

Data Intensive Question Answering with Large Language Models

Jiaoyan Chen

Lecturer at the Department of Computer Science, University of Manchester

Workshop on Large Language Models for Knowledge Engineering

31st Oct, 2024

University of Stuttgart

Data Intensive QA

LLM reasoning with local data storage

Data Intensive QA

LLM reasoning with the support of local data storage

Towards LLM limitations

- Changes & data freshness

- Data privacy & intellectual property

- Explanation

- Domain knowledge

- Costs from huge LLM size

- Hallucination

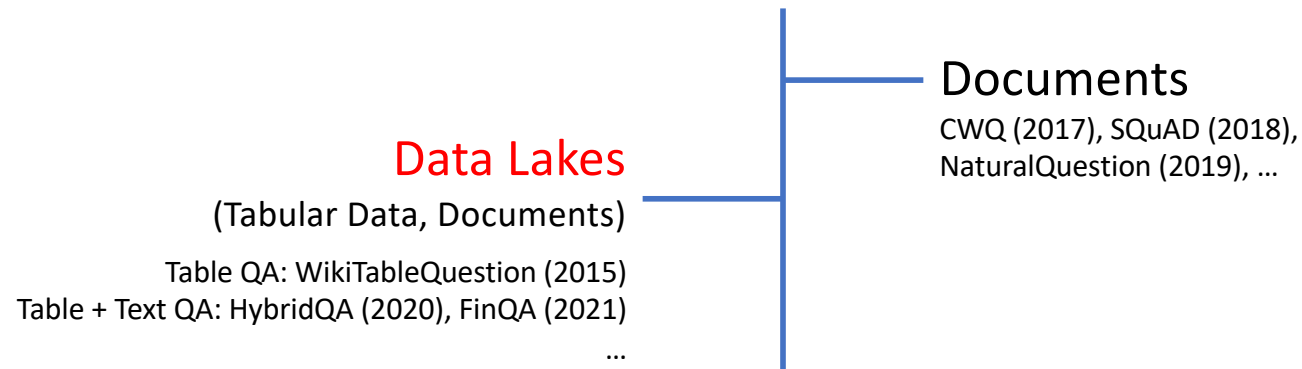
Data Intensive QA



Documents

CWQ (2017), SQuAD (2018),
NaturalQuestion (2019), ...

Data Intensive QA



Min, Dehai, et al. "Exploring the Impact of Table-to-text Methods on Augmenting LLM-based Question Answering with Domain Hybrid Data." NAACL 2024 (Industry Track)

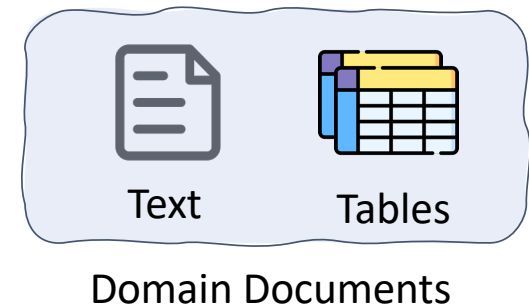
Benchmarking Table-to-Text in LLM-based QA

- Real-World Data Consists of **Hybrid Data (Text and Tables)**

Common in : **Scientific Literature , Medical Reports, etc.**

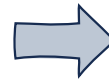
Tables alongside text provide :

- **Supplementary or complementary information**
- **Enhancing the understanding of the content**



- Enhancing LLMs in Domain-Specific Question Answering

- Domain-Specific Fine-Tuning (DSFT)
- Retrieval-Augmented Generation (RAG)

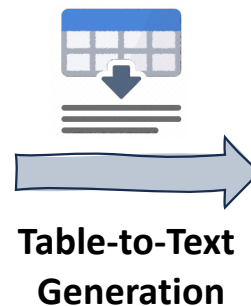


Both rely on domain-specific corpus

Table-to-Text Generation

- Generates natural language statements that faithfully describe the information in the provided table
- Transforms hybrid data into a **unified natural language representation**
- Preserves the **semantic connections** between the data

Frequency Band	Channel Bandwidth	Peak Data Rate
6 GHz	320 MHz	11.53 Gbps
5 GHz	160 MHz	5.765 Gbps
2.4 GHz	40 MHz	1.376 Gbps
...		



The 6 GHz band offers a channel bandwidth of 320 MHz. It can reach a peak data rate of 11.53 Gbps (gigabits per second). The 5 GHz band has a channel bandwidth of 160 MHz. Its peak data rate is 5.765 Gbps ...

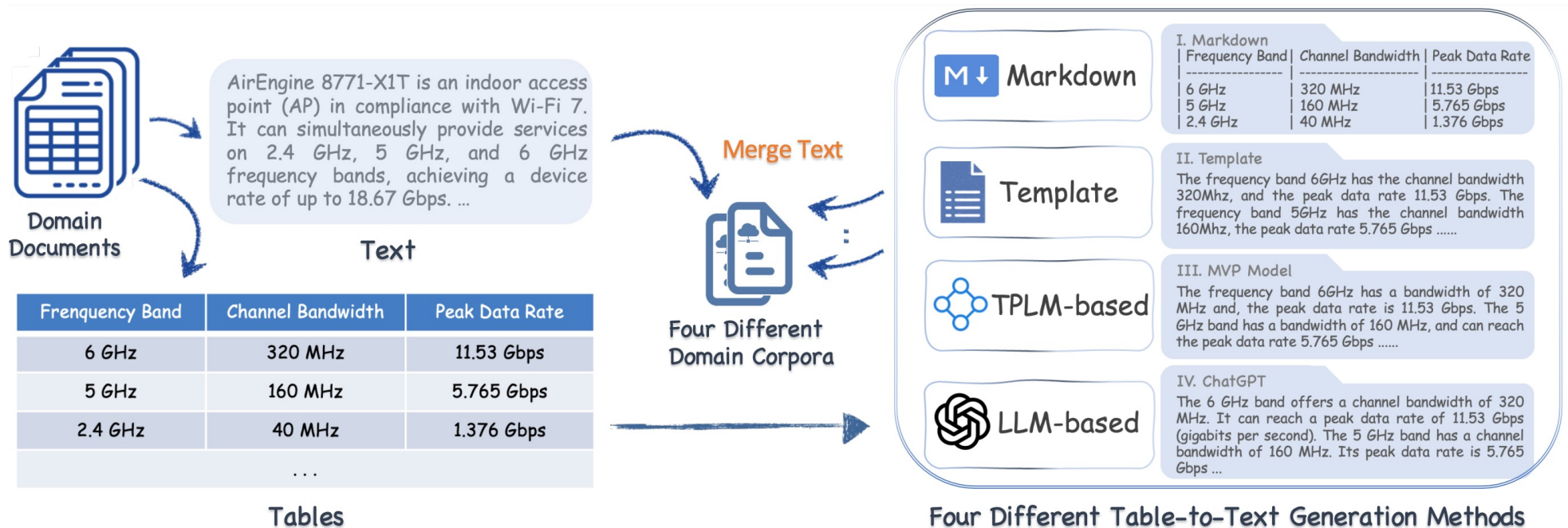
Research Gap

- The lack of comparative analysis on **how different table-to-text methods affect** the performance of domain-specific QA systems.

Address this research gap:

- **Step 1: Innovatively integrates table-to-text generation into the LLM-based Domain QA framework**
- **Step 2: Conducts extensive experiments with different table-to-text methods on two types of QA systems**

Building Domain Corpora with Table-to-text

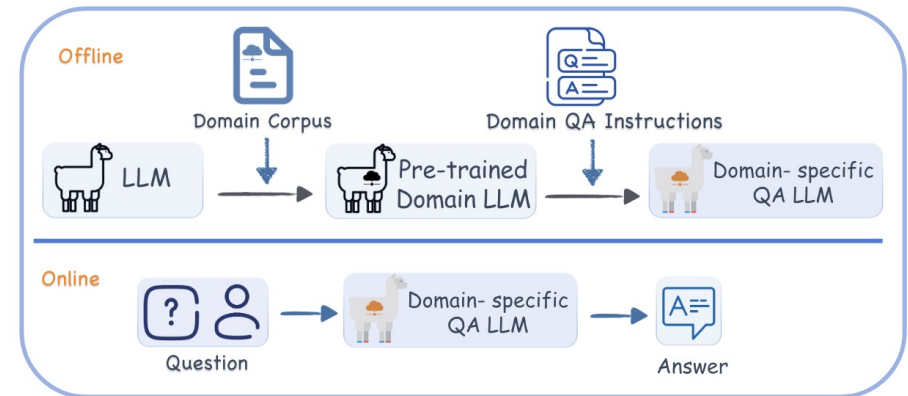


- TPLM-based method: fine-tuning Traditional Pre-trained Language Models (TPLMs), such as BART.
- LLM-based method: utilize LLMs in a in-context learning setting.

