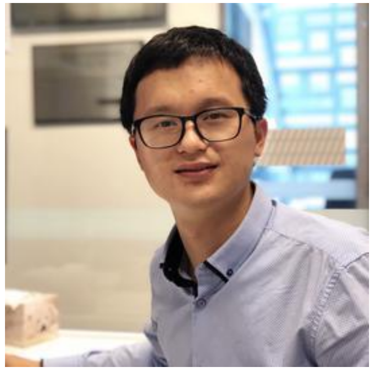
THE WEB CONFERENCE ACM 2025

Integrating Large Language Models and Knowledge Graphs for Next-level AGI

Lecture-style Tutorial

The Web Conference 2025



Linhao Luo — MONASH University

Carl Yang — EMORY UNIVERSITY

Evgeny Kharlamov — BOSCH

Shirui Pan — Griffith UNIVERSITY

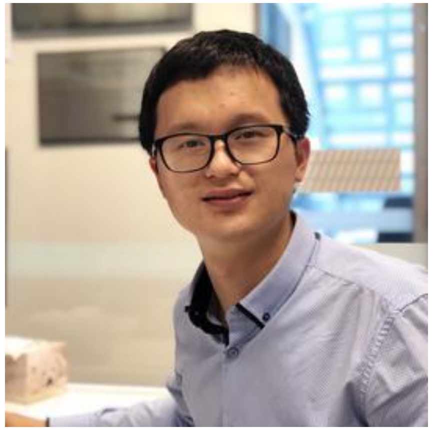
Room C2.5 April 29 1:30 pm - 5:00 pm, 2025 (GMT +10)

Presenters

Linhao Luo is a final-year Ph.D. candidate at Monash University. His research interests mainly focus on LLMs and KGs. He is the author of various leading research in the field LLMs and KGs, e.g., roadmap of unifying LLMs and KGs, and reasoning-on-graph (RoG).

Carl Yang is an Assistant Professor at Emory University. His research focuses on graph data mining, knowledge graphs, with applications in healthcare. He has published over 70 papers in top venues and leading author of various leading research in the field LLMs and KGs, e.g., knowledge graphs for healthcare, ClinGen, and GuardAgent.

Evgeny Kharlamov is a Senior Expert at the Bosch Centre for AI. His research interests encompass various AI topics, bridging knowledge representation and reasoning with learning, recently in LLM-centered and agentic scenarios. His research focuses on both theoretical aspects and practical applications across diverse industrial sectors.



Shirui Pan is a Professor and an ARC Future Fellow at Griffith University. His research interests include artificial intelligence and machine learning. He has made contributions to advance graph machine learning methods for solving hard AI problems for real-life applications, including anomaly detection, recommender systems, and time series forecasting. He has published 150+ papers at top conferences and journals.

Tutorial outline

Time	Content	Part	Presenter
30 min	**•Introduction and background** • Artificial general intelligence (AGI) • Large language models (LLMs) and knowledge graphs (KGs) • Challenges and opportunities	Part 1	Shirui Pan
60 min	**•Knowledge graph-enhanced large language models** • KG-enhanced LLM Training • KG-enhanced LLM Reasoning • Unified KG+LLM Reasoning	Part 2	Linhao Luo
30 min break			
50 min	**•Large language model-enhanced knowledge graphs** • LLM-enhanced KG integrations • LLM-enhanced KG construction and completion • LLM-enhanced Multi-modality KG	Part 3	Carl Yang
30 min	**•Applications of synergized KG-LLM systems** • QA system • Recommender system	Part 4	Evgeny Kharlamov
10 min	**•Future directions and conclusion**	Part 5	Linhao Luo

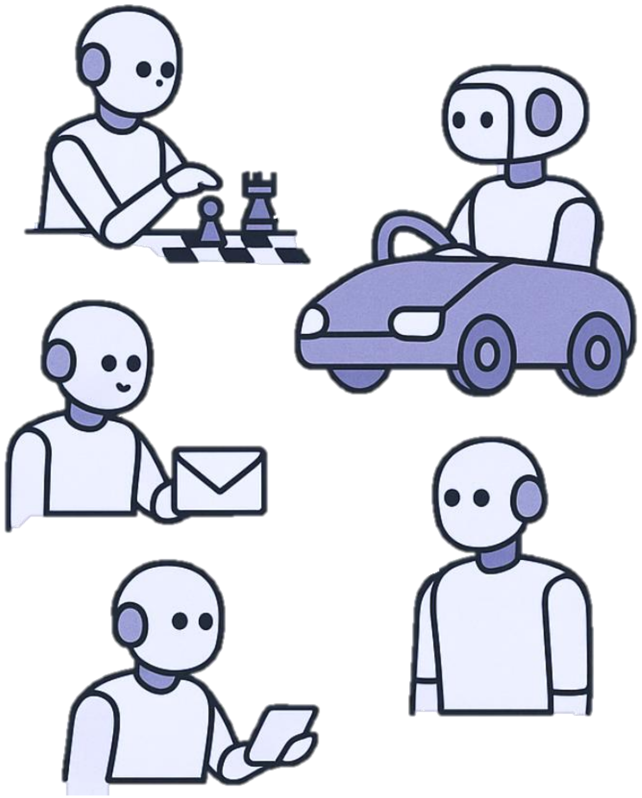
Part 1:Introduction and background

- Artificial general intelligence (AGI)
- Large language models (LLMs).
- Knowledge graphs (KGs).
- Limitations and opportunities toward AGI.

What is AGI?

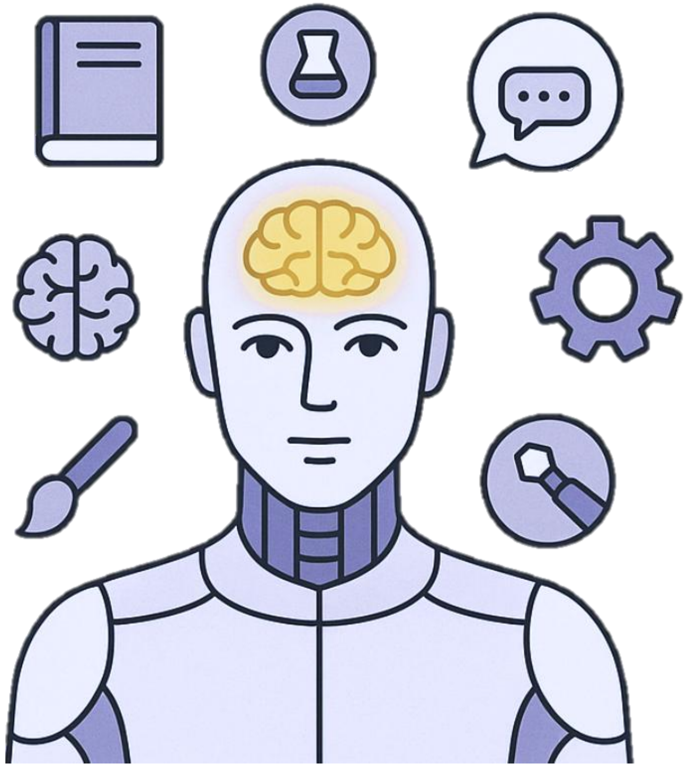
Artificial Intelligence (AI)

Specialized intelligence



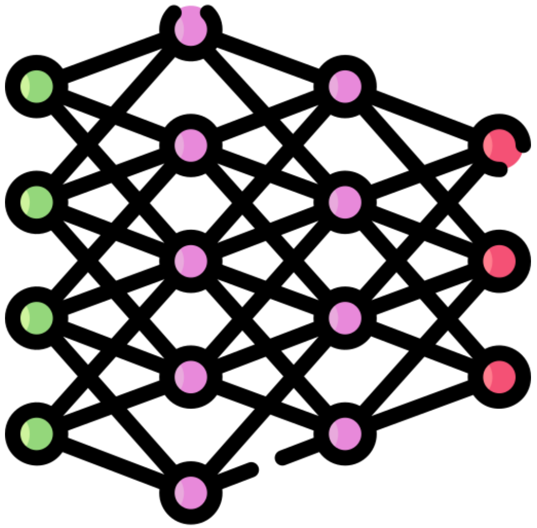
Artificial General Intelligence (AGI)

Human-level general intelligence



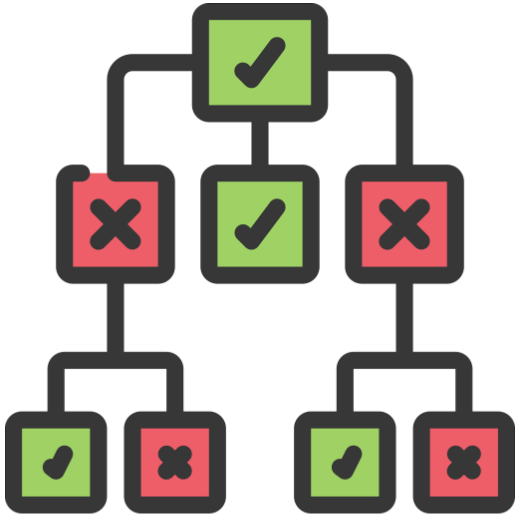
Roads to achieve AGI

Connectionism



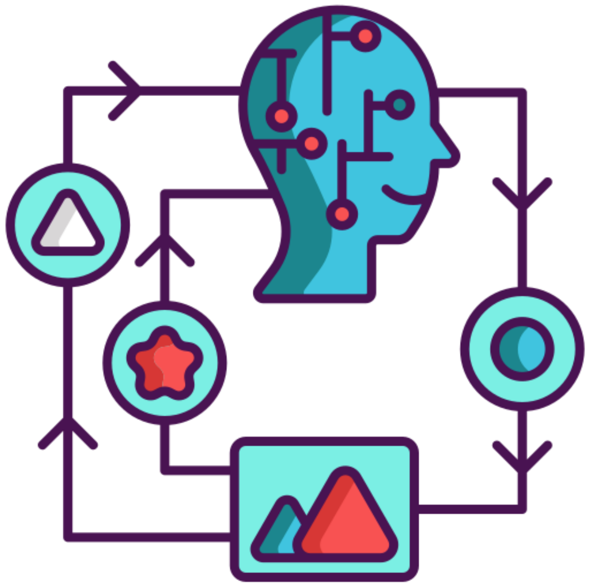
e.g., neural networks

Symbolicism



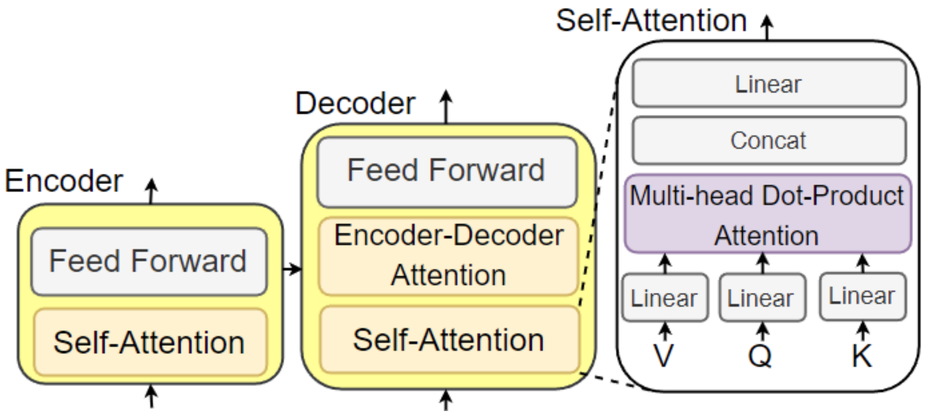
e.g., decision tree, KG, FOL

Actionism



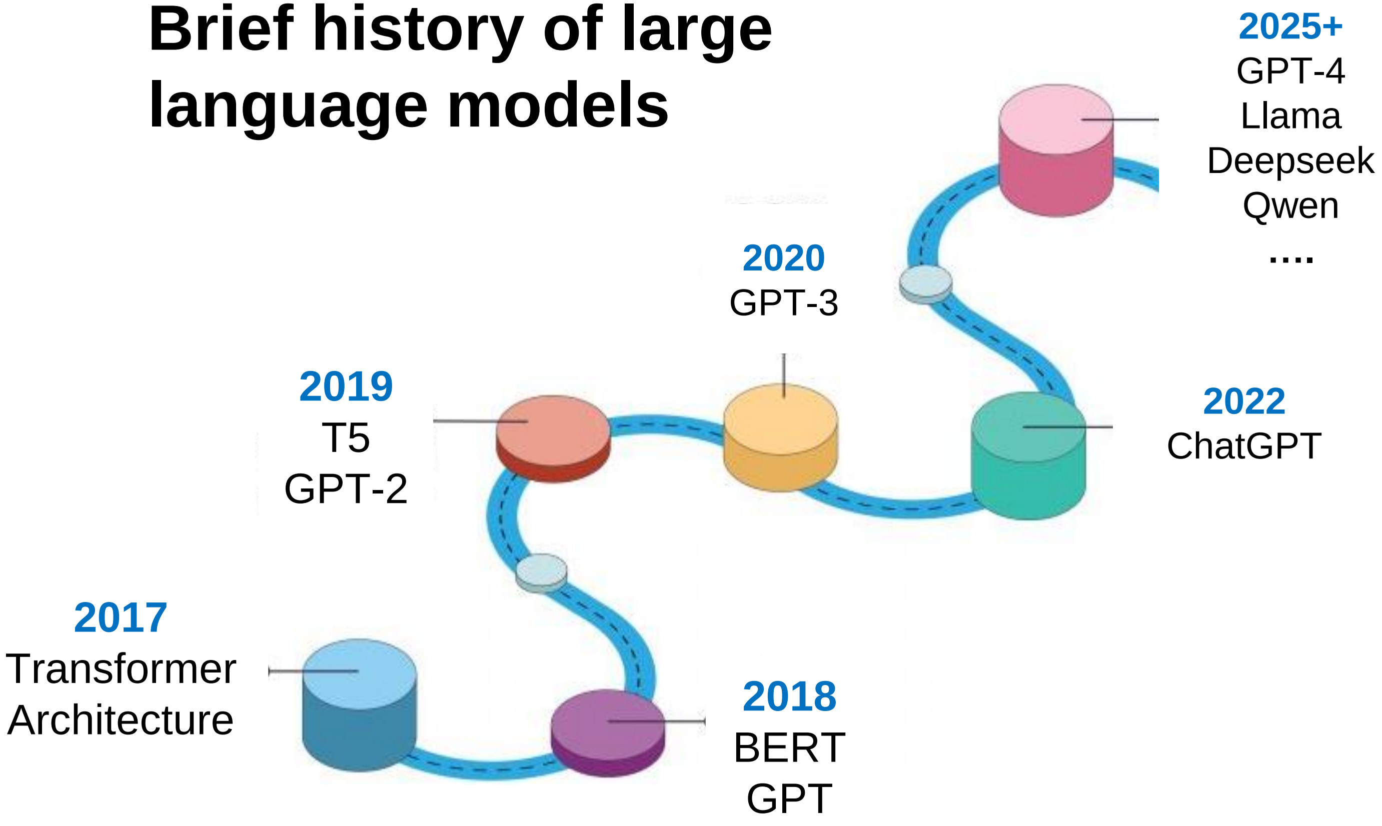
e.g., RL

LLMs as AGI



Transformer architecture

Brief history of large language models

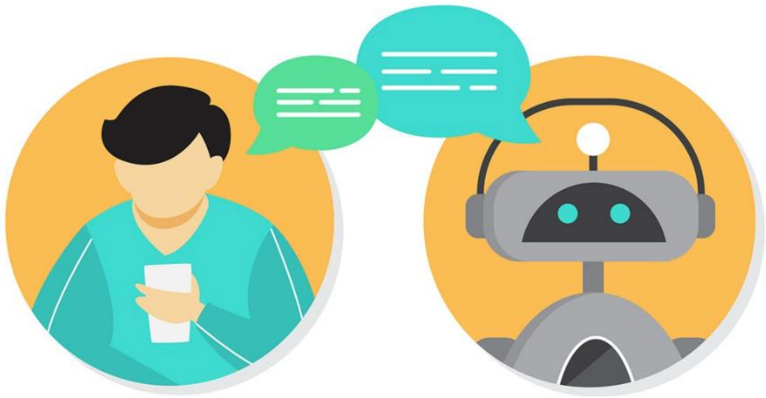


LLMs as AGI

- LLMs achieve surprising performance across many tasks.

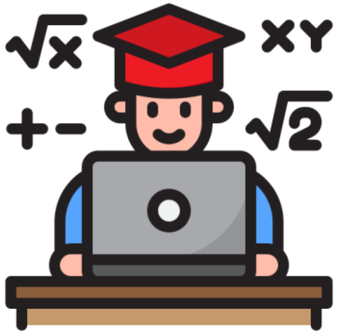
LLMs

Translation



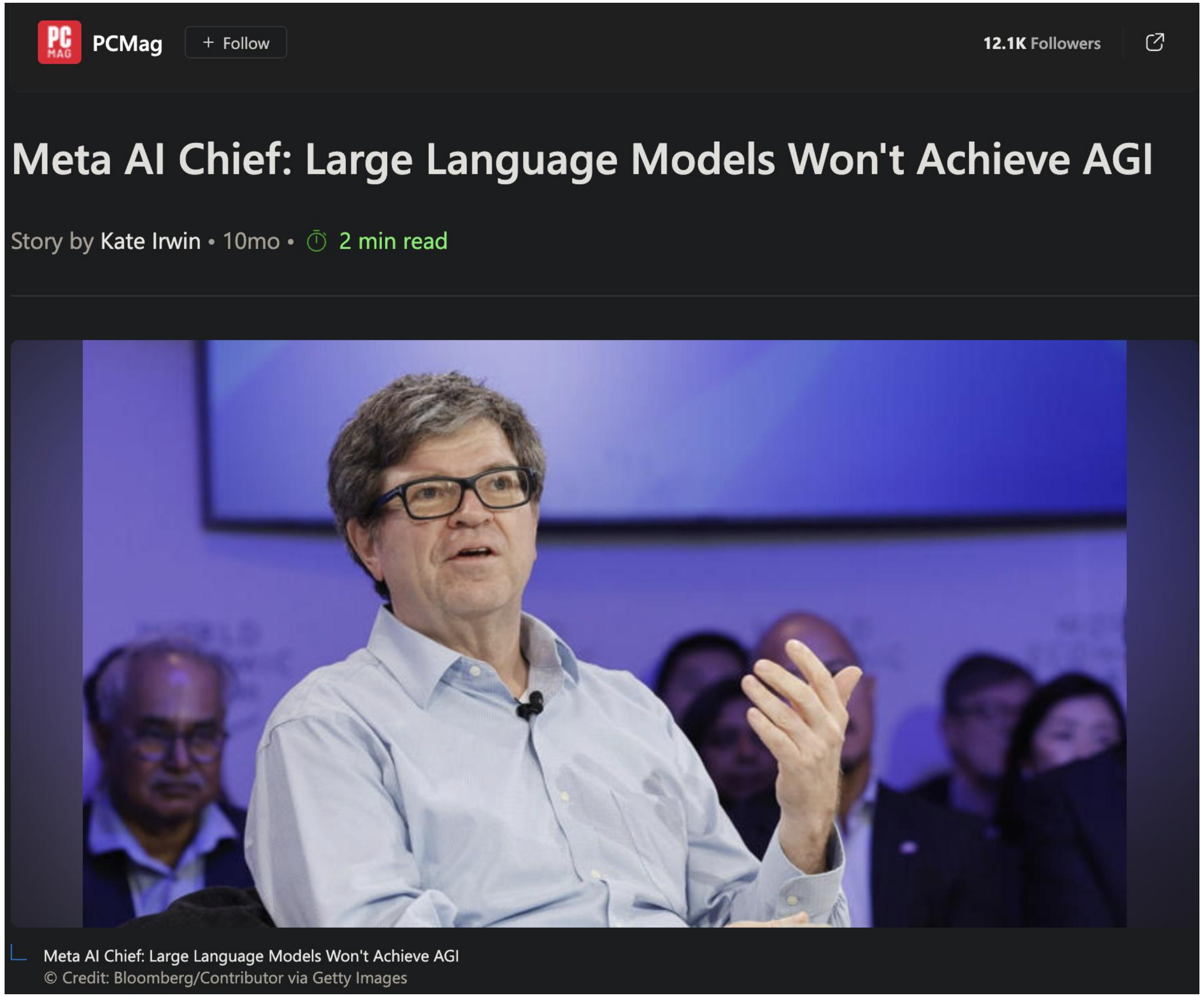
Conversation

Question Answering



Math Solver

Can LLMs achieve AGI?



PCMag + Follow 12.1K Followers

Meta AI Chief: Large Language Models Won't Achieve AGI

Story by Kate Irwin • 10mo • 2 min read

Meta AI Chief: Large Language Models Won't Achieve AGI
© Credit: Bloomberg/Contributor via Getty Images

Limitations of LLMs

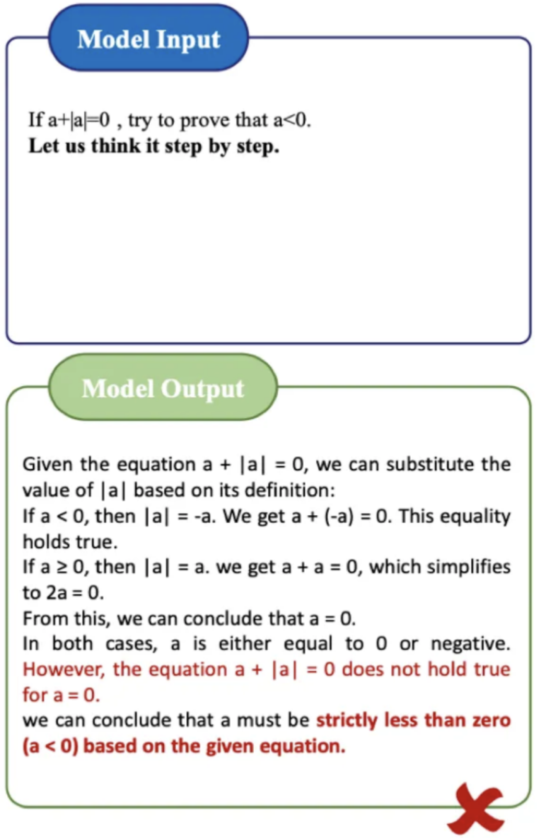
- LLMs suffer from **hallucination problems** during reasoning.

Q. When did Einstein discover gravity?

LLMs

Factual Error

A. Einstein discovered gravity in 1687



Reasoning Error

Hallucination impairs the trustworthiness of LLMs.

[1] Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., ... & Fung, P. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.
[2] https://bernardmarr.com/chatgpt-what-are-hallucinations-and-why-are-they-a-problem-for-ai-systems/

Limitations of LLMs

- LLMs limit in accessing **up-to-date knowledge**.

Apr 2 Mr. Trump added a 34 percent tariff on imports from China, to take effect on April 9, on top of two earlier rounds of 10 percent tariffs he had already imposed.

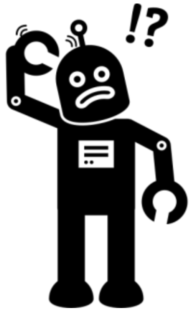
Apr 8 **Trump Threatens to Slap an Additional 50% Tariff on China**

By Alyssa Lukpat, Reporter

Donald J. Trump
@realDonaldTrump

Apr 10 Based on the lack of respect that China has shown to the World's Markets, I am hereby raising the Tariff charged to China by the United States of America to 125%, effective immediately. At some

Q. What is the current tariff on China?



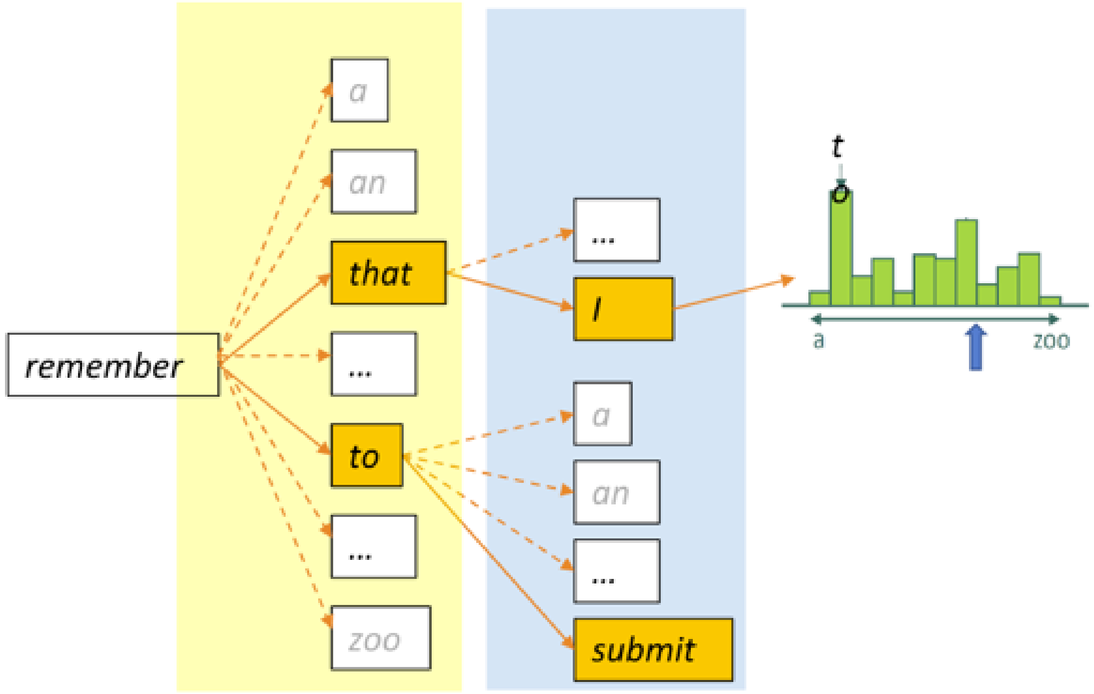
LLMs

Limitations of LLMs

- LLMs lack **interpretability.**
  - How to represent knowledge?
  - Why make such a decision?

Limitations of LLMs

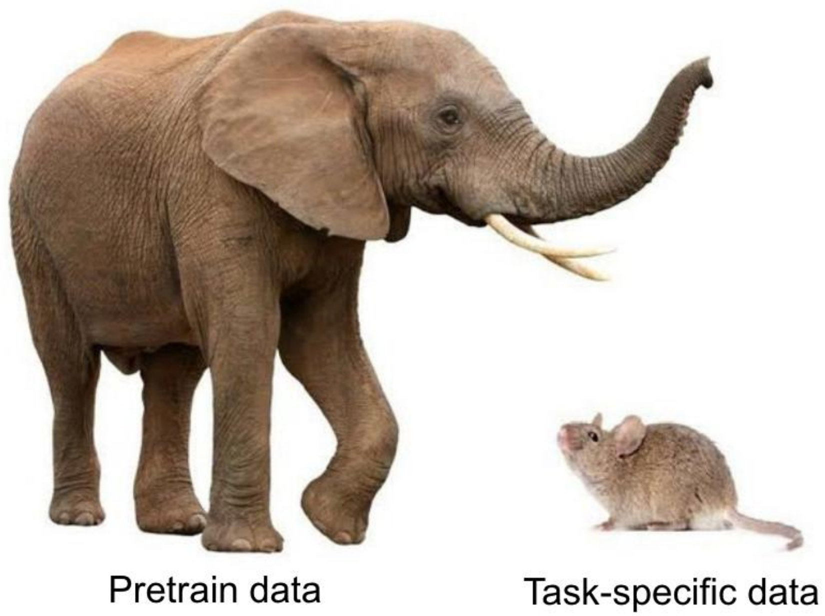
- LLMs are **indecisive.**
  - LLMs reason by probability.



Yixin Cao, et al. Trustworthy Natural Language Processing with Knowledge Guidance, WSDM-2023 Workshop

Limitations of LLMs

- LLMs are **heavy**
  - More data more parameters.
  - Cannot generalize to a specific domain.



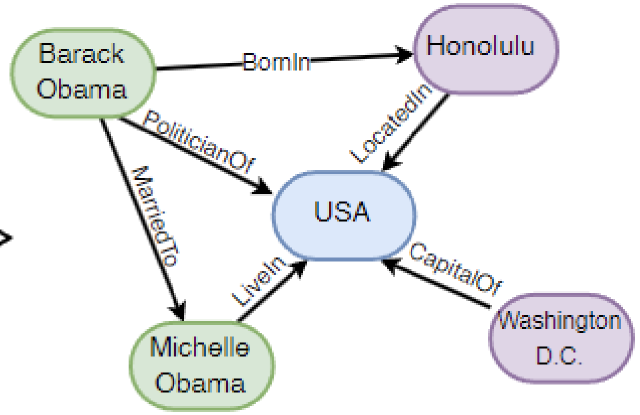
Yixin Cao, et al. Trustworthy Natural Language Processing with Knowledge Guidance, WSDM-2023 Workshop

LLMs are becoming supermen...

Who will be the watchmen?

Knowledge Graphs (KGs)

- Knowledge graphs (KGs), storing enormous facts in the way of triples, i.e., **(head entity, relation, tail entity)**



- KGs store facts in a structural manner.

KGs help AGI

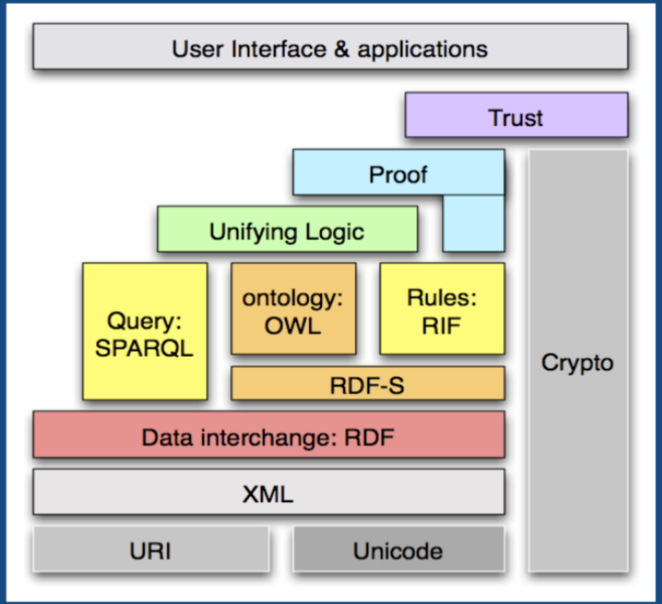
- KGs are **transparent**.

LLM is black-box

- How to represent knowledge?
- Why make such a decision?

KG is transparent

- Ontology and semantic definition
- Visible to users, e.g., nodes, edges
- Systematic store/exchange/update



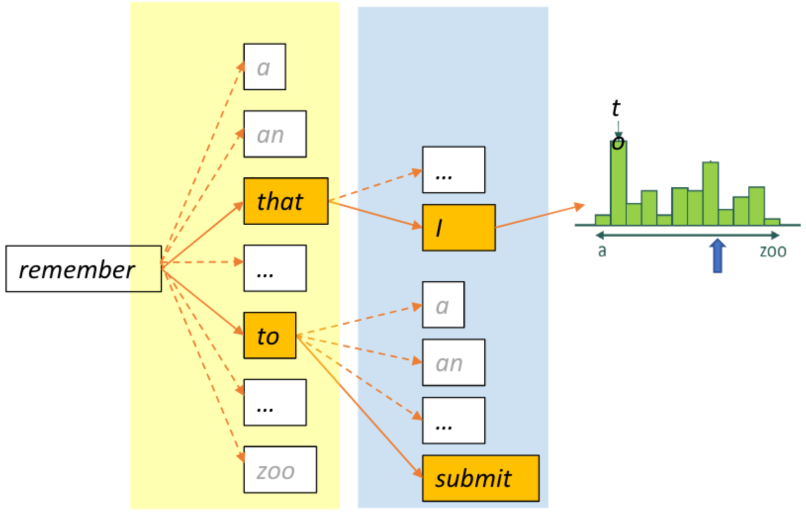
Yixin Cao, et al. Trustworthy Natural Language Processing with Knowledge Guidance, WSDM-2023 Workshop

KGs help AGI

- KGs are **adamant.**

LLM is indecisive

- Easily swayed
- Anything with a probability



KG is adamant

- Mostly black and white facts
- Photographic memory

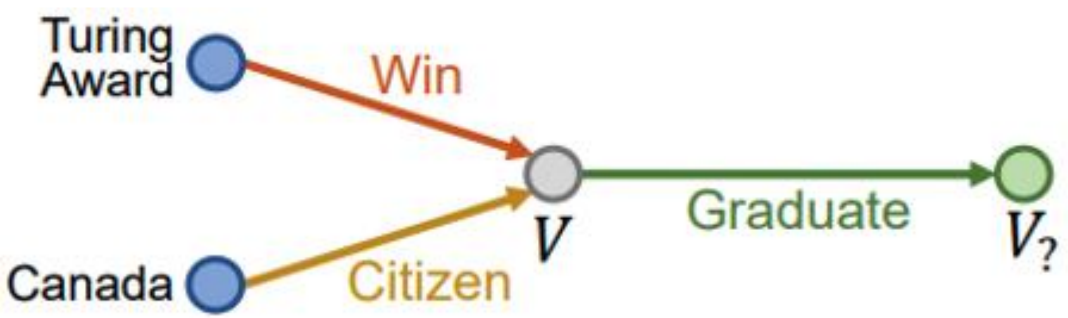
Capital	Singapore (city-state) 1°17′N 103°50′E
Official languages	English · Malay · Mandarin · Tamil
National language	Malay
Ethnic groups (2020)[a]	74.3% Chinese 13.5% Malay 9.0% Indian 3.2% Others
Religion (2020)[b]	31.1% Buddhism 20.0% No religion 18.9% Christianity 15.6% Islam 8.8% Taoism 5.0% Hinduism 0.6% Others
Demonym(s)	Singaporean
Government	Unitary dominant-party parliamentary republic
• President	Halimah Yacob
• Prime Minister	Lee Hsien Loong
Legislature	Parliament

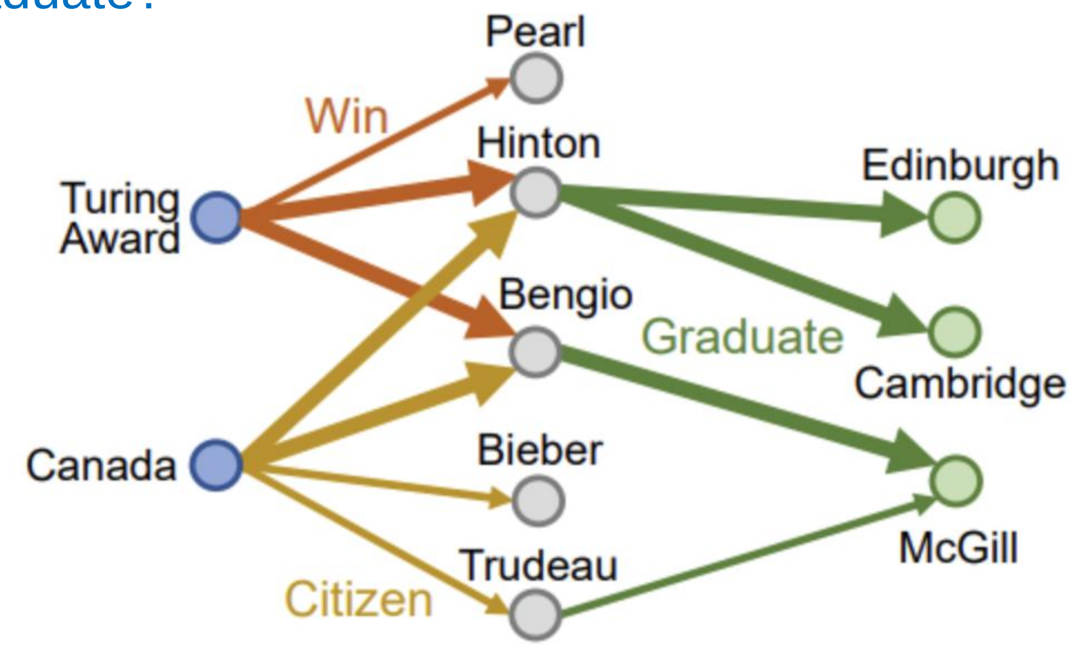
Yixin Cao, et al. Trustworthy Natural Language Processing with Knowledge Guidance, WSDM-2023 Workshop

KGs help AGI

- KGs power **symbolic reasoning**.

Q. Where did Canadian citizens with Turing Award Graduate?





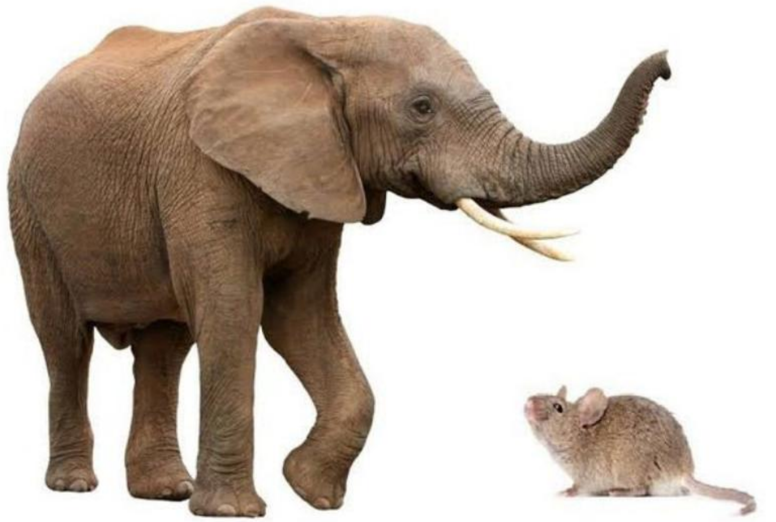
Ren, H., et al. Query2box: Reasoning over Knowledge Graphs in Vector Space Using Box Embeddings. ICLR 2020

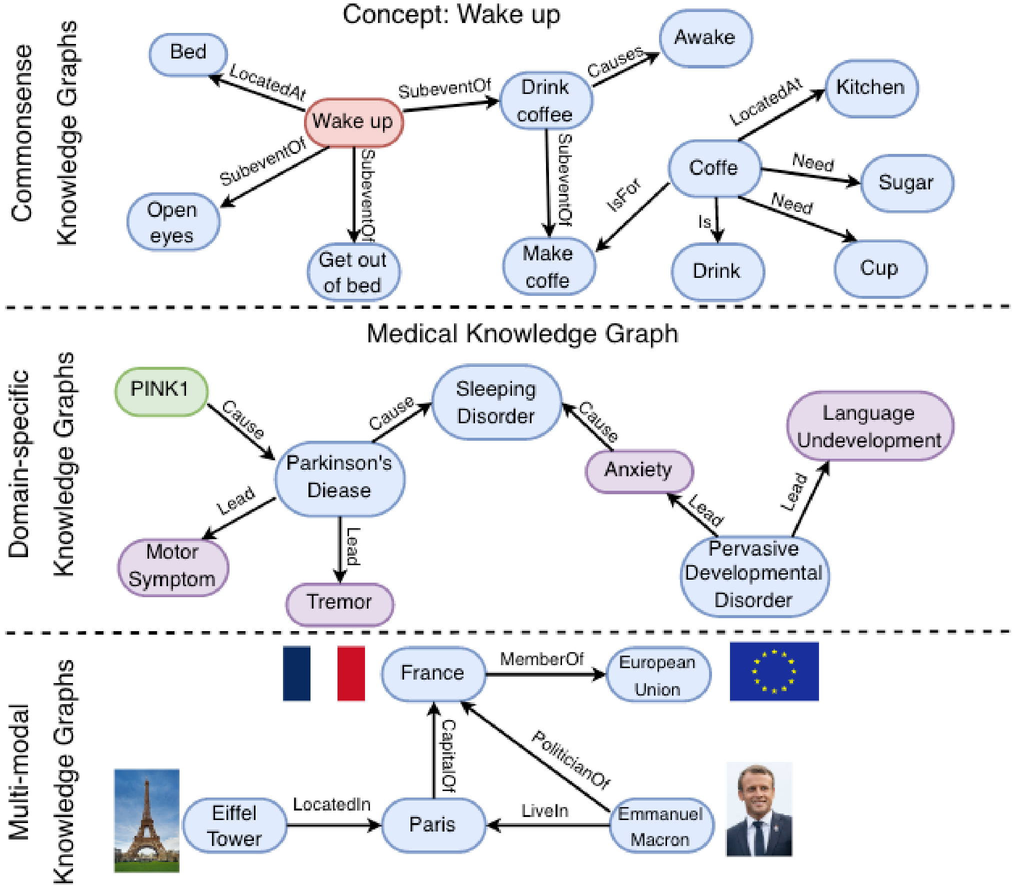
KGs help AGI

- KGs can provide **domain-specific knowledge.**

LLM is hungry

- More data more parameters
- Learn new knowledge inefficiently

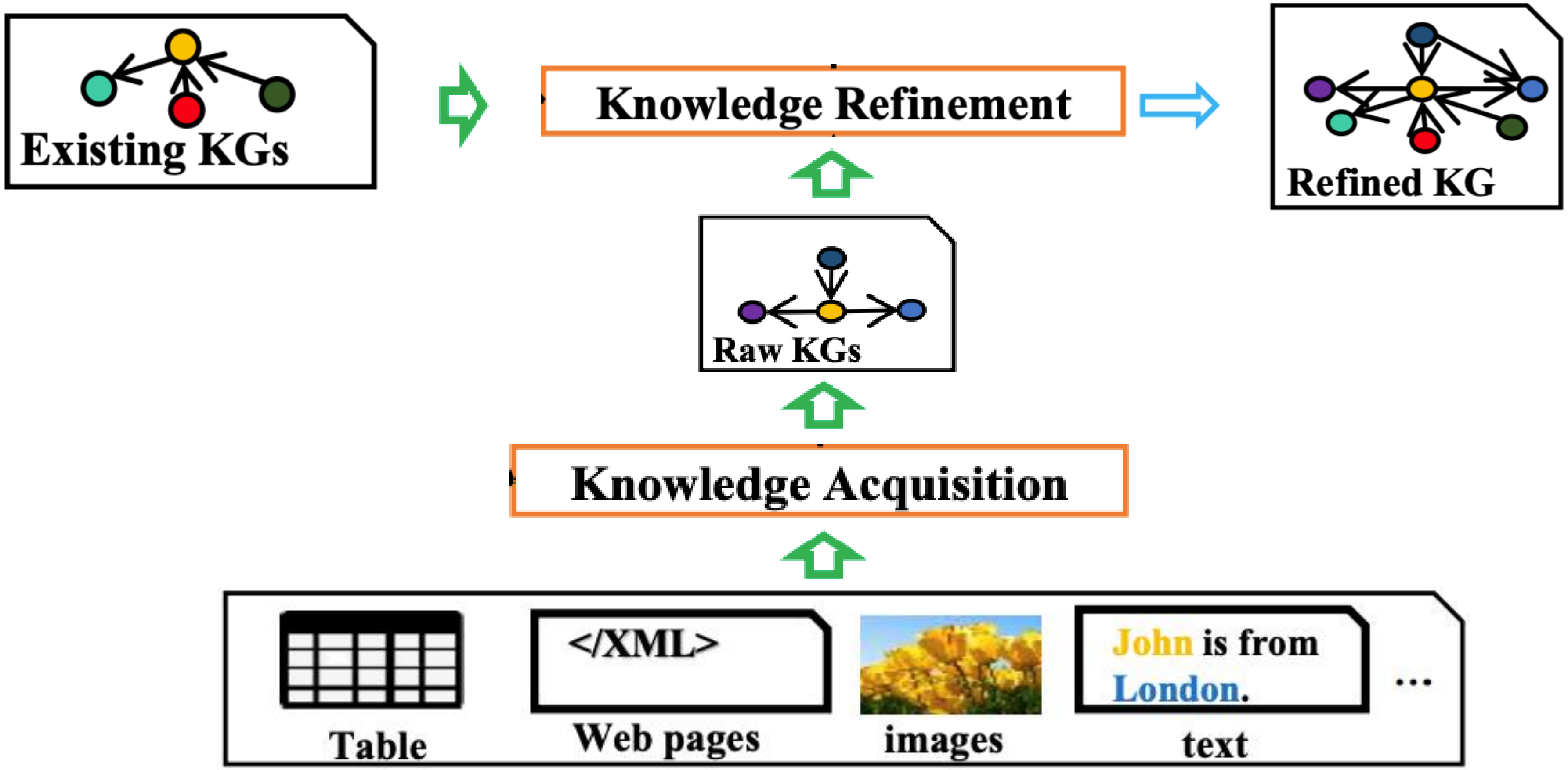




[1] Wang, J., Hu, X., Hou, W., Chen, H., Zheng, R., Wang, Y., ... & Xie, X. (2023). On the robustness of chatgpt: An adversarial and out-of-distribution perspective. arXiv preprint arXiv:2302.12095.
[2] Domain-specific knowledge graphs: A survey | Elsevier Enhanced Reader. (n.d.).
[3] Yixin Cao, et al. Trustworthy Natural Language Processing with Knowledge Guidance, WSDM-2023 Workshop

Limitations of KGs

- KGs are difficult to **construct.**

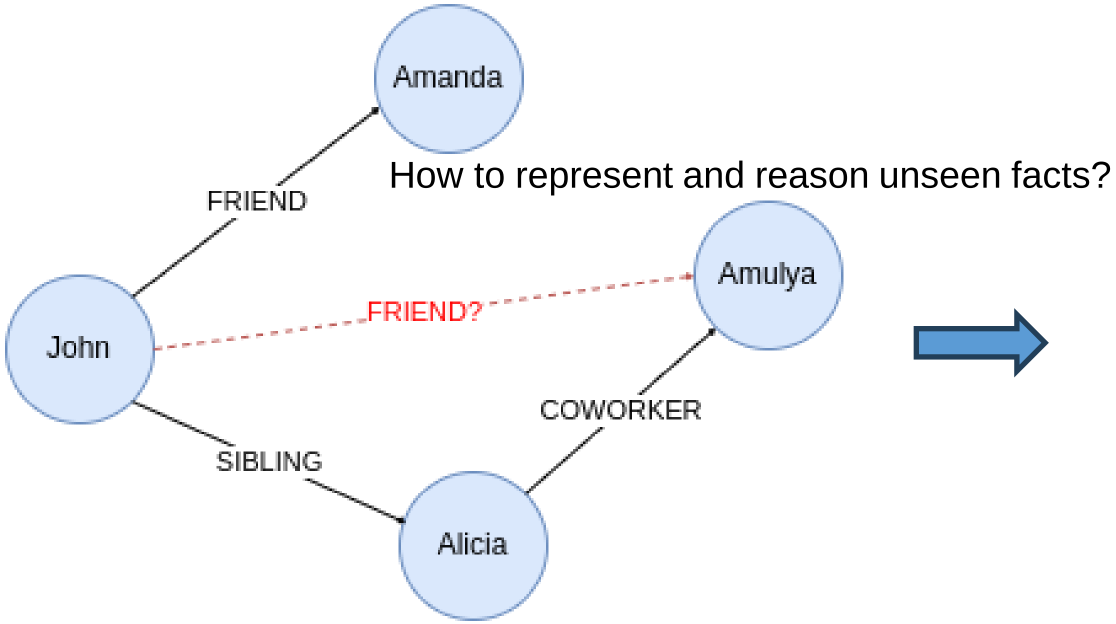


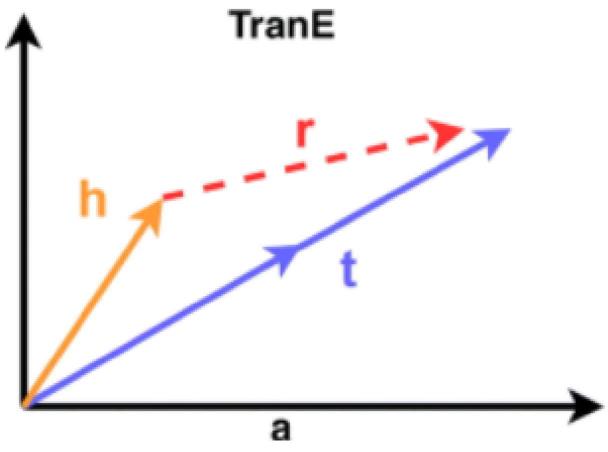
Heterogenous Data Source

Zhong, L., Wu, J., Li, Q., Peng, H., & Wu, X. (2023). A comprehensive survey on automatic knowledge graph construction. arXiv preprint arXiv:2302.05019.

