Collaborative Solutions for Complex Task Reasoning Using Large Models and Knowledge Graphs

Yongrui Chen

Institute of Cognitive Science

Southeast University

Contents

What is Knowledge?

The information, understanding, and skills that you gain through education or experience.

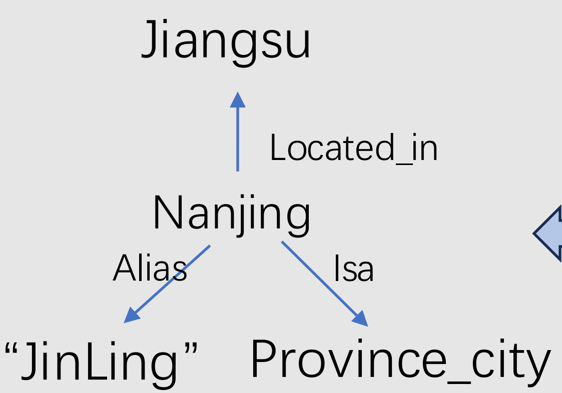
—— Oxford Dictionary

- The ability to learn and apply knowledge is the fundamental ability to determine whether artificial intelligence has human intelligence
- The following can be considered as knowledge
  - **Fact knowledge:** China is a country
  - **Description of information:** text or image
  - **Skills obtained by practice:** skill to open a bottle
- **Knowledge Base (KB):** a collection of knowledge, including documents, images, triples, rules or parameters of neural networks, etc.

Knowledge Graph

A knowledge graph (KG) is a data structure for representing knowledge using a graph

- Nodes in the graph can be either entities or literals
- Edges are relations between entities and entities or literals
- Semantics of KG is based on ontology languages such as RDFS[1] or OWL[2]



subgraph

<Subject, Predicate, Object>

<Nanjing, Located_in, Jiangsu>
<Nanjing, Isa, Province_city>

Triple

Knowledge Graph

DBpedia

WIKIDATA

Freebase

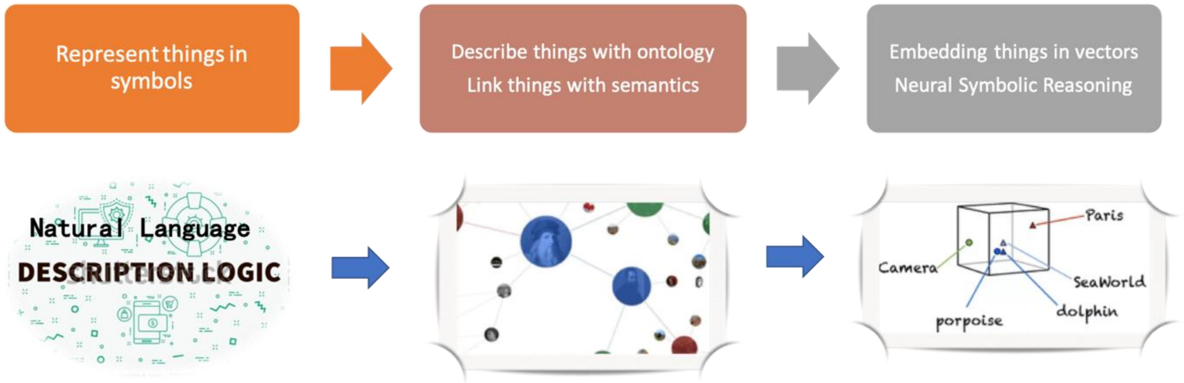
yago select knowledge

famous KGs

1. https://www.w3.org/TR/rdf-schema/
2. https://www.w3.org/OWL/

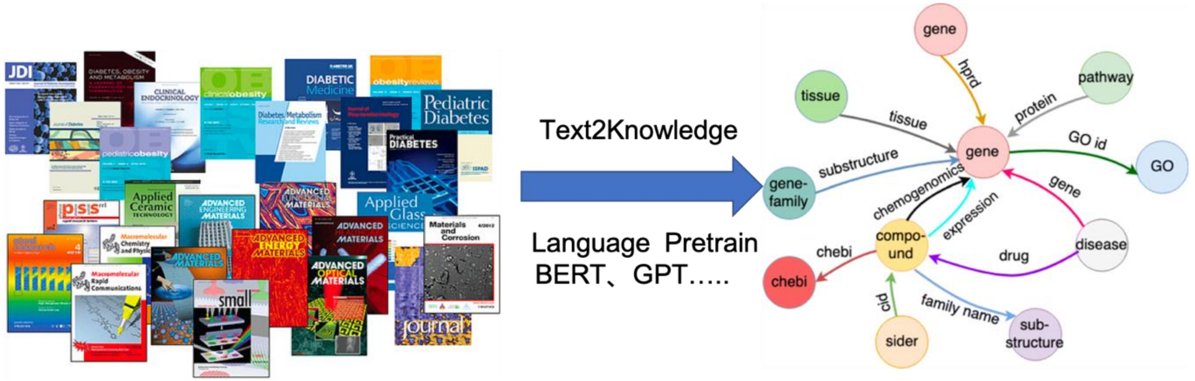
KG as Knowledge Base

KG as a World Model

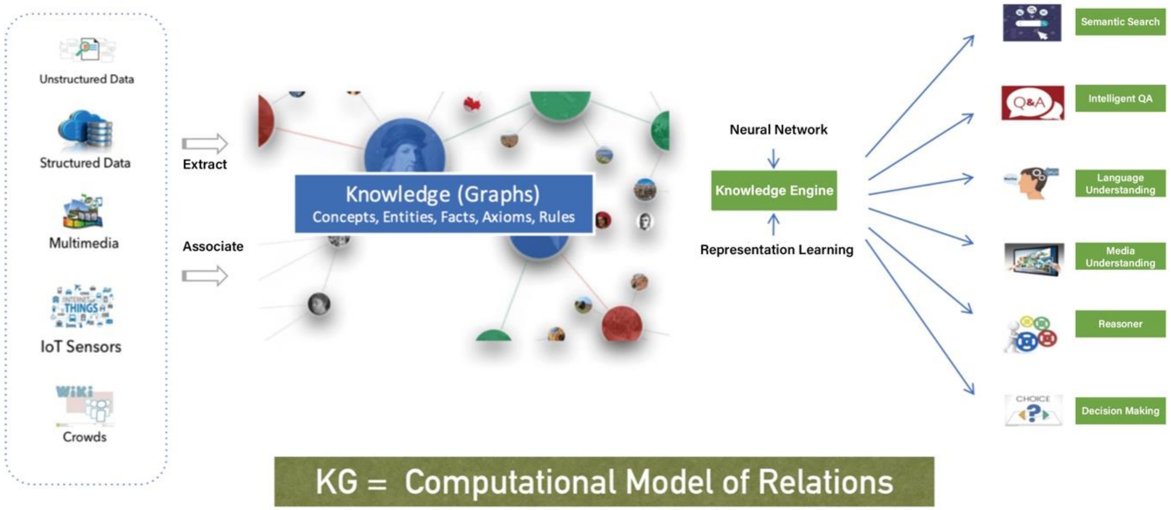


The Good Old Fashioned AI   The Semantic Web & Linked Knowledge   The Knowledge Graph

Text to Knowledge Graph



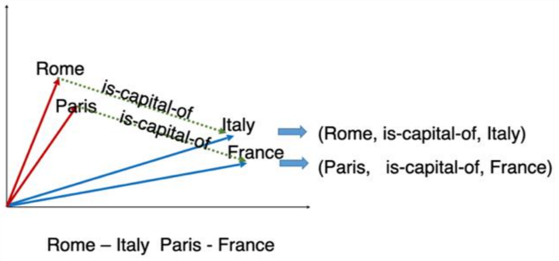
Graph Structure as Knowledge Base



KG Embeddings as Knowledge Base

Embeddings：Distributed Vector Representation

- Text：Learn a vector of each word in a sentence
- KG: Learn a vector for each entity or property
- Image/Video：Learn a vector for each visual object

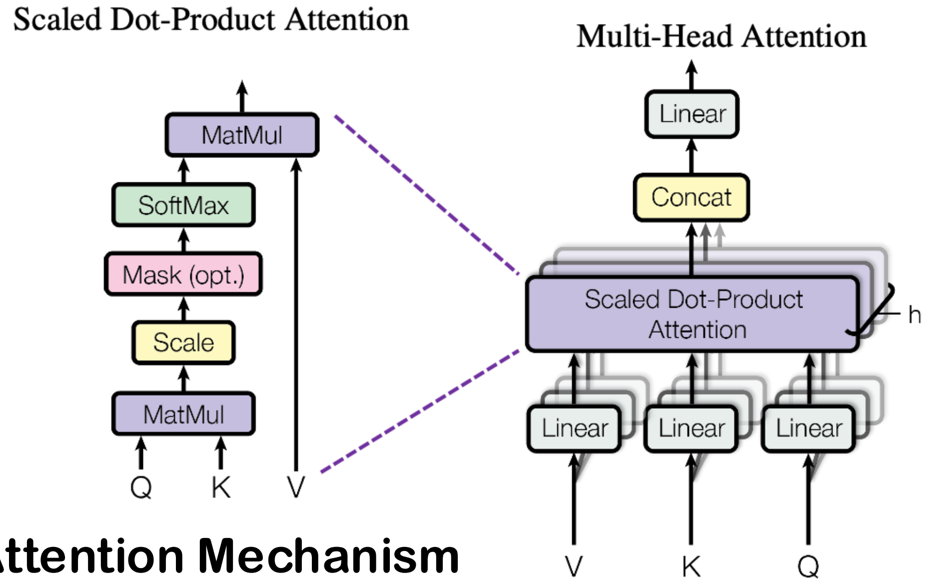


What is Language Model?

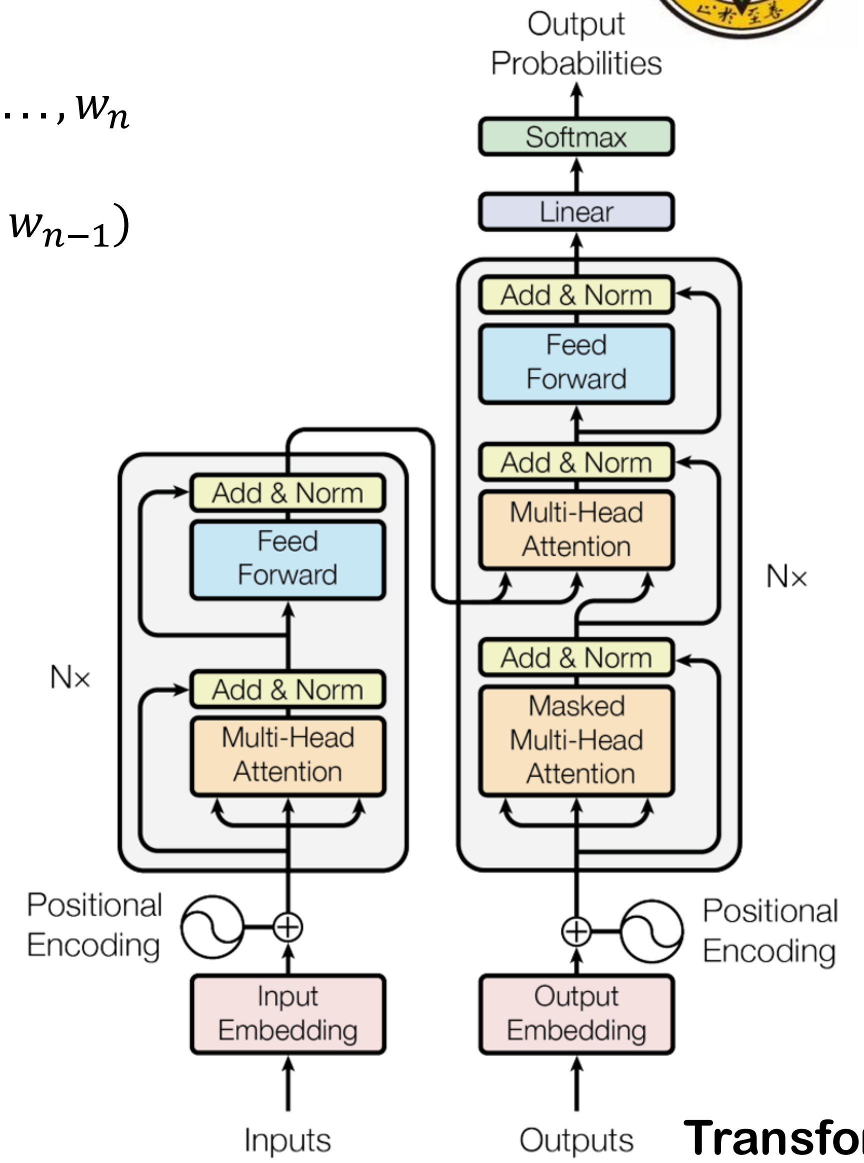
Calculate the probability of a word sequence： $w_1, w_2, \ldots, w_n$

$$P(w_1, w_2, \ldots, w_n) = P(w_1) \times P(w_2|w_1) \times \ldots \times P(w_n|w_1, \ldots, w_{n-1})$$

- **Transformer**, a most popular neural network；
- Encoder – Decoder architecture；
- Attention Mechanism；



Attention Mechanism

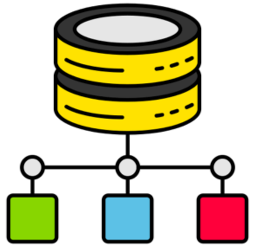


Transformer

Pre-training & Large Language Model

Pre-training

Train the model (Transformer) on a generic large-scale dataset to learn some fundamental, common features or patterns.



Pre-training

Transformer

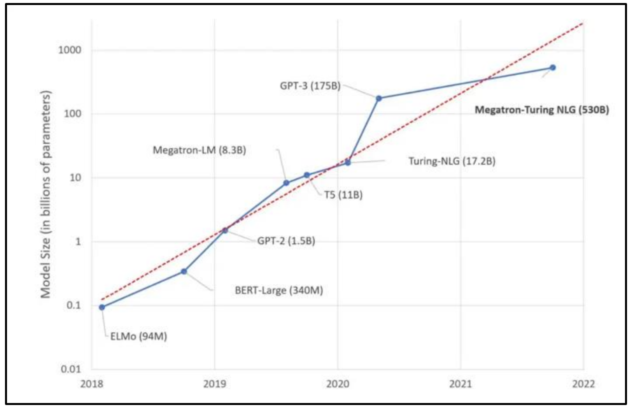
Training Object

Casual Language Model

Predict the n-th word using the previous $n-1$ words.

Large Language Model (LLM)

As the number of parameters gradually increases, when it reaches a certain scale (typically over one billion), it is referred to as an LLM.



Pre-trained LM → LLM

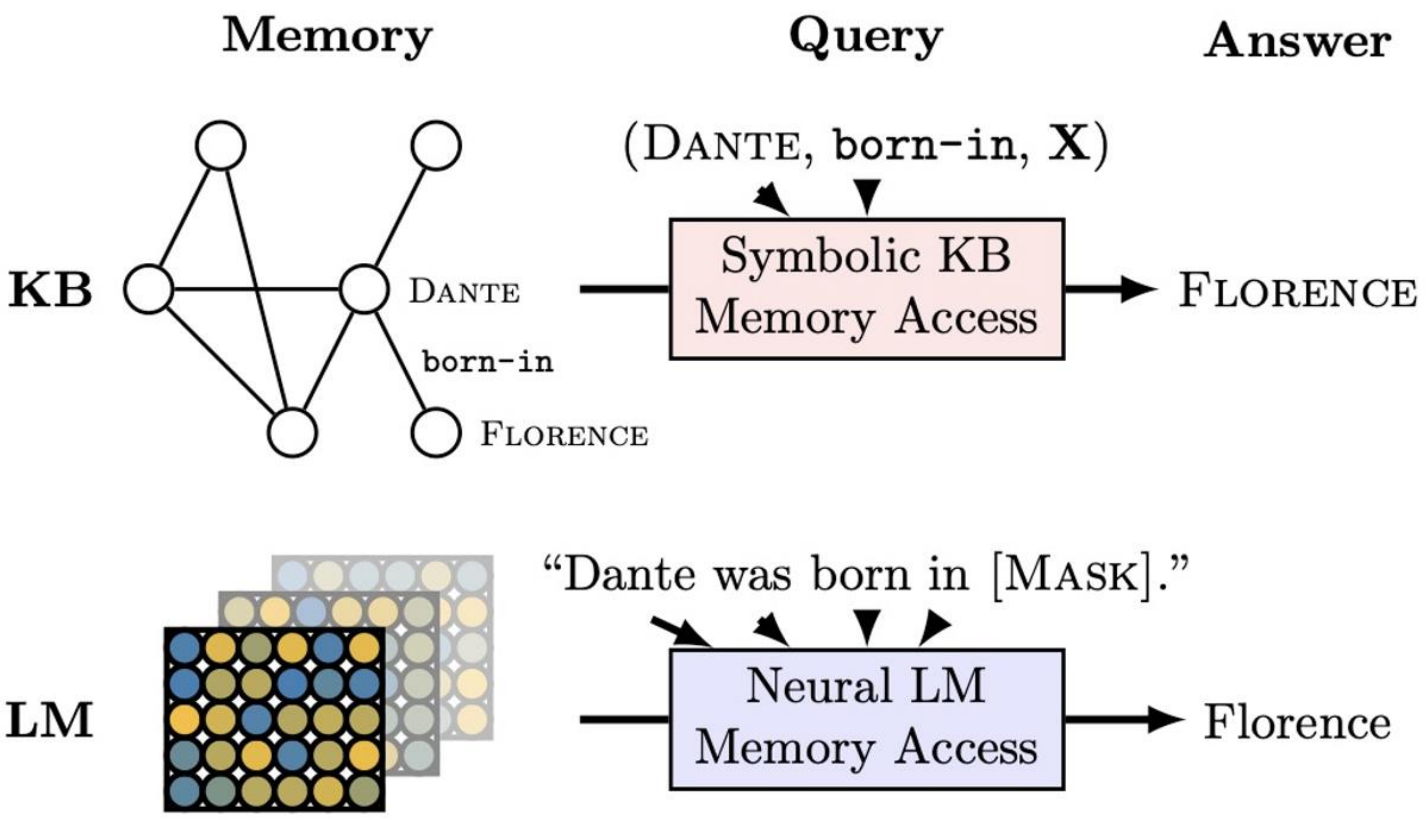
ELMo (94M), BERT (340M) …

GPT-3 (175B),Llama 2 (70B) …

As the scale of model increases, the performance of the model significantly improves!