Building LLM-based QA Systems with Domain Corpora

System1 - DSFT:

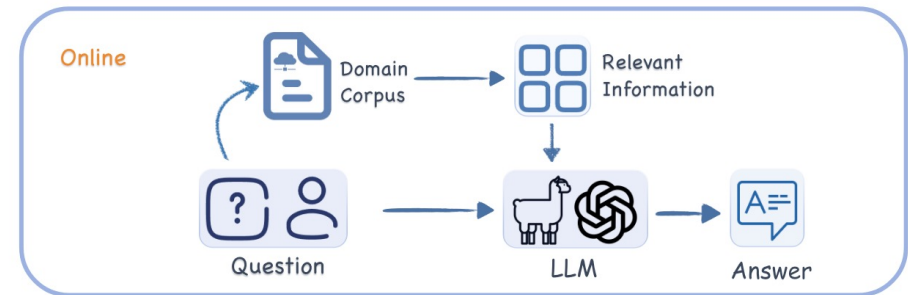
- Step 1: Incrementally pre-train the LLM on the domain corpus
- Step 2: Instruction tuning on the QA task



(a) Domain-Specific Fine-Tuning QA system

System 2 – RAG:

- LangChain framework
- Dense Passage Retriever (DPR) method for information retrieval



(b) Retrieval-Augmented Generation QA system

Dataset

ICT-DATA:

- Real-world industry hybrid dataset
- Based on 170 technical documents related to ICT (Information and Communication Technology) products
- 178 million words, 6GB text storage size
- Table data accounts for about 18% of the total word count

ICTQA:

- 9k questions with long-form answers
- Test set: 500 questions, whose answers involve knowledge from both tables and text.

Evaluation Metrics

Automated Evaluation:

- GPT-4 as an evaluator
- In-context learning: one demonstration
- Range: 0 to 5, discrete values. larger denotes better
- Based on helpfulness and similarity to the golden answer

Human Evaluation:

- 3 evaluators with domain knowledge
- Same scoring criteria with GPT-4

Experiment Settings

DSFT Paradigm:

- Meta's OPT (1.3B to 13B)
- Llama2-base (7B, 13B)
- QLoRA for pre-training and instruction fine-tuning

RAG Paradigm:

- Llama2-chat (7B, 13B, and 70B)
- GPT-3.5-turbo
- BGE model for DPR embedding
- Top-3 relevant text chunks based on similarity

Fair Comparison: the same setting on four different corpora (table-to-text generation methods).

Results and Analysis

Metrics	Table-to-Text Method	Domain-Specific Fine-Tuning						Retrieval-Augmented Generation			
		OPT-1.3B	OPT-2.7B	OPT-6.7B	OPT-13B	Llama2-7B	Llama2-13B	GPT-3.5-turbo	Llama2-7B	Llama2-13B	Llama2-70B
Human Eval.	Markdown	2.05	2.41	2.38	2.51	2.82	3.05	3.29	3.72	<u>3.98</u>	<u>3.94</u>
	Template	2.04	2.40	2.26	2.47	2.82	3.04	<u>3.36</u>	3.44	3.96	3.76
	TPLM-based	<u>2.12</u>	<u>2.43</u>	<u>2.43</u>	<u>2.58</u>	3.20	<u>3.13</u>	3.26	3.27	3.92	3.64
	LLM-based	2.18	2.57	2.51	2.62	<u>2.96</u>	3.19	3.62	<u>3.71</u>	4.26	4.09
	RSD(%)	2.80	3.40	5.00	3.00	7.60	3.00	7.20	9.00	6.80	9.00
GPT-4 Eval.	Markdown	1.74	2.16	2.27	2.25	2.7	3.06	3.28	3.66	<u>3.67</u>	3.74
	Template	1.81	2.22	2.39	2.34	2.84	3.08	3.27	3.06	3.38	3.37
	TPLM-based	<u>2.33</u>	<u>2.46</u>	<u>2.45</u>	<u>2.53</u>	3.20	<u>3.19</u>	<u>3.28</u>	2.9	3.41	3.30
	LLM-based	2.57	2.69	2.73	2.86	<u>3.06</u>	3.30	3.64	<u>3.59</u>	3.69	<u>3.54</u>
	RSD(%)	16.60	10.60	9.20	12.20	10.00	4.80	7.40	15.20	6.20	8.80

Relative Score Differences (RSD):

- 2.8% to 9.0% in human evaluation
- 4.8% to 16% in GPT4 evaluation

significantly impact the performance of systems

Performs well in DSFT paradigm:

- LLM-based method
- TPLM-based method


Performs well in RAG paradigm:

- LLM-based method
- Markdown format (surprise!)

Results and Analysis

RQ: What are the potential reasons for their different performances?

In DSFT Paradigm:



Freq (k)	C_1 · Markdown	C_2 · Template	C_3 · TPLM-based	C_4 · LLM-based
Term	821	1040	2358	2254
Verbs	313	315	682	1207

Absolute frequency of verbs and terms contained in the corpora C_i generated by different methods.

higher frequency of domain-specific terms and verbs leads to better system performance.

- *LM-based methods tend to supplement the domain entities as subjects/objects.
- Template methods use more pronouns, and monotonous predicates.
- Markdown format only retains the original content in the tables.

Results and Analysis

RQ: What are the potential reasons for their different performances?

In RAG Paradigm:

Under the same LLM reader setup:

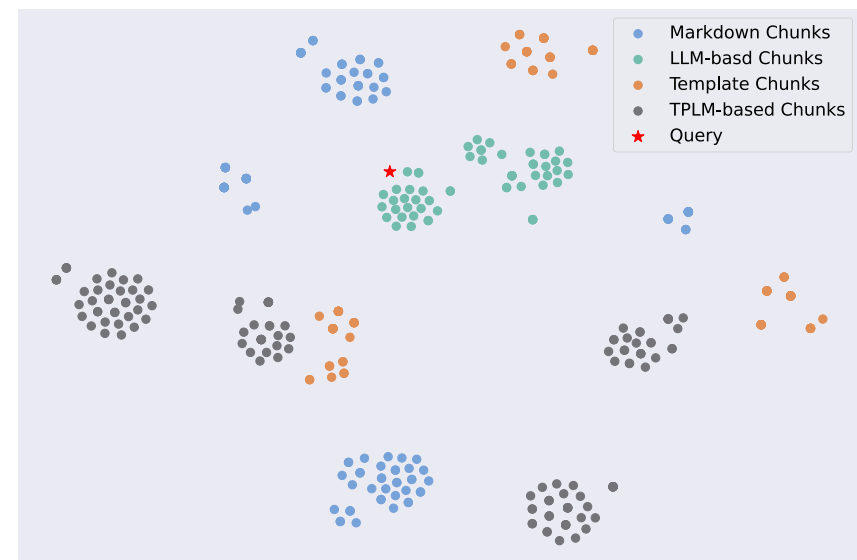
Semantic representations quality



Retrieval accuracy



RAG performance



A t-SNE visualization of chunk clusters in the embedding space.

