Title

StructRAG: Boosting Knowledge Intensive Reasoning of LLMs via Inference-time Hybrid Information Structurization

ICLR 2025

Author

Zhuoqun Li (李卓群)

Ph.D. Student (from 2022.9)

Chinese Information Processing, ISCAS

Interest LLM and retrieval-augmented generation (RAG)

Supervised by Le Sun, Xianpei Han, Hongyu Lin, and Yaojie Lu

* This work is completed in **Tongyi Lab**, supervised by 水德 and 翼飞

ISCAS 中国科学院软件研究所 The Institute of Software Chinese Academy of Sciences

中文信息处理实验室 Chinese Information Processing Laboratory

Alibaba

Background

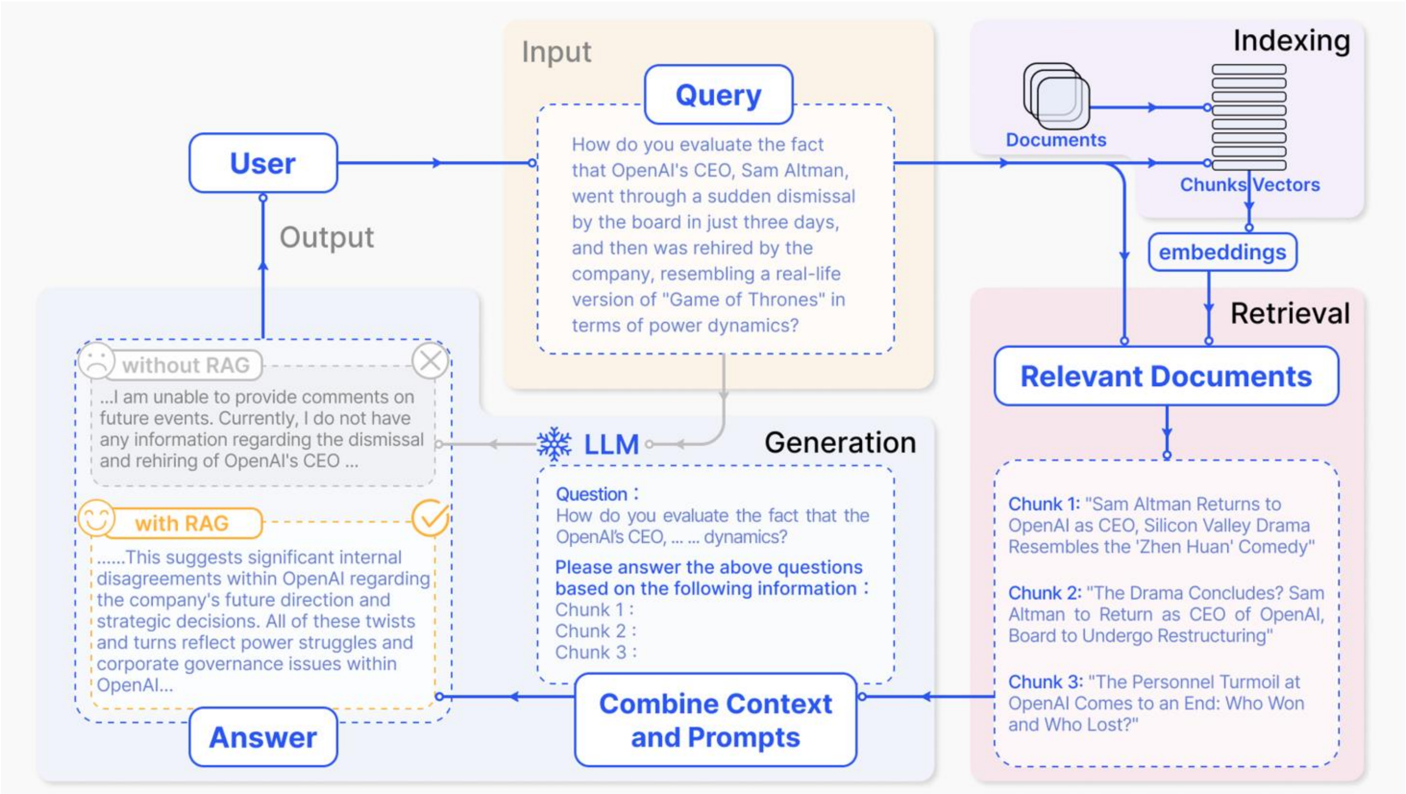
- LLM Factuality Hallucination
  - The content generated by the model is inconsistent with the facts of the real world.