Limitations of KGs

- KGs are **incomplete** and **noisy**.





Embedding model

How to represent and reason unseen facts?

Synergy of LLMs and KGs towards AGI

Knowledge Graphs (KGs)

Cons:
- Implicit Knowledge
- Hallucination
- Indecisiveness
- Black-box
- Lacking Domain-specific/New Knowledge

Pros:
- Structural Knowledge
- Accuracy
- Decisiveness
- Interpretability
- Domain-specific Knowledge
- Evolving Knowledge

Pros:
- General Knowledge
- Language Processing
- Generalizability

Cons:
- Incompleteness
- Lacking Language Understanding
- Unseen Facts

Large Language Models (LLMs)

Unifying Large Language Models and Knowledge Graphs: A Roadmap

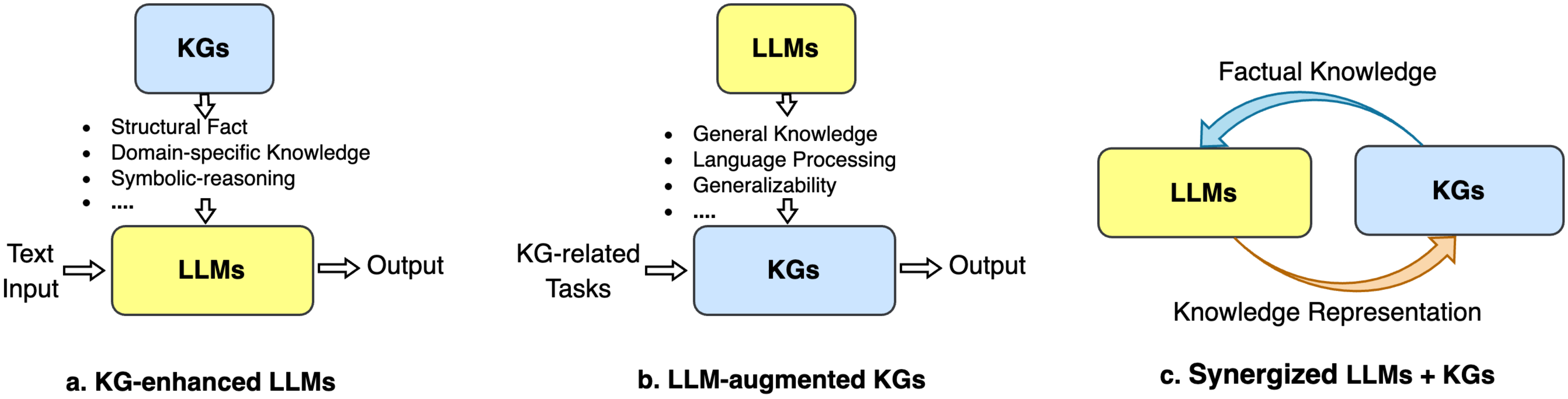
Shirui Pan, *Senior Member, IEEE*, Linhao Luo,
Yufei Wang, Chen Chen, Jiapu Wang, Xindong Wu, *Fellow, IEEE*

Abstract—Large language models (LLMs), such as ChatGPT and GPT4, are making new waves in the field of natural language processing and artificial intelligence, due to their emergent ability and generalizability. However, LLMs are black-box models, which often fall short of capturing and accessing factual knowledge. In contrast, Knowledge Graphs (KGs), Wikipedia and Huapu for example, are structured knowledge models that explicitly store rich factual knowledge. KGs can enhance LLMs by providing external knowledge for inference and interpretability. Meanwhile, KGs are difficult to construct and evolve by nature, which challenges the existing methods in KGs to generate new facts and represent unseen knowledge. Therefore, it is complementary to unify LLMs and KGs together and simultaneously leverage their advantages. In this article, we present a forward-looking roadmap for the unification of LLMs and KGs. Our roadmap consists of three general frameworks, namely, *1) KG-enhanced LLMs,* which incorporate KGs during the pre-training and inference phases of LLMs, or for the purpose of enhancing understanding of the knowledge learned by LLMs; *2) LLM-augmented KGs,* that leverage LLMs for different KG tasks such as embedding, completion, construction, graph-to-text generation, and question answering; and *3) Synergized LLMs + KGs*, in which LLMs and KGs play equal roles and work in a mutually beneficial way to enhance both LLMs and KGs for bidirectional reasoning driven by both data and knowledge. We review and summarize existing efforts within these three frameworks in our roadmap and pinpoint their future research directions.

Index Terms—Natural Language Processing, Large Language Models, Generative Pre-Training, Knowledge Graphs, Roadmap, Bidirectional Reasoning.

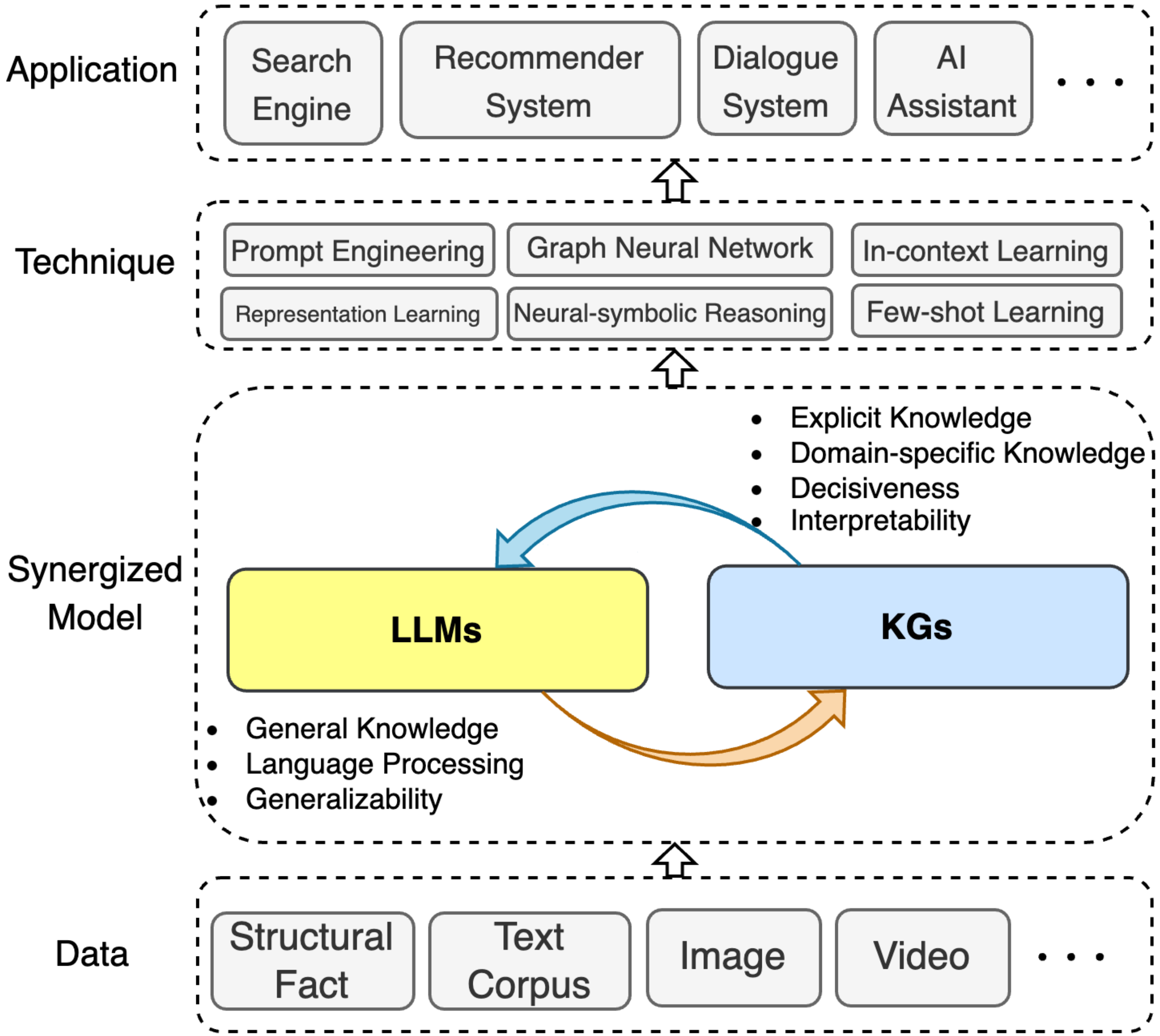
Pan, S., Luo, L., et al. (2024). Unifying large language models and knowledge graphs: A roadmap. *TKDE*.

Roadmaps



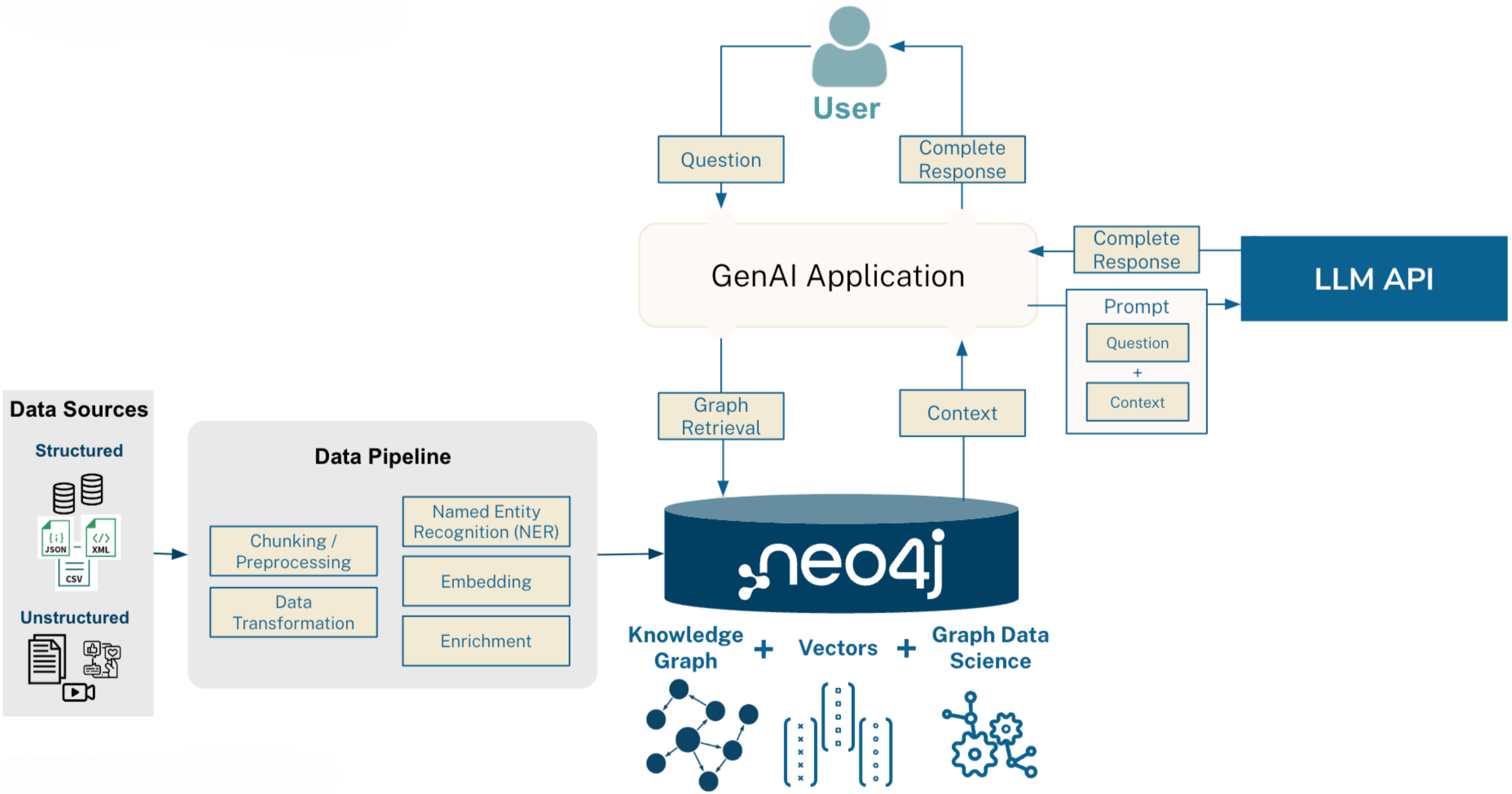
KGs
Structural Fact
Domain-specific Knowledge
Symbolic-reasoning
....
Text Input
LLMs
Output
a. KG-enhanced LLMs
LLMs
General Knowledge
Language Processing
Generalizability
....
KG-related Tasks
KGs
Output
b. LLM-augmented KGs
Factual Knowledge
LLMs
KGs
Knowledge Representation
c. Synergized LLMs + KGs

Roadmaps


Application
Search Engine
Recommender System
Dialogue System
AI Assistant
. . .
Technique
Prompt Engineering
Graph Neural Network
In-context Learning
Representation Learning
Neural-symbolic Reasoning
Few-shot Learning
Synergized Model
Explicit Knowledge
Domain-specific Knowledge
Decisiveness
Interpretability
LLMs
KGs
General Knowledge
Language Processing
Generalizability
Data
Structural Fact
Text Corpus
Image
Video
. . .

Applications

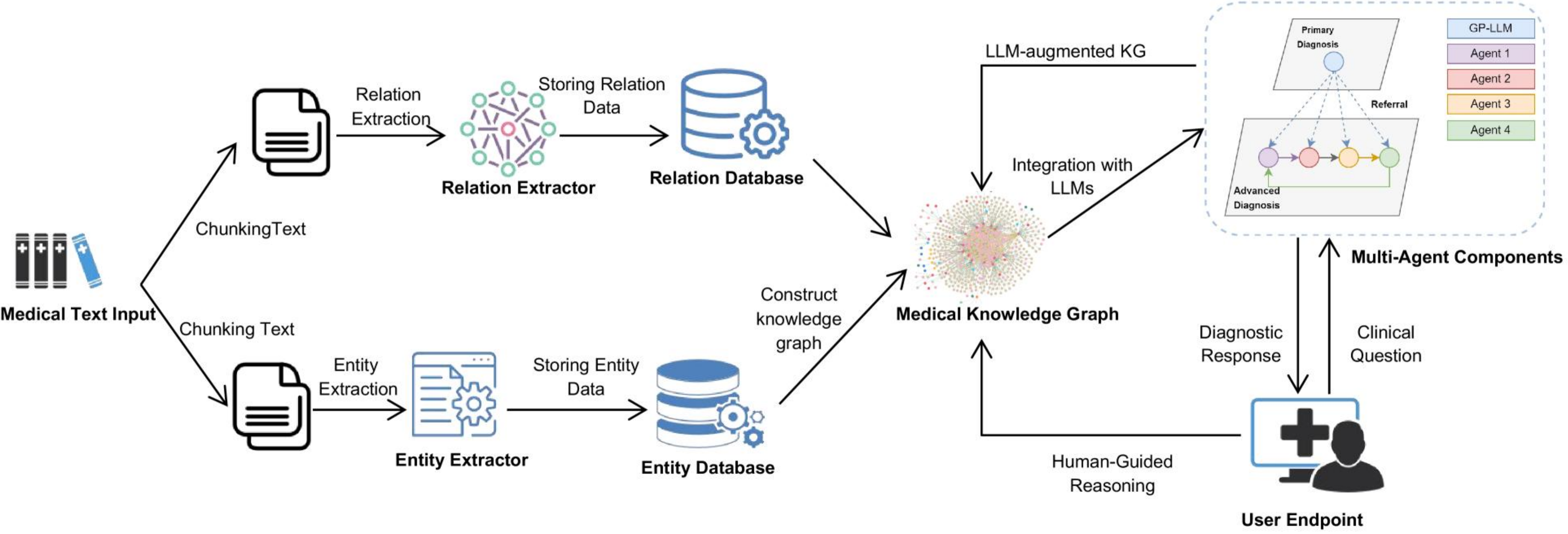
- KG-enhanced LLM retrieval augmented generation (GraphRAG).



https://neo4j.com/blog/genai/graphrag-manifesto/

Applications

- KG+LLM for medical diagnosis.



Zuo, Kaiwen, et al. "KG4Diagnosis: A Hierarchical Multi-Agent LLM Framework with Knowledge Graph Enhancement for Medical Diagnosis." *arXiv preprint arXiv:2412.16833* (2024).

Tutorial outline

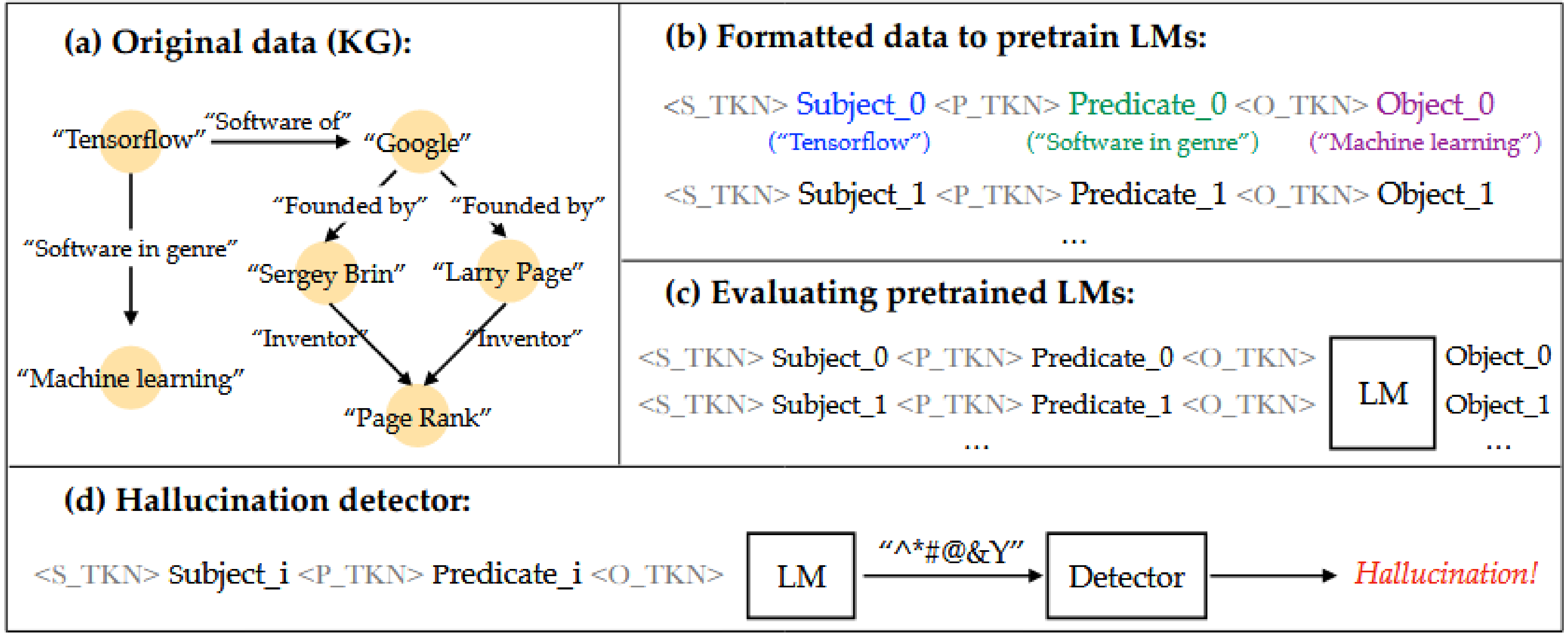
Time	Content	Part	Presenter
30 min	**•Introduction and background** • Artificial general intelligence (AGI) • Large language models (LLMs) and knowledge graphs (KGs) • Challenges and opportunities	Part 1	Shirui Pan
60 min	**•Knowledge graph-enhanced large language models** • KG-enhanced LLM Training • KG-enhanced LLM Reasoning • Unified KG+LLM Reasoning	Part 2	Linhao Luo
30 min break			
50 min	**•Large language model-enhanced knowledge graphs** • LLM-enhanced KG integrations • LLM-enhanced KG construction and completion • LLM-enhanced Multi-modality KG	Part 3	Carl Yang
30 min	**•Applications of synergized KG-LLM systems** • QA system • Recommender system	Part 4	Evgeny Kharlamov
10 min	**•Future directions and conclusion**	Part 5	Linhao Luo

Part 2:Knowledge graph-enhanced LLMs

- KG-enhanced LLM Training
- KG-enhanced LLM Reasoning
- Unified KG+LLM Reasoning

KG-enhanced LLM Training

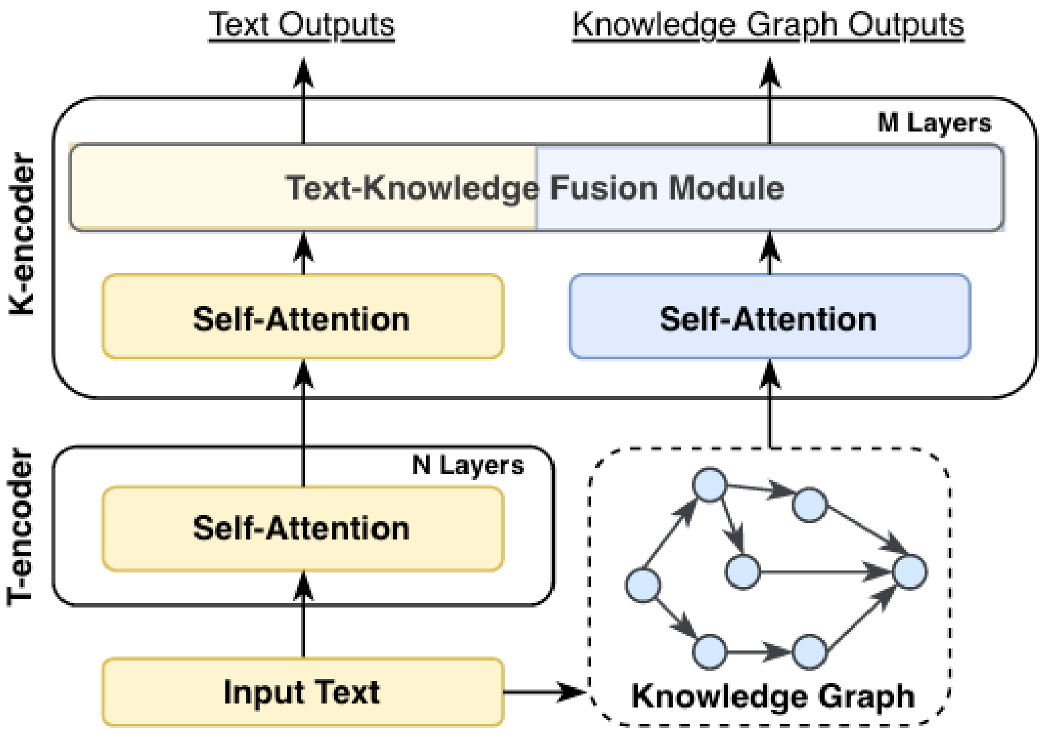
- **Generate data from KGs for LLM training.**



Hron, Jiri, et al. "Training Language Models on the Knowledge Graph: Insights on Hallucinations and Their Detectability." First Conference on Language Modeling.

KG-enhanced LLM Training

- **Integrating KGs by Additional Fusion Modules**
  - Additional modules to better capture the structure knowledge of KGs.

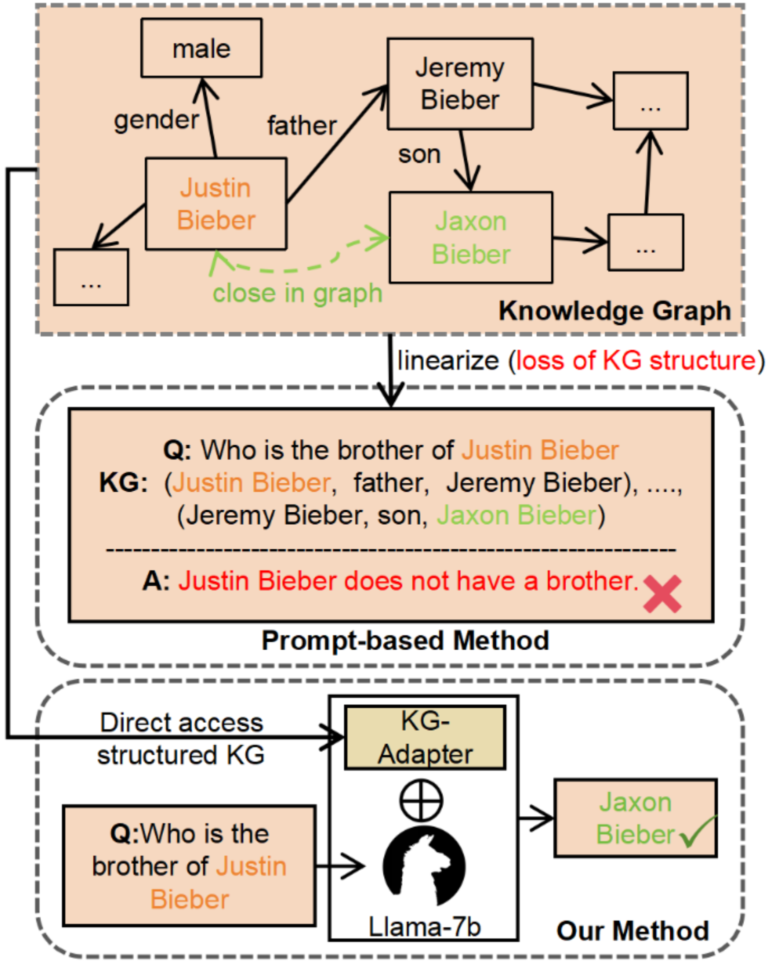


GNN …

Su, Y., Han, X., Zhang, Z., Lin, Y., Li, P., Liu, Z., ... & Sun, M. (2021). Cokebert: Contextual knowledge selection and embedding towards enhanced pre-trained language models. *AI Open*, *2*, 127-134.

KG-enhanced LLM Training

- **Integrating KGs by Additional Fusion Modules**



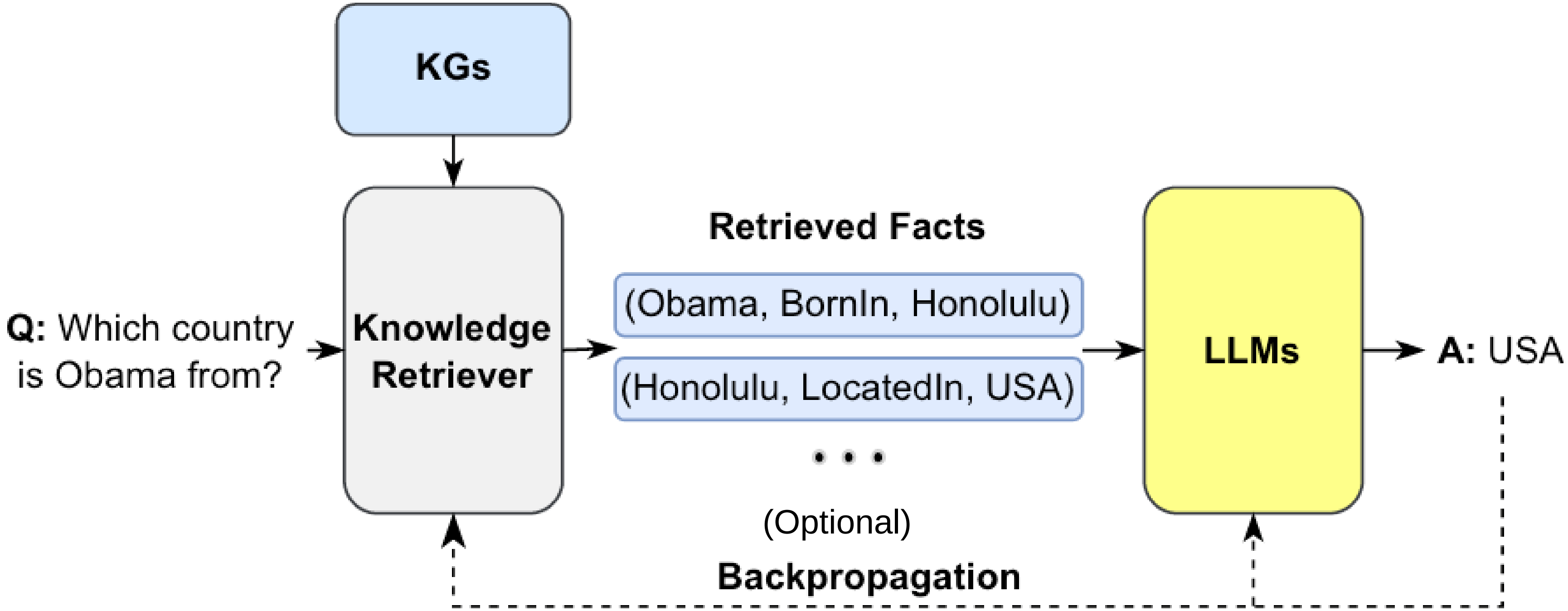
Tian, Shiyu, et al. "KG-adapter: Enabling knowledge graph integration in large language models through parameter-efficient fine-tuning." Findings of the Association for Computational Linguistics ACL 2024. 2024.

KG-enhanced LLM Reasoning

- KG-enhanced LLM training could fuse knowledge into LLMs.
- However, real-world knowledge is subject to change, and the pre-training approaches **cannot update knowledge without retraining the model.**
- **KG-enhanced LLM Reasoning** aims to separate the knowledge and text and inject the structural knowledge while LLM reasoning.

KG-enhanced LLM Reasoning

- **Retrieval-augmented Knowledge Fusion**
  - Retrieve-then-reasoning.
  - Parameters-free.
  - Can be applied to closed-source LLMs (e.g., ChatGPT).
  - Widely used in applications.



Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474.

KG-enhanced LLM Reasoning

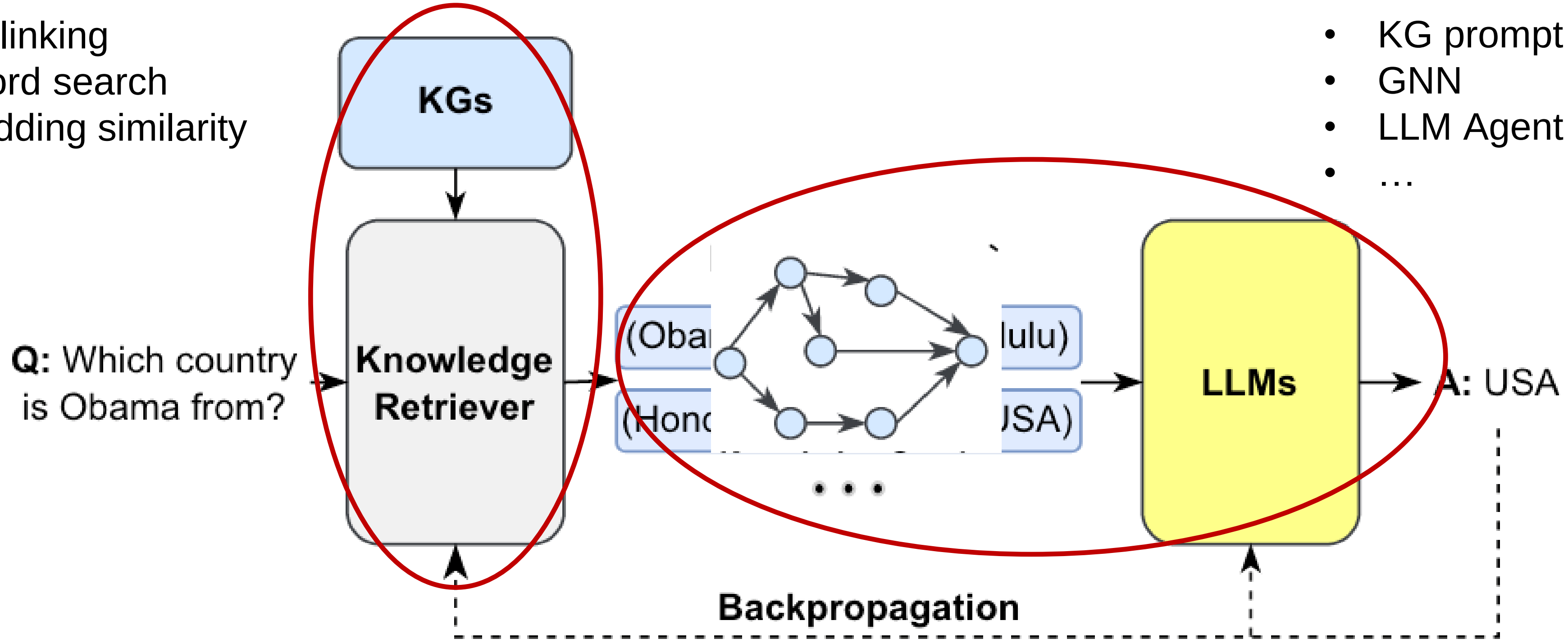
- **Retrieval-augmented Knowledge Fusion**
  - Techniques and challenges.

How to effectively retrieve on KGs?

- Entity linking
- Keyword search
- Embedding similarity
- …

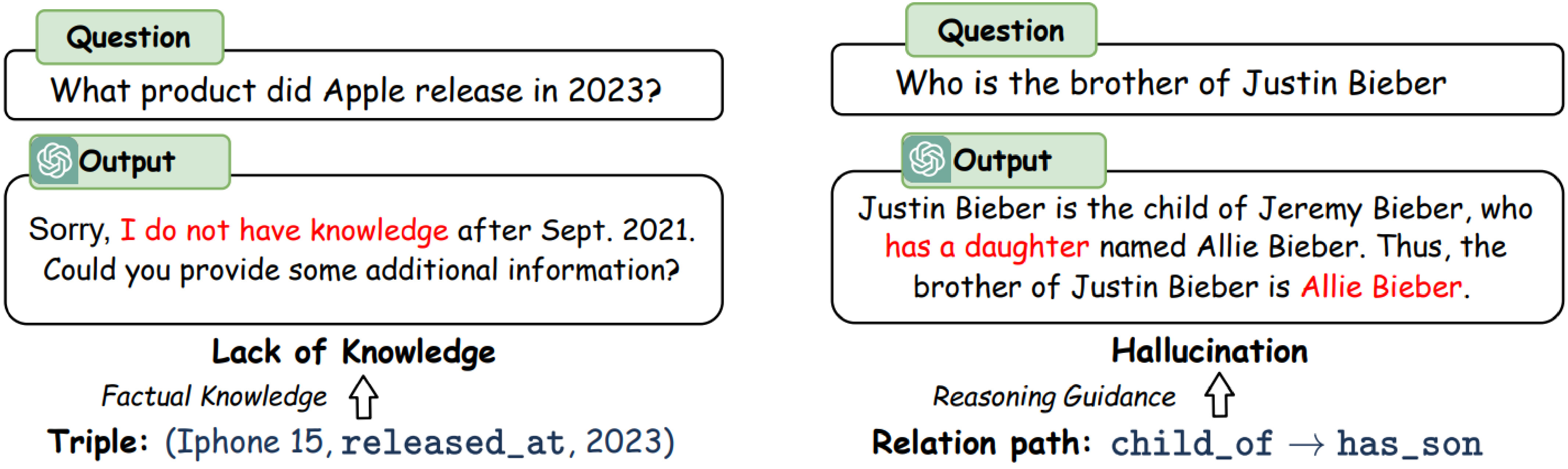
How to reason on retrieved KG structure?

- KG prompt
- GNN
- LLM Agent
- …



Reasoning on Graphs (RoG)

- **Aims to address two questions: Lack of Knowledge and Reasoning Hallucination**



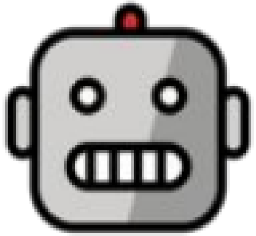
How does ROG work?

How to reason on graphs?

- **Plan-and-solve reasoning**
  - The plan is a hidden logic that can guide the reasoning.

LLMs

Q: Who is the child of Alice?



1. Planning

Step 1. Find the person that “Alice” is married to.
Step 2. Find the child of that person.

2. Reasoning

Step 1. “Alice” is married to “Bob”
Step 2. “Bob” is the father of “Charlie”
Thus, the answer is “Charlie”.

Wang, L., et al. (2023). Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. ACL 2023.

Reasoning on Graphs (RoG)

- **Relation paths as plans**
  - Relation paths are a **sequence of relations** that can serve as faithful plans for reasoning on graphs.
- **Example:**
  - **Question:**
    - Who is the child of Alice?
  - **Relation path z:**

$$z = \texttt{marry\_to} \to \texttt{father\_of}$$

Plan:

Step 1. Find the person that “Alice” is married to.
Step 2. Find the child of that person.

Execute the plan on KGs to **retrieve** reasoning paths.

  - **Reasoning paths w_z:**

$$w_z = \text{Alice} \xrightarrow{\texttt{marry\_to}} \text{Bob} \xrightarrow{\texttt{father\_of}} \boxed{\text{Charlie}}$$

Answer

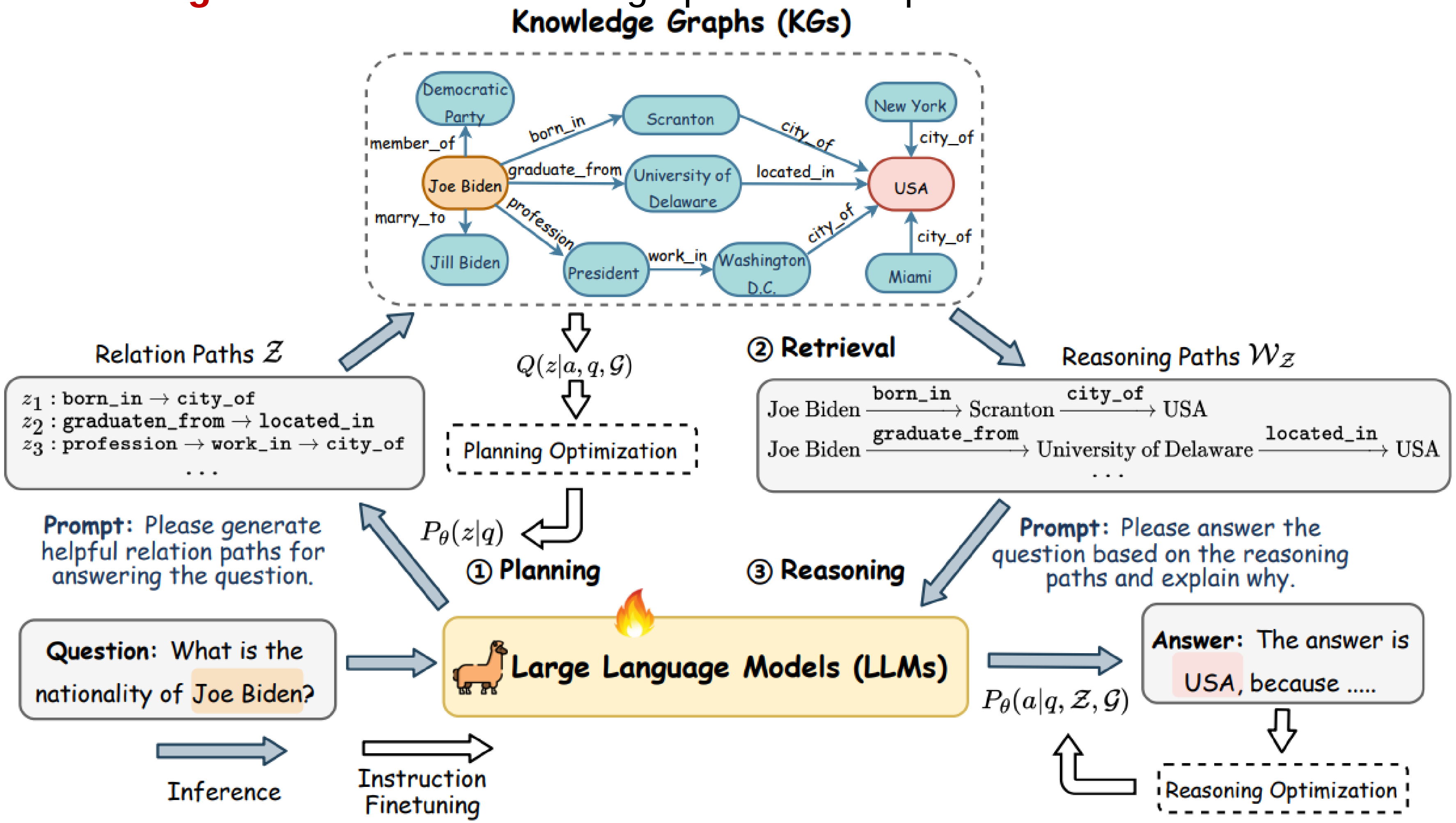
Luo Linhao, Li Yuan-Fang, Haffari Gholamreza, Pan Shirui, Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning, ICLR 2024

Reasoning on Graphs (RoG)

- **Planning-retrieval-reasoning.**

Planning: generate faithful relation paths as plans.