LLM as Knowledge Base

- An LLM is a parametric knowledge base



Fabio Petroni, et.al., Language Models as Knowledge Bases? EMNLP-IJCNLP, 2019.

KG vs LLM: Reasoning Capability Comparison

LLM Reasoning

- **Code Pre-training:** enhance LLM reasoning during training
- **Prompt Engineering:** eliciting LLM reasoning during inference

KG Reasoning

- Graph computing
- Rule-based reasoning
- Ontology reasoning
- Spatial-temporal reasoning
- KG embedding/GNN

LLM Reasoning

- zero-shot prompting
- Few-shot prompting
- CoT prompting
- Instruction

+

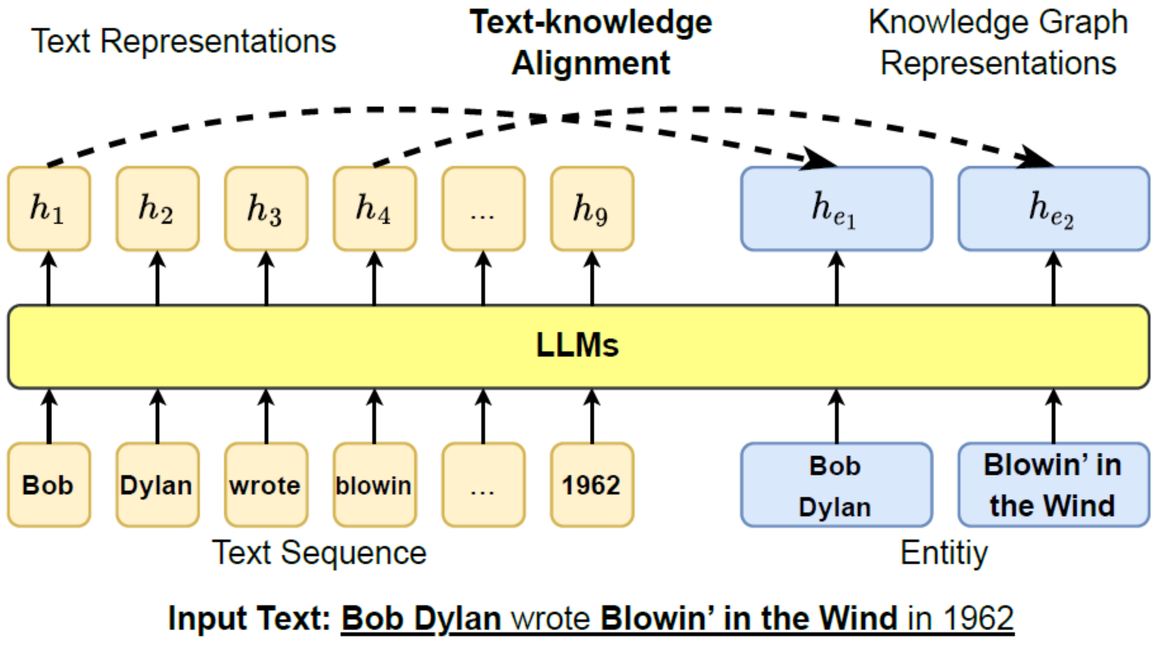
KG Reasoning

- Graph computing
- Rule-based reasoning
- Ontology reasoning
- Spatial-temporal reasoning
- KG embedding/GNN

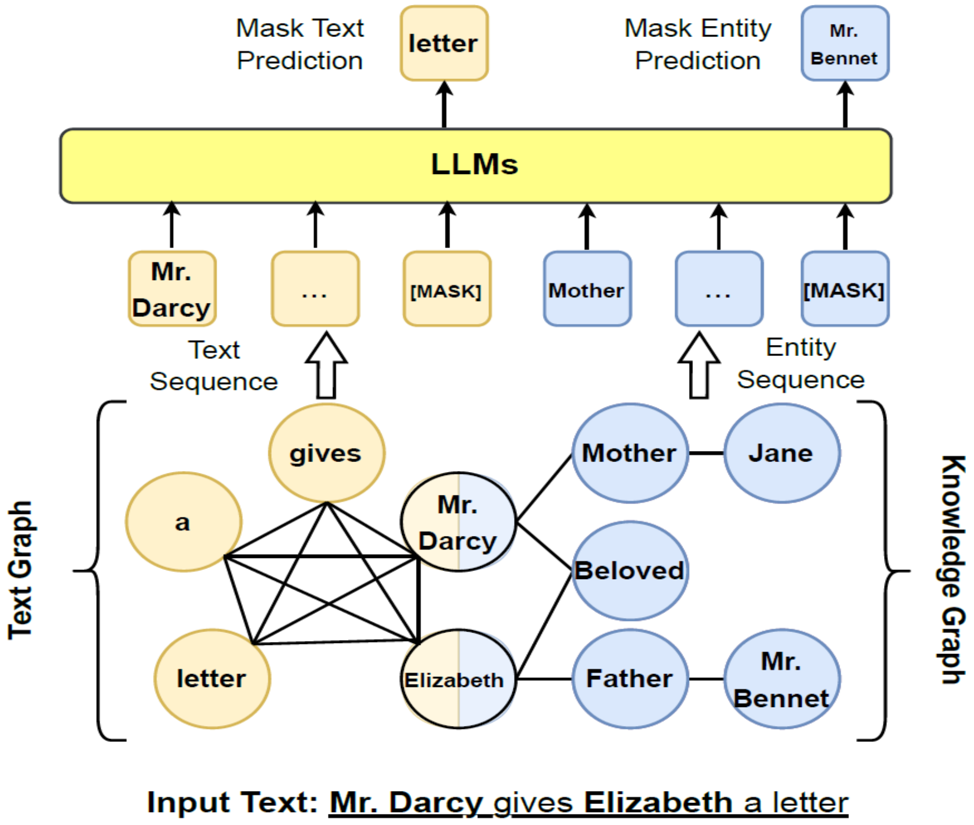
Contents

KG for LLM: Pre-training

- Designing pre-training objective to incorporate KG components
- Integrate KG with text as LLM training input



Aligned Pre-training Object (ERNIE …)

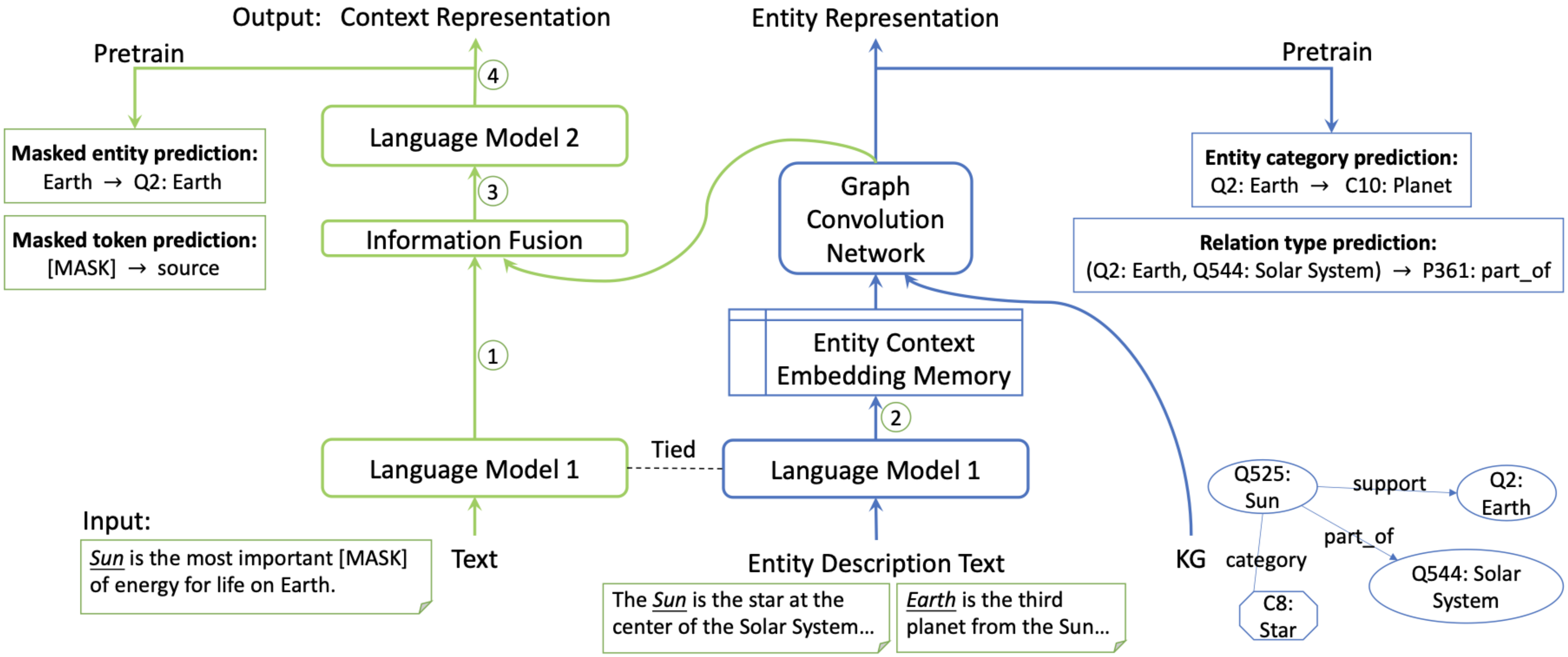


Combined Training Input (CoLake…)

ERNIE: Enhanced language representation with informative entities, ACL 2019.
CoLAKE: Contextualized language and knowledge embedding, 2020.

KG for LLM: Pre-training

- Integrating KGs into additional fusion modules

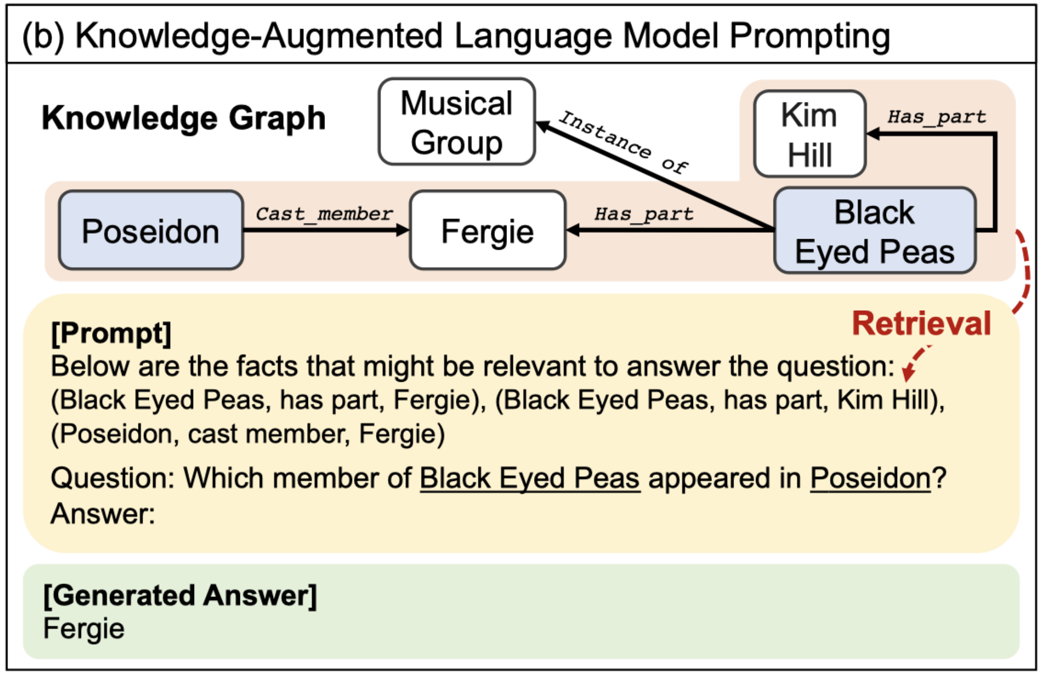


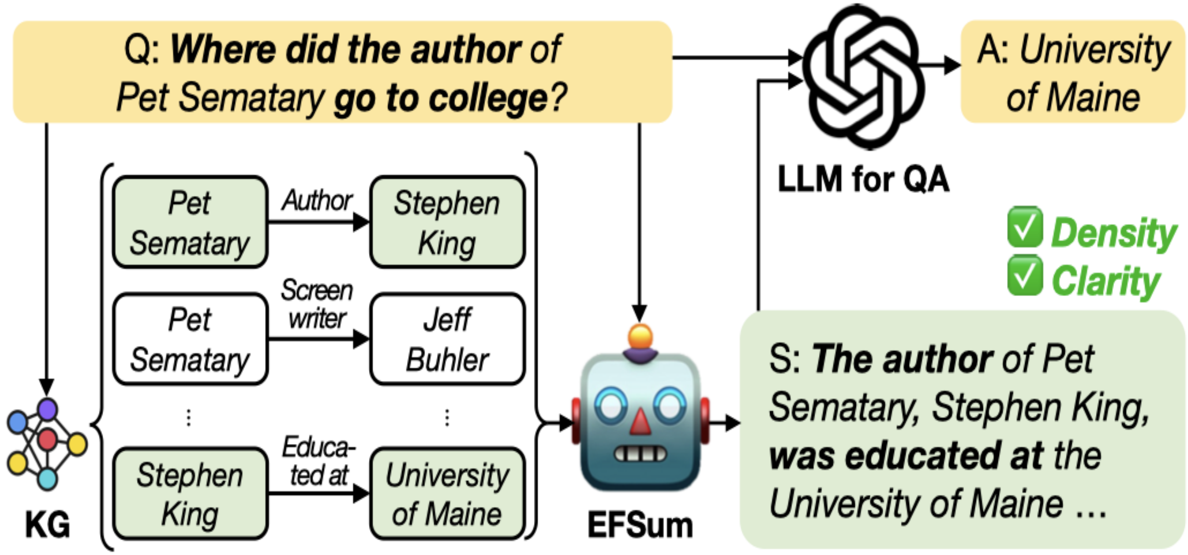
Fused module (JAKET ...)

JAKET: Joint Pre-training of Knowledge Graph and Language Understanding, AAAI 2022.

KG for LLM: KG as Prompt

- Knowledge graphs are directly utilized by LLMs as prompts without training





KAPING

Retrieve subgraph triples as prompt

EFSum

Summarize the related triples

Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering. ACL 2023 workshop
Evidence-Focused Fact Summarization for Knowledge-Augmented Zero-Shot Question Answering. Preprint, 2024

KG for LLM: KG as Prompt

- Experimental results of KAPING

Table 1: **Main results of language model prompting**, where we report the generation accuracy. The number inside the parentheses in the first row denotes the parameter size of language models, and best scores are emphasized in bold.

