
TORONTO
KDD2025

August 3–7, 2025
KDD2025

KET-RAG: A Cost-Efficient Multi-Granular Indexing Framework for Graph-RAG

Yiqian Huang, Shiqi Zhang, Xiaokui Xiao

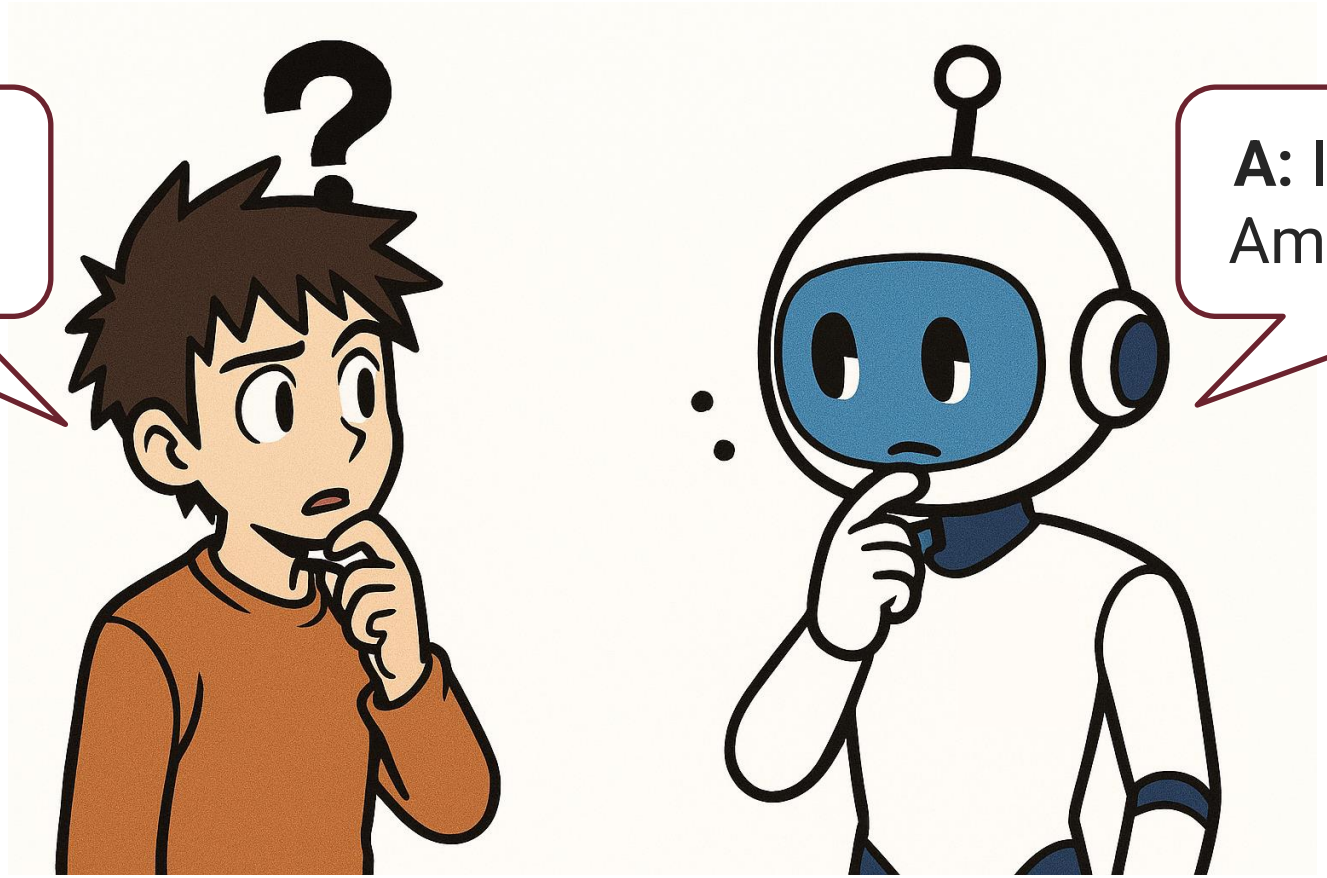
NUS National University of Singapore

PyroWis

Agenda

- ***Background:*** RAG and Graph-RAG
- ***Motivation:*** Graph-RAG's expense
- ***Our Method:*** KET-RAG
- ***Experiment***
- ***Future Work***