Retrieval-Reasoning: reason the answer on graphs with the plans.



Reasoning on Graphs (RoG)

- **Planning-retrieval-reasoning.**
  - **Planning:** generate faithful relation paths as plans.
  - **Retrieval-Reasoning:** reason the answer on graphs with the plans.
- **Challenges:**
  1. LLMs have zero knowledge of the relations in KGs.
  2. LLMs cannot understand the reasoning paths.

$$P_\theta(a|q,\mathcal{G}) = \sum_{z\in\mathcal{Z}} \underbrace{P_\theta(a|q,z,\mathcal{G})}_{\text{Reasoning}}\,\underbrace{P_\theta(z|q)}_{\text{Planning}},$$

ELBO Loss

$$\log P(a|q,\mathcal{G}) \geq \underbrace{\mathbb{E}_{z\sim Q(z)}[\log P_\theta(a|q,z,\mathcal{G})]}_{\text{Reasoning Optimization}} - \underbrace{D_{\mathrm{KL}}(Q(z)\|P_\theta(z|q))}_{\text{Planning Optimization}},$$

Reasoning on Graphs (RoG)

Planning Optimization:

- Distil the knowledge from KGs to generate faithful relation paths

- Estimate the posterior distribution of faithful relation paths with the **shortest path** connecting question and answer entities on KGs.

$$Q(z) \simeq Q(z|a,q,\mathcal{G}) = \begin{cases} \frac{1}{|\mathcal{Z}|}, \exists w_z(e_q, e_a) \in \mathcal{G}, \\ 0, else, \end{cases}$$

$$\Downarrow$$

$$\mathcal{L}_{\text{plan}} = D_{\text{KL}}(Q(z) \| P_\theta(z|q)) = D_{\text{KL}}(Q(z|a,q,\mathcal{G}) \| P_\theta(z|q)),$$
$$\simeq -\frac{1}{|\mathcal{Z}^*|} \sum_{z \in \mathcal{Z}^*} \log P_\theta(z|q),$$

Reasoning on Graphs (RoG)

Reasoning Optimization:

- Enable LLMs to conduct reasoning based on the retrieved reasoning paths

$$\mathcal{L}_{\text{reason}} = \mathbb{E}_{z \sim Q(z|a,q,\mathcal{G})}[\log P_\theta(a|q,z,\mathcal{G})] = \log P_\theta(a|q,\mathcal{Z}_K^*,\mathcal{G}).$$

Two instruction tunning tasks:

$$\mathcal{L} = \log \underbrace{P_\theta(a|q,\mathcal{Z}_K^*,\mathcal{G})}_{\text{Retrieval-reasoning}} + \underbrace{\frac{1}{|\mathcal{Z}^*|}\sum_{z\in Z^*}\log P_\theta(z|q)}_{\text{Planning}}$$

Planning-retrieval-reasoning

- **Planning:** generate faithful relation paths as plans.

> **Planning Prompt Template**
>
> Please generate a valid relation path that can be helpful for answering the following question: `<Question>`

- **Retrieval-Reasoning:** reason the answer on graphs with the plans.

> **Reasoning Prompt Template**
>
> Based on the reasoning paths, please answer the given question. Please keep the answer as simple as possible and return all the possible answers as a list.
>
> Reasoning Paths:
> `<Reasoning Paths>`
>
> Question:
> `<Question>`

Experiments

Table 2: Performance comparison with different baselines on the two KGQA datasets.

Type	Methods	WebQSP		CWQ	
		Hits@1	F1	Hits@1	F1
Embedding	KV-Mem (Miller et al., 2016)	46.7	34.5	18.4	15.7
	EmbedKGQA (Saxena et al., 2020)	66.6	-	45.9	-
	NSM (He et al., 2021)	68.7	62.8	47.6	42.4
	TransferNet (Shi et al., 2021)	71.4	-	48.6	-
	KGT5 Saxena et al. (2022)	56.1	-	36.5	-
Retrieval	GraftNet (Sun et al., 2018)	66.4	60.4	36.8	32.7
	PullNet (Sun et al., 2019)	68.1	-	45.9	-
	SR+NSM (Zhang et al., 2022)	68.9	64.1	50.2	47.1
	SR+NSM+E2E (Zhang et al., 2022)	69.5	64.1	49.3	46.3
Semantic Parsing	SPARQL (Sun et al., 2020)	-	-	31.6	-
	QGG (Lan & Jiang, 2020)	73.0	73.8	36.9	37.4
	ArcaneQA (Gu & Su, 2022)	-	75.3	-	-
	RnG-KBQA (Ye et al., 2022)	-	76.2	-	-
LLMs	Flan-T5-xl (Chung et al., 2022)	31.0	-	14.7	-
	Alpaca-7B (Taori et al., 2023)	51.8	-	27.4	-
	LLaMA2-Chat-7B (Touvron et al., 2023)	64.4	-	34.6	-
	ChatGPT	66.8	-	39.9	-
	ChatGPT+CoT	75.6	-	48.9	-
LLMs+KGs	KD-CoT (Wang et al., 2023b)	68.6	52.5	55.7	-
	UniKGQA (Jiang et al., 2022)	77.2	72.2	51.2	49.1
	DECAF (DPR+FiD-3B) (Yu et al., 2022a)	82.1	**78.8**	-	-
	RoG	**85.7**	70.8	**62.6**	**56.2**

Performance on KGQA tasks.

Table 1: Statistics of datasets.

Datasets	#Train	#Test	Max #hop
WebQSP	2,826	1,628	2
CWQ	27,639	3,531	4

Table 4: Effects of integrating the planning module of RoG with different LLMs for reasoning.

Methods	WebQSP		CWQ	
	Hits@1	Recall	Hits@1	Recall
ChatGPT	66.77	49.27	39.90	35.07
ChatGPT + RoG Planning	**81.51**	**71.60**	**52.68**	**48.51**
Alpaca-7B	51.78	33.65	27.44	23.62
Alpaca-7B + RoG Planning	**56.16**	**74.20**	**44.04**	**38.46**
LLaMA2-Chat-7B	64.37	44.61	34.60	29.91
LLaMA2-Chat-7B + RoG Planning	**74.20**	**56.16**	**56.41**	**51.99**
Flan-T5-xl	30.95	17.08	14.69	12.25
Flan-T5-xl + RoG Planning	**67.87**	**44.93**	**37.81**	**32.57**

Plug-and-play with other LLMs

Faithful reasoning and interpretable results

- **Understand the structure and explain.**

Table 7: Example of the *hallucination* of ChatGPT + CoT and faithful reasoning of RoG.

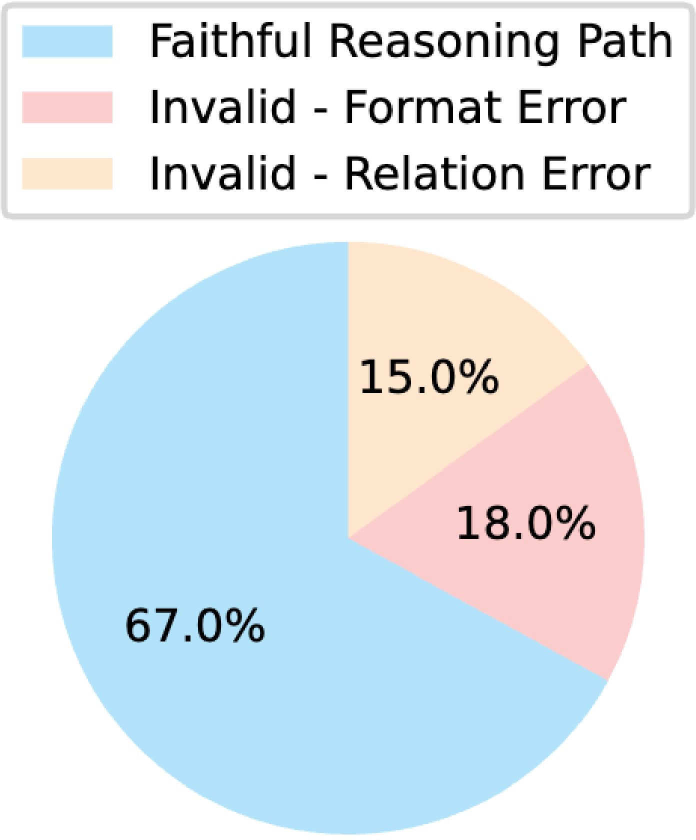
Question	Where is the home stadium of the team who won the 1946 World Series championship?
Answer	Busch Stadium
ChatGPT + CoT	Step 1: Identify the team that won the 1946 World Series championship. The team that won the 1946 World Series championship was the St. Louis Cardinals. Step 2: Determine the home stadium of the St. Louis Cardinals during that time. The St. Louis Cardinals' home stadium during the 1946 season was Sportsman's Park. Therefore, the possible answer to the question "Where is the home stadium of the team who won the 1946 World Series championship?" is: - Sportsman's Park.
Reasoning Paths	1946 World Series → sports.sports_team.championships → St. Louis Cardinals → sports.sports_team.arena_stadium → **Busch Stadium** 1946 World Series → sports.sports_team.championships → St. Louis Cardinals → sports.sports_team.arena_stadium → Roger Dean Stadium
RoG	Based on the reasoning paths provided, the home stadium of the team who won the 1946 World Series championship is **Busch Stadium**. This conclusion is based on the following path: **1. 1946 World Series → sports.sports_team.championships → St. Louis Cardinals → sports.sports_team.arena_stadium → Busch Stadium** The reasoning path leads us to the St. Louis Cardinals as the team that won the 1946 World Series, and Busch Stadium is the stadium associated with the St. Louis Cardinals. Therefore, **Busch Stadium** is the home stadium of the team that won the 1946 World Series championship.

Table 6: Example of *lack of knowledge* for ChatGPT + CoT and interpretable predictions of RoG.

Question	What type of government is used in the country with Northern District?
Answer	Parliamentary system
ChatGPT + CoT	To determine what type of government is used in a country with a Northern District, we need more specific information. The term "Northern District" does not specify the name of any particular country or provide enough context to identify its government type. Could you please provide additional details or clarify the specific country you are referring to?
Reasoning Paths	Northern District → location.administrative_division.first_level_division_of → Israel → government.form_of_government.countries → **Parliamentary system**
RoG	The type of government used in the country with the Northern District is a **Parliamentary system**. Explanations: 1. Northern District → location.administrative_division.first_level_division_of → Israel: The Northern District is a part of Israel. 2. Israel → government.form_of_government.countries → Parliamentary system: Israel has a Parliamentary system of government. So, by following the reasoning paths provided, we can conclude that the country with the Northern District, which is Israel, uses a Parliamentary system of government.

Graph-constrained Reasoning (GCR)

- **Findings:** Existing KG-enhanced reasoning methods (RoG) still cannot 100% ensure the faithful reasoning of LLMs.
- **Reason:** There are no constrains on the reasoning path generation. LLMs can generate paths that do not exist in the KGs.
- **Solution:** we introduce graph-constrained reasoning (GCR), a novel KG-guided reasoning paradigm to eliminate hallucinations and ensure accurate reasoning.

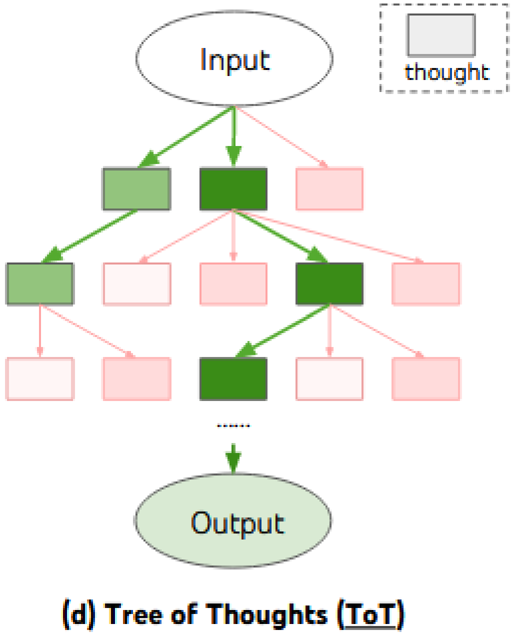


Reasoning Errors in RoG

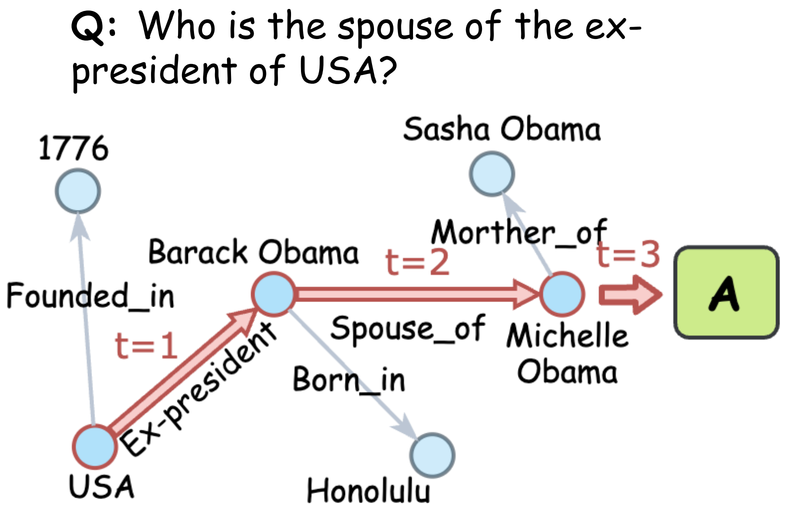
Luo, Linhao, et al. "Graph-constrained reasoning: Faithful reasoning on knowledge graphs with large language models." arXiv preprint arXiv:2410.13080 (2024).

From CoT to Graph-constrained Reasoning (GCR)

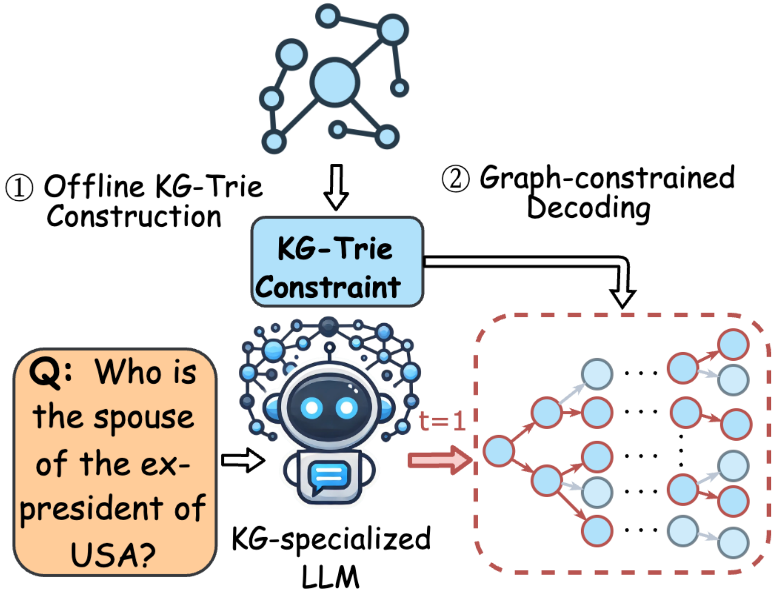
- **Graph-constrained Reasoning (GCR):**
  - Incorporates KGs into the decoding process of LLMs to achieve KG-grounded faithful reasoning (decoding on graphs)



LLM reasoning



Graph Reasoning

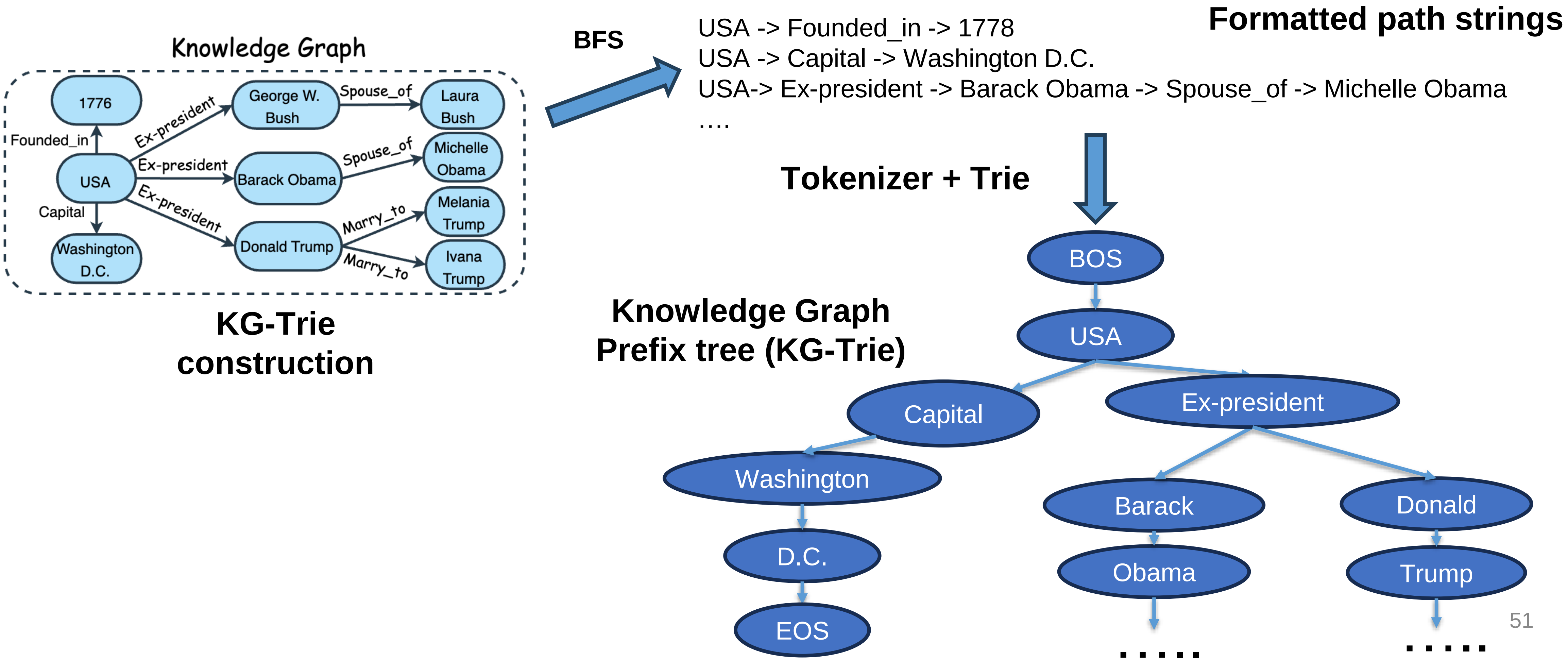


Graph-constrained Reasoning

Luo, Linhao, et al. "Graph-constrained reasoning: Faithful reasoning on knowledge graphs with large language models." arXiv preprint arXiv:2410.13080 (2024).

KG-Trie Construction

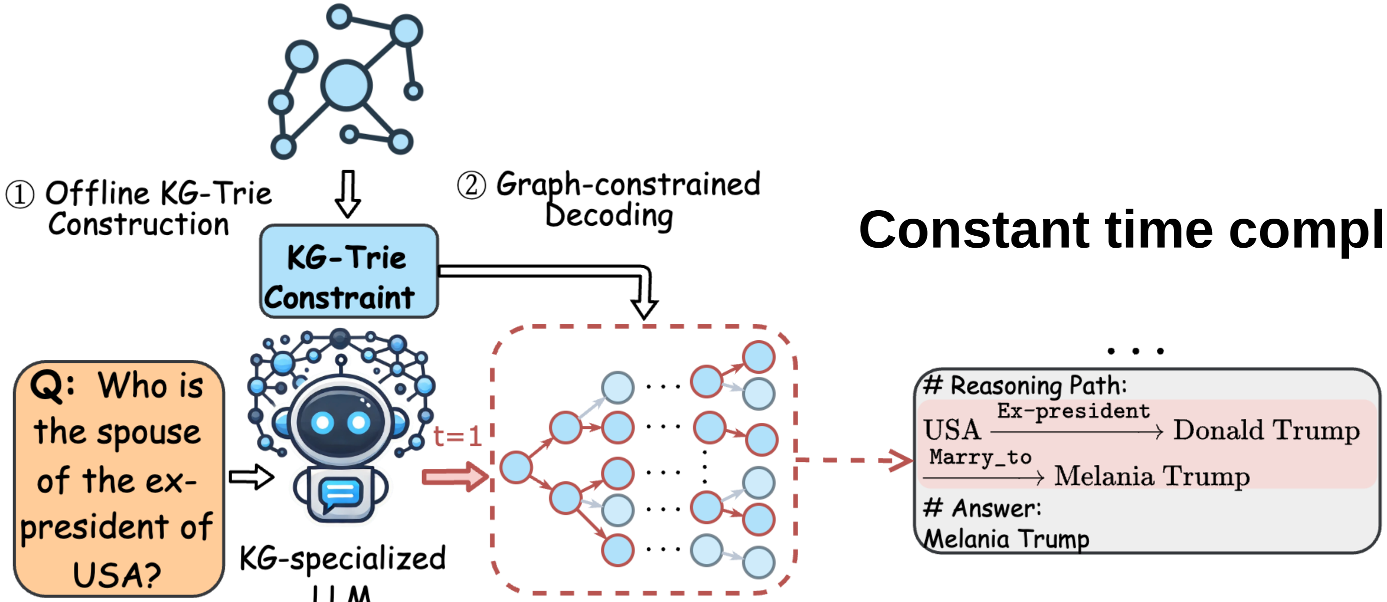
- We convert KGs into KG-Tries to facilitate efficient reasoning on KGs.



KG-Trie construction

Graph-constrained decoding

- We adopt KG-Trie as constraints to guide the decoding process of LLMs and only generate reasoning paths that are valid in KGs.



Constant time complexity: $O(|\mathcal{W}_z|)$

$$P_\phi(a, w_z|q) = \underbrace{P_\phi(a|q, w_z)}_{\text{Regular decoding}} \overbrace{\prod_{i=1}^{n} P_\phi(w_{z_i}|q, w_{z_{1,\dots,i-1}}) \mathcal{C}_\mathcal{G}(w_{z_i}|w_{z_{1,\dots,i-1}})}^{\text{Graph-constrained decoding}}, \quad (6)$$

$$\mathcal{C}_\mathcal{G}(w_{z_i}|w_{z_{1,\dots,i-1}}) = \begin{cases} 1, \exists \text{prefix}(w_{z_{1,\dots,i}}, w_z), \exists w_z \in \mathcal{W}_z, \\ 0, else, \end{cases} \quad (7)$$

Results

Table 1: Performance comparison with different baselines on the two KGQA datasets.

Types	Methods	WebQSP Hit	WebQSP F1	CWQ Hit	CWQ F1
LLM Reasoning	Qwen2-0.5B (Yang et al., 2024a)	26.2	17.2	12.5	11.0
	Qwen2-1.5B (Yang et al., 2024a)	41.3	28.0	18.5	15.7
	Qwen2-7B (Yang et al., 2024a)	50.8	35.5	25.3	21.6
	Llama-2-7B (Touvron et al., 2023)	56.4	36.5	28.4	21.4
	Llama-3.1-8B (Meta, 2024)	55.5	34.8	28.1	22.4
	GPT-4o-mini (OpenAI, 2024a)	63.8	40.5	63.8	40.5
	ChatGPT (OpenAI, 2022)	59.3	43.5	34.7	30.2
	ChatGPT+Few-shot (Brown et al., 2020)	68.5	38.1	38.5	28.0
	ChatGPT+CoT (Wei et al., 2022)	73.5	38.5	47.5	31.0
	ChatGPT+Self-Consistency (Wang et al., 2024)	83.5	63.4	56.0	48.1
Graph Reasoning	GraftNet (Sun et al., 2018)	66.7	62.4	36.8	32.7
	NSM (He et al., 2021)	68.7	62.8	47.6	42.4
	SR+NSM (Zhang et al., 2022)	68.9	64.1	50.2	47.1
	ReaRev (Mavromatis & Karypis, 2022)	76.4	70.9	52.9	47.8
KG+LLM	KD-CoT (Wang et al., 2023)	68.6	52.5	55.7	-
	EWEK-QA (Dehghan et al., 2024)	71.3	-	52.5	-
	ToG (ChatGPT) (Sun et al., 2024)	76.2	-	57.6	-
	ToG (GPT-4) (Sun et al., 2024)	82.6	-	68.5	-
	EffiQA (Dong et al., 2024)	82.9	-	69.5	
	RoG (Llama-2-7B) (Luo et al., 2024)	85.7	70.8	62.6	56.2
	GNN-RAG (Mavromatis & Karypis, 2024)	85.7	71.3	66.8	59.4
	GNN-RAG+RA (Mavromatis & Karypis, 2024)	90.7	73.5	68.7	60.4
	GCR (Llama-3.1-8B + ChatGPT)	**92.6**	73.2	72.7	60.9
	GCR (Llama-3.1-8B + GPT-4o-mini)	92.2	**74.1**	**75.8**	**61.7**

KGQA Performance

Table 2: Efficiency and performance comparison of different methods on WebQSP.

Types	Methods	Hit	Avg. Runtime (s)	Avg. # LLM Calls	Avg. # LLM Tokens
Retrieval-based	S-Bert	66.9	0.87	1	293
	BGE	72.7	1.05	1	357
	OpenAI-Emb.	79.0	1.77	1	330
	GNN-RAG	85.7	1.52	1	414
	RoG	85.7	2.60	2	521
Agent-based	ToG	75.1	16.14	11.6	7,069
	EffiQA	82.9	-	7.3	-
Ours	GCR	**92.6**	3.60	2	231

Efficiency and performance comparison

Findings:

- GCR achieves state-of-the-art performance
- GCR balances well between efficiency and effectiveness.

Results

Table 5: Examples of the faithful reasoning conducted by GCR. Red denotes the incorrect reasoning paths and answers, while **bold** denotes the correct paths and answers.

Case 1: Incorrect answers and hallucinated reasoning paths without constraints.	
Question	Who is niall ferguson 's wife?
Answer	Ayaan Hirsi Ali
GCR *w/o* constraint	# Reasoning Path: Niall Ferguson → people.person.children → Mabel Rose Ferguson → people.person.parents → Alyssa Mastromonaco #Answer: Alyssa Mastromonaco
GCR	# Reasoning Path: Niall Ferguson → people.person.children → Thomas Ferguson → people.person.parents → **Ayaan Hirsi Ali** #Answer: **Ayaan Hirsi Ali**
Case 2: Correct answers but hallucinated reasoning paths without constraints.	
Question	Where is jamarcus russell from?
Answer	Mobile
GCR *w/o* constraint	# Reasoning Path: JaMarcus Russell → people.person.place_of_birth → Tampa #Answer: **Mobile, Alabama**
GCR	# Reasoning Path: JaMarcus Russell → people.person.place_of_birth → **Mobile** #Answer: **Mobile**

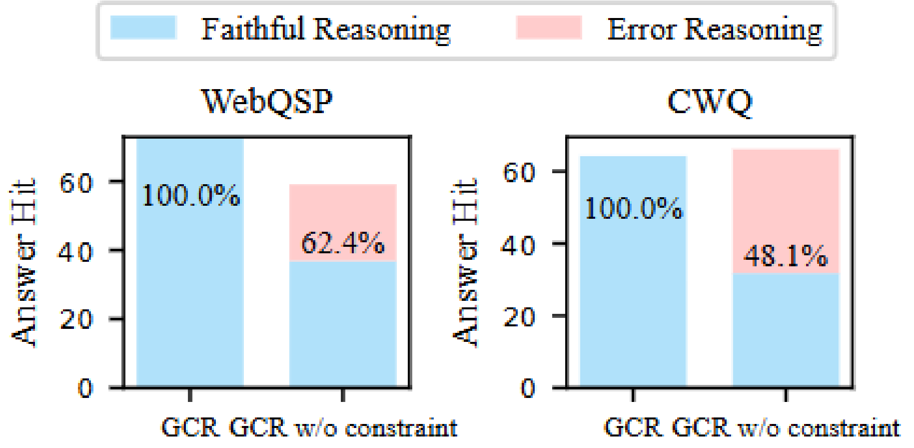


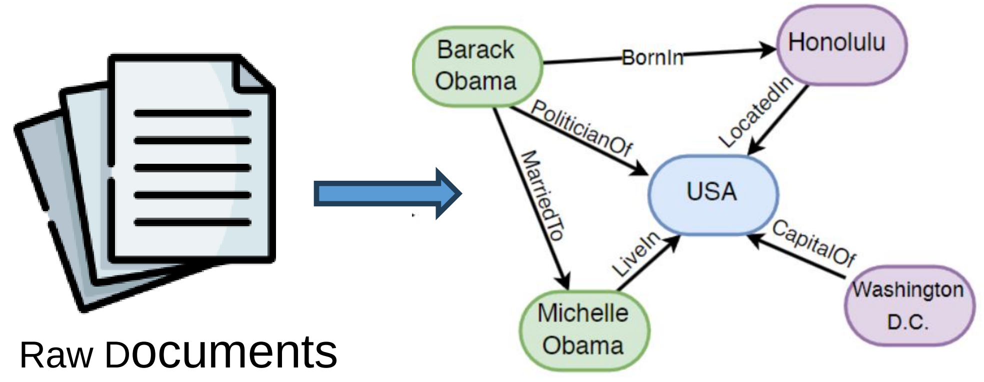
Figure 5: Analysis of performance and reasoning errors in GCR.

Faithful LLM reasoning with graph-constrained decoding

- The correct final answer may not result from a faithful reasoning of LLMs.
- Graph-constrained decoding can **eliminate** the hallucination when reasoning on KGs.
- Graph-constrained decoding can reduce the reasoning complexity and reach better performance.

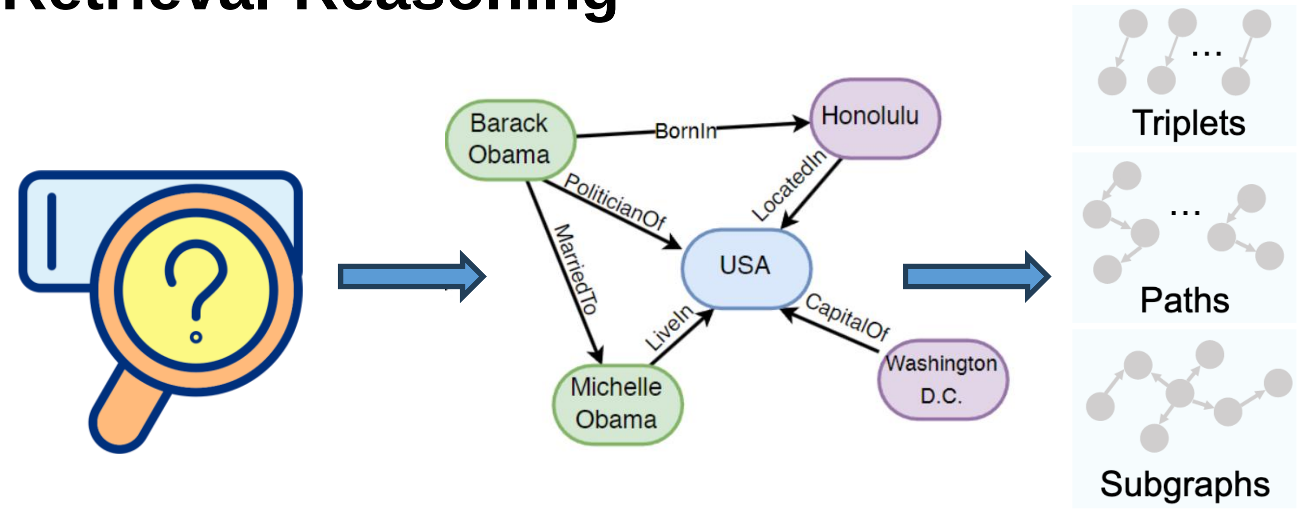
Limitation of KG-RAG

KG Construction



KGs are constructed from raw documents, which are often noisy and incomplete.

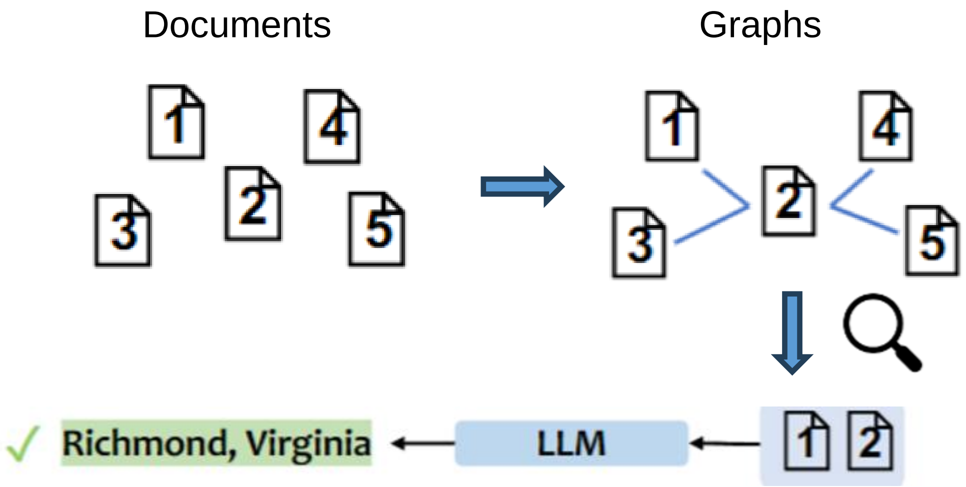
Retrieval-Reasoning



The construction of KGs leads to loss of information in the original documents.

Graph-enhanced RAG

- GraphRAG constructs a graph structure to explicitly model relationships between documents, allowing for more effective and efficient retrieval based on it.



1. Graph Construction

- hyperlinks
- reference
- …

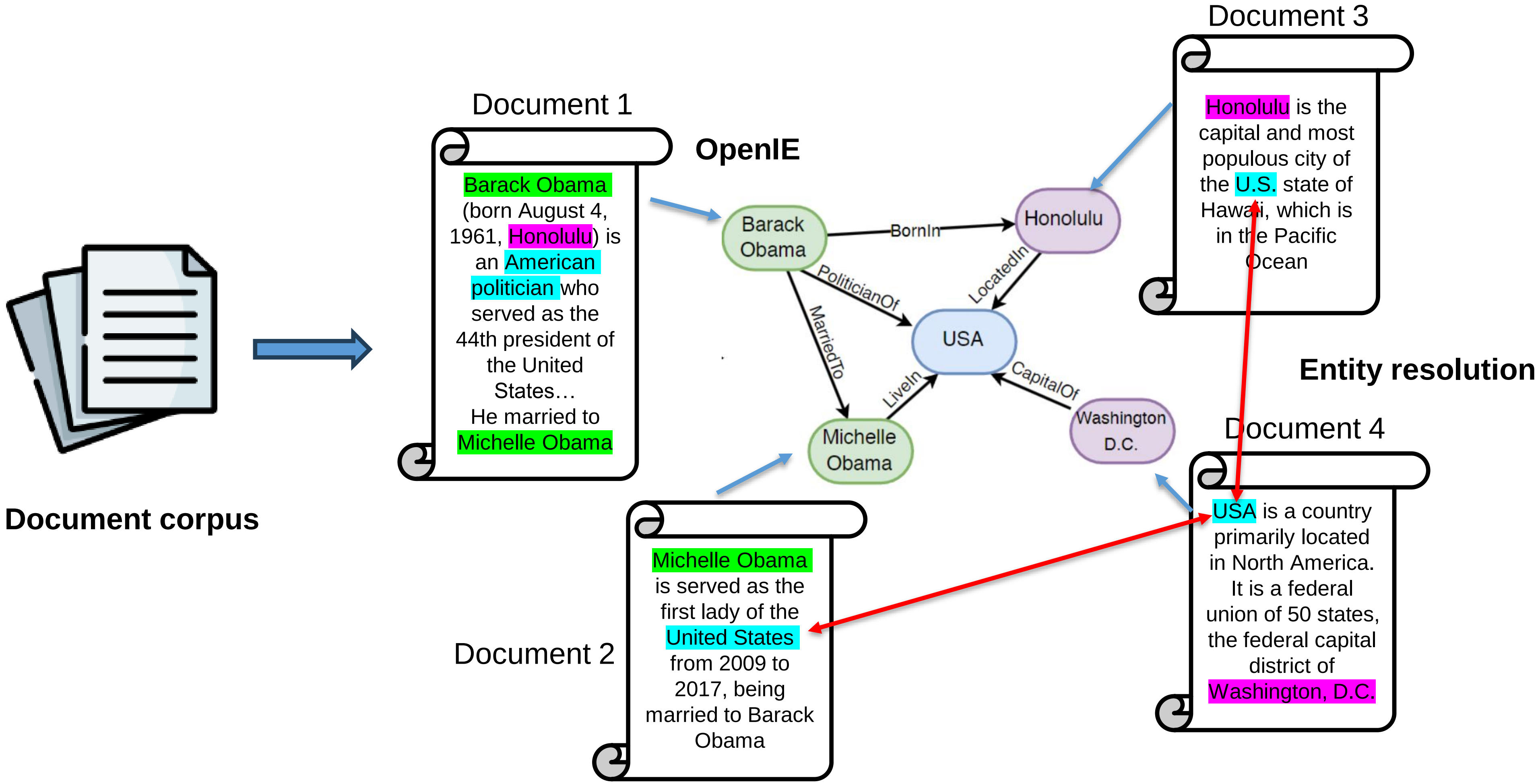
(Coarse relationships)

2. Graph-enhanced Retrieval

Graph-enhanced RAG

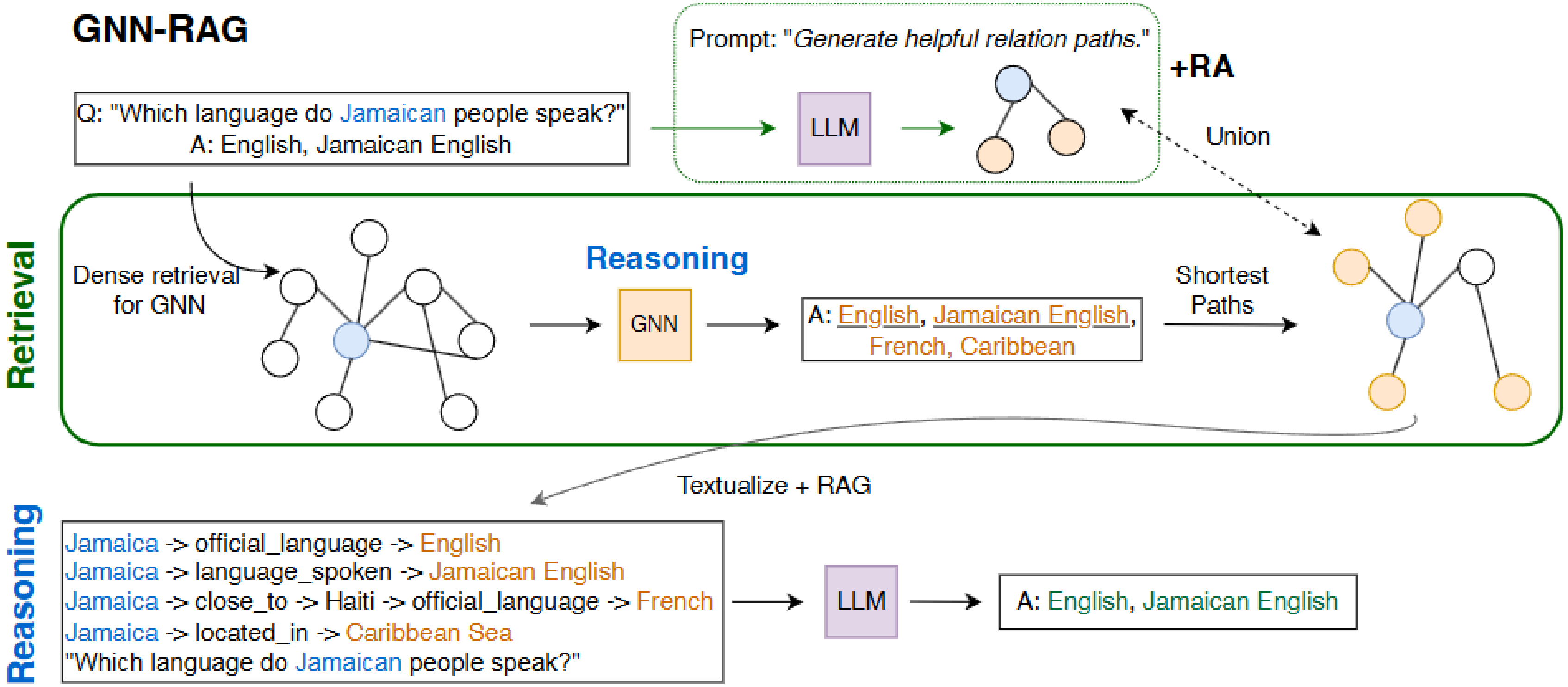
Knowledge Graph Index

- KGs can be used as a structural index of knowledge across multiple documents for accurate document retrieval.



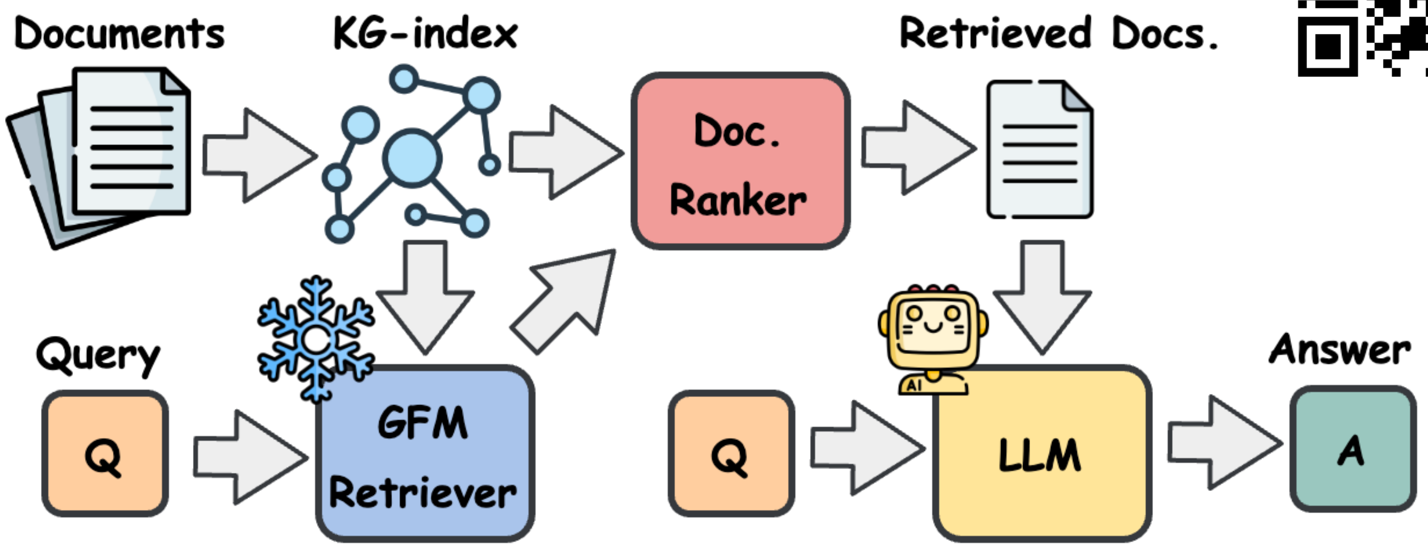
Unified KG+LLM Reasoning

- **GNNs** have demonstrated impressive performance in GraphRAG due to the powerful graph reasoning ability.
  - These methods still limit in generalizability as they need to be training from scratch in new datasets.