Datasets	Methods	T5 (0.8B)	T5 (3B)	T5 (11B)	OPT (2.7B)	OPT (6.7B)	OPT (13B)	T0 (3B)	T0 (11B)	GPT-3 (6.7B)	GPT-3 (175B)	AlexaTM (20B)	Average
WebQSP w/ Freebase	No Knowledge	6.95	13.40	9.48	19.85	29.77	28.38	21.43	40.77	44.63	63.59	46.79	29.55
	Random Knowledge	21.55	19.15	17.57	28.07	31.73	33.31	32.62	51.20	51.01	65.87	57.37	37.22
	Popular Knowledge	15.30	16.88	18.39	28.32	28.13	24.21	27.05	47.22	45.58	62.26	54.91	33.48
	Generated Knowledge	6.19	7.84	6.76	7.46	11.50	8.22	19.41	38.81	45.89	62.14	35.13	22.67
	KAPING (Ours)	**34.70**	**25.41**	**24.91**	**41.09**	**43.93**	**40.20**	**52.28**	**62.85**	**60.37**	**73.89**	**67.67**	**47.94**
WebQSP w/ Wikidata	No Knowledge	10.30	18.42	15.21	23.94	33.77	32.40	24.56	44.20	48.50	67.60	42.41	32.85
	Random Knowledge	17.94	22.78	24.28	37.24	35.61	38.27	28.85	47.68	52.05	60.64	55.63	38.27
	Popular Knowledge	15.35	20.80	20.74	30.83	30.01	27.83	24.83	48.02	47.41	63.37	53.92	34.83
	Generated Knowledge	11.94	13.30	12.28	11.26	17.53	14.19	22.92	41.34	48.77	65.89	31.16	26.42
	KAPING (Ours)	**23.67**	**40.38**	**35.47**	**49.52**	**53.34**	**51.57**	**49.86**	**58.73**	**60.44**	**69.58**	**65.04**	**50.69**
Mintaka w/ Wikidata	No Knowledge	11.23	14.25	17.06	19.76	27.19	26.83	14.75	23.74	34.65	56.33	41.97	26.16
	Random Knowledge	17.59	18.19	18.83	28.11	26.58	28.36	16.10	26.15	32.98	51.56	46.02	28.22
	Popular Knowledge	17.56	18.09	18.73	26.97	27.08	23.10	16.74	27.15	32.48	53.16	46.41	27.95
	Generated Knowledge	13.61	14.61	14.29	11.87	14.96	16.24	14.46	23.13	33.12	55.65	34.58	22.41
	KAPING (Ours)	**19.72**	**22.00**	**22.85**	**32.94**	**32.37**	**33.37**	**20.68**	**29.50**	**35.61**	**56.86**	**49.08**	**32.27**

KG for LLM: KG as Prompt

- Experimental results of EFSUM

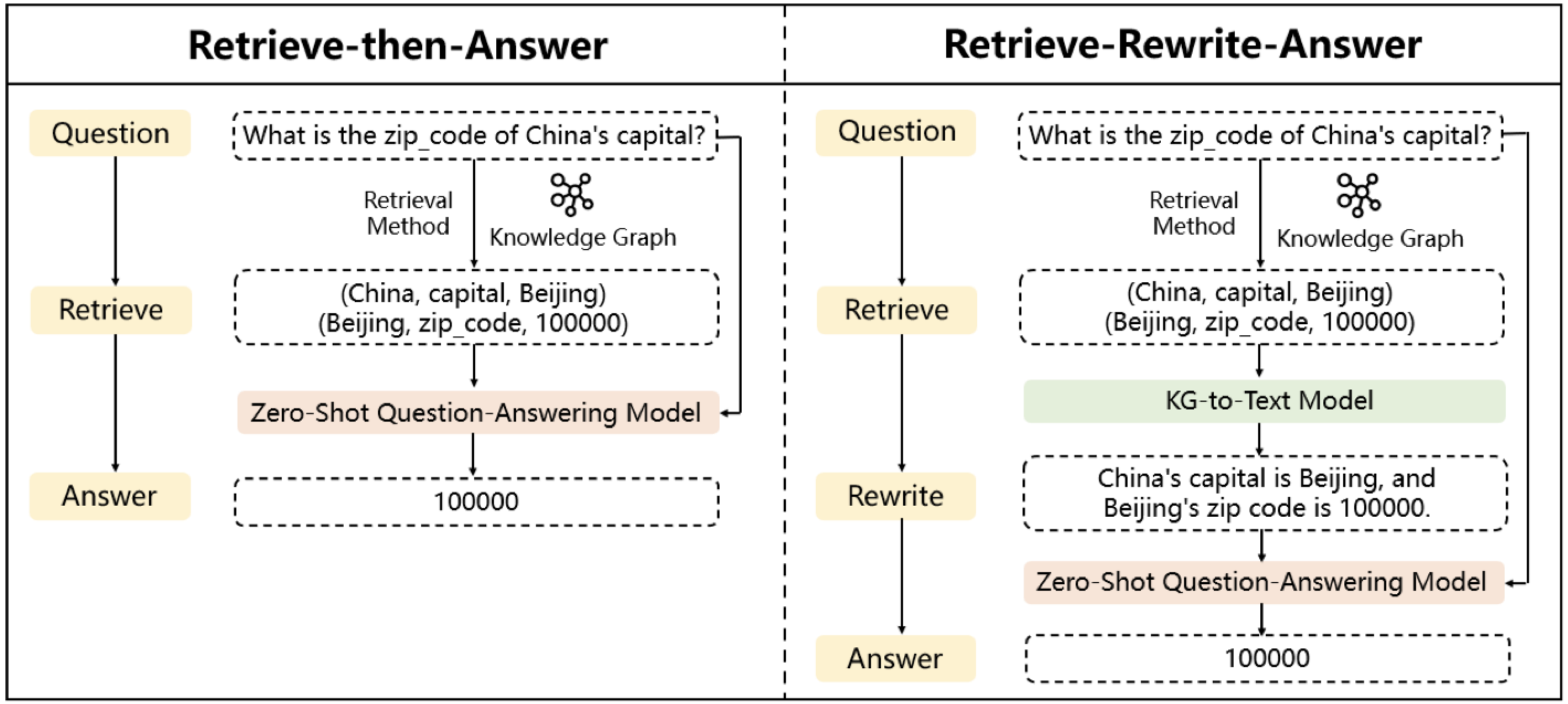
Datasets	Methods	GPT-3.5-turbo			Flan-T5-XL			Llama2-7B-Chat		
		Random	Popular	MPNet	Random	Popular	MPNet	Random	Popular	MPNet
WebQSP	No knowledge	0.506	0.506	0.506	0.409	0.409	0.409	0.539	0.539	0.539
	KAPING (Baek et al., 2023a)	0.441	0.437	0.538	0.297	0.329	0.439	0.476	**0.490**	**0.519**
	KG2Text (Ribeiro et al., 2021)	0.469	0.468	0.476	0.317	0.276	0.321	0.465	0.451	0.481
	Rewrite (Wu et al., 2023)	0.473	0.445	0.525	0.323	0.348	0.431	0.458	0.439	0.511
	EFSUM$_{prompt}$ (Ours)	**0.542**	0.534	0.538	0.443	0.442	0.468	**0.477**	0.472	0.491
	EFSUM$_{distill}$ (Ours)	0.475	**0.539**	**0.569**	**0.500**	**0.505**	**0.500**	0.457	0.488	0.497
Mintaka	No knowledge	0.540	0.540	0.540	0.228	0.228	0.228	0.440	0.440	0.440
	KAPING (Baek et al., 2023a)	**0.553**	0.516	**0.539**	0.201	0.198	0.279	0.417	**0.398**	0.407
	KG2Text (Ribeiro et al., 2021)	0.505	0.500	0.492	0.220	0.235	0.234	**0.421**	0.389	0.378
	Rewrite (Wu et al., 2023)	0.527	**0.524**	0.515	0.230	0.224	0.288	0.393	0.374	0.386
	EFSUM$_{prompt}$ (Ours)	0.454	0.492	0.496	0.213	0.215	0.321	0.390	0.392	**0.418**
	EFSUM$_{distill}$ (Ours)	0.427	0.425	0.474	**0.292**	**0.243**	**0.338**	0.397	0.393	0.406

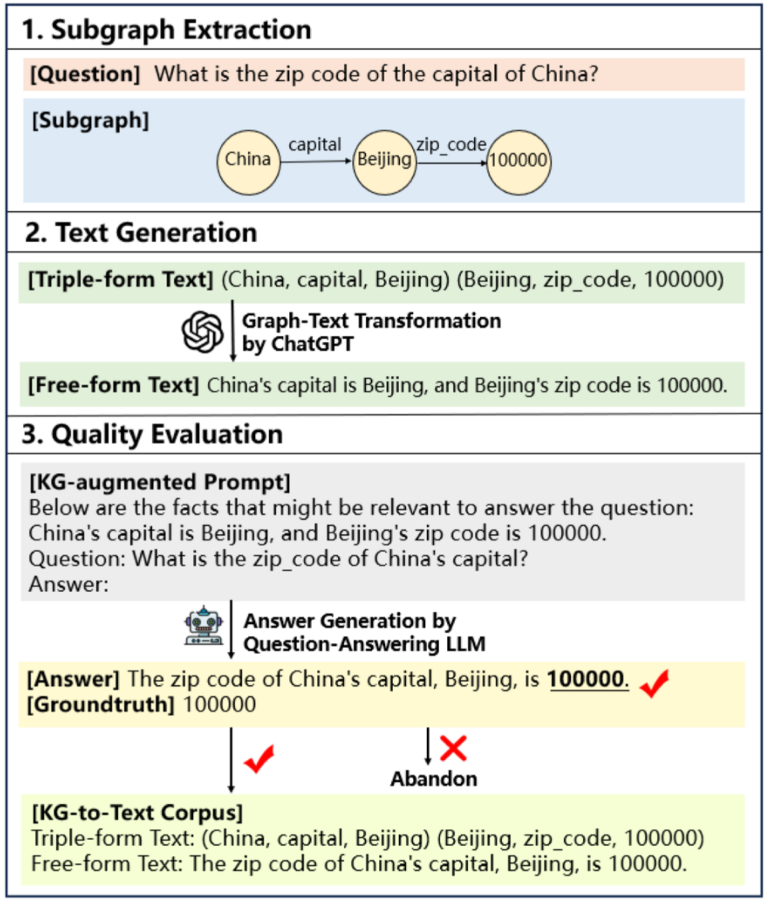
Table 2: QA accuracy of the LLMs based on various fact verbalization, with different fact retrieval strategies (i.e., random facts, popular facts, and question-relevant facts). We limit the maximum token length of contextual knowledge to $L = 400$. The best and second-best results are in **bold** and underlined, respectively.

Evidence-Focused Fact Summarization for Knowledge-Augmented Zero-Shot Question Answering Preprint, 2024

KG for LLM: KG-to-text Prompt

- Transform KG knowledge into well-textualized statements most informative

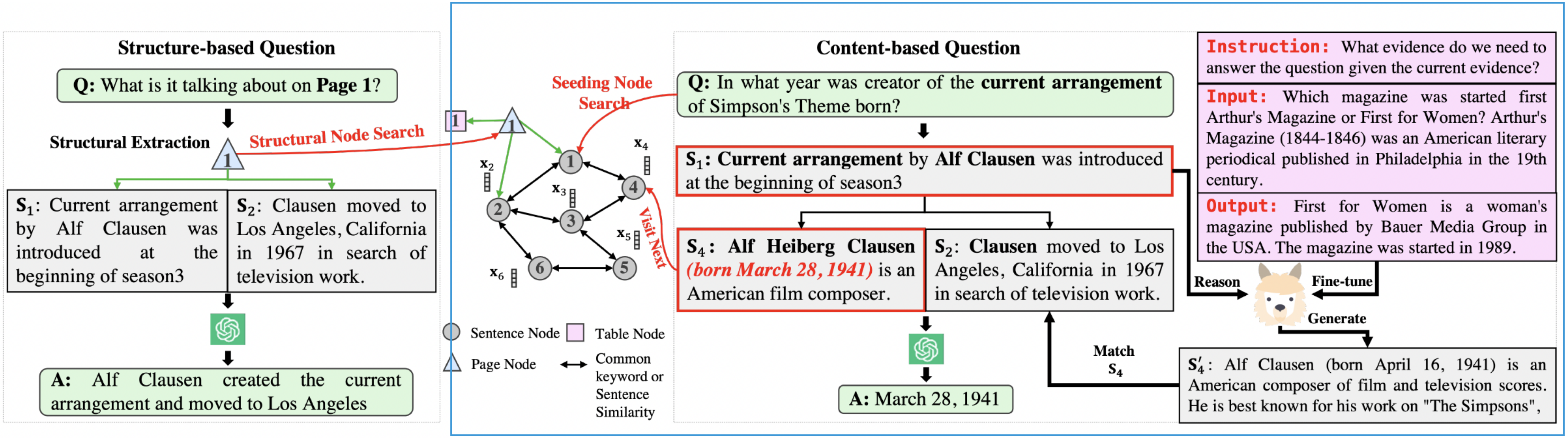




Retrieve-Rewrite-Answer: A KG-to-Text Enhanced LLMs Framework for Knowledge Graph Question Answering. IJCKG 2023

KG for LLM: Enhanced LLM Reasoning

- Knowledge graph prompting for LLM reasoning on multi-documents



Knowledge graph prompting (KGP)

For questions on document content, concatenate it with the currently retrieved context and prompt the LLM to generate the next evidence to answer the question.

Knowledge graph prompting for multi-document question answering. AAAI 2024

KG for LLM: Enhanced LLM Reasoning

- Experimental results of KGP

Method	HotpotQA			IIRC			2WikiMQA			MuSiQue			PDF-T	Rank	
	Acc	EM	F1	Acc	EM	F1	Acc	EM	F1	Acc	EM	F1	Struct-EM	w PDF-T	w/o PDF-T
None	41.80	19.00	30.50	19.50	8.60	13.17	44.40	18.60	25.07	30.40	4.60	10.58	0.00	8.53	9.00
KNN	71.57	40.73	57.97	43.82	25.15	37.24	52.40	31.20	42.13	44.70	18.86	30.04	–	7.00	7.33
TF-IDF	**76.64**	<u>45.97</u>	64.64	47.47	27.22	40.80	58.40	34.60	44.50	44.40	21.59	32.50	–	4.85	5.00
BM25	71.95	41.46	59.73	41.93	23.48	35.55	55.80	30.80	40.55	44.47	21.11	31.15	–	6.92	7.25
DPR	73.43	43.61	62.11	48.11	26.89	<u>41.85</u>	62.40	35.60	51.10	44.27	20.32	31.64	–	5.31	5.50
MDR	75.30	45.55	<u>65.16</u>	**50.84**	<u>27.52</u>	**43.47**	<u>63.00</u>	36.00	<u>52.44</u>	<u>48.39</u>	<u>23.49</u>	<u>37.03</u>	–	<u>3.07</u>	<u>3.08</u>
IRCoT	74.36	45.29	64.12	49.78	**27.73**	41.65	61.81	37.75	50.17	45.14	22.46	34.21	–	4.00	4.08
KGP-T5	<u>76.53</u>	**46.51**	**66.77**	48.28	26.94	41.54	**63.50**	**39.80**	**53.50**	**50.92**	**27.90**	**41.19**	**67.00**	**2.69**	**2.75**
Golden	82.19	50.20	71.06	62.68	35.64	54.76	72.60	40.20	59.69	57.00	30.60	47.75	100.00	1.00	1.00

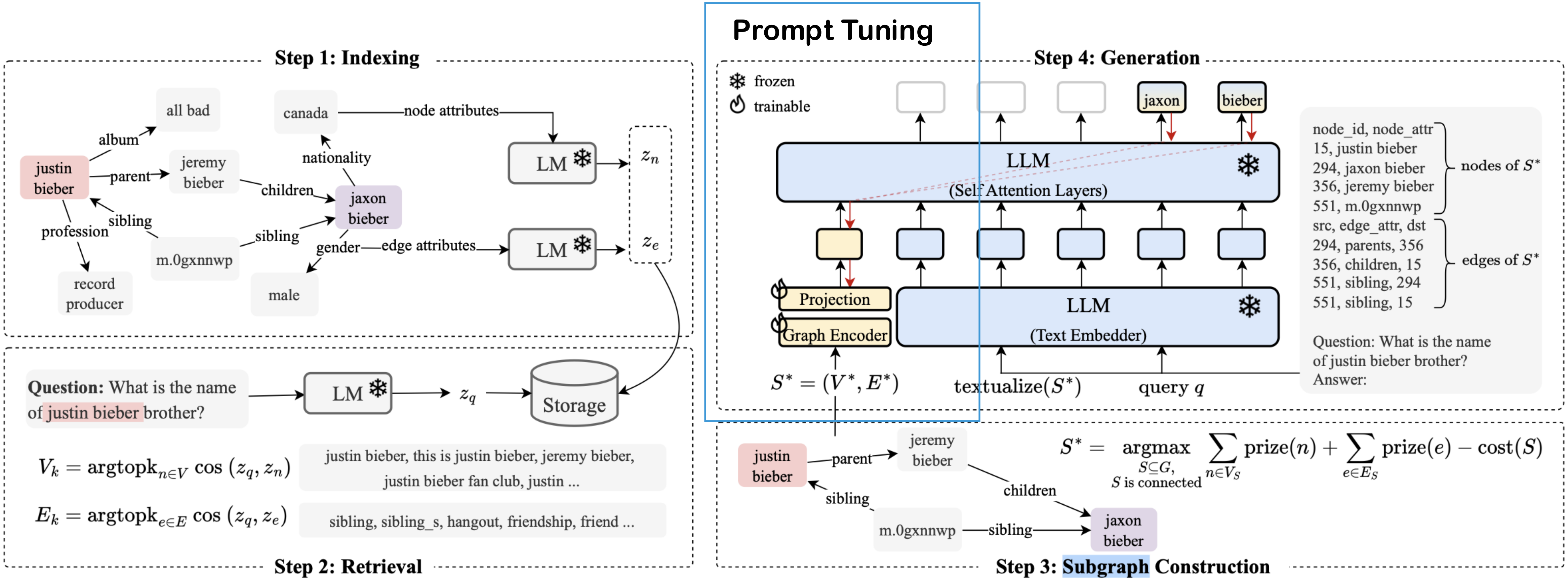
Table 1: MD-QA Performance (%) of different baselines. The best and runner-up are in bold and underlined. None: no passages but only the question is provided. Golden: supporting facts are provided along with the question. PDF-T stands for PDFTriage.

