Towards Low-Distortion Graph Representation Learning: Advanced Directions

Yuan Haonan @ *MAGIC Group*
yuanhn@buaa.edu.cn

August 29th, 2025

Outlines

I. Recap of Tutorial

- Low-Distortion GRL – Motivation and Key Concept
- Key Approaches to Reduce Distortion

II. Future Directions

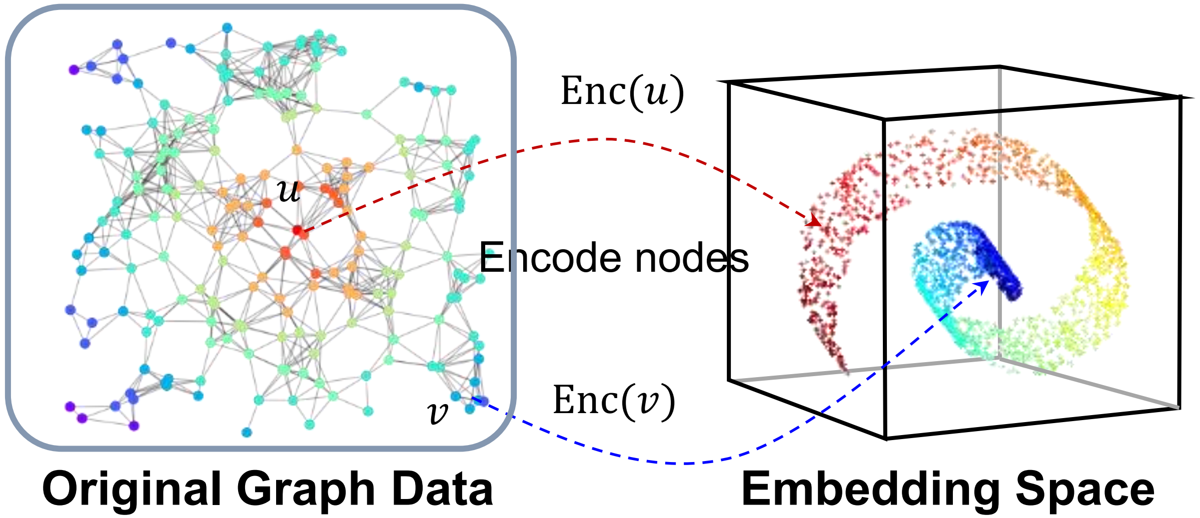
- Benchmarks for Low-Distortion GRL
- Graph Foundation Model (GFM)
- Graph Retrieval-Augmented Generation (GraphRAG)
- Graph World Model (GWM)

III. Open Challenges and Outlook

IV. Discussion

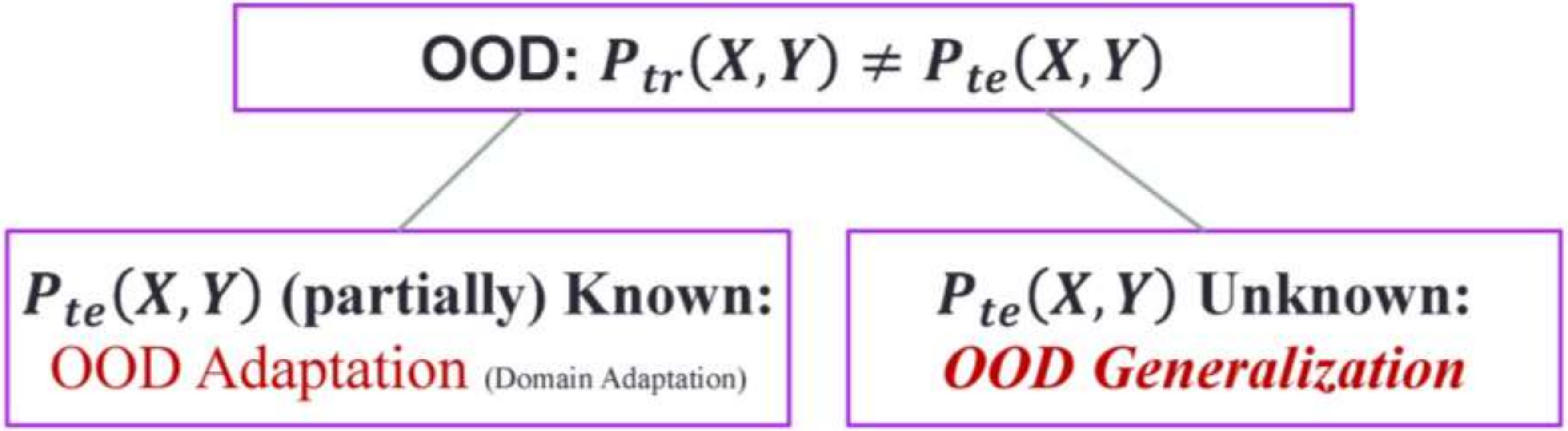
■ Low-Distortion GRL – Motivation and Key Concept

- □ **Graph distortion:** loss of intrinsic structure info in embeddings (noisy edges, missing links, altered topology)
- □ **Goal:** preserve essential graph properties in low-dim representations
- □ Complex topologies make embeddings sensitive to small perturbations, risking major information loss



Key Approaches to Reduce Distortion

- **Invariance-guided (causality-based) methods:** enforce stable representations under graph perturbations or interventions, isolating causal structure from noise

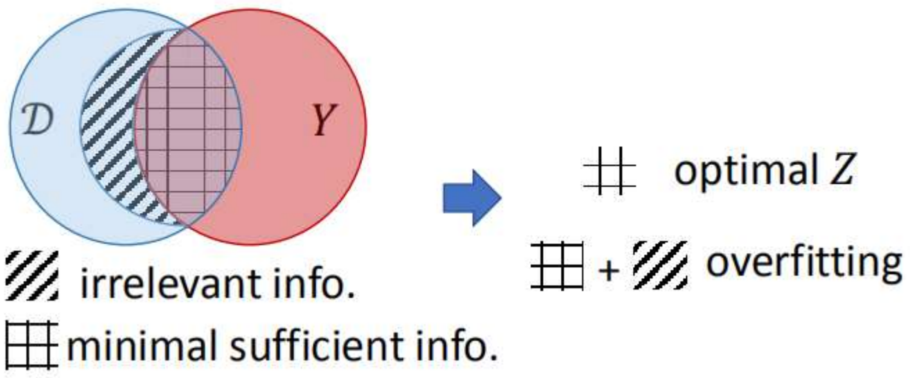


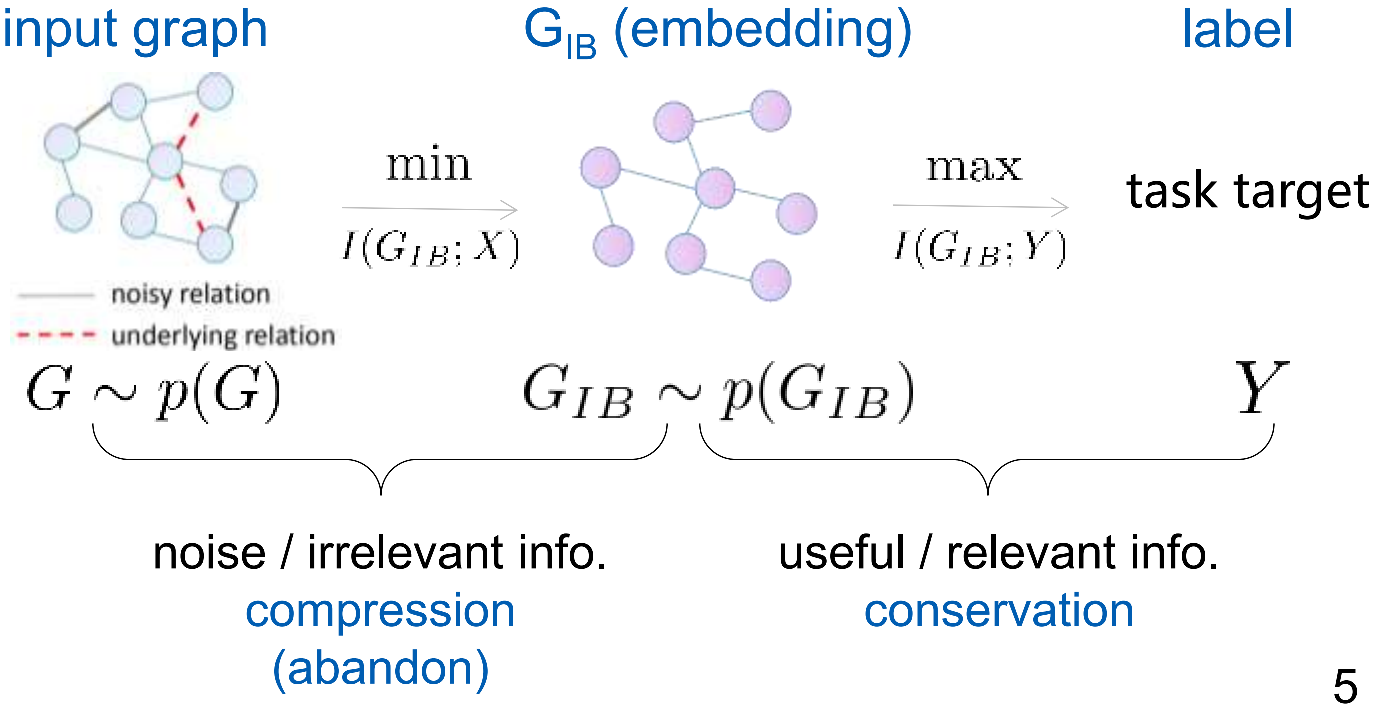
[Picture Credit] Stable Learning for Out-of-Distribution Generalization and the NICO Challenge. *Peng Cui*.

Key Approaches to Reduce Distortion

- **Information-theoretic methods:** maximize mutual information and minimize info loss between graphs and embeddings (retain as much original signal as possible)

$$\max_{G_{IB}} I(G_{IB};Y) - \beta I(G_{IB};G)$$

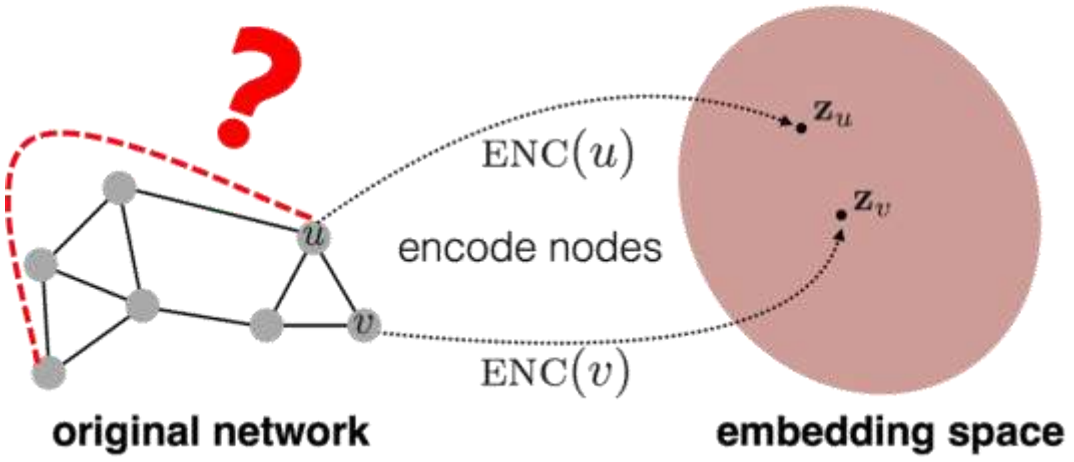


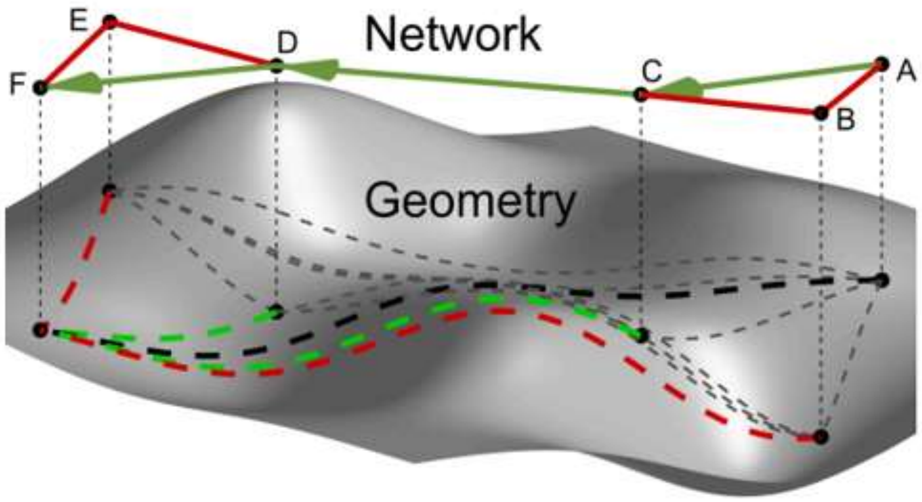


[Picture Credit] Graph information bottleneck. *Tailin Wu*, et al.

■ Key Approaches to Reduce Distortion

- **Geometry-guided methods:** embed graphs in non-Euclidean spaces (e.g. hyperbolic) better suited to graph structure, reducing embedding distortion for hierarchical or complex topologies





[Picture Credit] Efficient and Effective Edge-wise Graph Representation Learning. *Hewen Wang*, et al.
Hyperbolic mapping of complex networks. *Dmitri Krioukov*, et al.

Outlines

- I. Recap of Tutorial
  - Low-Distortion GRL – Motivation and Key Concept
  - Key Approaches to Reduce Distortion
- II. Future Directions
  - Benchmarks for Low-Distortion GRL
  - Graph Foundation Model (GFM)
  - Graph Retrieval-Augmented Generation (GraphRAG)
  - Graph World Model (GWM)
- III. Open Challenges and Outlook
- IV. Discussion

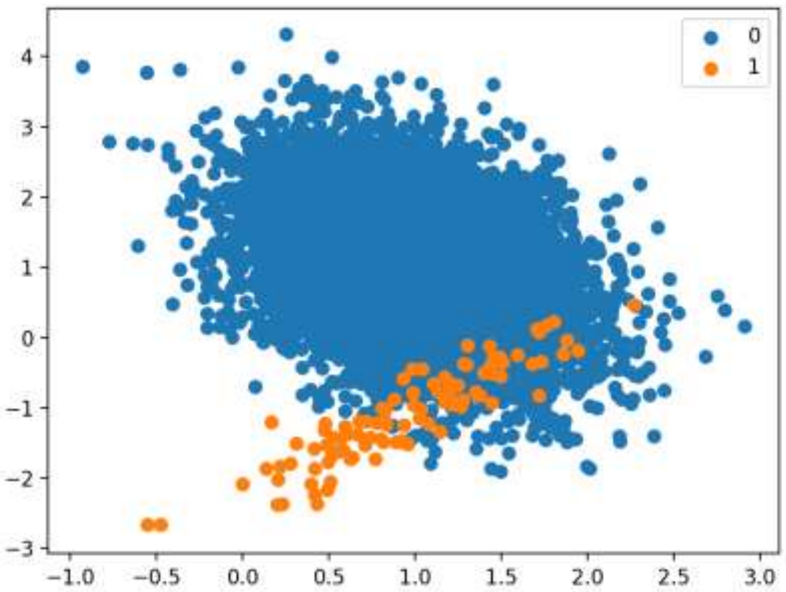
Benchmarks for Low-Distortion GRL

- **Need for structural fidelity benchmarks:** Current evaluations often don't reveal the advantages of low-distortion methods
- Develop new metrics (e.g. curvature, hyperbolicity) and tasks to test how well embeddings preserve topology and features
- **Community benchmarks & leaderboards:** Build shared datasets and standardized protocols to fairly compare methods and drive progress

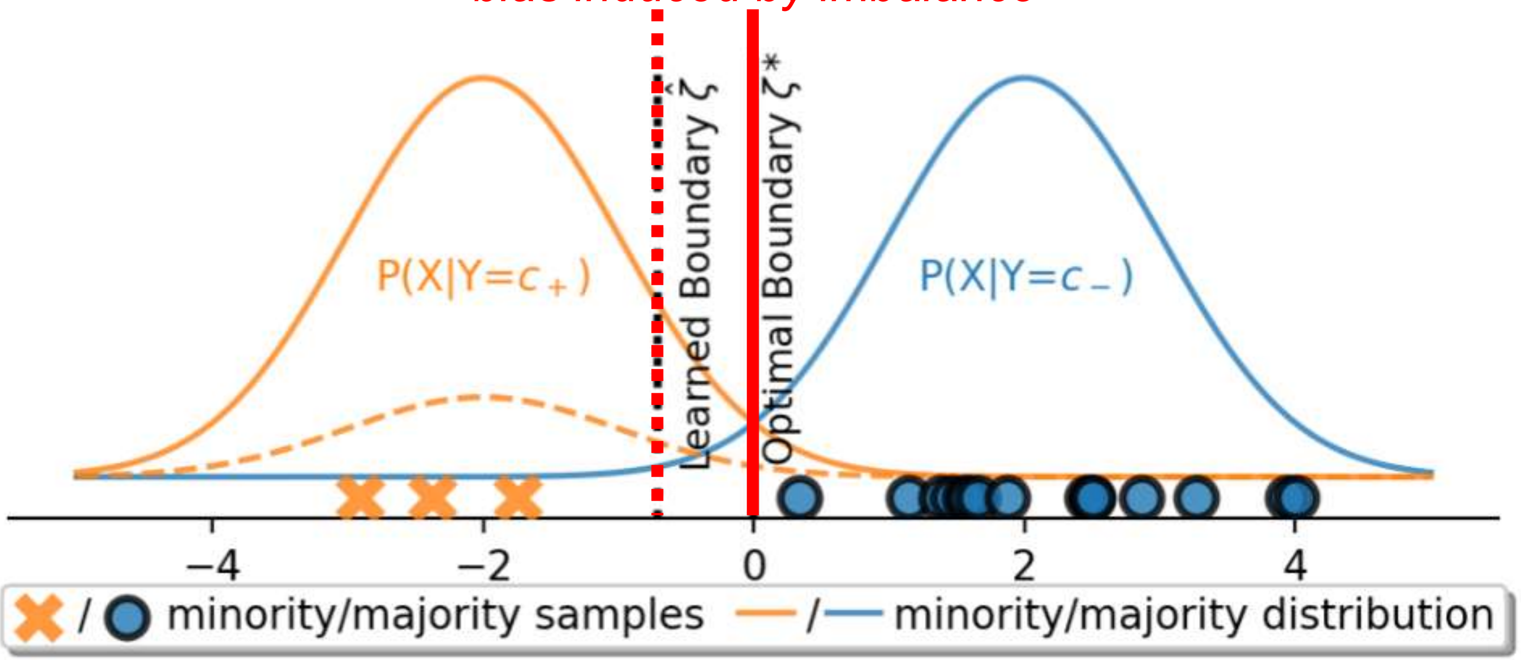
Benchmarks for Low-Distortion GRL

Imbalanced Machine Learning

- Data imbalance leads to decision boundary shift
- decision boundary shift → high-distortion GRL



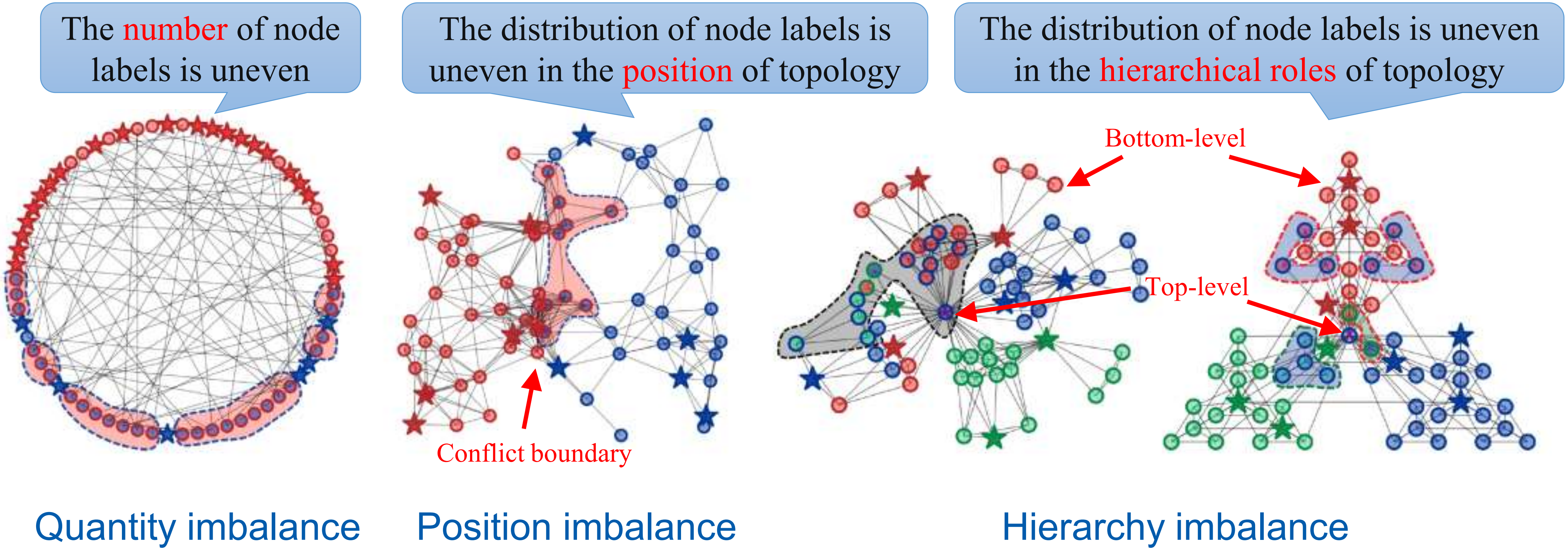
bias induced by imbalance



[Picture Credit] Handling inter-class and intra-class imbalance in class-imbalanced learning. *Zhining Liu*, et al.