Retrieval-friendly method:

LLM-based

Markdown format

Observations

Some practical suggestions for choosing table-to-text methods

Ready-to-use tips

DSFT Paradigm:

- LLM-based method (Pros: best performance; Cons: GPU/API cost, Data leakage risks)
- TPLM-based (Can well-tuned on this task. A good alternative for LLM)

RAG Paradigm:

- LLM-based method
 - best performance
- Markdown format (viable substitute)
 - ✓ easy-to-use
 - ✓ GPU-Free

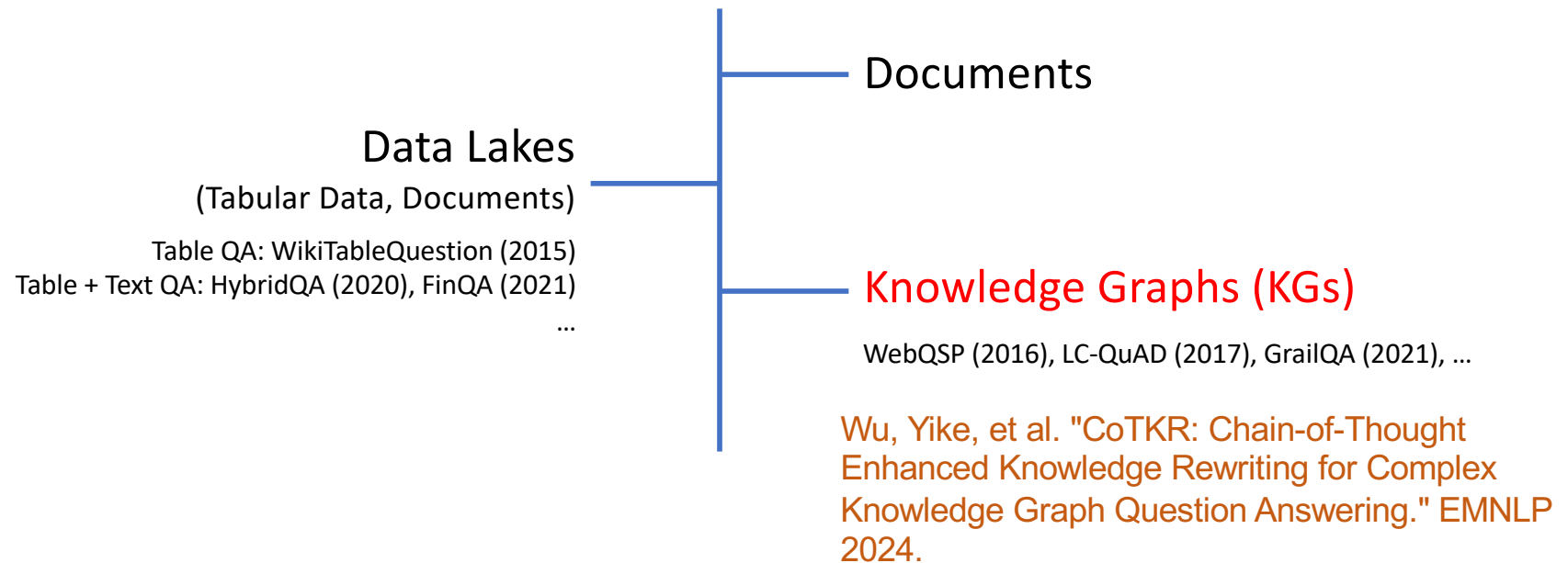
Freq (Avg.)	Markdown	Template	TPLM-based	LLM-based
Text Len	998	1259	1138	897

More Concise Text



Less Memory Costs
Less GPU Costs

Data Intensive QA



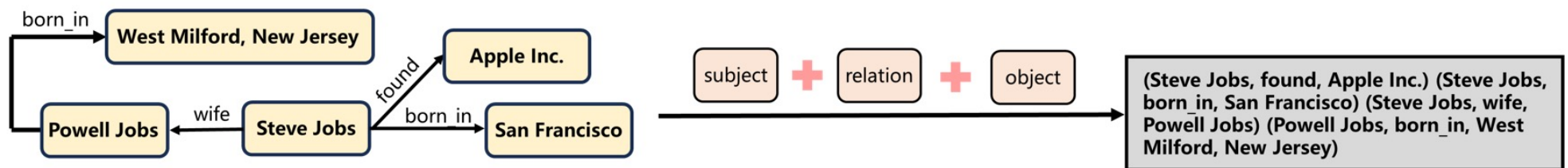
KG-augmented LLM for QA

- KG QA with Retrieval-augmented Generation
- **Challenge:** transform question-related subgraphs into natural language (**knowledge rewriting**) that LLMs can understand while preserving the structural information
- **This work:** design a novel knowledge rewriting method for KGQA

Previous Knowledge Rewriting Methods

Triple

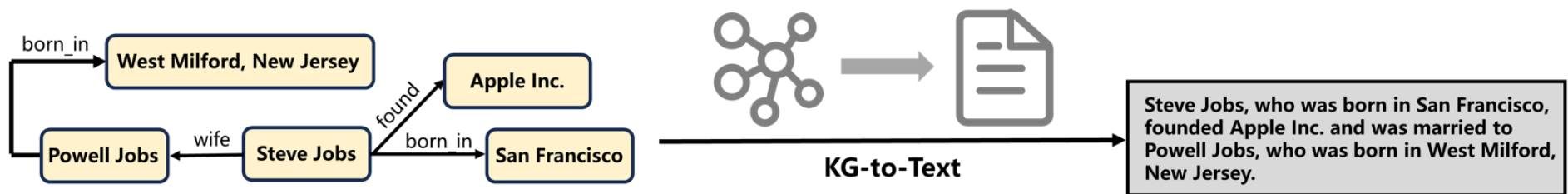
- Concatenate the subject, predicate, and object of a triple
- No need for additional models for knowledge rewriting



Previous Knowledge Rewriting Methods

KG-to-Text

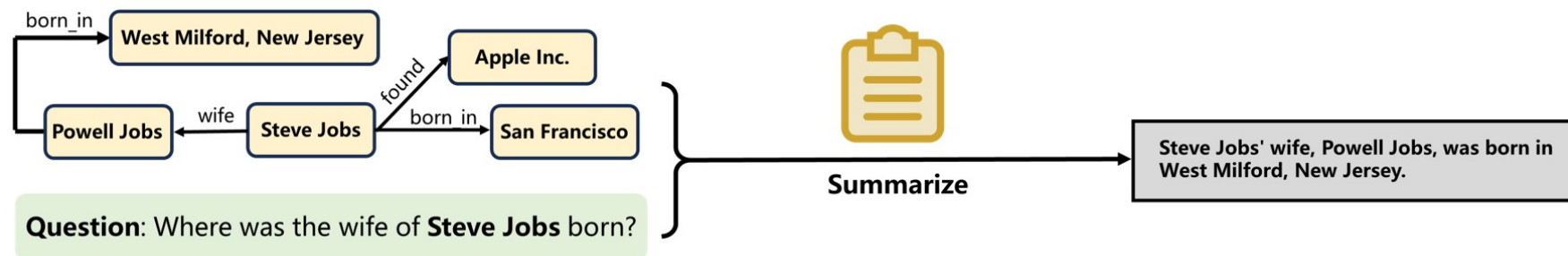
- Transform facts into free-form text with a KG-to-Text model
- Address the limitations of LLMs in understanding structured triple-form text