Category	Type	Example Dialog	Explanation
Factuality Hallucination	Factual Contradiction	**Instruction:** What are Thomas Edison's main contributions to science and technology? **Response:** Thomas Edison developed the first practical telephone and invented the light bulb.	The response is factually incorrect. In reality, **Edison improved the light bulb, building on earlier designs by others**, and **Alexander Graham Bell invented the telephone**.
	Factual Fabrication	**Instruction:** What are the major environmental impacts of the construction of the Eiffel Tower? **Response:** The construction of the Eiffel Tower in 1889 led to the extinction of the Parisian tiger, a species that played a crucial role in the region's ecosystem. Additionally, it is widely recognized as the event that sparked the global green architecture movement.	The response erroneously states that the construction of the Eiffel Tower led to the extinction of the 'Parisian tiger' and sparked the global green architecture movement. In fact, **there is no evidence of a 'Parisian tiger' ever existing, making this a fabricated claim.** Moreover, **attributing the origins of green architecture to the Eiffel Tower is an exaggeration**, as this movement has diverse roots and cannot be traced back to a single event.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Background

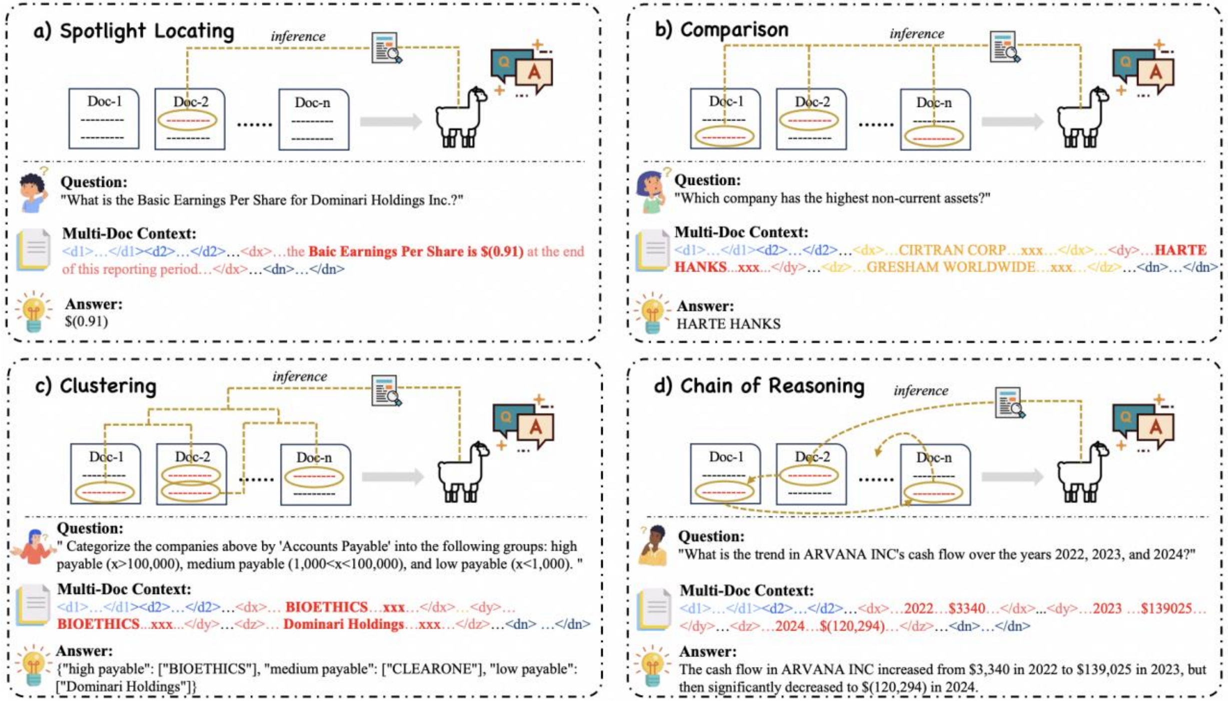
- RAG: Retrieval-Augmented Generation
  - Split documents into chunks and retrieve relevant chunks as external augmented knowledge.



Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey.

Task

- Knowledge-intensive Reasoning Task
  - Relevant information is scattered across multiple locations in the document library.
  - Need to make complex inferences based on valid information (e.g., *Financial Report Analysis*).



Minzheng Wang, Longze Chen, Cheng Fu, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang, and Yongbin Li. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa.

Task

- Existing chunk-based RAG **cannot** effectively solve knowledge-intensive reasoning tasks.

- High-noise External Knowledge
  - Due to scattered information, the retrieved chunks contain a significant amount of text noise.

- Hard to Inference
  - Based on high noise external knowledge, LLM cannot establish connections between information and therefore cannot make the effective inference.

Motivation

- How do human beings solve knowledge-intensive reasoning tasks ?

- Cognitive Load Theory
  - Transform scattered information into structured knowledge.
  - Using structured knowledge to assist in completing reasoning tasks.

- Cognitive Fit Theory
  - Use different types of structures for different types of tasks.
  - For example: Chunk, Table, Graph, Catalogue, and Algorithm.

Motivation

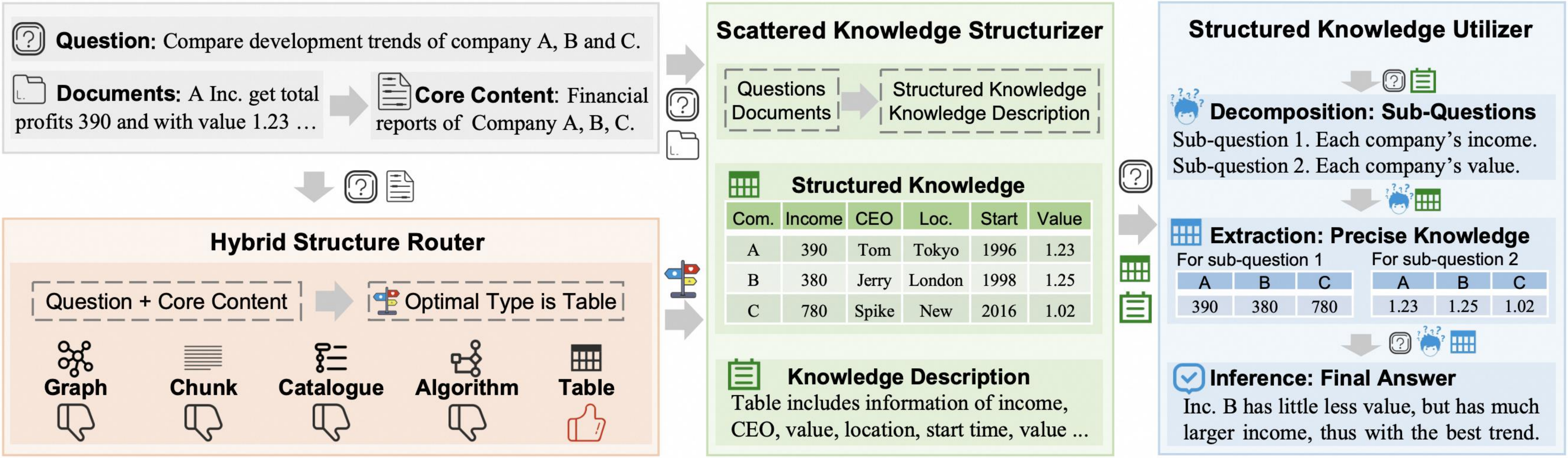
- Can LLMs learn from human beings to solve knowledge-intensive reasoning tasks ? **Yes !**

- Cognitive Similarity of LLMs and Human Beings
  - Chain of thought: *let's think step by step.*
  - OpenAI o1, DeepSeek R1: let LLMs do deep thinking as human beings.

- Powerful Structurization Capability of LLMs
  - Text-to-Table: A New Way of Information Extraction.
  - UIE (Unified Structure Generation for Universal Information Extraction).

