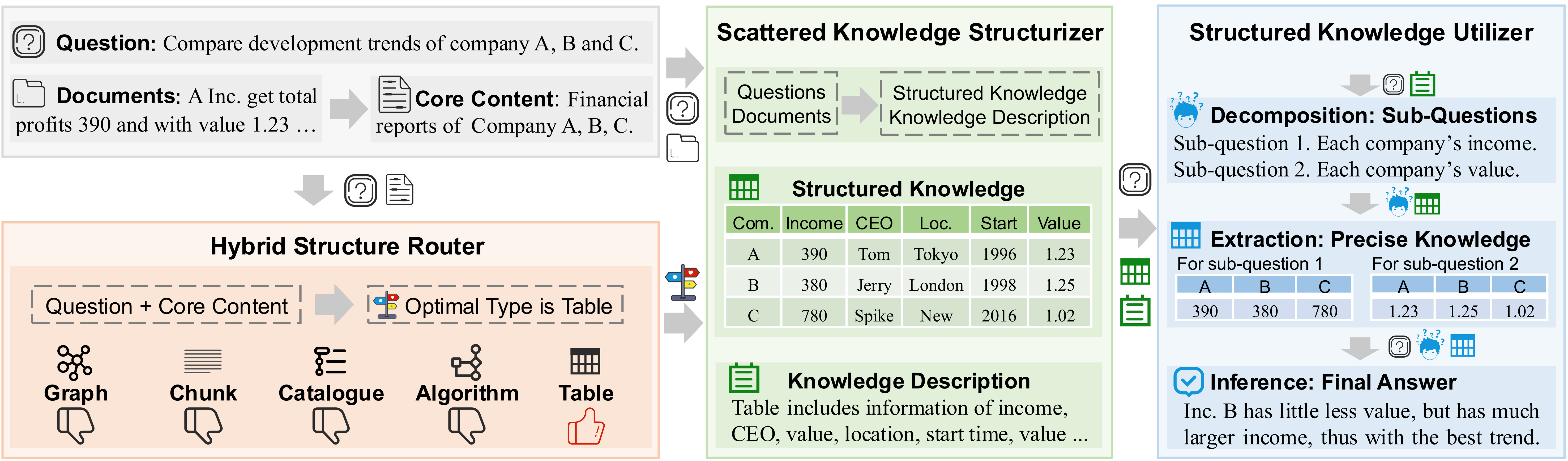
StructRAG: 面向知识密集任务的混合结构化知识增强

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

李卓群, 陈轩昂, 余海洋, 林鸿宇, 陆垚杰, 唐乔裕, 黄非, 韩先培, 孙乐, 李永彬

Thirteenth International Conference on Learning Representations (ICLR 2025)

联系人：李卓群　　邮箱：lizhuoqun2021@iscas.ac.cn



1. Background

- Knowledge-intensive Reasoning
  - Relevant information is highly dispersed
  - Deep reasoning based on retrieved information
- Previous Methods Perform Poorly
  - Retrieved chunks contain too much noise
  - Failing to identify relationships of information pieces

2. Motivation

- How Do Human Beings Solve Such Tasks?
  - Cognitive Load: Employ Knowledge **structurization**
  - Cognitive Fit: Choosing the **optimal type** of structure
- Can LLMs Learn from Human? Yes!
  - Showing similarities to human cognition, e.g., *cot*
  - Having ability for powerful and flexible structurization

3. StructRAG Framework

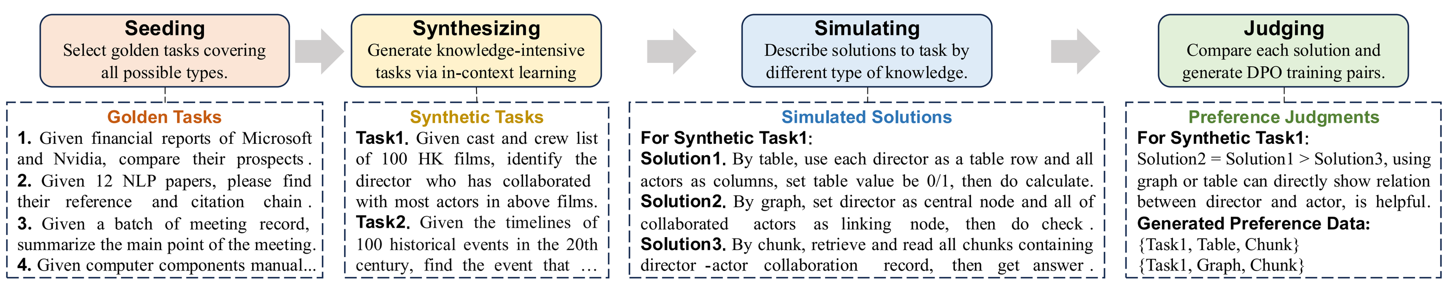
- Hybrid Structure Router
  - Determine the **optimal structure type**
- Scattered Knowledge Structurizer
  - Transform original information to **structured knowledge**
- Structured Knowledge Utilizer
  - Decompose complex questions and do deep reasoning

4. Router Training

- DPO Training: enhance determining optimal structure type

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

- Training Data Constructing



5. Experiments & Conclusion

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**

- SOTA performance on the Loong benchmark
- More noticeable performance gains as complexity increases

Project Page
(HuggingFace)