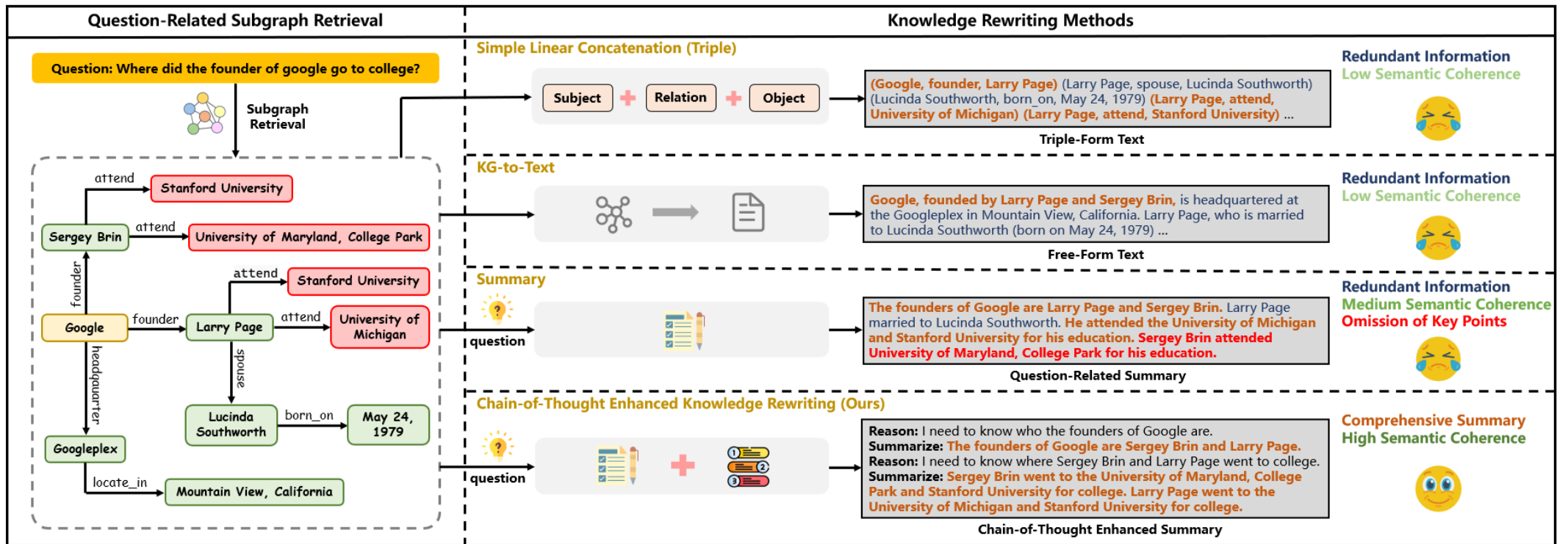
Previous Knowledge Rewriting Methods

Summary

- Convert triples into a question-relevant summary
- Alleviate the issue of redundant contextual knowledge

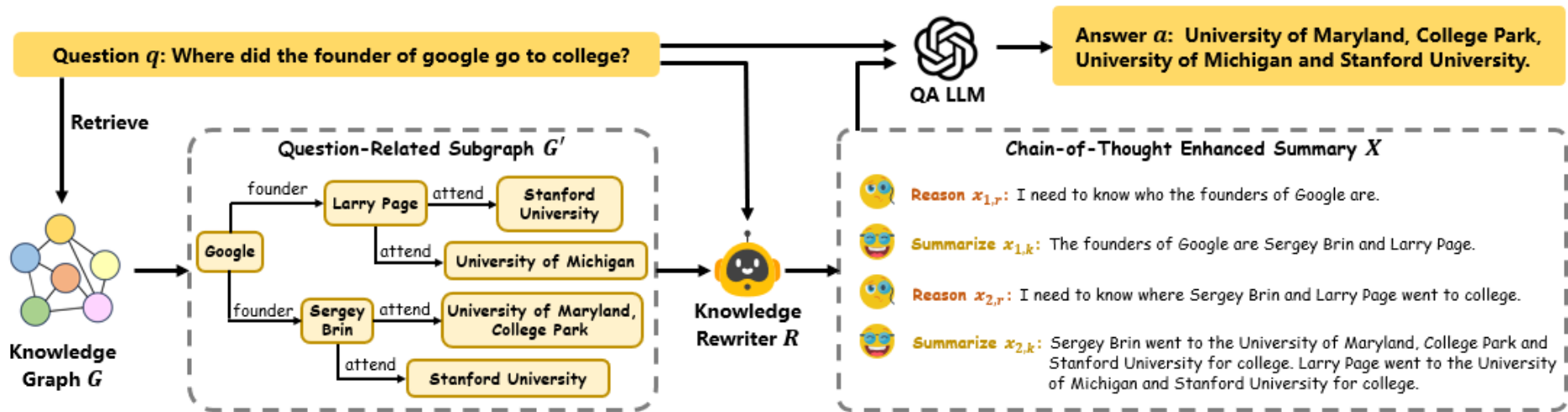


Motivation



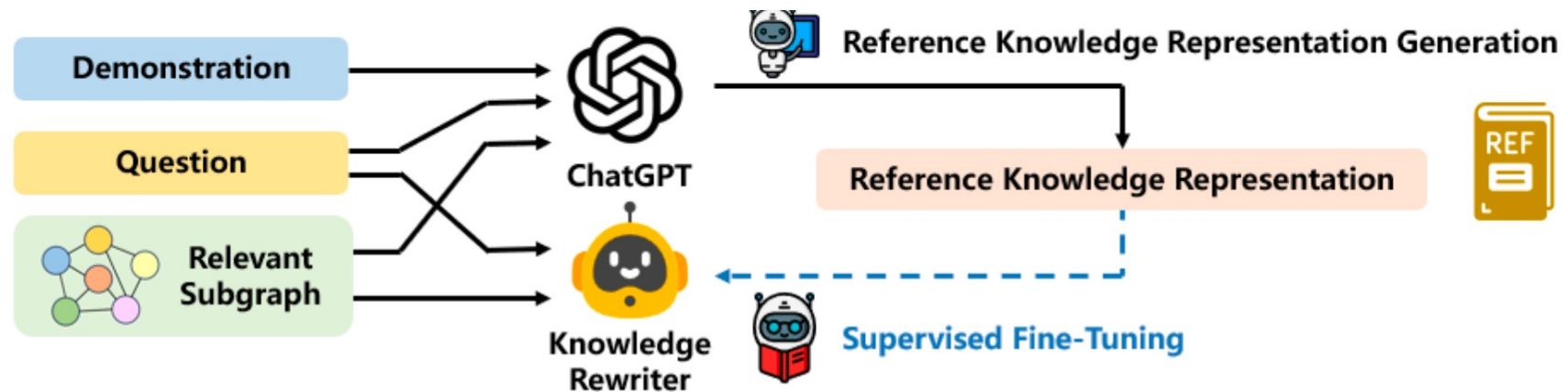
CoTKR: Chain-of-Thought Enhanced Knowledge Rewriting

Our knowledge rewriter alternatively conducts reasoning and summarizing.



Extending CoTKR with Training

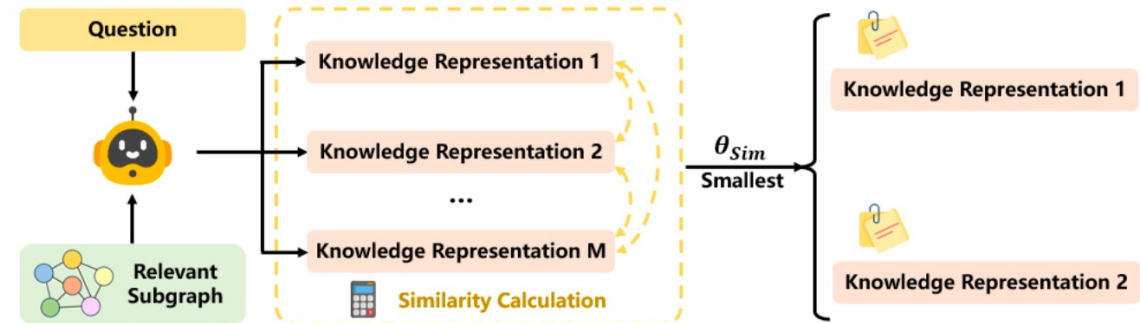
- **Stage 1: Supervised Learning with knowledge distilled from ChatGPT**



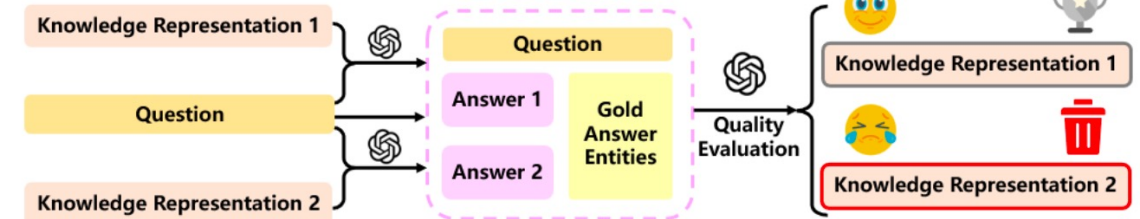
Extending CoTKR

- **Stage 2: PAQAF:**
Preference alignment
from Question
Answering Feedback

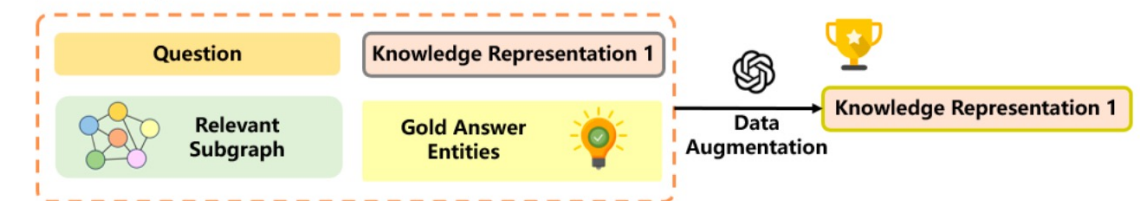
Step 1: Candidate Knowledge Representation Sampling