Knowledge graph prompting (KGP)

Knowledge graph prompting for multi-document question answering. AAAI 2024

KG for LLM: Enhanced RAG

- KG can help LLMs reduce hallucinations with Retrieval Augment Generation (RAG).



Prize-Collecting Steiner Tree (PCST)

G-retriever

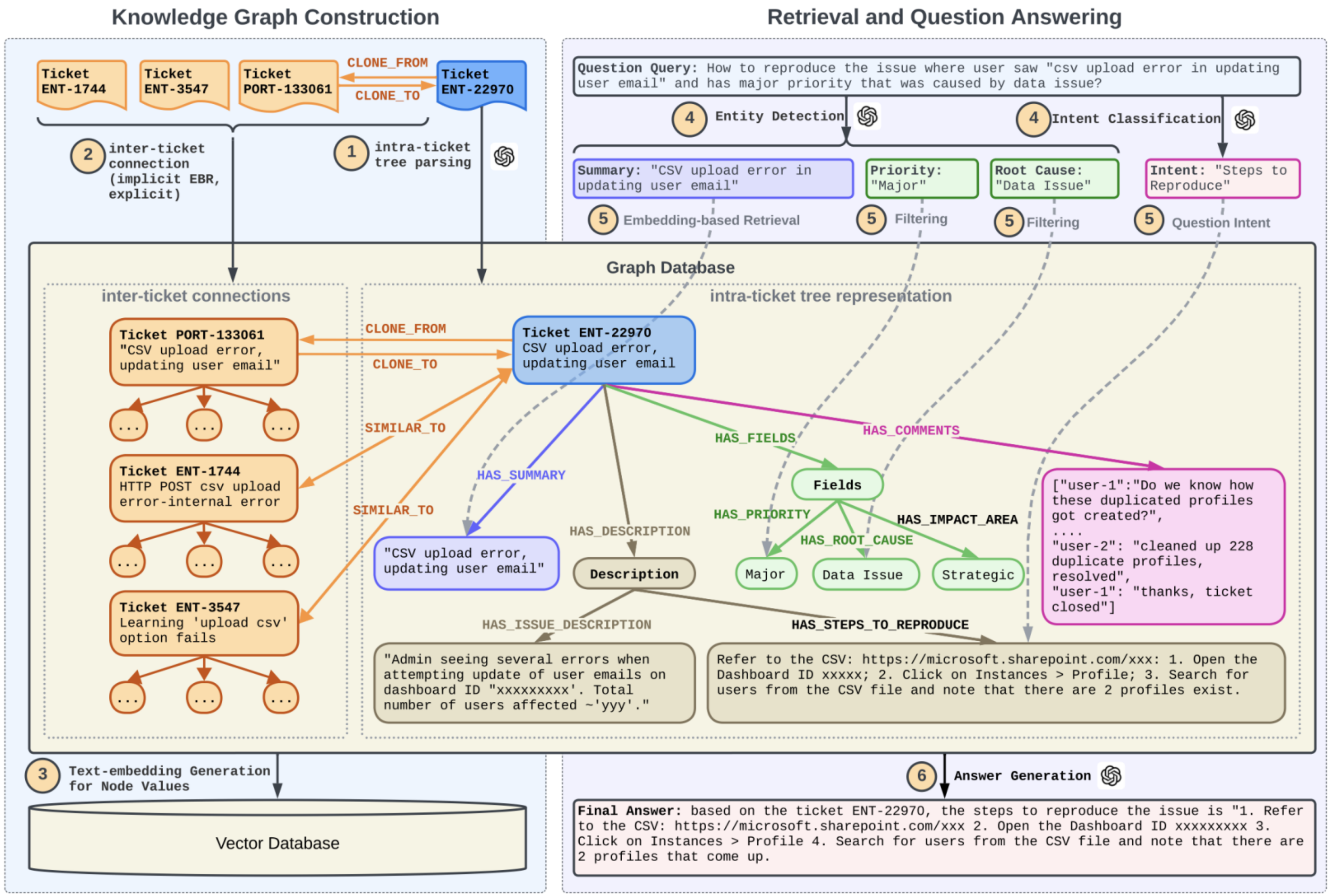
G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering . Preprint 2024.

KG for LLM: Enhanced RAG

- RAG on KG is more likely to capture intra-question structure and inter-question relationships

a) Build an KG from historical records.

b) Parsing consumer queries to identify named entities and intents. then navigates within the KG to identify related sub-graphs for generating answers

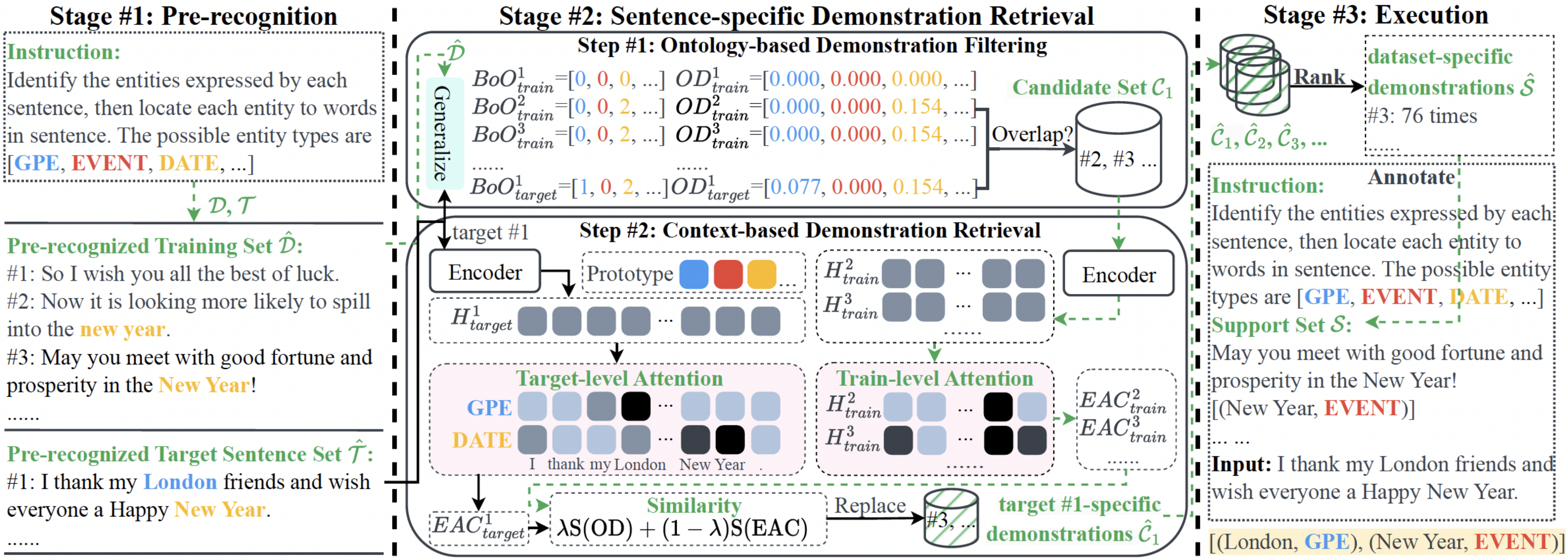


Retrieval-augmented generation with knowledge graphs for customer service question answering. SIGIR 2024.

KG for LLM: Enhanced ICL

- KG can help retrieve high-correlated demonstrations during inference for In-Context Learning (ICL).

ConsistNER

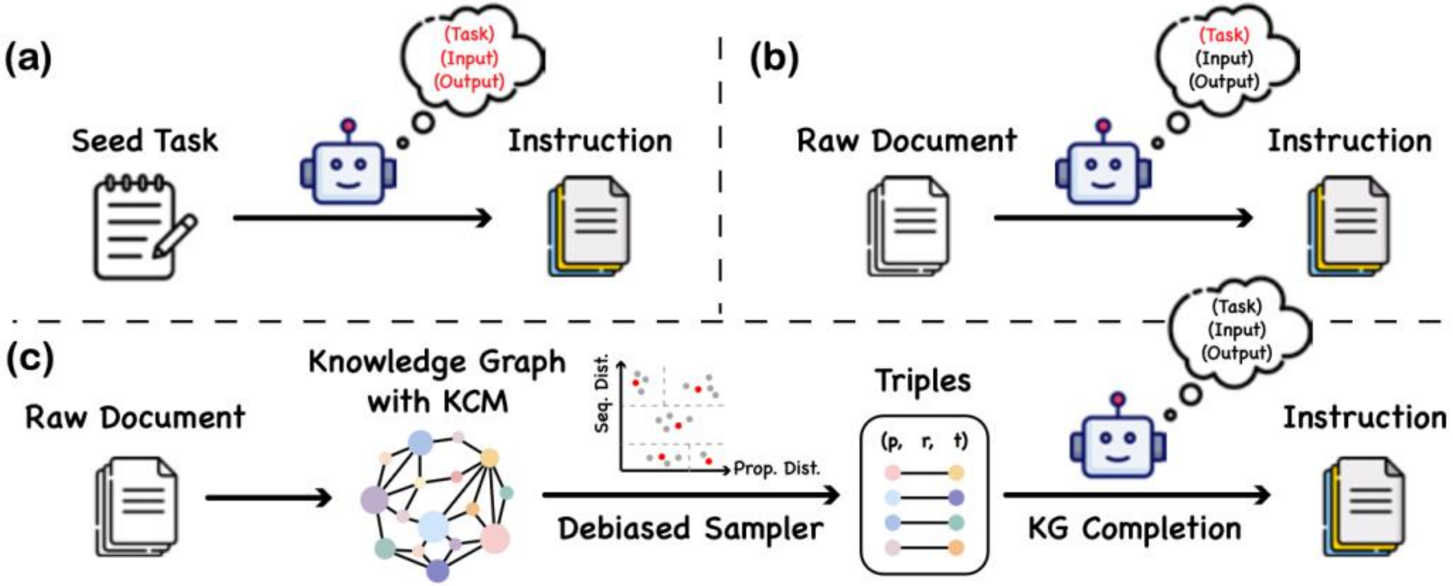


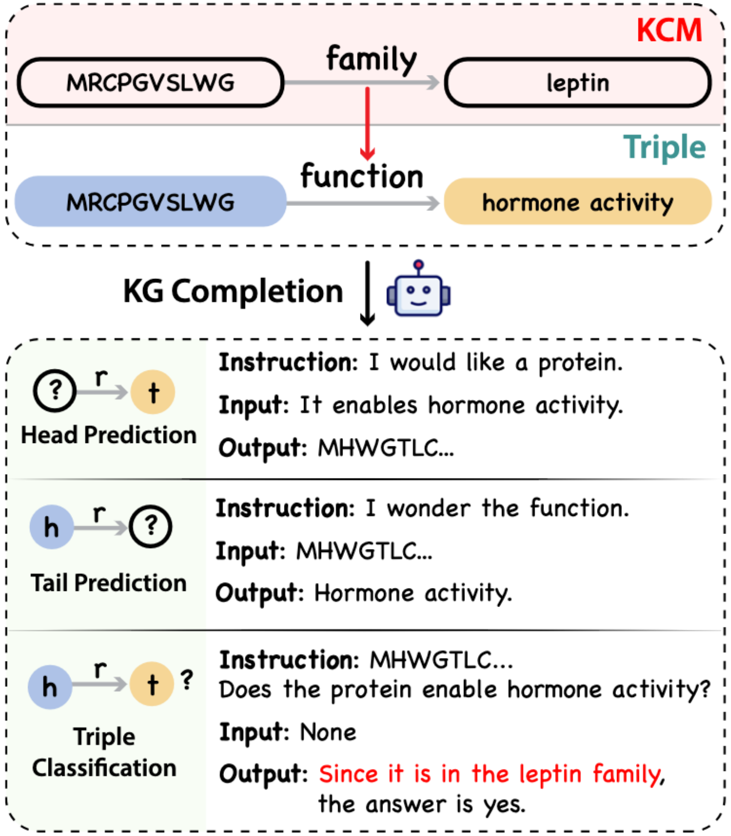
ConsistNER: Towards Instructive NER Demonstrations for LLMs with the Consistency of Ontology and Context. AAAI 2024.

KG for LLM: Instruction Construction

- KG can guide the construction of instruction datasets.

Knowledge to Instruction



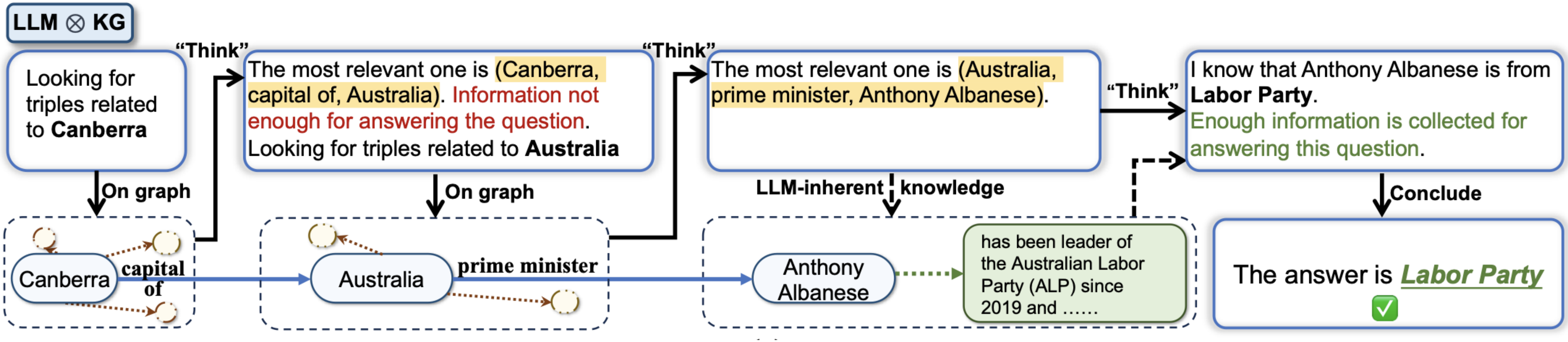


Using an LLM cooperated with KG completion tasks, to generate factual, logical, and diverse instructions.

InstructProtein: Aligning Human and Protein Language via Knowledge Instruction. ACL 2024

KG for LLM: Knowledge Fusion

- LLM provides internal knowledge through its parameters, while the KG provides external knowledge.



Think-on-Graph (ToG)

Think-on-graph: Deep and responsible reasoning of large language model with knowledge graph. ICLR, 2024.

KG for LLM: Knowledge Fusion

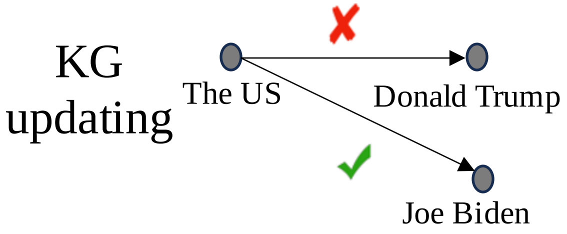
- Experimental results of TOG

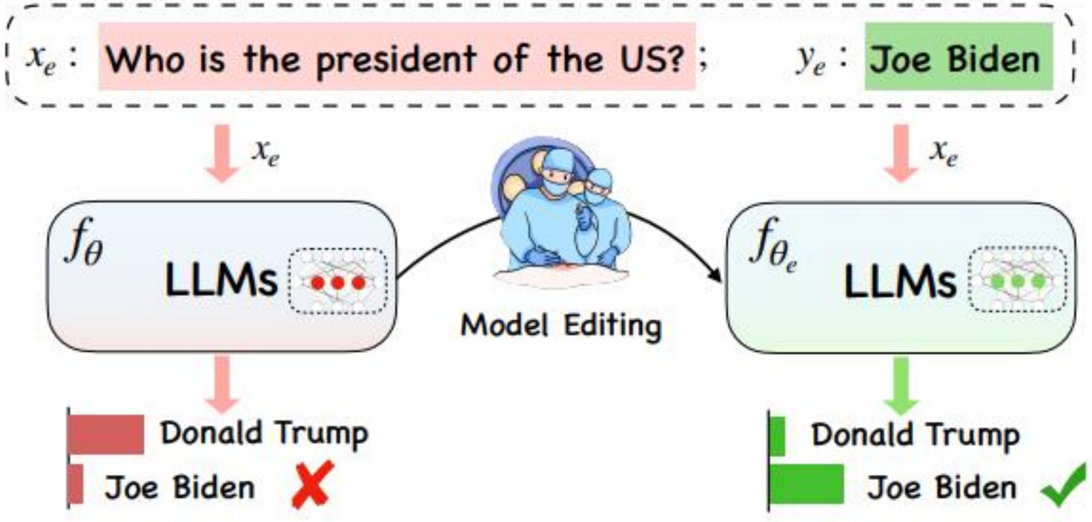
Method	Multi-Hop KBQA				Single-Hop KBQA	Open-Domain QA	Slot Filling		Fact Checking
	CWQ	WebQSP	GrailQA	QALD10-en	Simple Questions	WebQuestions	T-REx	Zero-Shot RE	Creak
Without external knowledge									
IO prompt w/ChatGPT	37.6	63.3	29.4	42.0	20.0	48.7	33.6	27.7	89.7
CoT w/ChatGPT	38.8	62.2	28.1	42.9	20.3	48.5	32.0	28.8	90.1
SC w/ChatGPT	45.4	61.1	29.6	45.3	18.9	50.3	41.8	45.4	90.8
With external knowledge									
Prior FT SOTA	70.4^{α}	82.1^{β}	75.4^{γ}	45.4^{δ}	85.8^{ϵ}	56.3^{ζ}	87.7^{η}	74.6^{θ}	88.2^{ι}
Prior Prompting SOTA	-	74.4^{κ}	53.2^{κ}	-	-	-	-	-	-
ToG-R (Ours) w/ChatGPT	58.9	75.8	56.4	48.6	45.4	53.2	75.3	86.5	93.8
ToG (Ours) w/ChatGPT	57.1	76.2	68.7	50.2	53.6	54.5	76.8	88.0	91.2
ToG-R (Ours) w/GPT-4	**69.5**	81.9	80.3	**54.7**	58.6	57.1	75.5	86.9	95.4
ToG (Ours) w/GPT-4	67.6	**82.6**	**81.4**	53.8	**66.7**	**57.9**	**77.1**	**88.3**	**95.6**

Think-on-graph: Deep and responsible reasoning of large language model with knowledge graph. ICLR, 2024.