Mavromatis, Costas, and George Karypis. "Gnn-rag: Graph neural retrieval for large language model reasoning." arXiv preprint arXiv:2405.20139 (2024).

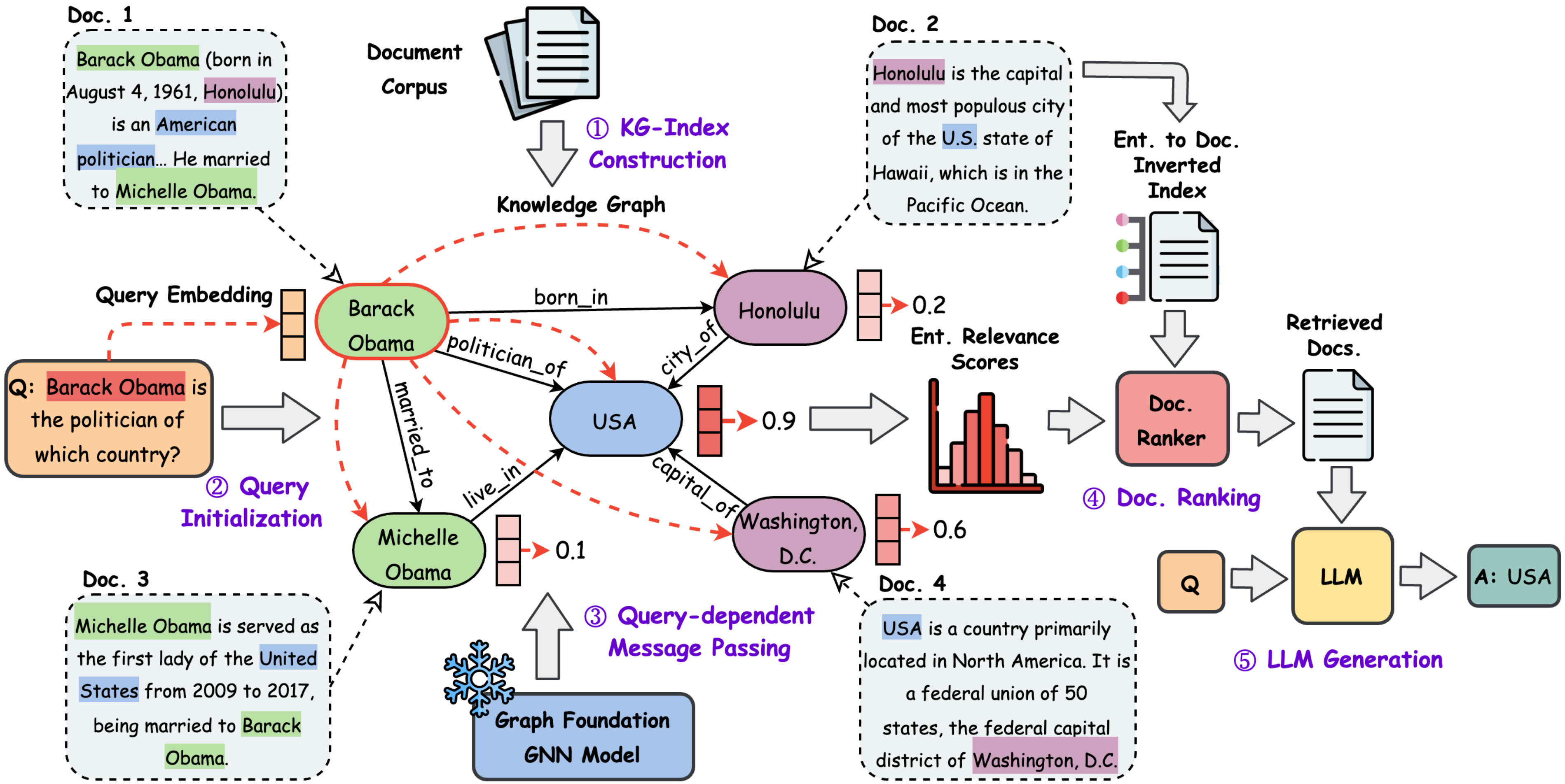
Graph Foundation Model for Retrieval Augmented Generation



- We propose a novel graph foundation model (GFM), powered by GNN for retrieval-augmented generation (GFM-RAG).
- We conducted large-scale training of GFM with 8M parameters on 60 KGs with over 14M triples derived from 700k documents across diverse datasets, allowing it to be directly applied to various unseen datasets.
- We achieves state-of-the-art performance across all datasets while demonstrating high efficiency, generalizability, and alignment with the neural scaling law, underscoring its potential for further enhancement.

Luo, Linhao, et al. "GFM-RAG: Graph Foundation Model for Retrieval Augmented Generation." *arXiv preprint arXiv:2502.01113* (2025).

Graph Foundation Model for Retrieval Augmented Generation



Query Initialization

$$\boldsymbol{q} = \text{SentenceEmb}(q),\ \boldsymbol{q} \in \mathbb{R}^d,$$

$$H^0 = \begin{cases} \boldsymbol{q}, & e \in \mathcal{E}_q, \\ \boldsymbol{0}, & \text{otherwise.} \end{cases}$$

Query-dependent Message Passing

Triple-level:

$$h_r^0 = \text{SentenceEmb}(r),\ h_r^0 \in \mathbb{R}^d,$$

$$m_e^{l+1} = \text{Msg}(h_e^l, g^{l+1}(h_r^l), h_{e'}^l), (e, r, e') \in \mathcal{G},$$

Entity-level:

$$h_e^{l+1} = \text{Update}(h_e^l, \text{Agg}(\{m_{e'}^{l+1} | e' \in \mathcal{N}_r(e), r \in \mathcal{R}\})),$$

Document Ranking

$$\mathcal{E}_q^T = \arg\text{top-}T(P_q),\ \mathcal{E}_q^T = \{e_1, \ldots, e_T\}.$$

$$F_e = \begin{cases} \frac{1}{\sum_{d \in \mathcal{D}} M[e,d]}, & e \in \mathcal{E}_q^T, \\ 0, & \text{otherwise,} \end{cases}$$

$$P_d = M^\top F_e,\ P_d \in \mathbb{R}^{|\mathcal{D}| \times 1}.$$

KG-index Construction

- **OpenIE:** gpt-4o-mini
- **Entity resolution:** colbert
  - Calculate the entities' embedding similarities and link entities with similar semantics by threshold σ.

$$s = h_{e_1}^T h_{e_2}, s > \sigma$$

Open Information Extraction

Instruction:

Your task is to construct an RDF (Resource Description Framework) graph from the given passages and named entity lists.
Respond with a JSON list of triples, with each triple representing a relationship in the RDF graph.
Pay attention to the following requirements:
- Each triple should contain at least one, but preferably two, of the named entities in the list for each passage.
- Clearly resolve pronouns to their specific names to maintain clarity.

Convert the paragraph into a JSON dict, it has a named entity list and a triple list.

One-Shot Demonstration:

Paragraph:

```
Radio City
Radio City is India's first private FM radio station and was started on 3 July 2001. It plays Hindi, English and regional songs. Radio City recently forayed into New Media in May 2008 with the launch of a music portal - PlanetRadiocity.com that offers music related news, videos, songs, and other music-related features.
```

```
{"named_entities": ["Radio City", "India", "3 July 2001", "Hindi","English", "May 2008", "PlanetRadiocity.com"]}

{"triples":
  [
    ["Radio City", "located in", "India"],
    ["Radio City", "is", "private FM radio station"],
    ["Radio City", "started on", "3 July 2001"],
    ["Radio City", "plays songs in", "Hindi"],
    ["Radio City", "plays songs in", "English"],
    ["Radio City", "forayed into", "New Media"],
    ["Radio City", "launched", "PlanetRadiocity.com"],
    ["PlanetRadiocity.com", "launched in", "May 2008"],
    ["PlanetRadiocity.com", "is", "music portal"],
    ["PlanetRadiocity.com", "offers", "news"],
    ["PlanetRadiocity.com", "offers", "videos"],
    ["PlanetRadiocity.com", "offers", "songs"]
  ]
}
```

Input:

Convert the paragraph into a JSON dict, it has a named entity list and a triple list.
Paragraph:

```
PASSAGE TO INDEX
```

{"named_entities": [NER LIST]}

Figure 9: Prompt for OpenIE during indexing.

Training Graph Foundation Model

- GFM is trained to predict **the target entities** given the query.

$$P_q = \sigma(\text{MLP}(H_q^L)),\ P_q \in \mathbb{R}^{|\mathcal{E}|\times 1}. \quad (10)$$

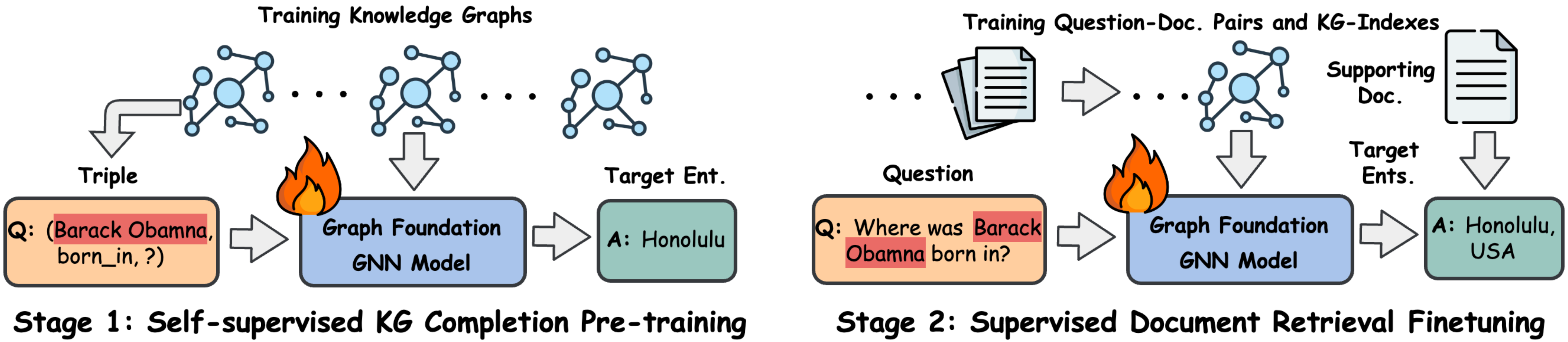
$$\mathcal{L}_{\text{BCE}} = -\frac{1}{|\mathcal{A}_q|}\sum_{e\in\mathcal{A}_q}\log P_q(e) - \frac{1}{|\mathcal{E}^-|}\sum_{|\mathcal{E}^-|}\log(1-P_q(e)), \quad (11)$$

$$\mathcal{L}_{\text{RANK}} = -\frac{1}{|\mathcal{A}_q|}\sum_{e\in\mathcal{A}_q}\frac{P_q(e)}{\sum_{e'\in\mathcal{E}^-}P_q(e')}. \quad (12)$$

$$\mathcal{L} = \alpha\mathcal{L}_{\text{BCE}} + (1-\alpha)\mathcal{L}_{\text{RANK}}. \quad (13)$$

Lin, Zhutian, et al. "Understanding the ranking loss for recommendation with sparse user feedback." KDD 2024.

Training Graph Foundation Model



Synthetic query-target pairs

Labeled query-target pairs

Experiments

- **Datasets:**
  - HotpotQA
  - MuSiQue
  - 2Wiki
- **Training:** 8 A100s
  - **Pre-training:** 1 epoch (15 hours)
  - **Fine-tuning:** 10 epoch (5 hours, 30 mins per epoch.)

Table 1. Statistics of the query-doc pairs and KGs used for training.

Dataset	#Q-doc Pair	#Document	#KG	#Entity	#Relation	#Triple
HotpotQA	20,000	204,822	20	1,930,362	967,218	6,393,342
MuSiQue	20,000	410,380	20	1,544,966	900,338	4,848,715
2Wiki	20,000	122,108	20	916,907	372,554	2,883,006
Total	**60,000**	**737,310**	**60**	**4,392,235**	**2,240,110**	**14,125,063**

Retrieval Performance

Table 2. Retrieval performance comparison.

Category	Method	HotpotQA		MuSiQue		2Wiki	
		R@2	R@5	R@2	R@5	R@2	R@5
Single-step	BM25	55.4	72.2	32.3	41.2	51.8	61.9
	Contriever	57.2	75.5	34.8	46.6	46.6	57.5
	GTR	59.4	73.3	37.4	49.1	60.2	67.9
	ColBERTv2	64.7	79.3	37.9	49.2	59.2	68.2
	RAPTOR	58.1	71.2	35.7	45.3	46.3	53.8
	Proposition	58.7	71.1	37.6	49.3	56.4	63.1
	LightRAG	38.8	54.7	24.8	34.7	45.1	59.1
	HippoRAG (Contriever)	59.0	76.2	41.0	52.1	71.5	89.5
	HippoRAG (ColBERTv2)	60.5	77.7	40.9	51.9	70.7	89.1
Multi-step	IRCoT + BM25	65.6	79.0	34.2	44.7	61.2	75.6
	IRCoT + Contriever	65.9	81.6	39.1	52.2	51.6	63.8
	IRCoT + ColBERTv2	67.9	82.0	41.7	53.7	64.1	74.4
	IRCoT + HippoRAG (Contriever)	65.8	82.3	43.9	56.6	75.3	93.4
	IRCoT + HippoRAG (ColBERTv2)	67.0	83.0	45.3	57.6	75.8	93.9
Single-step	GFM-RAG	**78.3**	**87.1**	**49.1**	**58.2**	**90.8**	**95.6**

naive methods

Graph-based

Findings:

- Graph-based method (HippoRAG) > naïve methods.
- Multi-step framework can improve the performance
- GFM-RAG can effectively conduct the multi-hop reasoning in a single step.

QA Performance

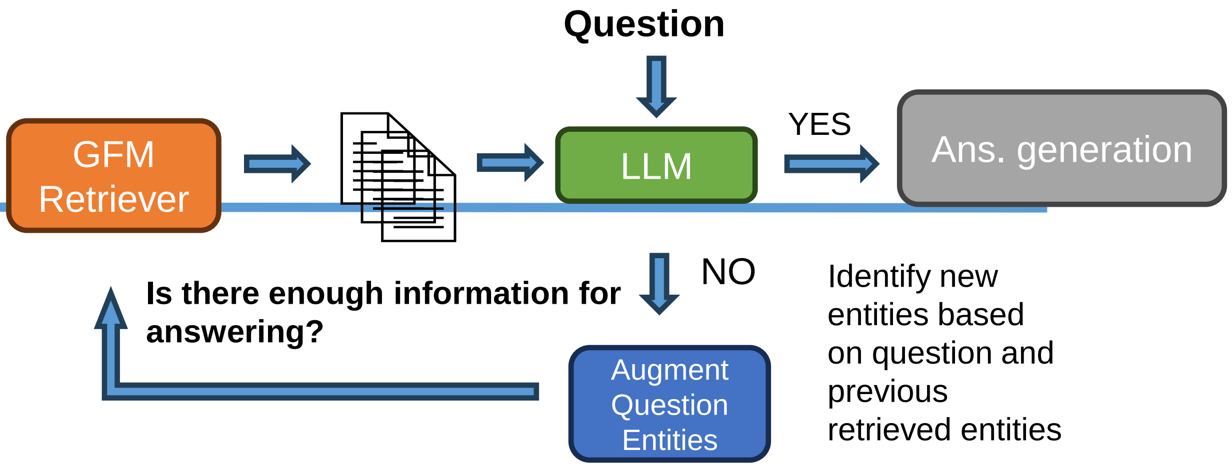


Table 3. Question answering performance comparison.

Category	Retriever	HotpotQA		MuSiQue		2Wiki	
		EM	F1	EM	F1	EM	F1
Single-step	None	30.4	42.8	12.5	24.1	31.0	39.0
	ColBERTv2	43.4	57.7	15.5	26.4	33.4	43.3
	HippoRAG (ColBERTv2)	41.8	55.0	19.2	29.8	46.6	59.5
Multi-step	IRCoT (ColBERTv2)	45.5	58.4	19.1	30.5	35.4	45.1
	IRCoT + HippoRAG (ColBERTv2)	45.7	59.2	21.9	33.3	47.7	62.7
Single-step	GFM-RAG	51.6	66.9	30.2	40.4	69.8	77.7
Multi-step	IRCoT + GFM-RAG	**56.0**	**71.8**	**36.6**	**49.2**	**72.5**	**80.8**

Findings:

- STOA performance.
- Compatibility with multi-step agent framework in multi-hop reasoning tasks. (Joint reasoning with LLMs)

Efficiency

Table 4. Retrieval efficiency and performance comparison.

Method	HotpotQA		MuSiQue		2Wiki	
	Time (s)	R@5	Time (s)	R@5	Time (s)	R@5
ColBERTv2	**0.035**	79.3	**0.030**	49.2	**0.029**	68.2
HippoRAG	0.255	77.7	0.251	51.9	0.158	89.1
IRCoT + ColBERTv2	1.146	82.0	1.152	53.7	2.095	74.4
IRCoT + HippoRAG	3.162	83.0	3.104	57.6	3.441	93.9
GFM−RAG	<u>0.107</u>	**87.1**	<u>0.124</u>	**58.2**	<u>0.060</u>	**95.6**

Findings:

- GFM-RAG achieves a great efficiency in performing multi-step reasoning in a single step.

Path Interpretations

Table 5. Path interpretations of GFM for multi-hop reasoning, where r^{-1} denotes the inverse of original relation.

Question	What *football club* was owned by the singer of "*Grow Some Funk of Your Own*"?
Answer	Watford Football Club
Sup. Doc.	["Grow Some Funk of Your Own", "Elton John"]
Paths	1.095: (grow some funk of your own, is a song by, elton john) → (elton john, equivalent, sir elton hercules john) → (sir elton hercules john, named a stand after^{-1}, **watford football club**) 0.915: (grow some funk of your own, is a song by, elton john) → (elton john, equivalent, sir elton hercules john) → (sir elton hercules john, owned, **watford football club**)
Question	When was the judge born who made notable contributions to the trial of the man who tortured, raped, and murdered eight student nurses from *South Chicago Community Hospital* on the night of *July 13-14, 1966*?
Answer	June 4, 1931
Sup. Doc.	["Louis B. Garippo", "Richard Speck"]
Paths	0.797: (south chicago community hospital, committed crimes at^{-1}, richard speck) → (richard speck, equivalent, trial of richard speck) → (trial of richard speck, made contributions during^{-1}, **louis b garippo**) 0.412: (south chicago community hospital, were from^{-1}, eight student nurses) → (eight student nurses, were from, south chicago community hospital) → (south chicago community hospital, committed crimes at^{-}1, **richard speck**)

The path's importance to the final prediction can be quantified by the partial derivative of the prediction score with respect to the triples at each layer.

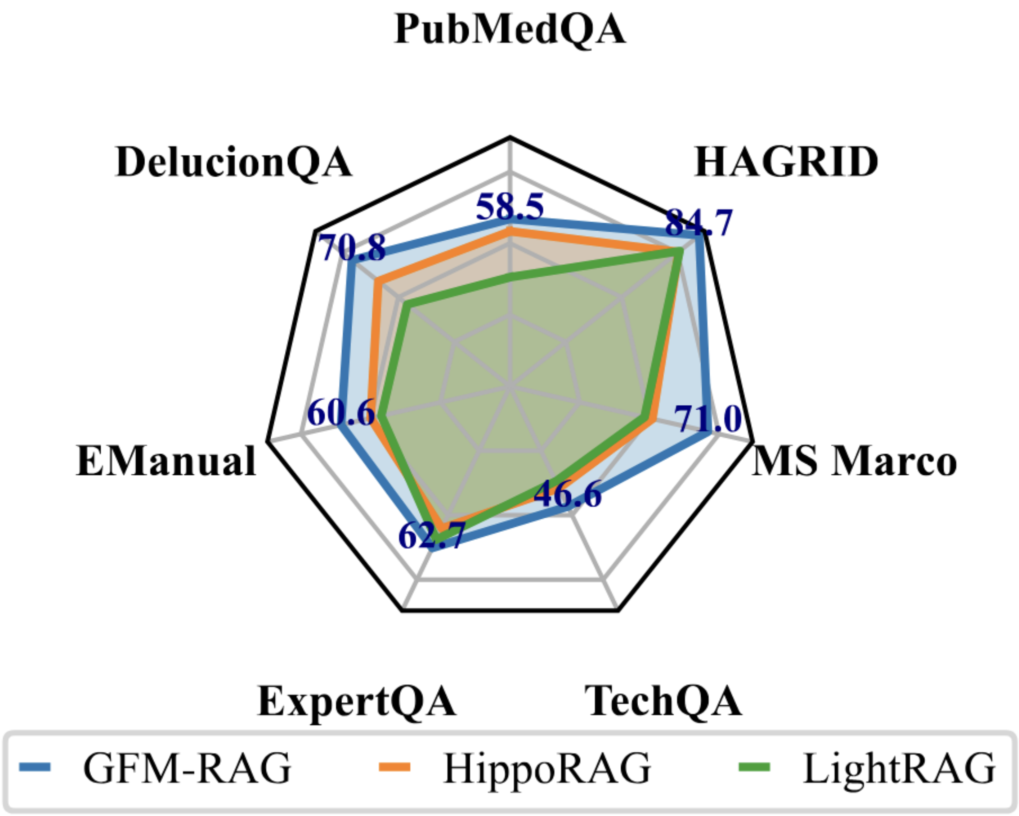
$$s_1, s_2, \ldots, s_L = \text{top-}k \frac{\partial p_e(q)}{\partial s_*}.$$

Generalizability

- Zero-shot transfer to new datasets

Table 6. Statistics of the dataset and constructed KG-index used for testing.

Dataset	Domain	#Test	#Document	#Entity	#Relation	#Triple
HotpotQA	Multi-hop	1,000	9,221	87,768	45,112	279,112
MuSiQue	Multi-hop	1,000	6,119	48,779	20,748	160,950
2Wiki	Multi-hop	1,000	11,656	100,853	55,944	319,618
PubMedQA	Biomedical	2,450	5,932	42,389	20,952	149,782
DelucionQA	Customer Support	184	235	2,669	2,298	6,183
TechQA	Customer Support	314	769	10,221	4,606	57,613
ExpertQA	Customer Support	203	808	11,079	6,810	16,541
EManual	Customer Support	132	102	695	586	1,329
MS Marco	General Knowledge	423	3,481	24,740	17,042	63,995
HAGRID	General Knowledge	1,318	1,975	23,484	18,653	48,969



Model Neural Scaling Law

- Performance of the foundation GNN model scales with the data and parameters.

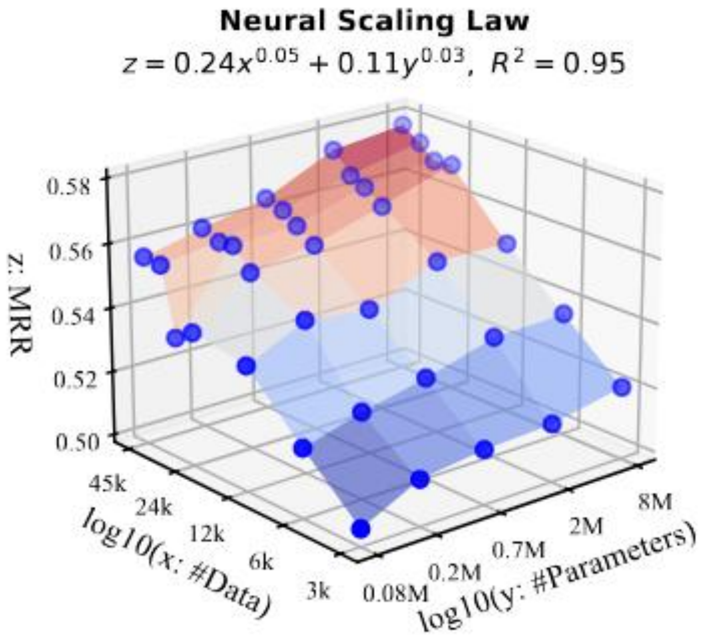


Figure 4. Neural scaling law of GFM−RAG.

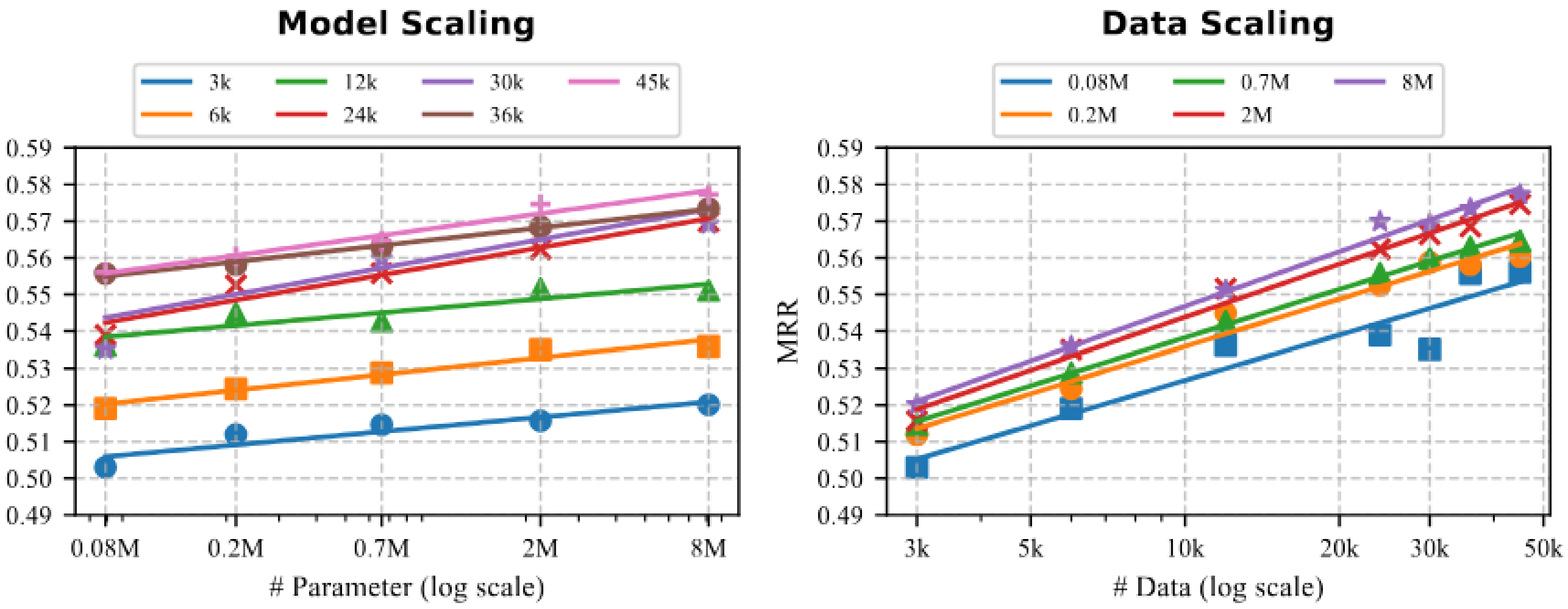


Figure 5. The illustration of the model and data scaling law of GFM−RAG.

Summary

- **Knowledge graph-enhanced large language models**

- KG-enhanced LLM Training
  - Generate training data from KGs
  - Inject KGs with additional modules

- KG-enhanced LLM reasoning
  - Reasoning on Graph (RoG)
  - Graph-constrained Reasoning (GCR)

- Unified KG+LLM Reasoning
  - Graph Foundation Model for Retrieval Augmented Generation (GFM-RAG)

Tutorial outline

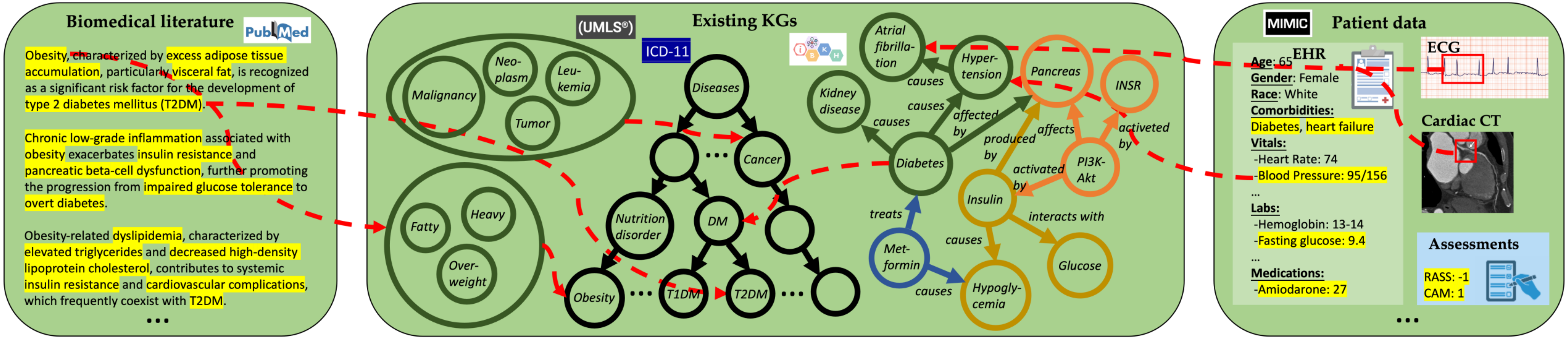
Time	Content	Part	Presenter
30 min	**•Introduction and background** • Artificial general intelligence (AGI) • Large language models (LLMs) and knowledge graphs (KGs) • Challenges and opportunities	Part 1	**Shirui Pan**
60 min	**•Knowledge graph-enhanced large language models** • KG-enhanced LLM Training • KG-enhanced LLM Reasoning • Unified KG+LLM Reasoning	Part 2	**Linhao Luo**
30 min break			
50 min	**•Large language model-enhanced knowledge graphs** • LLM-enhanced KG integrations • LLM-enhanced KG construction and completion • LLM-enhanced Multi-modality KG	Part 3	**Carl Yang**
30 min	**•Applications of synergized KG-LLM systems** • QA system • Recommender system	Part 4	**Evgeny Kharlamov**
10 min	**•Future directions and conclusion**	Part 5	**Linhao Luo**

Part 3:Large Language Models-enhanced Knowledge Graphs

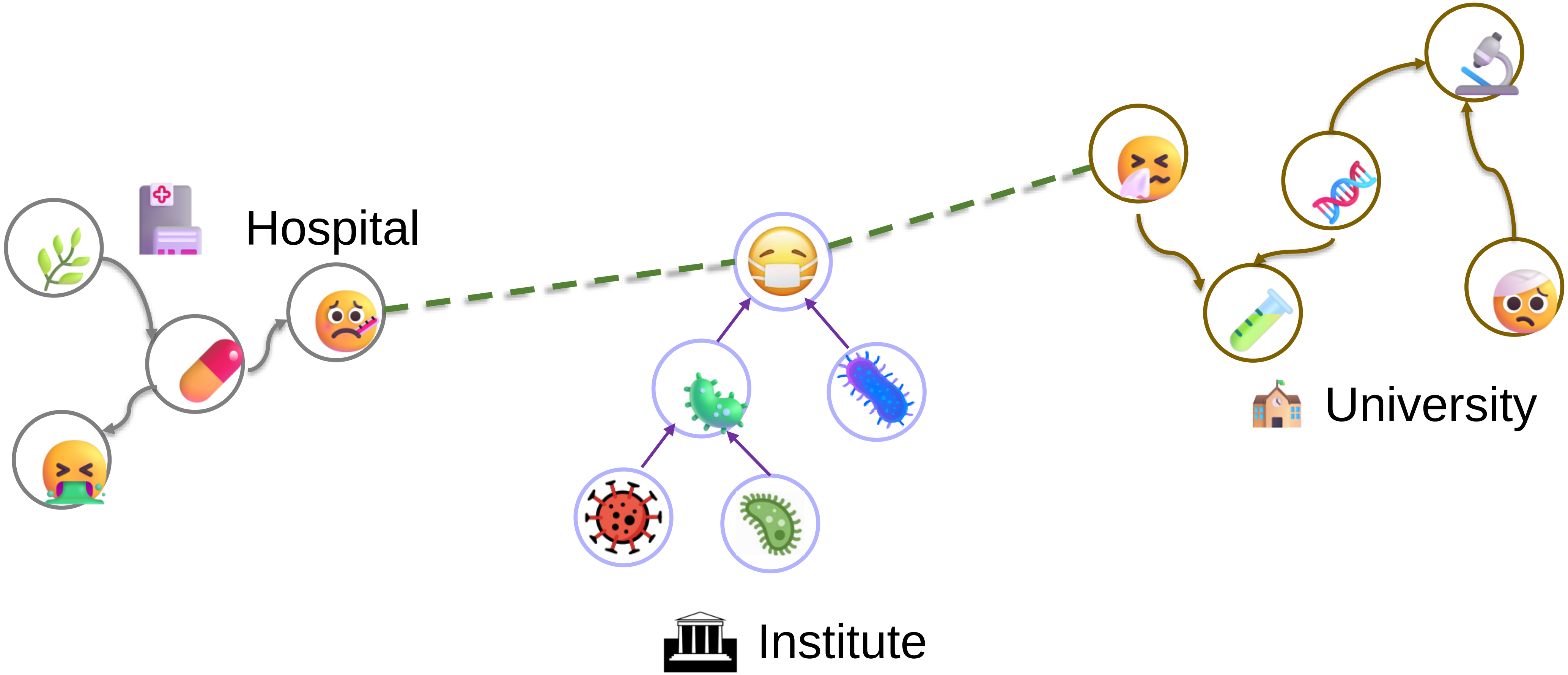
- Integrating existing KGs
- Constructing and Completing KGs
- Enriching KGs with multi-modality data

Integrating existing KGs

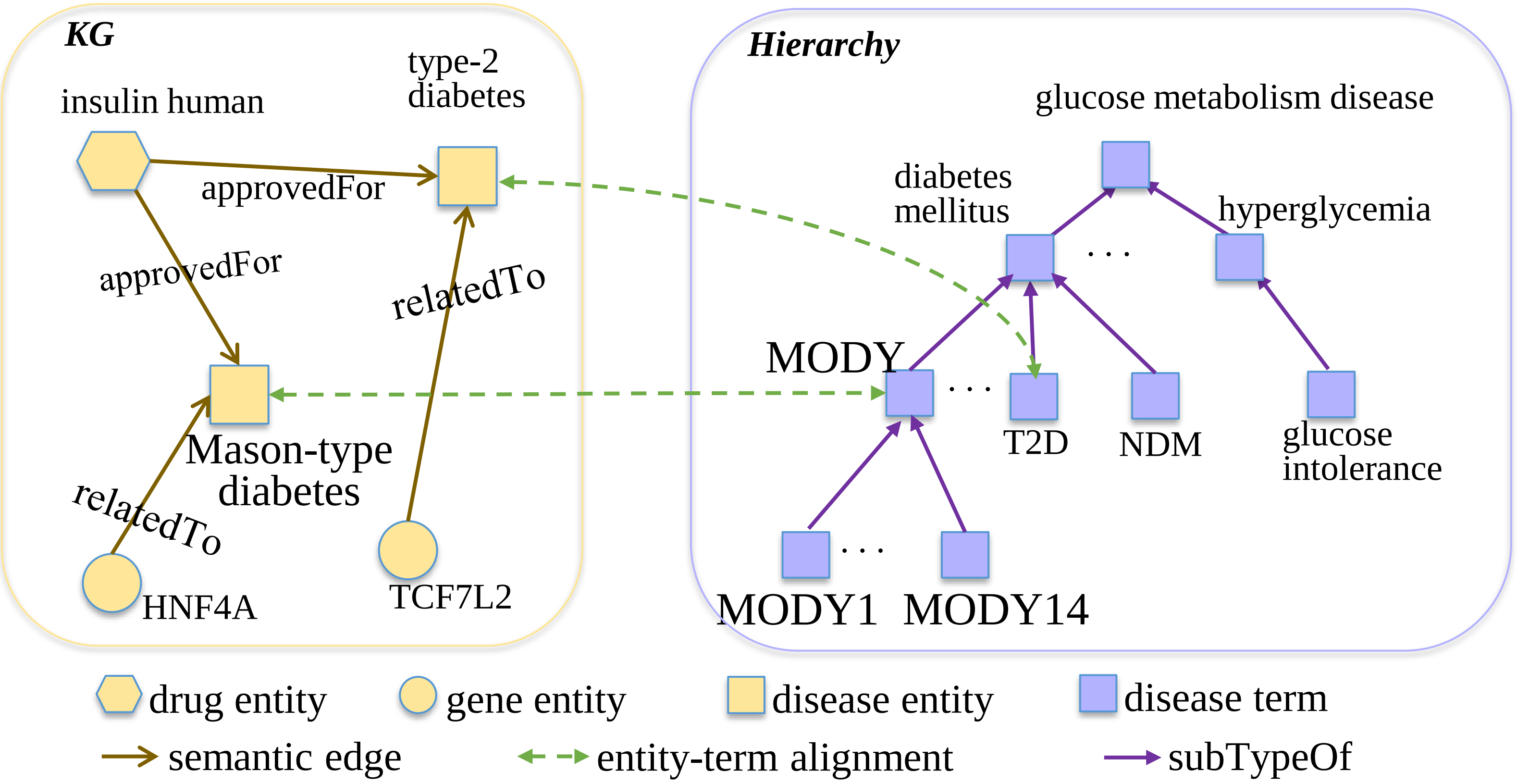
- KG integration (knowledge fusion or alignment) involves merging KGs from diverse sources and formats



Integrating existing KGs



Problem Definition



Definition of Biomedical Knowledge Fusion (BKF)

Given a KG $\mathcal{G} = (E, R, ET)$, and a hierarchy $\mathcal{H} = (T, TP)$, a set of pre-aligned entity-term pairs $[e_a, t_a]_{a=1}^{M}$, and a set of unaligned entities $[e_1, e_2, \dots, e_N] \in \mathcal{G}$.
The goal is to link each unaligned entity to the hierarchy:

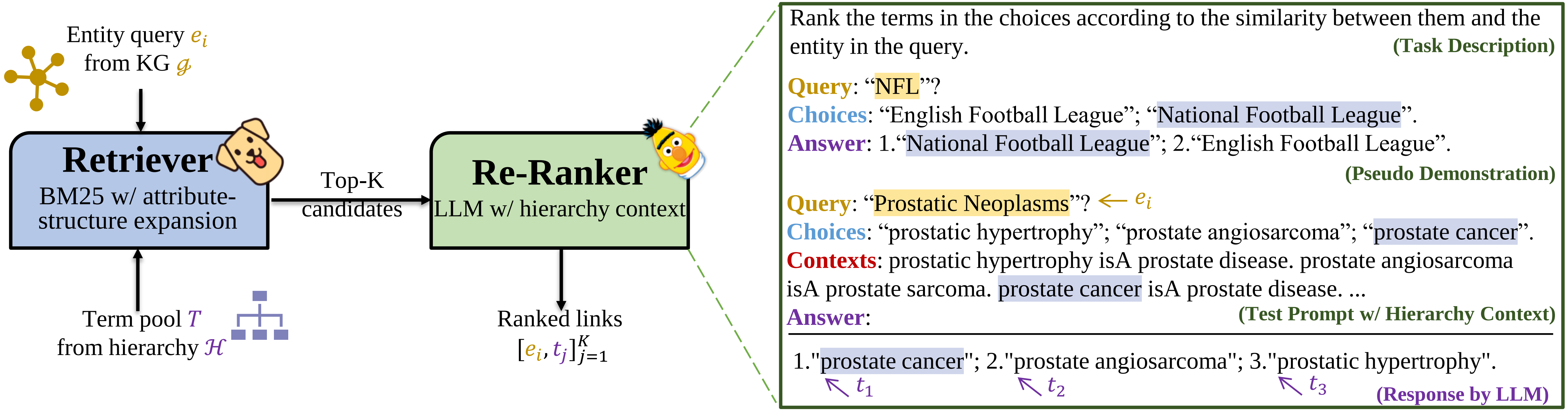
$$LK = \{(e_i, t_j) \mid e_i \in \mathcal{G}, t_j \in \mathcal{H}\}.$$

$M \ll N$ in BKF!

Lu, Jiaying, et al. "Hiprompt: Few-shot biomedical knowledge fusion via hierarchy-oriented prompting." Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023.

HiPrompt

- A few-shot BKF framework via Hierarchy-Oriented Prompting
- We formulate the BKF problem as a ranking problem, and utilize the classic retrieve and re-rank approach
  - unsupervised retriever
  - few-shot re-ranker



Lu, Jiaying, et al. "Hiprompt: Few-shot biomedical knowledge fusion via hierarchy-oriented prompting." Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023.

Benchmark Datasets

Dataset	Source	#Disease	#Entities	#Links
SDKG-DzHi	SDKG	841	19,416	635
	DzHi	11,159	11,159	635
repoDB-DzHi	repoDB	2,074	3,646	709
	DzHi	11,159	11,159	709

Table 2: Statistics of the KG-Hi-BKF benchmark.

Scan to download

Lu, Jiaying, et al. "Hiprompt: Few-shot biomedical knowledge fusion via hierarchy-oriented prompting." Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023.

Main Experimental Results

Setting	Model	SDKG-DzHi						repoDB-DzHi					
		Hits@1	Hits@3	nDCG@1	nDCG@3	WuP	MRR	Hits@1	Hits@3	nDCG@1	nDCG@3	WuP	MRR
Zero-shot	Edit Dist	65.51	70.39	68.08	50.82	85.53	68.69	68.69	71.37	71.71	54.15	85.21	70.71
	BM25	73.07	87.40	77.56	63.01	91.97	81.06	59.38	74.75	70.33	64.51	90.71	68.84
	LogMap	75.75	79.06	76.97	54.82	85.06	77.38	86.60	87.73	87.38	60.79	91.68	87.09
	PARIS	22.68	22.68	23.15	16.13	43.85	22.68	6.35	6.35	6.42	4.44	32.28	6.35
	AML	OOM	OOM	OOM	OOM	OOM	OOM	78.00	78.56	78.67	54.90	86.02	78.26
	SapBERT	69.61	87.24	76.38	63.86	93.78	78.97	75.04	90.69	81.24	73.51	94.25	83.61
	SelfKG	57.95	69.45	58.98	47.29	74.25	64.70	72.78	81.10	75.95	63.78	88.41	77.71
	HiPrompt	**90.79**	**93.08**	**91.57**	**77.00**	**96.74**	**92.13**	**88.01**	**91.26**	**90.70**	**82.85**	**97.06**	**90.64**
One-shot	SapBERT	69.56	87.22	76.34	63.84	93.29	78.93	75.00	90.68	81.21	73.51	94.13	83.59
	MTransE	0.0	0.16	0.0	0.05	35.09	0.16	0.0	0.28	0.14	0.27	28.89	0.37
	HiPrompt	**92.11**	**95.11**	**93.53**	**77.63**	**97.25**	**93.91**	**88.28**	**91.53**	**90.61**	**81.31**	**96.39**	**90.28**

Table 1: Main experiment results (in percentages).

Lu, Jiaying, et al. "Hiprompt: Few-shot biomedical knowledge fusion via hierarchy-oriented prompting." Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023.