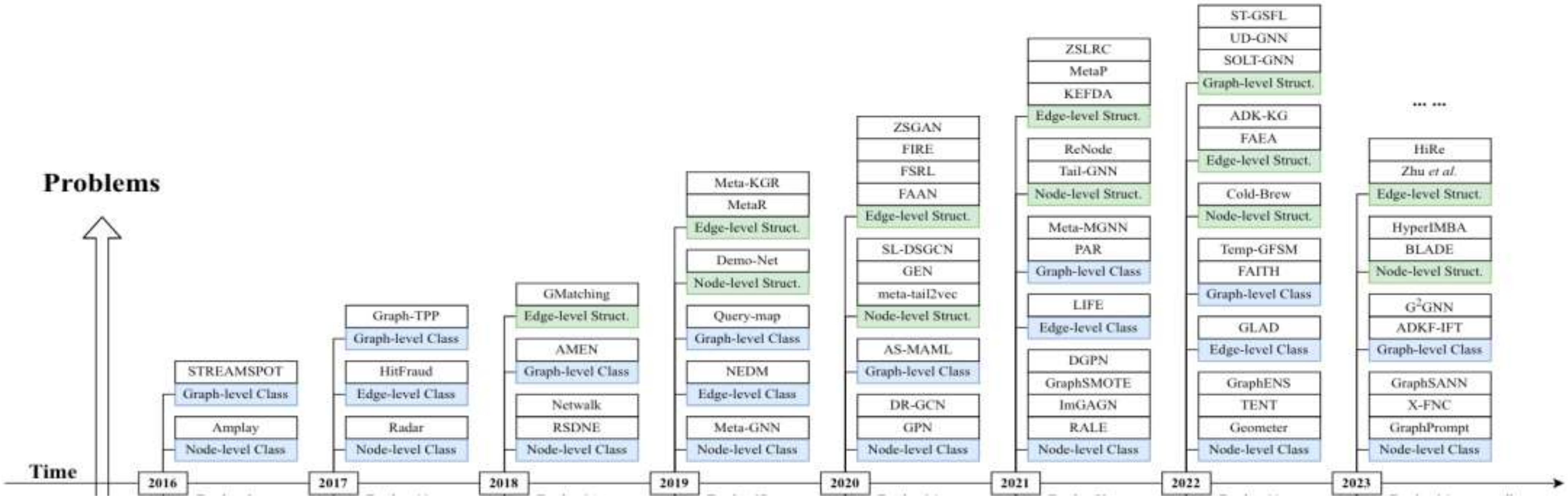
Benchmarks for Low-Distortion GRL

Imbalanced Graph Learning (IGL)



[Picture Credit] Hyperbolic geometric graph representation learning for hierarchy-imbalance node classification. *Xingcheng Fu*, et al.

■Benchmarks for Low-Distortion GRL



- □ There lacks a comprehensive benchmark for Imbalanced Graph Learning (IGL), which significantly impedes the understanding and progress of IGL

[Picture Credit] A survey of imbalanced learning on graphs: Problems, techniques, and future directions. *Zemin Liu*, et al.

Benchmarks for Low-Distortion GRL

Why Benchmark for IGL?

- **Node-level** imbalanced graph learning
  - Class-imbalance: the disproportionate distribution of labeled nodes across classes
  - Topology-imbalance: the positional distribution of labeled nodes on the graph
- **Graph-level** imbalanced graph learning
  - Class-imbalance: the disproportionate distribution of labeled graphs across classes
  - Size-imbalance: the great disparity in graph sizes between multiple graphs
- ……

Algorithms (Node-level)	
Resampling	LTE4G, ALLIE……
Reweighting	TAM, HyperIMBA, Renode, TOPOAUC……
Reconstruction	GraphENS, PASTEL, GraphSANN, GraphSHA……

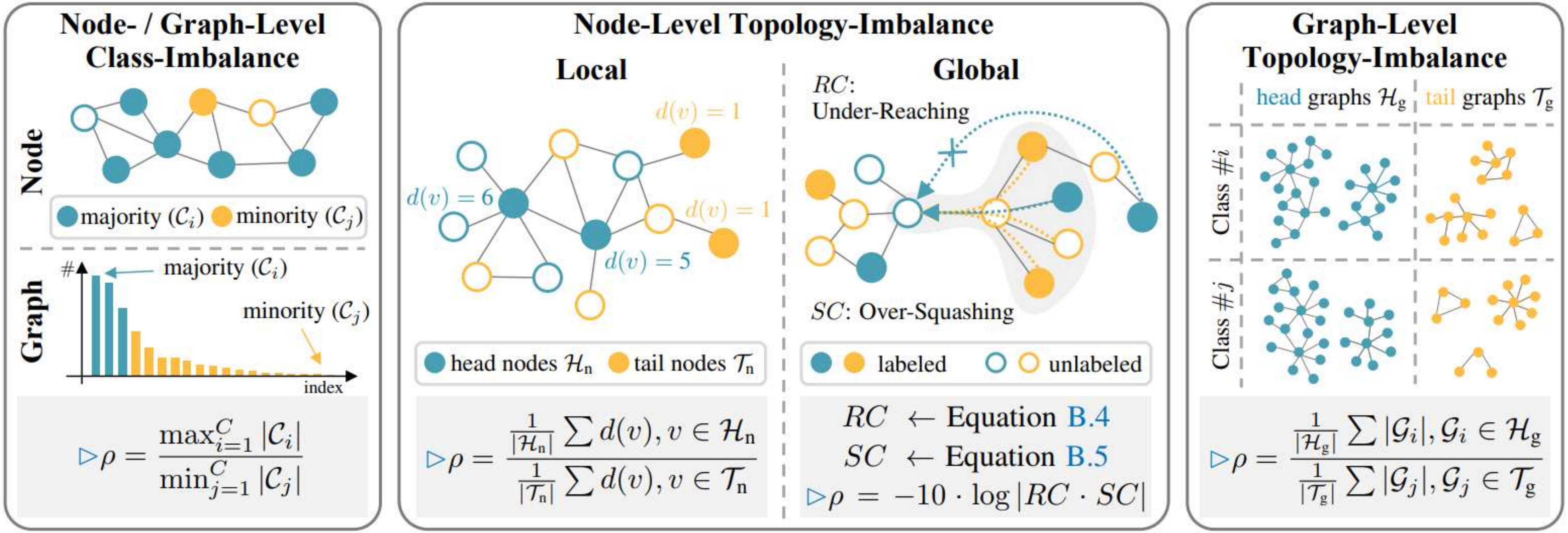
Datasets (Node-level)	
Manual imbalanced datasets	Cora, Citeseer, Pubmed, Chameleon, Squirrel, Actor
Natural imbalanced datasets	Amazon-Photo, Amazon-Computers, ogbn-arXiv

⇩

Comparison | Backbones | Homophily | …

Benchmarks for Low-Distortion GRL

- **IGL-Bench:** Establishing the comprehensive benchmark for IGL
- **17** diverse graph datasets and **24** distinct IGL algorithms



[Picture Credit] IGL-Bench: Establishing the comprehensive benchmark for imbalanced graph learning (ICLR 2025 Spotlight). *Jiawen Qin*, et al.

■ Graph Foundation Models (GFM)

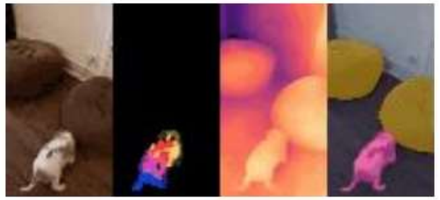
- **Foundation Models:** A foundation model is a model that is trained on broad data and be adapted to a wide range of downstream tasks.
  - Pretrain-then-finetune
  - Revolutionize many research domains

Language

OpenAI × GPT4

Language foundation models show initial signs of universal AI capabilities.

Vision



Meta × DINOv2

Vision foundation models exhibit strong image understanding capabilities.

Speech

Google × USM

Speech foundation models show the capability to recognize hundreds of languages.

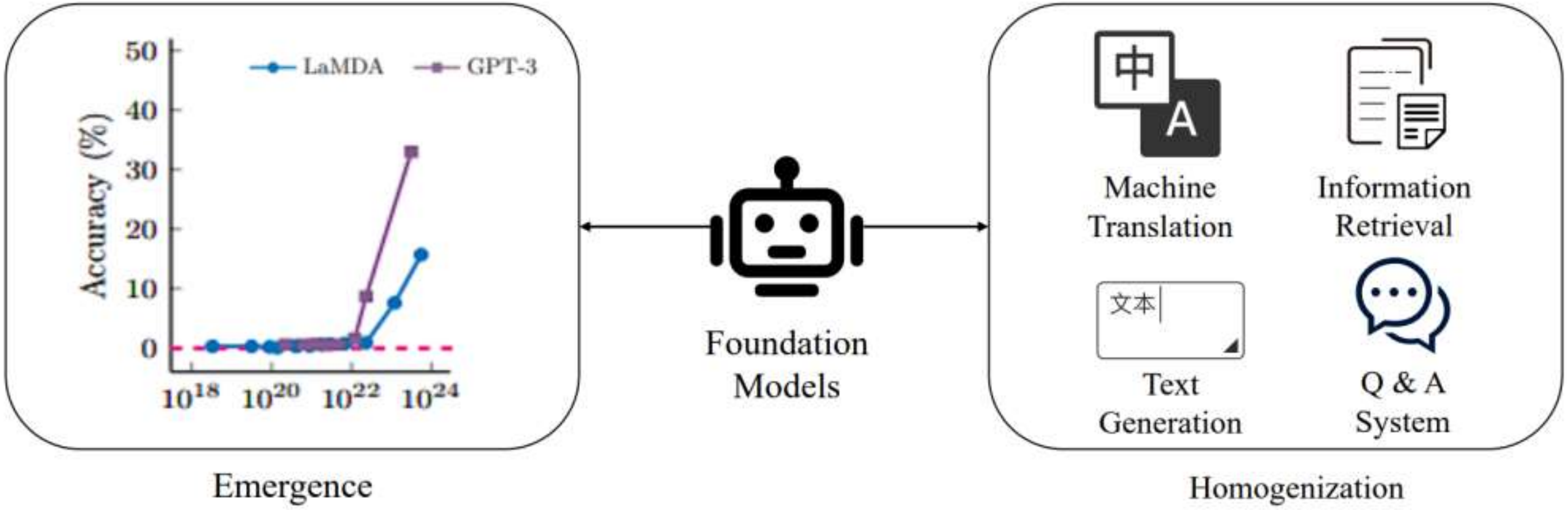
[Picture Credit] Lecture-style tutorial: Towards graph foundation models. *Chuan Shi*, et al.

Graph Foundation Models (GFM)

Foundation Models

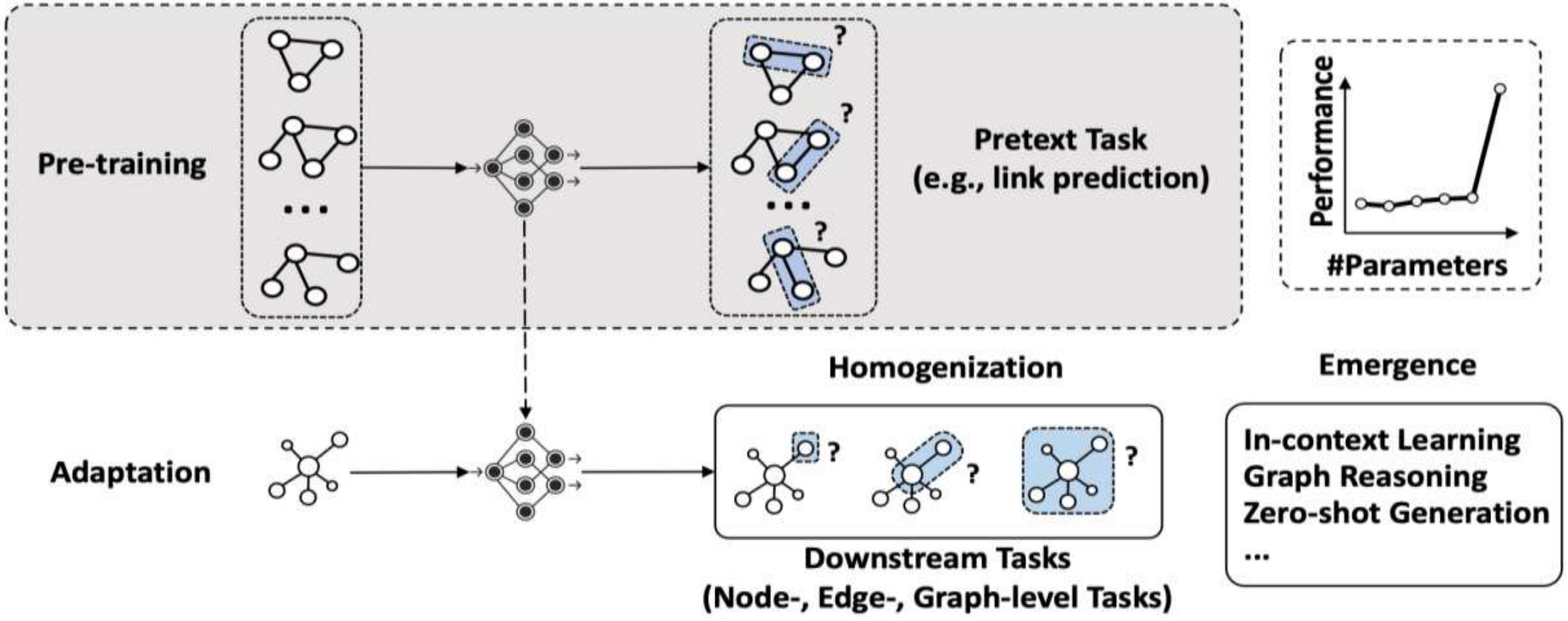
Two Characteristics of Foundation Models:

- **Emergence:** As scales up, it spontaneously manifests novel capabilities.
- **Homogenization:** Enables its deployment across diverse applications.



■ Graph Foundation Models (GFM)

- □ **Graph Foundation Models:** A **graph** foundation model (GFM) is a model pre-trained on extensive **graph** data, adapted for diverse downstream **graph** tasks.



[Picture Credit] Graph foundation models: Concepts, opportunities and challenges. *Jiawei Liu*, et al.

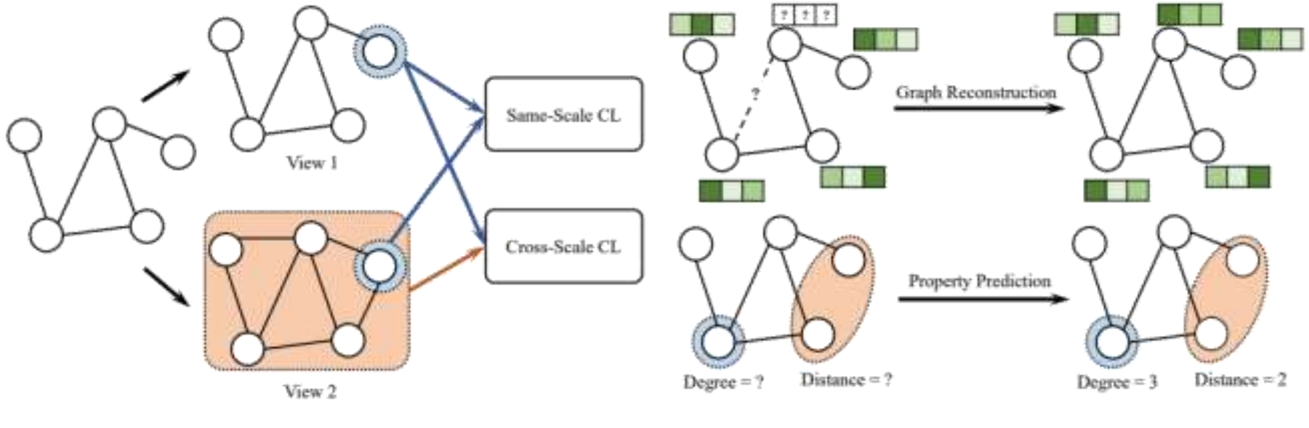
Graph Foundation Models (GFM)

- **Graph Foundation Models:** A **graph** foundation model (GFM) is a model pre-trained on extensive **graph** data, adapted for diverse downstream **graph** tasks.
- **Two Characteristics of GFMs:**
  - **Emergence:** Scaling up GNNs, redesign graph neural network architectures (e.g. deeper, Transformer-based) with far more parameters to unlock emergent capabilities.
  - **Homogenization:** Leverage large unlabeled graph datasets for self-supervised learning, then adapt one model to diverse downstream graph tasks

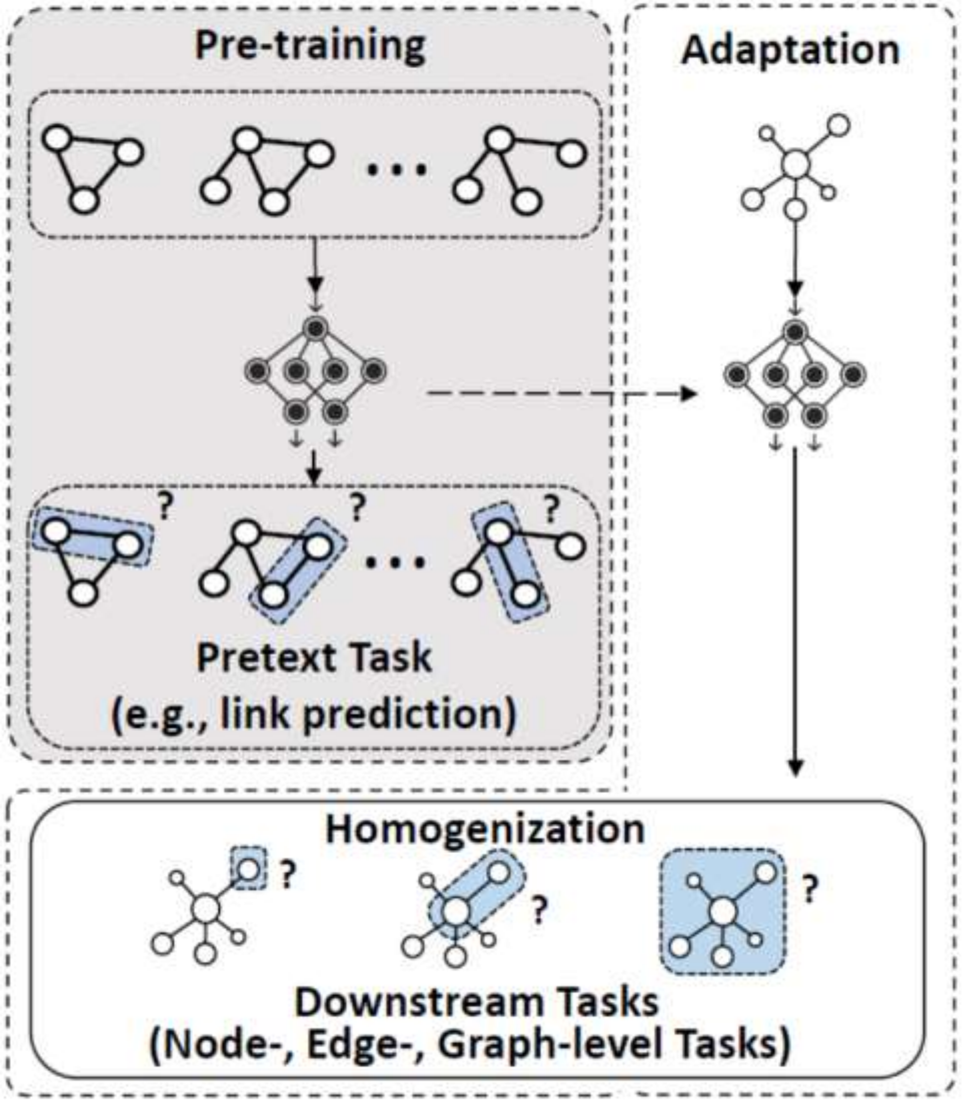
■ Graph Foundation Models (GFM)

□ Key Techniques of GFMs:

- **Pre-training:** neural networks are trained on a large graph dataset in a self-supervised manner (contrastive / generative)



(a) Contrastive Methods. (b) Generative Methods.