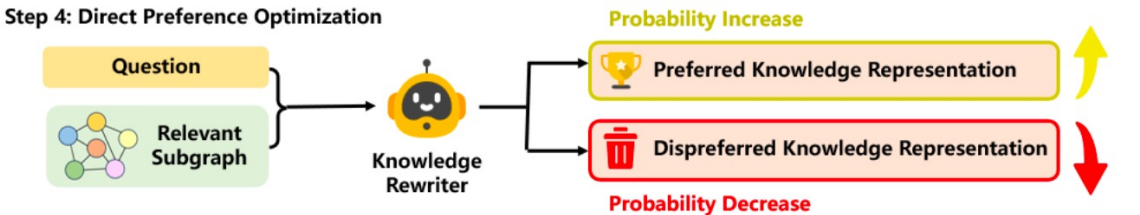
Step 2: Preference Annotation based on Question Answering Feedback



Step 3: Data Augmentation based on ChatGPT



Step 4: Direct Preference Optimization



Experiment Settings

- Benchmark

GrailQA, GraphQuestions

- Large Language Models

Knowledge Rewriter: Llama-2-7B-Chat, Llama-3-8B-Instruct, ChatGPT.

QA model: ChatGPT, Mistral-7B-Instruct-v0.3.

Experiment Settings

Retrieval Methods

2-Hop: Retain 30 triples from the 2-hop subgraph of the head entity, prioritizing those with higher semantic similarity to the question.

BM25: Linearize the 1-hop subgraph of the entity as the article. We take the top 30 triples corresponding to the candidate documents as the retrieval results.

Ground Truth Subgraph (GS): We modify the SPARQL queries from the datasets to obtain the ground truth subgraphs. These subgraphs represent the results of an ideal retriever.

Experiment Settings

Evaluation Metrics

Accuracy (Acc) measures whether the generated answer includes at least one correct answer entity.

Recall measures the proportion of correct entities present in the model's response.

Exact Match (EM) evaluates whether the response contains all the answer entities.

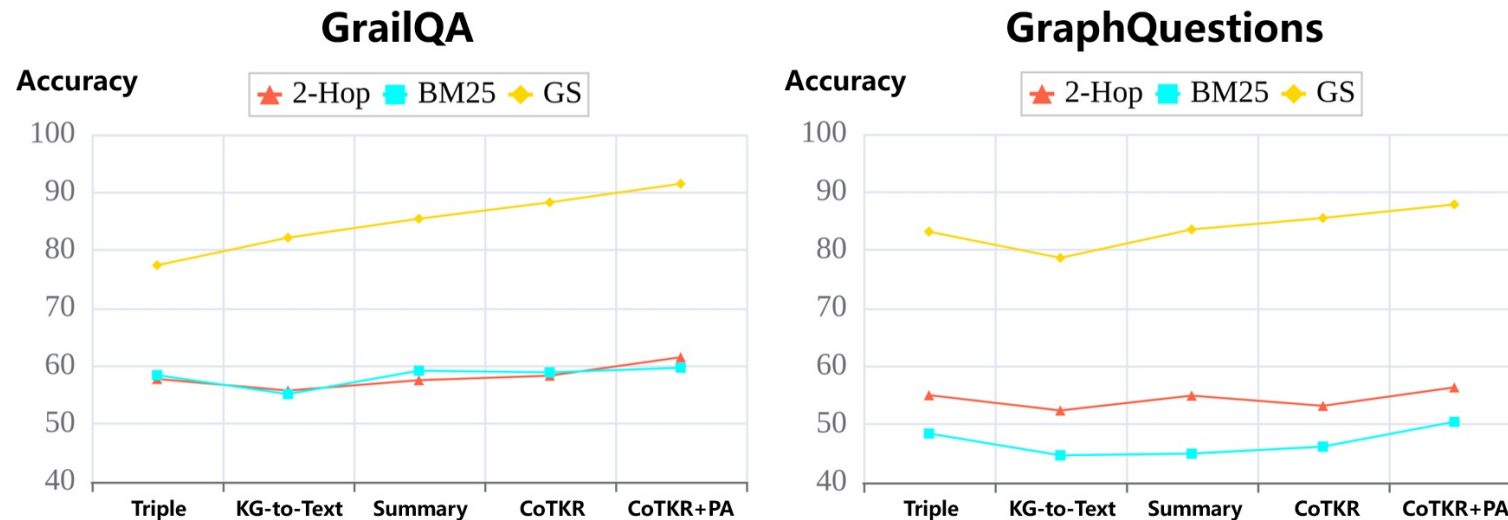
Main Results

- CoTKR surpasses the baselines, demonstrating its effectiveness.
- CoTKR+PA matches or outperforms ChatGPT as a knowledge rewriter, demonstrating the effectiveness of the training framework.
- A well-crafted knowledge representation is crucial for LLM used in KGQA.

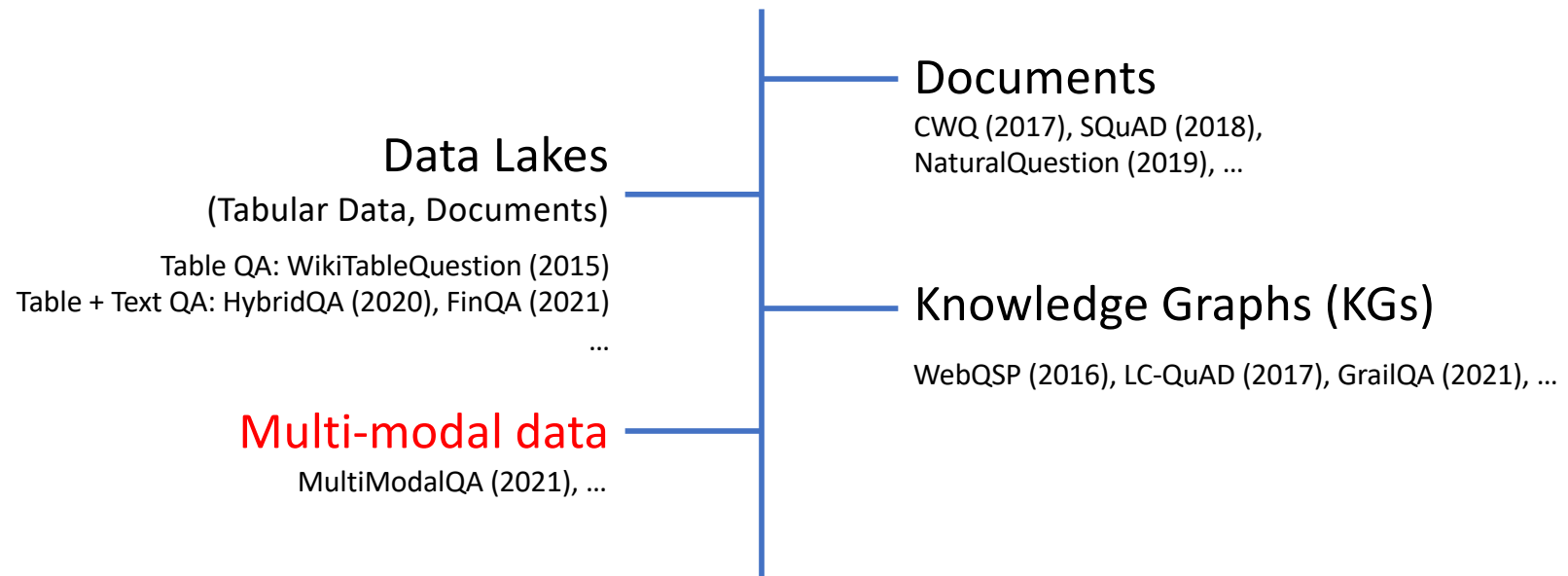
KR LLMs	Methods	GrailQA			GraphQuestions		
		Acc	Recall	EM	Acc	Recall	EM
ChatGPT as QA model							
None	No Knowledge	28.91	22.81	20.14	35.87	25.76	22.09
	Triple	57.76	49.67	44.73	55.03	46.65	41.63
Llama-2	KG-to-Text	54.75	47.35	42.44	49.73	40.00	33.74
	Summary	58.14	51.38	46.38	52.94	44.70	38.41
	CoTKR	58.64	52.33	47.88	51.36	45.20	39.96
	CoTKR+PA	59.25	53.52	49.64	56.78	47.99	42.46
Llama-3	KG-to-Text	55.76	48.41	43.90	52.40	45.06	39.83
	Summary	57.55	51.06	46.80	54.95	46.86	40.75
	CoTKR	58.33	52.55	48.65	53.19	47.23	43.17
	CoTKR+PA	61.51	56.08	52.67	56.37	49.31	45.26
ChatGPT	KG-to-Text	56.32	49.05	44.73	53.53	45.59	41.17
	Summary	58.54	51.81	47.29	55.62	48.93	44.97
	CoTKR	59.87	53.19	49.02	54.28	48.18	44.68
Mistral as QA model							
None	No Knowledge	29.44	23.13	20.30	38.20	26.92	22.13
	Triple	54.47	47.78	43.25	51.32	45.97	41.67
Llama-2	KG-to-Text	49.49	42.91	38.41	44.59	37.98	32.82
	Summary	54.10	47.79	43.15	49.85	42.33	36.45
	CoTKR	56.75	51.10	46.71	50.19	43.73	38.54
	CoTKR+PA	58.15	52.98	49.13	55.07	47.02	41.71
Llama-3	KG-to-Text	50.64	44.32	40.13	49.06	43.04	38.25
	Summary	53.84	47.71	43.49	52.03	44.30	38.50
	CoTKR	56.47	51.33	47.36	52.65	46.48	42.21
	CoTKR+PA	59.31	54.13	50.24	54.82	47.76	43.09
ChatGPT	KG-to-Text	51.04	44.87	40.97	49.14	43.04	38.83
	Summary	54.44	48.16	43.97	52.28	47.10	43.30
	CoTKR	57.28	51.14	47.09	52.82	47.13	43.55