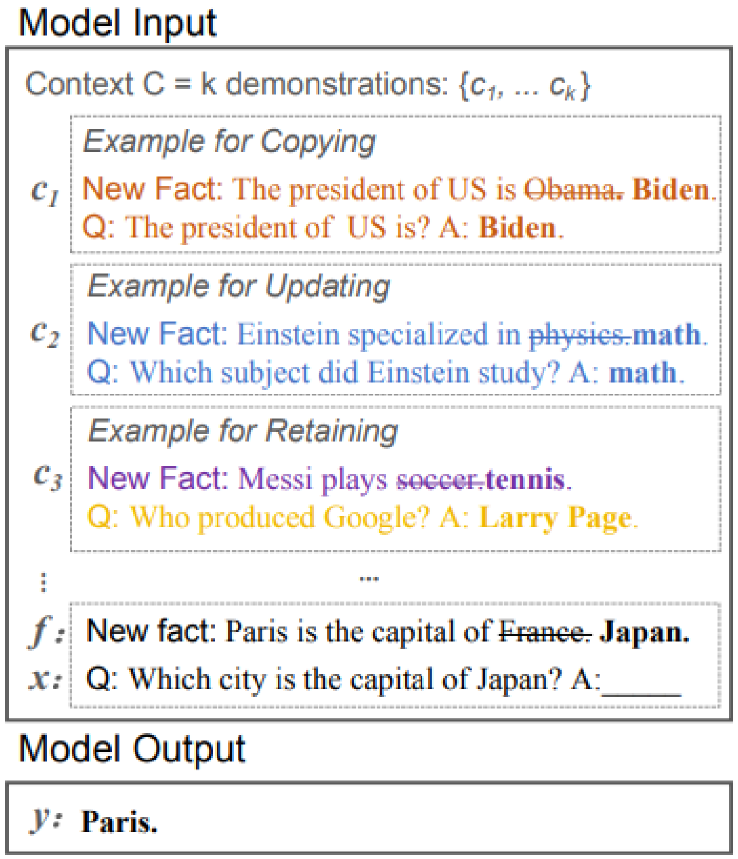
KG for LLM: Knowledge Editing

- Extracting updating knowledge from KG as In-Context Learning examples for knowledge editing





EasyEdit

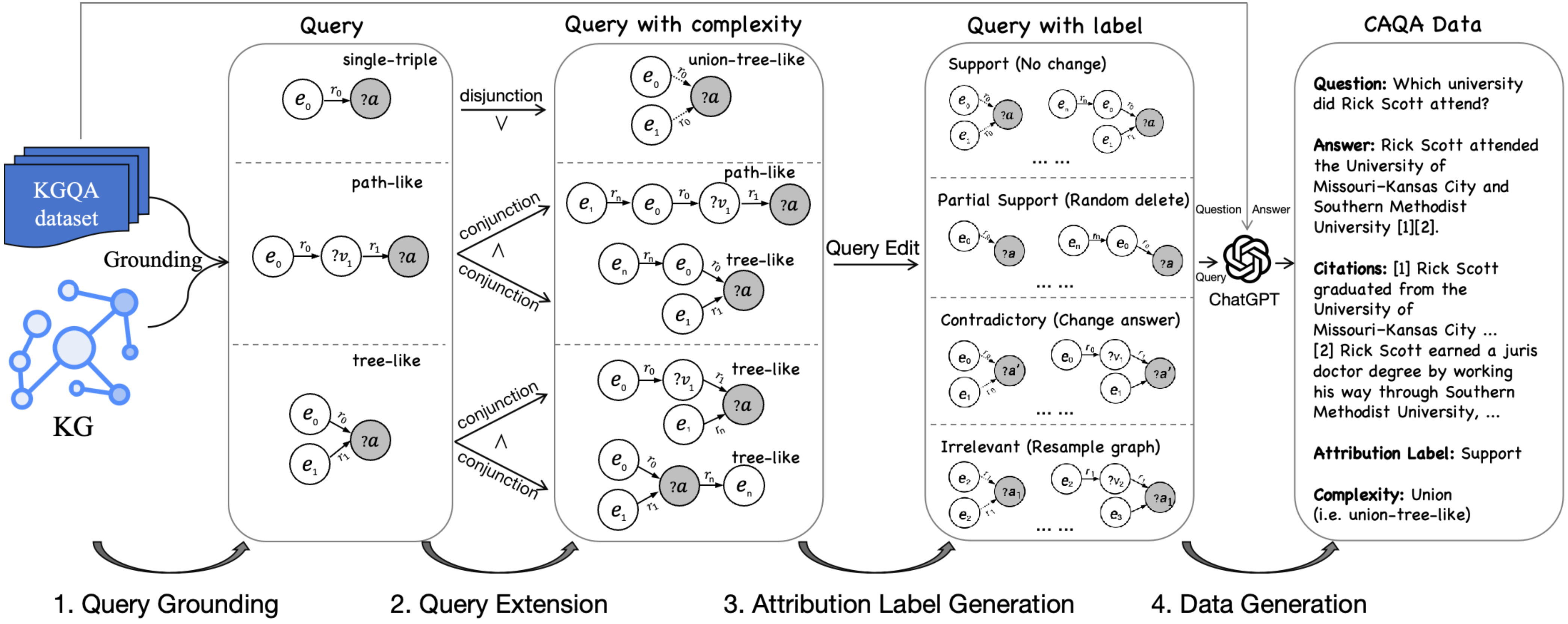


In-Context Knowledge Editing (IKE)

Editing Large Language Models: Problems, Methods, and Opportunities. EMNLP 2023.
Can We Edit Factual Knowledge by In-Context Learning? EMNLP 2023.

KG for LLM: Knowledge Validation

- Evaluating the attribution: verifying whether the generated answer is fully supported by the citation.



CAQA benchmark

Benchmarking Large Language Models in Complex Question Answering Attribution using Knowledge Graphs. Preprint. 2024

KG for LLM: Knowledge Validation

- Experimental results on CAQA dataset.

Evaluators (Size)	CAQA				
	Sup.	Ins.	Con.	Irr.	Overall
LLaMA-2 (7B)	0.423	0.121	0.057	0.170	0.279
LLaMA-2-chat (7B)	0.462	0.158	0.058	0.053	0.183
Mistral (7B)	0.456	0.178	0.191	0.153	0.305
Mistral-Instruct (7B)	0.591	0.189	0.159	0.016	0.324
Vicuna (7B)	0.437	0.007	0.001	0.000	0.111
LLaMA-2 (13B)	0.418	0.164	0.161	0.125	0.279
LLaMA-2-chat (13B)	0.469	0.171	0.173	0.103	0.224
Vicuna (13B)	0.485	0.049	0.000	0.000	0.143
GPT-3.5-turbo	0.592	0.150	0.616	0.497	0.506
GPT-4	**0.829**	**0.430**	**0.776**	**0.628**	**0.687**
AUTOIS (11B)	0.609	-	-	-	0.152
ATTRSCORE (13B)	0.667	-	0.611	-	0.320
LLaMA-2 (7B)	0.922	0.897	**0.944**	**0.933**	0.926
LLaMA-2-chat (7B)	0.925	0.903	0.943	0.927	0.930
Mistral (7B)	0.927	0.908	**0.944**	0.849	0.882
Vicuna (7B)	0.937	0.907	0.940	0.906	0.932
LLaMA-2 (13B)	0.929	0.907	0.938	0.923	0.925
Vicuna (13B)	**0.942**	**0.923**	0.939	0.923	**0.933**

Table 5: The performance of the different attribution evaluators on our CAQA benchmark. Evaluators of the first (resp. second) part follow the zero-shot (resp. fine-tuning) setting.

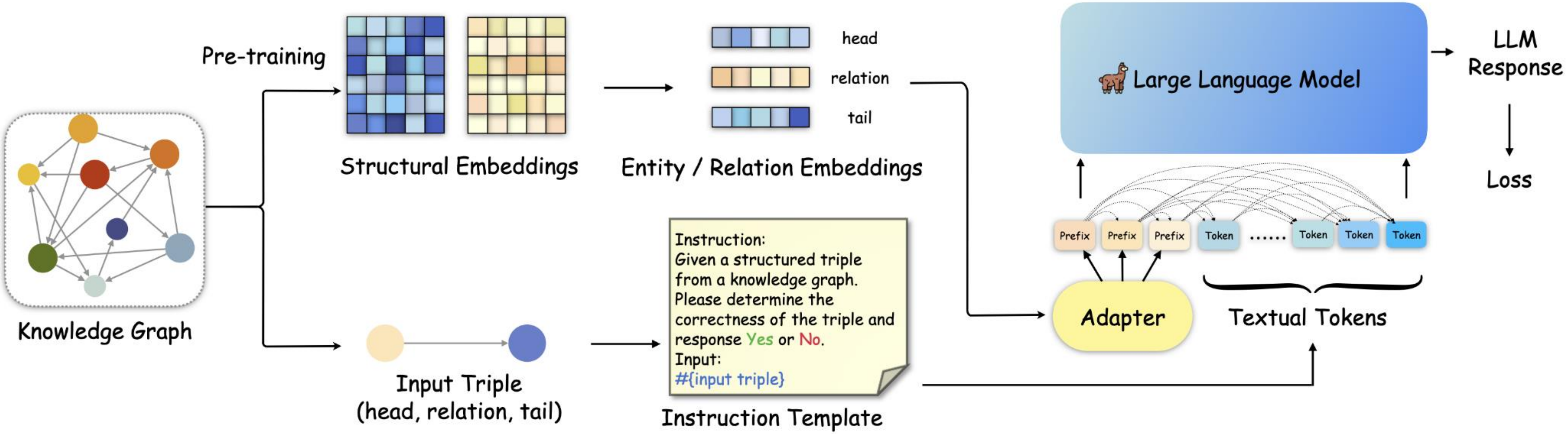
Evaluators (Size)	CAQA			
	S.	C.	I.	U.
LLaMA-2 (7B)	0.286	0.249	0.282	0.260
LLaMA-2-chat (7B)	0.281	0.235	0.291	0.290
Mistral (7B)	0.315	0.281	0.294	0.265
Mistral-Instruct (7B)	0.339	0.278	0.300	0.271
Vicuna (7B)	0.341	0.268	0.290	0.285
LLaMA-2 (13B)	0.314	0.270	0.303	0.253
LLaMA-2-chat (13B)	0.338	0.279	0.305	0.278
Vicuna (13B)	0.339	0.257	0.296	0.288
GPT-3.5	0.551	0.323	0.346	0.525
GPT-4	**0.743**	**0.416**	**0.501**	**0.787**
AUTOIS (11B)	0.403	0.171	0.272	0.281
ATTRSCORE (13B)	0.473	0.333	0.308	0.303
LLaMA-2 (7B)	0.923	0.815	0.931	0.921
LLaMA-2-chat (7B)	0.935	0.820	0.930	0.924
Mistral (7B)	0.935	0.831	0.921	0.905
Vicuna (7B)	**0.956**	0.823	**0.936**	0.939
LLaMA-2 (13B)	0.954	0.824	**0.936**	0.939
Vicuna (13B)	0.950	**0.847**	0.935	**0.940**

Table 6: Performance of all evaluators on various level of attribution complexity. Evaluators of the first (resp. second) part follow the zero-shot (resp. fine-tuning) setting.

Contents

LLM for KG: KG Completion

- Knowledge Prefix Adapter: structure-aware reasoning with structure embedding.



KoPA

Making Large Language Models Perform Better in Knowledge Graph Completion. Preprint 2024.

LLM for KG: KG Completion

- Experimental results of CAQA dataset.

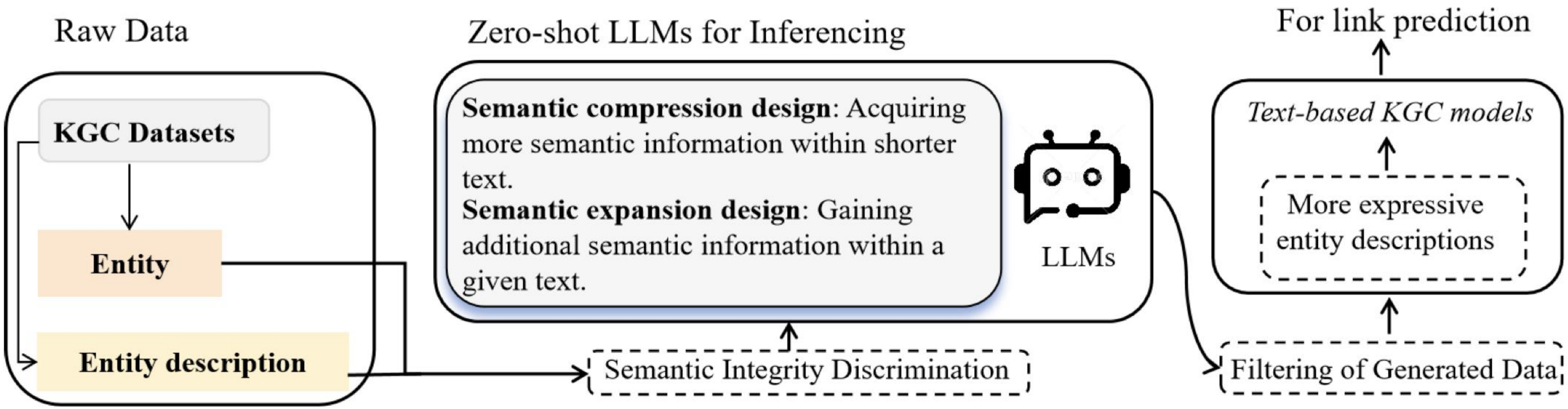
	Model	UMLS				CoDeX-S				FB15K-237N			
		Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
Embedding-based	TransE [3]	84.49	86.53	81.69	84.04	72.07	71.91	72.42	72.17	69.71	70.80	67.11	68.91
	DistMult [38]	86.38	87.06	86.53	86.79	66.79	69.67	59.46	64.16	58.66	58.98	56.84	57.90
	ComplEx [34]	90.77	89.92	91.83	90.87	67.64	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 [21]	-	-	-	-	-	-	-	-	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**

Making Large Language Models Perform Better in Knowledge Graph Completion. Preprint 2024.

LLM for KG: KG Completion

- Does the texts optimized by LLMs are more effective for text-based KGC models?