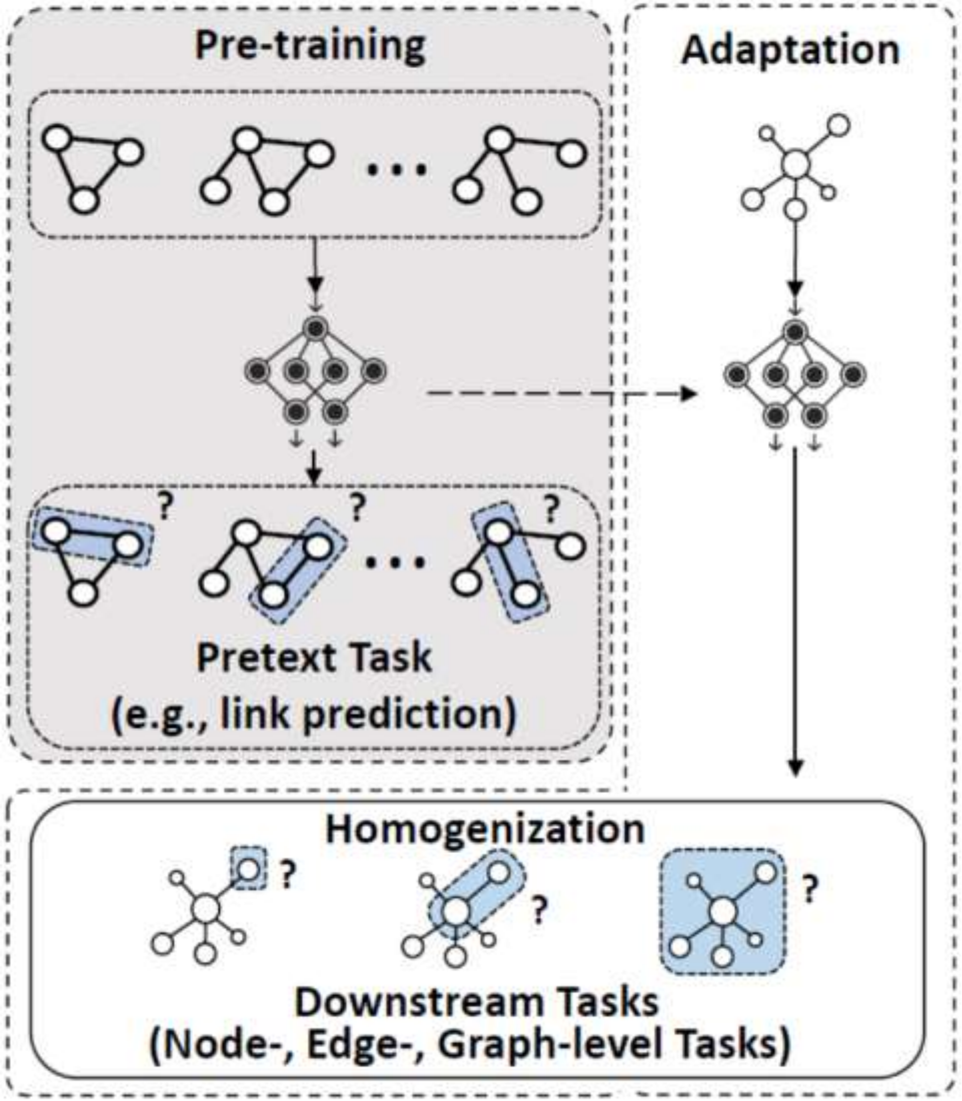
[Picture Credit] Graph foundation models: Concepts, opportunities and challenges. *Jiawei Liu*, et al.
Can large language models analyze graphs like professionals? a benchmark, datasets and models. *Xin Li*, et al.

■ Graph Foundation Models (GFM)

□ Key Techniques of GFMs:

- □ **Pre-training:** neural networks are trained on a large graph dataset in a self-supervised manner (contrastive / generative)

- □ **Adaptation:** adapt pre-trained models to specific downstream tasks or domains to enhance their performance



[Picture Credit] Graph foundation models: Concepts, opportunities and challenges. *Jiawei Liu*, et al.

Graph Foundation Models (GFM)

GFMs V.S. LLMs

		Language Foundation Model	Graph Foundation Model
Similarities	Goal	Enhancing the model's expressive power and its generalization across various tasks	
	Paradigm	Pre-training and Adaptation	
Intrinsic differences	Data	Euclidean data (text)	Non-Euclidean data (graphs) or a mixture of Euclidean (e.g., graph attributes) and non-Euclidean data
	Task	Many tasks, similar formats	Limited number of tasks, diverse formats
Extrinsic differences	Backbone Architectures	Mostly based on Transformer	No unified architecture
	Homogenization	Easy to homogenize	Difficult to homogenize
	Domain Generalization	Strong generalization capability	Weak generalization across datasets
	Emergence	Has demonstrated emergent abilities	No/unclear emergent abilities as of the time of writing

Graph Foundation Models (GFM)

Text-free GFMs

- single-domain pre-training, cross-task adaptation
- multi-domain pre-training, cross-domain / task adaptation

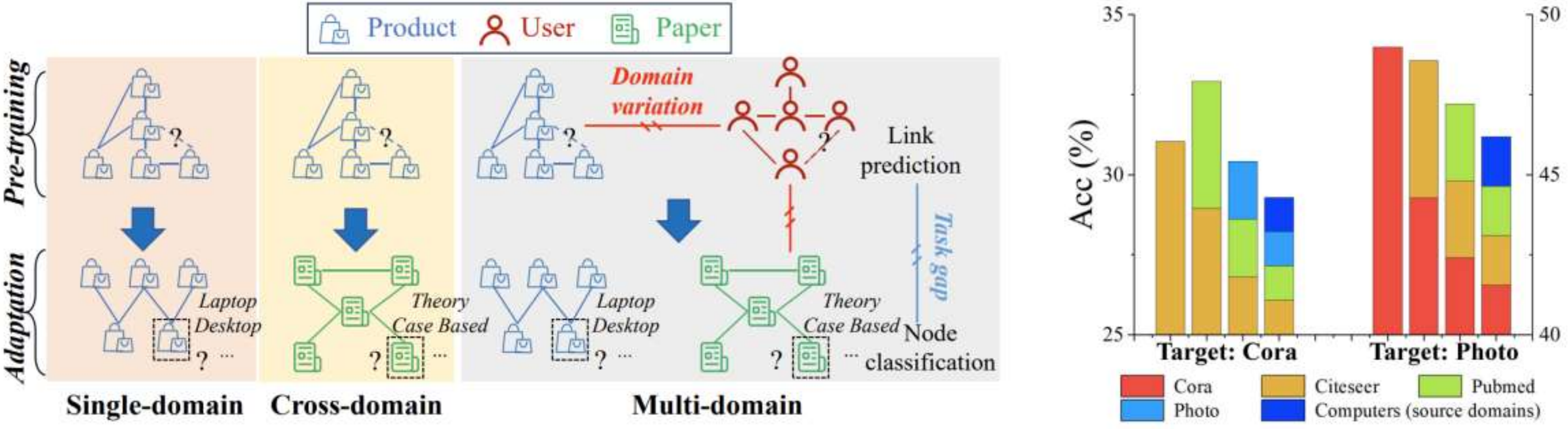
- homogeneous and heterogeneous GFMs
- robust GFMs (noise, adversarial attacks)
- stable GFMs (few-shot, fine-tuning)
- scalable GFMs

- GFMs with theoretical guarantees in knowledge transfer
- …

Graph Foundation Models (GFM)

Text-free GFMs

Challenge: Multi-domain conflicts & dimension inconsistency



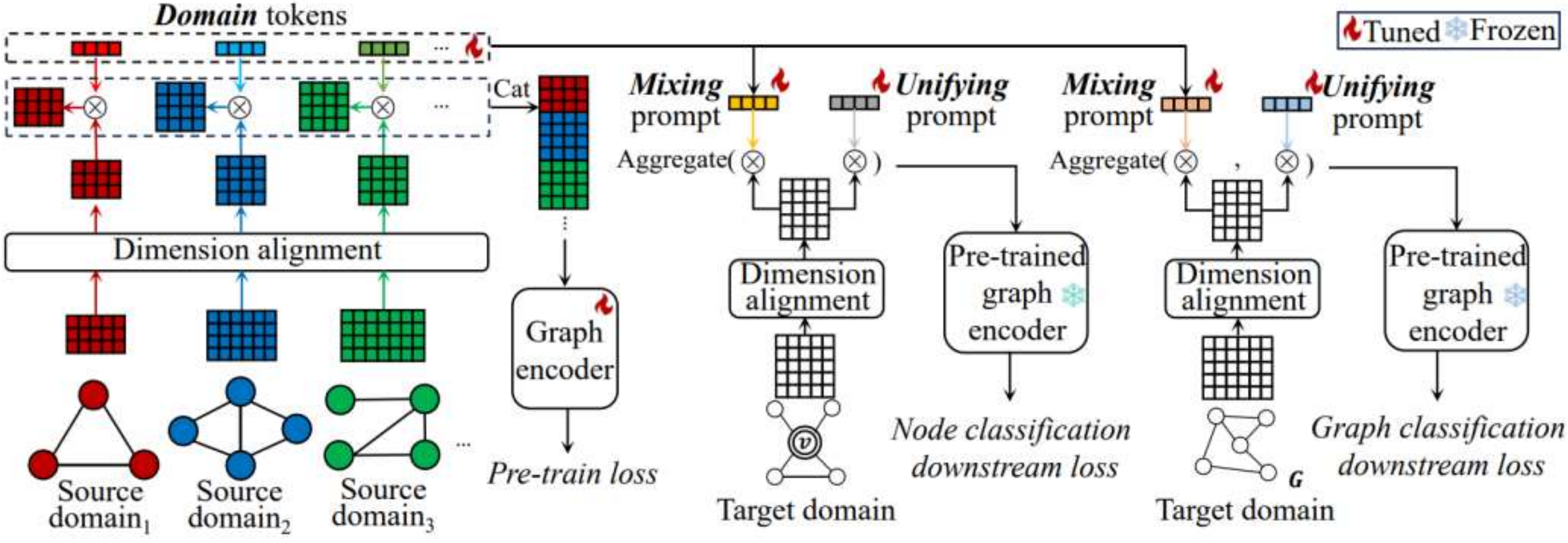
(a) Various transfer scenarios (b) Observation of domain conflicts

[Picture Credit] Text-free multi-domain graph pre-training: Toward graph foundation models. *Xingtong Yu*, et al.

Graph Foundation Models (GFM)

Text-free GFMs

Challenge: Multi-domain conflicts & dimension inconsistency



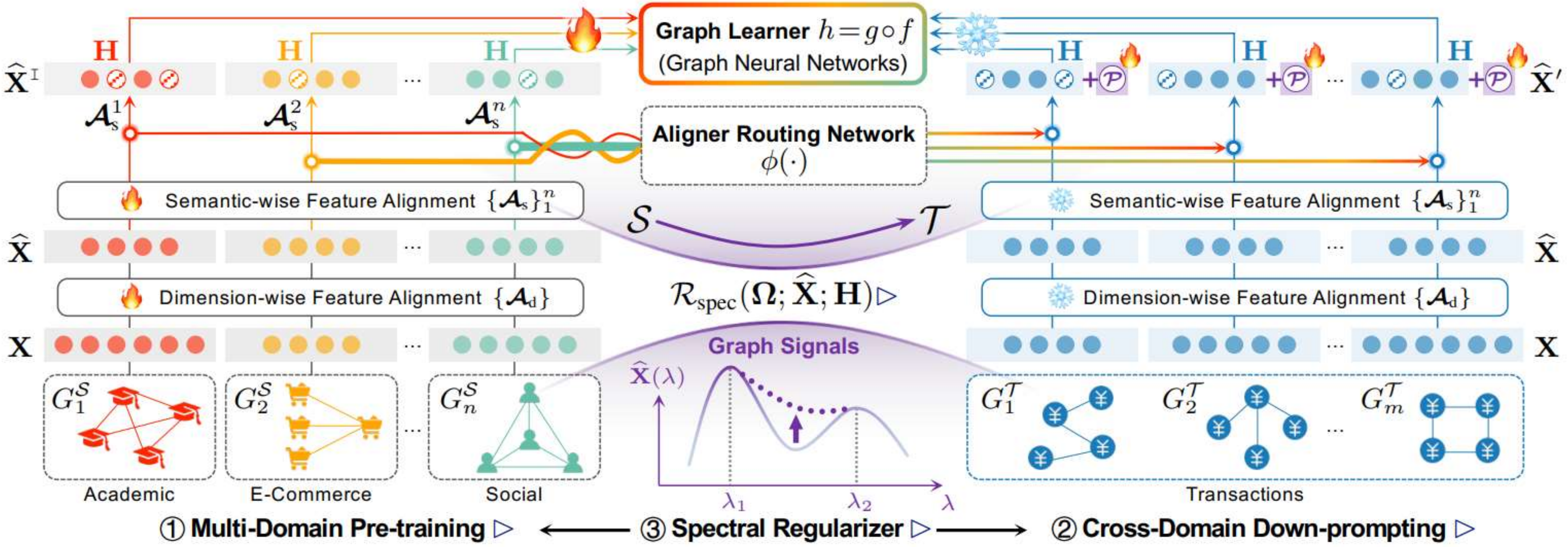
(a) Cross domain pre-training (b) Downstream node classification / graph classification

[Picture Credit] Text-free multi-domain graph pre-training: Toward graph foundation models. *Xingtong Yu*, et al.

■Graph Foundation Models (GFM)

□ Text-free GFMs

□ Challenge: Lack of theoretical guarantees in knowledge transfer

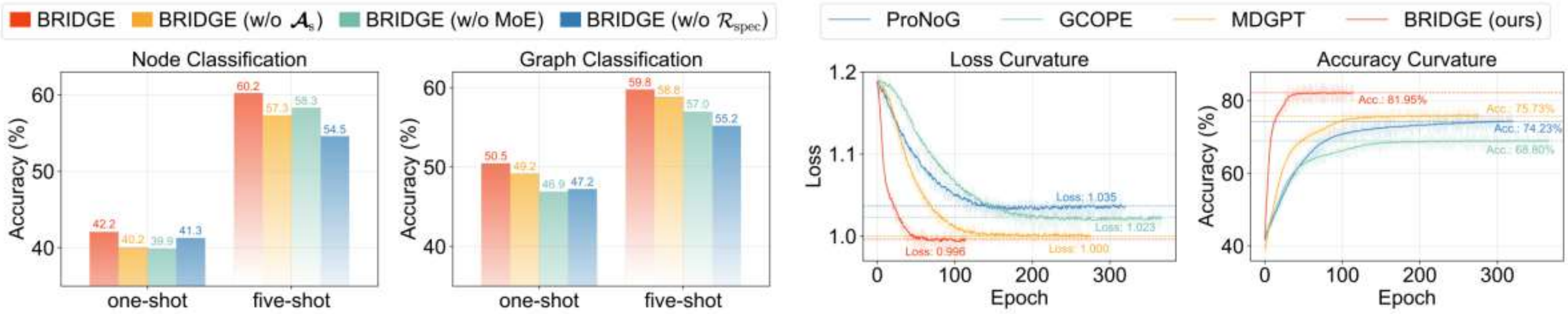


[Picture Credit] How much can transfer? BRIDGE: Bounded multi-domain graph foundation model with generalization guarantees. *Haonan Yuan*, et al.

■Graph Foundation Models (GFM)

□ Text-free GFMs

□ Challenge: Lack of theoretical guarantees in knowledge transfer

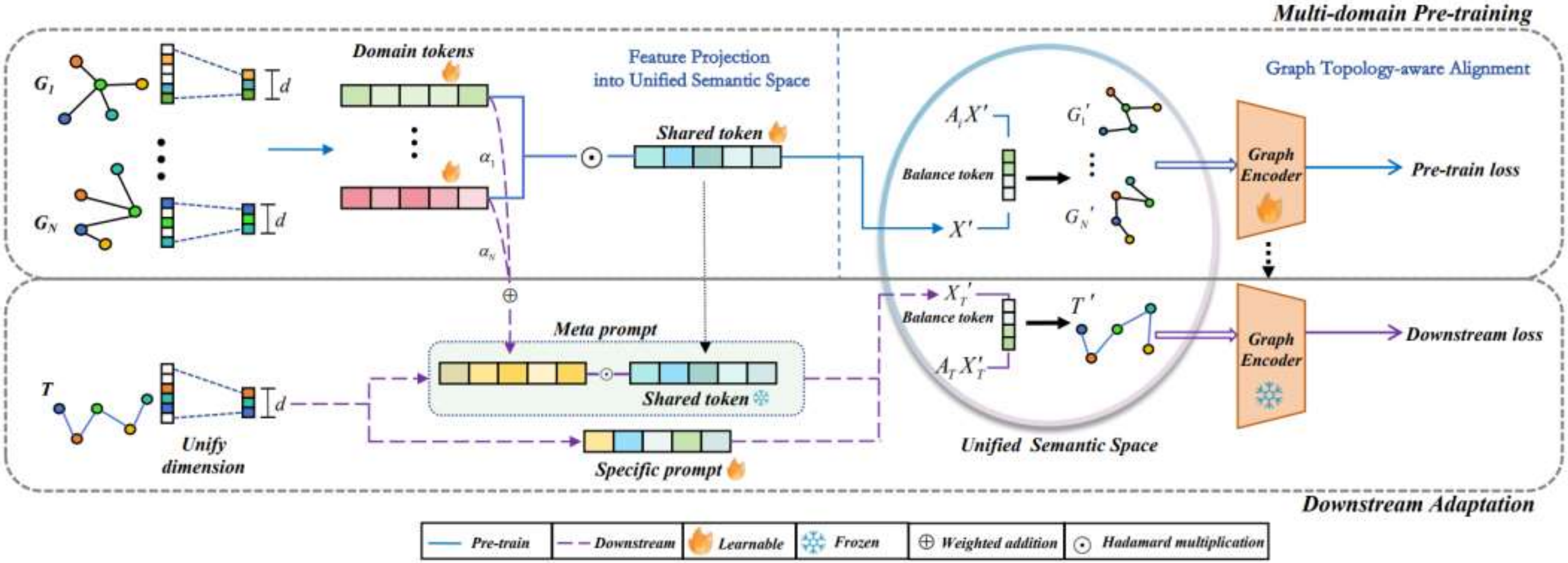


[Picture Credit] How much can transfer? BRIDGE: Bounded multi-domain graph foundation model with generalization guarantees. *Haonan Yuan*, et al.