Ablation Studies

Expan.	*SDKG-DzHi* Hits@5	Hits@10	Hits@20	*repoDB-DzHi* Hits@5	Hits@10	Hits@20
Name	88.66	89.61	90.55	85.05	88.72	90.27
+Atr.	94.96	96.85	98.11	89.00	92.52	95.20
+Str.	90.08	90.71	91.81	88.15	90.27	92.24
+Atr.+Str.	**96.85**	**97.64**	**98.74**	**91.11**	**93.65**	**95.63**

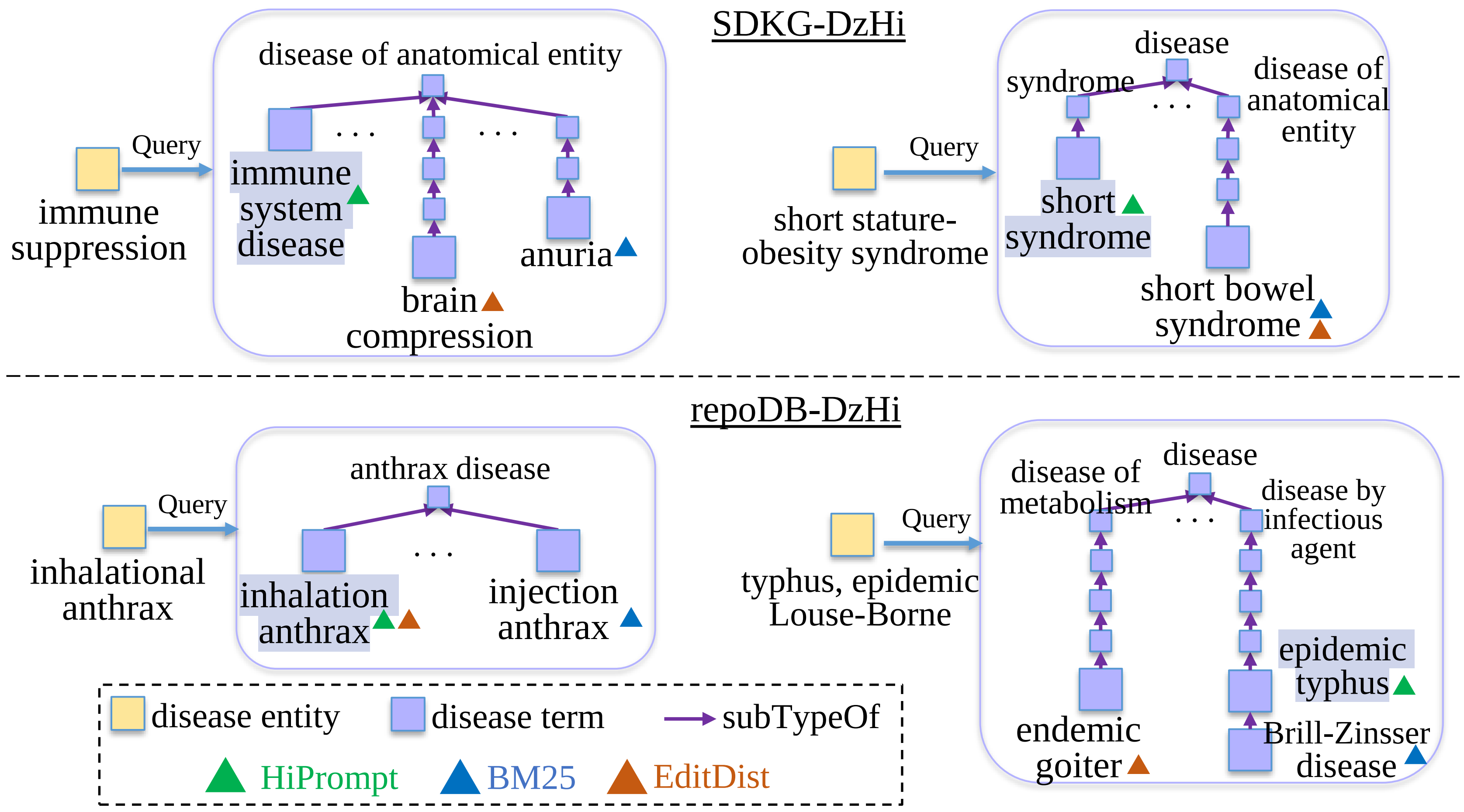
Table 3: Retriever with various expansion strategies.

LLMs	SDKG-DzTaxo Hits@1	Hits@3	MRR	repoDB-DzTaxo Hits@1	Hits@3	MRR
	One-shot (prompt w/o Hi. Context)					
GPT-3	**91.80**	**94.32**	**93.45**	**87.85**	**91.24**	**89.92**
GPT-JT	75.08	86.44	81.80	58.33	69.77	66.42
OPT-6.7B	68.93	80.44	76.38	60.73	73.59	69.33
	One-shot (prompt w/ Hi. Context)					
GPT-3	**92.11**	**95.11**	**93.91**	**88.28**	**91.53**	**90.28**
GPT-JT	80.76	93.69	87.45	69.07	82.91	77.24
OPT-6.7B	72.40	84.86	79.64	63.70	77.68	72.41

Table 4: Re-ranker with various LLMs and prompts.

Lu, Jiaying, et al. "Hiprompt: Few-shot biomedical knowledge fusion via hierarchy-oriented prompting." Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023.

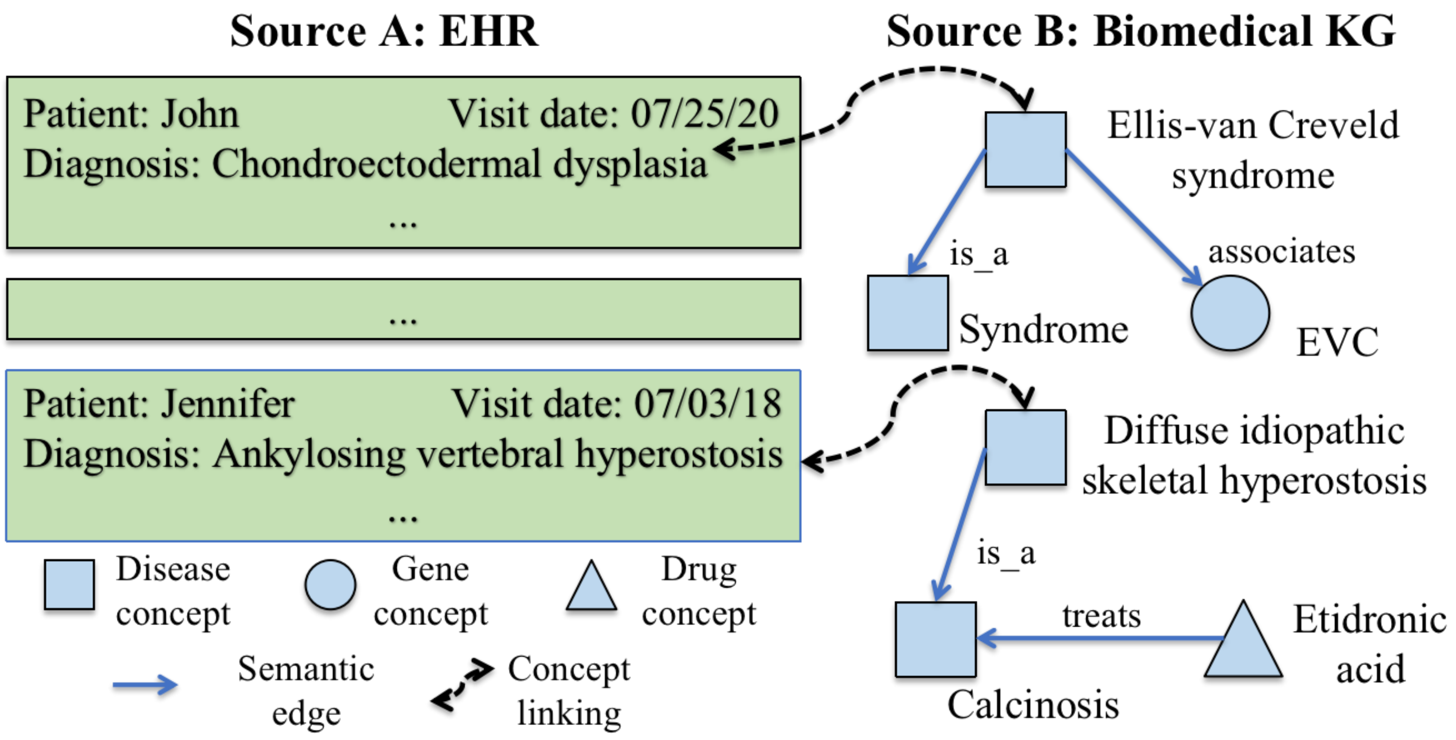
Case Study



Lu, Jiaying, et al. "Hiprompt: Few-shot biomedical knowledge fusion via hierarchy-oriented prompting." Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023.

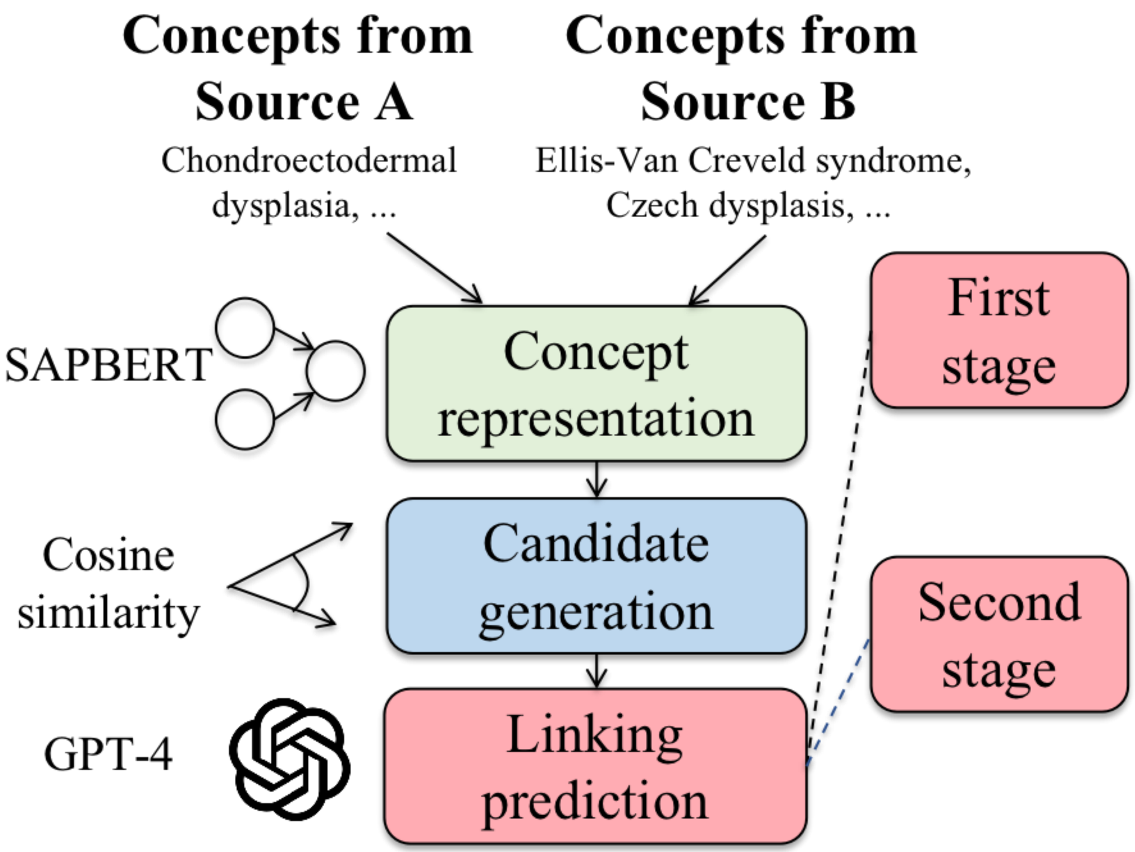
Biomedical Concept Link

- The cross-source biomedical linking task is challenging due to discrepancies in the biomedical naming conventions used in different systems.



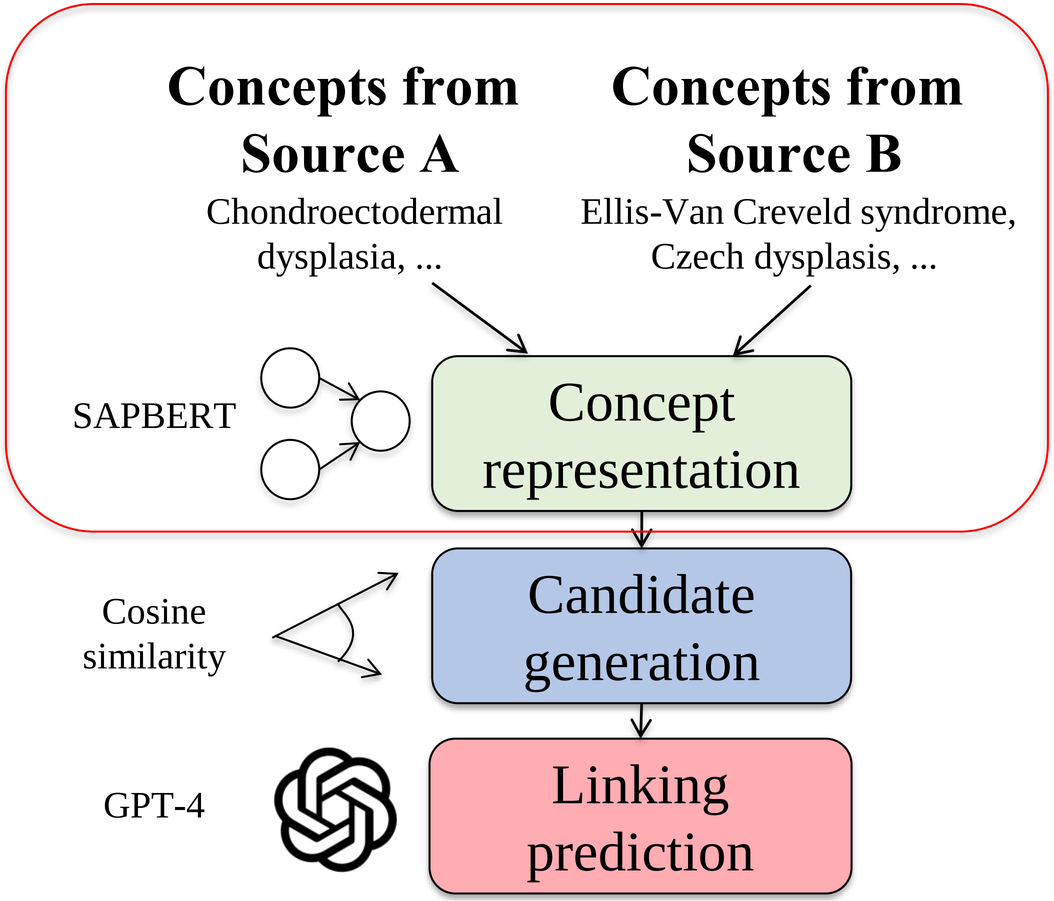
Xie, Yuzhang, et al. "PromptLink: Leveraging large language models for cross-source biomedical concept linking." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

PromptLink



Xie, Yuzhang, et al. "PromptLink: Leveraging large language models for cross-source biomedical concept linking." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

Concept Representation

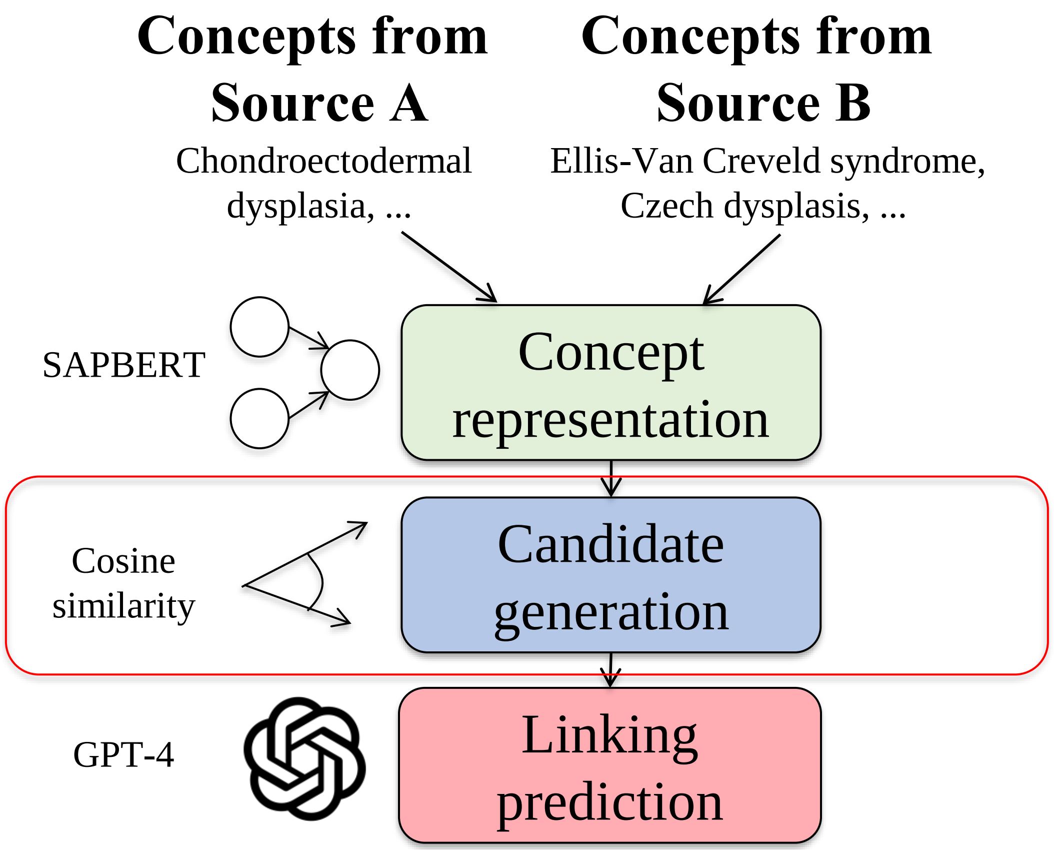


- After pre-processing text by lowercasing and removing punctuation, we use a pre-trained LM (specifically SapBERT), to create embeddings for concepts.
- For concepts that span multiple tokens, the token-level embeddings are averaged to create the concept embedding.

$h_m = PLM(m),\ m$ as EHR concept.

$h_c = PLM(c), c$ as KG concept.

Candidate Generation

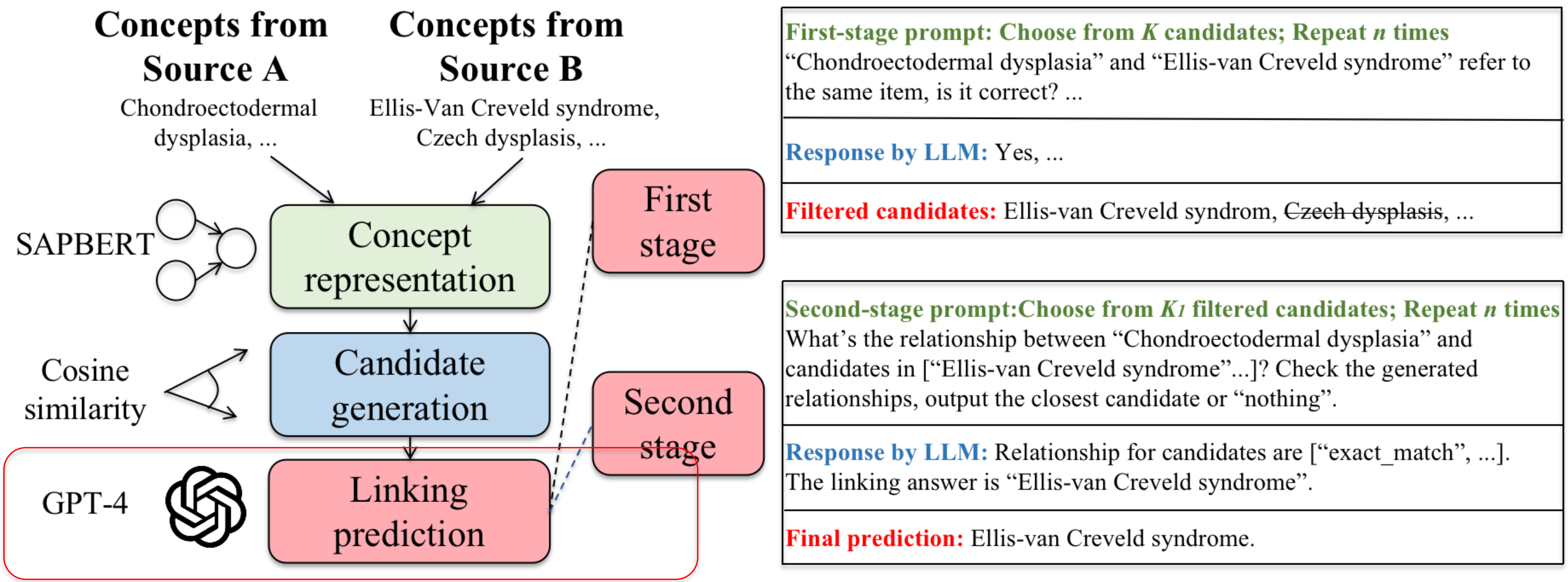


- For candidate generation, we compute cosine similarity S between pairs of EHR concept embedding h_m and KG concept embedding h_c.
- Given each input query EHR concept m, We select the top-K (K=10) KG concepts $[c_1, c_2, \dots, c_K]$ with the highest similarities as candidates for further GPT-based linking prediction.

$$S = cos(h_m, h_c)$$

Xie, Yuzhang, et al. "PromptLink: Leveraging large language models for cross-source biomedical concept linking." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

Link Prediction



Xie, Yuzhang, et al. "PromptLink: Leveraging large language models for cross-source biomedical concept linking." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

Prompt Design

- In the first stage, the LLM is prompted to check if a concept pair (m_i, c_j) should be linked.
- To improve the prompt response quality, we adopt the self-consistency prompting strategy that repeatedly prompts the same question to the LLM multiple (n=5) times, thus obtaining the belief score $B_{i,j}$.

$$B_{i,j} = \frac{number\ of\ \text{“}yes\text{”}}{n}$$

First-stage prompt: Choose from *K* candidates; Repeat *n* times “Chondroectodermal dysplasia” and “Ellis-van Creveld syndrome” refer to the same item, is it correct? ...
Response by LLM: Yes, ...
Filtered candidates: Ellis-van Creveld syndrom, ~~Czech dysplasis~~, ...

Second-stage prompt:Choose from K_1 filtered candidates; Repeat *n* times What’s the relationship between “Chondroectodermal dysplasia” and candidates in [“Ellis-van Creveld syndrome”...]? Check the generated relationships, output the closest candidate or “nothing”.
Response by LLM: Relationship for candidates are [“exact_match”, ...]. The linking answer is “Ellis-van Creveld syndrome”.
Final prediction: Ellis-van Creveld syndrome.

Xie, Yuzhang, et al. "PromptLink: Leveraging large language models for cross-source biomedical concept linking." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

Prompt Design

- Considering the belief scores across different candidates, we derive a comprehensive filter strategy to exclude irrelevant candidates, using parameter τ (set as $0.8 \times n$).
- If $max(B\_(i,1), \ldots, B\_(i,K)) \geq \tau$, this indicates some candidates closely align with the query concept.

First-stage prompt: Choose from *K* candidates; Repeat *n* times “Chondroectodermal dysplasia” and “Ellis-van Creveld syndrome” refer to the same item, is it correct? ...
Response by LLM: Yes, ...
Filtered candidates: Ellis-van Creveld syndrom, ~~Czech dysplasis~~, ...

Second-stage prompt:Choose from K_1 filtered candidates; Repeat *n* times What’s the relationship between “Chondroectodermal dysplasia” and candidates in [“Ellis-van Creveld syndrome”...]? Check the generated relationships, output the closest candidate or “nothing”.
Response by LLM: Relationship for candidates are [“exact_match”, ...]. The linking answer is “Ellis-van Creveld syndrome”.
Final prediction: Ellis-van Creveld syndrome.

Xie, Yuzhang, et al. "PromptLink: Leveraging large language models for cross-source biomedical concept linking." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

Prompt Design

- In the second stage, the LLM evaluates the candidates $[c_1, c_2, \ldots, c_{K_1}]$ retained from the first stage's filtering process.
- In this stage, we also use the self-consistency strategy that prompts one question for the same $n = 5$ times.

First-stage prompt: Choose from *K* candidates; Repeat *n* times
"Chondroectodermal dysplasia" and "Ellis-van Creveld syndrome" refer to the same item, is it correct? ...

Response by LLM: Yes, ...

Filtered candidates: Ellis-van Creveld syndrom, ~~Czech dysplasis~~, ...

Second-stage prompt:Choose from *K1* filtered candidates; Repeat *n* times
What's the relationship between "Chondroectodermal dysplasia" and candidates in ["Ellis-van Creveld syndrome"...]? Check the generated relationships, output the closest candidate or "nothing".

Response by LLM: Relationship for candidates are ["exact_match", ...]. The linking answer is "Ellis-van Creveld syndrome".

Final prediction: Ellis-van Creveld syndrome.

Xie, Yuzhang, et al. "PromptLink: Leveraging large language models for cross-source biomedical concept linking." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

Prompt Design

- We calculate the occurrence frequency $f_{i,j}$ for answers in $[c_1, c_2, \ldots, c_{K_1}] \cup [NIL]$ and retrieve the final linking result for query EHR concept mi.
- If $f_{i,NIL}$,*0.5* < this indicates a high probability that none of the candidates are appropriate.
- Otherwise, the candidate c_j with the highest frequency $f\_(i,j)$ is decided as the final linking result

First-stage prompt: Choose from *K* candidates; Repeat *n* times
"Chondroectodermal dysplasia" and "Ellis-van Creveld syndrome" refer to the same item, is it correct? ...

Response by LLM: Yes, ...

Filtered candidates: Ellis-van Creveld syndrom, ~~Czech dysplasis~~, ...

Second-stage prompt:Choose from *K₁* filtered candidates; Repeat *n* times
What's the relationship between "Chondroectodermal dysplasia" and candidates in ["Ellis-van Creveld syndrome"...]? Check the generated relationships, output the closest candidate or "nothing".

Response by LLM: Relationship for candidates are ["exact_match", ...]. The linking answer is "Ellis-van Creveld syndrome".

Final prediction: Ellis-van Creveld syndrome.

Xie, Yuzhang, et al. "PromptLink: Leveraging large language models for cross-source biomedical concept linking." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

Concept Linking Experiment Results

Table 1: Comparison of the zero-shot accuracy for different methods on MIID and CISE.

Method	Acc-MIID	Acc-CISE
Cosine Distance	0.2981	0.2907
Jaccard Distance	0.2123	0.3280
Levenshtein Distance	0.1995	0.3033
Jaro-Winkler Distance	0.3141	0.3693
BM25	0.4722	0.3993
BioBERT	0.3423	0.5280
BioClinicalBERT	0.3007	0.5007
BioGPT	0.3530	0.5093
BioDistilBERT	0.4240	0.5293
KrissBERT	0.5265	0.5787
ada002	0.5968	0.6773
SAPBERT	0.7213	0.8167
PromptLink	**0.7756**	**0.8880**

- PromptLink outperforms competing approaches across both datasets in terms of zero-shot accuracy, underscoring the superiority of our LLM-based concept linking methodology.

Xie, Yuzhang, et al. "PromptLink: Leveraging large language models for cross-source biomedical concept linking." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

Case Studies

- Three scenarios are presented: (1) concepts assessed by both ground-truth labels and a clinician; (2) concepts evaluated by a clinician due to missing ground-truth labels; (3) irrelevant concepts judged by a clinician. Overall, PromptLink could link biomedical concepts more accurately and appropriately.

Table 3: Analyzed cases.

ID	EHR Concept	PromptLink's Prediction	SAPBERT's Prediction
I	Chondroectodermal dysplasia	Ellis-van Creveld syndrome 😃🏷️	Cranioectodermal dysplasia
II	Dermatophytosis of hand	Tinea manuum 😃🏷️	Hand dermatosis
III	Late syphilis, unspecified	Tertiary syphilis 😃🏷️	Secondary syphilis
IV	Hypopotassemia	Hypokalemia 😃🏷️	Hypocupremia nos
V	Epidemic vertigo	Vestibular neuronitis 😃🏷️	Vertigo
VI	Postprocedural fever	Postoperative complications 😃	Postcardiotomy syndrome
VII	Acquired cardiac septal defect	Heart septal defect 😃	Atrial heart septal defect
VIII	Height of bed	NIL 😃	Binge eating disorder
IX	Level one	NIL 😃	Glaucoma 1 open angle

Note: "😃" indicates this prediction is justified by the clinician. "🏷️" indicates this prediction is justified by the ground-truth label.

Xie, Yuzhang, et al. "PromptLink: Leveraging large language models for cross-source biomedical concept linking." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

Constructing and Completing KGs

- KGs require both quality and coverage
  - KGs can include erroneous and inconsistent knowledge
  - KGs are “small”
  - New knowledge is constantly generated, making existing knowledge inaccurate and incomplete

Traditional Text-based Prompting

Zhu, Yuqi, et al. "Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities." *World Wide Web* 27.5 (2024): 58.

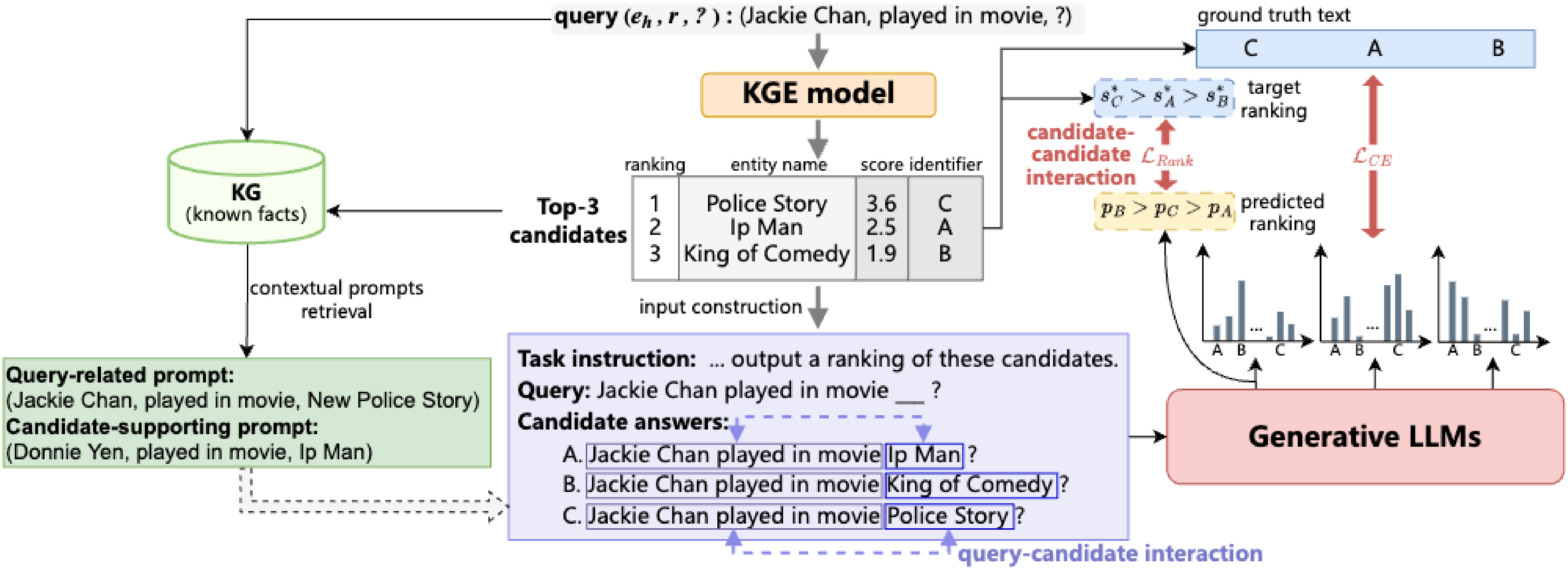
Traditional Text-based Prompting

Model	Knowledge Graph Construction				Knowledge Graph Reasoning			
	DuIE2.0	Re-TACRED	SciERC	MAVEN	FB15K-237	ATOMIC2020	FreebaseQA	MetaQA
Fine-Tuned SOTA	69.42	91.4	53.2	68.8	32.4	46.9	79.0	100
Zero-shot								
text-davinci-003	11.43	9.8	4.0	30.0	16.0	15.1	95.0	33.9
ChatGPT	10.26	15.2	4.4	26.5	24.0	10.6	95.0	52.7
GPT-4	31.03	15.5	7.2	34.2	32.0	16.3	95.0	63.8
One-shot								
text-davinci-003	30.63	12.8	4.8	25.0	32.0	14.1	95.0	49.5
ChatGPT	25.86	14.2	5.3	34.1	32.0	11.1	95.0	50.0
GPT-4	41.91	22.5	9.1	30.4	40.0	19.1	95.0	56.0

Zhu, Yuqi, et al. "Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities." *World Wide Web* 27.5 (2024): 58.

Traditional Text-based Prompting

- KC-GenRe re-ranks Top-3 candidates predicted by the first-stage KGE model through LLMs for a given query $(e_h, r, ?)$.



Wang, Yilin, et al. "KC-GenRe: A knowledge-constrained generative re-ranking method based on large language models for knowledge graph completion." *arXiv preprint arXiv:2403.17532* (2024).

Traditional Text-based Prompting

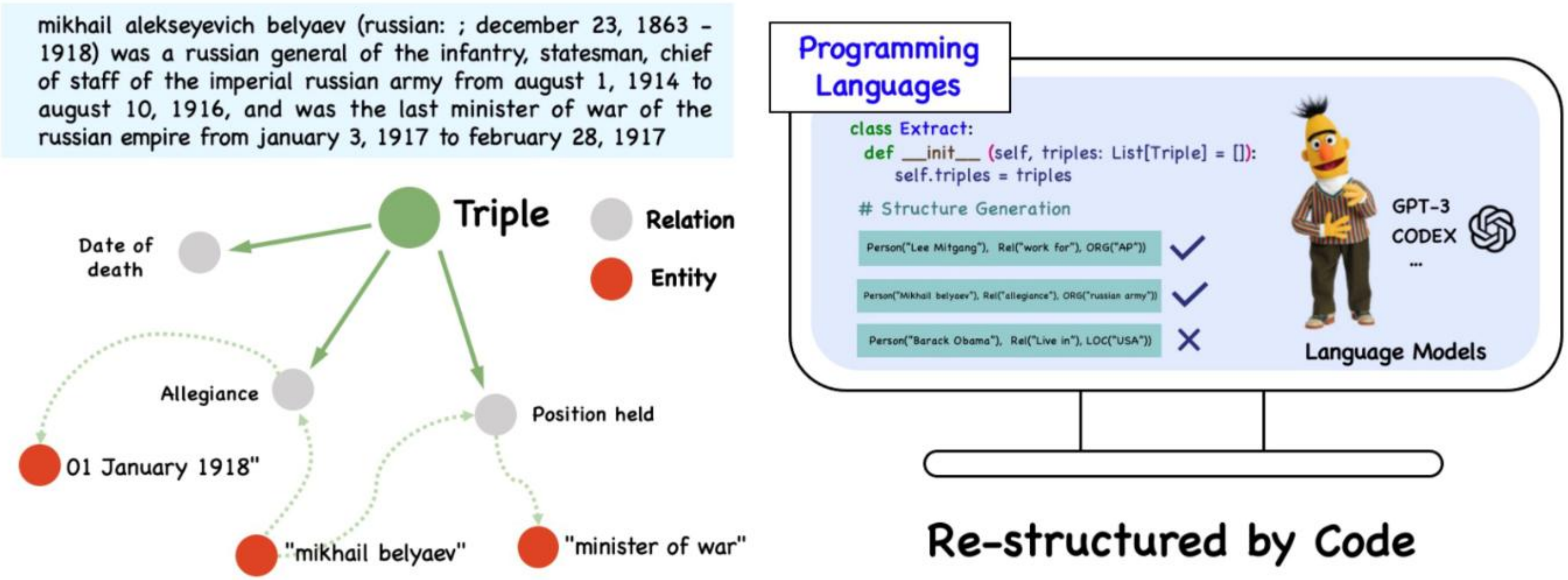
Model	Wiki27K				FB15K-237-N			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
TransE† (Bordes et al., 2013)	0.155	0.032	0.228	0.378	0.255	0.152	0.301	0.459
TransC† (Lv et al., 2018)	0.175	0.124	0.215	0.339	0.233	0.129	0.298	0.395
ConvE† (Dettmers et al., 2018)	0.226	0.164	0.244	0.354	0.273	0.192	0.305	0.429
WWV† (Veira et al., 2019)	0.198	0.157	0.237	0.365	0.269	0.137	0.287	0.443
TuckER (Balazevic et al., 2019)	0.249	0.185	0.269	0.385	0.309	0.227	0.340	0.474
RotatE† (Sun et al., 2019)	0.216	0.123	0.256	0.394	0.279	0.177	0.320	0.481
KG-BERT† (Yao et al., 2019)	0.192	0.119	0.219	0.352	0.203	0.139	0.201	0.403
LP-RP-RR† (Kim et al., 2020)	0.217	0.138	0.235	0.379	0.248	0.155	0.256	0.436
PKGC† (Lv et al., 2022)	0.285	0.230	0.305	**0.409**	0.332	0.261	0.346	0.487
KC-GenRe	**0.317**	**0.274**	**0.330**	0.408	**0.399**	**0.338**	**0.427**	**0.505**

Model	ReVerb20K					ReVerb45K				
	MRR	MR	Hits@1	Hits@3	Hits@10	MRR	MR	Hits@1	Hits@3	Hits@10
TransE (Bordes et al., 2013)	0.138	1150.5	0.034	0.201	0.316	0.202	1889.5	0.122	0.243	0.346
ComplEx (Trouillon et al., 2016)	0.038	4486.5	0.017	0.043	0.071	0.068	5659.8	0.054	0.071	0.093
R-GCN (Schlichtkrull et al., 2018)	0.122	1204.3	-	-	0.187	0.042	2866.8	-	-	0.046
ConvE (Dettmers et al., 2018)	0.262	1483.7	0.203	0.287	0.371	0.218	3306.8	0.166	0.243	0.314
KG-BERT (Yao et al., 2019)	0.047	420.4	0.014	0.039	0.105	0.123	1325.8	0.070	0.131	0.223
RotatE (Sun et al., 2019)	0.065	2861.5	0.043	0.069	0.108	0.141	3033.4	0.110	0.147	0.196
PairRE (Chao et al., 2021)	0.213	1366.2	0.166	0.229	0.296	0.205	2608.4	0.153	0.228	0.302
ResNet (Lovelace et al., 2021)	0.224	2258.4	0.188	0.240	0.292	0.181	3928.9	0.150	0.196	0.242
BertResNet-ReRank (Lovelace et al., 2021)	0.272	1245.6	0.225	0.294	0.347	0.208	2773.4	0.166	0.227	0.281
CaRe (Gupta et al., 2019)	0.318	973.2	-	-	0.439	0.324	1308.0	-	-	0.456
OKGIT (Chandrahas and Talukdar, 2021)	0.359	527.1	0.282	0.394	0.499	0.332	773.9	0.261	0.363	0.464
OKGSE (Xie et al., 2022a)	0.372	487.3	0.291	0.408	0.524	0.342	**771.1**	0.274	0.371	0.473
CEKFA (Wang et al., 2023b)	0.387	416.7	0.310	0.427	0.515	0.369	884.5	0.294	0.409	0.502
KC-GenRe	**0.408**	**410.8**	**0.331**	**0.450**	**0.547**	**0.404**	874.1	**0.332**	**0.444**	**0.534**

Wang, Yilin, et al. "KC-GenRe: A knowledge-constrained generative re-ranking method based on large language models for knowledge graph completion." *arXiv preprint arXiv:2403.17532* (2024).

Code-based Instructions

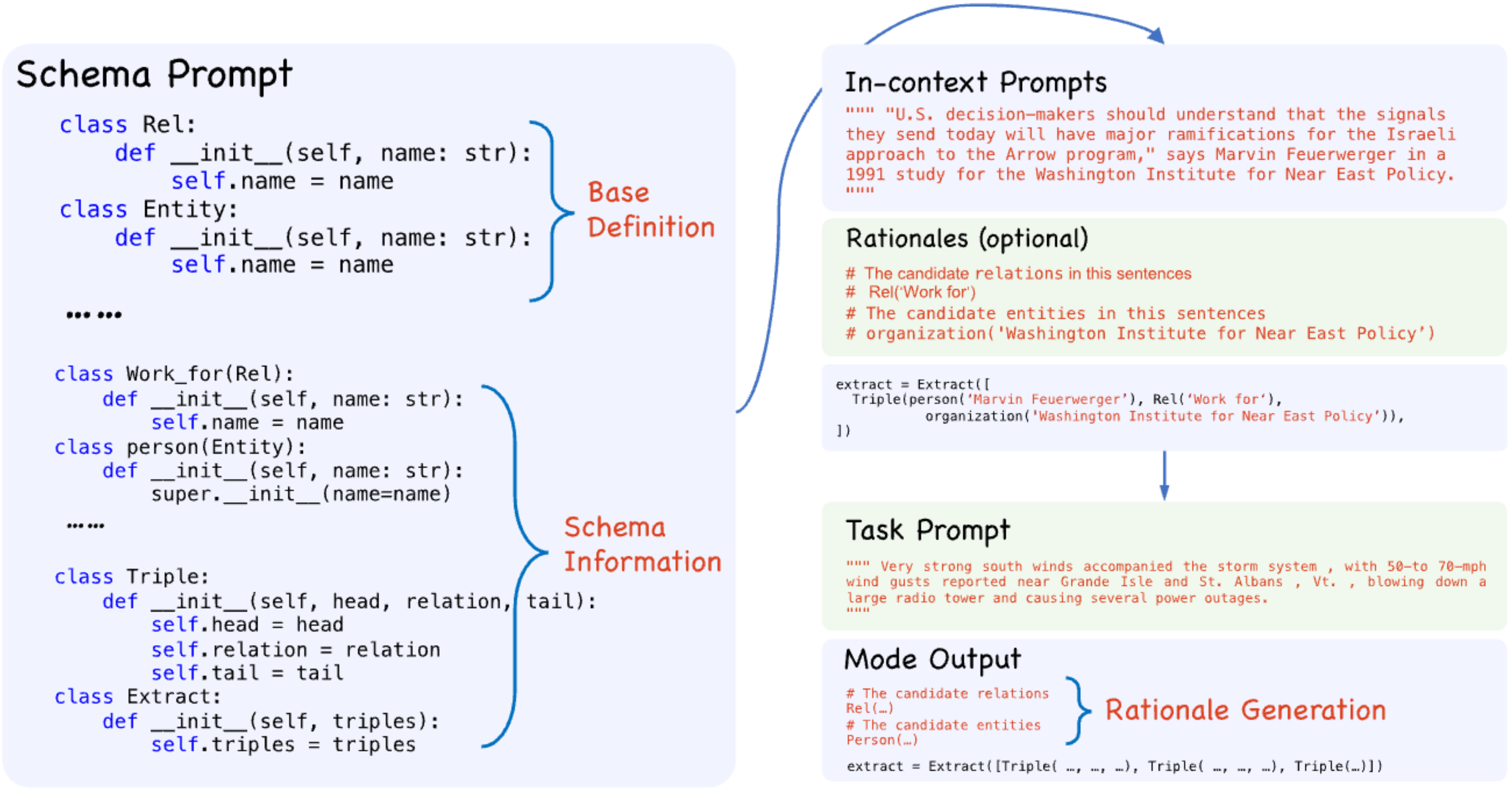
- Code LLMs, designed for processing structured data like programming code, naturally align with the hierarchical and relational nature of KGs



Bi, Zhen, et al. "Codekgc: Code language model for generative knowledge graph construction." *ACM Transactions on Asian and Low-Resource Language Information Processing* 23.3 (2024): 1-16.