Background: Asking questions to LLMs

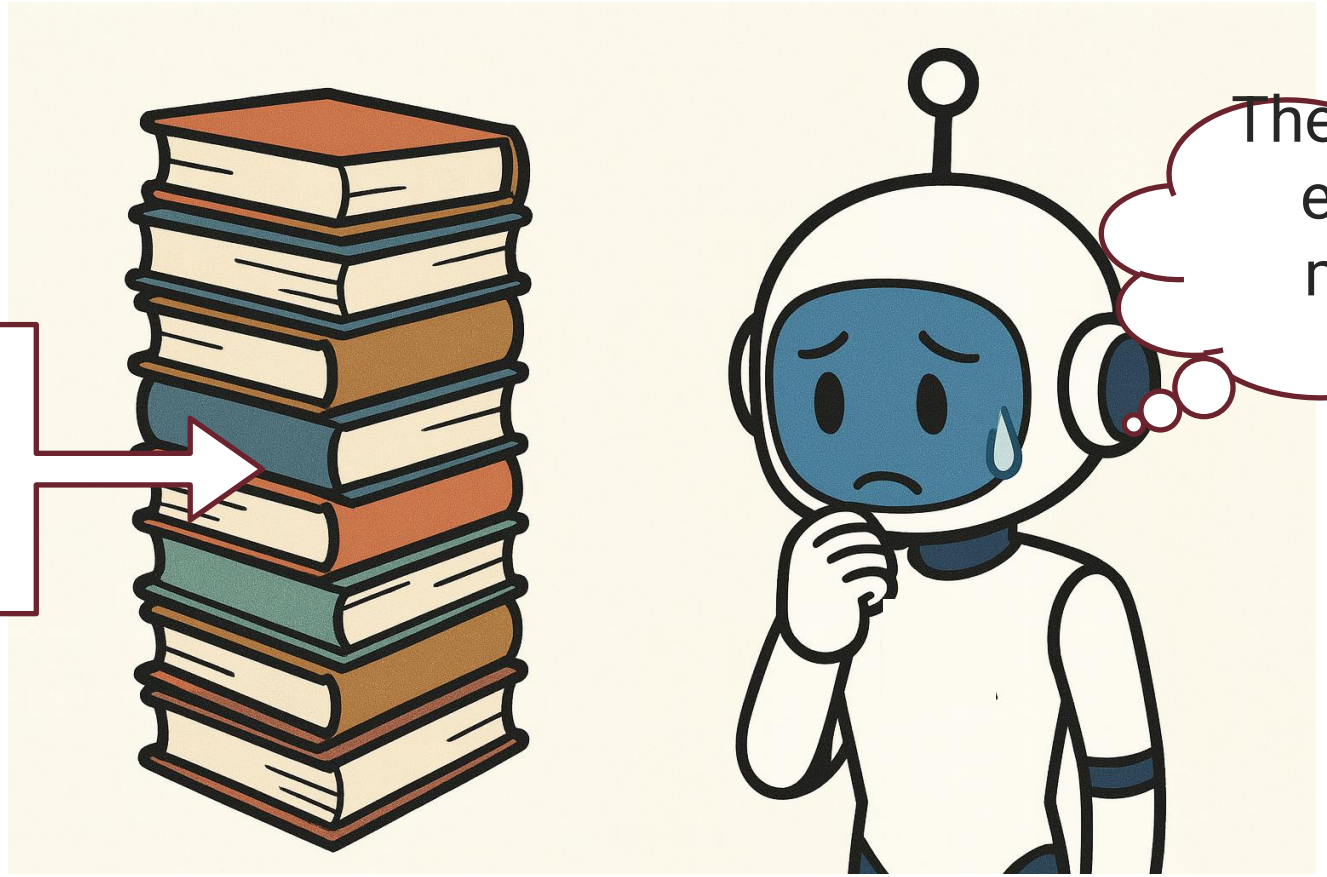


Without external knowledge, large language models (LLMs) can only guess the answer.

Background: Asking questions to LLMs

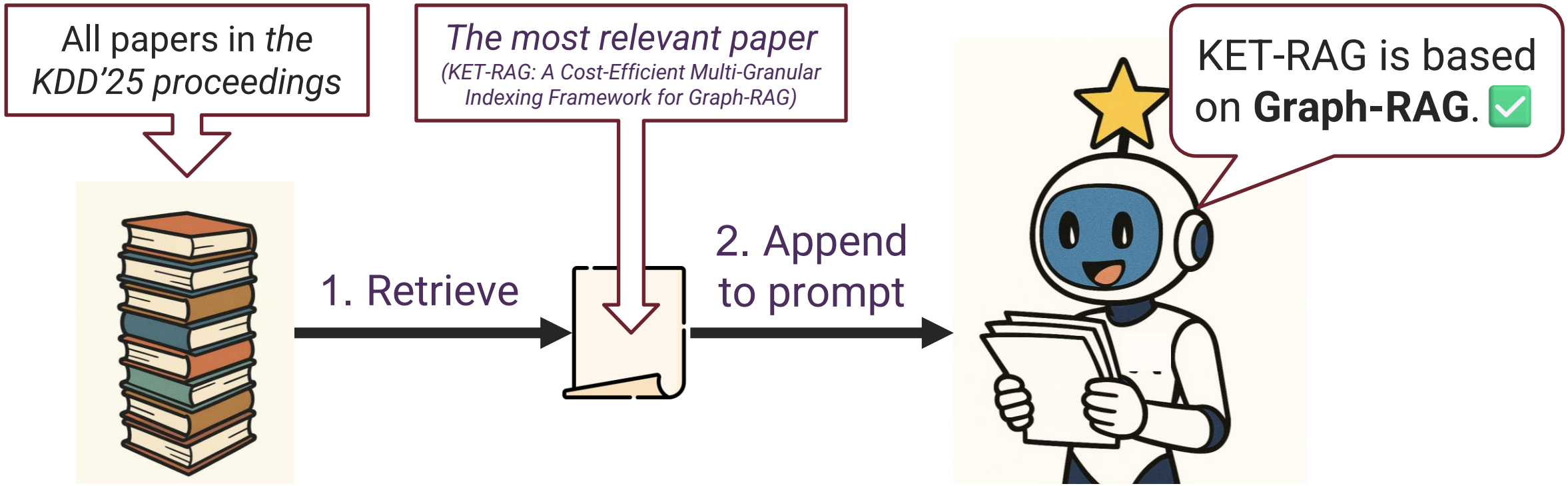
All papers in *the KDD'25 proceedings*

($\approx$ 5,000,000 tokens)



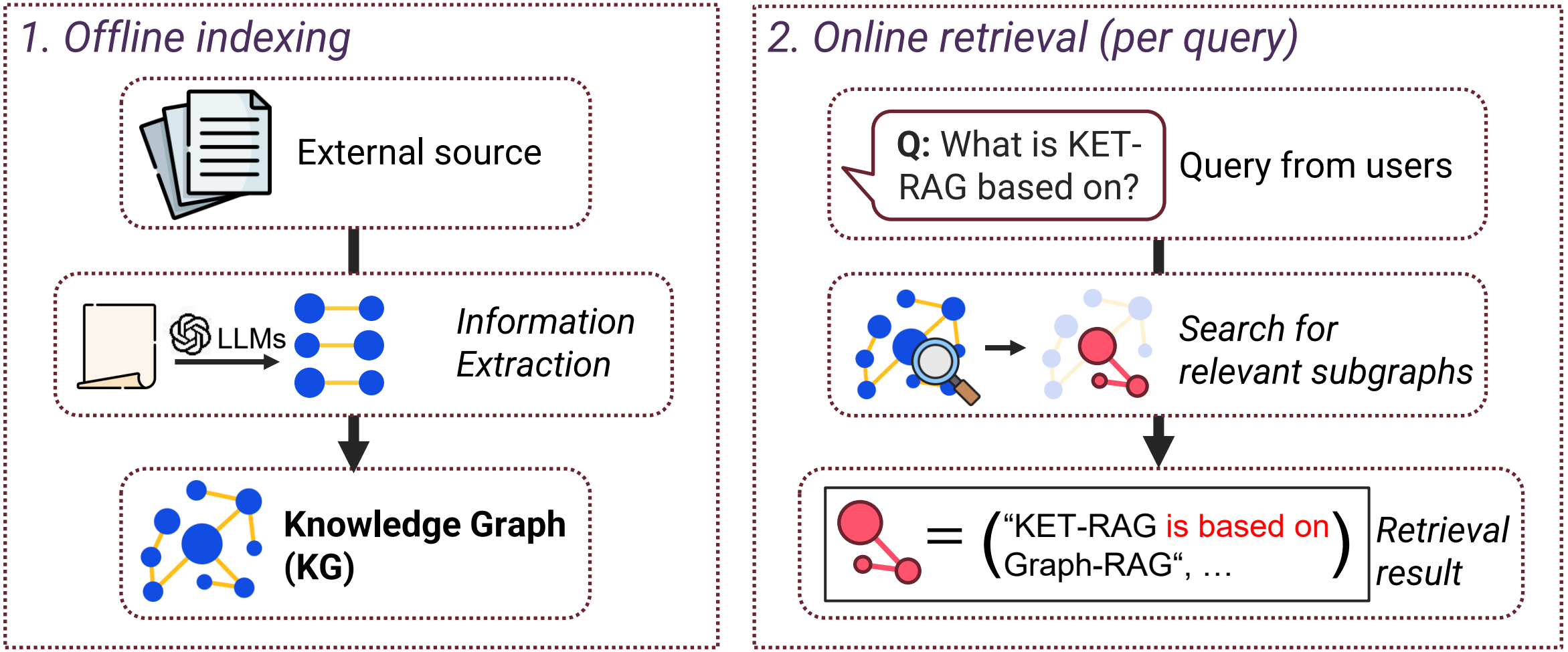
Even when we have an external text source, they are too large to provide in full to LLMs.