Impact of Retrieval Methods

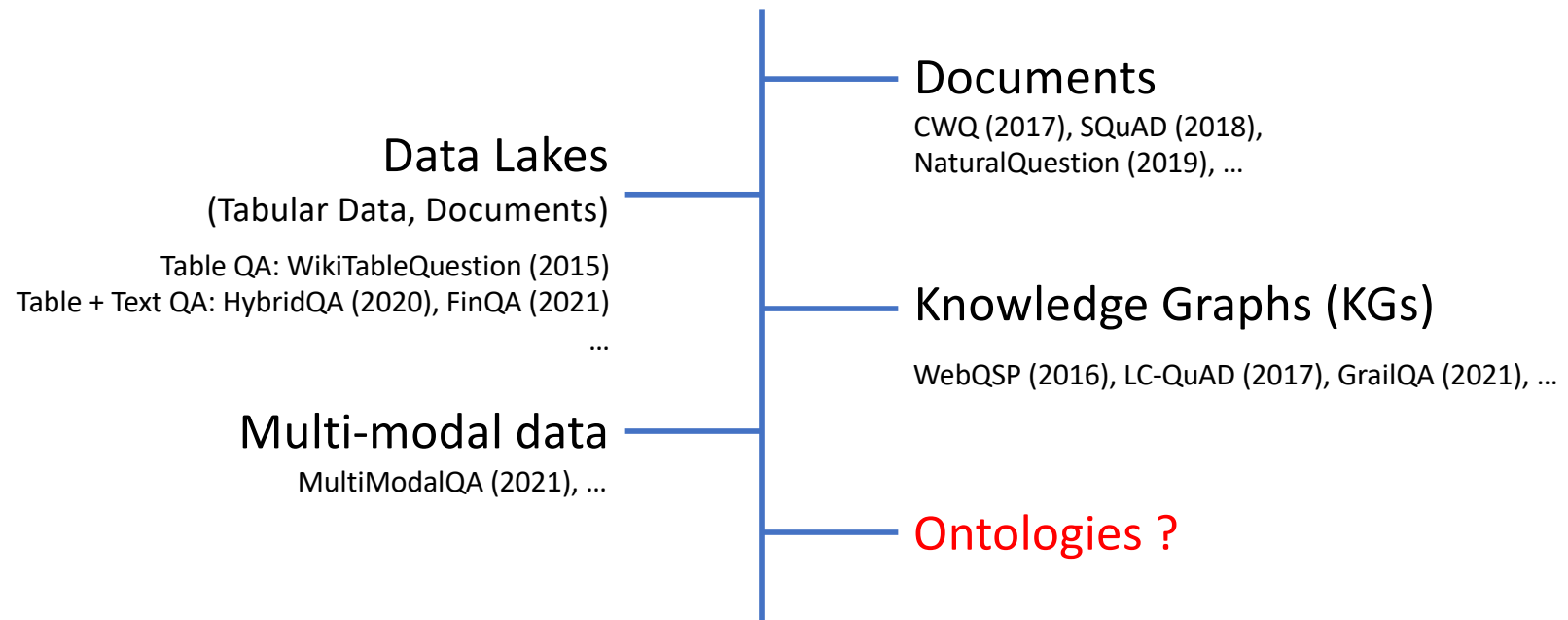
- 2-Hop: may be insufficient for challenging questions, but suitable for simpler ones.
- The design of a high-quality retriever remains an open problem.
- CoTKR performs best across various retrieval methods



Data Intensive QA



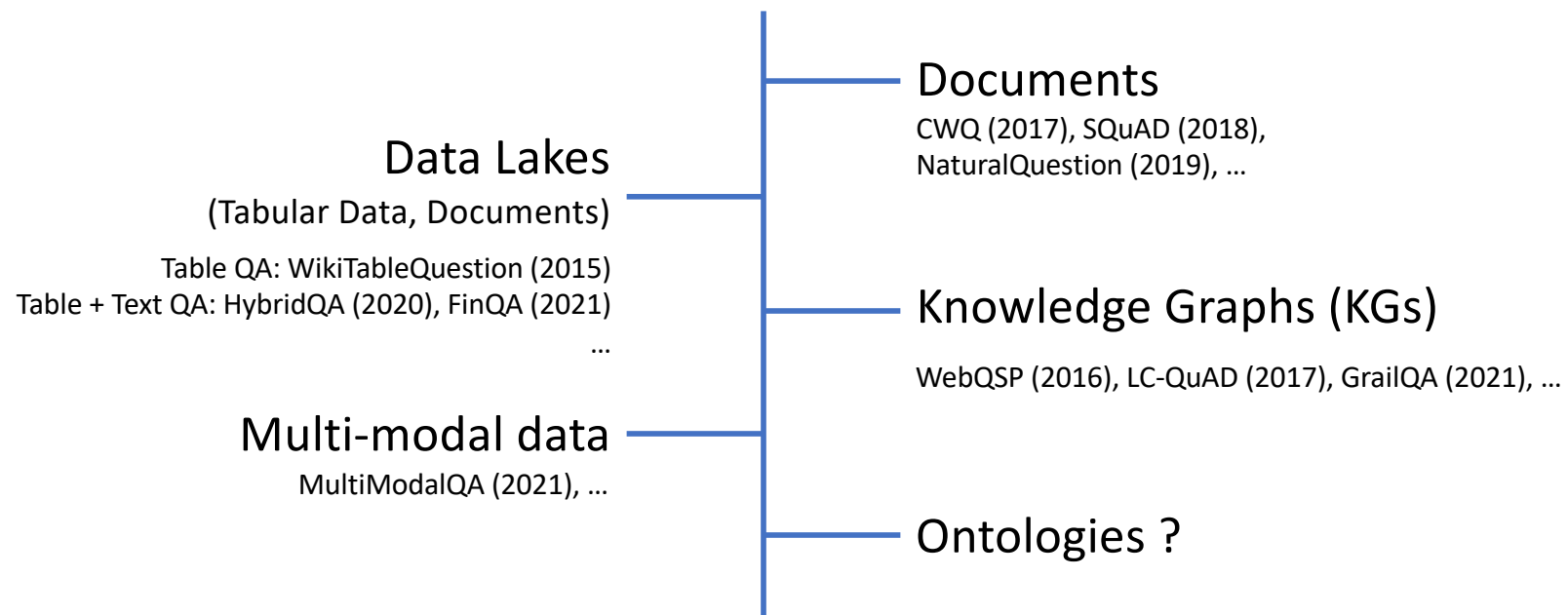
Data Intensive QA



Ontology in RAG (Open Questions)

- How to support knowledge retrieval with **ontology embedding**?
- How to combine ontologies with data lakes and/or KGs?
- Can we use ontology to manage and integrate the data sources and snippets for RAG?
 - How to design and construct such an ontology from data?