Code-based Instructions

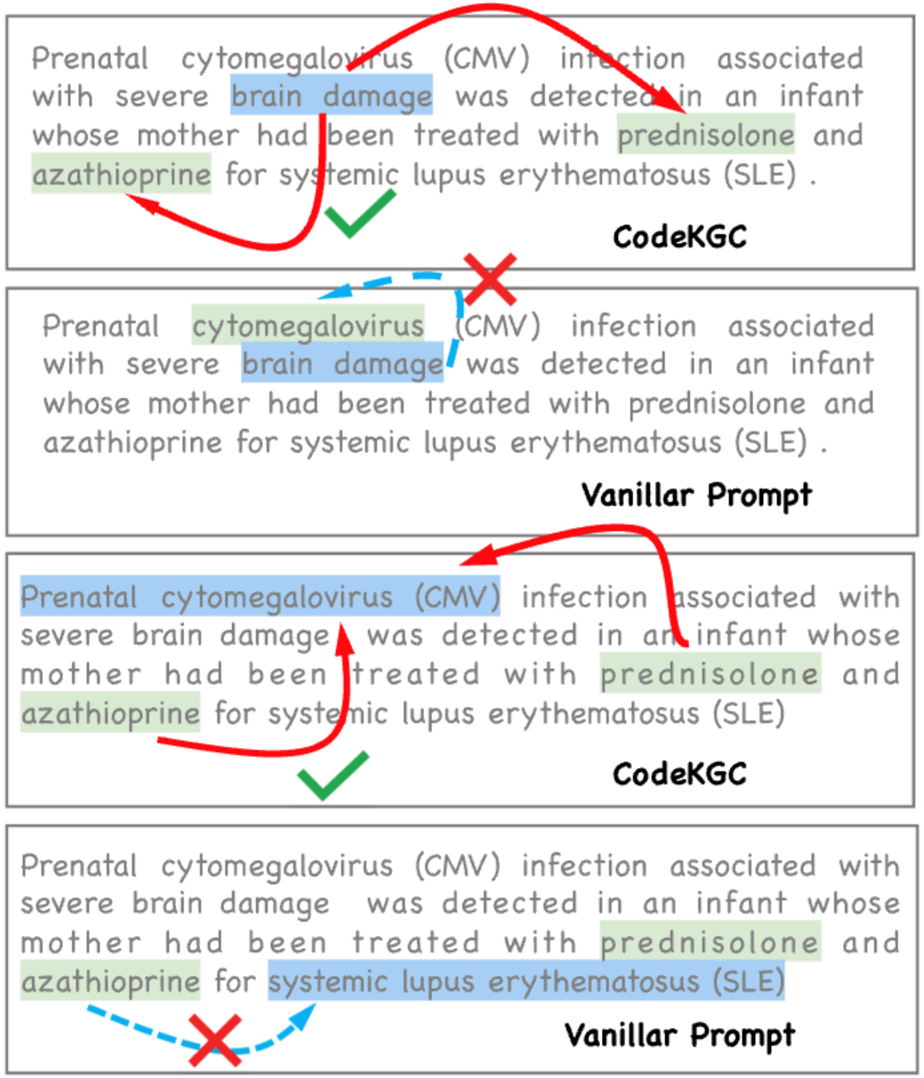
- The original natural language is converted into code formats and then fed into the code LM which is guided by a specified task prompt. They use schema-aware prompt to preserve the relations, properties, and constraints in the knowledge graph.



Bi, Zhen, et al. "Codekgc: Code language model for generative knowledge graph construction." *ACM Transactions on Asian and Low-Resource Language Information Processing* 23.3 (2024): 1-16.

Code-based Instructions

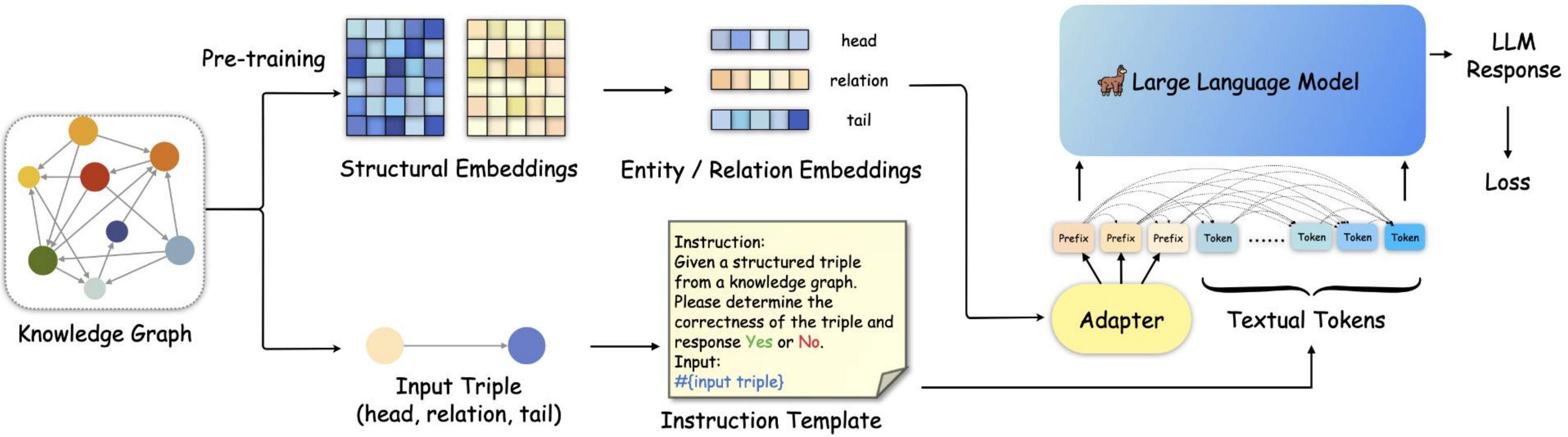
	Comparable SOTA	Dataset ADE	Dataset CONLL04	Dataset SciERC
Zero-Shot	UIE [16]	24.3	16.1	10.3
	Vanilla Prompt (text-davinci-002)	41.2	18.4	12.2
	Vanilla Prompt (text-davinci-003)	41.7	30.5	18.1
	CodeKGC (text-davinci-002)	42.5 (↑1.3)	35.8 (↑17.4)	15.0 (↑2.8)
	CodeKGC (text-davinci-003)	**43.7** (↑2.0)	**41.6** (↑11.1)	**19.5** (↑1.4)
Few-Shot	UIE [16]	50.3	39.0	19.2
	Vanilla Prompt (text-davinci-002)	45.7	28.2	14.1
	Vanilla Prompt (text-davinci-003)	58.8	43.2	18.8
	CodeKGC (text-davinci-002)	61.5 (↑15.8)	42.7 (↑14.5)	18.5 (↑4.4)
	CodeKGC (text-davinci-003)	**64.2** (↑5.4)	**49.6** (↑6.4)	**24.7** (↑5.9)



Bi, Zhen, et al. "Codekgc: Code language model for generative knowledge graph construction." *ACM Transactions on Asian and Low-Resource Language Information Processing* 23.3 (2024): 1-16.

LLM Fine-tuning for KG Completion

- The knowledge prefix adapter (KoPA) model first pre-trains structural embeddings for the entities and relations in the given KG and then instruction fine-tune the LLM.
- The structural embeddings of the given input triple will be projected into the textual space of the LLM by the adapter and serve as prefix tokens in the front of the input sequence.



Zhang, Yichi, et al. "Making large language models perform better in knowledge graph completion." *Proceedings of the 32nd ACM International Conference on Multimedia*. 2024.

LLM Fine-tuning for KG Completion

Paradigm	Model	UMLS				CoDeX-S				FB15K-237N			
		Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
Embedding-based	TransE [3]	84.49	86.53	81.69	84.04	72.07	71.91	72.42	72.17	69.71	70.80	67.11	68.91
	DistMult [38]	86.38	87.06	86.53	86.79	66.79	69.67	59.46	64.16	58.66	58.98	56.84	57.90
	ComplEx [34]	90.77	89.92	91.83	90.87	67.64	67.84	67.06	67.45	65.70	66.46	63.38	64.88
	RotatE [31]	92.05	90.17	94.41	92.23	75.68	75.66	75.71	75.69	68.46	69.24	66.41	67.80
PLM-based	KG-BERT [40]	77.30	70.96	92.43	80.28	77.30	70.96	92.43	80.28	56.02	53.47	97.62	67.84
	PKGC [23]	-	-	-	-	-	-	-	-	79.60	-	-	79.50
LLM-based Training-free	Zero-shot(Alpaca)	52.64	51.55	87.69	64.91	50.62	50.31	99.83	66.91	56.06	53.32	97.37	68.91
	Zero-shot(GPT-3.5)	67.58	88.04	40.71	55.67	54.68	69.13	16.94	27.21	60.15	86.62	24.01	37.59
	ICL(1-shot)	50.37	50.25	75.34	60.29	49.86	49.86	50.59	50.17	54.54	53.67	66.35	59.34
	ICL(2-shot)	53.78	52.47	80.18	63.43	52.95	51.54	98.85	67.75	57.81	56.22	70.56	62.58
	ICL(4-shot)	53.18	52.26	73.22	60.99	51.14	50.58	99.83	67.14	59.29	57.49	71.37	63.68
	ICL(8-shot)	55.52	55.85	52.65	54.21	50.62	50.31	99.83	66.91	59.23	57.23	73.02	64.17
LLM-based Fine-tuning	KG-LLaMA [41]	85.77	87.84	83.05	85.38	79.43	78.67	80.74	79.69	74.81	67.37	96.23	79.25
	KG-Alpaca [41]	86.01	94.91	76.10	84.46	80.25	79.38	81.73	80.54	69.91	62.71	98.28	76.56
	Vanilla IT	86.91	95.18	77.76	85.59	81.18	77.01	88.89	82.52	73.50	65.87	97.53	78.63
	Structure-aware IT	89.93	93.27	86.08	89.54	81.27	77.14	88.40	82.58	76.42	69.56	93.95	79.94
KoPA		**92.58**	90.85	94.70	**92.70**	**82.74**	77.91	91.41	**84.11**	77.65	70.81	94.09	**80.81**

Zhang, Yichi, et al. "Making large language models perform better in knowledge graph completion." *Proceedings of the 32nd ACM International Conference on Multimedia*. 2024.

Enriching KGs with multi-modality data

- Besides textual KGs and online literature, the world is multi-modality and knowledge should be multi-modality as well
- Aligning general multi-modality foundation models (MMFMs) to real domain-specific data (e.g., medical data) is challenging due to the lack of high-quality fine-grained pairs of *X-and-text* labeled data such as for instruction tuning

Visual Knowledge Extraction

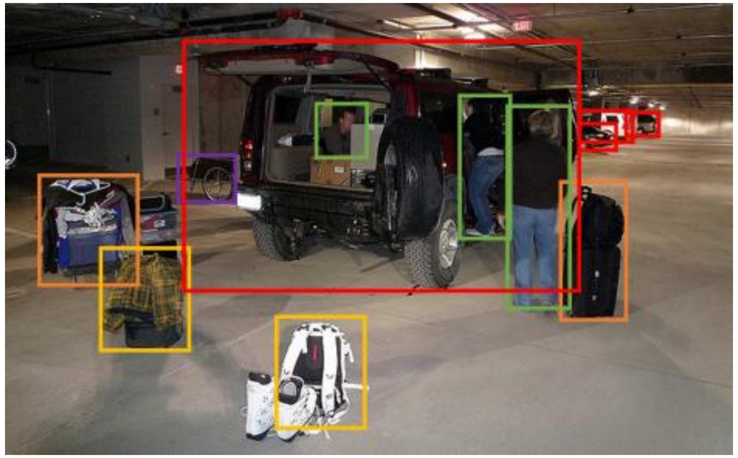
- Images contain rich fine-grained knowledge that complements the textual knowledge documented in literature
- Existing method on visual knowledge extraction reply on pre-defined formats or vocabularies, restricting the expressiveness of the extracted knowledge
- We aim to explore a new paradigm of open visual knowledge extraction

	Limitations of Existing Approaches	Our Target
Format	Restricted by a fixed knowledge format (e.g., sub-verb-obj tuples)	Format-free knowledge
Vocabulary	Limited by predefined sets of objects or relations	Open-world w/o predefined set
Language	Produced knowledge is often limited in language richness to capture fine-grained information	Reflect real-word diverse language variety and capture nuanced details

Cui, Hejie, et al. "Open visual knowledge extraction via relation-oriented multimodality model prompting." *Advances in neural information processing systems* 36 (2023): 23499-23519.

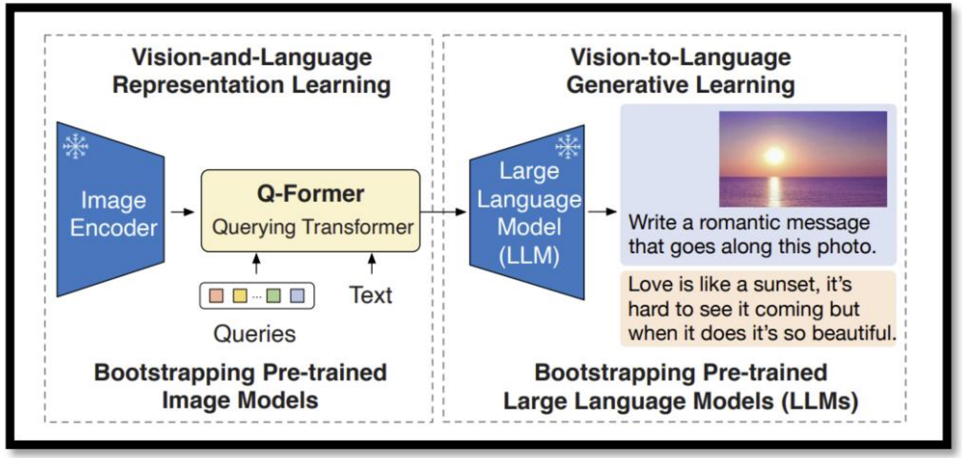
OpenViK: A New Paradigm

- Leverage pre-trained large multi-modality models by eliciting open visual knowledge through relation-oriented visual prompting



Input Data

Visually Grounded

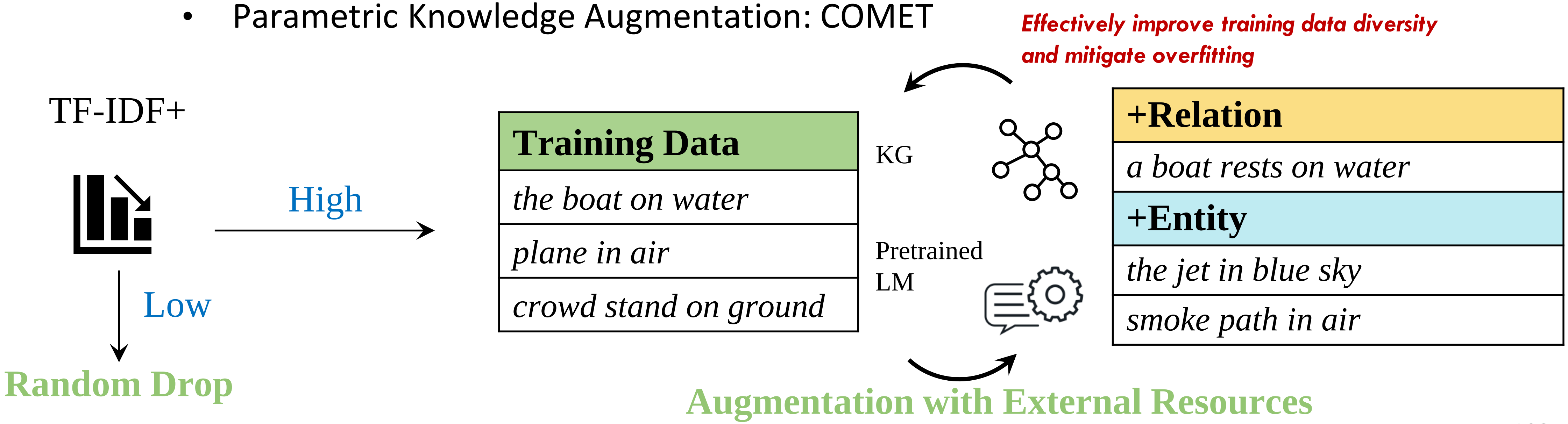


Pre-Trained Large Vision-Language Model

Cui, Hejie, et al. "Open visual knowledge extraction via relation-oriented multimodality model prompting." *Advances in neural information processing systems* 36 (2023): 23499-23519.

Diversity-Driven Data Enhancement

- Challenge: long-tail distribution biased to more prevalent relations and entities
- Two strategies based on an adapted TF-IDF+ score: $\mathcal{S}_r = (\log(\frac{N}{1 + f_r * \alpha_1}))^{\alpha_2}$
  - Random dropping on low-quality data
  - Data augmentation with external knowledge resources
    - Non-parametric Knowledge Augmentation: ConceptNet
    - Parametric Knowledge Augmentation: COMET

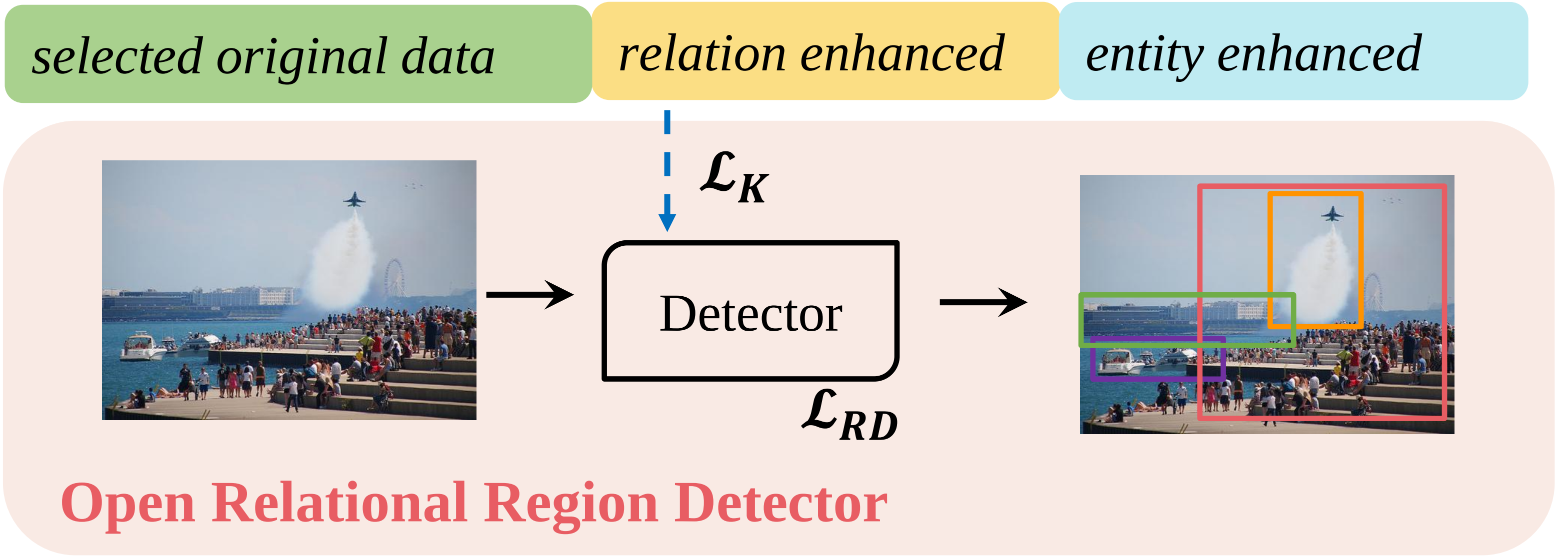


Cui, Hejie, et al. "Open visual knowledge extraction via relation-oriented multimodality model prompting." *Advances in neural information processing systems* 36 (2023): 23499-23519.

Open Relational Region Detector

- Unique challenge: detecting regions potentially containing relational knowledge
- Two adaptations on FasterRCNN:
  - Region regression $\mathcal{L}_{\mathrm{RD}}$: object-centric region → higher-order knowledge-centric regions
  - Knowledge supervision $\mathcal{L}_{\mathrm{K}}$: replace object-centric label classification with regional knowledge supervision
- Training objective:

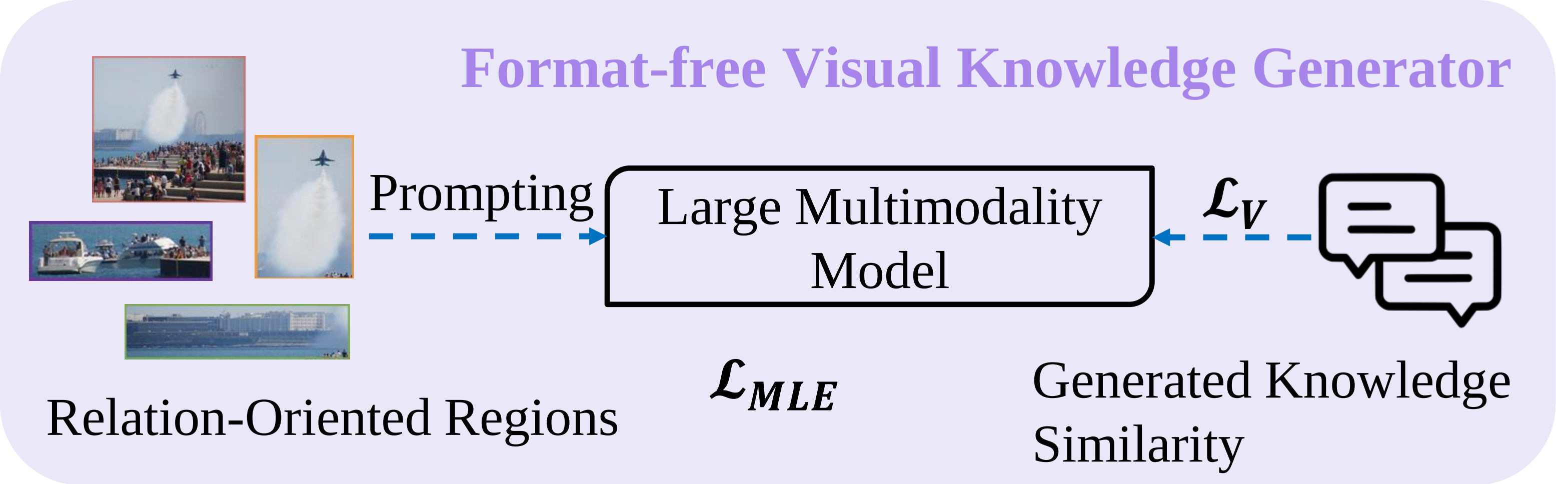
$$\mathcal{L}_v = \mathcal{L}_{\mathrm{RD}} + \mathcal{L}_{\mathrm{K}}$$



Cui, Hejie, et al. "Open visual knowledge extraction via relation-oriented multimodality model prompting." *Advances in neural information processing systems* 36 (2023): 23499-23519.

Format-free Visual Knowledge Generator

- Conditioning the generator on the detected relational region for better knowledge grounding
- Architecture: pre-trained vision transformer ViT-B + image-grounded text decoder of BLIP
- Decoder leverages the detected regional mask as a binary visual prompt
- Language model loss: $\mathcal{L}_{\text{MLE}}$
- Penalty term $\mathcal{L}_V$ to improve information variety $\mathcal{L}_{\text{V}} = \frac{1}{N_i} \sum_{N_i} \text{ReLU}\left(-\log\left(1 - \left(s\left(T_a, T_b\right) - \phi\right)\right)\right)$



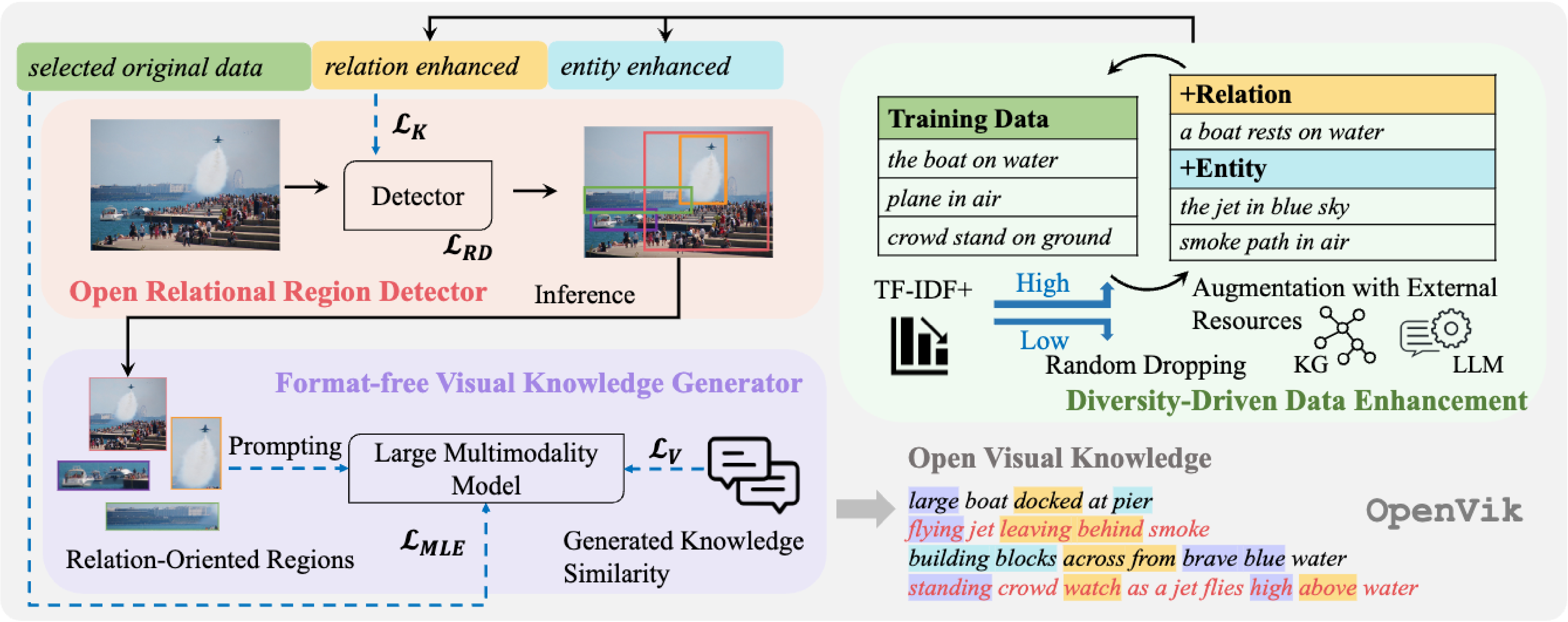
Training objective:

$$\mathcal{L}_l = \alpha \times \mathcal{L}_{\text{MLE}} + (1 - \alpha) \times \mathcal{L}_{\text{V}}$$

Cui, Hejie, et al. "Open visual knowledge extraction via relation-oriented multimodality model prompting." *Advances in neural information processing systems* 36 (2023): 23499-23519.

OpenVik Framework Overview

- **OpenVik is designed to extract format-free open visual knowledge with novel entities, diverse relations, and nuanced descriptive details**



Cui, Hejie, et al. "Open visual knowledge extraction via relation-oriented multimodality model prompting." *Advances in neural information processing systems* 36 (2023): 23499-23519.

Evaluation on Generated Knowledge

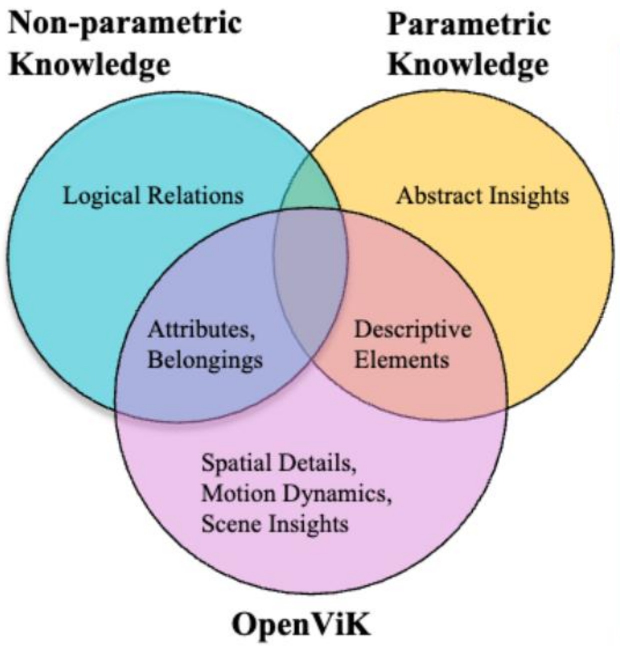
- Generation Performance
  - Scene graph generation
  - Relational captioning
  - Region captioning
  - Language generation metrics: BLEU, ROUGE, METEOR
- In-Depth Knowledge Quality
  - Validity
  - Conformity
  - Freshness
  - Diversity

Method	Generation Performance			In-Depth Knowledge Quality			
	BLEU↑	ROUGE-L↑	METEOR↑	Validity↑	Conformity↑	Freshness↑	Diversity↑
Closed/Open Scene Graph Generation							
IMP [52]	0.075	0.123	0.118	0.800	0.823	0.676	0.316
Neural Motifs [63]	0.229	<u>0.283</u>	**0.273**	0.822	0.767	0.667	0.349
UnbiasSGG [44]	0.217	0.258	0.194	0.739	0.733	0.666	0.357
Ov-SGG [17]	0.167	0.210	0.183	0.712	0.633	0.693	0.413
Dense Relational Captioning							
MTTSNet+REM [22]	0.240	0.226	0.228	<u>0.897</u>	0.852	0.754	0.375
Region Captioning							
DenseCap [20]	0.248	0.245	0.196	0.883	0.843	0.790	0.543
Sub-GC [65]	0.272	0.263	0.221	0.892	<u>0.871</u>	<u>0.795</u>	<u>0.547</u>
BLIP [27]	0.264	0.266	0.252	0.886	0.855	0.760	0.531
BLIP2 [26]	0.275	**0.285**	0.257	0.892	0.871	0.766	0.535
Open Visual Knowledge Extraction							
OpenVik	**0.280**	<u>0.283</u>	0.250	**0.907**	**0.883**	**0.809**	**0.619**

Cui, Hejie, et al. "Open visual knowledge extraction via relation-oriented multimodality model prompting." *Advances in neural information processing systems* 36 (2023): 23499-23519.

Comparison with Existing Knowledge Sources

- Non-parametric knowledge in knowledge graphs
- Parametric knowledge in pre-trained language models

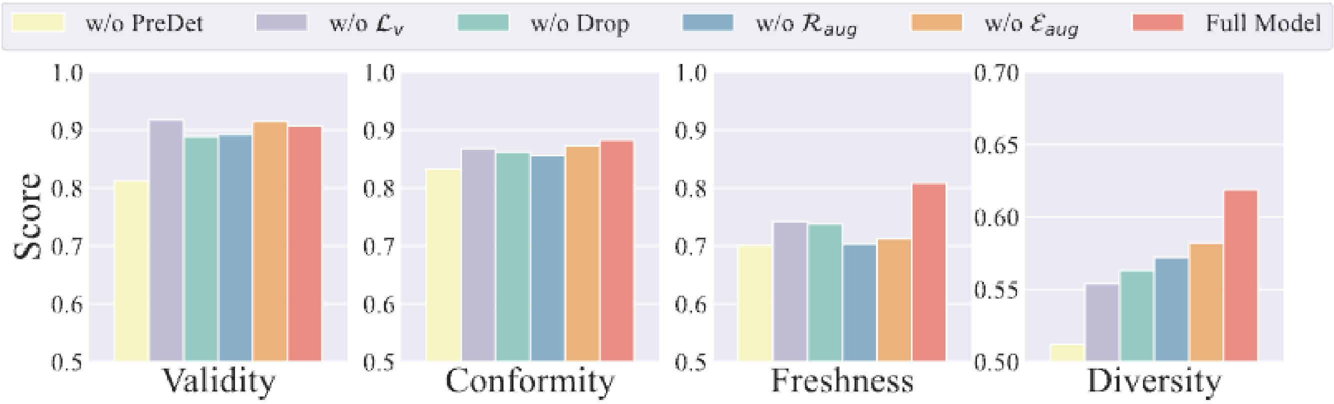


Knowledge Source	Examples
Non-parametric Knowledge	dog IsA animal (ConceptNet); conceptnet IsA knowledge graph (ConceptNet);
	dog HasProperty black (ConceptNet); dog and brown fur covering black (OpenVik)
	computer HasA keyboard (ConceptNet); keyboard with computer (OpenVik)
Parametric Knowledge	people using light bulbs to illuminate the room; (LLM)
	yellow sign in corner (both); black seat attached to bike (both);
Open Visual Knowledge (OpenViK)	three layer cake on table; blue trash can full of garbage next to brown dresser; blue box sitting beside a sneaky garage; (OpenVik)
	people wearing fashionable black hats are skiing; baby elephants walking around adventurous wood; (OpenVik)
	the light shining from bright black background; hanging fan are above tall shelf; brown chair in the background of the room; (OpenVik)

Cui, Hejie, et al. "Open visual knowledge extraction via relation-oriented multimodality model prompting." *Advances in neural information processing systems* 36 (2023): 23499-23519.

Ablation Study

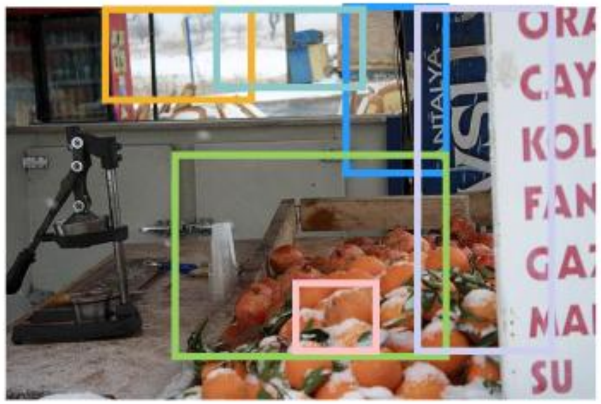
- Influence of information variety regularization and diversity-drive data enhancement
- Influence of pre-training for the open relational region detector
- Influence of data enhancement strategies for training dataset diversity



Metrics	Training Dataset			Generate Knowledge
	Visual Genome [24]	*Relational Caps [22]*	*Diversity Enhanced (Ours)*	OpenVik *(Ours)*
Diversity	0.589	0.604	0.632	0.619

Cui, Hejie, et al. "Open visual knowledge extraction via relation-oriented multimodality model prompting." *Advances in neural information processing systems* 36 (2023): 23499-23519.

Case Studies

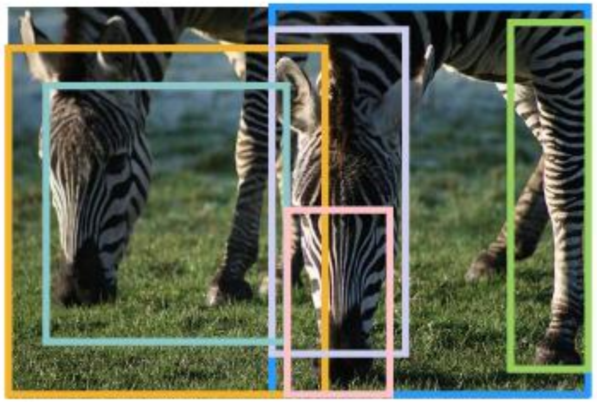


OpenVik:
- blue post attached to wall with white letter
- the open window to snowy ground
- wood box full of different size of orange
- white banner on a building with letter o
- blue box sitting beside a sneaky garage
- a orange covered with ice and green leaves

Visual Genome-Scene Graph:
<drink, in, cooler>
<orange, in, box>
<banner, on, building>
<item, on, table>

Visual Genome-Region Descriptions:
oranges in a wood thing
green leaves on oranges
red writing on a white sign
drink in red cooler

Relational Caps:
snow-covered oranges in wood thing
the frost on snow-covered oranges
green leaves on snow-covered oranges
red writing on white sign



OpenVik:
- striped mane belongs to grazing zebra
- zebra with striped ears eating green grass
- white stripe adorning leg
- dark brown mane growing behind head
- grass everywhere surround standing zebra
- black nose above green lively grass

Visual Genome-Scene Graph:
<hair, on, head>
<zebra, eat, grass>
<eye, on, zebra>
<grass, on, ground>

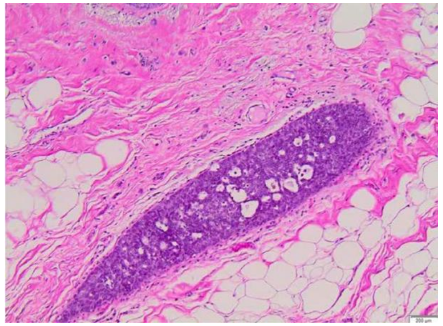
Visual Genome-Region Descriptions:
black and white striped leg
light shining on the zebra
thin line of black hair
two zebras grazing in the grass

Relational Caps:
sticking up ear of grazing zebra
black eye of eating zebra
grazing zebra in green grass
the muzzle of grazing zebra

Cui, Hejie, et al. "Open visual knowledge extraction via relation-oriented multimodality model prompting." *Advances in neural information processing systems* 36 (2023): 23499-23519.

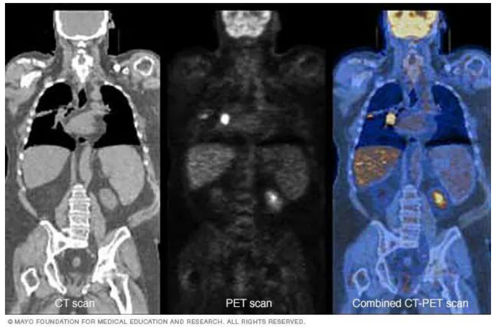
Visual Knowledge Extraction for Healthcare

Medical Images contain rich details in assisting disease diagnosis, interpretation and intervention.



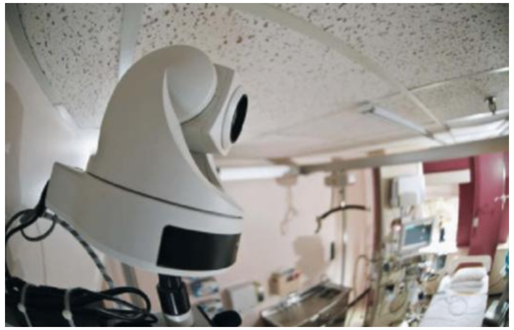
Pathology Slide

Cellular Detail



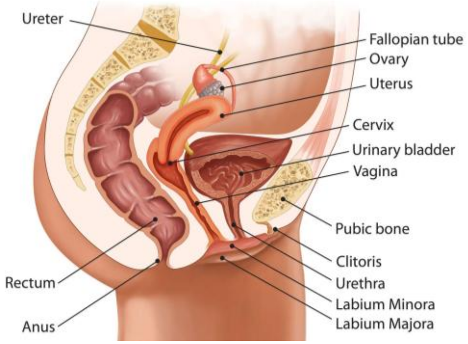
PET Scan

Metabolic Information



ICU Camera

Patient Behavior



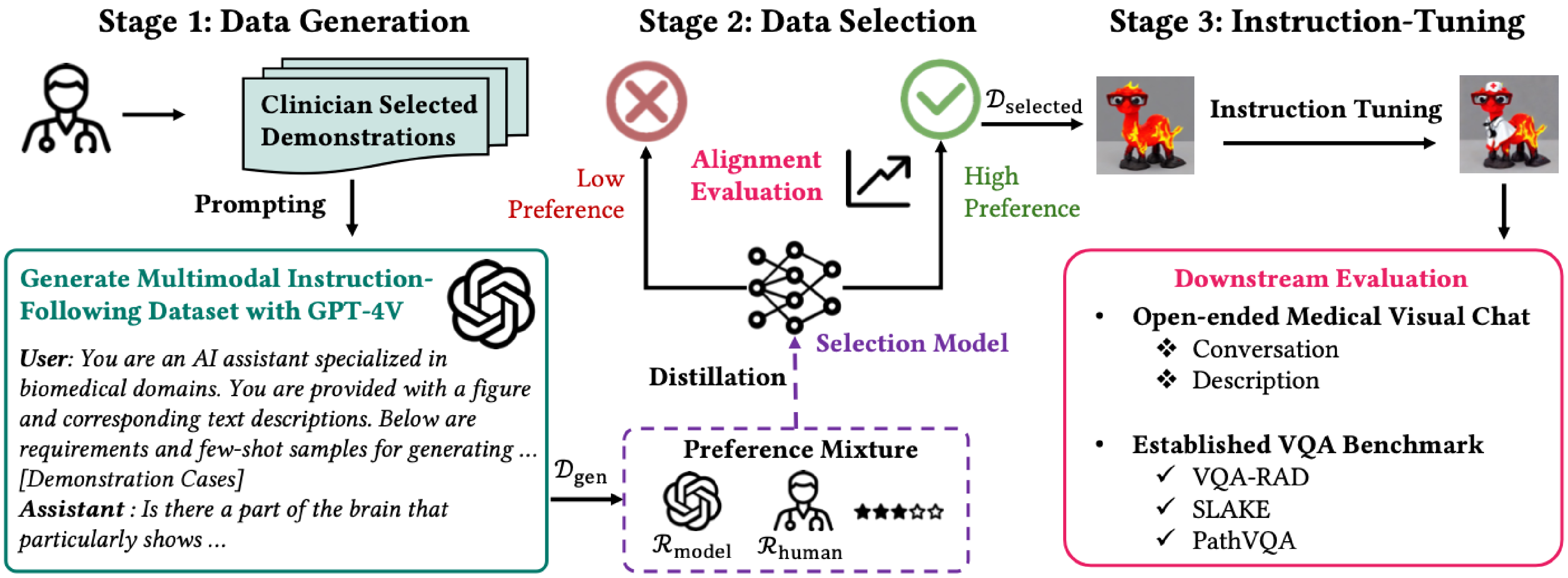
Anatomy Diagram

Anatomical Structure

By applying open visual knowledge extraction to the medical domain, we can unlock new insights and support clinical decision-making in powerful ways.

Cui, Hejie, et al. "Biomedical visual instruction tuning with clinician preference alignment." *Advances in neural information processing systems* (2024).

Biomed-VITAL



Cui, Hejie, et al. "Biomedical visual instruction tuning with clinician preference alignment." *Advances in neural information processing systems* (2024).

Stage 1: Data Generation

Diverse few-shot demonstration selection

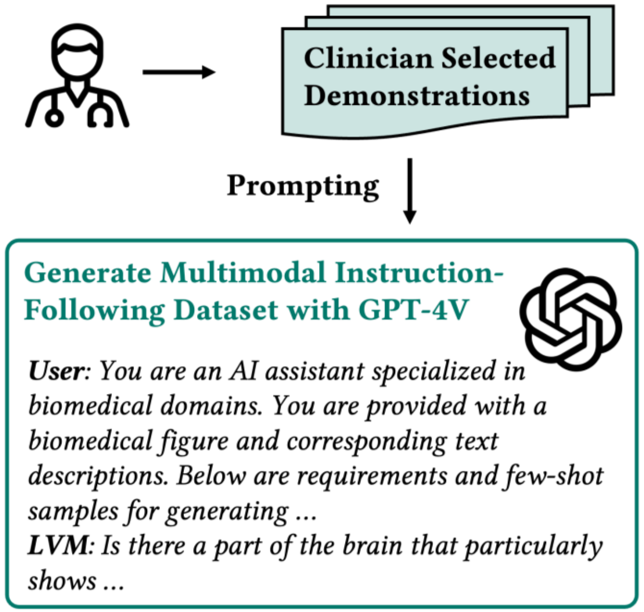
- Sample from K clusters to ensure the diversity of the clinician annotation for the generator

Instruction-following data generation with GPT-4V

- Incorporate visual input and clinician-annotated few-shot demonstrations

Raw dataset

- Image-text pairs from the PMC-15M dataset to generate multi-round QA instructional data



Cui, Hejie, et al. "Biomedical visual instruction tuning with clinician preference alignment." *Advances in neural information processing systems* (2024).

