G-reasoner: Foundation Models for Unified Reasoning on Graph-structured Knowledge with LLMs

Linhao Luo[1], Zicheng Zhao[2], Junnan Liu[1], Zhangchi Qiu[4], Junnan Dong[5], Serge Panev[6], Chen Gong[3], Thuy-Trang Vu[1], Gholamreza Haffari[1], Dinh Phung[1], Alan Wee-Chung Liew[4], Shirui Pan[4]

[1]Monash University, [2]Nanjing University of Science and Technology, [3]Shanghai Jiao Tong University, [4]Griffith University, [5]Tencent Youtu Lab, [6]NVIDIA

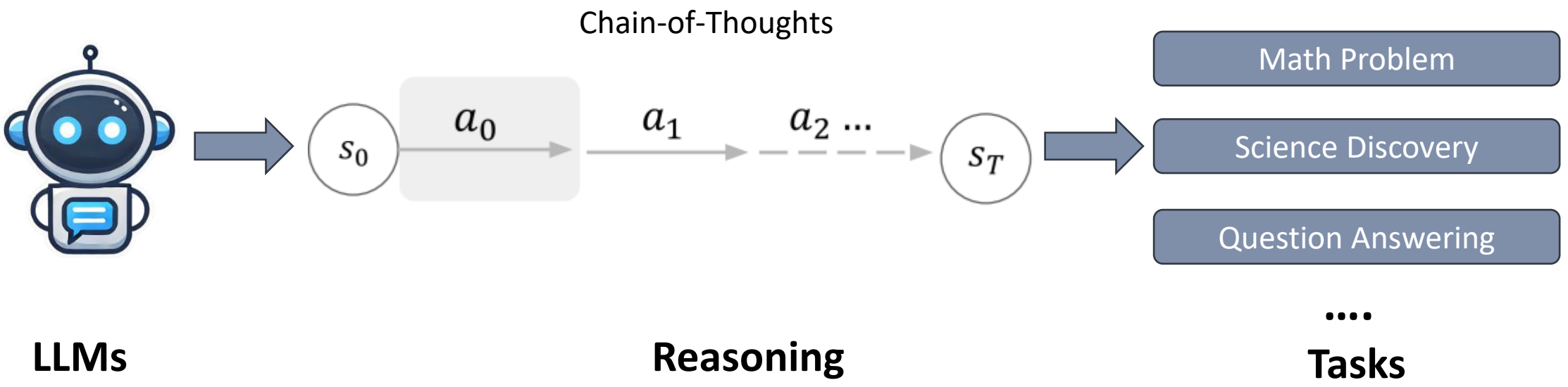
Presenter: Linhao Luo

Paper

Code

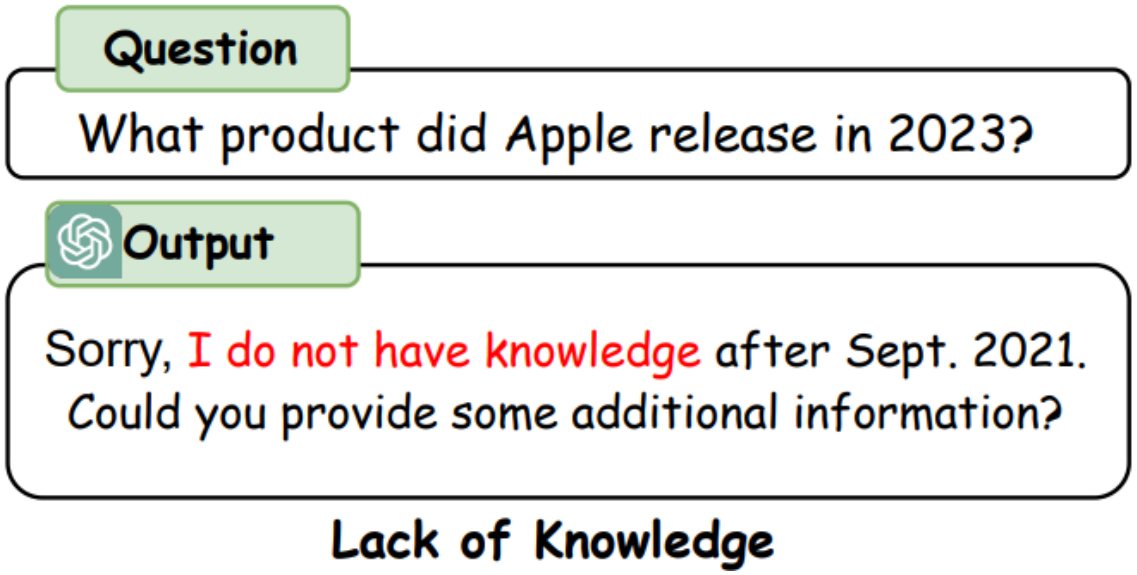
Large Language Models

- LLMs exhibit great reasoning ability to solve many complex tasks.