Graph Foundation Models (GFM)

Text-free GFMs

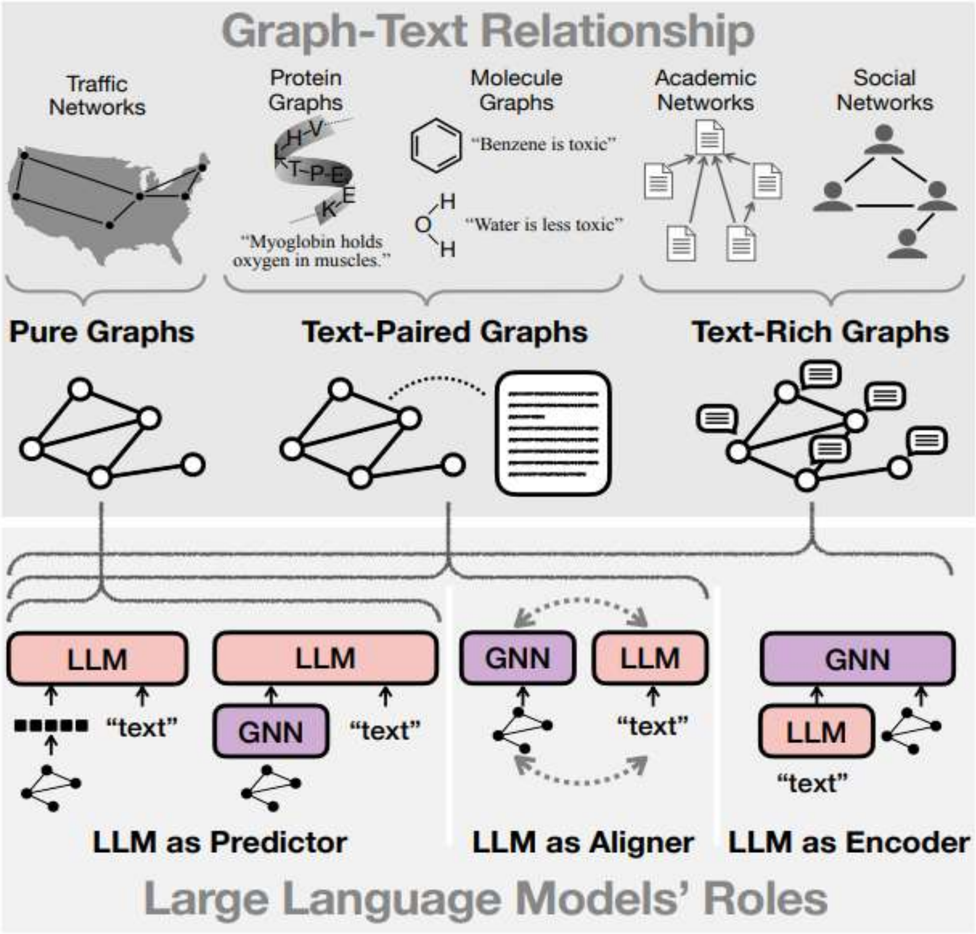
Challenge: Graph topology differences across domains

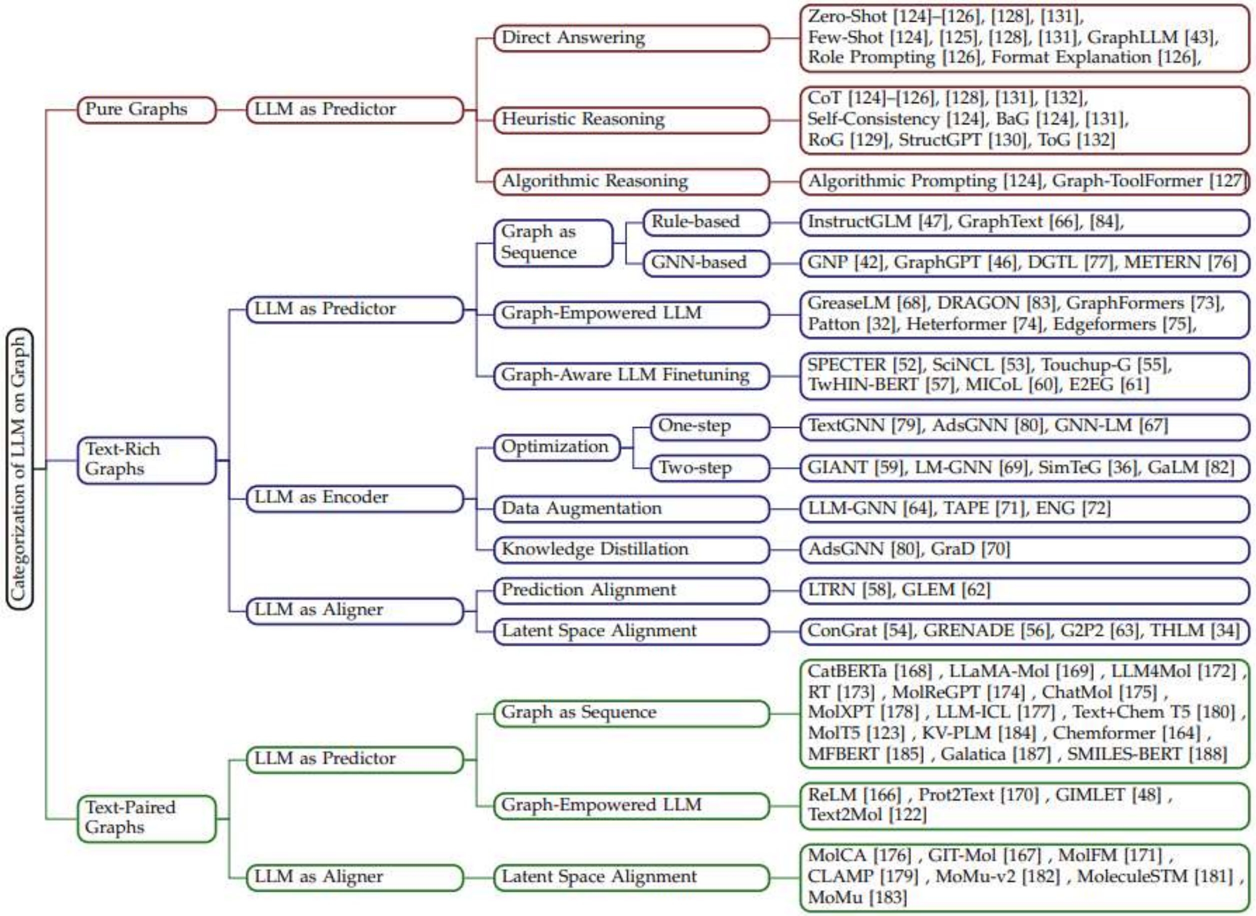


[Picture Credit] Multi-domain graph foundation models: Robust knowledge transfer via topology alignment. *Shuo Wang*, et al.

Graph Foundation Models (GFM)

Text-attributed GFMs

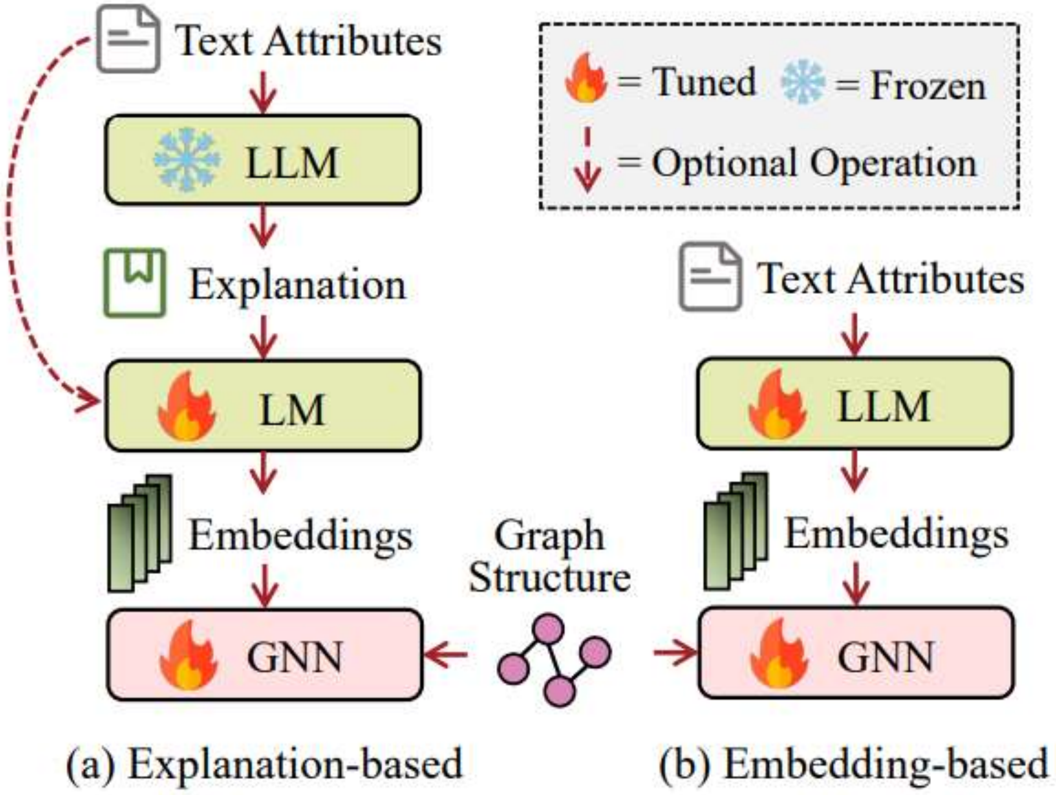




Graph Foundation Models (GFM)

Text-attributed GFMs

LLM-as-enhancer / encoder



Enhancement: $e_i = f_{\mathrm{LLM}}(t_i, p),\ \mathbf{x}_i = f_{\mathrm{LM}}(e_i, t_i),$

Graph Learning: $\mathbf{H} = f_{\mathrm{GNN}}(\mathbf{X}, \mathbf{A}),$

Enhancement: $\mathbf{x}_i = f_{\mathrm{LLM}}(t_i),$

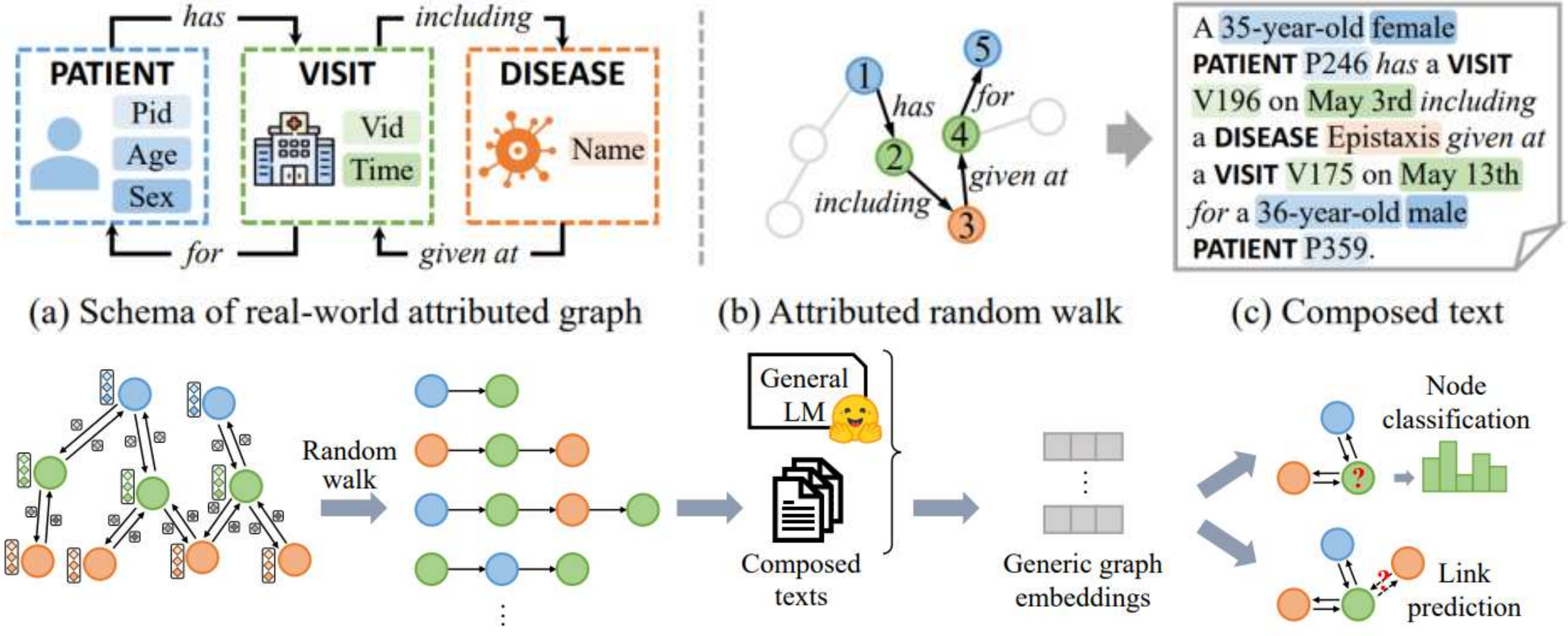
Graph Learning: $\mathbf{H} = f_{\mathrm{GNN}}(\mathbf{X}, \mathbf{A}).$

[Picture Credit] A survey of graph meets large language model: progress and future directions. *Yuhan Li*, et al.

Graph Foundation Models (GFM)

Text-attributed GFMs

LLM-as-enhancer / encoder



(a) Schema of real-world attributed graph (b) Attributed random walk (c) Composed text

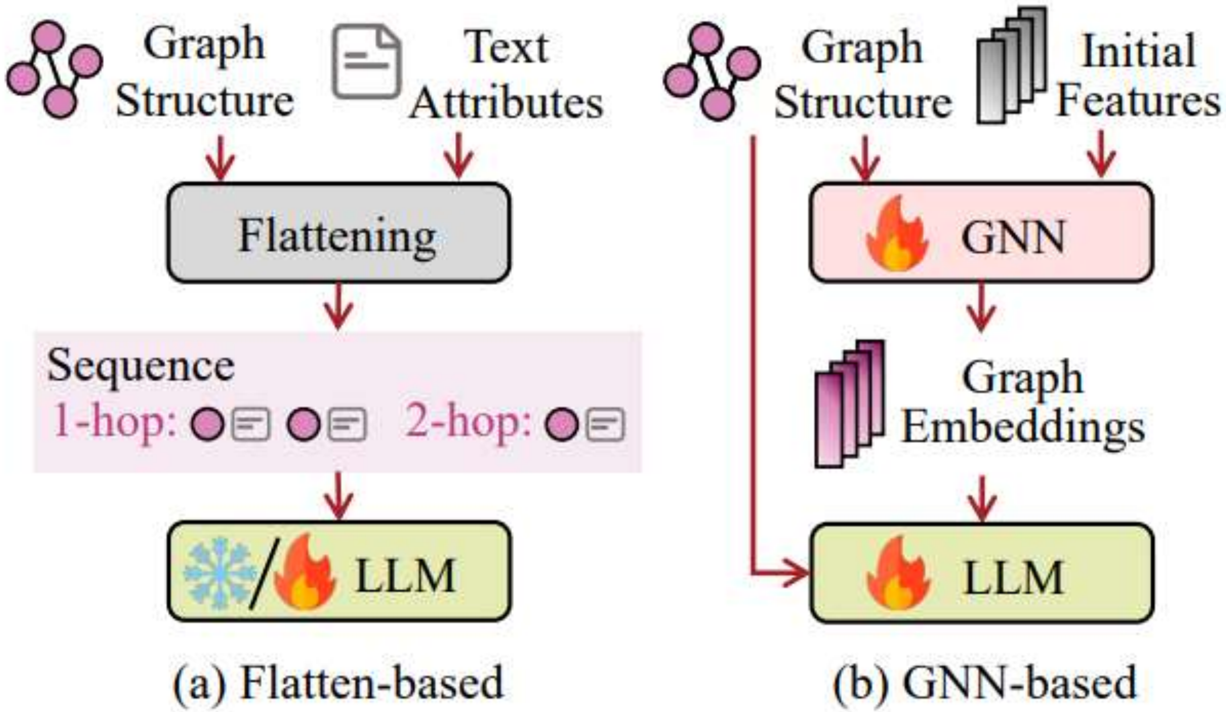
(a) Attributed RW-based textualization program (b) Graph-aware LM fine-tuning (c) Downstream tasks

[Picture Credit] A survey of graph meets large language model: progress and future directions. *Yuhan Li*, et al.

Graph Foundation Models (GFM)

Text-attributed GFMs

LLM-as-predictor



Graph Flattening: $G_{seq} = \texttt{Flat}(\mathcal{V}, \mathcal{E}, \mathcal{T}, \mathcal{J}),$

Prediction: $\tilde{Y} = \texttt{Parse}(f_{\text{LLM}}(G_{seq}, p)),$

Graph Learning: $\mathbf{H} = f_{\text{GNN}}(\mathbf{X}, \mathbf{A}),$

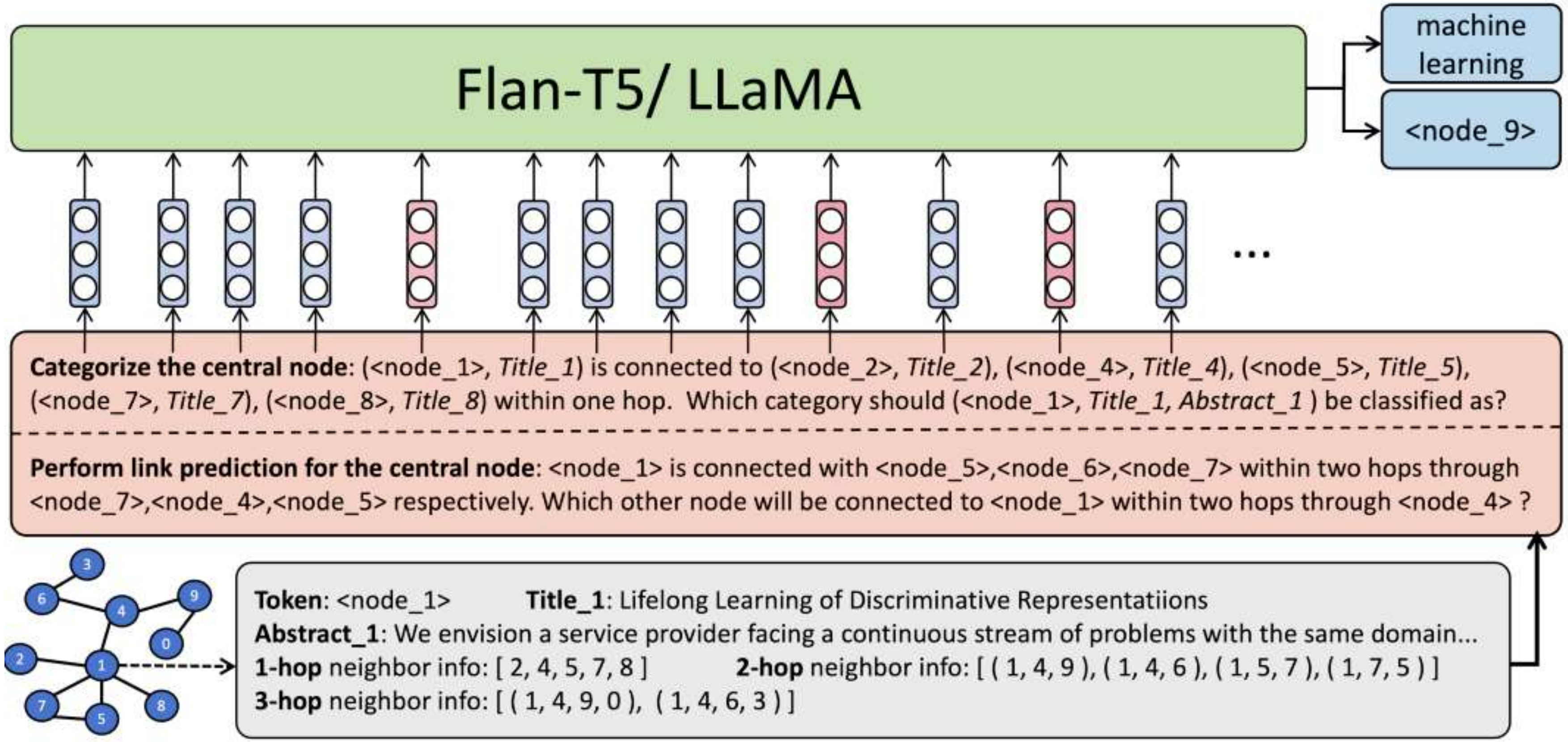
Prediction: $\tilde{Y} = \texttt{Parse}(f_{\text{LLM}}(\mathbf{H}, p)),$

[Picture Credit] A survey of graph meets large language model: progress and future directions. *Yuhan Li*, et al.