Background: **RAG** (**R**etrieval-**A**ugmented **G**eneration)



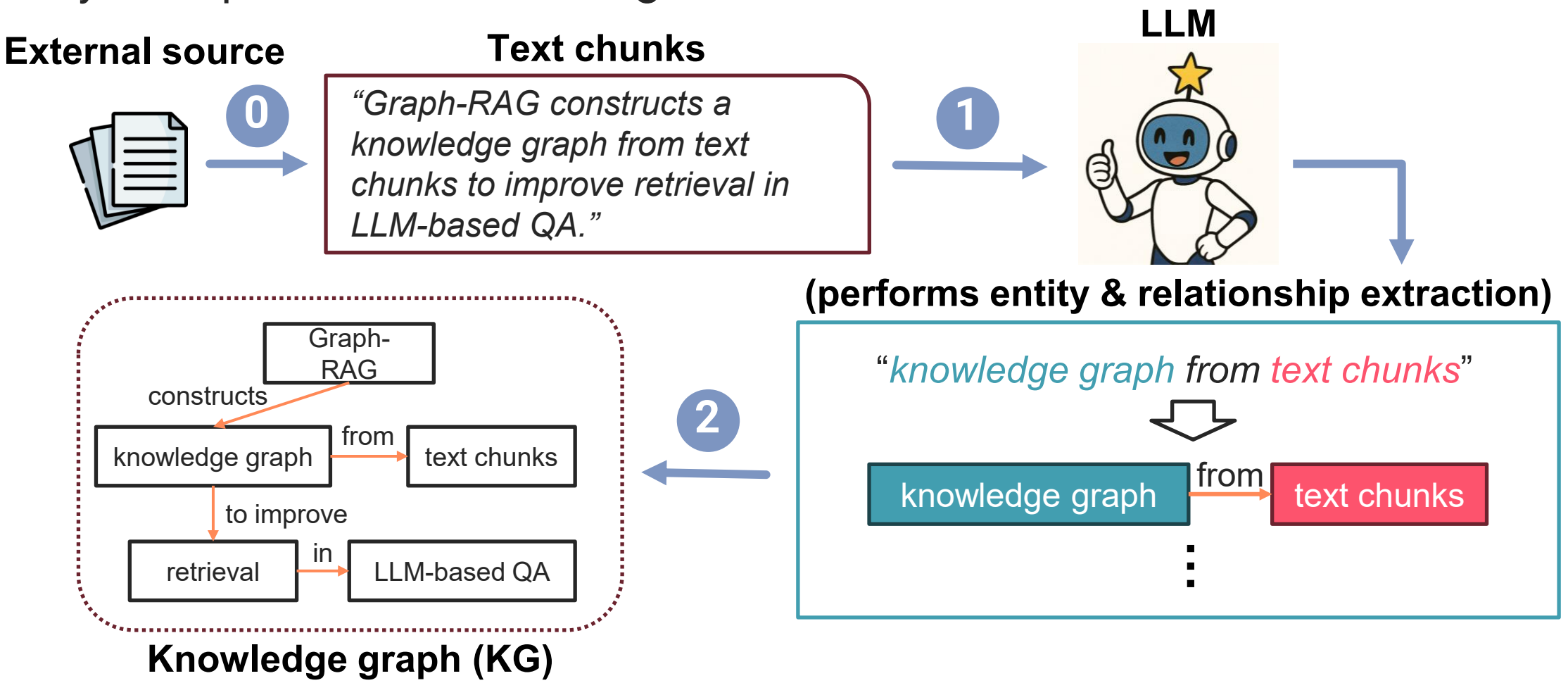
Given an external text source, RAG retrieves relevant context to augment the LLM's generation (e.g., answer).