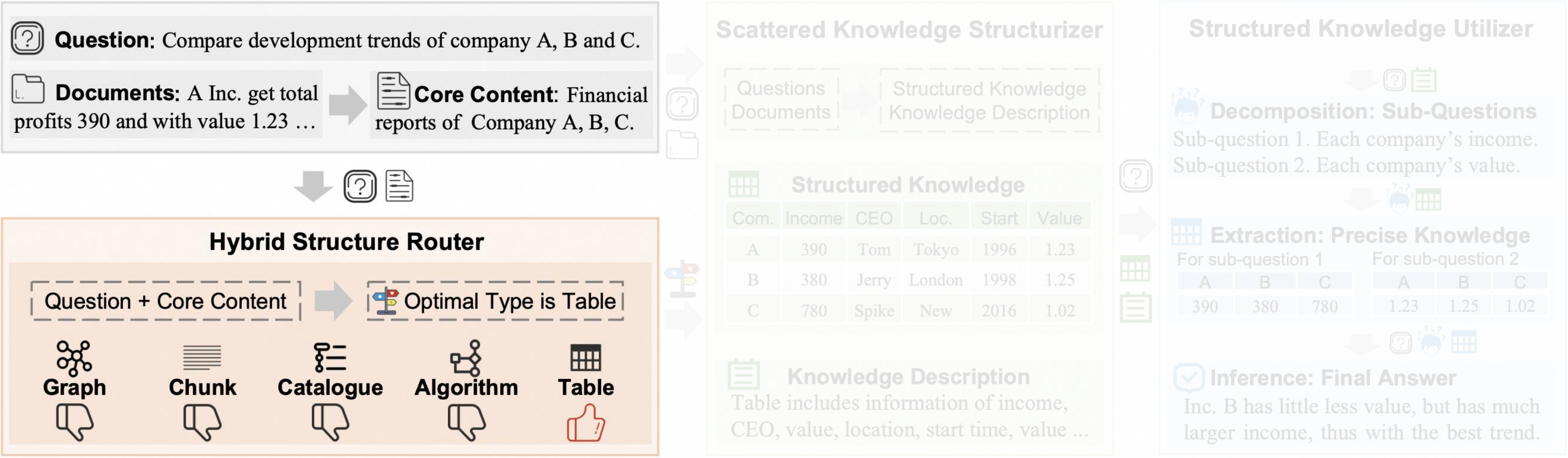
StructRAG

- Router → Structurizer → Utilizer
  - Boosting Knowledge Intensive Reasoning of LLMs via Inference-time Hybrid Information Structurization