Graph Foundation Models (GFM)

Text-attributed GFMs

LLM-as-predictor

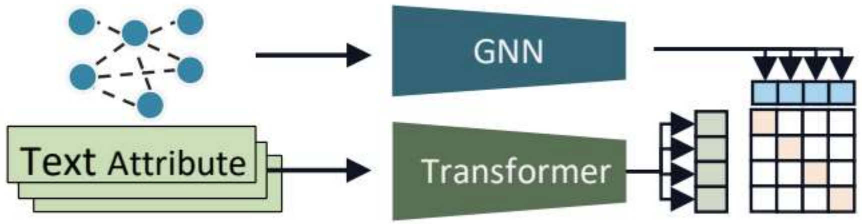


[Picture Credit] Language is all a graph needs. *Ruosong Ye*, et al.

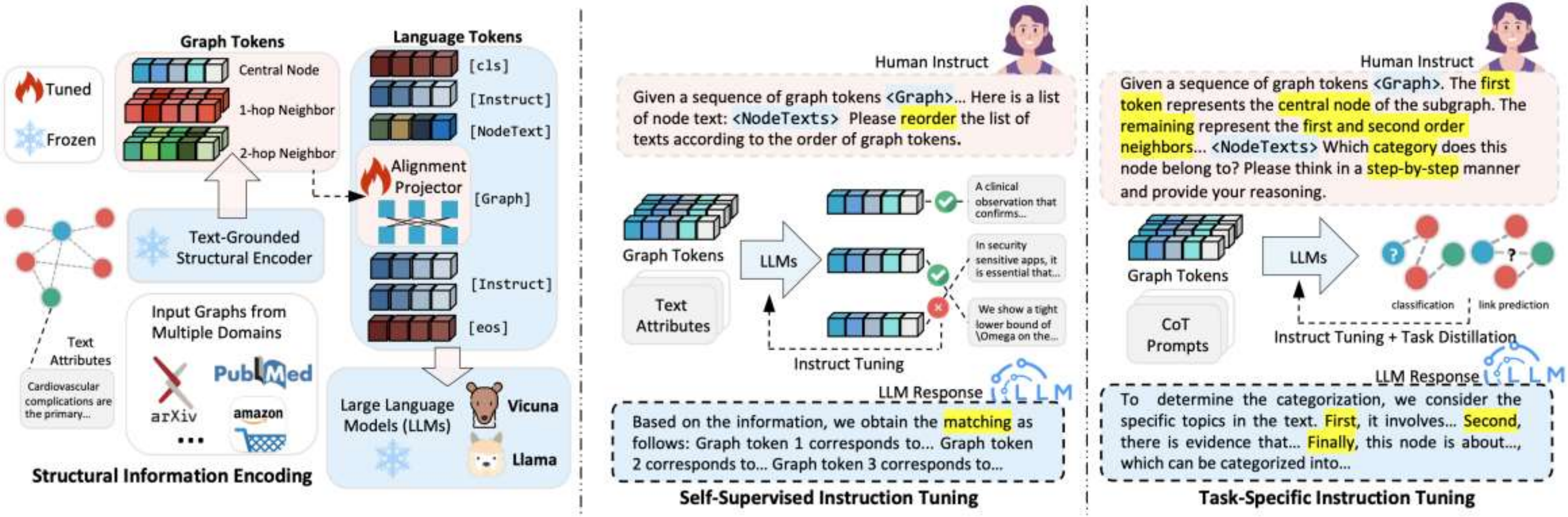
■ Graph Foundation Models (GFM)

□ Text-attributed GFMs

□ LLM-as-predictor



Workflow of text-structure alignment



[Picture Credit] GraphGPT: Graph instruction tuning for large language models. *Jiabin Tang*, et al.

Graph Foundation Models (GFM)

Text-attributed GFMs

LLM-as-predictor

Graph Matching

Graph Information: <graph>: **Central Node**: 68442, **Edge index**: [[…src node…],[…dst node…]], **Node list**: […]
Human Question: Given a sequence of graph tokens **<graph>** that constitute a subgraph of a citation graph, …. Here is a list of paper titles: 1. … 2. …, please reorder the list of papers according to the order of graph tokens.
GraphGPT Response: Based on the given graph tokens and the list of paper titles, we obtain the matching of graph tokens and papers: Graph token 1 corresponds to smt based induction methods for timed systems. Graph token 2 corresponds to …

Node Classification

Graph Information: <graph>: **Central Node**: 2, **Edge index**: [[…src node…],[…dst node…]], **Node list**: […]
Human Question: Given a citation graph: **<graph>** where the 0th node is the target paper, with the following information: Abstract: … Title: … Question: Which arXiv CS sub-category does this paper belong to? …
GraphGPT Response: cs.IT, cs.LG, cs.SP, cs.CV, cs.NA. The paper discusses the Restricted Isometry …. So, it is likely to belong to cs.IT…

Link Prediction

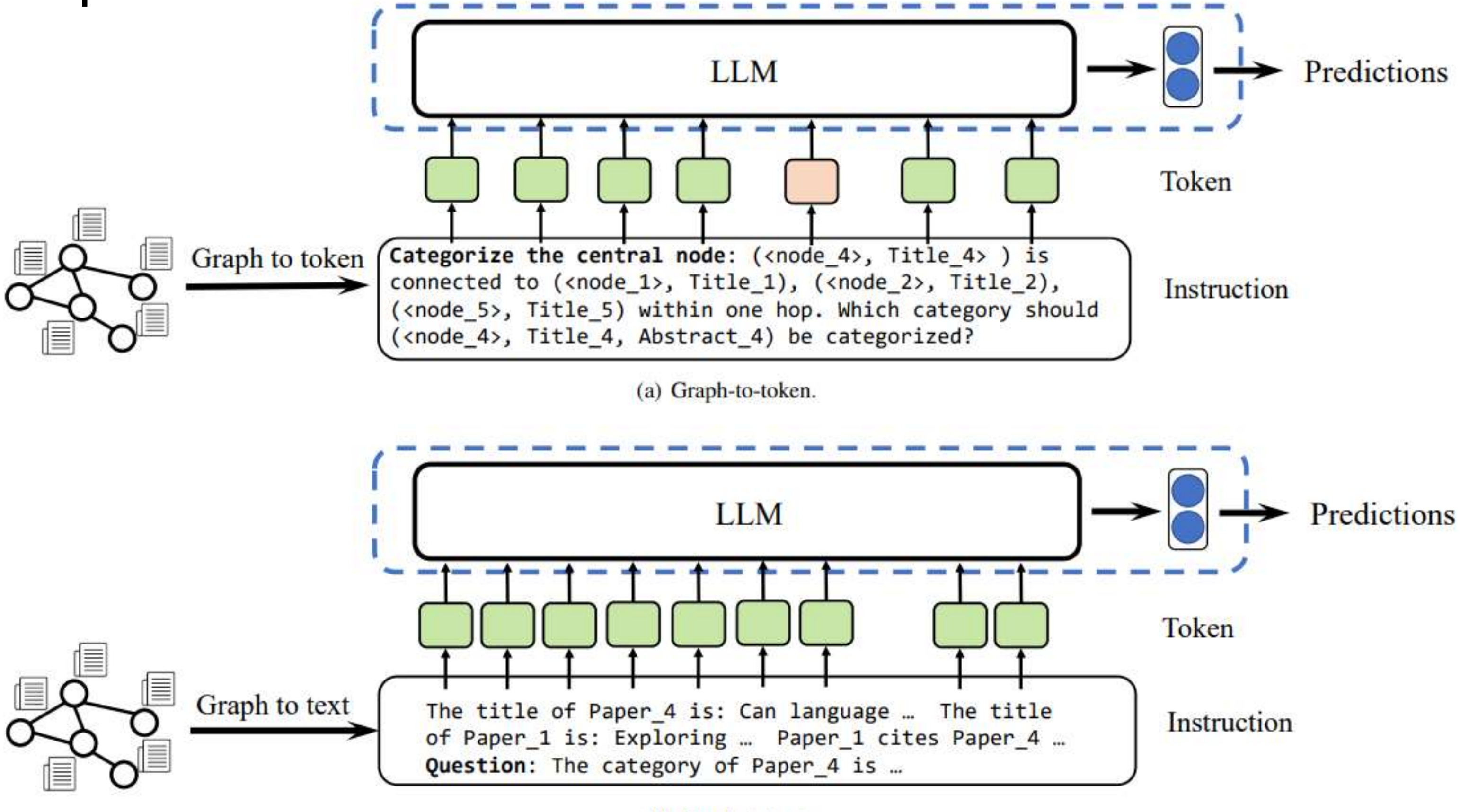
Graph Information: <graph>: **Central Node 1**: 8471, **Edge index 1**: [[…src node…],[…dst node…]], **Node list 1**: […]
<graph>: **Central Node 2**: 19368, **Edge index 2**: [[…src node…],[…dst node…]], **Node list 2**: […]
Human Question: Given a sequence of graph tokens: **<graph>** that constitute a subgraph of a citation graph, …. Abstract: … Titile: … and the other sequence of graph tokens: **<graph>**, … Abstract: … Title: …, are these two central nodes connected? Give me an answer of "yes" or "no".
GraphGPT Response: Yes, they are connected. Based on the first paper, …. And the second paper proposes ….

[Picture Credit] GraphGPT: Graph instruction tuning for large language models. *Jiabin Tang*, et al.

Graph Foundation Models (GFM)

Text-attributed GFMs

LLM-as-predictor



(a) Graph-to-token.

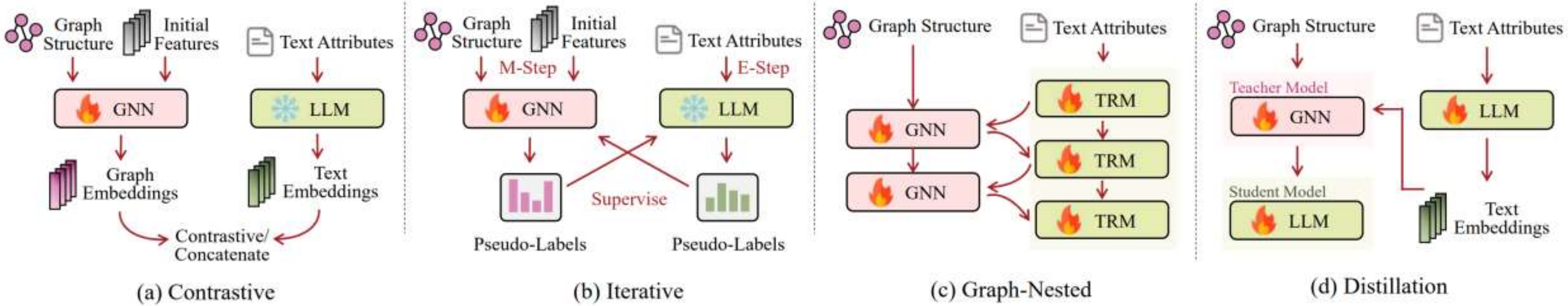
(b) Graph-to-text.

[Picture Credit] Graph foundation models: Concepts, opportunities and challenges. *Jiawei Liu*, et al.

■Graph Foundation Models (GFM)

☐ Text-attributed GFMs

☐ GNN-LLM-Alignment

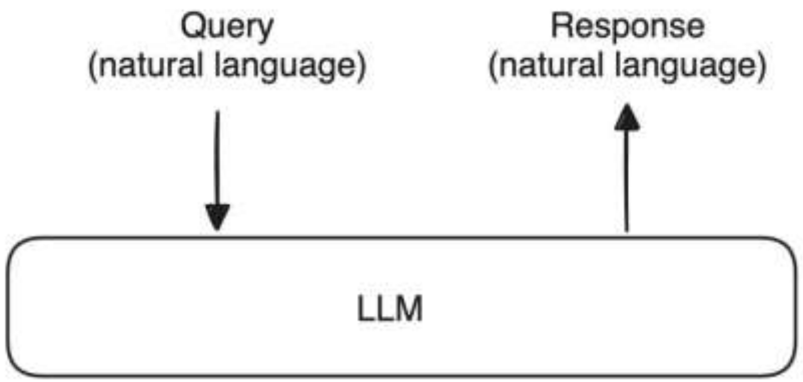


[Picture Credit] A survey of graph meets large language model: progress and future directions. *Yuhan Li*, et al.

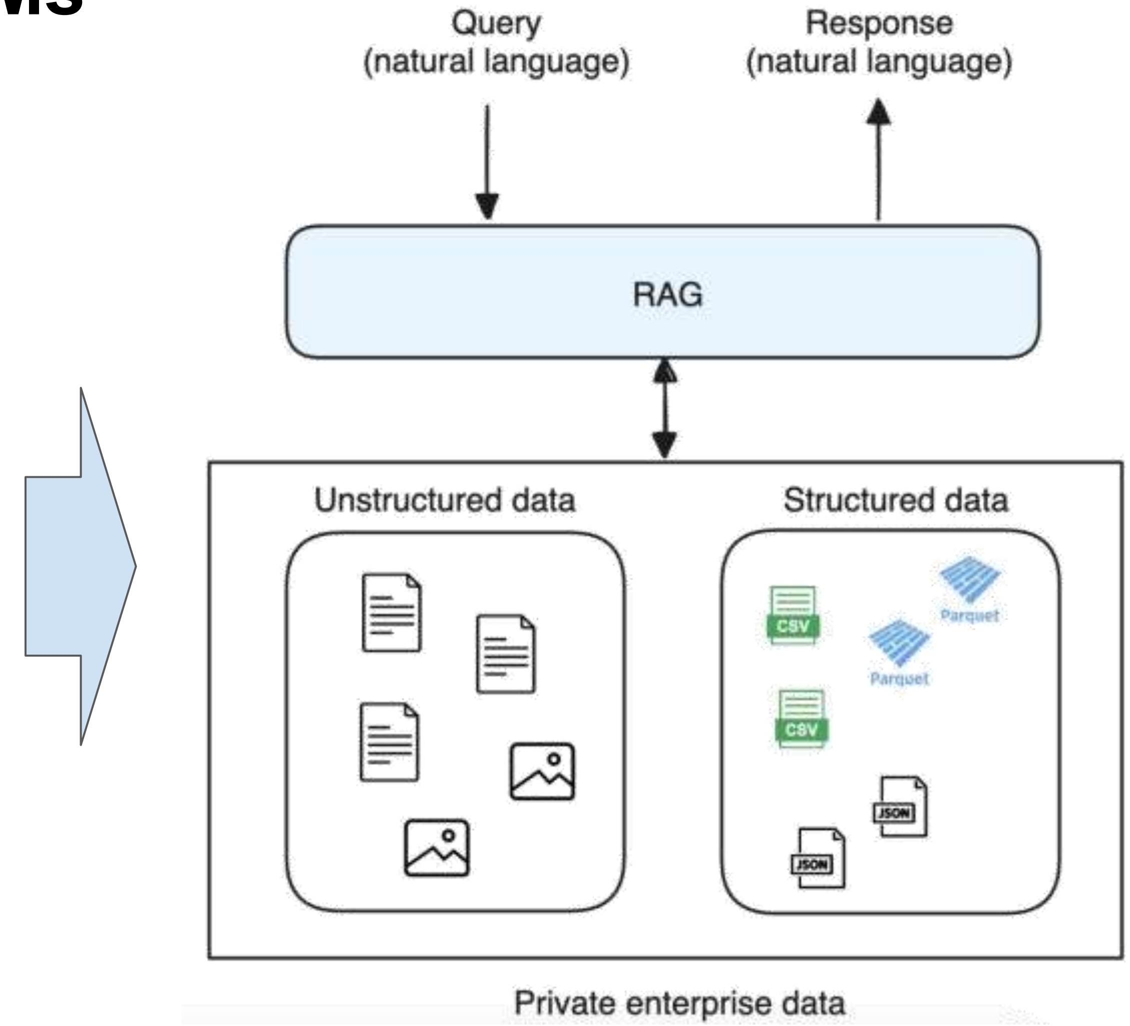
Graph Retrieval-Augmented Generation (GraphRAG)

Retrieval in the age of LLMs

"Chat with an LLM"

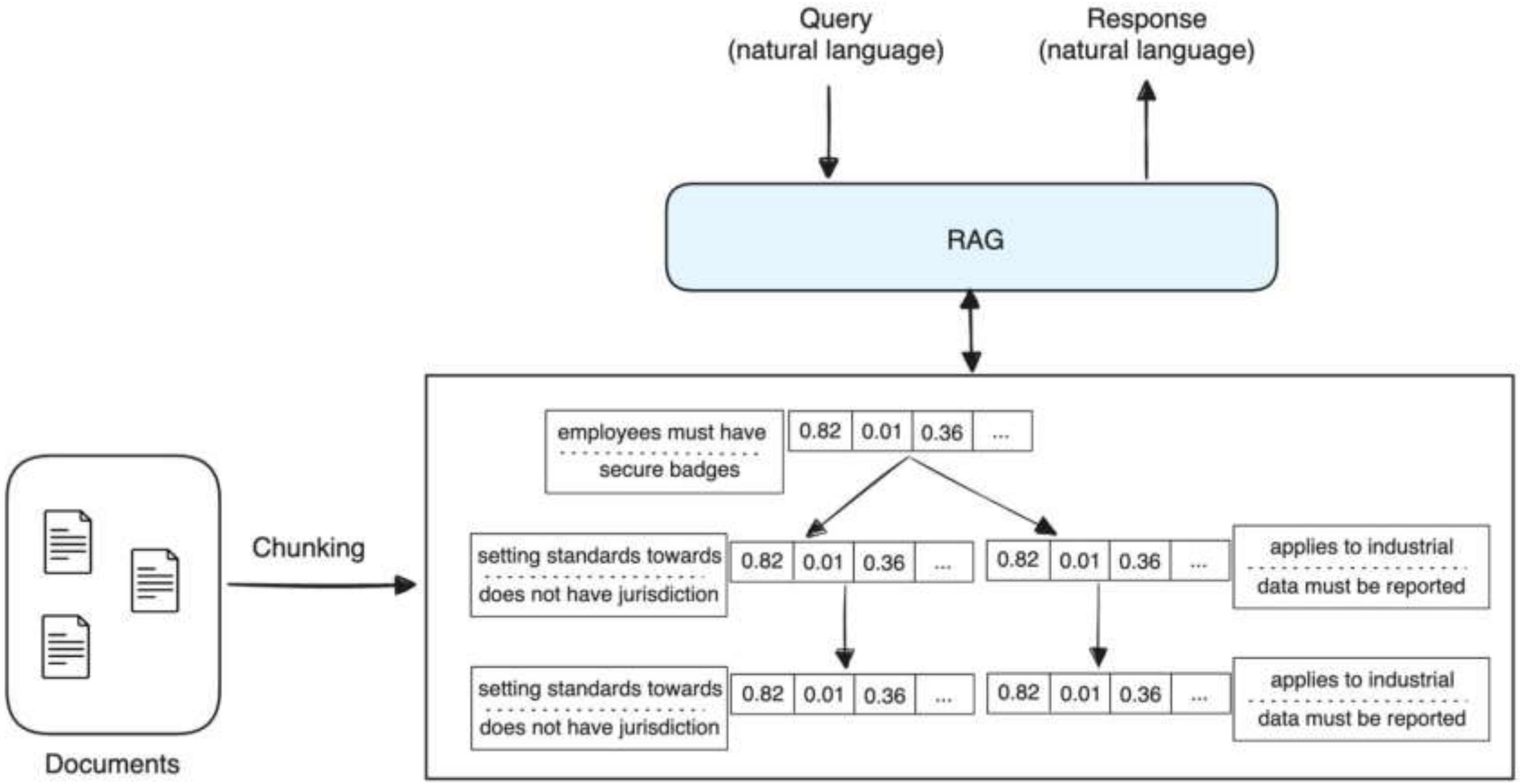


Cannot easily retrieve from private enterprise data



■Graph Retrieval-Augmented Generation (GraphRAG)