Background: Graph-RAG



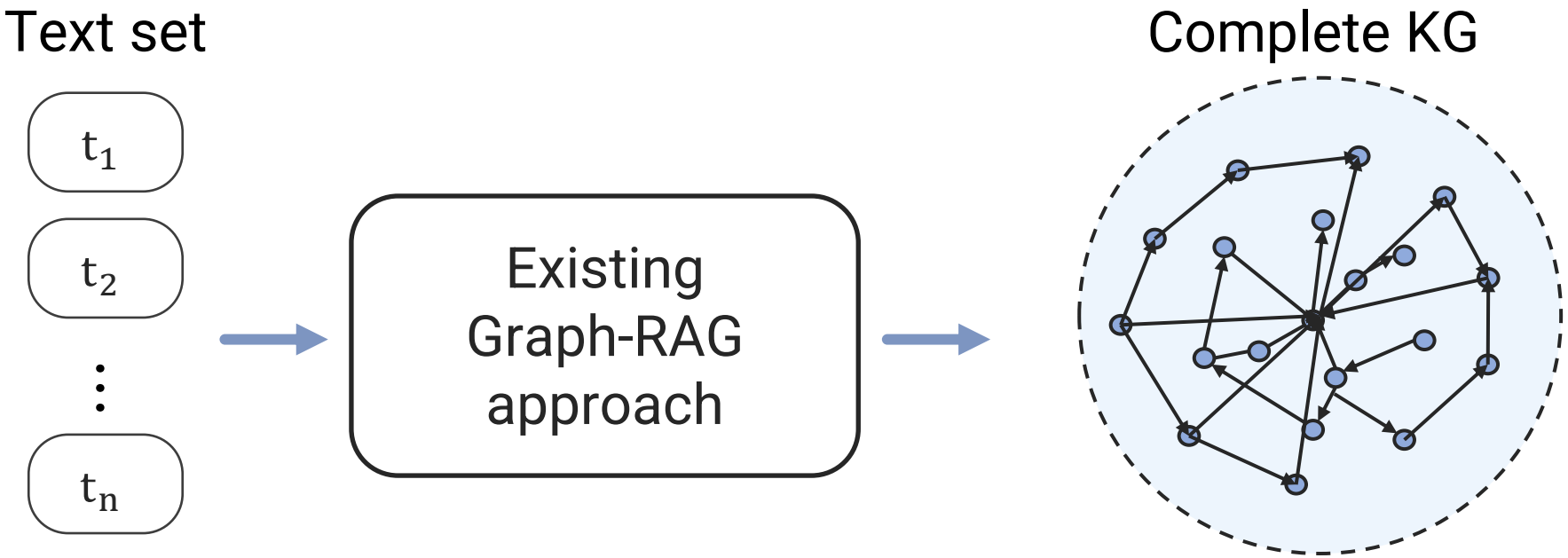
Background: Graph-RAG

- Toy example of **offline indexing**



Motivation: Graph-RAG's expense

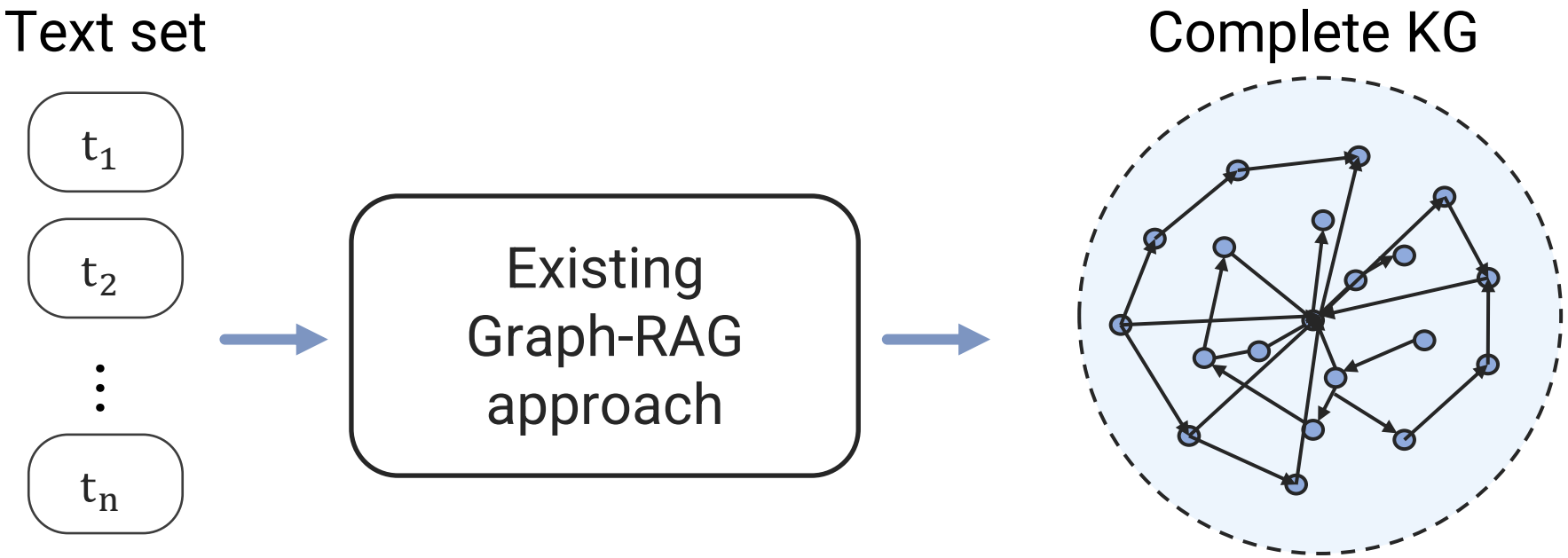
- **Main cost:** constructing the knowledge graph



Whole corpus → **fully** read by LLMs → complete KG

Motivation: Graph-RAG's expense

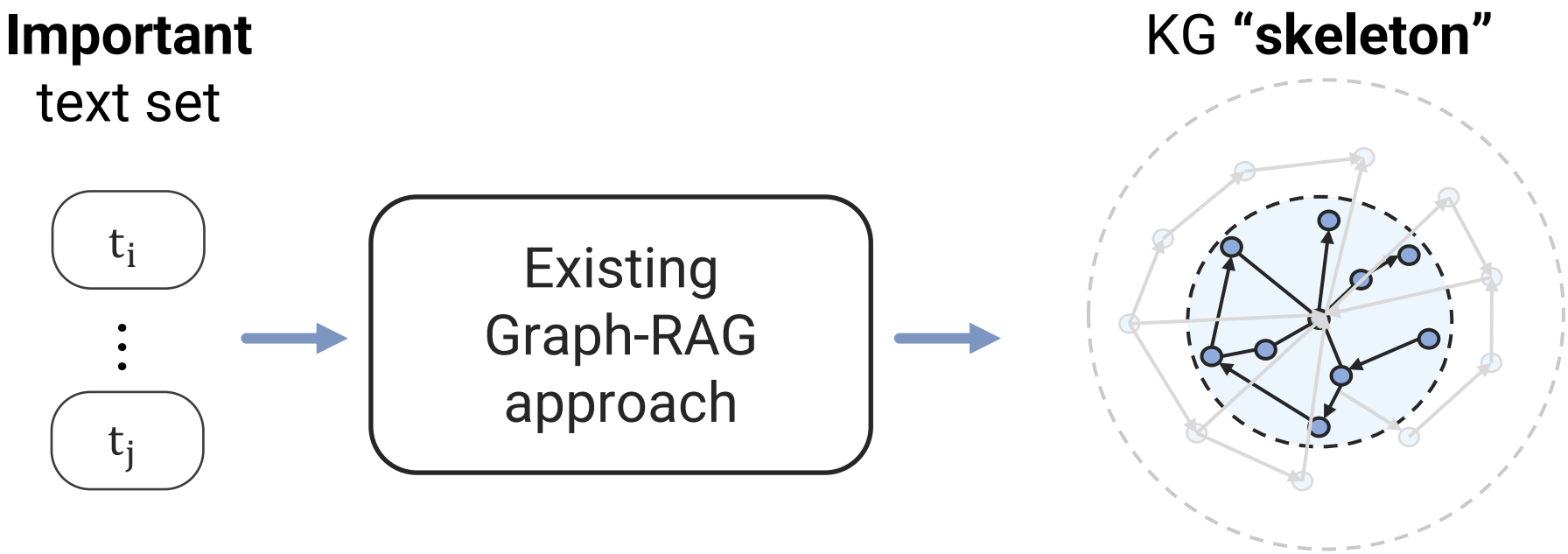
- **Main cost:** constructing the knowledge graph