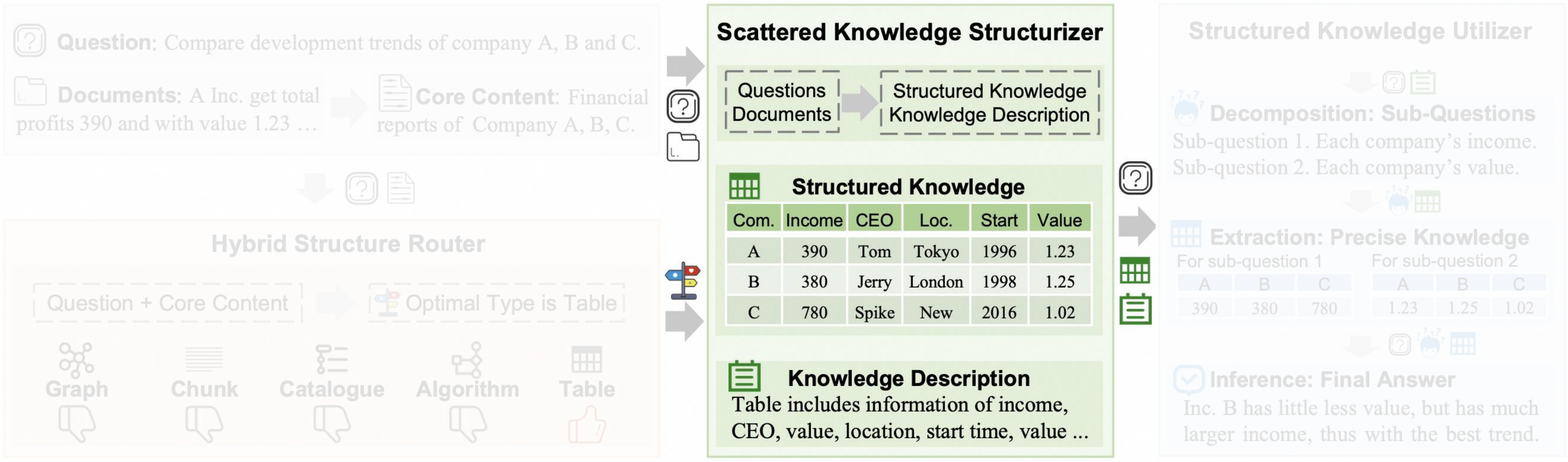
StructRAG

- **Router**: determine the optimal structure type based on the current task scenario.
  - Using LLM as the foundation, train a Router (which will be discussed in the next section).
  - **Input**: the query and the core content description of documents.
  - **Output**: the optimal structure types (currently including five possible structure types).



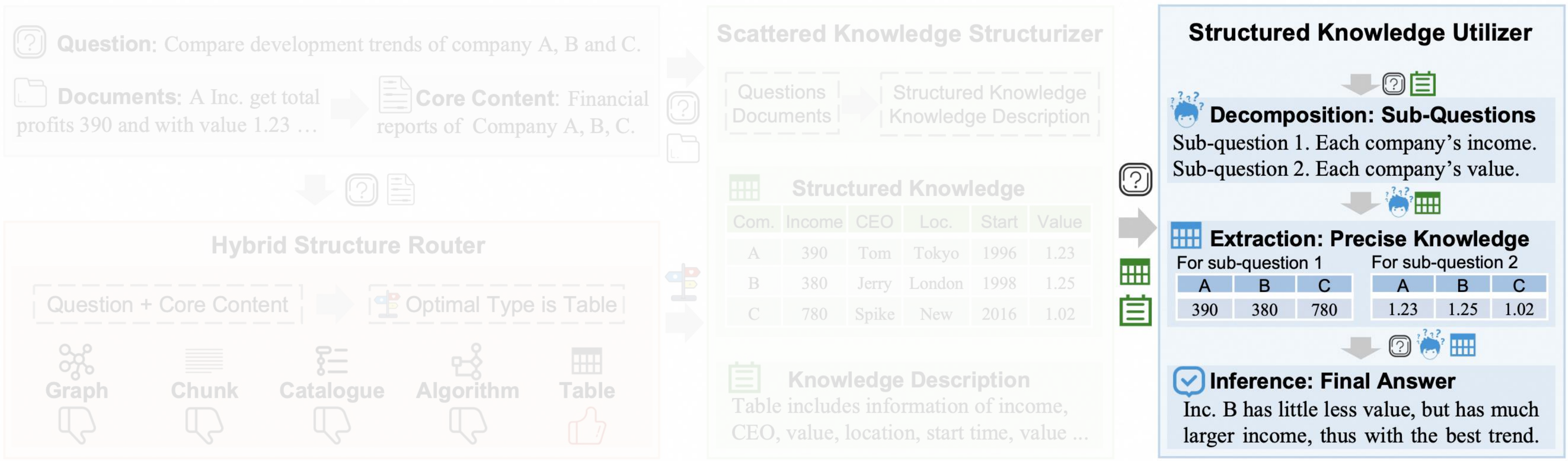
StructRAG

- **Structurizer**: transform scattered information into structured knowledge.
  - Leverage the capabilities of LLMs to achieve structured process based on manually designed prompts.
  - **Input**: the query and raw documents.
  - **Output**: selected type of structured knowledge and a brief description of the structured knowledge.



StructRAG

- **Utilizer**: use structured knowledge to reason out the answer.
  - **Decomposition**: decompose complex problem into subproblems based on structured knowledge descriptions.
  - **Extraction**: extract precise structured knowledge based on each subproblem.
  - **Inference**: Integrate all subproblems and precise knowledge to generate the final answer.



Router Training

- The router accuracy is the core factor of StructRAG performance.