□A deeper look at traditional RAG

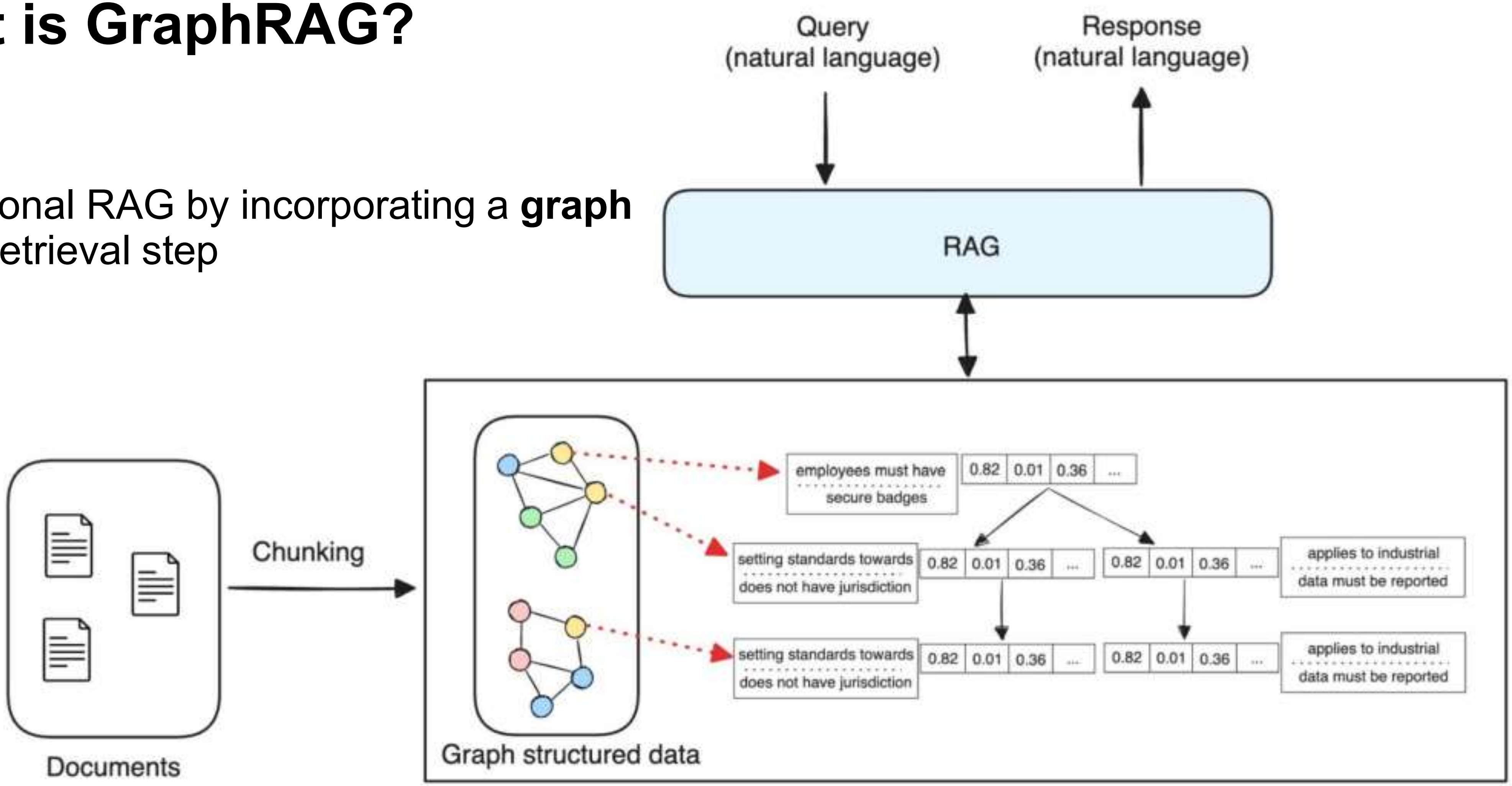


Vector database (retrieve top-k documents)

Graph Retrieval-Augmented Generation (GraphRAG)

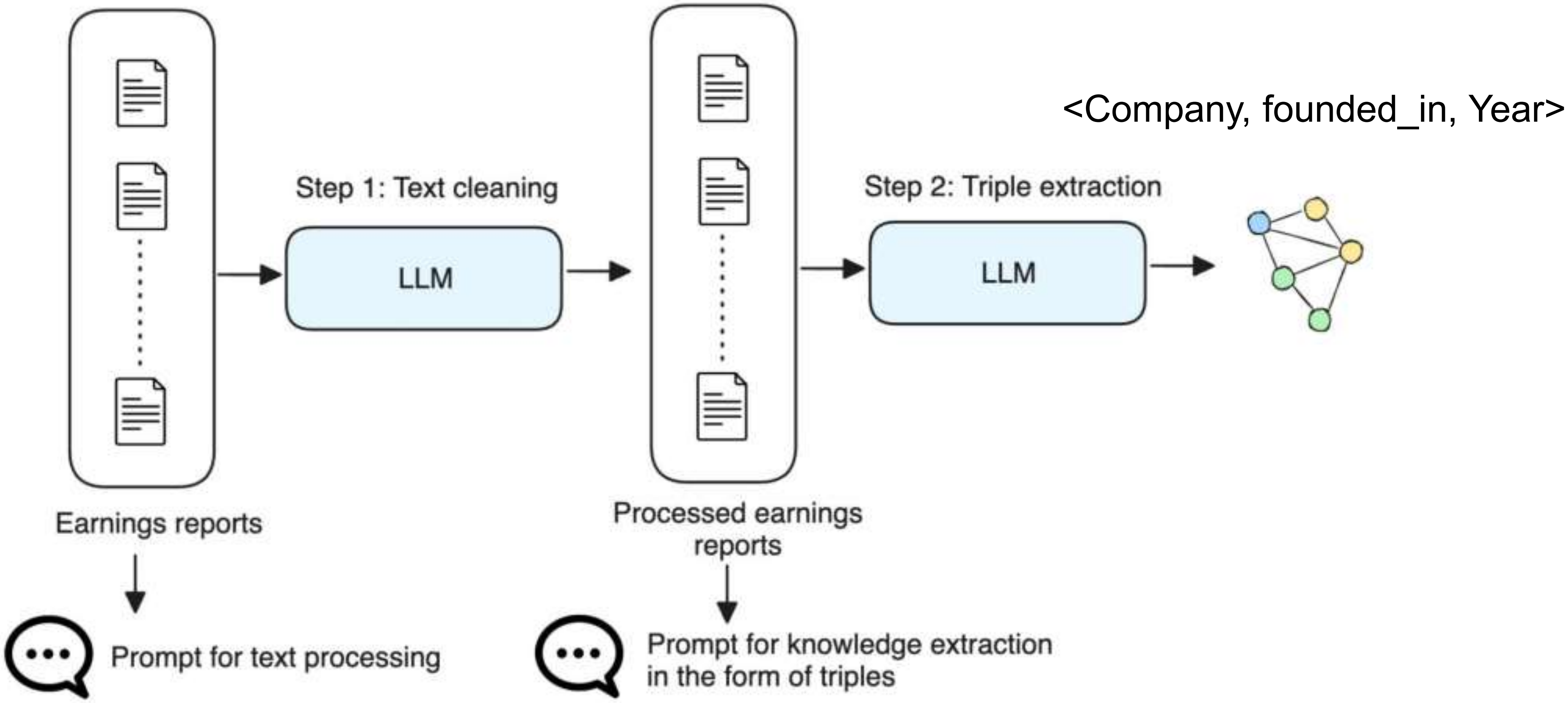
What is GraphRAG?

Extends traditional RAG by incorporating a **graph** as part of the retrieval step



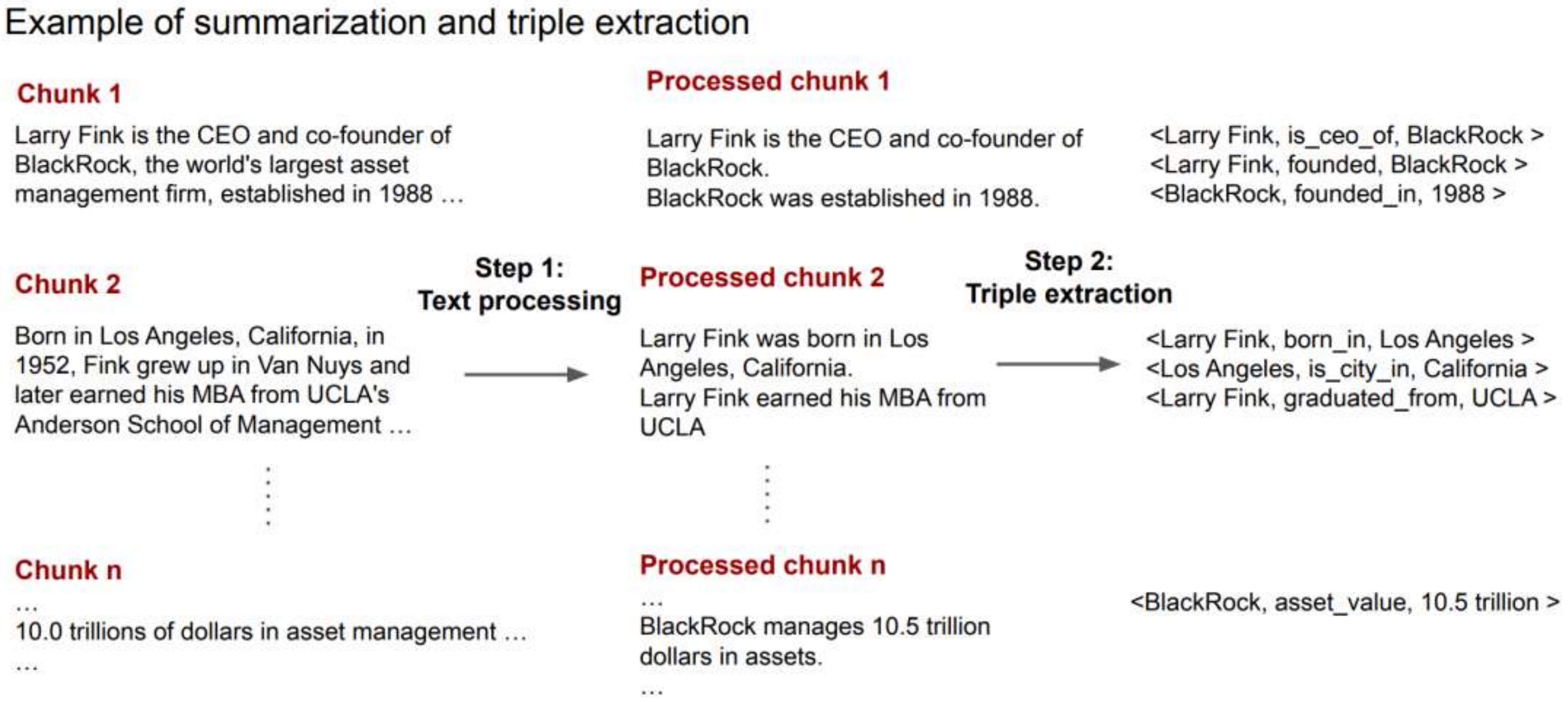
Graph Retrieval-Augmented Generation (GraphRAG)

GraphRAG pipeline



Graph Retrieval-Augmented Generation (GraphRAG)

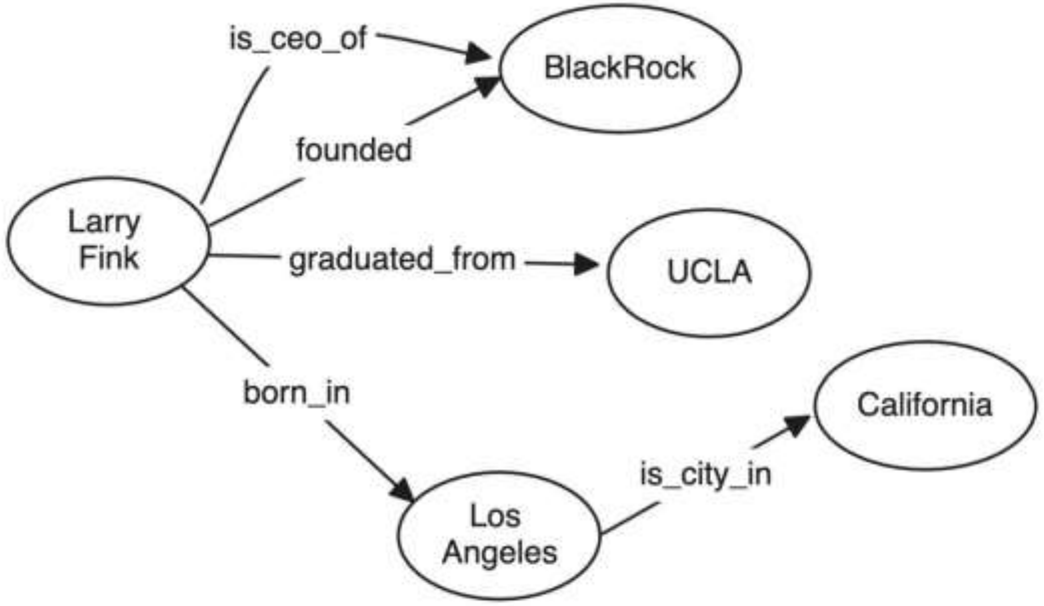
GraphRAG pipeline



■ Graph Retrieval-Augmented Generation (GraphRAG)

□ GraphRAG pipeline

Recall: Graphs can model simple sentences



Chunk 1

<Larry Fink, is_ceo_of, BlackRock >
<Larry Fink, founded, BlackRock >

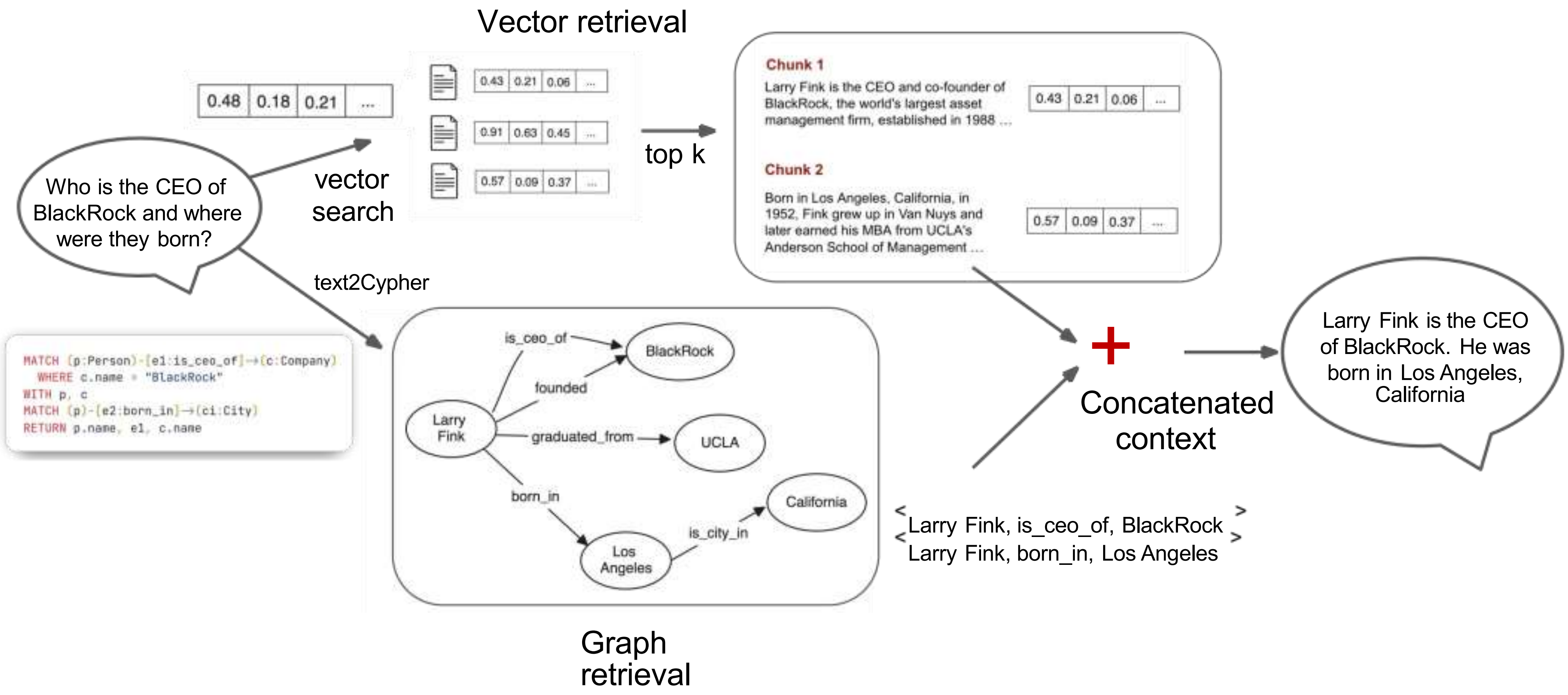
Chunk 2

<Larry Fink, born_in, Los Angeles >
<Los Angeles, is_city_in, California >
<Larry Fink, graduated_from, UCLA >

- Benefit 1: Information in disparate chunks are now **directly connected**
- Benefit 2: Triples are a form of capturing the **essence** of text chunks in very simple sentences
- Benefit 3: Can now put the triples into a graph DB where you can query it using a **query language**

■Graph Retrieval-Augmented Generation (GraphRAG)