Data Intensive QA



Attribution

Hu, Nan, et al. "Benchmarking large language models in complex question answering attribution using knowledge graphs." 2024.

Attribution of QA

Attribution Category	Examples
Supportive	<p>Question: Who plays Fruma Sarah in Fiddler on the Roof?</p> <p>Answer: Fruma Sarah is a character in the musical “Fiddler on the Roof”, and Ruth Madoc played the role [1].</p> <p>Citations: [1] ... In 1971 Ruth Madoc played Fruma Sarah in the film version of the musical “Fiddler on the Roof”, and in 1972 she appeared as ...</p>
Partially Supportive	<p>Question: Who plays Patrick in 10 Things I Hate About You?</p> <p>Answer: Patrick is played by actor Heath Ledger in the film 10 Things I Hate About You [1].</p> <p>Citations: [1] 10 Things I Hate About You is a 1999 American teen romantic comedy-drama film directed by Gil Junger and starring Heath Ledger, Julia Stiles, Joseph Gordon-Levitt, and Larisa Oleynik. The screenplay, written by ...</p>
Contradictory	<p>Question: Who directed a George Pal’s production?</p> <p>Answer: George Pal directed a production called Puppertoons [1].</p> <p>Citations: [1] ... The Puppetoon Movie is a 1987 animated film written, produced, and directed by Arnold Leibovit ...</p>
Irrelevant	<p>Question: Who played the weasley brothers in Harry Potter?</p> <p>Answer: James and Oliver Phelps, identical twin actors, played the roles of Fred and George Weasley in the Harry Potter film series [1].</p> <p>Citations: [1] Chris Rankin plays of “Bugsy Malone”, “The Lion, The Witch and The Wardrobe” and Harry Potter series ... he plays a brother of Harry Potter’s best friend, ...</p>

Research Problem

- How to evaluate the attribution of QA systems?

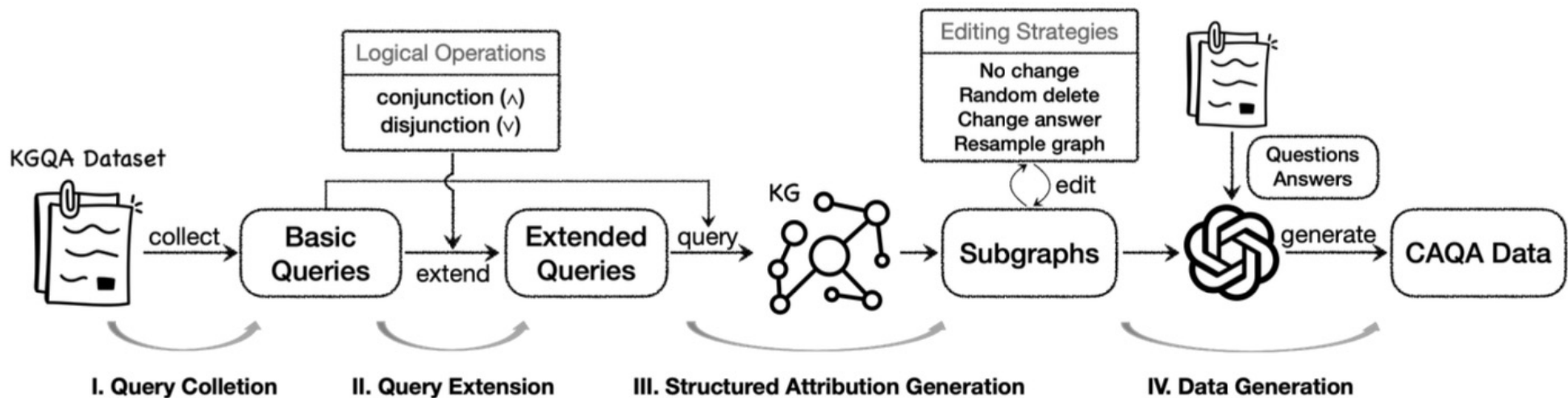
Manul evaluator

Automatic evaluator

What is the reliability of an automatic evaluator?

Complex Attributed QA (CAQA) Benchmark

- Expect: each answer associated with a gold attribution of **supportive**, **partially supportive**, **contradictory**, and **irrelevant**
- Input: a KGQA dataset (question annotated by SPARQL query)



Query types: Single-triple, path-like, tree-like

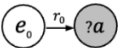
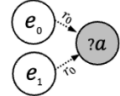

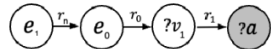
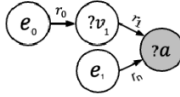
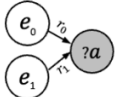
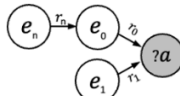
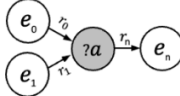
Complex Att

- Different complexities in reasoning

Complexity	Examples	Reasoning Graphs
Single	<p>Question: Which radio program episode appears in All Things Considered?</p> <p>Answer: The radio program episode in which All Things Considered appears is Remorse: The 14 Stories of Eric Morse [1].</p> <p>Citations: [1] <i>Remorse: The 14 Stories of Eric Morse is an episode of the radio program All Things Considered....</i></p>	
Union	<p>Question: Which university did Rick Scott attend?</p> <p>Answer: Rick Scott attended the University of Missouri–Kansas City and Southern Methodist University [1][2].</p> <p>Citations: [1] <i>Rick Scott graduated from the University of Missouri–Kansas City ...</i> [2] <i>Rick Scott earned a juris doctor degree by working his way through Southern Methodist University, ...</i></p>	
Intersection	<p>Question: The computer designer for Macintosh 128k and NeXT computer was whom?</p> <p>Answer: The computer designer for Macintosh 128k and NeXT computer was Steve Jobs [1][2].</p> <p>Citations: [1] <i>The computer designer for Macintosh 128k was Jerry Manock, who worked with Steve Jobs to develop the vertical body ...</i> [2] <i>Several former Apple employees followed Jobs to NeXT, including Joanna Hoffman, Bud Tribble, George Crow, Rich Page...</i></p>	
Concatenation	<p>Question: What are the official languages in the politician Mohammad Najibullah's country?</p> <p>Answer: Pashto and Dari are the official languages in the politician Mohammad Najibullah's country. [1][2].</p> <p>Citations: [1] <i>Mohammad Najibullah was the president of Afghanistan from 1986 to 1992 ...</i> [2] <i>Afghanistan s a multilingual country, where Pashto and Dari (a dialect of Persian) are the official languages with ...</i></p>	

Complex Attributed QA (CAQA) Benchmark

- Query Expansion

Original Query l			Extended Query l'		
Definitions	Structures	Examples	Definitions	Structures	Examples
$(e_0, r_0, ?a)$	S.		$(e_0, r_0, ?a)$ $\vee (e_1, r_0, ?a) \vee \dots \vee (e_m, r_0, ?a)$	U.	
$[e_0, r_0, ?v_1, \dots, ?v_{n-1}, r_{n-1}, ?a]$	P.		$[e_0, r_0, ?v_1, \dots, ?v_{n-1}, r_{n-1}, ?a]$ $\wedge (e_1, r_n, e_0)$	P.	
			$[e_0, r_0, ?v_1, \dots, ?v_{n-1}, r_{n-1}, ?a]$ $\wedge (e_1, r_n, ?a)$	T.	
$\wedge_{i=0}^{n-1} (e_i, r_i, ?a)$	T.		$\wedge_{i=0}^{n-1} (e_i, r_i, ?a), i \neq k$ $\wedge (e_n, r_n, e_k) \wedge (e_k, r_k, ?a)$	T.	
			$\wedge_{i=0}^{n-1} (e_i, r_i, ?a) \wedge (e_n, r_n, ?a)$	T.	