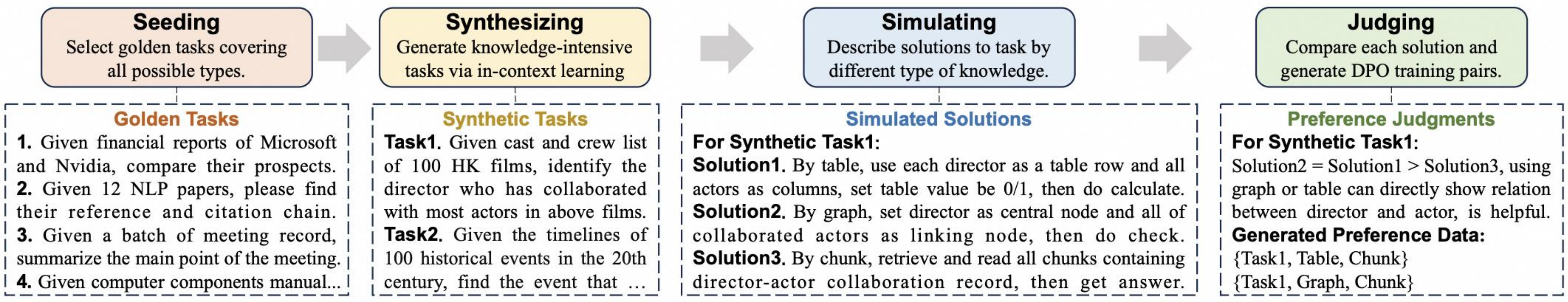
- Use DPO algorithm to train a high-performance router.
  - Set the question and core content of documents as input, the name of structure as output.

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(q,C,t_w,t_l)\sim D_{\text{synthetic}}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(t_w \mid q, C)}{\pi_{\text{ref}}(t_w \mid q, C)} - \beta \log \frac{\pi_\theta(t_l \mid q, C)}{\pi_{\text{ref}}(t_l \mid q, C)} \right) \right]$$

- However, how to construct the training dataset ?

Router Training

- LLM-based Training Data Construction
  - **Task Synthesis**: Generate more knowledge-intensive reasoning tasks based on the given seed tasks.
  - **Solution Simulation**: Simulate the process of solving problems by different types of structured knowledge.
  - **Preference Judgment**: Determine the superiority between different solutions and obtain training data for DPO.



Experiments

- Base Model
  - Qwen2-7B-Instruct (for Router) , Qwen2-72B-Instruct (for Structurizer and Utilizer)
- Dataset
  - Loong (Leave No Document Behind: Benchmarking Long-Context LLMs with Extended Multi-Doc QA)
- Baselines
  - **Long-context**: Directly input the entire document as external knowledge to the LLM.
  - **RAG**: Split document into multiple chunks, retrieve a small number of relevant chunks as external knowledge.
  - **RQ-RAG**: Based on RAG, iteratively perform query refinement based on the retrieval results.
  - **GraphRAG**: Construct the original document into a multi-level graph based on information extraction triples, and retrieve sub-graphs as external knowledge.

Experiments

- StructRAG achieves the state-of-the-art performance overall.