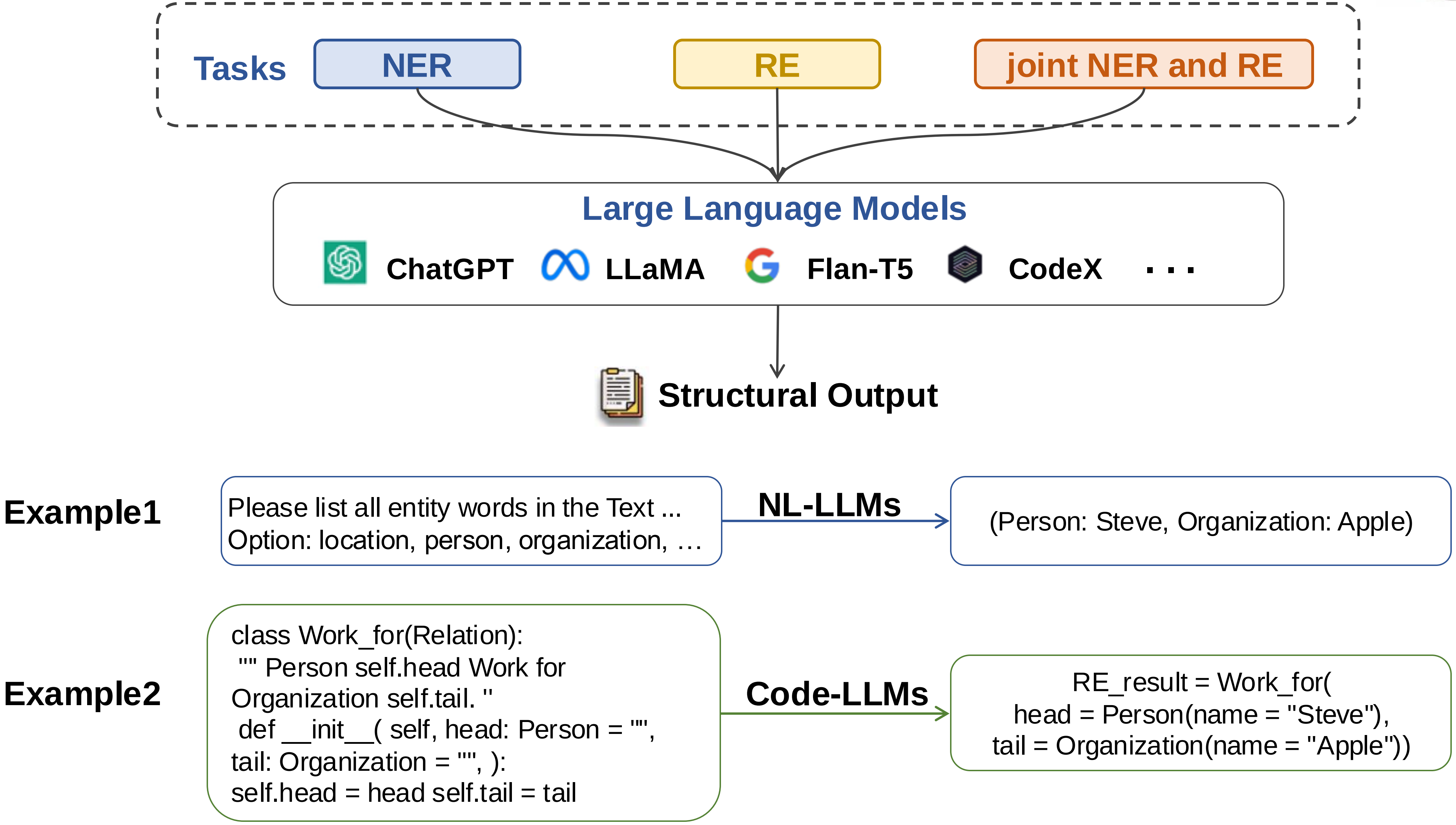
LLMs can add or remove content from entity descriptions.



Constrained Prompts for KGC (CP-KGC)

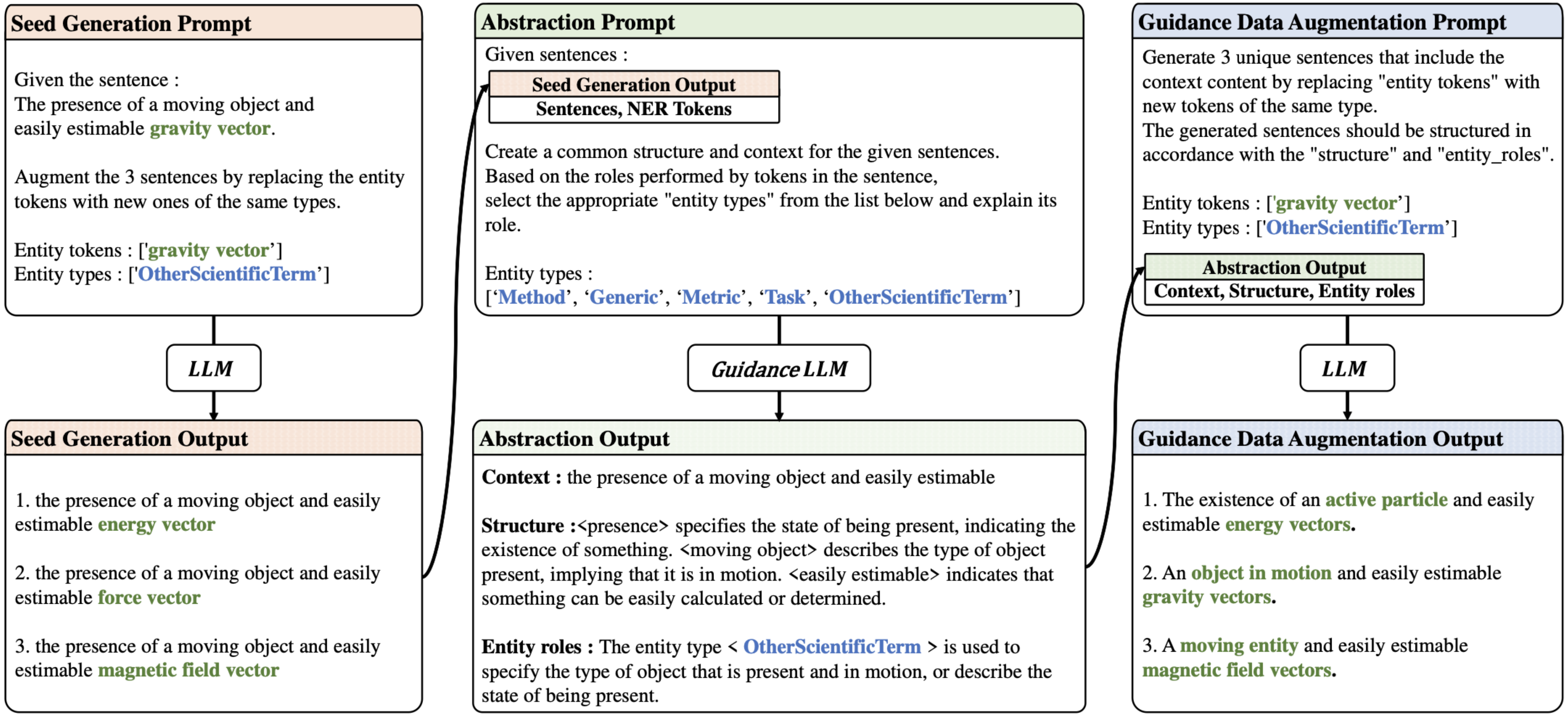
Enhancing text-based knowledge graph completion with zero-shot large language models: A focus on semantic enhancement, KBS 2024

LLM for KG: Entity and Relation Extraction



LLM for KG: Named Entity Recognition

- LLM can perform guidance data augmentation for NER tasks.

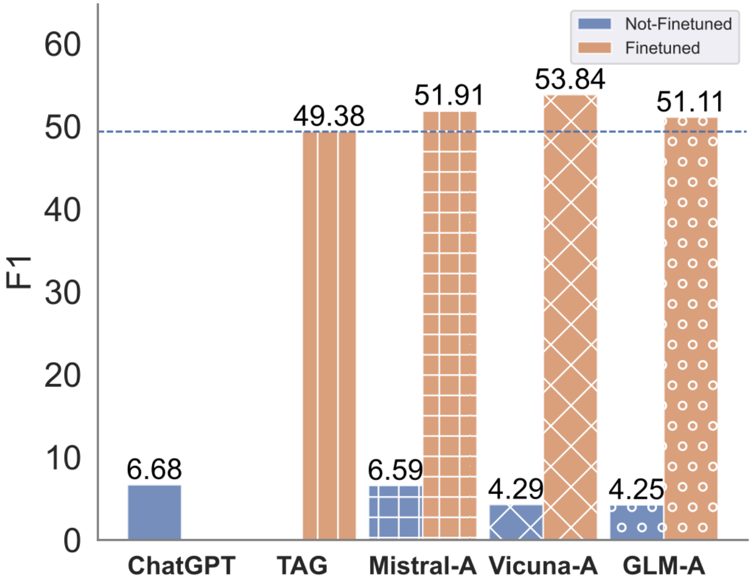


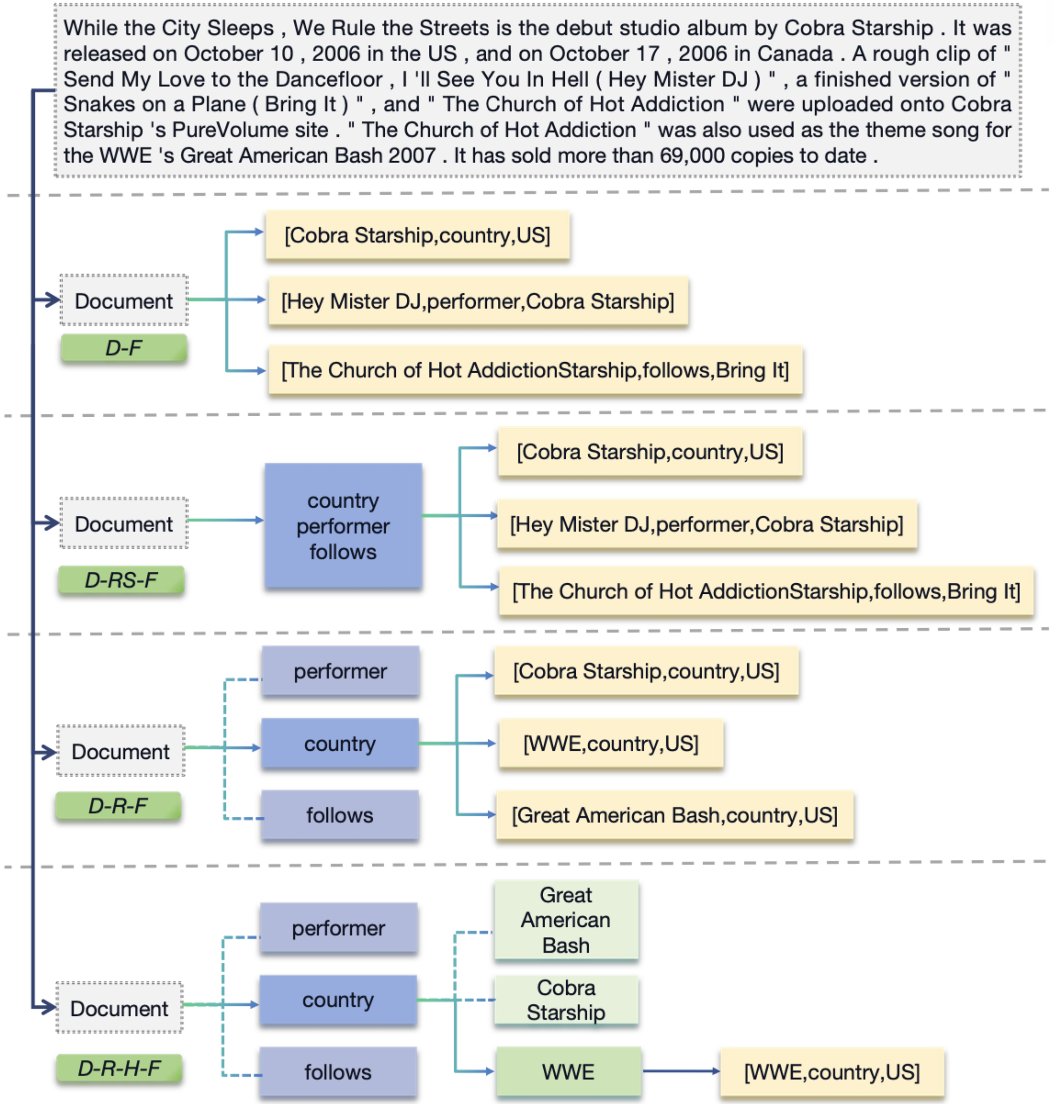
Guidance LLM Data Augmentation

Guidance-Based Prompt Data Augmentation in Specialized Domains for Named Entity Recognition, ACL 2024

LLM for KG: Relation Extraction

- Exploring LLM on different RE paradigms
- RHF (Relation-Head-Facts).





Document-facts

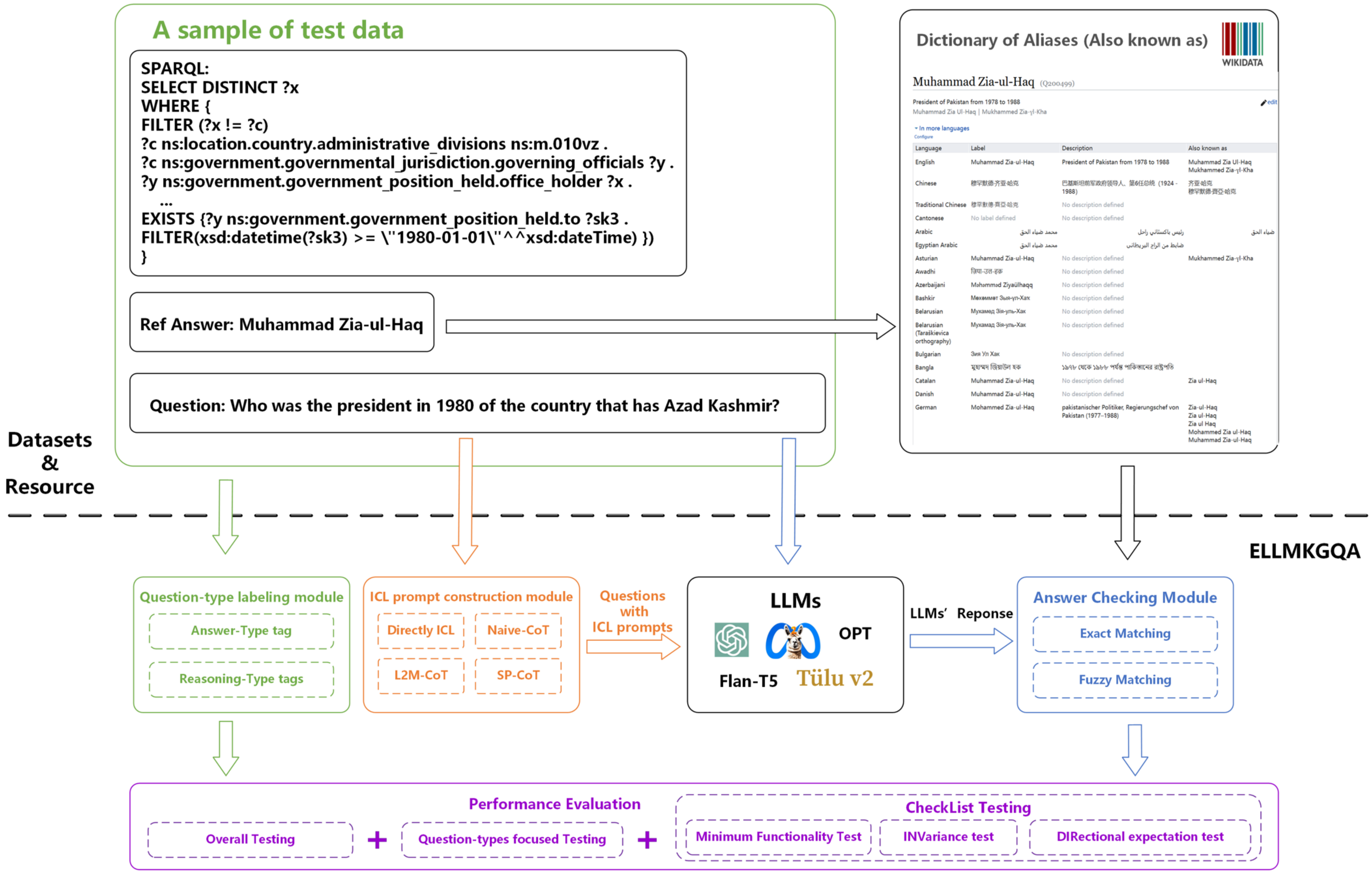
Document-relations-facts

Document-relation-facts

Document-relation-head-facts

AutoRE: Document-Level Relation Extraction with Large Language Models , ACL Demo 2024

LLM for KG: KBQA



ELLMKGQA framework:

The Question-type Labeling Module identifies the answer type of the input question and the reasoning type involved in answering the question (based on the **question text**, **reference answer**, and corresponding **SPARQL query**).

The ICL Prompt Construction Module converts the input question into various inquiry forms with different contextual learning strategies

The Answer Checking Module determines whether the LLM's response includes the correct answer to the input question by utilizing a combination of exact matching and fuzzy matching methods (employing an alias dictionary from **Wikidata** in exact matching to reduce false negatives).