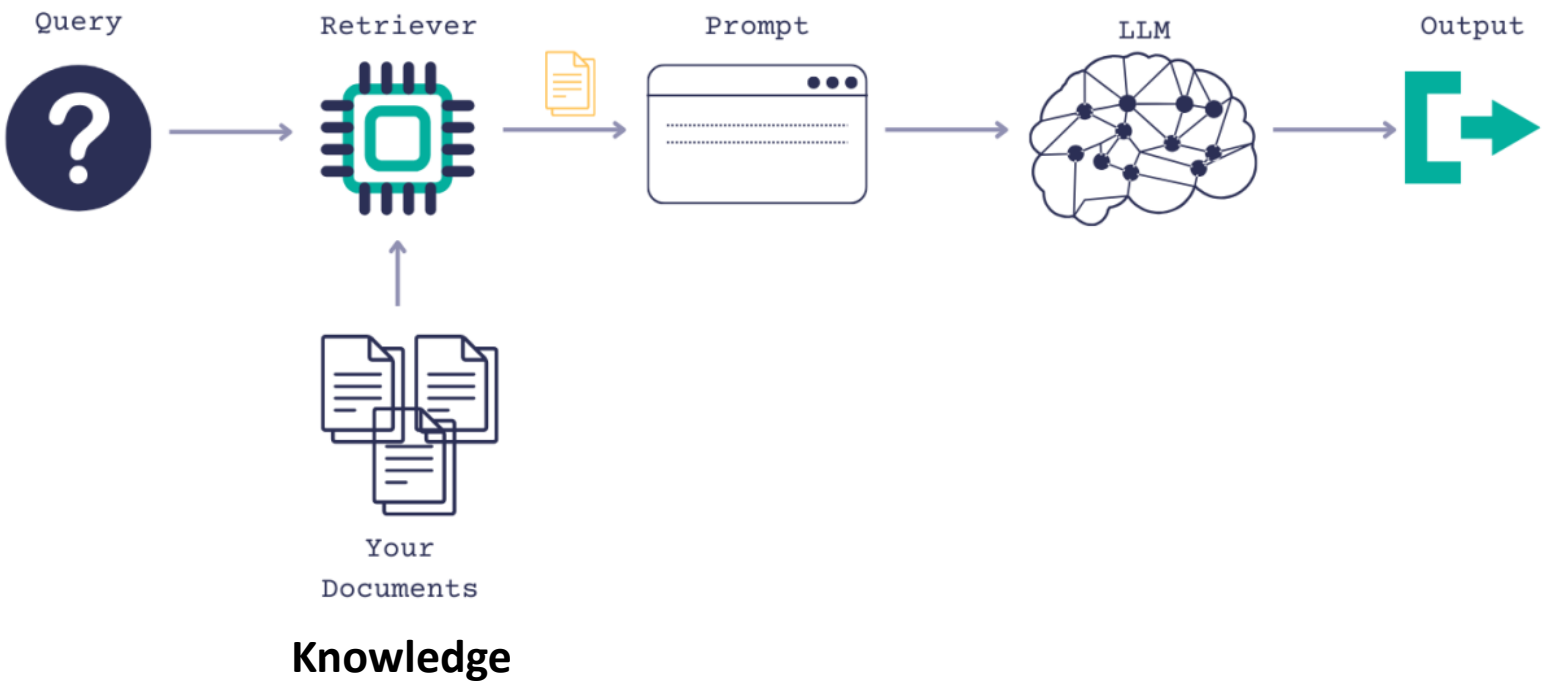
Large Language Models

- LLMs lack up-to-date or domain-specific knowledge, which limits their performance in knowledge-intensive tasks, such as medical diagnosis, legal judgment.



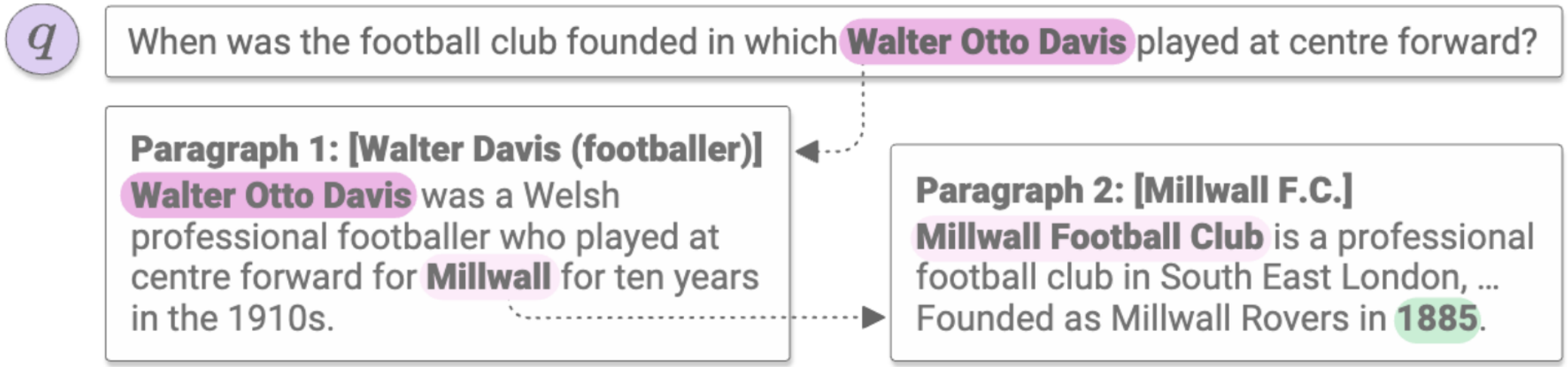
Retrieval-Augmented Generation (RAG)

- Retrieval-augmented generation (RAG) allows LLMs to reason in **external knowledge**, enhance its applicability.



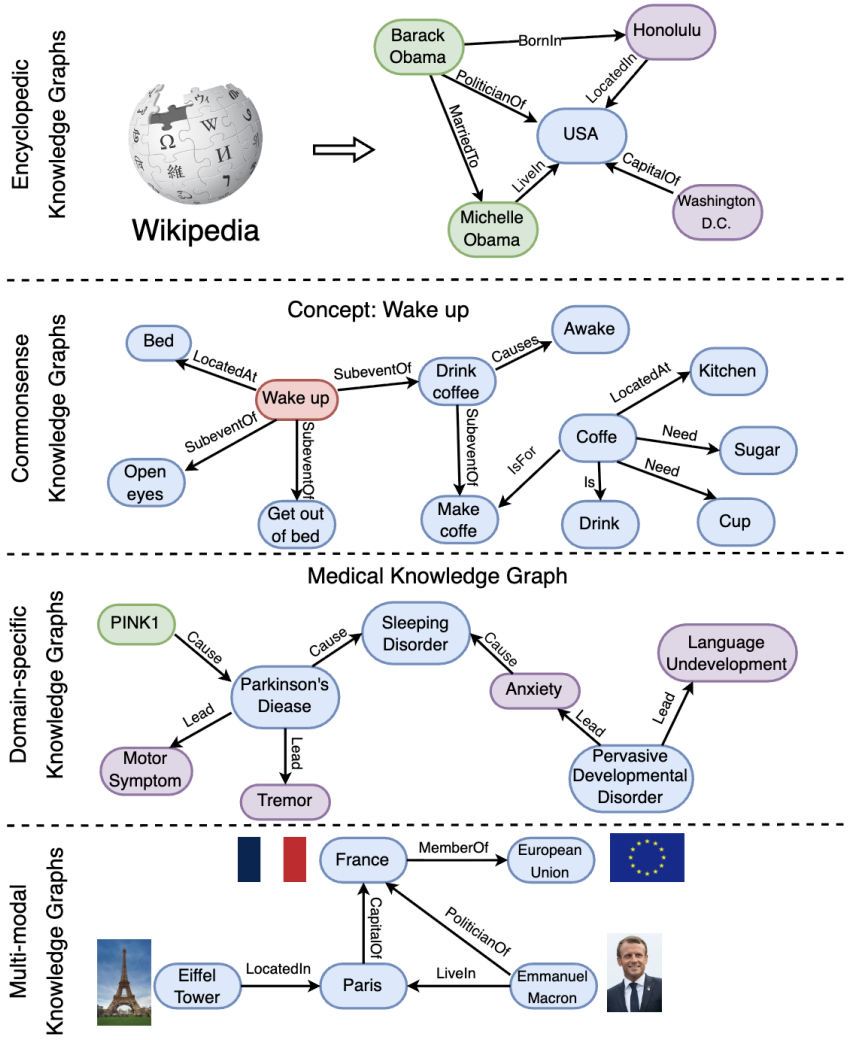
Reasoning on Structured Knowledge

- To better reason with knowledge, we need to capture the structural connections (associations) between knowledge.



Reasoning on Structured Knowledge

- Moreover, many real-world knowledge is structured like **knowledge graphs.**



Pan, Shirui, et al. "Unifying large language models and knowledge graphs: A roadmap." TKDE 2024

Reasoning on Structured Knowledge

- Unfortunately, LLMs, with their unstructured nature in architecture, cannot effectively reason on structured knowledge.