Stage 2: Data Selection

Preference data from two resources:

- Human preference: clinician annotation, limited but high quality
- Model preference: GPT-4V ratings based on clinician criteria, scalable complement

Preference distillation:

- Selection model training: pairwise ranking objective
- Adaptive preference mixing strategy

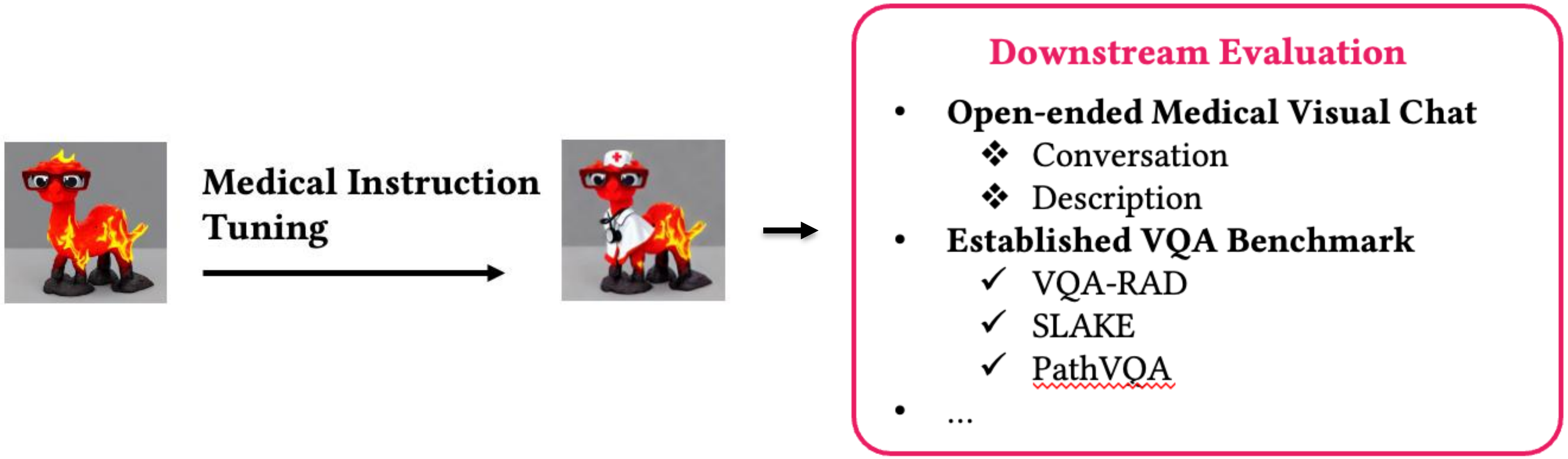
$$(z_i, z_j) = \begin{cases} (1,0), & \mathcal{R}_i \geq \mathcal{R}_j \\ (0,1), & \mathcal{R}_i < \mathcal{R}_j \end{cases}$$

$$\mathcal{L}_Q = -w_{i,j}\left(z_i \log \sigma\left(f(x_i)\right) + z_j \log \sigma\left(f(x_j)\right)\right)$$

Cui, Hejie, et al. "Biomedical visual instruction tuning with clinician preference alignment." *Advances in neural information processing systems* (2024).

Stage 3: Instruction-Tuning

- Continue training the LLaVA model on our curated instruction-following dataset



Cui, Hejie, et al. "Biomedical visual instruction tuning with clinician preference alignment." *Advances in neural information processing systems* (2024).

Experimental Results

Model	Data Size	Question Types: Conversation (#:143)	Question Types: Description (#: 50)	Domains: CXR (#: 37)	Domains: MRI (#: 38)	Domains: Histology (#: 44)	Domains: Gross (#: 34)	Domains: CT (#: 40)	Overall (#: 193)
LLaVA-Med	N	58.53	56.16	43.97	51.19	60.01	86.49	50.63	57.92
BioMed-VITAL	Top 10% $*N$	64.11	60.05	56.35	52.57	59.02	87.60	62.82	63.06
BioMed-VITAL	Top 50% $*N$	65.95	64.26	55.75	55.57	60.96	94.06	64.70	65.51
BioMed-VITAL	Top 80% $*N$	68.50	**67.65**	55.24	**58.73**	62.65	**101.88**	**67.05**	68.28
BioMed-VITAL	N	**69.73**	65.51	**59.22**	57.39	**67.15**	99.26	63.63	**68.63**
Model Ablation									
BioMed-VITAL^{A0}	N	65.38	60.63	63.48	53.82	57.32	92.30	58.16	64.15
BioMed-VITAL^{A1}	N	67.82	59.48	59.68	53.98	60.34	97.89	60.74	65.66
BioMed-VITAL^{A2}	N	67.53	62.78	60.64	54.62	61.07	98.27	61.21	66.30

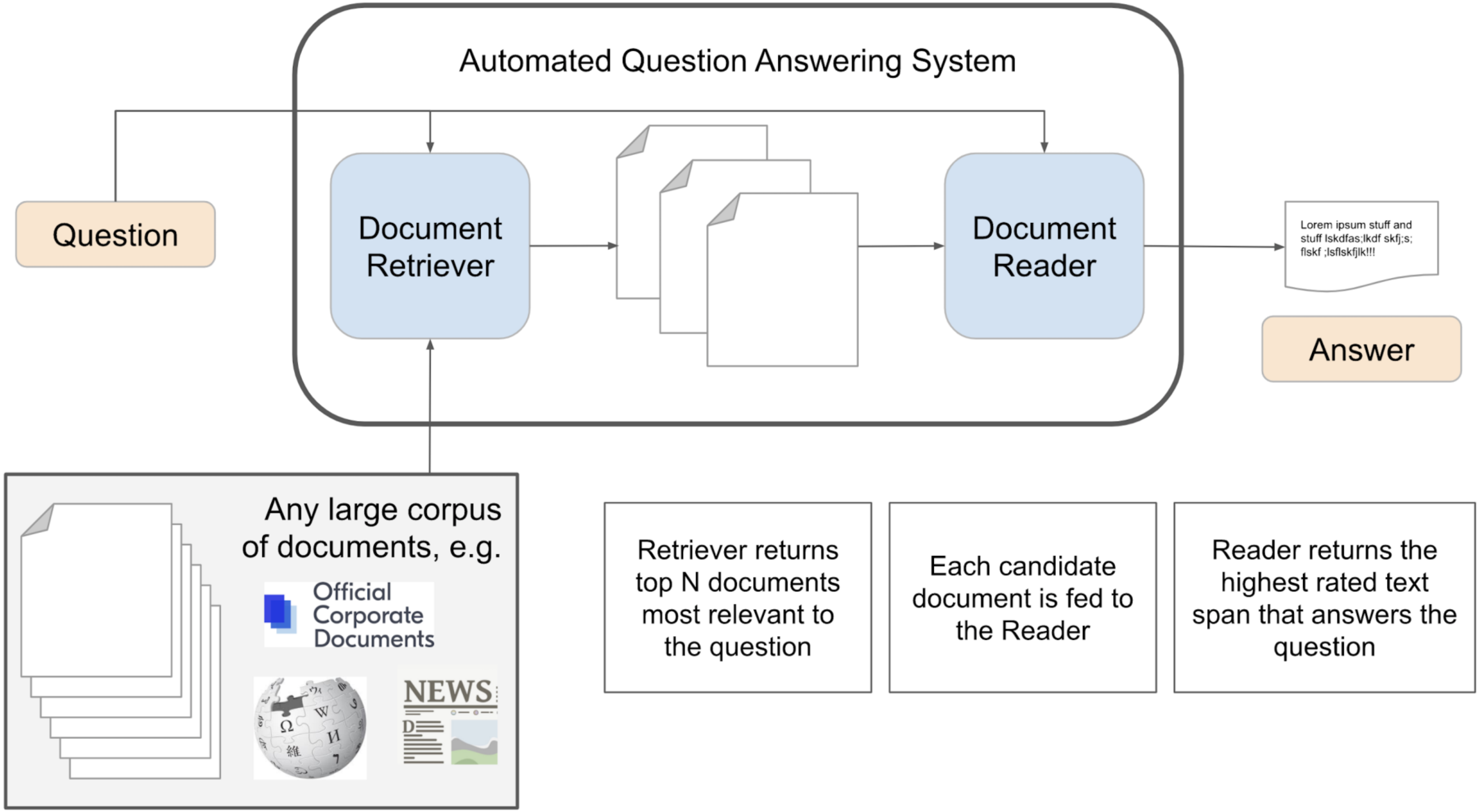
Model	VQA-RAD Ref	VQA-RAD Open	VQA-RAD Closed	SLAKE Ref	SLAKE Open	SLAKE Closed	PathVQA Ref	PathVQA Open	PathVQA Closed
Supervised fine-tuning results from models based on LLaVA (model size, training sample size)									
LLaVA (7B)		50.00	65.07		78.18	63.22		7.74	63.20
LLaVA-Med (7B, 60k)		61.52	84.19		83.08	85.34		37.95	91.21
LLaVA-Med (13B, 60k)		64.58	77.94		84.97	85.58		38.82	92.39
BioMed-VITAL (7B, 60k)		63.46	84.71		85.41	87.26		38.96	92.39
BioMed-VITAL (13B, 60k)		64.88	84.55		87.82	86.54		39.71	91.41
BioMed-VITAL (13B, 150k)		**69.72**	**84.86**		**91.69**	**90.70**		**39.89**	**92.42**
Literature-reported results from representative SoTA methods									
MMQ [9]	53.70		75.80				13.40		84.00
Prefix T. Medical LM [40]				84.30		82.01	40.00		87.00
PubMedCLIP [10]	60.10		80.00	78.40		82.50			
BiomedCLIP [46]	67.60		79.80	82.05		89.70			
M2I2 [21]	66.50		83.50	74.70		91.10	36.30		88.00
MUMC [20]	71.50		84.20	81.50		81.50	39.00		65.10
M3AE [5]	67.23		83.46	80.31		87.82			
CoQAH [16]	30.20		67.50	42.50		73.90			
PMC-CLIP [22]	67.00		84.00	81.90		88.00			

Cui, Hejie, et al. "Biomedical visual instruction tuning with clinician preference alignment." *Advances in neural information processing systems* (2024).

Tutorial outline

Time	Content	Part	Presenter
30 min	**•Introduction and background** • Artificial general intelligence (AGI) • Large language models (LLMs) and knowledge graphs (KGs) • Challenges and opportunities	Part 1	Shirui Pan
60 min	**•Knowledge graph-enhanced large language models** • KG-enhanced LLM Training • KG-enhanced LLM Reasoning • Unified KG+LLM Reasoning	Part 2	Linhao Luo
30 min break			
60 min	**•Large language model-enhanced knowledge graphs** • LLM-enhanced KG integrations • LLM-enhanced KG construction and completion • LLM-enhanced Multi-modality KG	Part 3	Carl Yang
20 min	**•Applications of synergized KG-LLM systems** • QA system • Recommender system	Part 4	Evgeny Kharlamov
10 min	**•Future directions and conclusion**	Part 5	Linhao Luo

KG+LLM QA System



https://qa.fastforwardlabs.com/methods/background/2020/04/28/Intro-to-QA.html

KG+LLM QA System

- Limitations and challenges of existing QA system.

- **Domain-specific knowledge understanding**
  - Structured data, unstructured data, and domain expert rules.

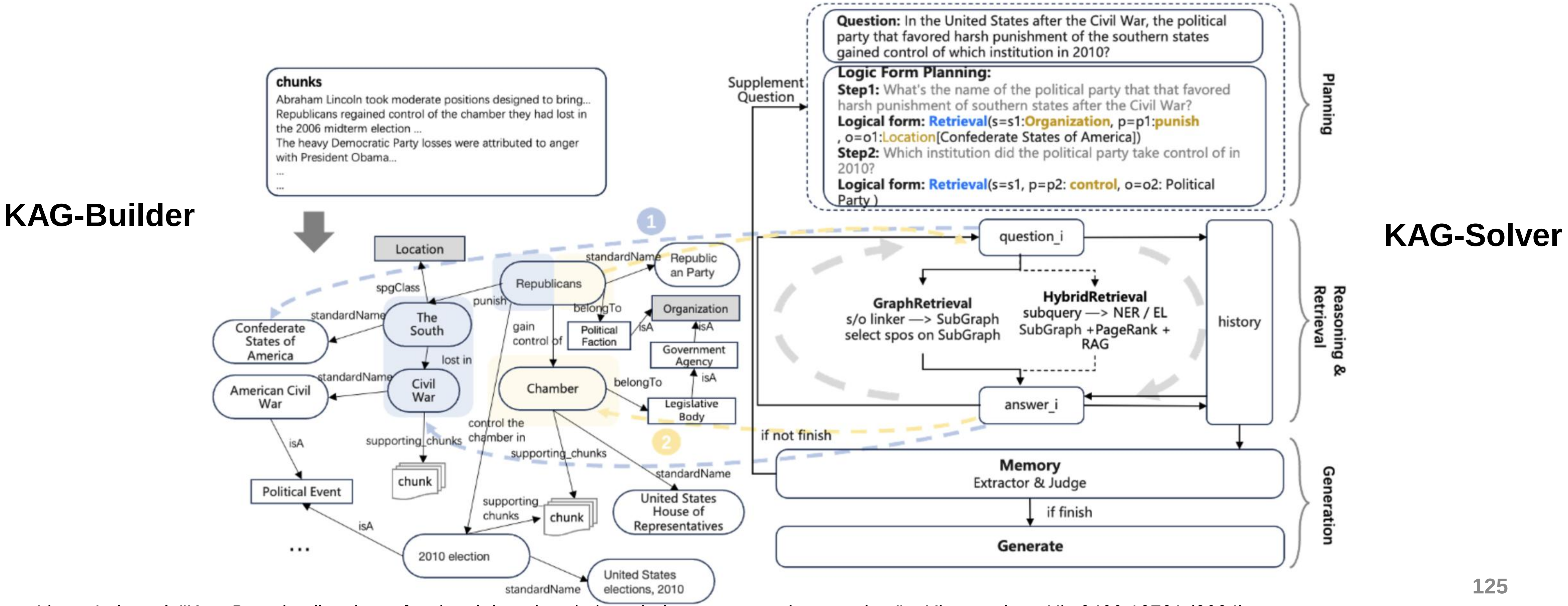
- **Lacking symbolic reasoning expression**
  - Retrieval strategy based on semantic similarity cannot handle complex reasoning, quantitative analysis.
  - For example, how many rivers pass through Indian and China?

- **LLM hallucinations**
  - LLM can still hallucinate even with retrieved documents.

Liang, Lei, et al. "Kag: Boosting llms in professional domains via knowledge augmented generation." arXiv preprint arXiv:2409.13731 (2024).

KAG – Joint KG + LLM Reasoning

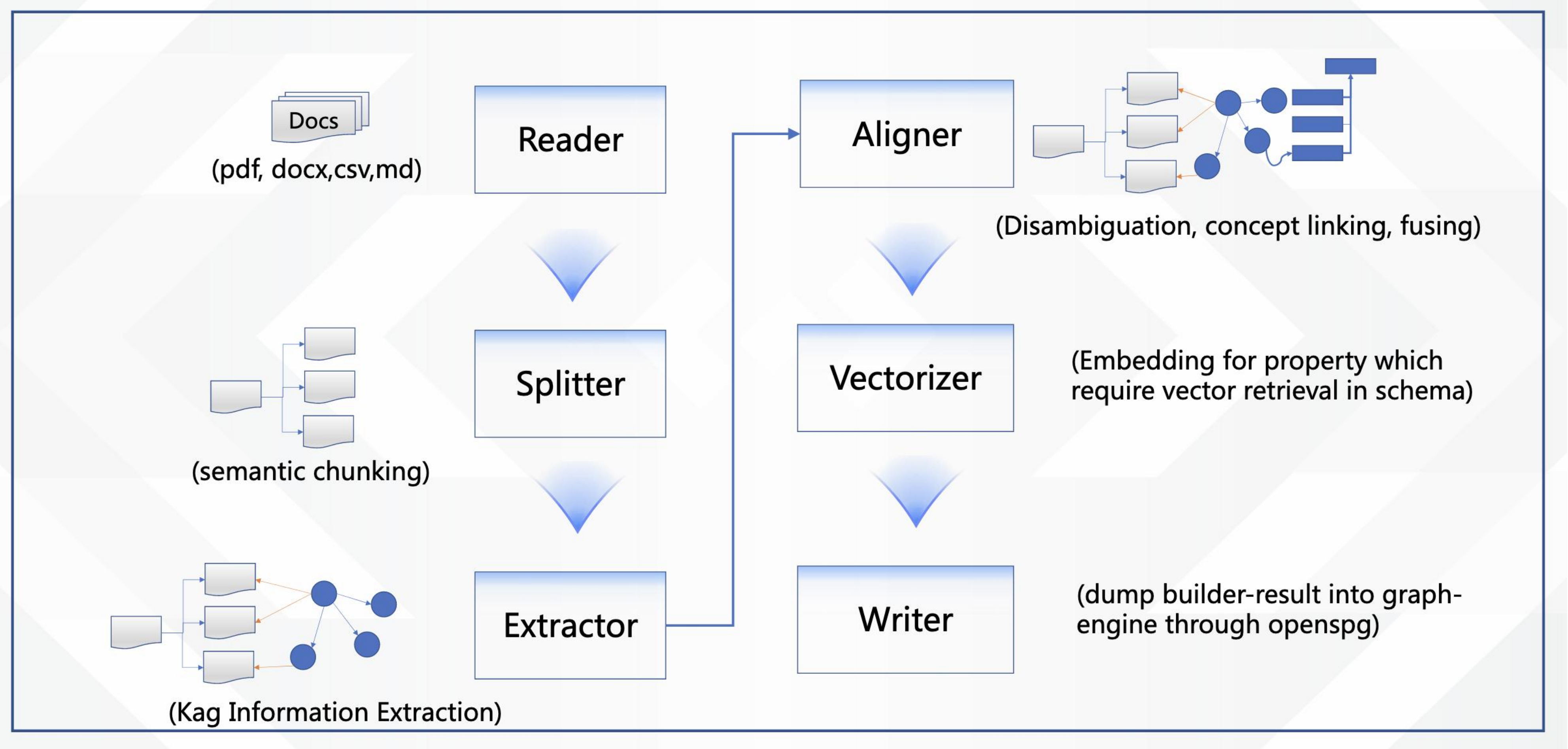
- **KG Symbolic Reasoning**: Logical Form Execution.
- **LLM Neural Reasoning**: CoT Reasoning and Planning.



Liang, Lei, et al. "Kag: Boosting llms in professional domains via knowledge augmented generation." arXiv preprint arXiv:2409.13731 (2024).

KAG – Joint KG + LLM Reasoning

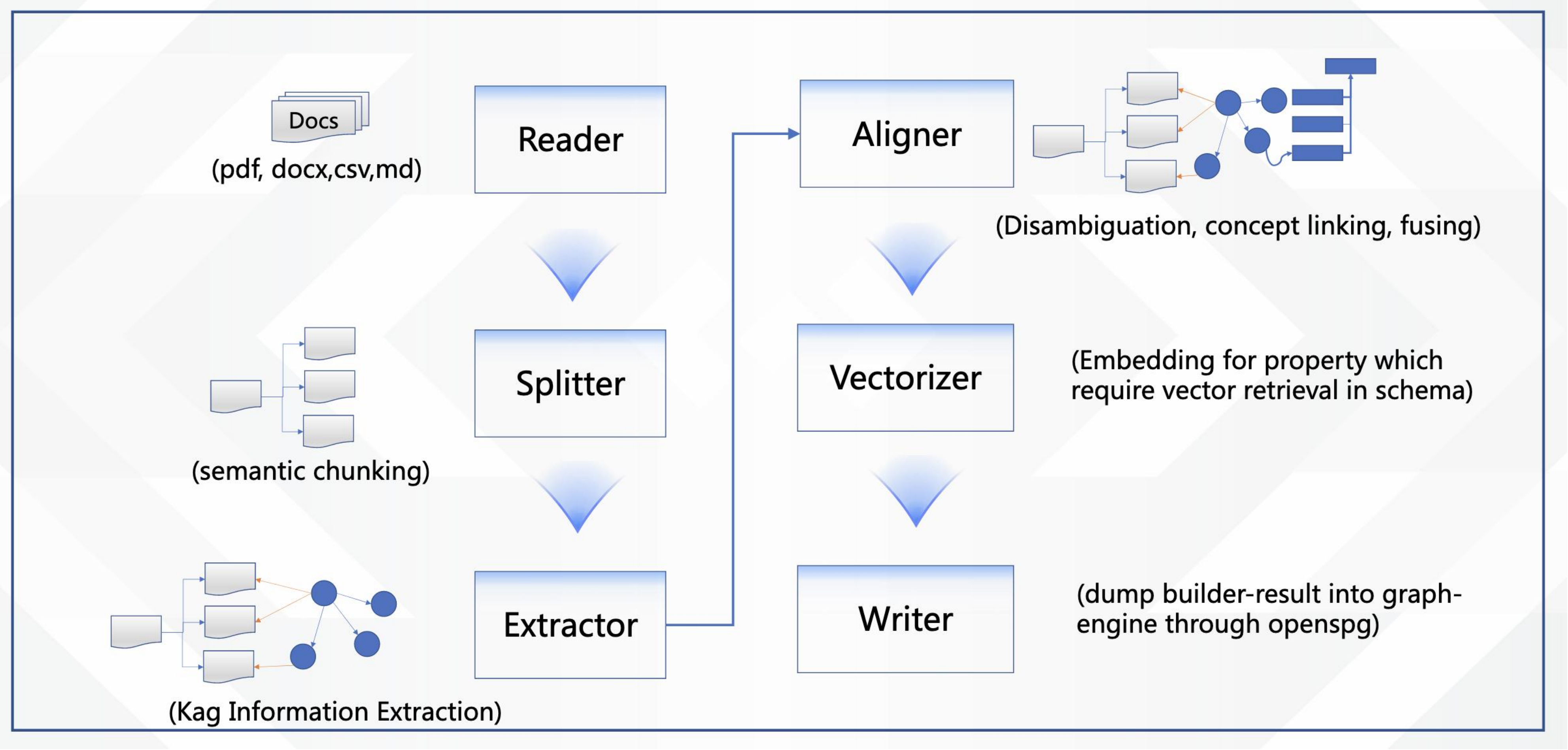
- **KAG-Builder**



Liang, Lei, et al. "Kag: Boosting llms in professional domains via knowledge augmented generation." arXiv preprint arXiv:2409.13731 (2024).

KAG – Joint KG + LLM Reasoning

- **KAG-Builder**

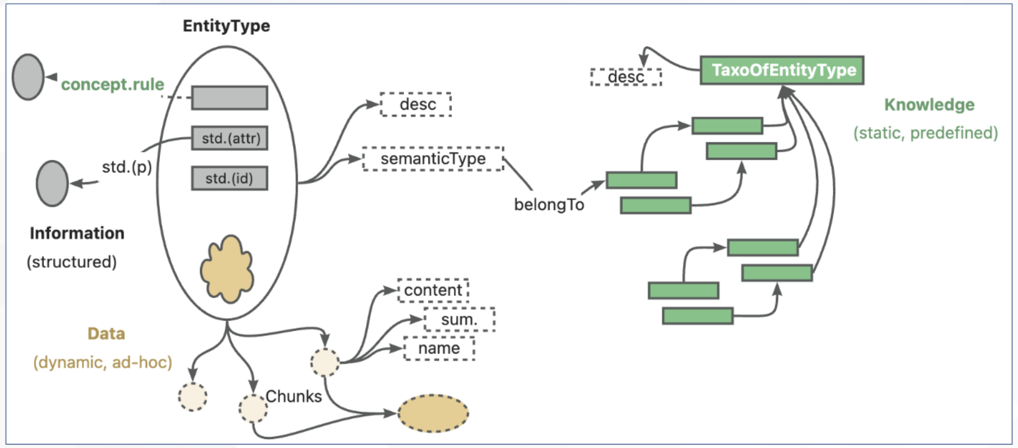


Liang, Lei, et al. "Kag: Boosting llms in professional domains via knowledge augmented generation." arXiv preprint arXiv:2409.13731 (2024).

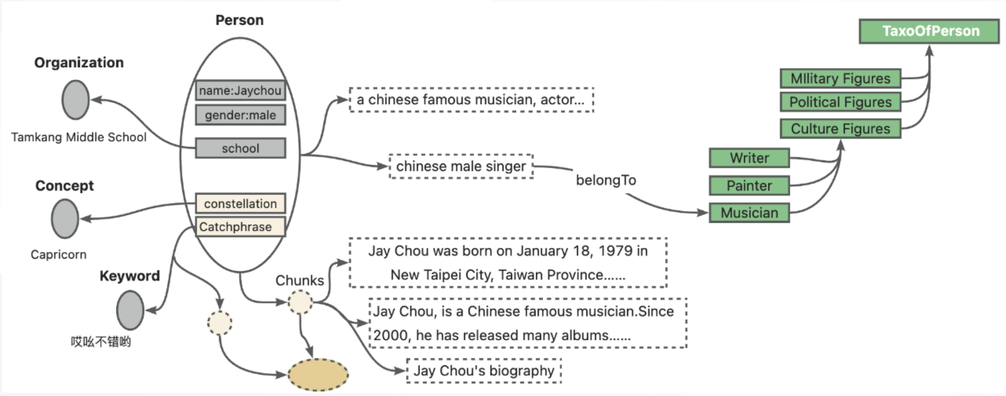
KAG – Joint KG + LLM Reasoning

- **KAG-Schema and text-KG joint index**

Kag – Indexing Structure



Kag – Indexing instance of Jay Chou



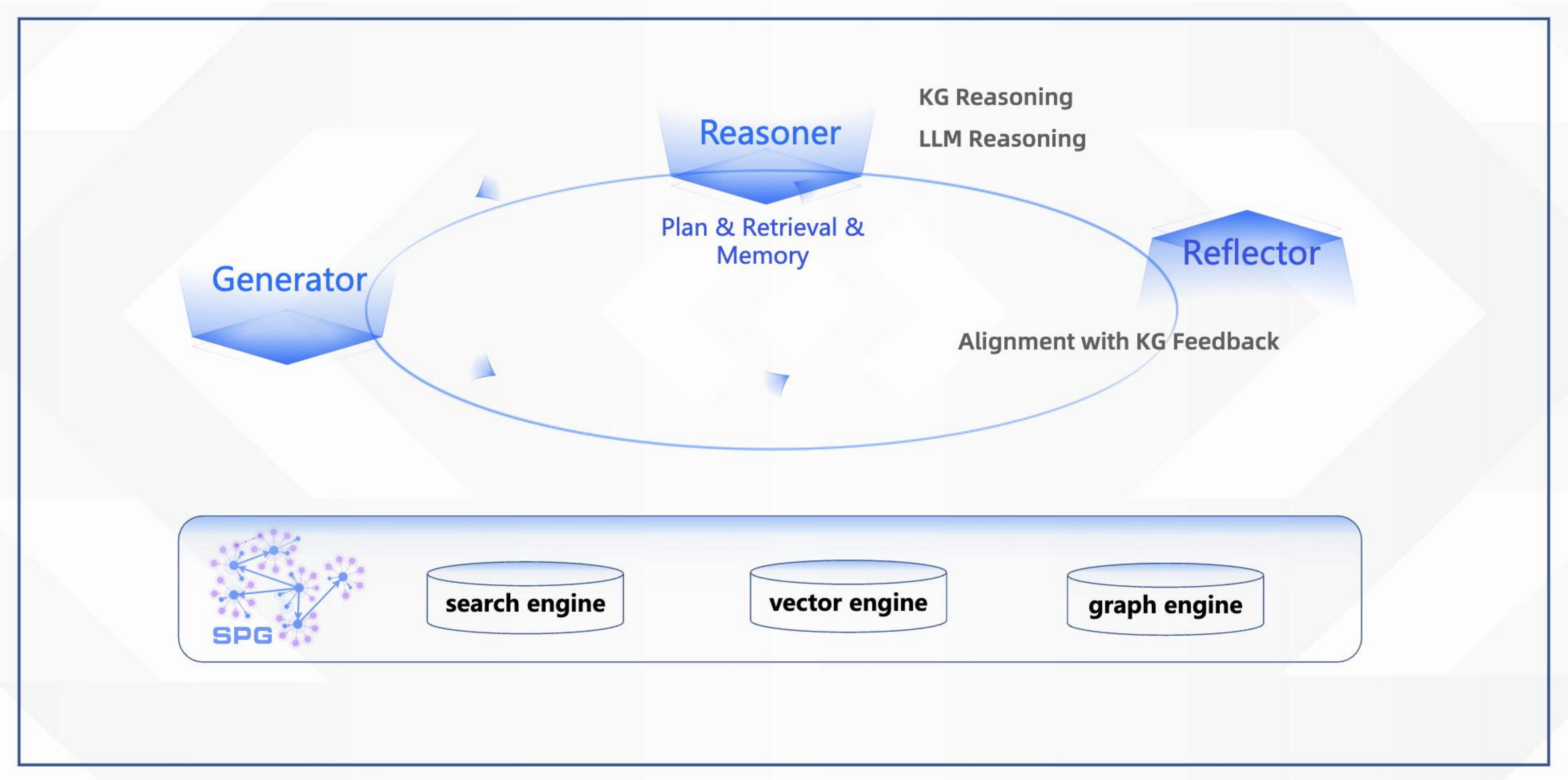
default.schema

```
Organization: EntityType
    properties:
      id: Text
        index: TextAndVector
      name: Text
        index: TextAndVector
      desc: Text
        index: TextAndVector
      semanticType: Text

Person: EntityType
    properties:
      id: Text
        index: TextAndVector
      name: Text
        index: TextAndVector
      desc: Text
        index: TextAndVector
      school: Organization
      gender: Text
      semanticType: Text

Works: EntityType
Concept: EntityType
GeoLocation: EntityType
......
Chunks: EntityType
Others: EntityType
```


Liang, Lei, et al. "Kag: Boosting llms in professional domains via knowledge augmented generation." arXiv preprint arXiv:2409.13731 (2024).

KAG – Joint KG + LLM Reasoning

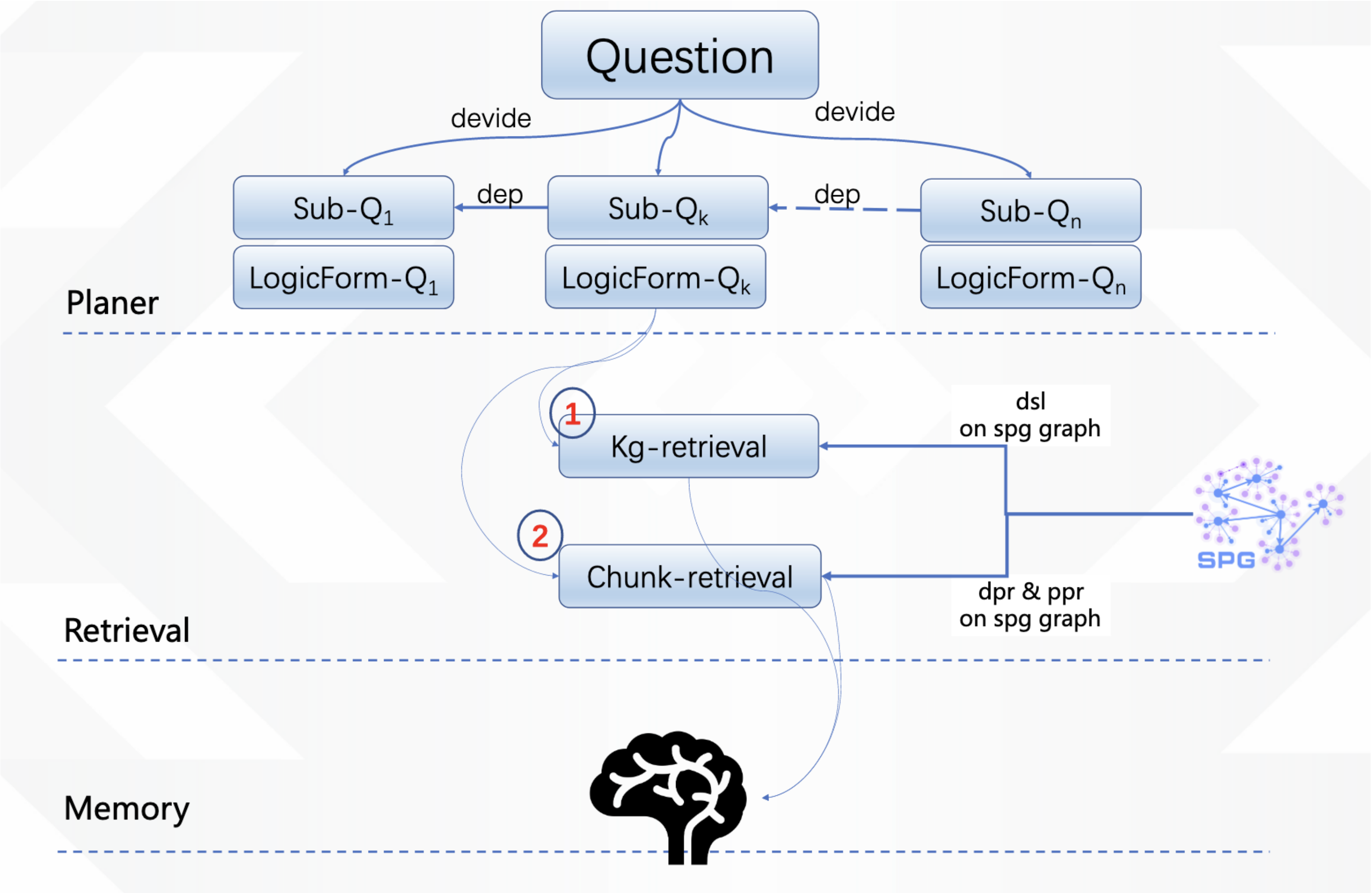
- **KAG-solver**



Liang, Lei, et al. "Kag: Boosting llms in professional domains via knowledge augmented generation." arXiv preprint arXiv:2409.13731 (2024).

KAG – Joint KG + LLM Reasoning

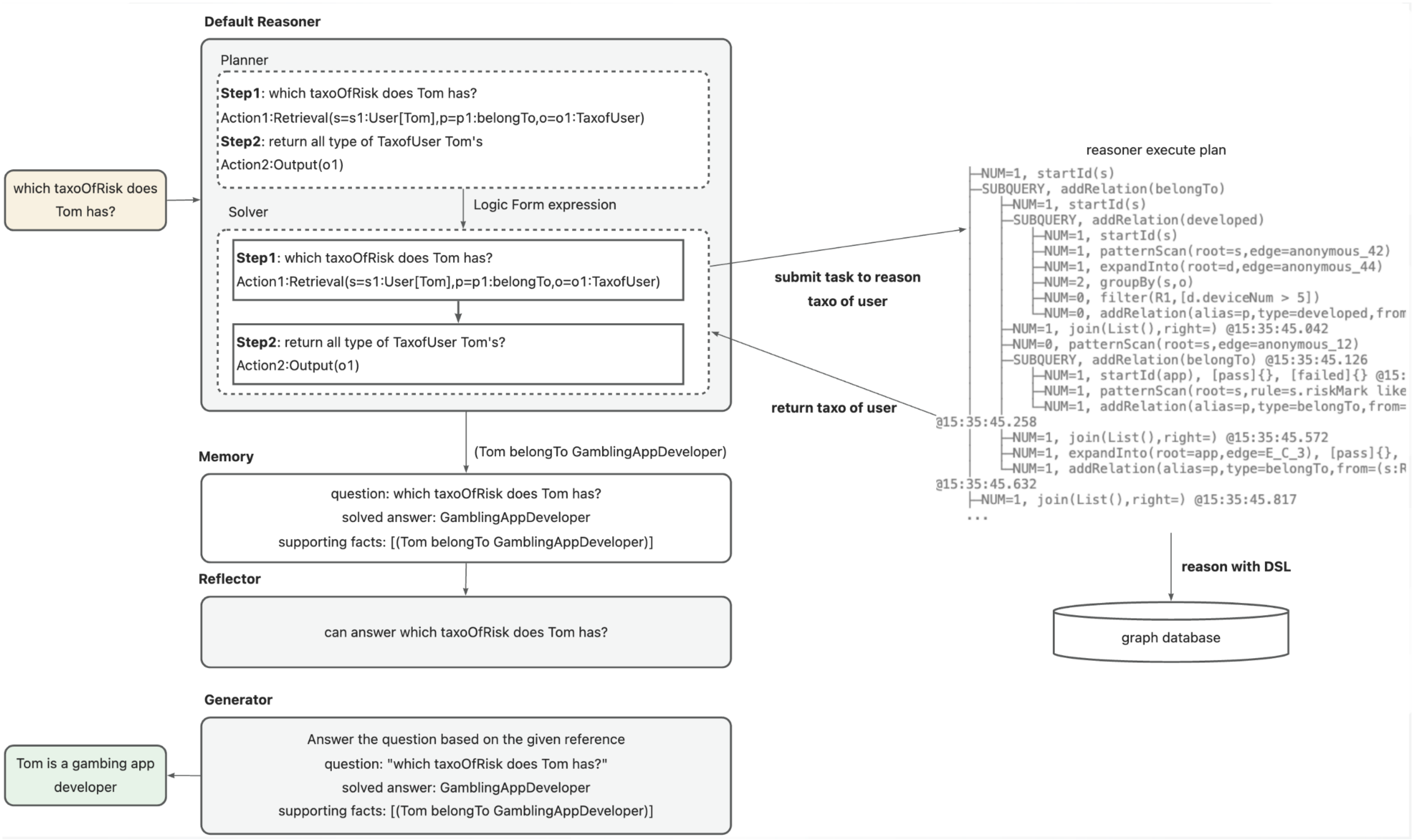
- **KAG-solver**



Liang, Lei, et al. "Kag: Boosting llms in professional domains via knowledge augmented generation." arXiv preprint arXiv:2409.13731 (2024).

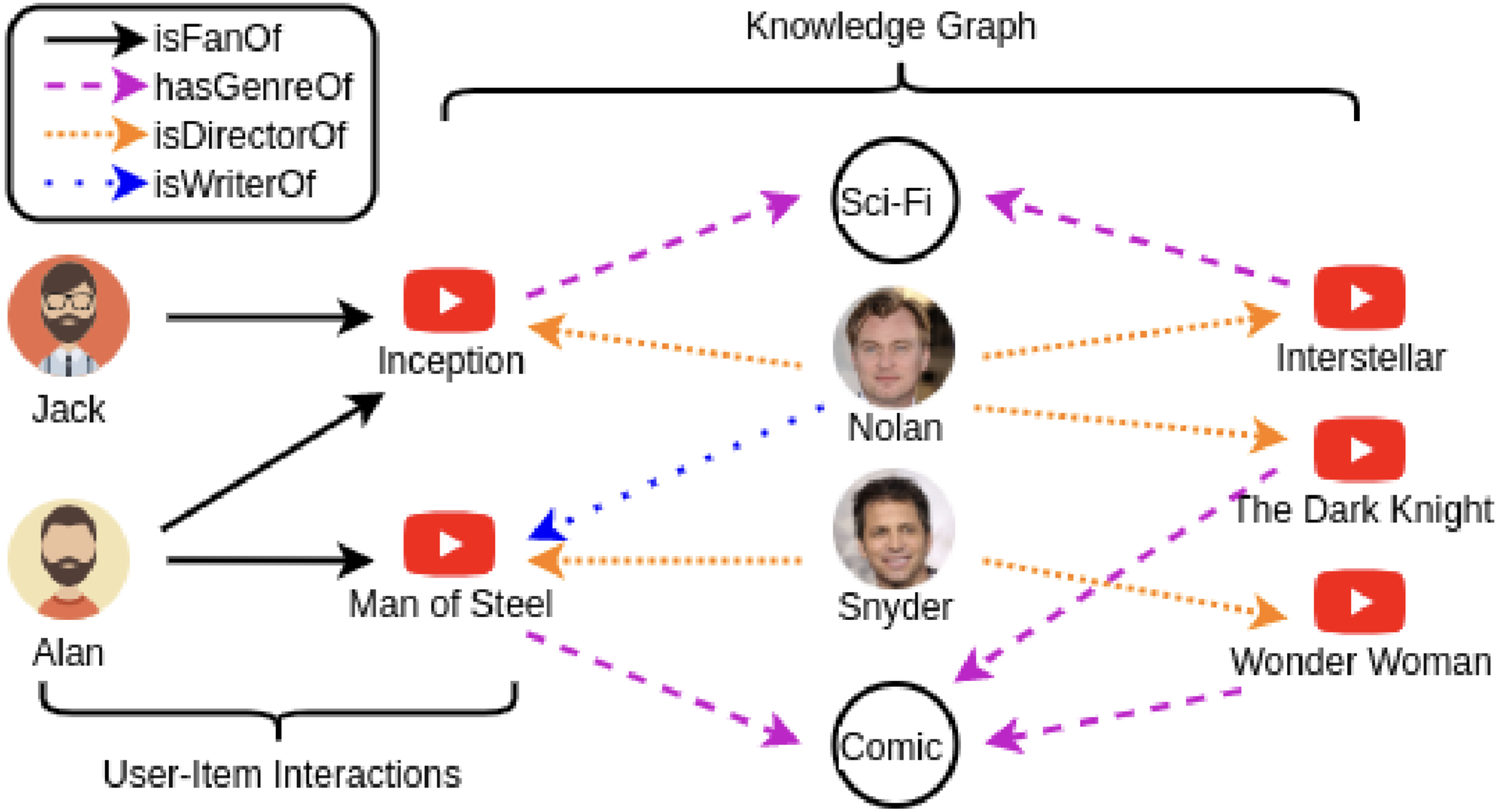
KAG – Joint KG + LLM Reasoning

- **KAG-solver**

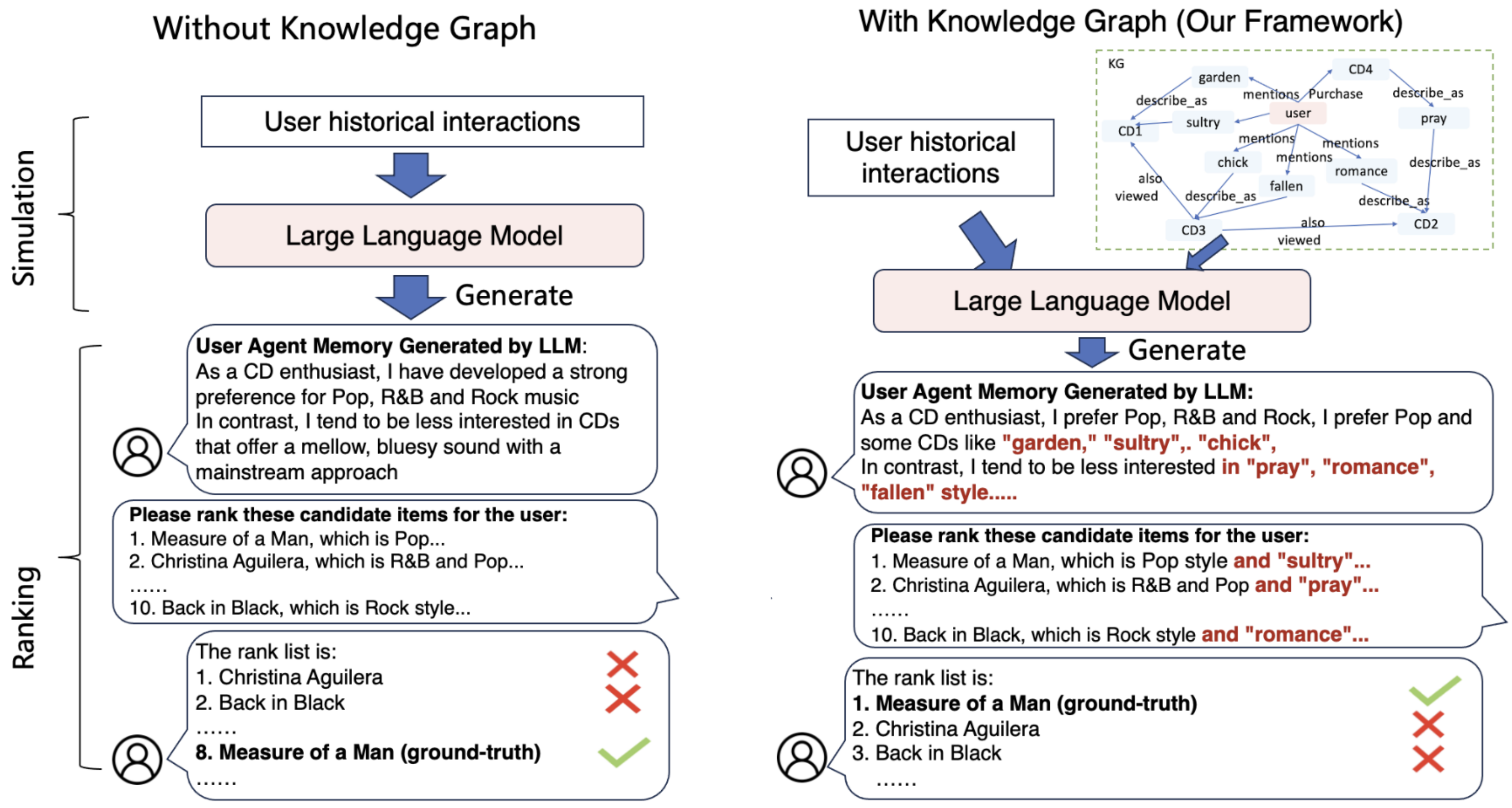


Liang, Lei, et al. "Kag: Boosting llms in professional domains via knowledge augmented generation." arXiv preprint arXiv:2409.13731 (2024).