Method	Spot.		Comp.		Clus.		Chain.		Overall	
	LLM Score	EM	LLM Score	EM	LLM Score	EM	LLM Score	EM	LLM Score	EM
Set 1 (10K-50K Tokens)										
Long-context (Yang et al., 2024a)	68.49	**0.55**	60.60	0.37	47.08	0.08	**70.39**	**0.36**	60.11	0.29
RAG (Lewis et al., 2020)	51.08	0.35	44.53	0.27	37.96	0.05	53.95	0.35	46.11	0.23
RQ-RAG (Chan et al., 2024)	72.31	0.54	48.16	0.05	47.44	0.07	58.96	0.25	53.51	0.17
GraphRAG (Edge et al., 2024)	31.67	0.00	27.60	0.00	40.71	0.14	54.29	0.43	40.82	0.18
StructRAG (Ours)	**74.53**	0.47	**75.58**	**0.47**	**65.13**	**0.23**	67.84	0.34	**69.43**	**0.35**
Set 2 (50K-100K Tokens)										
Long-context (Yang et al., 2024a)	64.53	0.43	42.60	0.21	38.52	0.05	51.18	0.20	45.71	0.17
RAG (Lewis et al., 2020)	66.27	**0.46**	46.28	0.31	38.95	0.05	46.15	**0.22**	45.42	0.19
RQ-RAG (Chan et al., 2024)	57.35	0.35	50.83	0.16	42.85	0.03	47.60	0.10	47.09	0.10
GraphRAG (Edge et al., 2024)	24.80	0.00	14.29	0.00	37.86	0.00	46.25	0.12	33.06	0.03
StructRAG (Ours)	**68.00**	0.41	**63.71**	**0.36**	**61.40**	**0.17**	**54.70**	0.19	**60.95**	**0.24**
Set 3 (100K-200K Tokens)										
Long-context (Yang et al., 2024a)	46.99	0.27	37.06	0.13	31.50	0.02	35.01	0.07	35.94	0.09
RAG (Lewis et al., 2020)	**73.69**	**0.55**	42.20	0.27	32.78	0.02	37.65	0.13	42.60	0.18
RQ-RAG (Chan et al., 2024)	50.50	0.13	44.62	0.00	36.98	0.00	36.79	0.07	40.93	0.05
GraphRAG (Edge et al., 2024)	15.83	0.00	27.40	0.00	42.50	0.00	43.33	0.17	33.28	0.04
StructRAG (Ours)	68.62	0.44	**57.74**	**0.35**	**58.27**	**0.10**	**49.73**	**0.13**	**57.92**	**0.21**
Set 4 (200K-250K Tokens)										
Long-context (Yang et al., 2024a)	33.18	0.16	26.59	0.08	29.84	**0.01**	25.81	0.04	28.92	0.06
RAG (Lewis et al., 2020)	52.17	**0.24**	24.60	0.10	26.78	0.00	17.79	0.00	29.29	0.07
RQ-RAG (Chan et al., 2024)	29.17	0.08	40.36	0.00	26.92	0.00	34.69	0.00	31.91	0.01
GraphRAG (Edge et al., 2024)	17.50	0.00	26.67	0.00	20.91	0.00	33.67	0.33	23.47	0.05
StructRAG (Ours)	**56.87**	0.19	**55.62**	**0.25**	**56.59**	0.00	**35.71**	**0.05**	**51.42**	**0.10**

Experiments

- The advancements of StructRAG are even more apparent in more complex scenarios.