□GraphRAG pipeline



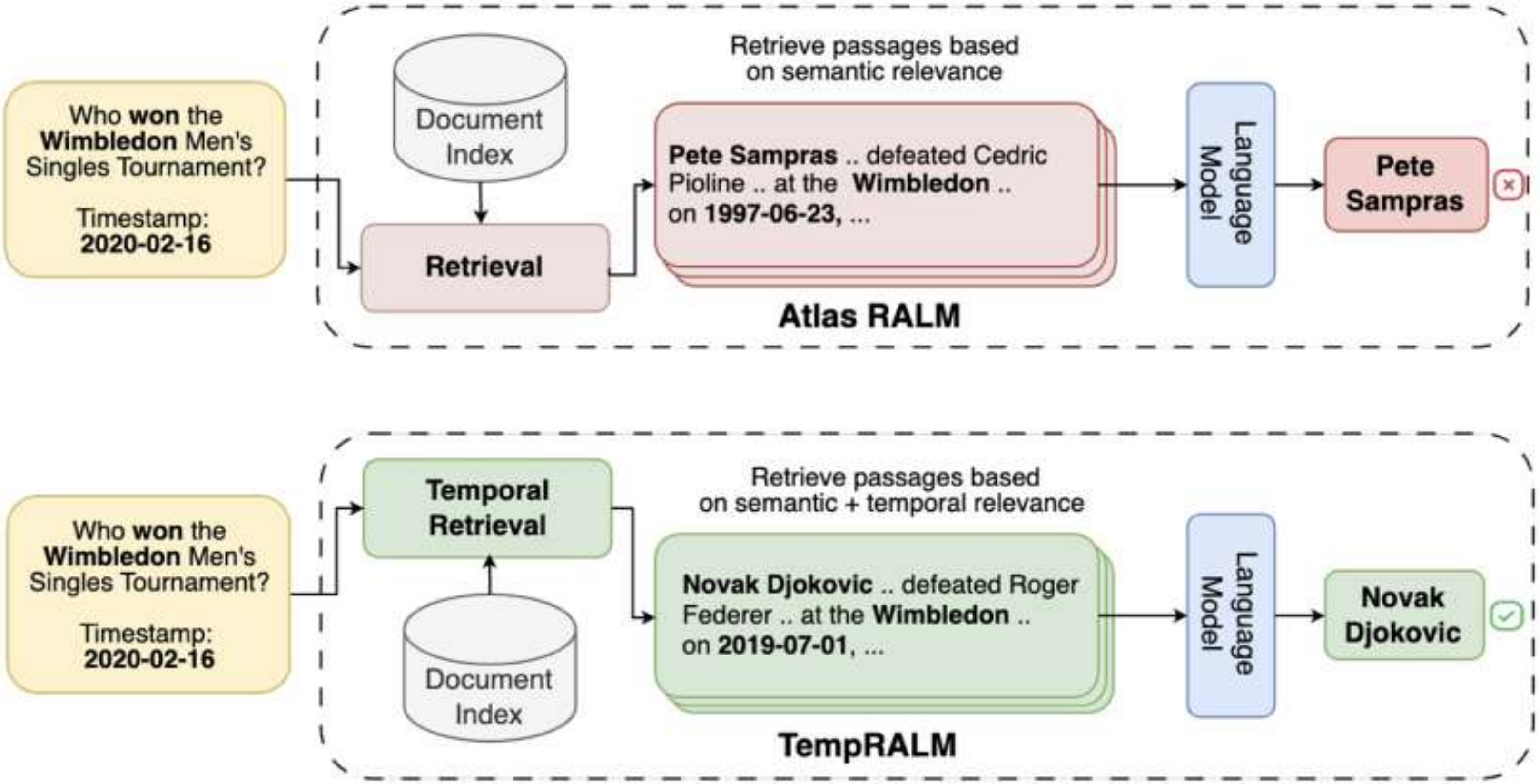
Graph Retrieval-Augmented Generation (GraphRAG)

Why Dynamic GraphRAG?

- The intrinsic structure of a graph (with nodes and edges) can model the temporal dynamics and evolution of events.
- Traversing that sub-graph lets the system return both semantic related and time closely event chains rather than scattered snippets.
- Presents retrieved events as a chronologically sorted timeline + a Time-CoT template that teaches the model inclusion, overlap & persistence rules

Graph Retrieval-Augmented Generation (GraphRAG)

Dynamic GraphRAGs



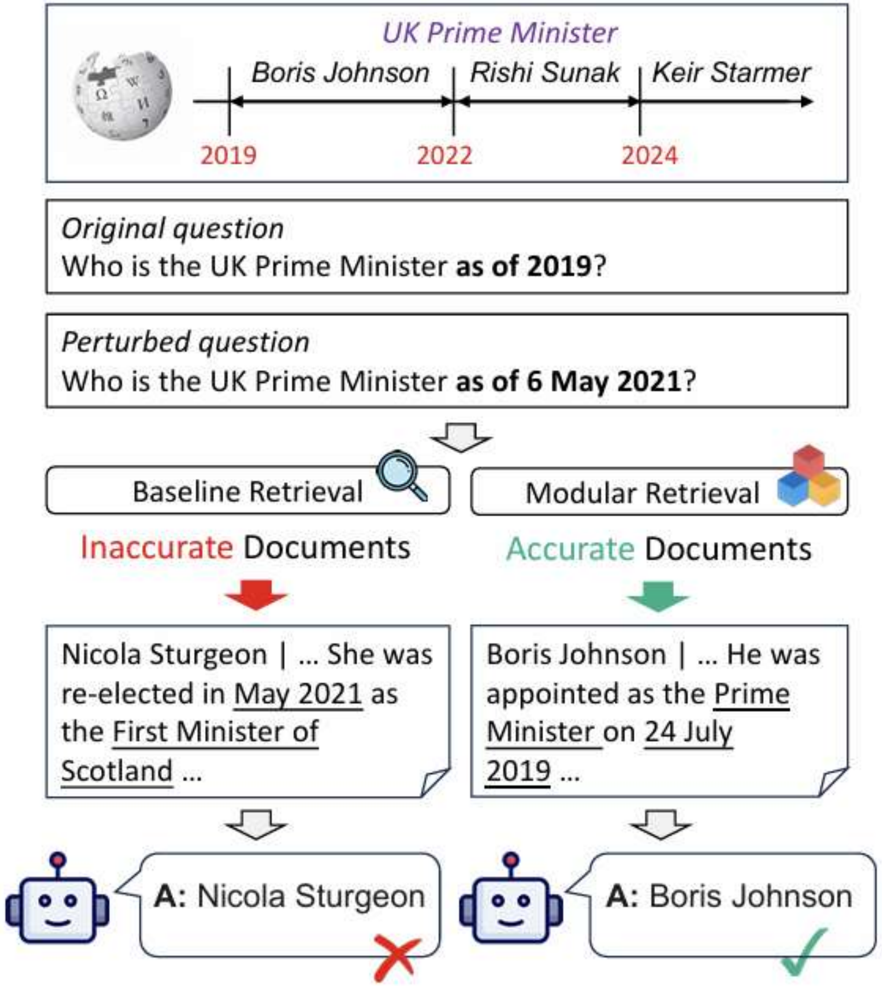
semantic score + temporal score

$$TempRet_t(q,d,qt,dt) = \begin{cases} s(q,d) + \tau(qt,dt) & \text{if } qt \geq dt \\ -\infty, & \text{otherwise} \end{cases}$$

[Picture Credit] It's about time: Incorporating temporality in retrieval augmented language models. *Anoushka Gade*, et al.

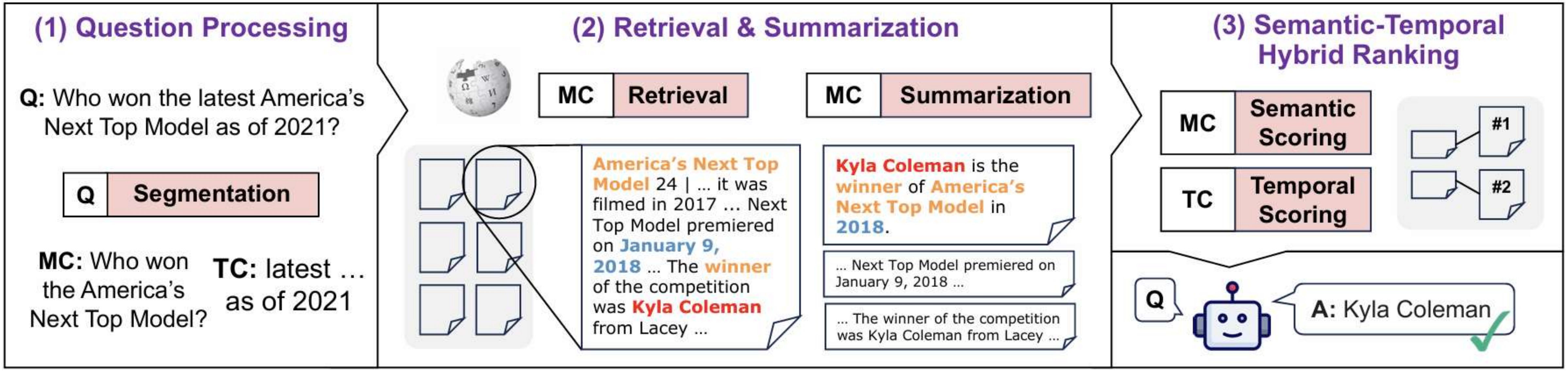
Graph Retrieval-Augmented Generation (GraphRAG)

Dynamic GraphRAGs



■Graph Retrieval-Augmented Generation (GraphRAG)

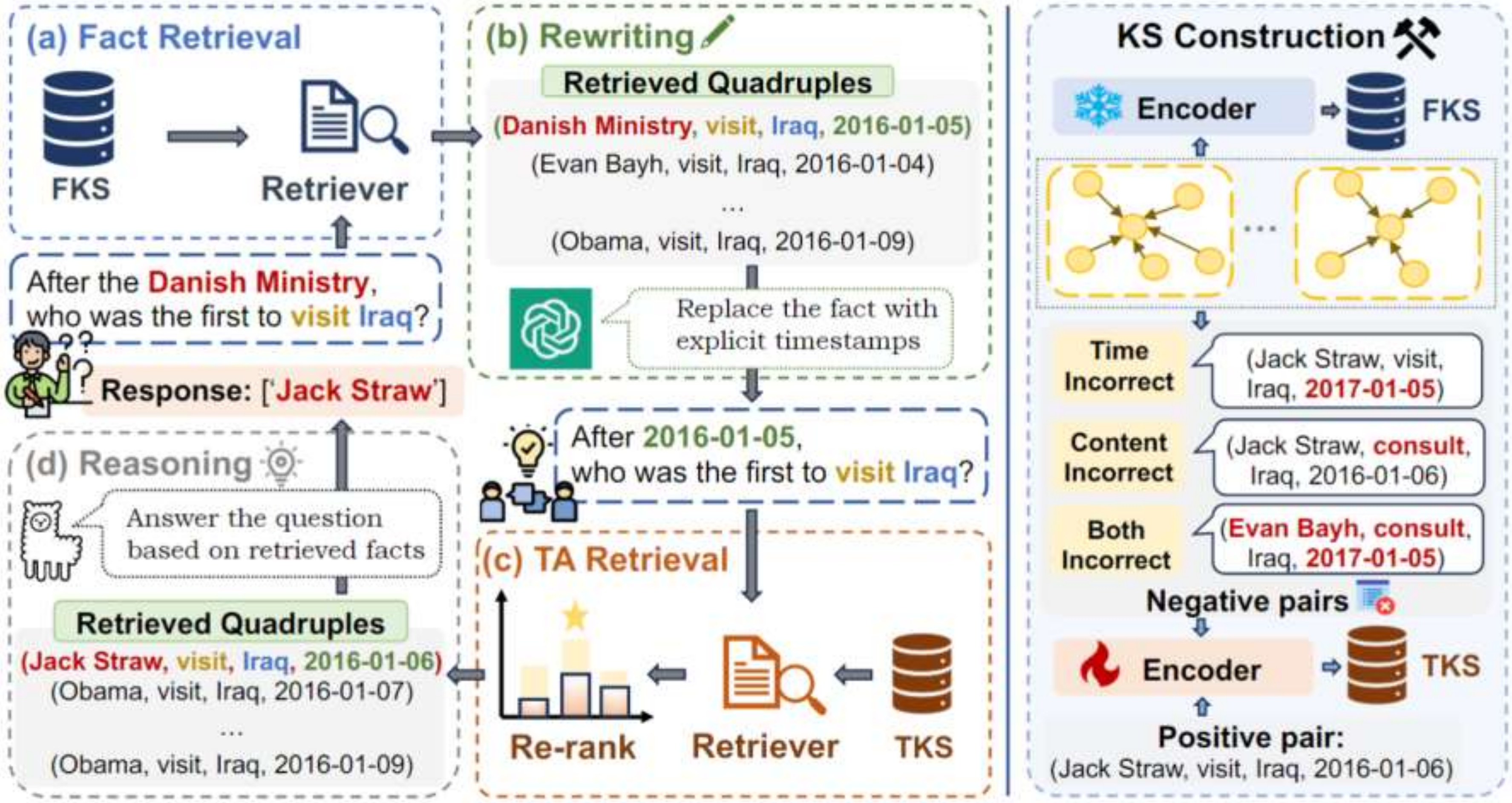
□Dynamic GraphRAGs



[Picture Credit] MRAG: A modular retrieval framework for time-sensitive question answering. *Siyue Zhang*, et al.

Graph Retrieval-Augmented Generation (GraphRAG)

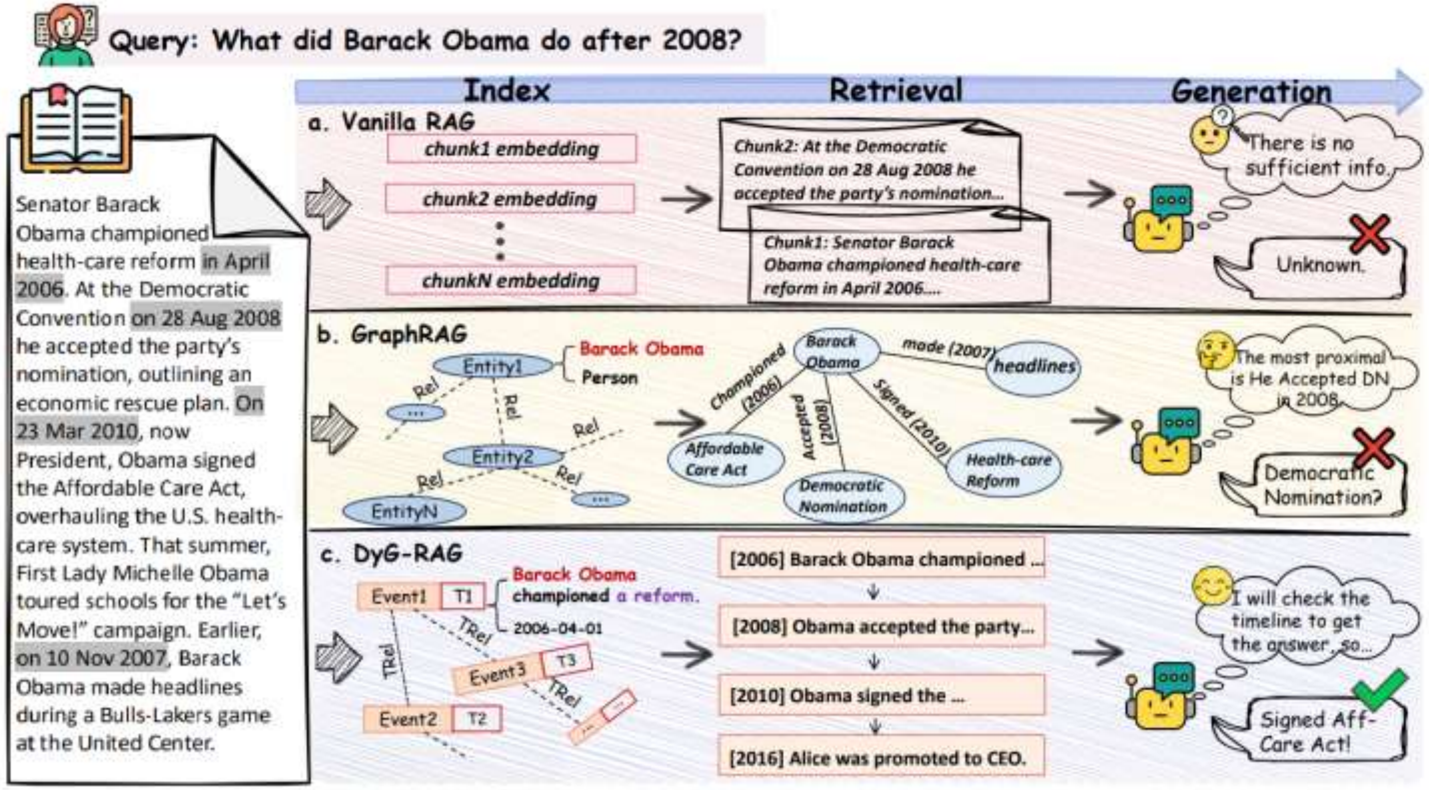
Dynamic GraphRAGs



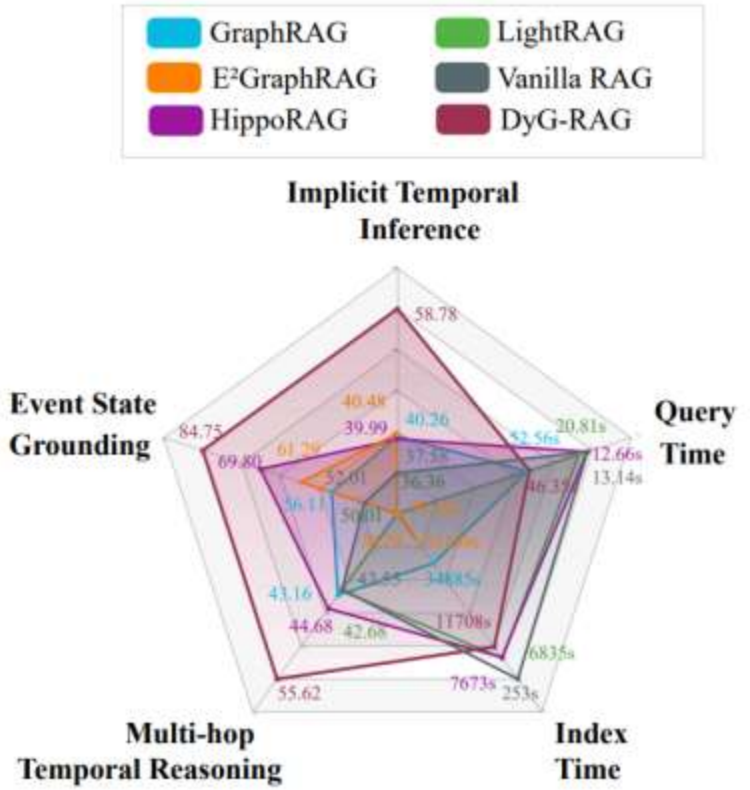
[Picture Credit] TimeR4: Time-aware retrieval-augmented large language models for temporal knowledge graph question answering. *Xinying Qian*, et al.

Graph Retrieval-Augmented Generation (GraphRAG)

Dynamic GraphRAGs



(a) Comparison of RAG pipeline



(b) Comparison of multi-dimensions

[Picture Credit] DyG-RAG: Dynamic graph retrieval-augmented generation with event-centric reasoning. *Qingyun Sun*, et al.

