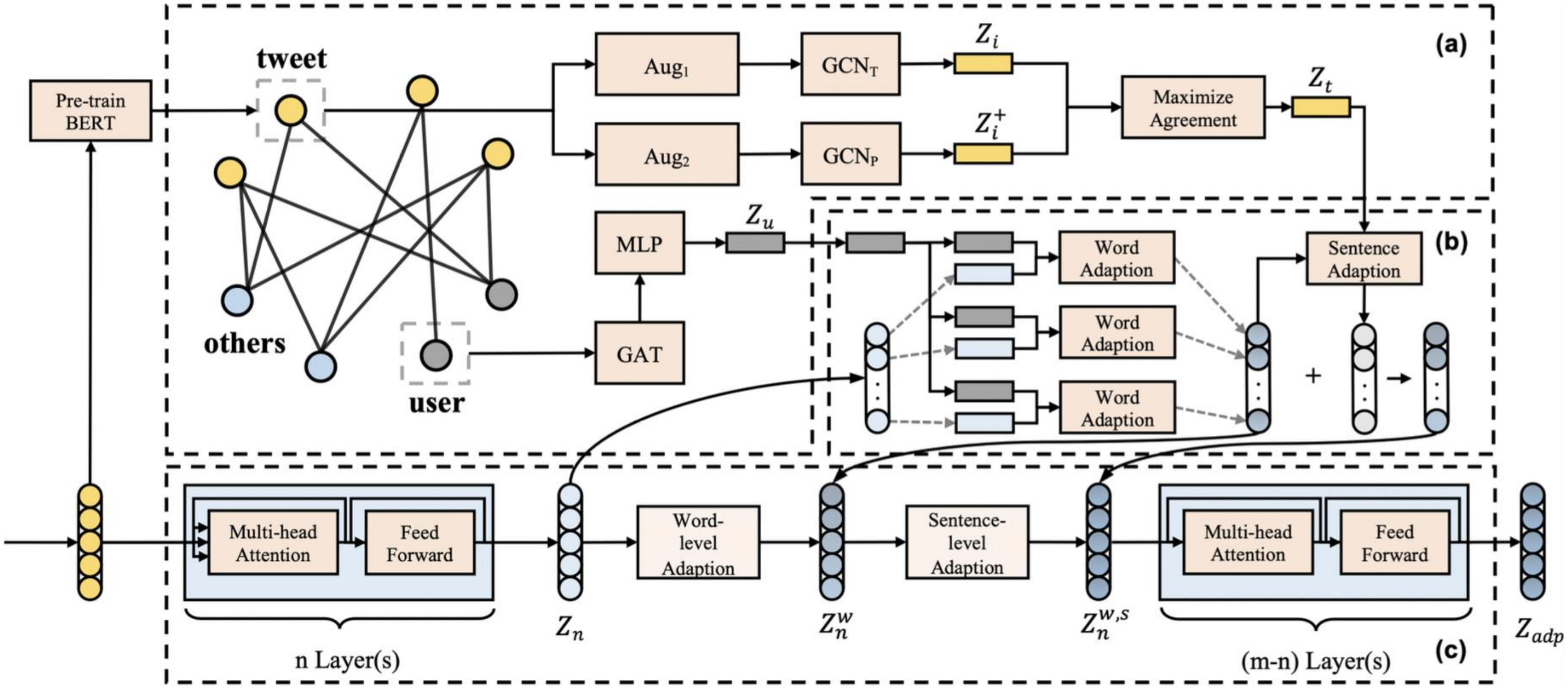
- **Healthcare: GNN and LLM as Multi-modal Data Encoder**



A Multi-Modality Framework for Drug-Drug Interaction Prediction by Harnessing Multi-source Data, CIKM 2023

Joint Graph and Language Models for Various Applications

- **Social Network Analysis: GNN and LLM as Multi-modal Data Encoder**



GraphBERT: Bridging Graph and Text for Malicious Behavior Detection on Social Media, ICDM 2022

Outline

On the Intersection of Language and Graph Models

1. Introduction and Background
2. Graph Data/Model Enhances Language Model
3. Language Data/Model Enhances Graph Model
4. Joint Graph and Language Models for Societal Applications
5. Conclusion

Conclusion

- Graph Data/Model Improves Language Model
  - MASS: LLM for Advanced Reasoning
  - Graph as Augmented Information/Data for LLMs in QA, Text Generation, etc.
- Language Data/Model Improves Graph Model
  - GP2M: Graph Foundation Model
  - LLM as Text/Attribute Generation or Data Augmentation for GNNs, etc.
- Joint Graph-Language Model for Societal Applications
  - Recommender Systems, Social Network Analysis, Healthcare, etc.

Q&A

- **Feel free to contact me for any questions!**
  - ☞ Contact email: chuxuzhang@gmail.com