Method	Spot.		Comp.		Clus.		Chain.		Overall	
	LLM Score	EM	LLM Score	EM	LLM Score	EM	LLM Score	EM	LLM Score	EM
Set 1 (10K-50K Tokens)										
Long-context (Yang et al., 2024a)	68.49	**0.55**	60.60	0.37	47.08	0.08	**70.39**	**0.36**	60.11	0.29
RAG (Lewis et al., 2020)	51.08	0.35	44.53	0.27	37.96	0.05	53.95	0.35	46.11	0.23
RQ-RAG (Chan et al., 2024)	72.31	0.54	48.16	0.05	47.44	0.07	58.96	0.25	53.51	0.17
GraphRAG (Edge et al., 2024)	31.67	0.00	27.60	0.00	40.71	0.14	54.29	0.43	40.82	0.18
StructRAG (Ours)	**74.53**	0.47	**75.58**	**0.47**	**65.13**	**0.23**	67.84	0.34	**69.43**	**0.35**
Set 2 (50K-100K Tokens)										
Long-context (Yang et al., 2024a)	64.53	0.43	42.60	0.21	38.52	0.05	51.18	0.20	45.71	0.17
RAG (Lewis et al., 2020)	66.27	**0.46**	46.28	0.31	38.95	0.05	46.15	**0.22**	45.42	0.19
RQ-RAG (Chan et al., 2024)	57.35	0.35	50.83	0.16	42.85	0.03	47.60	0.10	47.09	0.10
GraphRAG (Edge et al., 2024)	24.80	0.00	14.29	0.00	37.86	0.00	46.25	0.12	33.06	0.03
StructRAG (Ours)	**68.00**	0.41	**63.71**	**0.36**	**61.40**	**0.17**	**54.70**	0.19	**60.95**	**0.24**
Set 3 (100K-200K Tokens)										
Long-context (Yang et al., 2024a)	46.99	0.27	37.06	0.13	31.50	0.02	35.01	0.07	35.94	0.09
RAG (Lewis et al., 2020)	**73.69**	**0.55**	42.20	0.27	32.78	0.02	37.65	0.13	42.60	0.18
RQ-RAG (Chan et al., 2024)	50.50	0.13	44.62	0.00	36.98	0.00	36.79	0.07	40.93	0.05
GraphRAG (Edge et al., 2024)	15.83	0.00	27.40	0.00	42.50	0.00	43.33	0.17	33.28	0.04
StructRAG (Ours)	68.62	0.44	**57.74**	**0.35**	**58.27**	**0.10**	**49.73**	**0.13**	**57.92**	**0.21**
Set 4 (200K-250K Tokens)										
Long-context (Yang et al., 2024a)	33.18	0.16	26.59	0.08	29.84	**0.01**	25.81	0.04	28.92	0.06
RAG (Lewis et al., 2020)	52.17	**0.24**	24.60	0.10	26.78	0.00	17.79	0.00	29.29	0.07
RQ-RAG (Chan et al., 2024)	29.17	0.08	40.36	0.00	26.92	0.00	34.69	0.00	31.91	0.01
GraphRAG (Edge et al., 2024)	17.50	0.00	26.67	0.00	20.91	0.00	33.67	0.33	23.47	0.05
StructRAG (Ours)	**56.87**	0.19	**55.62**	**0.25**	**56.59**	0.00	**35.71**	**0.05**	**51.42**	**0.10**

Experiments

➢ All three modules contribute positively to the overall framework.

Method	Set 1		Set 2		Set 3		Set 4		Overall	
	LLM Score	EM	LLM Score	EM	LLM Score	EM	LLM Score	EM	LLM Score	EM
StructRAG	**69.43**	**0.35**	**60.95**	**0.24**	**57.92**	**0.21**	**51.42**	**0.10**	**60.38**	**0.23**
w/o router	51.09	0.28	48.28	0.17	39.52	0.13	41.83	**0.10**	45.33	0.17
w/o structurizer	64.97	0.29	52.17	0.17	53.18	0.19	44.24	**0.10**	53.92	0.19
w/o utilizer	68.23	0.29	59.73	**0.24**	53.29	0.19	35.77	**0.10**	55.94	0.22

➢ Using any single fixed type of knowledge cannot achieve good performance on diverse tasks.

Method	Set 1		Set 2		Set 3		Set 4		Overall	
	LLM Score	EM	LLM Score	EM	LLM Score	EM	LLM Score	EM	LLM Score	EM
StructRAG	**69.43**	**0.35**	**60.95**	**0.24**	**57.92**	**0.21**	**51.42**	0.10	**60.38**	**0.23**
w/ only table	48.00	0.23	55.19	**0.24**	50.35	0.19	38.44	**0.12**	49.66	0.21
w/ only graph	30.59	0.09	24.05	0.05	17.46	0.03	20.96	0.04	22.71	0.05
w/ only chunk	64.97	0.29	52.17	0.17	53.18	0.19	44.24	0.10	53.92	0.19
w/ only catalogue	30.49	0.10	36.36	0.13	36.77	0.12	23.75	0.03	33.26	0.10
w/ only algorithm	43.53	0.24	32.86	0.08	31.59	0.13	16.67	0.04	32.32	0.12

Limitations

- Structure Setting
  - Design a set of structural types to cover a wider range of tasks.
- Structurization Process
  - Reduce the time consumed in constructing structured knowledge.
  - Improve the LLM’s ability to do structurization by post-training, such as reinforcement learning.
- Structured Knowledge Utilization
  - Simultaneously use multiple types of structured knowledge in one knowledge-intensive reasoning task.

End

Thanks for your listening！

lizhuoqun2021@iscas.ac.cn

https://github.com/Li-Z-Q/StructRAG

https://iclr.cc/virtual/2025/poster/30265

https://huggingface.co/papers/2410.08815