- **We need a better way to enhance LLMs to conduct reasoning on structured knowledge.**

Reasoning on Structured Knowledge

- Existing works focus on two-fold:
  - Structured knowledge index construction:
    - Tree
    - Graph
    - ….
  - Structure-enhanced retrieval:
    - Graph search
    - Agent + tool calling
    - GNN

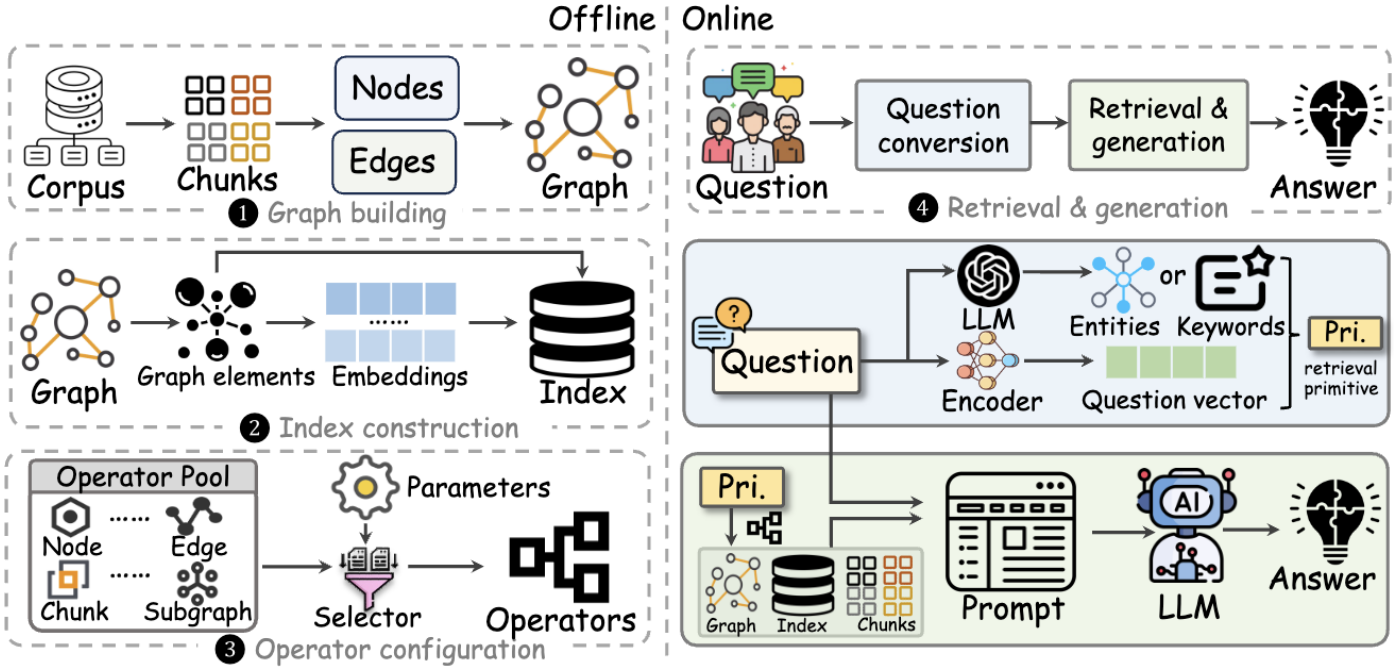


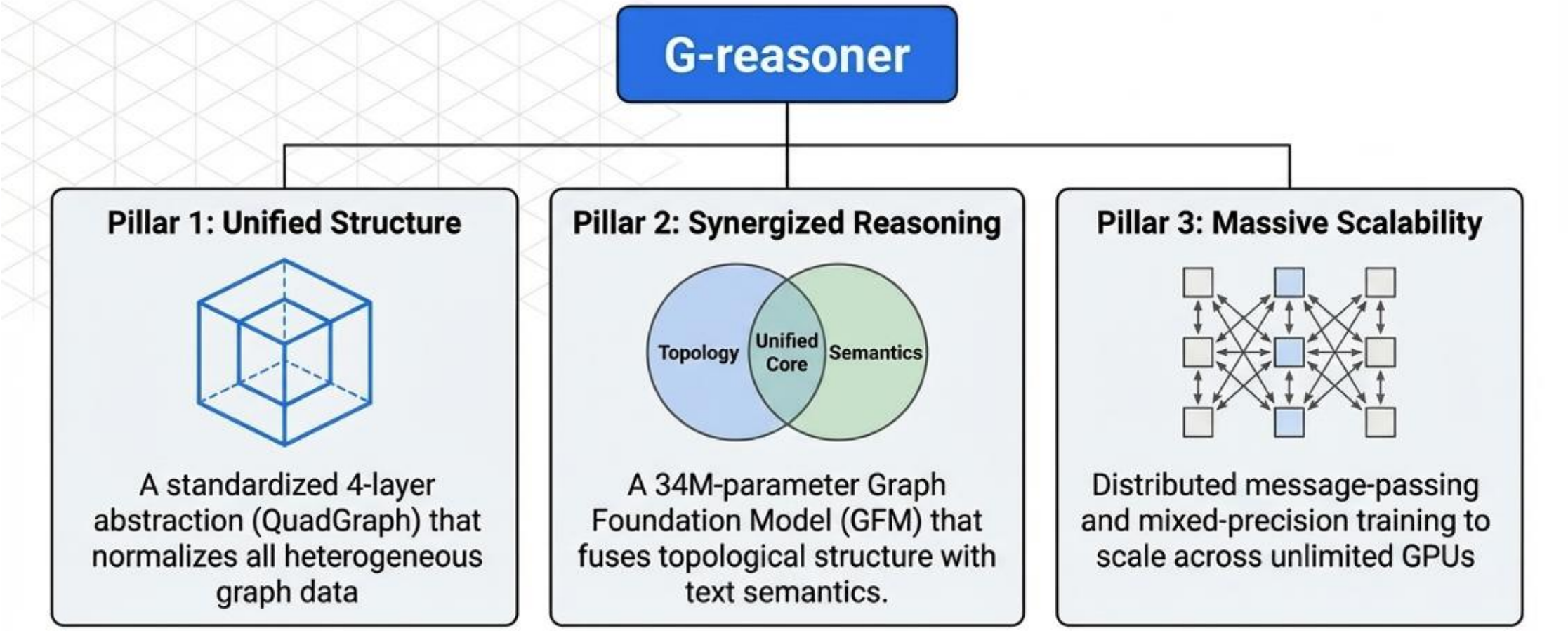
Figure 2: Workflow of graph-based RAG methods under our unified framework.