a **3.2MB** dictionary + GPT-4o-mini = **$21**
a single **5GB** legal case + GPT-4o-mini = **$33,000**

Motivation: Graph-RAG's expense

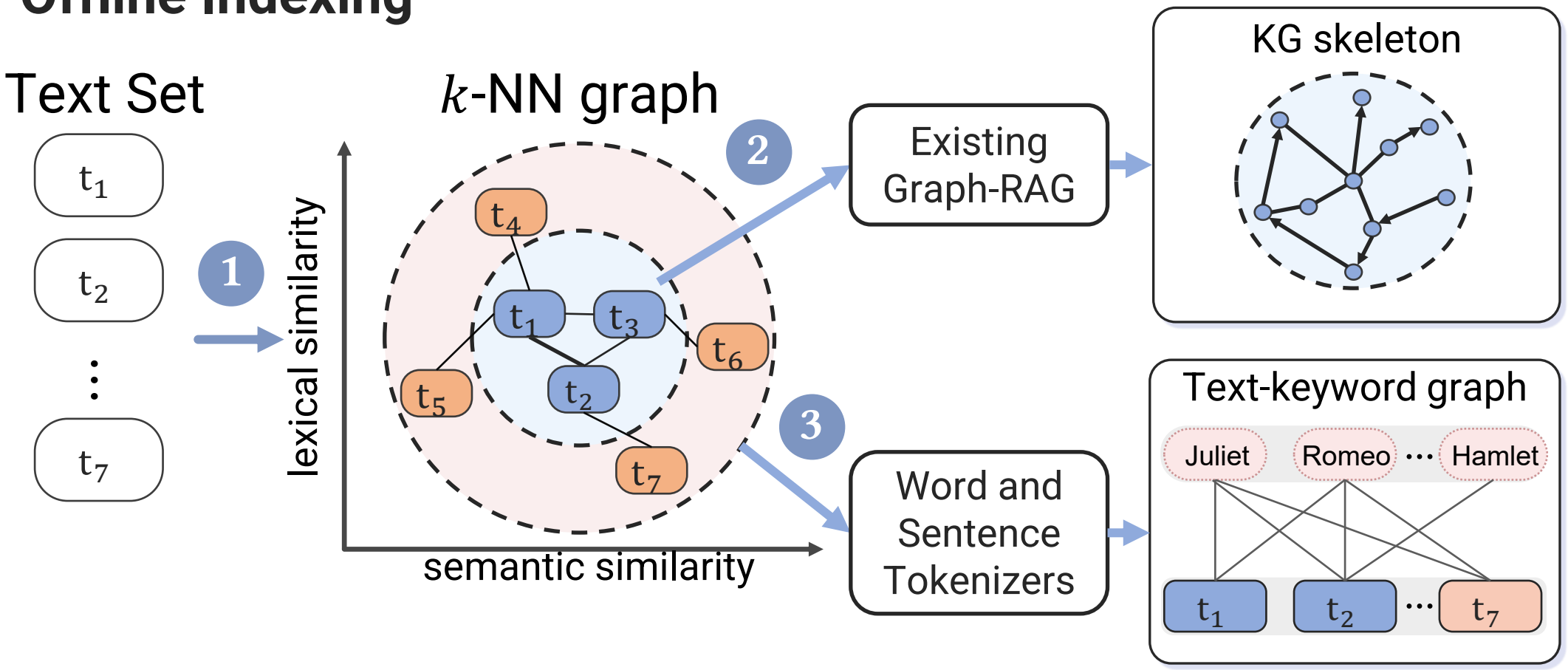
- **Idea:** we can only use **important** text



$$"skeleton"\ cost = \frac{|\text{Important text set}|}{|\text{overall text set}|} \times overall\ cost$$