■ Graph Retrieval-Augmented Generation (GraphRAG)

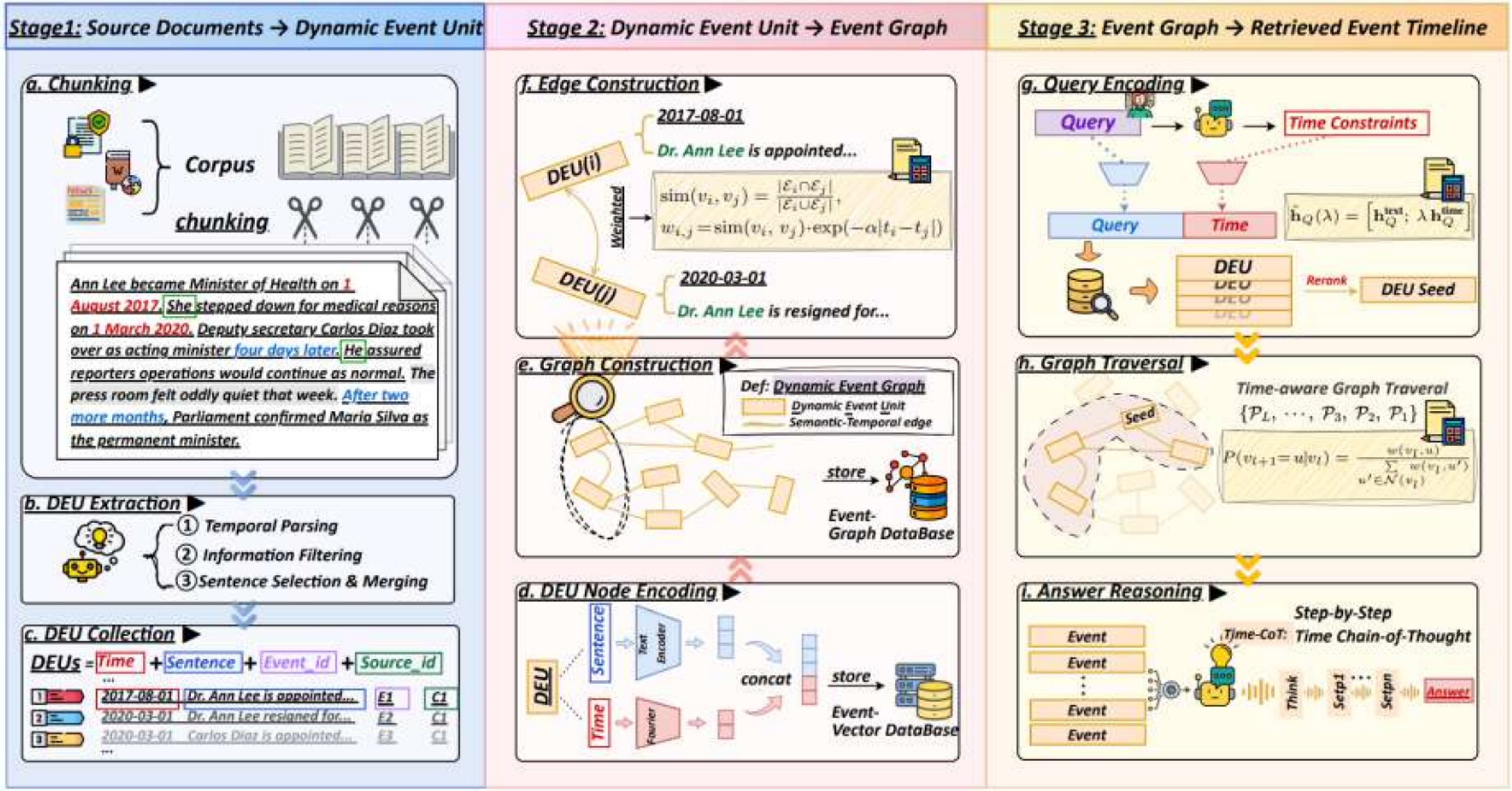
□ Dynamic GraphRAGs

Method	Graph Unit	Edge Type	Retrieval Strategy	Reasoning Mechanism	Dynamic
GraphRAG	Entity + Community	KG Relations + Community Links	Local + Global	Community summary	✗
LightRAG	Chunk Entities	Intra-chunk KG Relations	Dual-level keywords	Shallow path merge	✗
E^2GraphRAG	Summary Tree + Entity	Semantic + Hierarchical Links	Adaptive Local/Global	Chunk ranking	✗
HippoRAG	Concept Nodes	Concept Associations	PPR-guided multi-hop	PPR subgraph rank	✗
HybridRAG	KG + Chunks	KG relations	Hybrid merge	Evidence voting	✗
DyG-RAG	**Dynamic Event Units**	**Temporal–Semantic Links**	**Time-aware graph walk**	**Time-CoT**	✓

[Picture Credit] DyG-RAG: Dynamic graph retrieval-augmented generation with event-centric reasoning. *Qingyun Sun*, et al.

■Graph Retrieval-Augmented Generation (GraphRAG)

□Dynamic GraphRAGs



$$\mathrm{sim}(v_i, v_j) = \frac{|\mathcal{E}_i \cap \mathcal{E}_j|}{|\mathcal{E}_i \cup \mathcal{E}_j|},\quad w_{i,j} = \mathrm{sim}(v_i, v_j)\cdot \exp(-\alpha|t_i - t_j|)$$

$$\tilde{\mathbf{h}}_Q(\lambda) = [\mathbf{h}_Q^{\text{text}}; \lambda \mathbf{h}_Q^{\text{time}}]$$

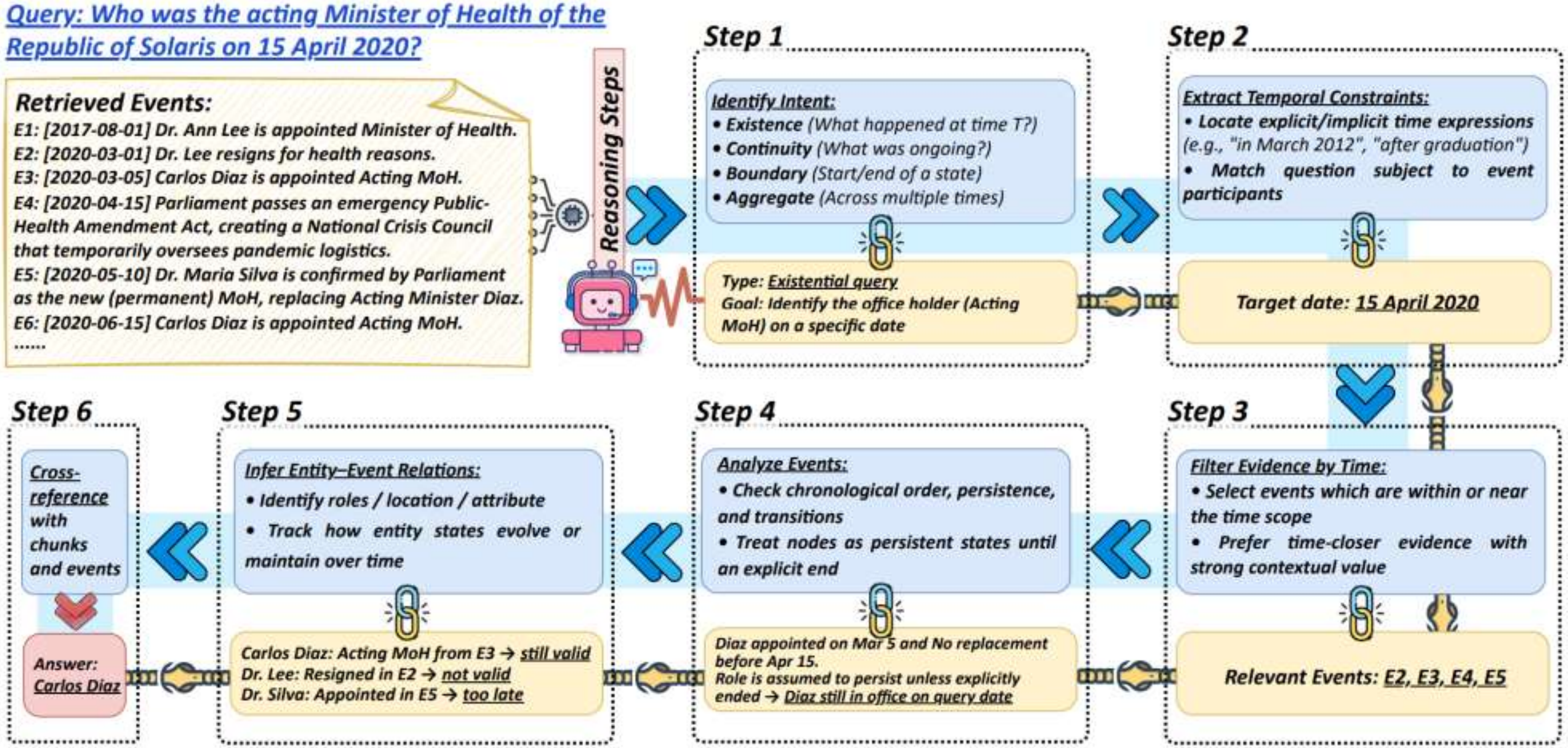
$$\{\mathcal{P}_L, \cdots, \mathcal{P}_3, \mathcal{P}_2, \mathcal{P}_1\}$$

$$P(v_{l+1} = u | v_l) = \frac{w(v_l, u)}{\sum_{u' \in \mathcal{N}(v_l)} w(v_l, u')}$$

[Picture Credit] DyG-RAG: Dynamic graph retrieval-augmented generation with event-centric reasoning. *Qingyun Sun*, et al.

Graph Retrieval-Augmented Generation (GraphRAG)

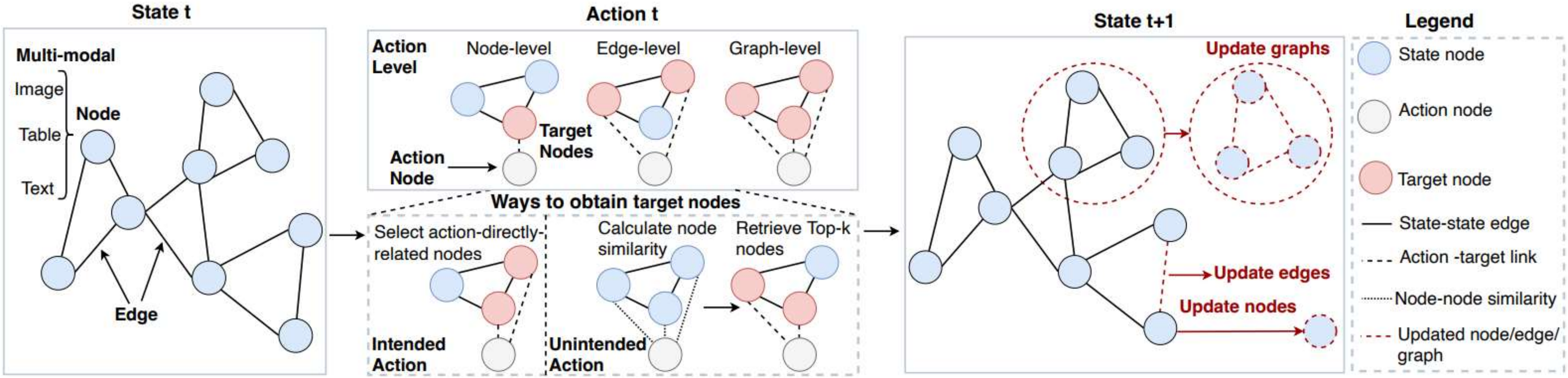
Dynamic GraphRAGs



[Picture Credit] DyG-RAG: Dynamic graph retrieval-augmented generation with event-centric reasoning. *Qingyun Sun*, et al.

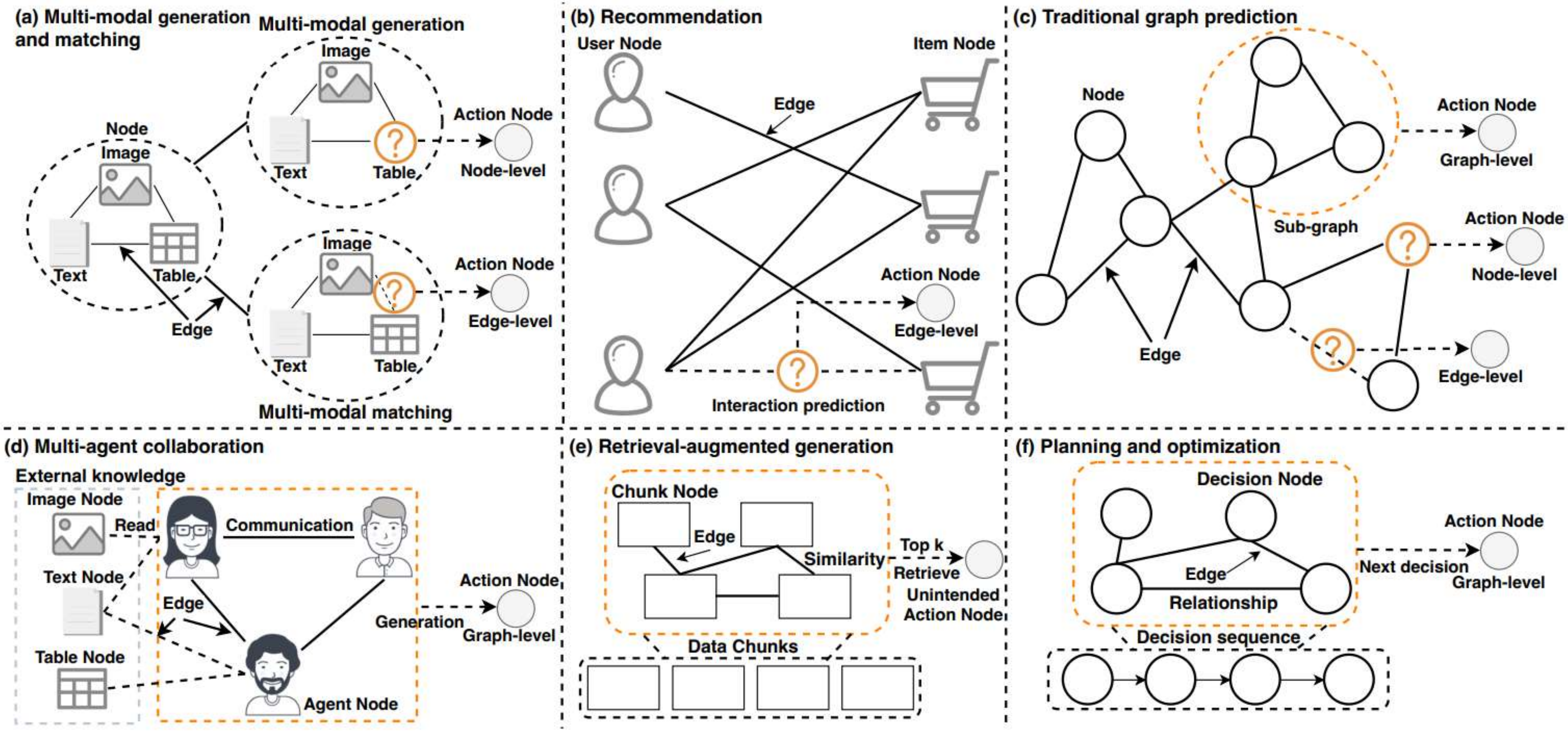
■ Graph World Model (GWM) ➔ Model the World as Graph

- ◻ unifies world-modeling with graph structure: represents the world state as a graph with multi-modal data and action nodes
- ◻ handle diverse tasks (generation, recommendation, multi-agent simulation, planning) by reasoning over a structured graph



[Picture Credit] Graph world model. *Tao Feng*, et al.

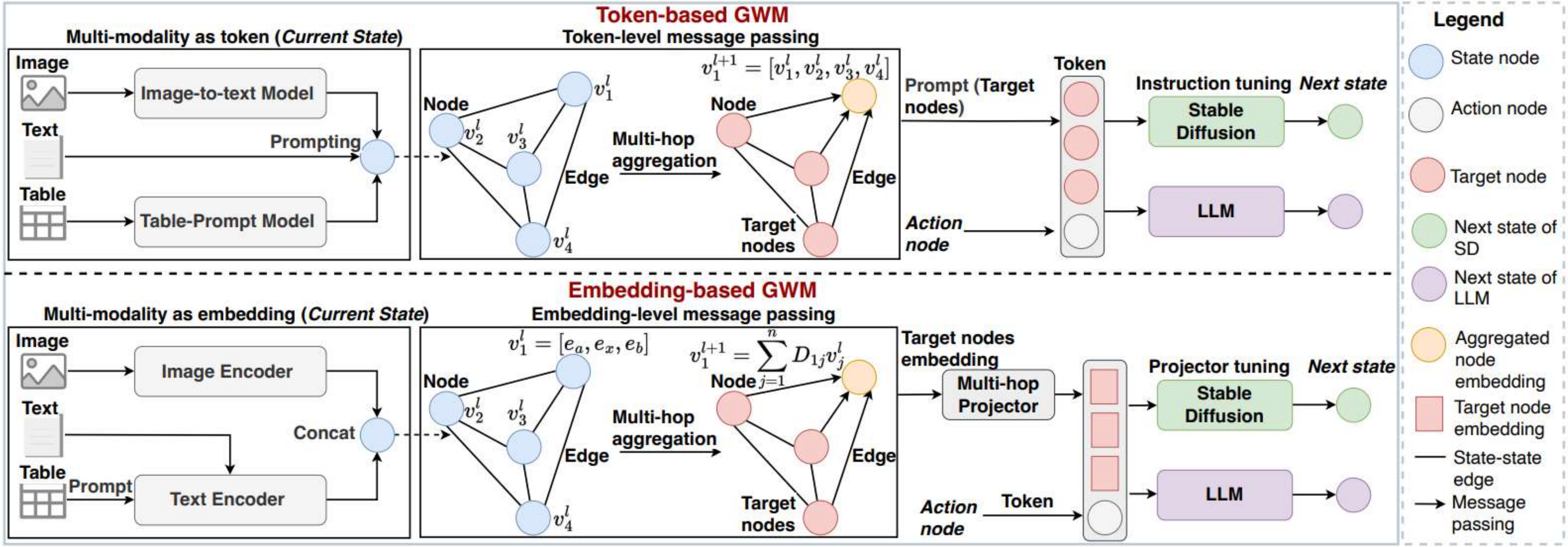
Graph World Model (GWM)



[Picture Credit] Graph world model. *Tao Feng*, et al.

Graph World Model (GWM)

Possible frameworks



Outlines

I. Recap of Tutorial

- Low-Distortion GRL – Motivation and Key Concept
- Key Approaches to Reduce Distortion

II. Future Directions

- Benchmarks for Low-Distortion GRL
- Graph Foundation Model (GFM)
- Graph Retrieval-Augmented Generation (GraphRAG)
- Graph World Model (GWM)

III. Open Challenges and Outlook

IV. Discussion

Data and Evaluation Challenges

- **Limited graph data scale:** Few large, high-quality graph datasets for pre-training; graph data often noisy or domain-specific
- **Heterogeneity:** Graphs vary greatly (social networks vs. molecules); a single model must handle diverse structures and feature types
- **No standard distortion metric:** Lacking unified measures of structural information loss – e.g. δ-hyperbolicity
- **Evaluation gaps:** Current benchmarks don't fully capture fidelity of structure preservation, complicating fair comparison of methods

■Model and Training Challenges

- □ **Architectural limits:** GNNs struggle with – scaling to billion-parameter models without losing local detail is unresolved
- □ **Training paradigms:** Unsupervised pre-training objectives for graphs are not as clear or universal as language modeling;
- □ **Efficiency:** Graph training doesn't scale easily – computing over large graphs or batches of graphs pushes memory and runtime limits
- □ **GNN-LLM integration:** Combining graph and language model components introduces complexity

Deployment and Trust Challenges

- **Lack of "killer app":** Needs a high-impact application (analogous to ChatGPT for LLMs) to drive broad adoption and investment
- **Domain adaptation:** Foundation graph models may not seamlessly transfer across domains – risk of distortion or failure
- **Trustworthiness:** Ensuring fairness, explainability, and robustness in graph models is vital – graph embeddings can inherit biases
- **Privacy and ethics:** Graph data often involve sensitive relationships; using them in large models raises privacy concerns and potential misuse if distortions lead to incorrect inferences

Outlines

- I. Recap of Tutorial
  - Low-Distortion GRL – Motivation and Key Concept
  - Key Approaches to Reduce Distortion
- II. Future Directions
  - Benchmarks for Low-Distortion GRL
  - Graph Foundation Model (GFM)
  - Graph Retrieval-Augmented Generation (GraphRAG)
  - Graph World Model (GWM)
- III. Open Challenges and Outlook
- **IV. Discussion**

Thank You!

Yuan Haonan @ *MAGIC Group*
yuanhn@buaa.edu.cn
August 29th, 2025

Towards Low-Distortion Graph Representation Learning

MAGIC Group

August 29th, 2025