Zhou, Yingli, et al. "In-Depth Analysis of Graph-Based RAG in a Unified Framework." VLDB 2025

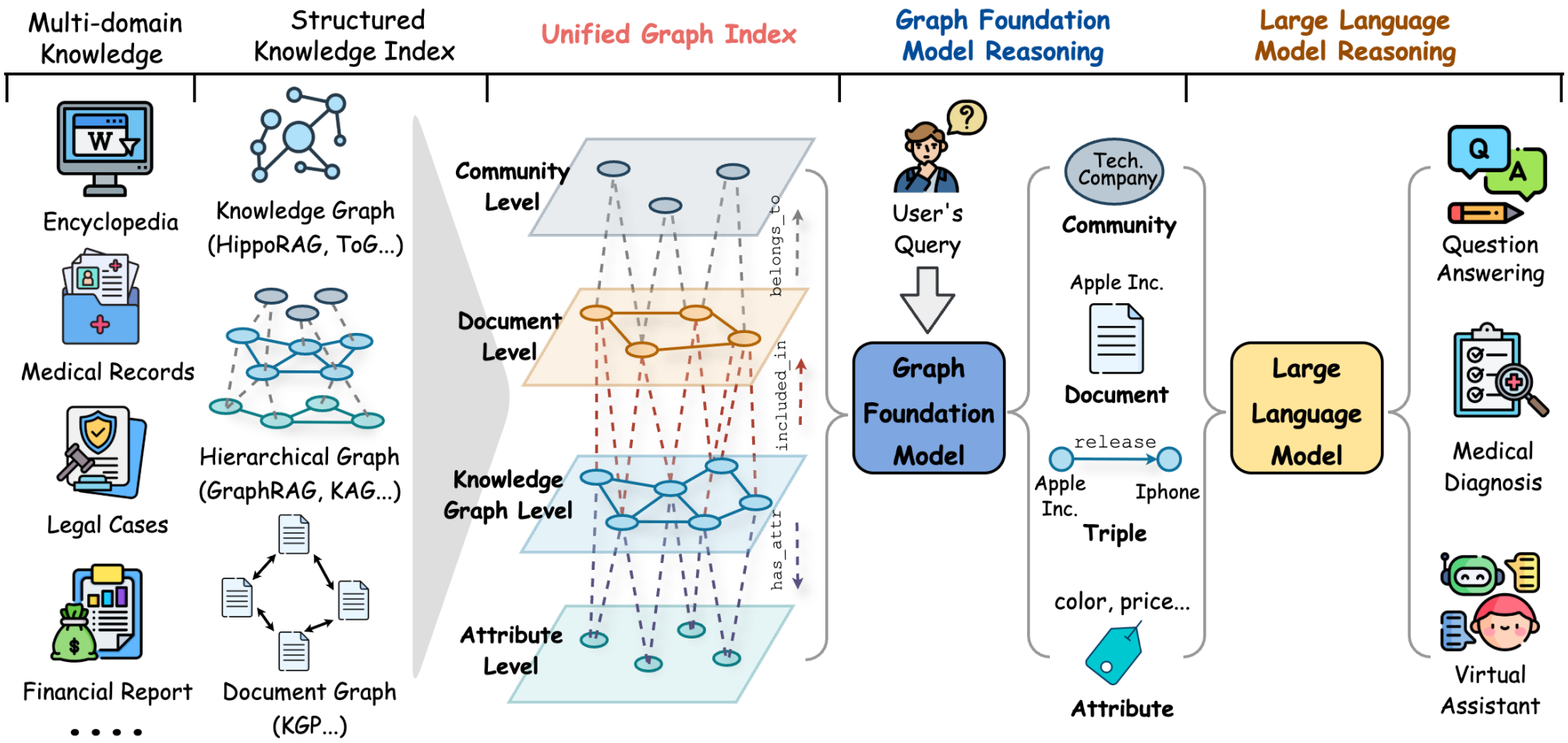
Reasoning on Structured Knowledge Trilemma Speed, Structure, or Scale?

Paradigm	Generalizability	Latency (Speed)	Computational Cost	Scalability	Multi-hop Accuracy
Graph Search (HippoRAG, LightRAG)	Low (ad-hoc structure)	Low (~2-20s)	Low	High	Moderate
Agentic Reasoning (ToG, KAG, Youtu-GraphRAG)	Low (ad-hoc structure)	Critical (> 70s)	Critical	Moderate	High
GNN-based Reasoning (G-retriever, GFM-RAG)	Low (KG specific)	Low (~2-10s)	Moderate	Low	Moderate

G-reasoner: Unified Reasoning Framework

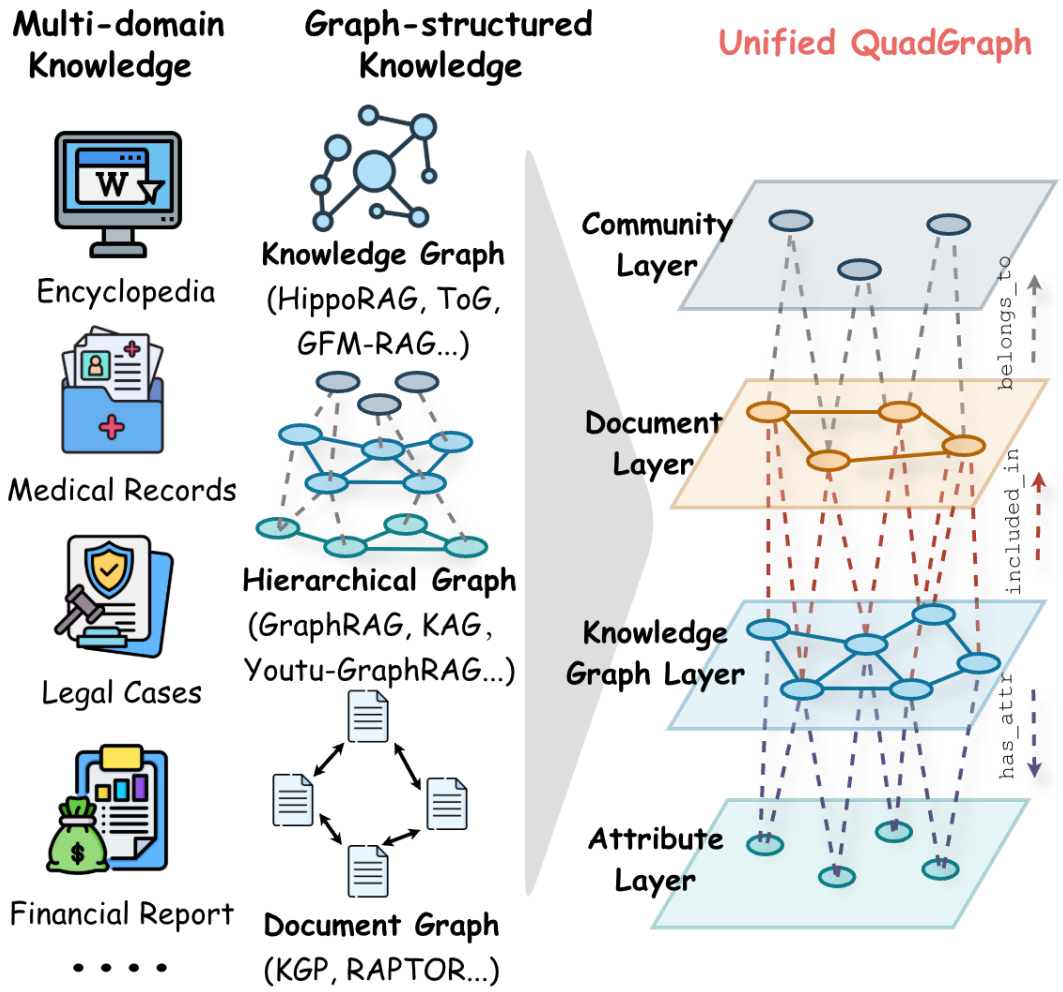


G-reasoner: Graph Foundation Models for Unified Reasoning on Structured Knowledge with LLMs