LLM for KG: KBQA

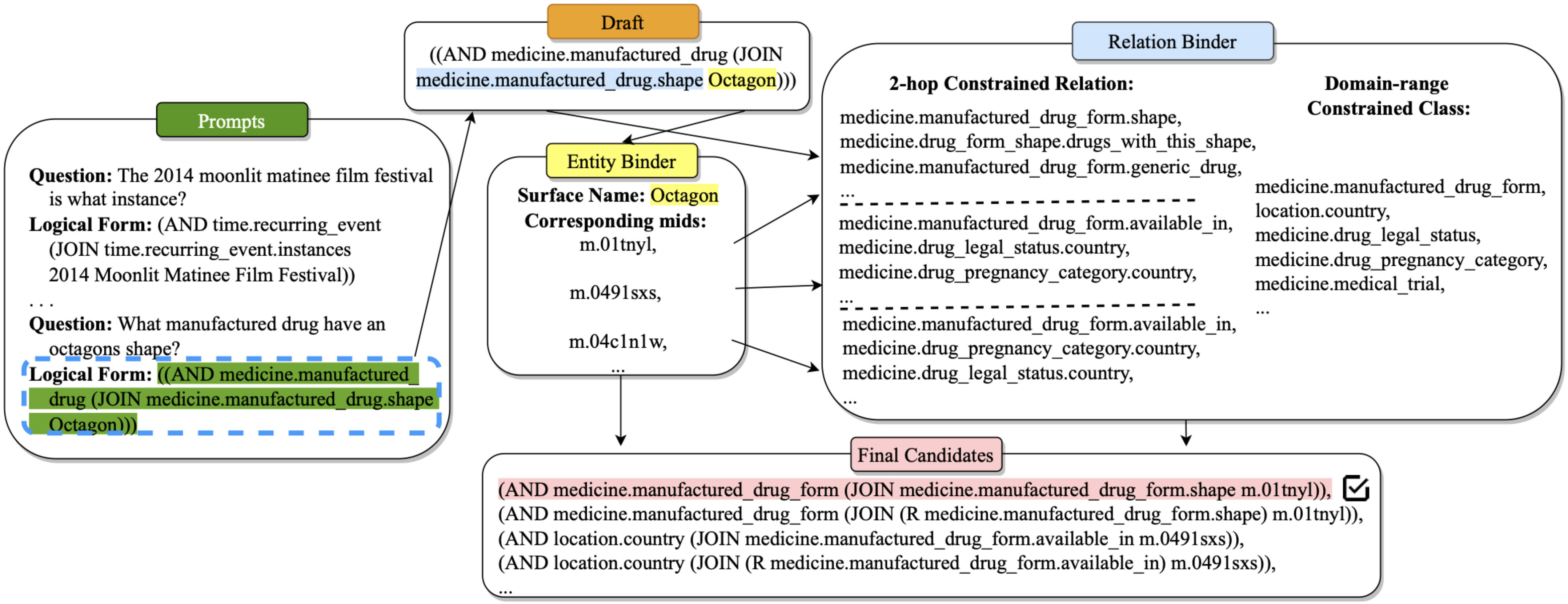
- Experimental results of LLM KBQA

Datasets	KQApro	LC-quad2.0	WQSP	CWQ	GrailQA	GraphQuestions
	Acc	F1	Acc	Acc	Acc	F1
SOTA(supervised)	95.32 [3]	83.45 [4]	82.10 Yu et al. (2022)	72.20 Hu et al. (2022)	76.31	31.8 Gu and Su (2022)
SOTA(unsupervised)	94.20 Nie et al. (2022)	-	62.98 Ye et al. (2022)	-	-	-
FLAN-T5-XXL	37.27	30.14	59.87	46.69	29.02	32.27
LLaMA2-7B	49.78	50.85	82.39	63.04	46.74	61.01
LLaMA2-7B-Direct	44.79	44.88	69.16	55.49	38.46	45.92
LLaMA2-7B-Naive	50.59 ↑	44.86	73.36	58.24	40.47	50.98
LLaMA2-7B-L2M	47.59	40.17	64.39	54.27	35.25	43.91
LLaMA2-7B-SP	45.53	41.22	58.18	53.98	33.79	40.70
LLaMA2-13B	48.42	48.92	80.66	59.14	45.22	61.18
LLaMA2-70B	51.82	51.83	85.81	63.85	48.88	63.15
LLaMA3-8B	49.08	51.51	84.29	62.91	45.53	62.42
LLaMA3-8B-Direct	41.82	40.90	76.12	52.79	34.11	49.85
LLaMA3-8B-Naive	50.50 ↑	51.30	69.73	58.12	38.07	53.54
LLaMA3-8B-L2M	18.95	25.58	56.79	43.23	26.07	39.51
LLaMA3-8B-SP	39.68	41.59	67.51	51.04	31.70	44.41
LLaMA3-70B	54.43	60.85	86.32	68.68	51.79	69.25
LLaMA3-70B-Direct	42.36	45.28	76.51	57.83	35.75	49.15
LLaMA3-70B-Naive	57.55 ↑	61.13 ↑	84.72	73.30 ↑	50.94	65.57
LLaMA3-70B-L2M	42.36	45.28	76.51	57.83	35.75	49.15
LLaMA3-70B-SP	44.81	47.74	78.68	56.60	36.13	51.70
GPT-4	50.19	54.53	83.49	65.57	43.96	60.38
GPT-4-Direct	41.60	43.58	77.74	54.06	35.28	48.49
GPT-4-Naive	46.42	49.72	74.15	60.57	37.26	49.53
GPT-4-L2M	50.09	51.32	77.74	64.34	42.92	51.51
GPT-4-SP	50.00	49.91	78.30	62.17	42.08	52.55
GPT-4o	56.98	61.51	85.85	74.06	53.96	67.17
GPT-4o-Direct	51.42	55.57	83.96	64.62	41.79	42.92
GPT-4o-Naive	48.49	53.40	79.43	61.79	38.77	49.81
GPT-4o-L2M	43.77	47.74	71.32	59.34	31.98	43.49
GPT-4-SP	46.60	49.53	72.26	58.77	36.98	46.51

Can ChatGPT Replace Traditional KBQA Models? An In-Depth Analysis of the Question Answering Performance of the GPT LLM Family. ISWC 2023

LLM for KG: KBQA

- LLMs help generate logical forms as the draft for a specific question by imitating a few demonstrations.

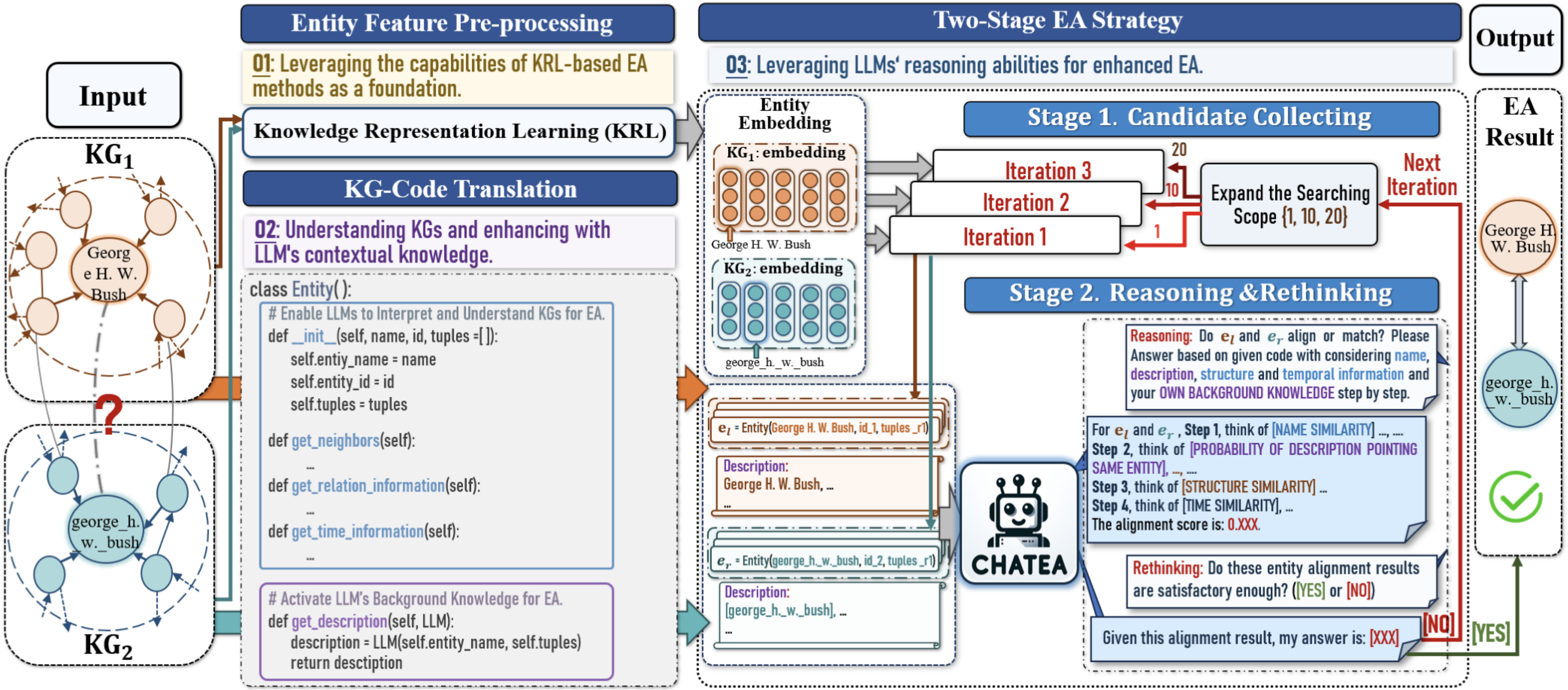


KB-Binder

Few-shot In-context Learning for Knowledge Base Question Answering , ACL Demo 2024

LLM for KG: Entity Alignment

- Leverage LLM to aligned the entities from two different KGs.



Chat Entity Alignment (ChatEA)

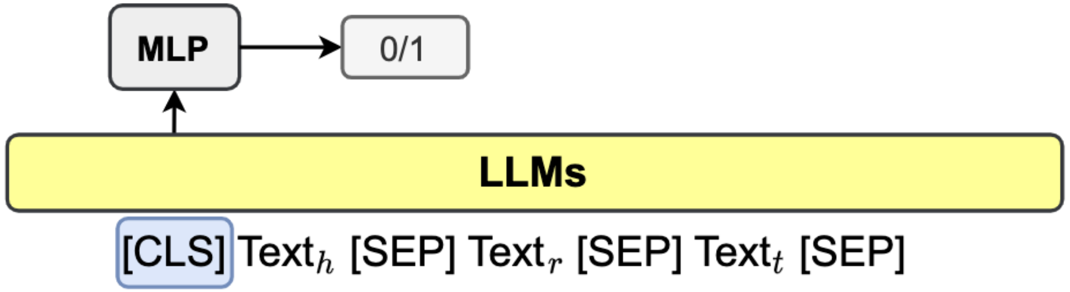
Unlocking the Power of Large Language Models for Entity Alignment , ACL 2024

LLM for KG: KG Reasoning

- By leveraging the context encoding capability of LLMs, the representation of the knowledge graph is enhanced using textual information from the knowledge graph.

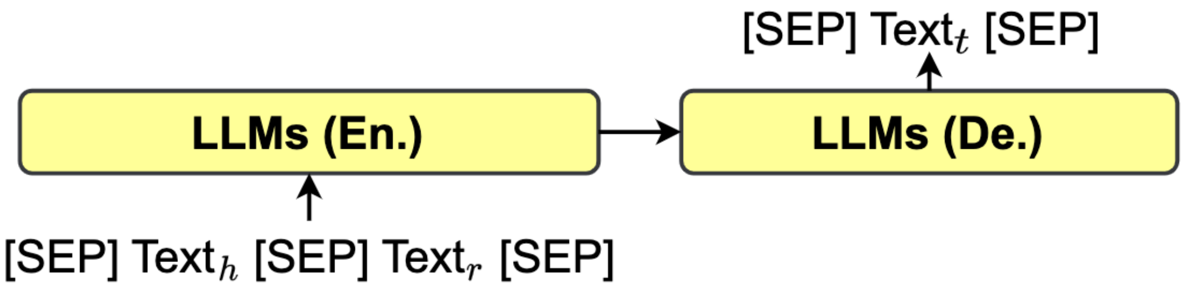
Discriminative Methods:

- Encoder-only PLMs (e.g., BERT)



Generative Methods:

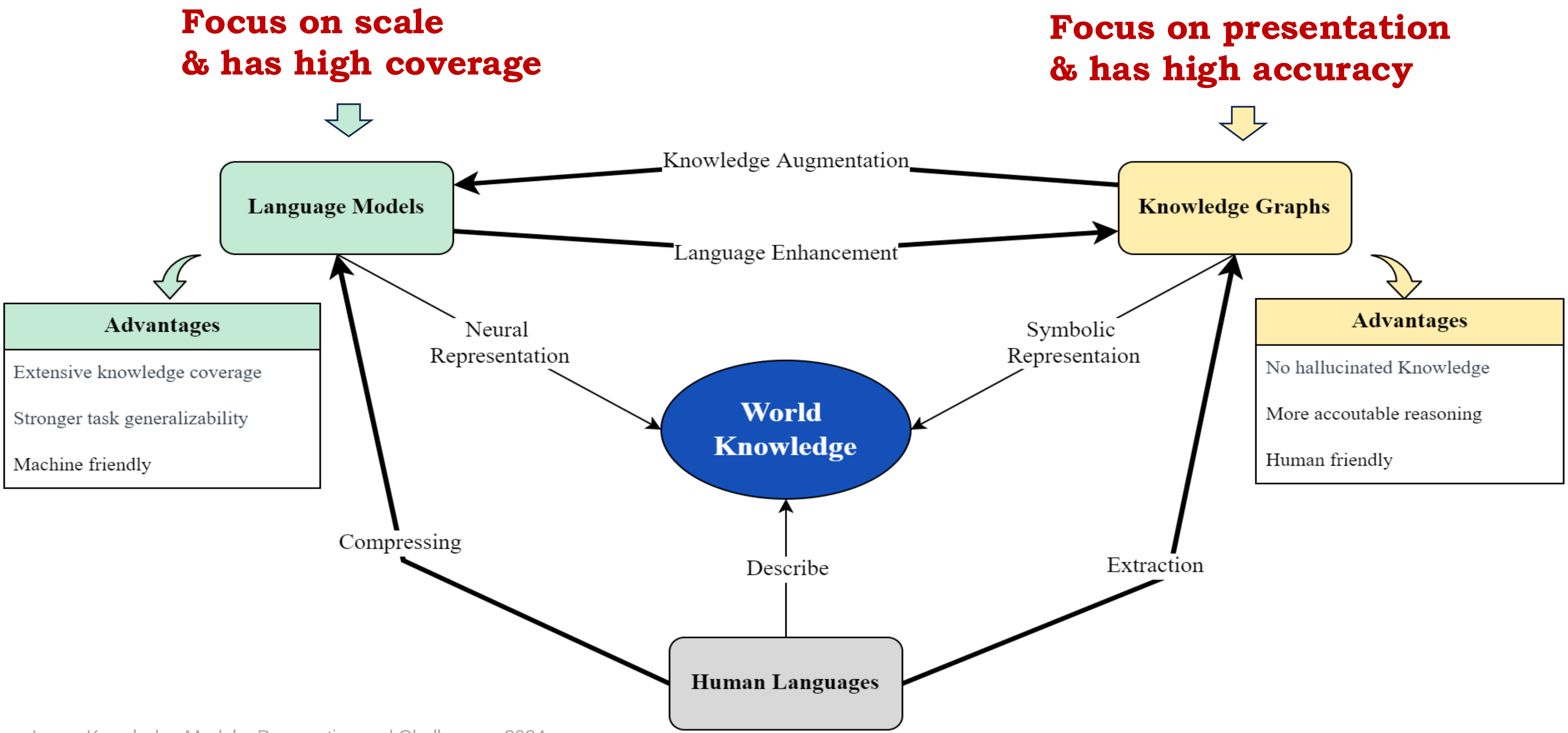
- Encoder-decoder or decoder-only PLMs



Unifying Large Language Models and Knowledge Graphs: A Roadmap. IEEE TKDE 2024

Contents

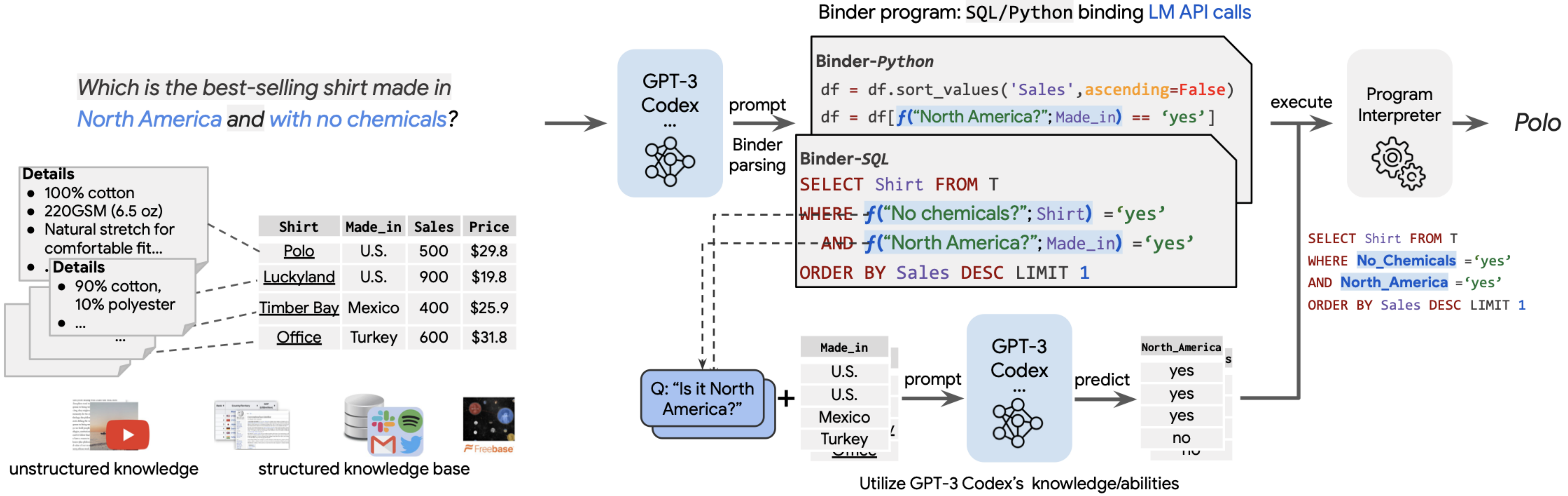
How do KG and LLM collaborate?