Our Method: KET-RAG

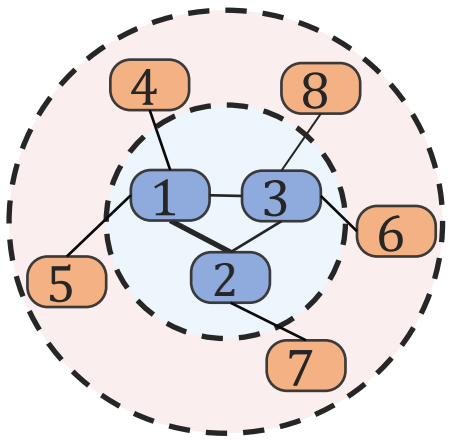
- **Offline Indexing**



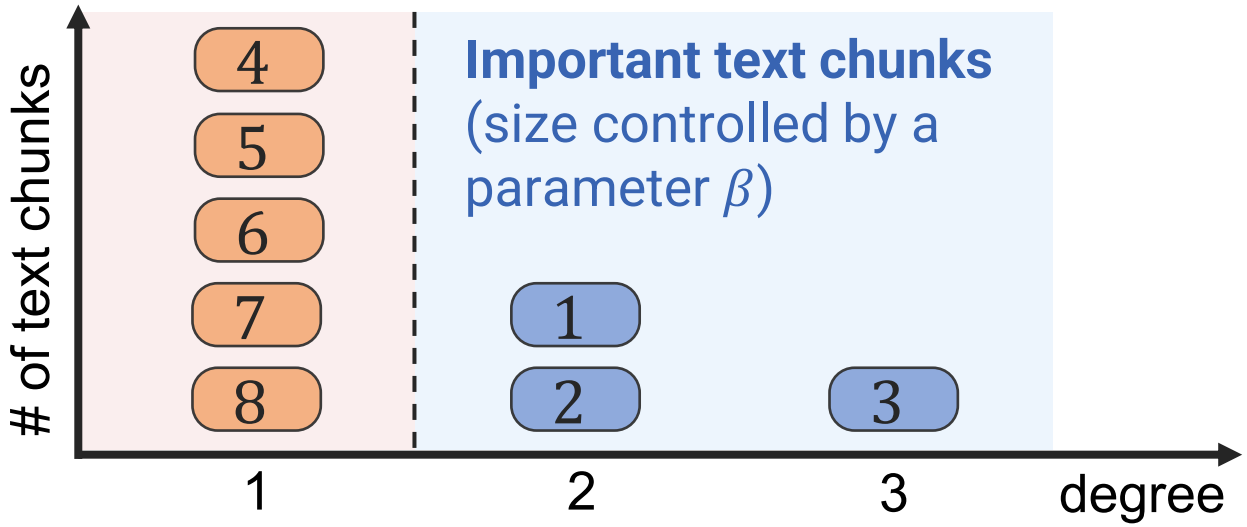
Our Method: KET-RAG

Offline Indexing – 1. k-NN graph

k-NN graph



Observation from *Fig.2* of our paper

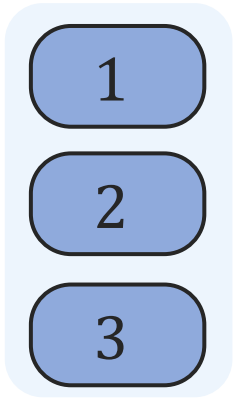


A small subset of text is **important**
(=highly relevant to others in k-NN graph)

Our Method: KET-RAG

Offline Indexing – 2. KG skeleton

Important text set
(size=$\beta \cdot original$)



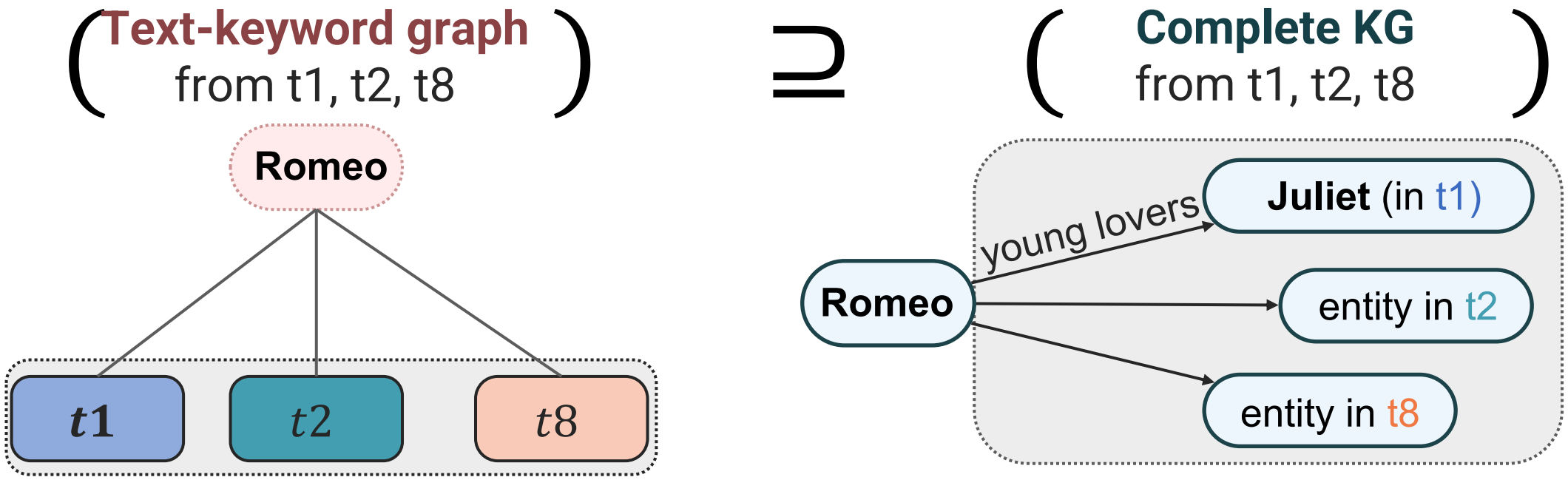
Existing Graph-RAG approach

KG skeleton

$$\text{cost}_{\text{KG skeleton}} = \beta \cdot cost_{KG}$$
$$\Rightarrow \mathbf{cost_{KET-RAG} \approx \beta \cdot cost_{Graph-RAG}}$$

Our Method: KET-RAG

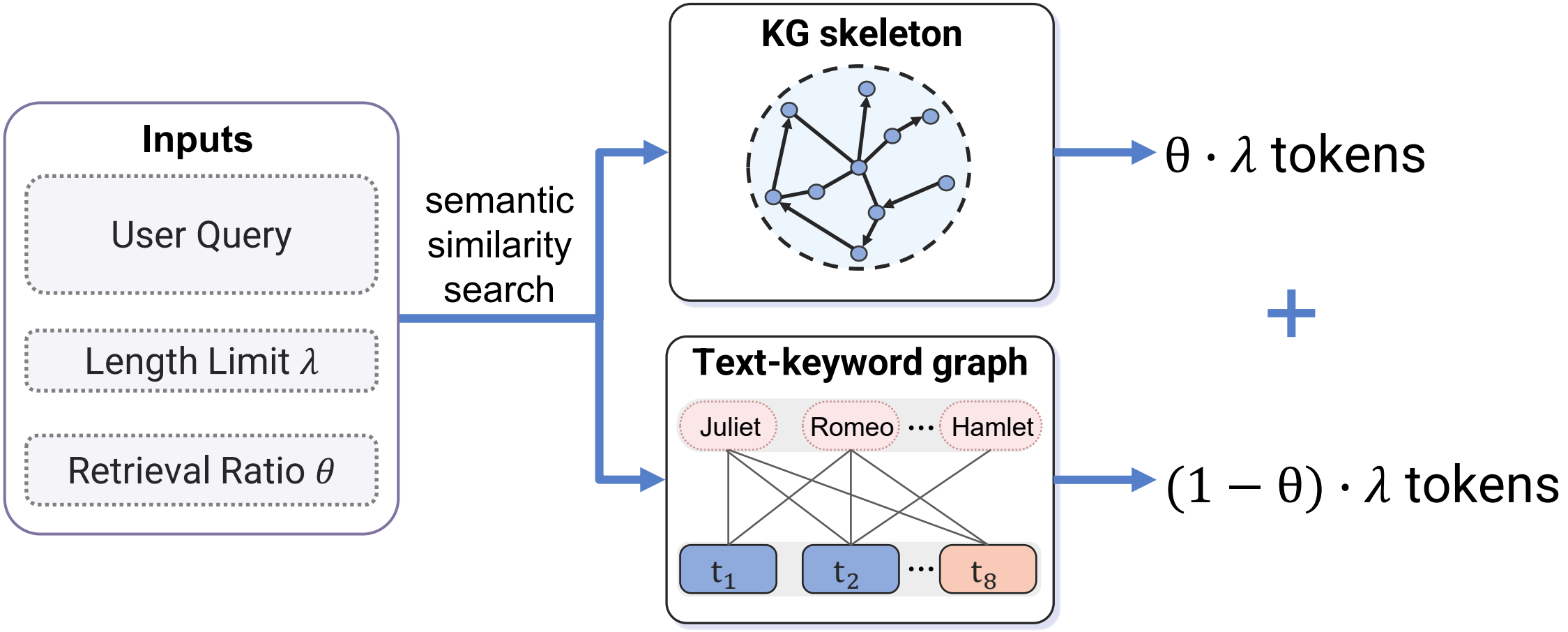
Offline Indexing – 3. Text-keyword graph