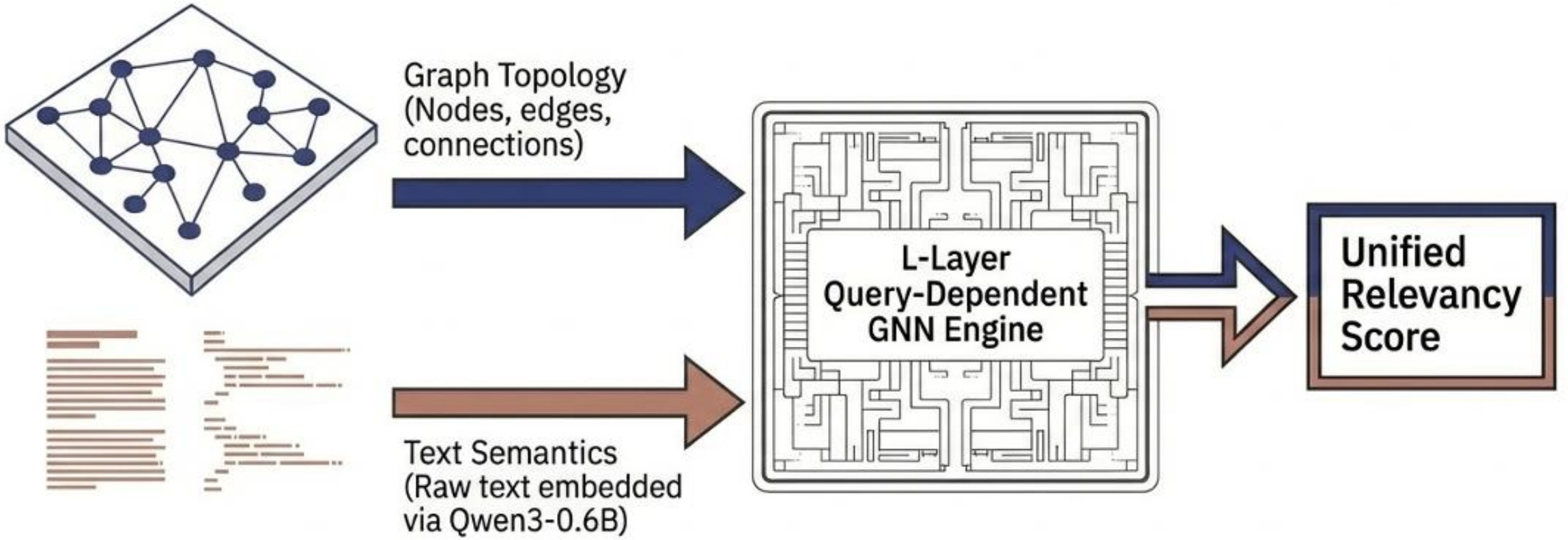
QuadGraph: An Universal Graph Interface

- To handle diverse graph structure for different domains, we created a standardized, four-layer abstraction.
- **QuadGraph Layers:**
  - **Attribute Layer:** Captures common properties of nodes (e.g., price, date).
  - **Knowledge Graph Layer:** Represents entities and their relationships as structured factual triples.
  - **Document Layer:** Contains unstructured text from documents and passages.
  - **Community Layer:** Groups related nodes to provide high-level, global structural information.
- This flexible structure can represent and unify the outputs of most existing GraphRAG construction methods.



Synergizing Topology and Semantics Reasoning

- G-reasoner adopts a 34M parameter GFM that jointly reason over the graph topology and text semantics to score node against user queries.



Synergizing Topology and Semantics Reasoning

G-reasoner adopts a 34M parameter GFM that jointly reason over the graph topology and text semantics to score node against user queries.

Synergizes Structure & Semantics:

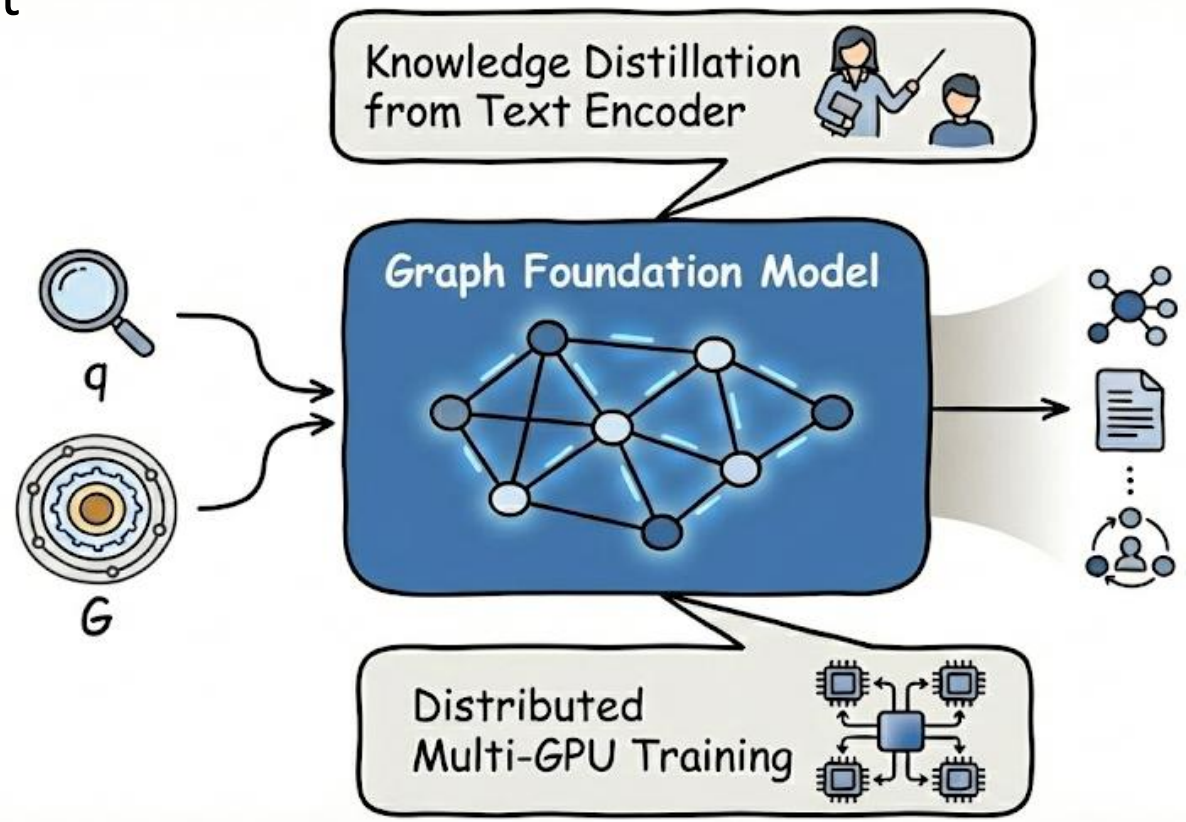
- Jointly reasons over the graph topology and rich text embedded in nodes and edges.