KG+LLM for Recommender System



KG+LLM for Recommender System



(a) Without KG

(b) With KG

Guo, Taicheng, et al. "Knowledge Graph Enhanced Language Agents for Recommendation." arXiv preprint arXiv:2410.19627 (2024).

G-Refer: Graph Retrieval-Augmented Large Language Model for Explainable Recommendation

TheWebConf 2025

Yuhan Li, Xinni Zhang, Linhao Luo, Heng Chang, Yuxiang Ren, Irwin King, Jia Li

Oral Presentation: Wednesday, 30 April 2025, 10:30 - 12:00, Session 4 Recsys1, C2.1 Room

Poster: Wednesday, 30 April, 16:30pm-17:15pm, Research@Parkside Ballroom, Posterboard-04

香港科技大学(广州) THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)
香港中文大學 The Chinese University of Hong Kong
MONASH University
HUAWEI

KG+LLM for Recommender System

- Explainable Recommendation

  - The primary goal of explainable recommendation is to **create clear textual explanations** that allow us to understand the rationale behind each recommendation. Specifically, for each interaction between a user u and an item i, the explanations generated can be described as follows:

$$explanation(u, i) = generate(u, i, \mathcal{X}_u, \mathcal{X}_i, \tau)$$

  - User/item profiles, interaction histories, and any side-information related to both users and items.

XRec: Large Language Models for Explainable Recommendation. EMNLP 2024 Findings.

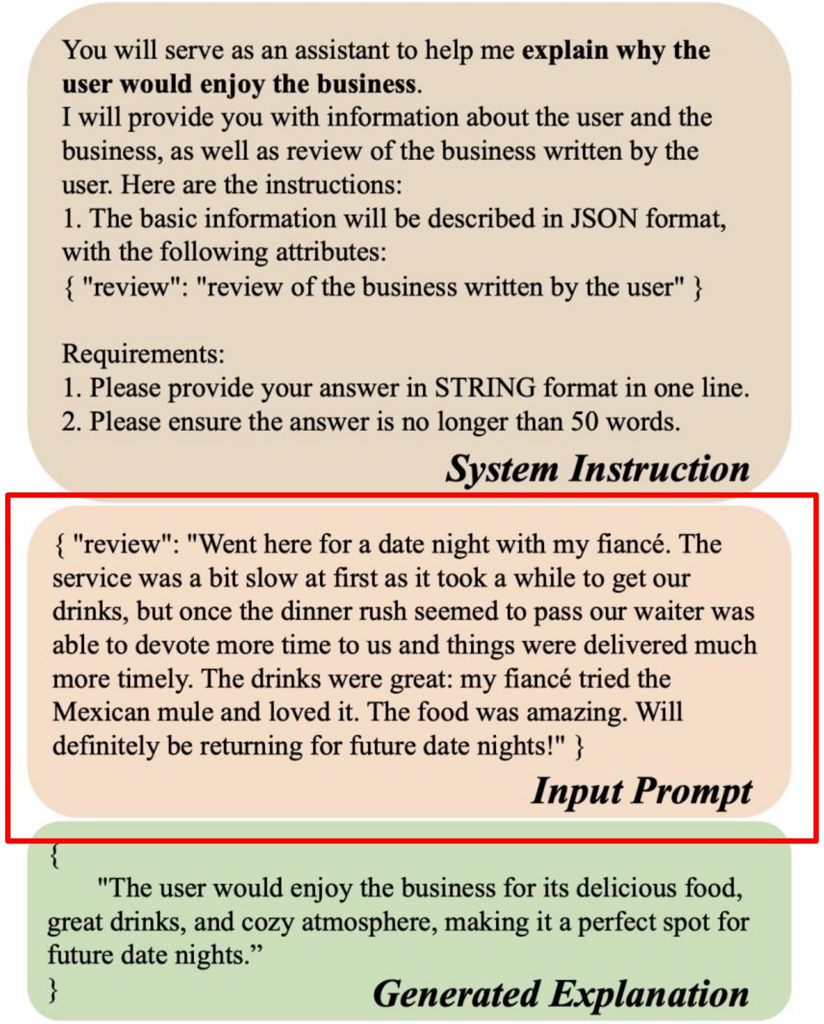
Existing Datasets & SoTA Solutions

- Datasets Proposed by Ma et al. (2024)

Table 2: Statistics of the experimental datasets.

Dataset	#Users	#Items	#Interactions
Amazon	15,349	15,247	360,839
Yelp	15,942	14,085	393,680
Google	22,582	16,557	411,840

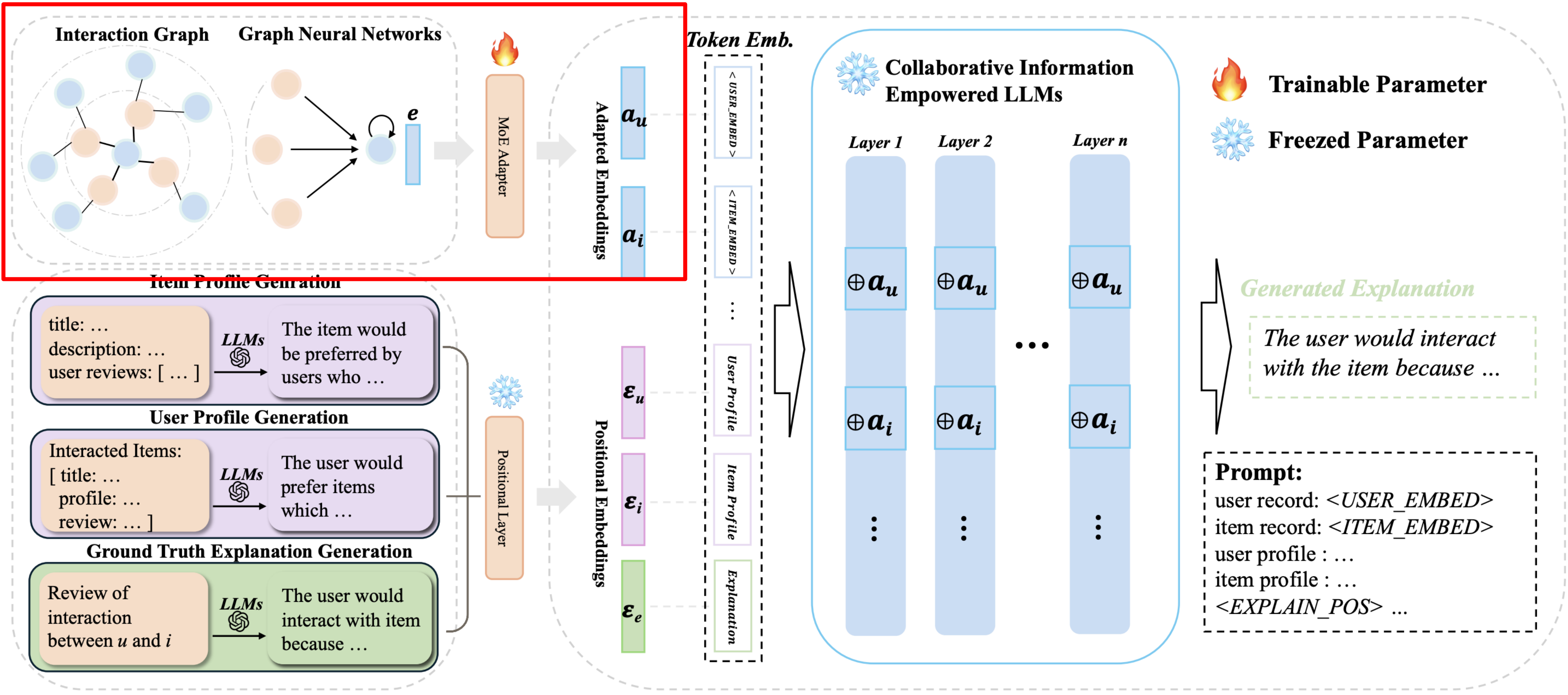
- Generating ground truth explanations by rephasing users' actual reviews.



XRec: Large Language Models for Explainable Recommendation. EMNLP 2024 Findings.

Existing Datasets & SoTA Solutions

- SoTA baseline - XRec



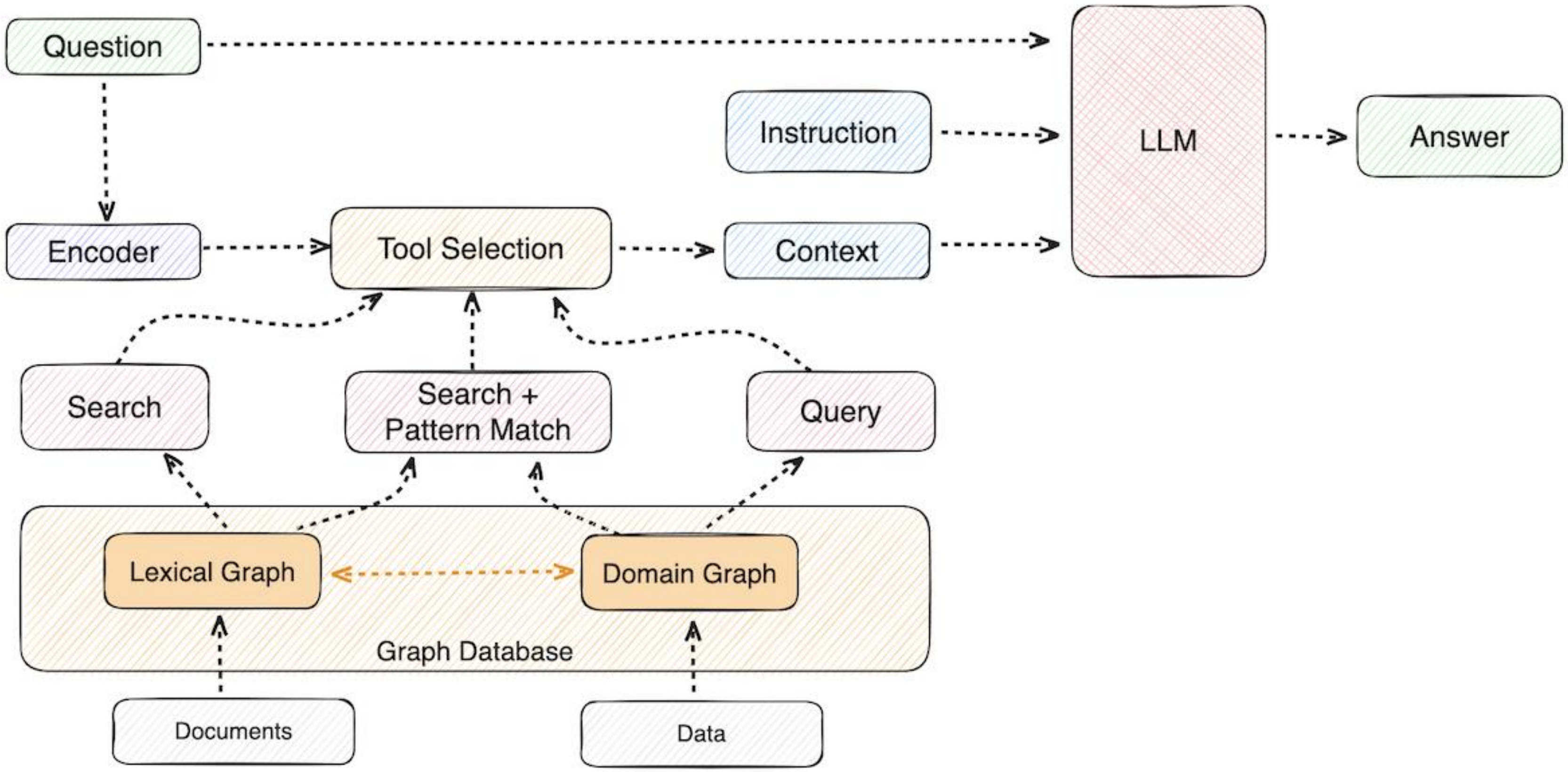
1) Implicit CF Signal.

2) Modality Gap.

XRec: Large Language Models for Explainable Recommendation. EMNLP 2024 Findings.

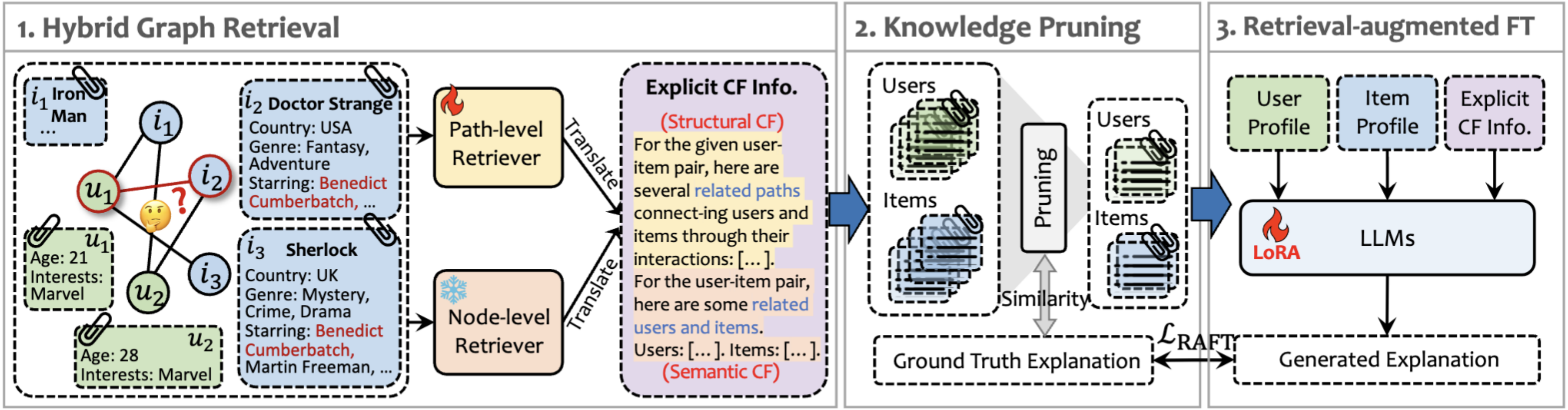
Our Solution

- Introducing GraphRAG
  - Implicit -> Explicit
  - Modality gap -> None



Overview of G-Refer

- Three key modules

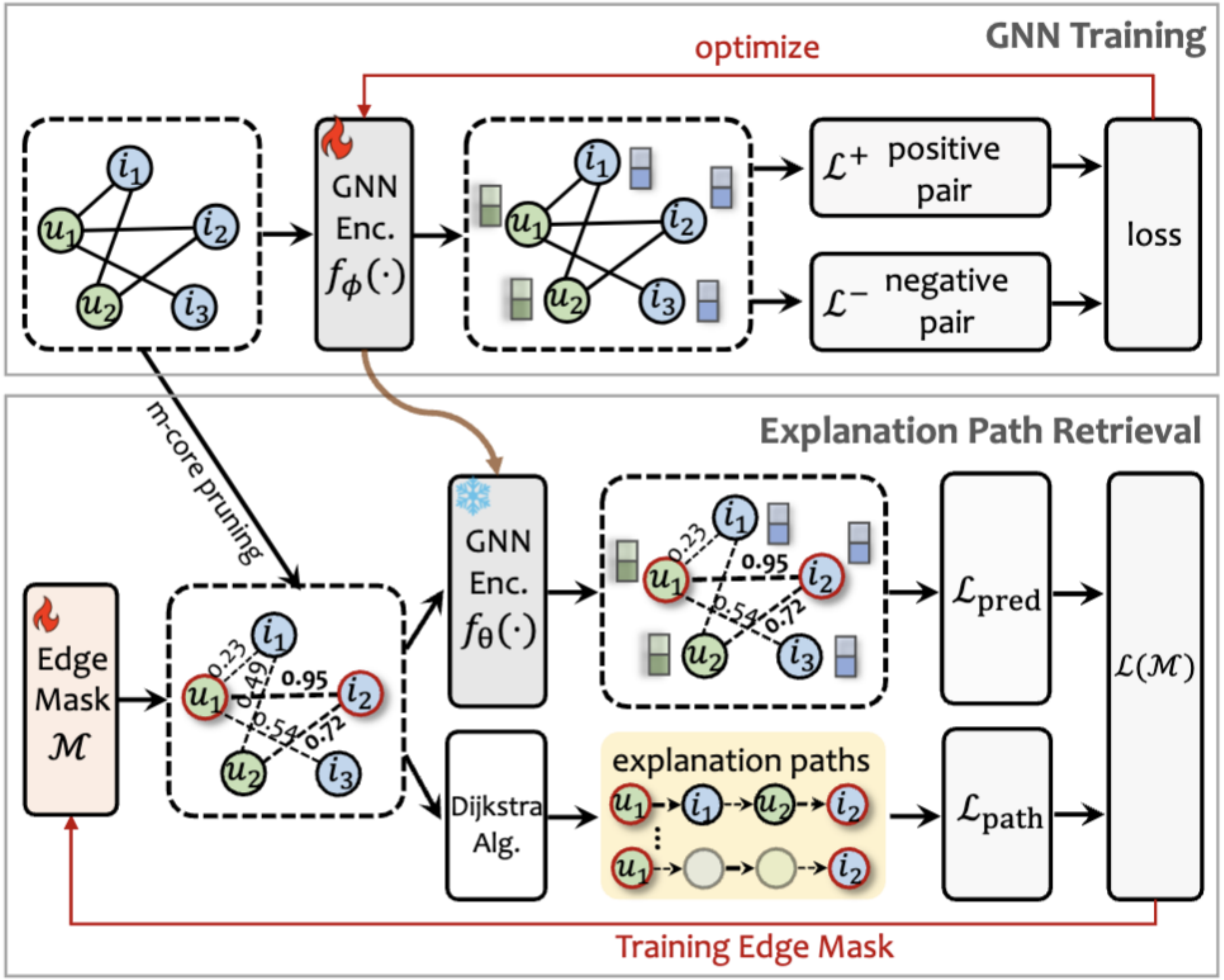


Hybrid Graph Retrieval

- Path-level Retriever
  - PaGE-Link (WWW'23)
  - GNN T raining
  - M-core Pruning
  - Explanation Path Retrieval

- Node-level Retriever
  - Dense Retrieval

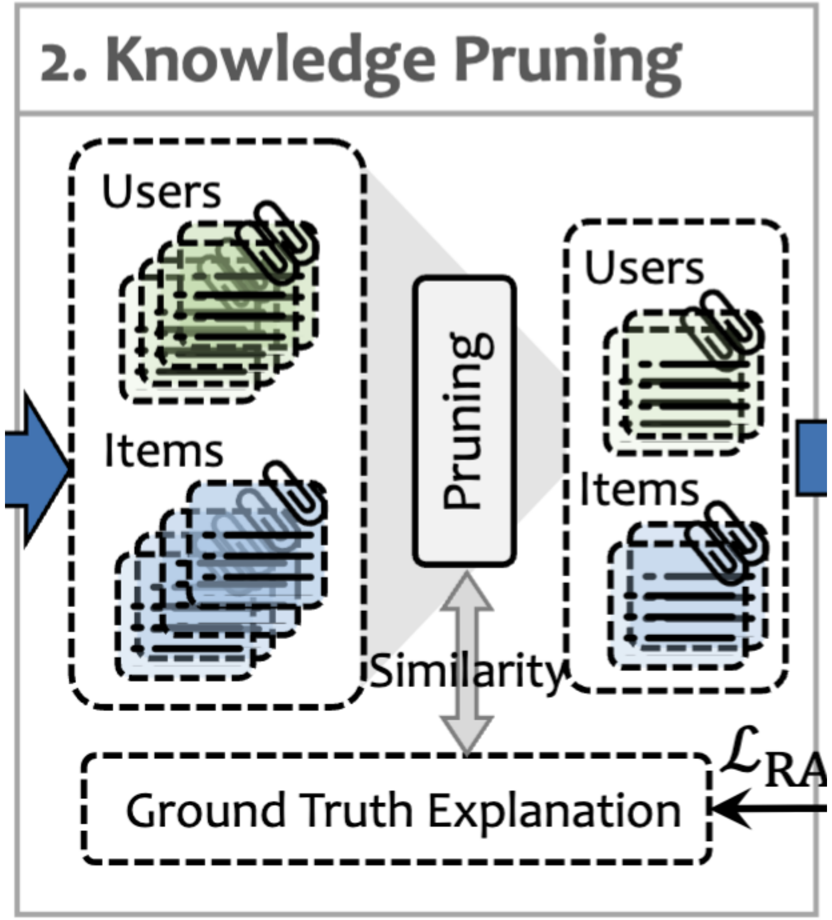
- Graph Translation



Knowledge Pruning

- Motivation
  - It is noticed that for some user-item pairs, a sufficient explanation can be derived solely from their profiles, without the need for additional CF information.
- Re-Ranking and Pruning

$$sim((u,i), \mathrm{Explain}_{(u,i)}) = \frac{f(b_u \oplus c_i) \cdot f(\mathrm{Explain}_{(u,i)})}{\|f(b_u \oplus c_i)\| \|f(\mathrm{Explain}_{(u,i)})\|}$$



Retrieval-augmented Fine-tuning

- Motivation
  - RAFT adapts the LLM to better utilize retrieved CF information to generate explanation, especially for the requirement of domain-specific knowledge it has never seen before.
  - By training the LLM to generate ground-truth responses even when irrelevant CF information is given, we enable the LLM to ignore misleading retrieval content and lean into its internal knowledge to reduce hallucination.
- RAFT Loss

$$\mathcal{L}_{\text{RAFT}} = -\sum_{(u,i)\in\mathcal{D}_{\text{prune}}} \log P(\text{Explain}_{(u,i)} | b_u, c_i, \mathcal{K}_{(u,i)}, Q; \theta),$$

Experiment Results

Models	Explainability ↑							Stability ↓					
	GPT_{score}	$BERT^{P}_{score}$	$BERT^{R}_{score}$	$BERT^{F1}_{score}$	$BART_{score}$	BLEURT	USR	GPT_{std}	$BERT^{P}_{std}$	$BERT^{R}_{std}$	$BERT^{F1}_{std}$	$BART_{std}$	$BLEURT_{std}$
Amazon-books													
NRT	75.63	0.3444	0.3440	0.3443	-3.9806	-0.4073	0.5413	12.82	0.1804	0.1035	0.1321	0.5101	0.3104
Att2Seq	76.08	0.3746	0.3624	0.3687	-3.9440	-0.3302	0.7757	12.56	0.1691	0.1051	0.1275	0.5080	0.2990
PETER	77.65	**0.4279**	0.3799	0.4043	-3.8968	-0.2937	0.8480	11.21	0.1334	0.1035	0.1098	0.5144	0.2667
PEPLER	78.77	0.3506	0.3569	0.3543	-3.9142	-0.2950	0.9563	11.38	0.1105	0.0935	0.0893	0.5064	0.2195
XRec	82.57	0.4193	0.4038	0.4122	-3.8035	**-0.1061**	**1.0000**	9.60	**0.0836**	0.0920	**0.0800**	0.4832	**0.1780**
G-Refer (7B)	82.70	0.4076	0.4476	0.4282	-3.3358	-0.1246	1.0000	9.04	0.0937	**0.0845**	0.0820	0.4009	0.1893
G-Refer (8B)	**82.82**	0.4073	**0.4494** (+4.56%)	**0.4289** (+1.67%)	**-3.3110**	-0.1203	1.0000	**8.95**	0.0945	0.0855	0.0825	**0.3983**	0.1912
Yelp													
NRT	61.94	0.0795	0.2225	0.1495	-4.6142	-0.7913	0.2677	16.81	0.2293	0.1134	0.1581	0.5612	0.2728
Att2Seq	63.91	0.2099	0.2658	0.2379	-4.5316	-0.6707	0.7583	15.62	0.1583	0.1074	0.1147	0.5616	0.2470
PETER	67.00	0.2102	0.2983	0.2513	-4.4100	-0.5816	0.8750	15.57	0.3315	0.1298	0.2230	0.5800	0.3555
PEPLER	67.54	0.2920	0.3183	0.3052	-4.4563	-0.3354	0.9143	14.18	0.1476	0.1044	0.1050	0.5777	0.2524
XRec	74.53	**0.3946**	0.3506	0.3730	-4.3911	-0.2287	**1.0000**	11.45	**0.0969**	0.1048	**0.0852**	0.5770	0.2322
G-Refer (7B)	74.91	0.3573	0.4264	0.3922	-3.7729	-0.1451	**1.0000**	10.88	0.1050	**0.0952**	0.0862	0.4815	0.2197
G-Refer (8B)	**75.16**	0.3629	**0.4373** (+8.67%)	**0.4003** (+2.73%)	**-3.6448**	-0.1336	**1.0000**	**10.76**	0.1068	0.0995	0.0885	**0.4743**	**0.2182**
Google-reviews													
NRT	58.27	0.3509	0.3495	0.3496	-4.2915	-0.4838	0.2533	19.16	0.2176	0.1267	0.1571	0.6620	0.3118
Att2Seq	61.31	0.3619	0.3653	0.3636	-4.2627	-0.4671	0.5070	17.47	0.1855	0.1247	0.1403	0.6663	0.3198
PETER	65.16	0.3892	0.3905	0.3881	-4.1527	-0.3375	0.4757	17.00	0.2819	0.1356	0.2005	0.6701	0.3272
PEPLER	61.58	0.3373	0.3711	0.3546	-4.1744	-0.2892	0.8660	17.17	0.1134	0.1161	0.0999	0.6752	0.2484
XRec	69.12	**0.4546**	0.4069	0.4311	-4.1647	-0.2437	0.9993	14.24	**0.0972**	0.1163	0.0938	0.6591	0.2452
G-Refer (7B)	71.47	0.4253	0.4873	0.4566	-3.3857	-0.1561	**1.0000**	13.46	0.1184	**0.0872**	0.0921	**0.4739**	**0.2415**
G-Refer (8B)	**71.73**	0.4245	**0.4935** (+7.48%)	**0.4592** (+2.81%)	**-3.3235**	**-0.1518**	**1.0000**	**13.23**	0.1175	0.0920	**0.0916**	0.4761	0.2511

Human Evaluation & Ablation

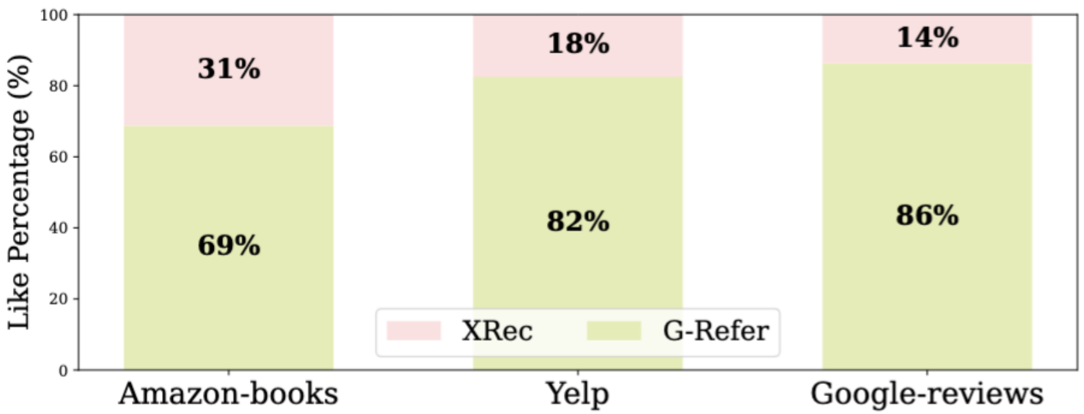


Figure 3: Human evaluation comparing XRec and G-Refer.

Table 2: Ablation study for G-Refer, with the best results highlighted red and the worst in blue for the component.

Datasets	Yelp		Google-reviews	
Ablations	$\text{BERT}^{F1}\uparrow$	$\text{BERT}_{std}\downarrow$	$\text{BERT}^{F1}\uparrow$	$\text{BERT}_{std}\downarrow$
Variants on graph retriever:				
w/o path-level	0.3927	0.0868	0.4560	0.0924
w/o node-level	0.3966	0.0894	0.4544	0.0922
w/o GraphRAG	0.3880	0.0896	0.4468	0.0940
Variants on link prediction model:				
w/ LightGCN	0.3941	0.0870	0.4589	0.0922
w/ R-GCN	0.4003	0.0885	0.4592	0.0916
Variants on LLMs with different scales:				
w/ Qwen-0.5B	0.3201	0.6530	0.4129	0.2171
w/ Qwen-1.5B	0.3557	0.3940	0.4451	0.0940
w/ Qwen-3B	0.3994	0.0861	0.4602	0.0903
w/ Qwen-7B	0.3991	0.0851	0.4582	0.0914
Knowledge pruning v.s. full training set:				
w/o pruning	0.4002	0.0892	0.4605	0.0909
G-Refer	0.4003	0.0885	0.4592	0.0916

Hyperparameter & Efficiency

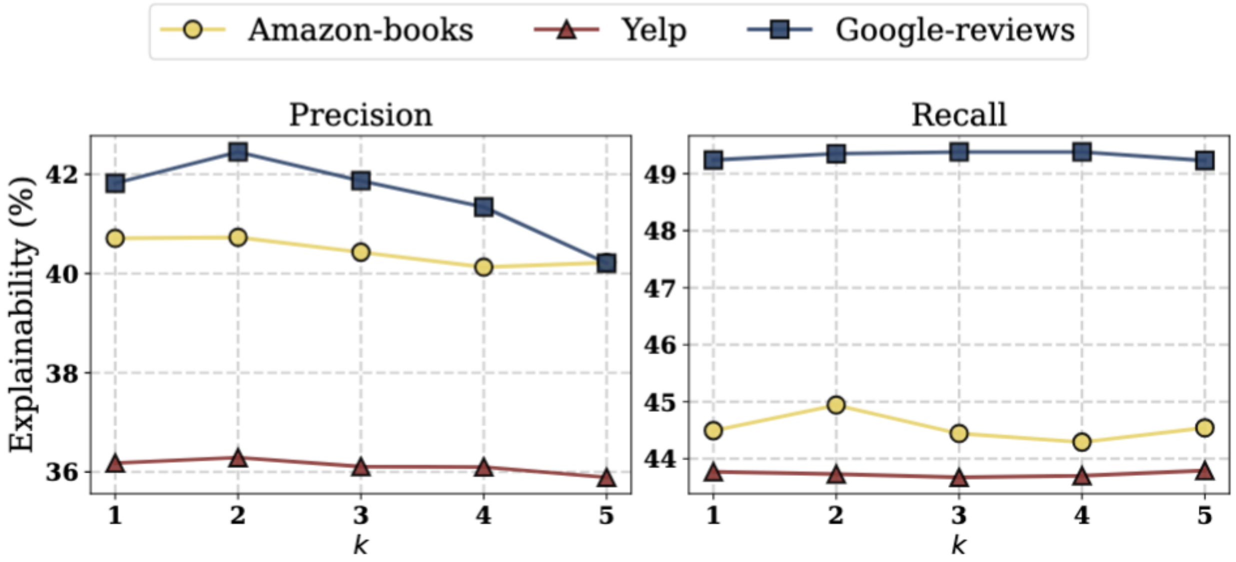


Figure 4: Performance of different retrieved number k.

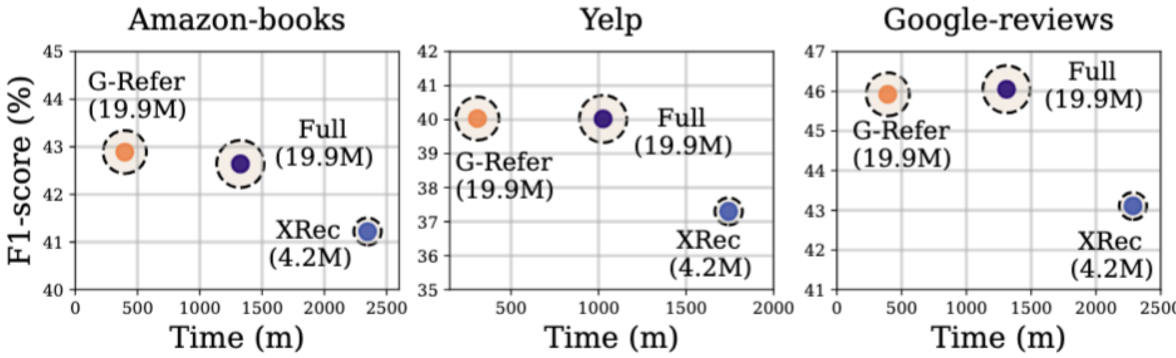


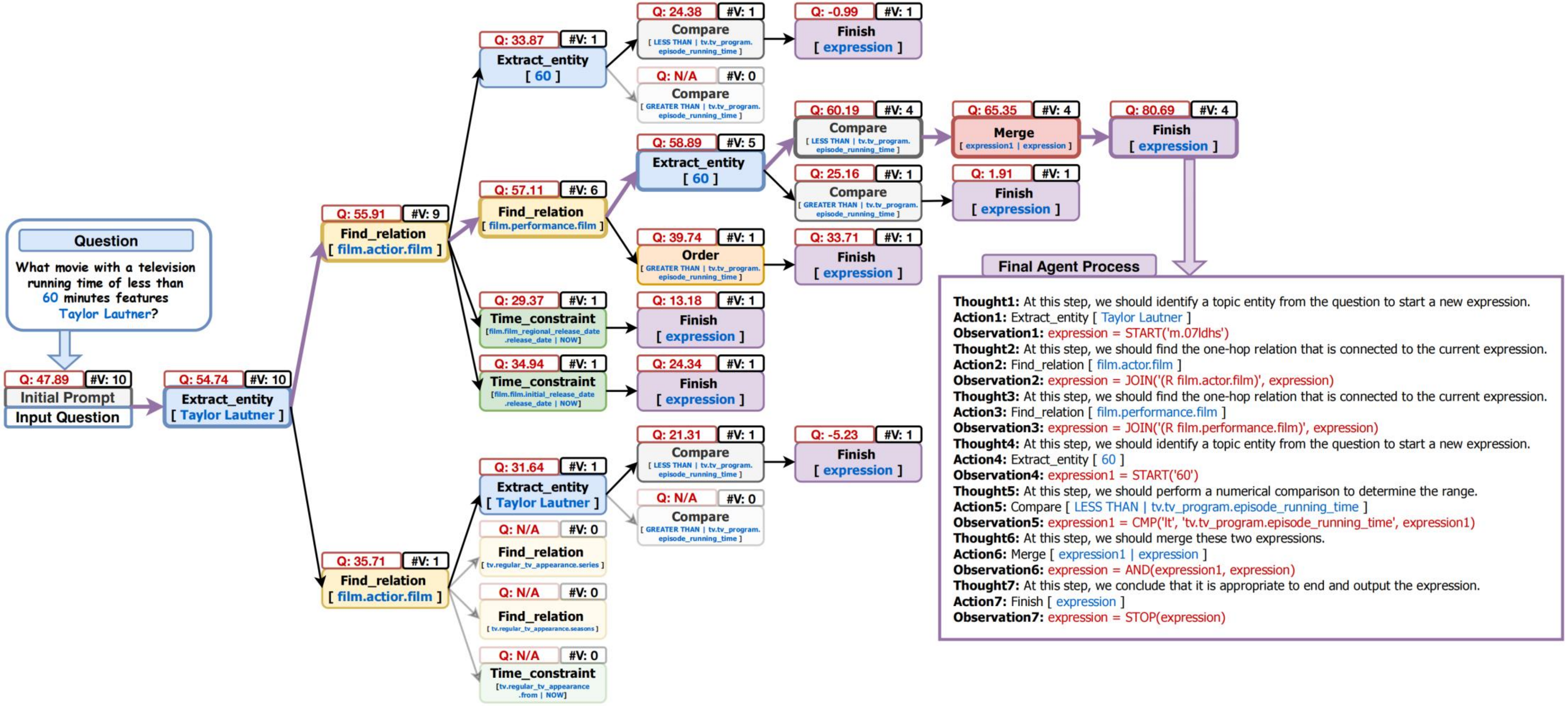
Figure 5: Efficiency Analysis of G-Refer.

Tutorial outline

	Content		Presenter
30 min	•**Introduction and background** • Artificial general intelligence (AGI) • Large language models (LLMs) and knowledge graphs (KGs) • Challenges and opportunities	Part 1	**Shirui Pan**
60 min	•**Knowledge graph-enhanced large language models** • KG-enhanced LLM Training • KG-enhanced LLM Reasoning • Unified KG+LLM Reasoning	Part 2	**Linhao Luo**
30 min break			
60 min	•**Large language model-enhanced knowledge graphs** • LLM-enhanced KG integrations • LLM-enhanced KG construction and completion • LLM-enhanced Multi-modality KG	Part 3	**Carl Yang**
20 min	•**Applications of synergized KG-LLM systems** • QA system • Recommender system	Part 4	**Evgeny Kharlamov**
10 min	•**Future directions and conclusion**	Part 5	**Linhao Luo**

Future Directions – **KG-guided O1 Reasoning**

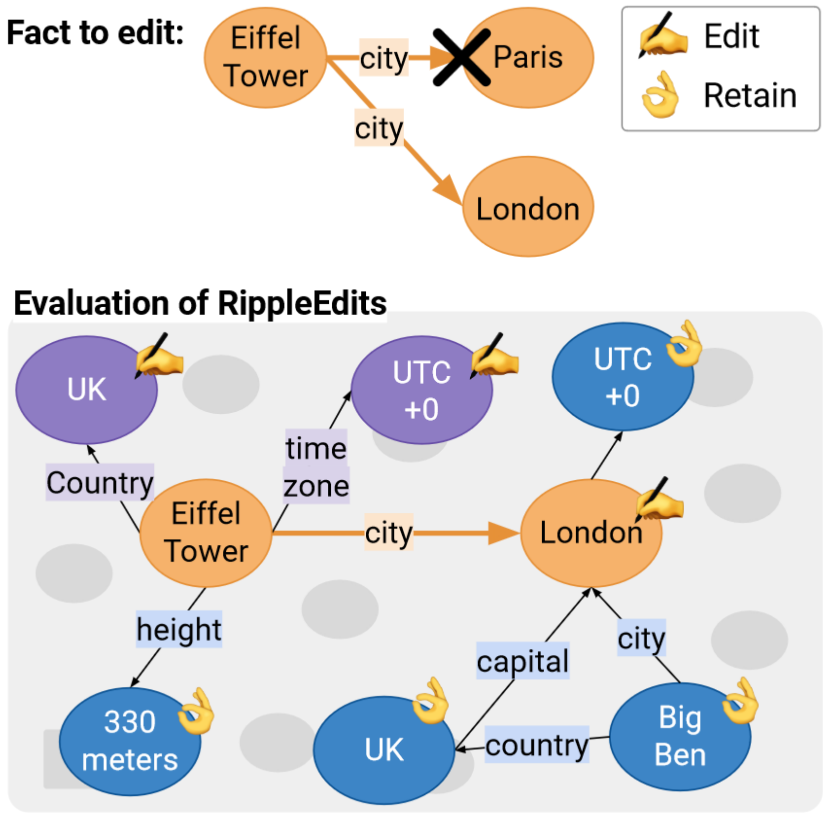
- **O1 reasoning on KGs.**



Luo, Haoran, et al. "KBQA-o1: Agentic Knowledge Base Question Answering with Monte Carlo Tree Search." arXiv preprint arXiv:2501.18922 (2025).

Future Directions – Knowledge Edit

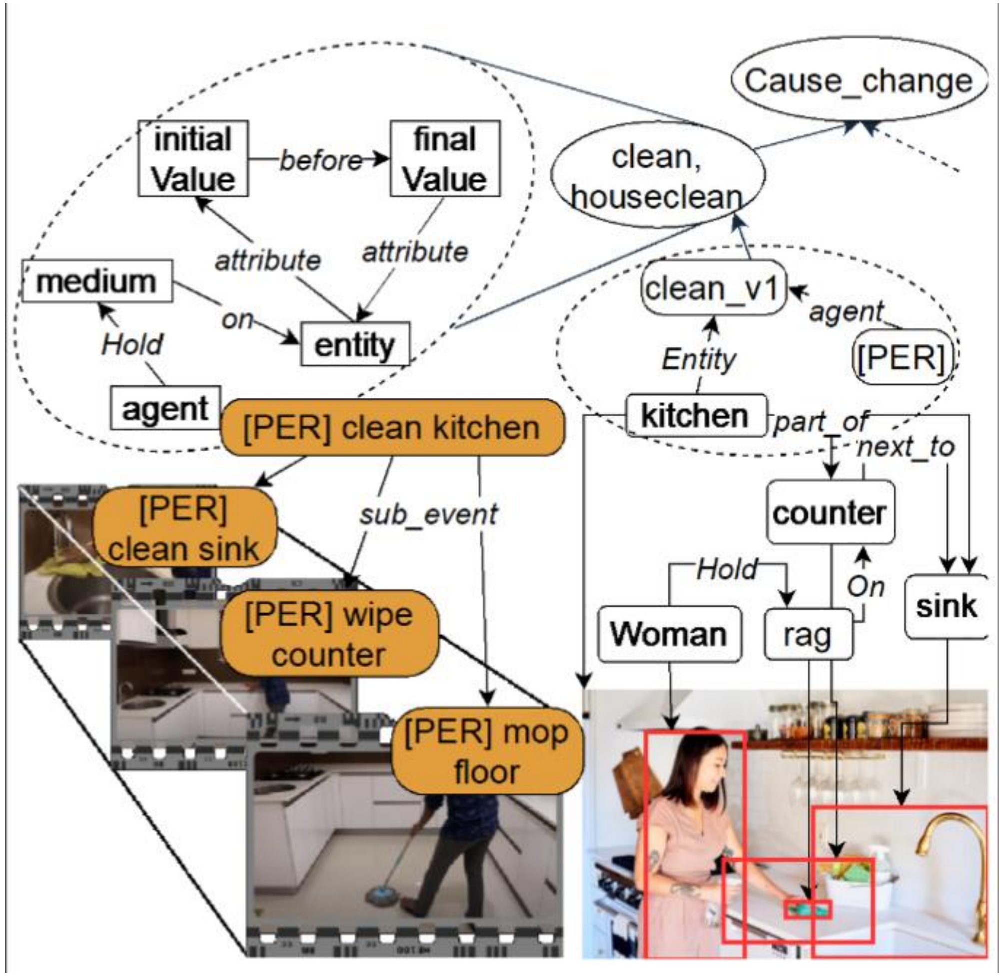
- **KGs for Editing Knowledge in LLMs.**
  - Add new or delete old knowledge stored in LLMs with KGs.



Cohen, R., Biran, E., Yoran, O., Globerson, A., & Geva, M. (2023). Evaluating the ripple effects of knowledge editing in language models. *arXiv preprint arXiv:2307.12976*.

Future Directions – **Multi-modal Reasoning**

- **Multi-Modal KG-enhanced Reasoning**



Zheng, Changmeng, et al. "A Picture Is Worth a Graph: A Blueprint Debate Paradigm for Multimodal Reasoning." ACM MM 2024

Thanks for listening!
Q&A