Complex Attributed QA (CAQA) Benchmark

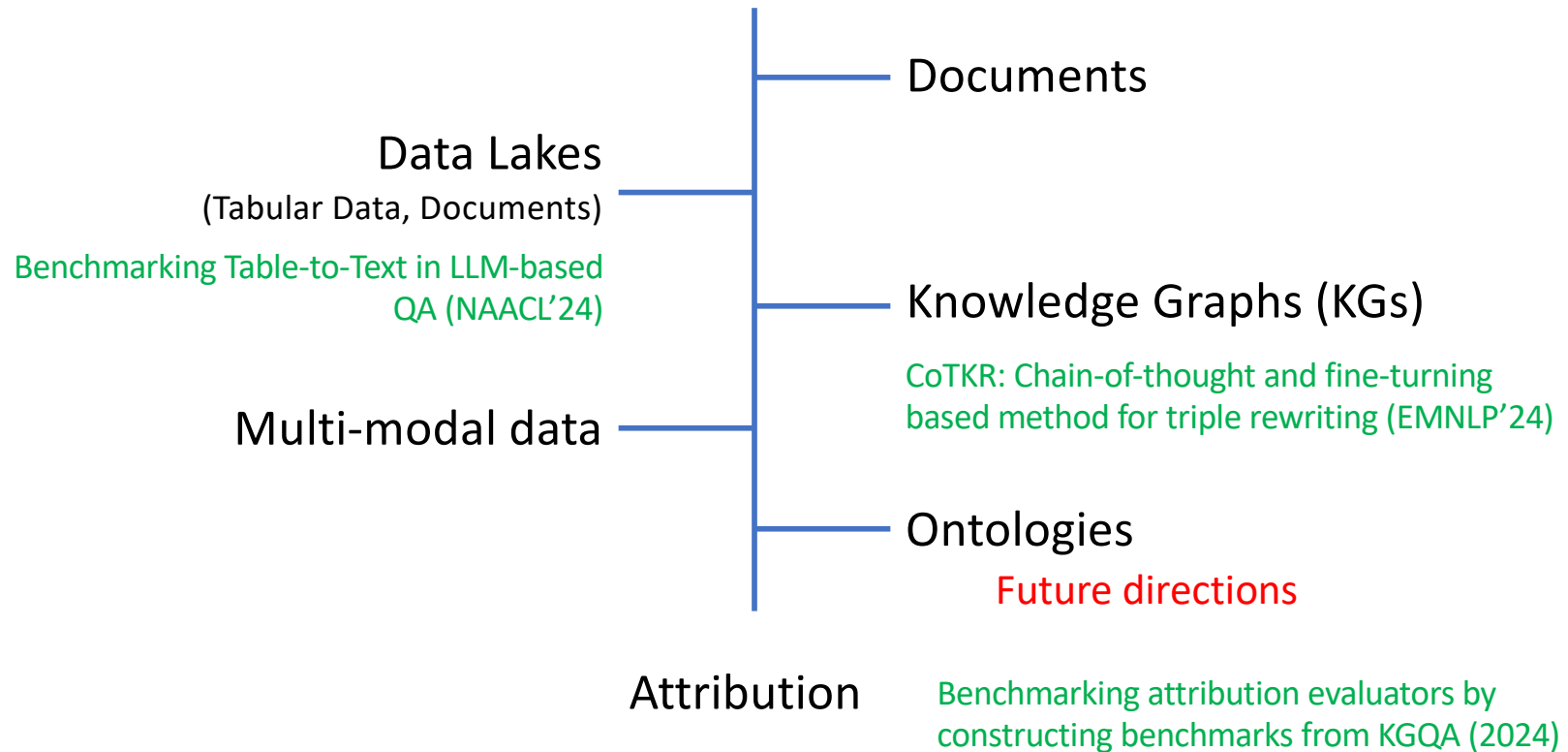
- GPT-3.5-turbo for text generation
- GrailQA & WebQuestionsSP
 - Freebase
- CAQA: 161K samples

Classes		Train	Test	Total
		137,211	23,963	161,174
Category	Sup.	39,489	6,668	46,157
	Ins.	28,868	5,065	33,933
	Con.	36,620	6,423	43,043
	Irr.	32,234	5,807	38,041
Complexity	S.	73,795	10,443	84,238
	C.	46,783	8,455	55,238
	U.	5,347	886	6,233
	I.	11,286	4,179	15,465

Results

Settings	Evaluators (Size)	Category					Complexity			
		Sup.	Par.	Con.	Irr.	Overall	S.	C.	I.	U.
Zero-Shot	LLaMA-2 (7B)	0.423	0.121	0.057	0.170	0.279	0.286	0.249	0.282	0.260
	LLaMA-2 (13B)	0.418	0.164	0.161	0.125	0.279	0.314	0.270	0.303	0.253
	LLaMA-3 (8B)	0.467	0.120	0.072	0.007	0.296	0.304	0.271	0.283	0.259
	Mistral (7B)	0.456	0.178	0.191	0.153	0.305	0.315	0.281	0.294	0.265
	Vicuna (7B)	0.513	0.100	0.064	0.199	0.327	0.343	0.273	0.312	0.256
	Vicuna (13B)	0.634	0.211	0.393	0.275	0.405	0.432	0.314	0.361	0.374
	LLaMA-3 (70B)	0.746	0.104	0.653	0.592	0.525	0.645	0.279	0.305	0.578
	GPT-3.5-turbo	0.583	0.017	0.598	0.512	0.497	0.555	0.321	0.363	0.363
	GPT-4	0.771	0.456	0.745	0.473	0.630	0.685	0.451	0.514	0.616
Few-Shot	LLaMA-2 (7B)	0.300	0.066	0.009	0.334	0.248	0.259	0.218	0.167	0.308
	LLaMA-2 (13B)	0.419	0.199	0.167	0.089	0.272	0.274	0.271	0.233	0.267
	LLaMA-3 (8B)	0.573	0.202	0.234	0.156	0.336	0.356	0.279	0.310	0.294
	Mistral (7B)	0.412	0.152	0.041	0.415	0.349	0.339	0.278	0.300	0.271
	Vicuna (7B)	0.578	0.183	0.081	0.324	0.325	0.337	0.272	0.354	0.311
	Vicuna (13B)	0.633	0.208	0.383	0.288	0.403	0.427	0.315	0.397	0.374
	LLaMA-3 (70B)	0.741	0.182	0.608	0.584	0.521	0.628	0.295	0.314	0.563
	GPT-3.5-turbo	0.602	0.031	0.340	0.604	0.467	0.512	0.324	0.384	0.368
	GPT-4	0.794	0.520	0.728	0.653	0.680	0.745	0.492	0.473	0.559
Fine-Tuning	LLaMA-2 (7B)	0.922	0.897	0.944	0.933	0.926	0.923	0.815	0.931	0.921
	LLaMA-2 (13B)	0.929	0.907	0.938	0.923	0.925	0.954	0.824	0.936	0.939
	LLaMA-3 (8B)	0.935	0.901	0.935	0.928	0.926	0.935	0.820	0.930	0.924
	Mistral (7B)	0.927	0.908	0.944	0.849	0.882	0.935	0.831	0.921	0.905
	Vicuna (7B)	0.937	0.907	0.940	0.906	0.932	0.956	0.823	0.936	0.939
	Vicuna (13B)	0.942	0.923	0.939	0.923	0.933	0.950	0.847	0.935	0.940

Summary



Summary (Ontology, LLM, RAG)

- How to support knowledge retrieval with **ontology embedding**?
e.g., HiTs (NeurIPS'24), OWL2Vec* (MLJ 2021), Box²EL (WWW'24)
Ontology Embedding Survey (2024, <https://arxiv.org/abs/2406.10964>)
- How to combine ontologies with data lakes and/or KGs?
- Can we use ontology to manage and integrate the data sources and snippets for RAG?
How to design and construct such an ontology from data?
E.g., the DeepOnto library (<https://github.com/KRR-Oxford/DeepOnto>), Data Lake Schema Inference (ongoing)

Thanks for your attention

Q&A

jiaoyan.chen@manchester.ac.uk