Weakly Supervised Training:

- Knowledge distillation from pre-trained text embedding model.

Scalable Implementations:

- Mixed precision training and a distributed message-passing to handle massive graphs across multiple GPUs.



Synergizing Topology and Semantics Reasoning

- **Semantic Encoding**

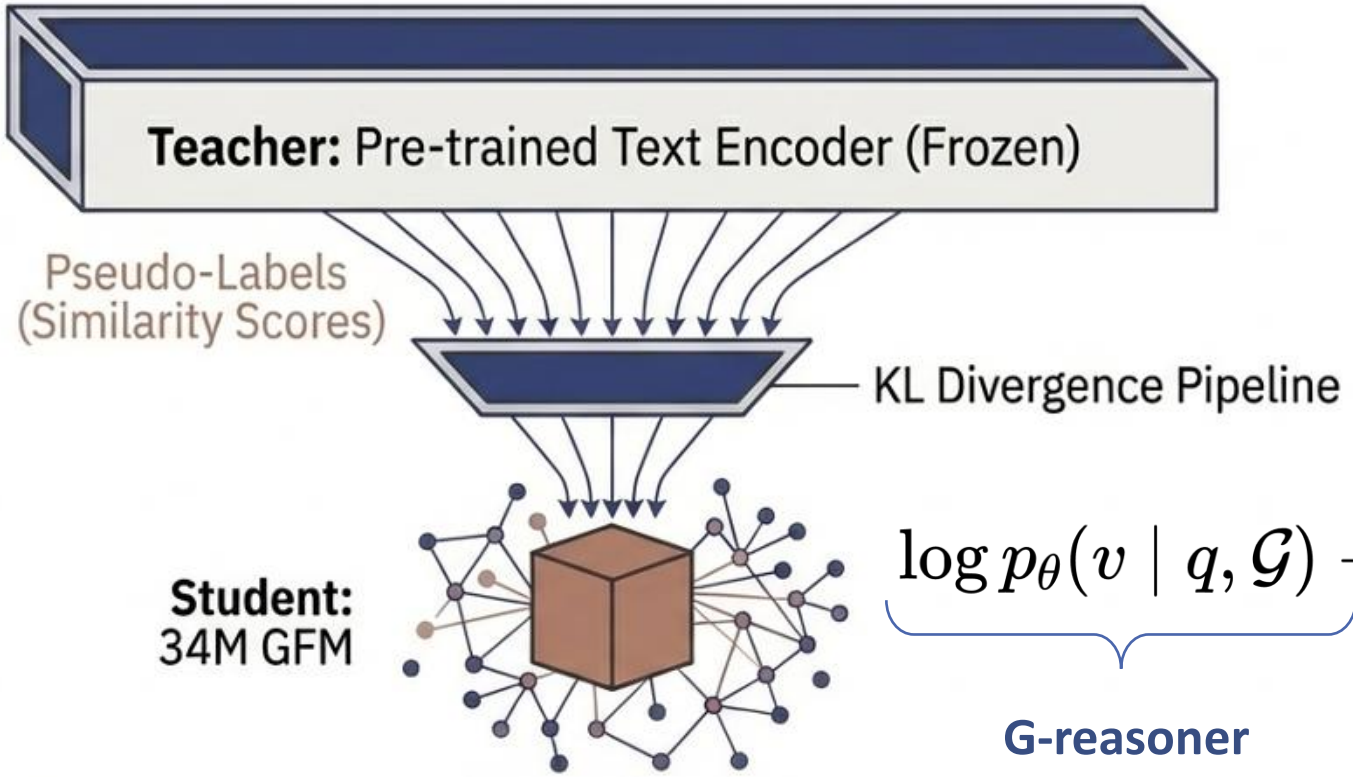
$$\mathbf{h}_q = \mathrm{SentenceEmb}(g), q \in \mathbb{R}^d,$$
$$\mathbf{H}_\mathcal{V} = \mathrm{SentenceEmb}(\mathcal{V}), \mathbf{h}_v \in \mathbf{H}_\mathcal{V}, \mathbf{h}_v \in \mathbb{R}^d.$$

- **GFM Reasoning**

$$\mathbf{h}_v^0 = \texttt{Init}(\mathbf{h}_v, \mathbf{1}_{v \in \mathcal{V}_q} \cdot \mathbf{h}_q),\ v \in \mathcal{V},$$
$$\mathbf{h}_v^l = \texttt{Update}\big(\mathbf{h}_v^{l-1}, \texttt{Agg}\big(\{\texttt{Msg}(\mathbf{h}_v^{l-1}, \mathbf{h}_r^l, \mathbf{h}_{v'}^{l-1}) \mid (v, r, v') \in \mathcal{E}\}\big)\big),\ l \in \{1, \ldots, L\},$$
$$p(v) = \texttt{Predictor}_{t_v}(\mathbf{h}_v^L, \mathbf{h}_v, \mathbf{h}_q).$$

Overcoming Data Scarcity via Weak Supervision

- To effectively train GFM without exhaustive human labels, we leverage the pre-trained text embedding model to provide weak supervision signals for training GFM.