Large Knowledge Models: Perspective and Challenges, 2024.

KG x LLM: Neural-symbolic Framework

- Binding a unified API of LLM functionalities to a programming language (e.g., SQL, Python, SPARQL ...) to extend its grammar coverage and thus tackle more diverse questions.

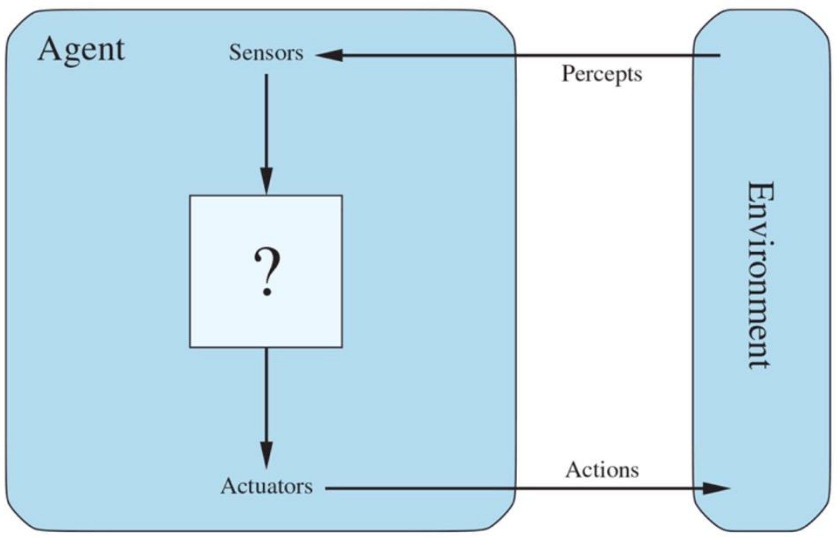


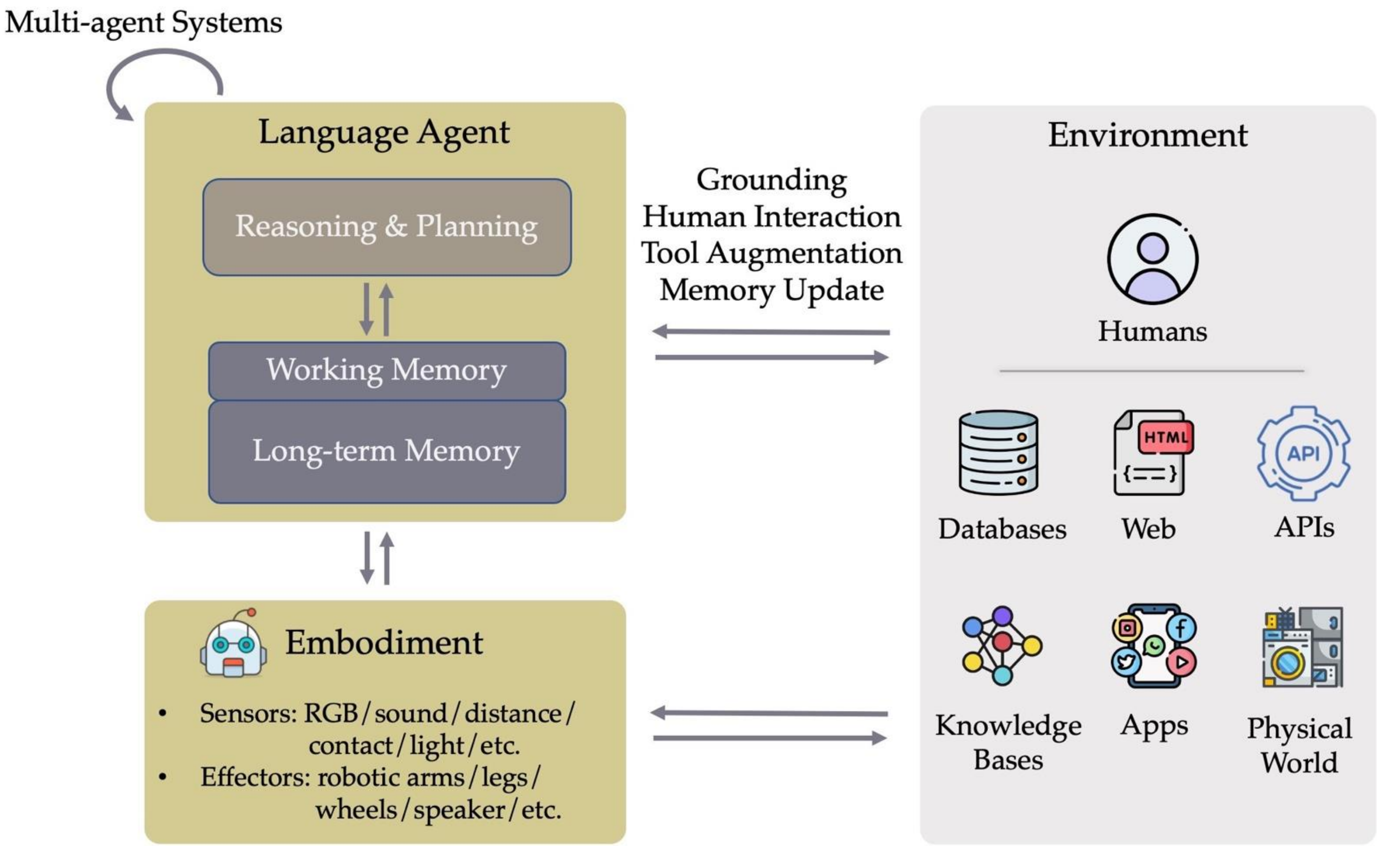
Binder

KG-Agent: An Efficient Autonomous Agent Framework for Complex Reasoning over Knowledge Graph. Preprint 2024

KG x LLM: Language Agent

- Contemporary agents use language for their thought process, which makes it much easier to incorporate heterogeneous external percepts and do multi-step (speculative) planning and reasoning, all in a non-programmed and explicit way.

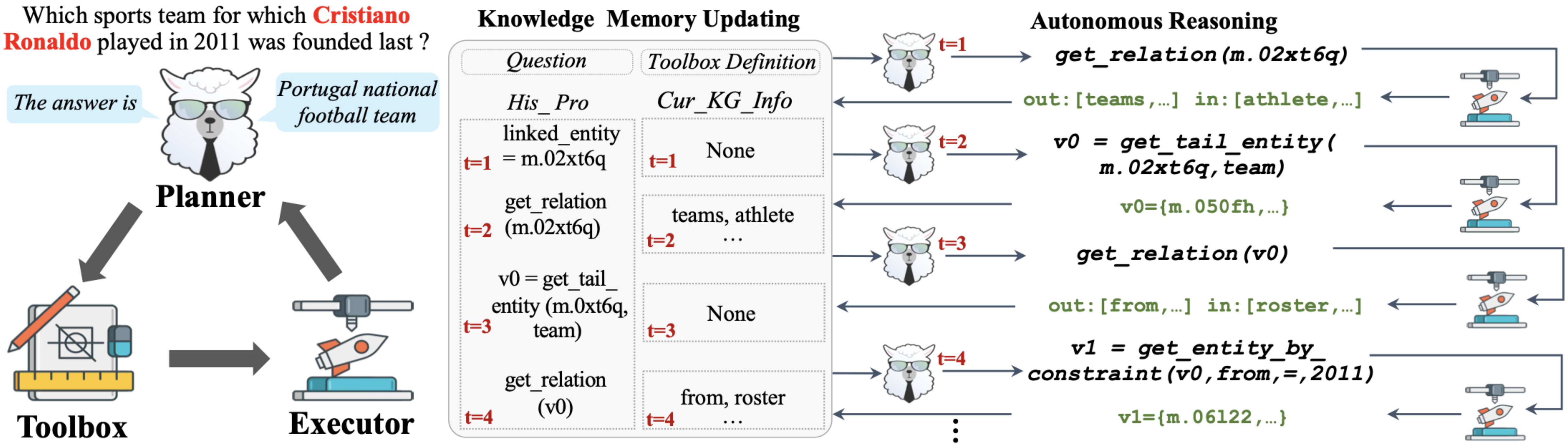




https://yusu.substack.com/p/language-agents

KG x LLM: KG Agent

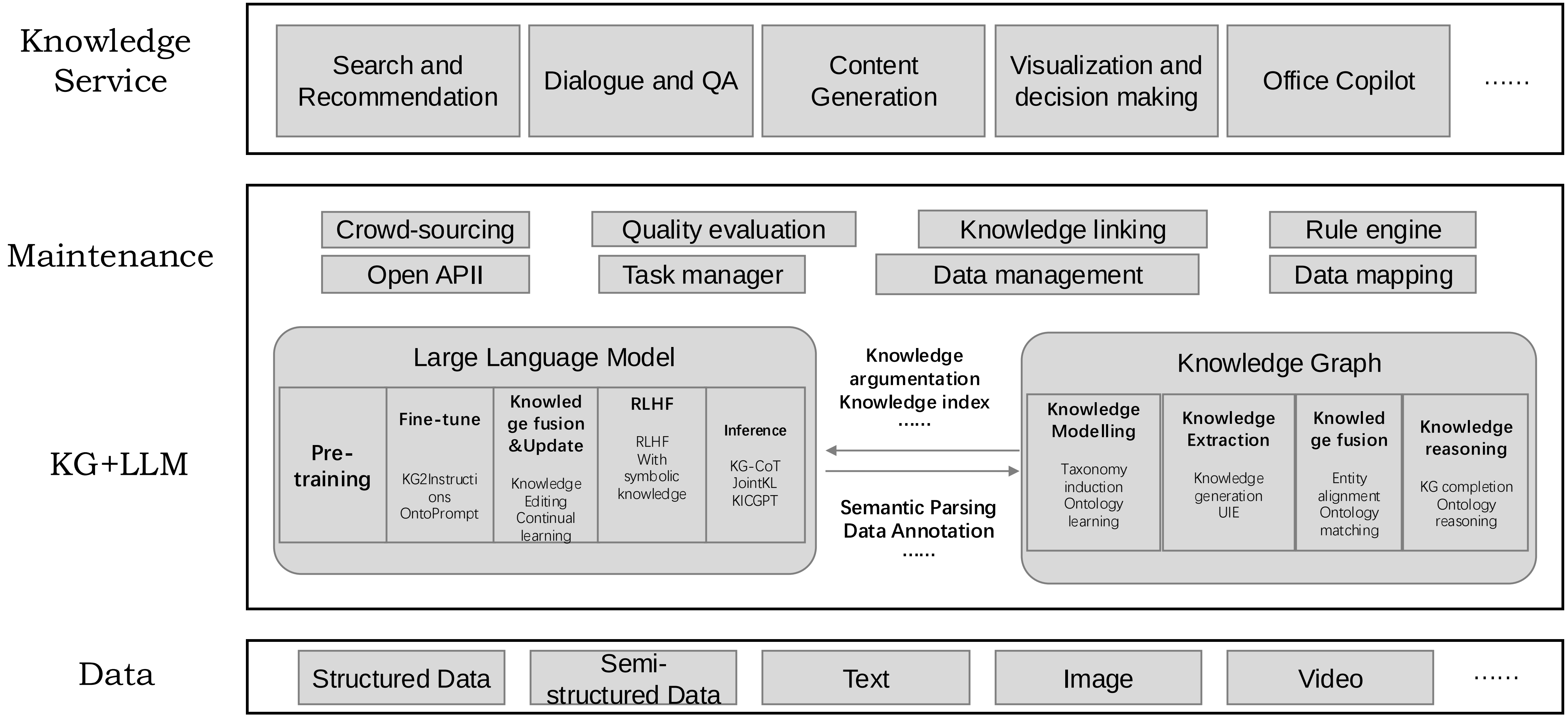
- Integrating the LLM, multifunctional toolbox, KG-based executor, and knowledge memory, and develop an iteration mechanism that autonomously selects the tool then updates the memory for reasoning over KG



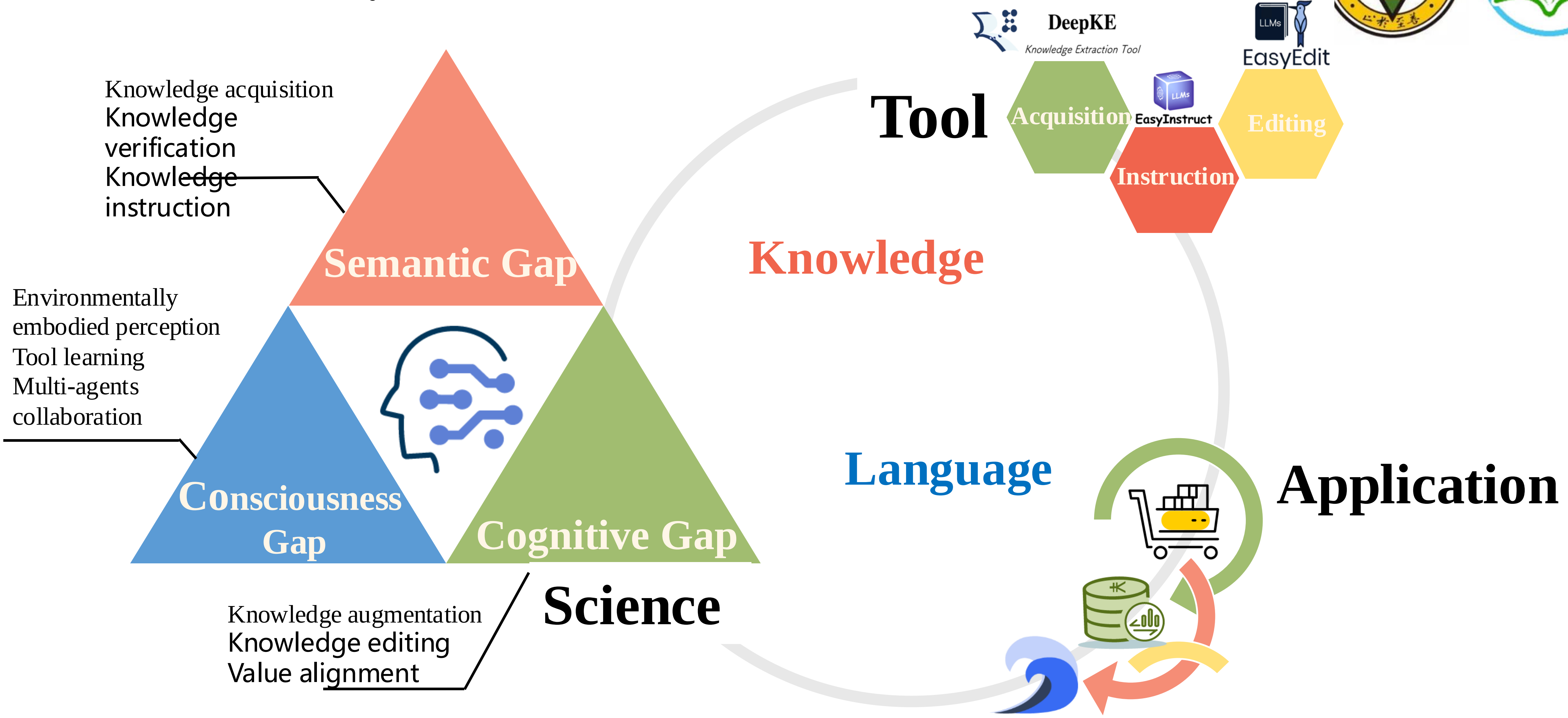
KG-Agent

KG-Agent: An Efficient Autonomous Agent Framework for Complex Reasoning over Knowledge Graph. Preprint 2024

KG x LLM: Knowledge Service Platform



KG x LLM: OpenKG



Language as "form", knowledge as "heart", graph as "skeleton"

Contents

Conclusion

- KG for LLM
  - ✓ KG can enhance pre-training, instruction-tuining, RAG, ICL, fusion, update, validation of LLM
- LLM for KG
  - ✓ LLM can knowledge graph completion, extraction, fusion, reasoning and validation of KG
- Integration of LLM and KG
  - ✓ New agents can be designed
  - ✓ OpenKG: **Language as "form", knowledge as "heart", graph as "skeleton"**

Future Work

- KG for LLM
  - ✓ Effective and efficient learning of symbolic knowledge in KGs
  - ✓ Benchmarks generated by KGs to validate LLMs
  - ✓ Improving (interpretable) reasoning ability of LLM using KGs

- LLM for KG
  - ✓ Automating KG engineering pipeline using agent based LLM
  - ✓ Tool-augmented LLM for symbolic reasoning of KG
  - ✓ Enhancing Knowledge services based on KGs by LLM

- Integration of LLM and KG
  - ✓ Newly designed unified agent
  - ✓ Generalizable, trustable and stable knowledge services
  - ✓ Programmable knowledge engine

Thank you!

Email address: yongruichen@seu.edu.cn