Text-keyword graph already encodes information from the **KG**

Our Method: KET-RAG

- **Retrieval**



Experiments

Scenario

- Multi-hop QA (datasets: MuSiQue, HotpotQA)
- open-ended QA (datasets: RAG-QA Arena)

Competitors

(1) Vanilla RAG (2) KNN-Graph-RAG (3) Microsoft's Graph-RAG

(4) Hybrid-RAG (5) HyDE (6) HippoRAG (7) LightRAG

Our proposals

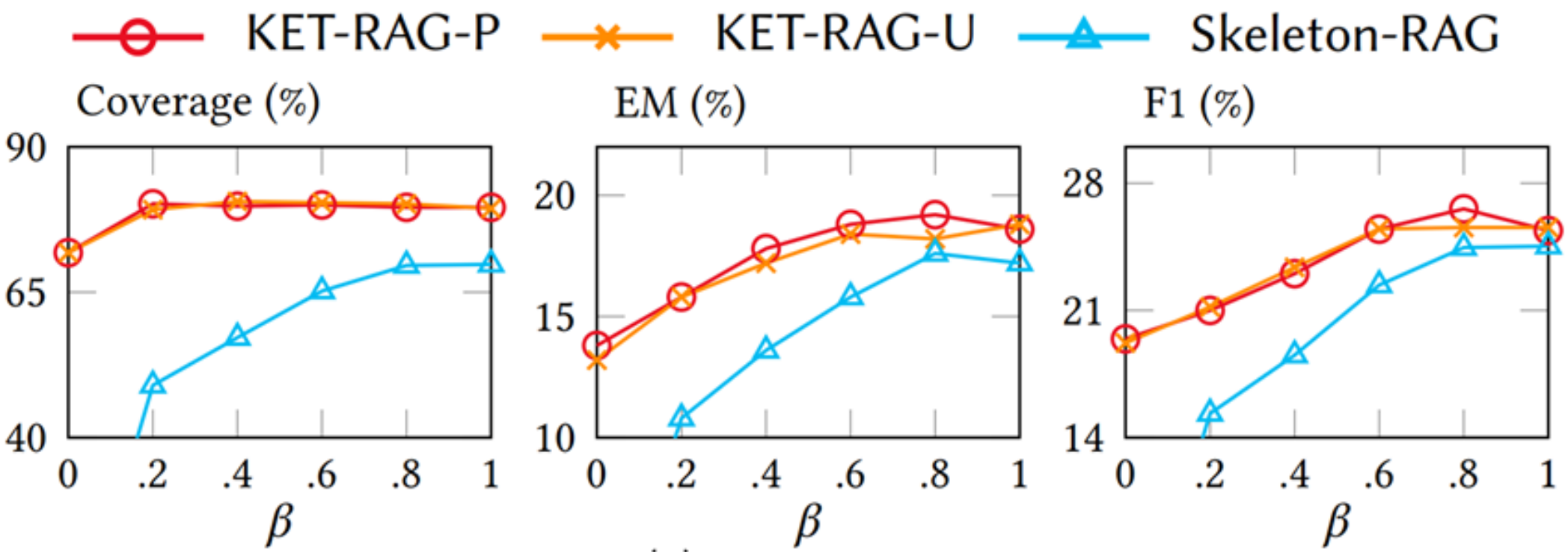
- Keyword-RAG (only the text-keyword graph), Skeleton-RAG (only the KG skeleton),
- **KET-RAG-U, KET-RAG-P** (final solutions)

Experiments: overall results

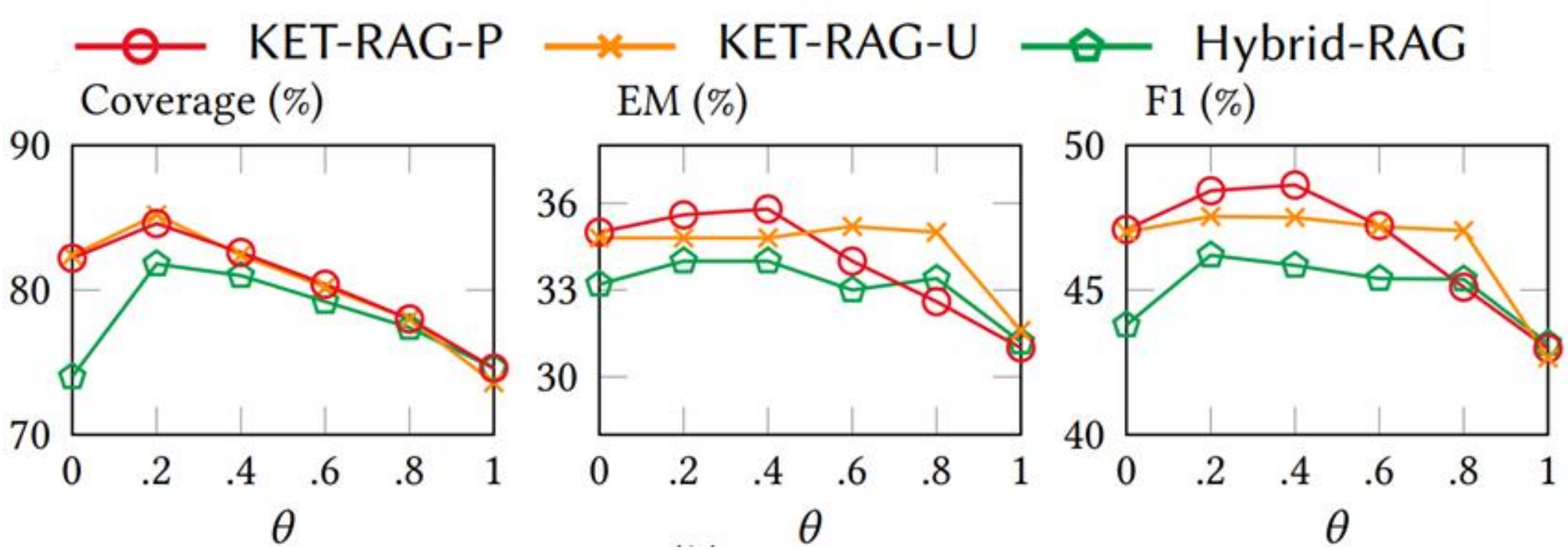
compared to Graph-RAG: superior performance, ≈80% of the cost