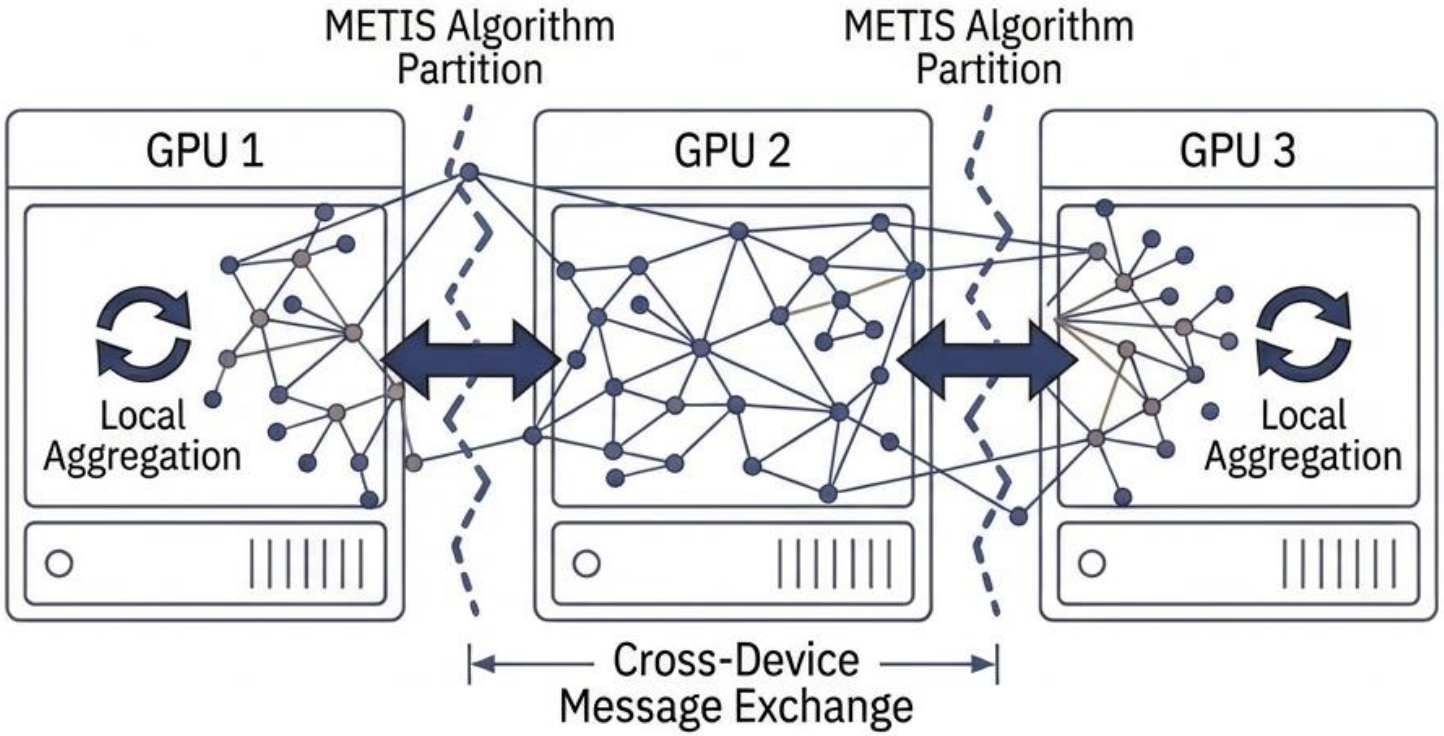


$$\log p_\theta(v \mid q, \mathcal{G}) - \lambda D_{\mathrm{KL}}(p_\phi(\mathcal{V} \mid q, \mathcal{S}) \,\|\, p_\theta(\mathcal{V} \mid q, \mathcal{G}))$$

G-reasoner **Distill from the text embedding models.**

Engineered for Unlimited Scale

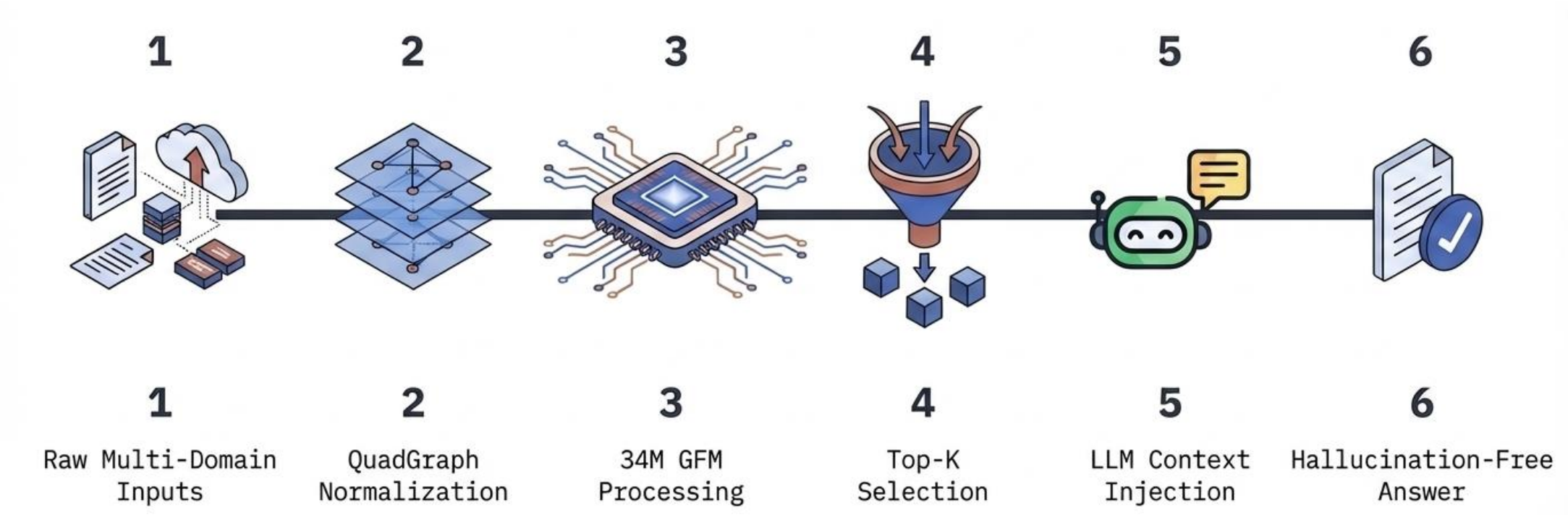
- G-reasoner partitions massive graphs across **distributed GPUs**. It computes locally and exchanges messages globally, allowing compute to scale linearly with data size via **mixed-precision training**.



PERFORMANCE METRICS:
Memory Reduced: -17.5%
Throughput Increased: +111%

Unified Reasoning Engine

- G-reasoner is an end-to-end framework for knowledge-enhanced reasoning.
- It processes chaos, restructures it, analyzes it with specialized GNNs, and provides the LLM only the precise structural and semantic context needed for a perfect answer.



QA Reasoning Performance

Dataset	# Query	# Document
HotpotQA (Yang et al., 2018)	1,000	9,221
MuSiQue (Trivedi et al., 2022)	1,000	6,119
2Wiki (Ho et al., 2020)	1,000	11,656
G-bench (Novel) (Xiang et al., 2025)	2,010	461
G-bench (Medical) (Xiang et al., 2025)	2,062	2,406
G-bench (CS) (Xiao et al., 2025)	1,018	24,534

Method	HotpotQA		MuSiQue		2Wiki		G-bench (Novel)	G-bench (Medical)	G-bench (CS)
	EM	F1	EM	F1	EM	F1	ACC	ACC	ACC
Non-structure Methods									
None (GPT-4o-mini) (OpenAI, 2024)	28.6	41.0	11.2	36.3	30.2	36.3	51.4	67.1	70.7
BM25 (Robertson & Walker, 1994)	52.0	63.4	20.3	28.8	47.9	51.2	56.5	68.7	71.7
ColBERTv2 (Santhanam et al., 2022)	43.4	57.7	15.5	26.4	33.4	43.3	56.2	71.8	71.9
Qwen3-Emb (8B) (Zhang et al., 2025b)	53.4	67.6	31.9	44.1	57.2	63.2	56.2	70.4	73.5
Graph-enhanced Methods									
RAPTOR (Sarthi et al., 2024)	50.6	64.7	27.7	39.2	39.7	48.4	43.2	57.1	73.6
GraphRAG (MS) (Edge et al., 2024)	51.4	67.6	27.0	42.0	34.7	61.0	50.9	45.2	72.5
LightRAG (Guo et al., 2024)	9.9	20.2	2.0	9.3	2.5	12.1	45.1	63.9	71.2
KAG (Liang et al., 2025)	59.5	72.2	33.8	46.0	67.3	75.1	-	-	-
HippoRAG (Jimenez Gutierrez et al., 2024)	46.3	60.0	24.0	35.9	59.4	67.3	44.8	59.1	72.6
HippoRAG 2 (Gutiérrez et al., 2025)	56.3	71.1	35.0	49.3	60.5	69.7	56.5	64.9	-
SubgraphRAG (Li et al., 2025a)	44.5	57.0	25.1	35.7	62.7	69.0	-	-	-
G-retriever (He et al., 2024)	41.4	53.4	23.6	34.3	33.5	39.6	-	-	69.8
GFM-RAG (Luo et al., 2025)	56.2	69.5	30.2	49.2	69.8	77.7	58.6	72.2	72.1
G-reasoner	**61.4**	**76.0**	**38.5**	**52.5**	**74.9**	**82.1**	**58.9**	**73.3**	**73.9**

Cross-graph Generalizability

- Because QuadGraph acts as a universal translator, the GFM can be deployed directly onto existing graph architectures and immediately outperform their native retrieval methods.

Retriever	Graph Structure	QuadGraph Layer				HotpotQA		MuSiQue		2Wiki	
		KG	Doc.	Attr.	Com.	EM	F1	EM	F1	EM	F1
Personalized PageRank	HippoRAG	✓	-	-	-	46.3	60.0	24.0	35.9	59.4	67.3
Embedding+ Graph Search	LightRAG	✓	✓	-	-	9.9	20.2	2.0	9.3	2.5	12.1
G-reasoner	HippoRAG	✓	-	-	-	**54.0**	**68.3**	28.9	41.0	**72.0**	**80.0**
	LightRAG	✓	✓	-	-	49.7	62.0	25.3	35.9	59.4	64.4
	Youtu-GraphRAG	✓	✓	✓	✓	52.3	65.9	**30.3**	**42.5**	69.7	77.7

Efficiency: Fast and Accurate

Method	G-bench (CS)	
	Time (s)	ACC
Agent-based Methods		
KGP (Wang et al., 2024)	89.4	71.9
ToG (Sun et al., 2024)	70.5	71.7
DALK (Li et al., 2024)	26.8	69.3
Graph Search Methods		
GraphRAG (MS) (Edge et al., 2024)	44.9	72.5
LightRAG (Guo et al., 2024)	14.0	71.2
HippoRAG (Jimenez Gutierrez et al., 2024)	2.4	72.6
GNN-based Methods		
G-retriever (He et al., 2024)	23.8	69.8
GFM-RAG (Luo et al., 2025)	2.0	72.1
G-reasoner	**0.2**	**73.9**

Efficiency: Compute Scaling

- G-reasoner can scale linearly with data size and model size to support large-scale training and inference.

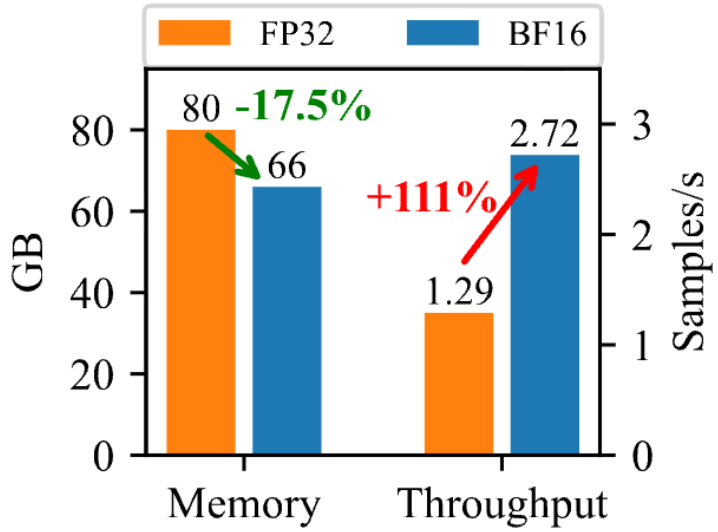


Figure 3: Memory and throughput gain brought by mixed precision training.

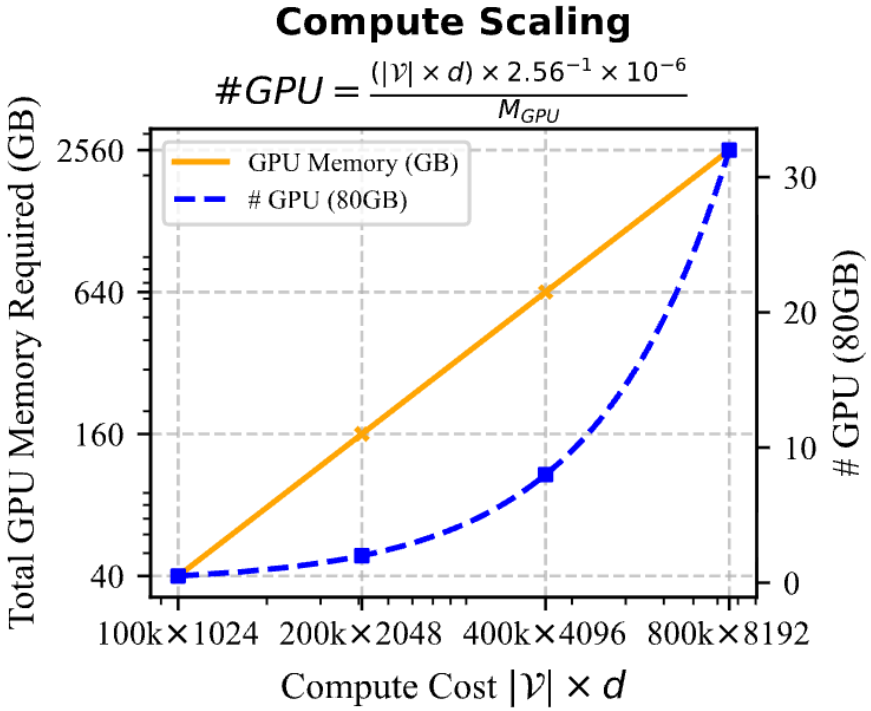


Figure 4: Compute scaling of G-reasoner.

Thanks for your listening!

Paper

Code