Dataset	MuSiQue					HotpotQA					RAG-QA Arena	
Metric	USD	Coverage	EM	F1	BERTScore	USD	Coverage	EM	F1	BERTScore	USD	Win Rate
Text-RAG	0.02	26.4	3.0	5.4	62.9	0.01	37.2	14.8	19.4	69.9	0.03	0.0
KNNG-RAG	0.02	21.2	2.8	4.6	62.5	0.01	28.6	13.4	16.9	68.8	0.03	17.8
MS-Graph-RAG	2.30	47.6	11.4	15.8	67.4	2.30	63.0	21.6	30.2	74.2	5.05	33.4
Hybrid-RAG	2.32	49.2	10.4	15.1	68.8	2.31	64.8	22.6	30.5	76.8	5.08	35.0
HyDE	0.03	33.0	4.2	6.7	63.5	0.02	40.8	16.2	21.3	70.9	0.03	33.2
HippoRAG	1.49	51.6	9.2	14.2	66.8	1.40	64.0	**29.2**	38.3	77.3	1.37	38.4
LightRAG-Local	1.80	39.0	9.4	12.9	66.6	1.77	57.6	22.4	28.1	73.1	4.00	29.0
LightRAG-Global	1.80	37.6	6.4	9.5	64.8	1.77	49.0	19.6	24.5	71.8	4.00	23.0
LightRAG-Hybrid	1.80	45.6	10.2	14.4	67.2	1.77	61.6	24.6	30.7	74.2	4.00	30.0
Keyword-RAG	0.03	50.8	7.0	11.8	68.1	0.03	60.2	24.4	33.5	**78.9**	0.07	37.0
Skeleton-RAG	1.86	43.4	11.0	14.1	66.8	1.84	57.8	20.0	26.7	73.2	4.04	28.6
KET-RAG-U	1.89	76.2	13.4	18.1	68.1	1.87	81.4	28.4	38.2	77.5	4.08	35.0
KET-RAG-P	1.89	**77.0**	**14.0**	**18.9**	**69.0**	1.87	**81.6**	28.6	**38.7**	77.5	4.08	**39.6**

Experiments: varying budget (β)



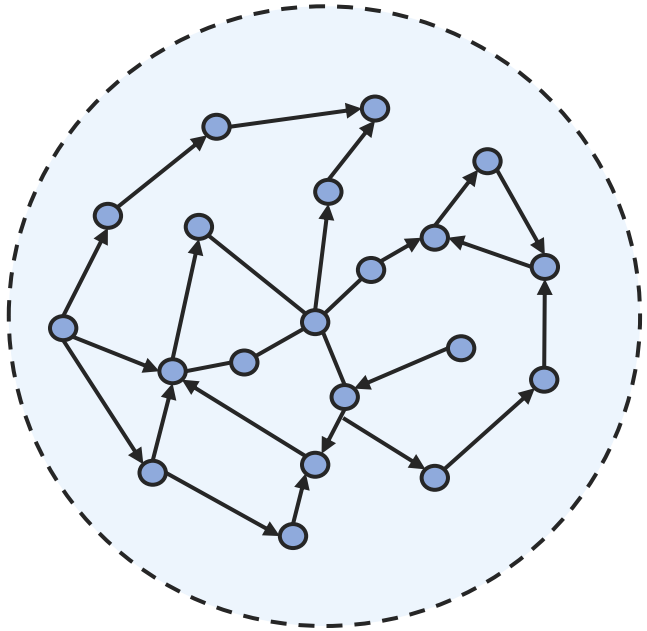
Experiments: varying retrieval ratio (θ)



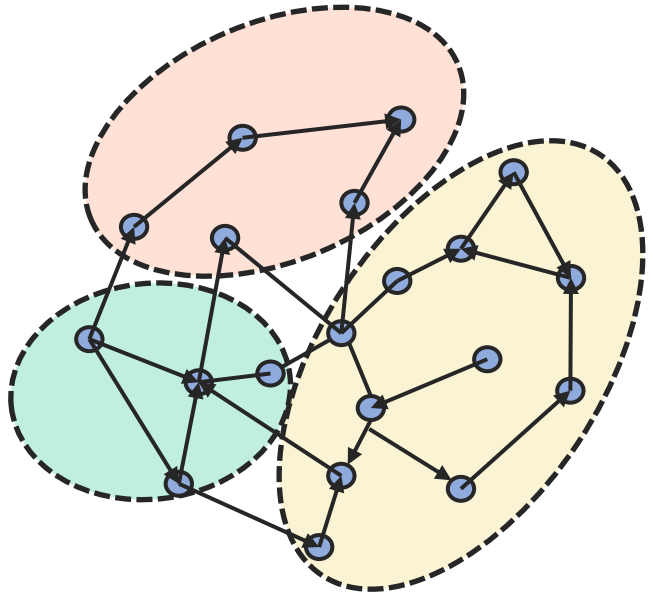
Future Work: handling global queries

Example: summarizing the main theme of the documents

Complete KG



Community summaries



THANK YOU!

Code: *github.com/waetr/KET-RAG*