Graph RAG & QFS

Graph Retrieval-Augmented Generation & Query-Focused Summarization

1. **From Local to Global: A Graph RAG Approach to Query-Focused Summarization.**

   Microsoft Research. arXiv '24

2. **Beyond Relevant Documents: A Knowledge-Intensive Approach for Query-Focused Summarization Using Large Language Models.**

   University of Amsterdam. ICPR '25

Presenter

Junseo, Yu

02.20.2025

DMAIS Lab at CAU

Contents From Local to Global

1. Introduction

- **QFS**
- **RAG**
- **Existing Problems**
- **Motivation**

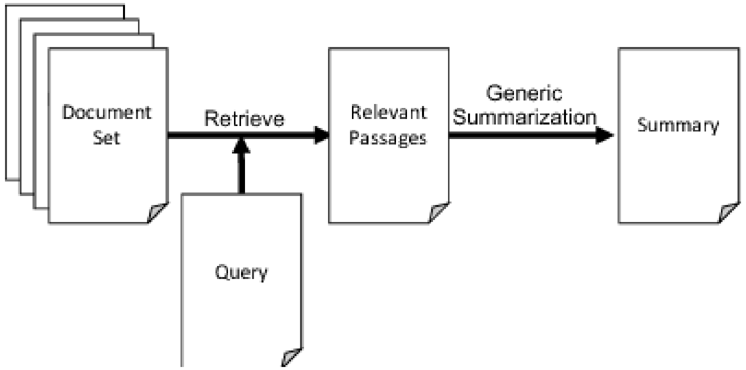
Introduction

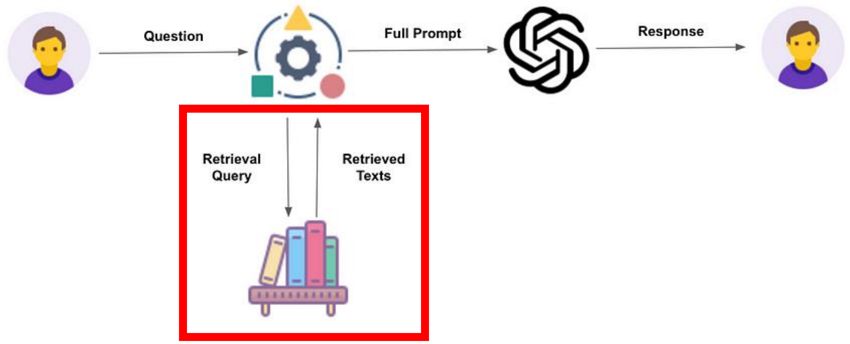
QFS

What is Query Focused Summarization (QFS)?

- **QFS**
  - **Generates** a summary that is directly **relevant** to a given query
  - **E.g., Q:** What were the main points of the president on climate change?

    A: The main points of the Trump…
- **QMDS (Q**uery-focused **M**ulti-**D**ocument **S**ummarization)
  - Generates a summary by **extracting** relevant information from **multiple documents** to a given query





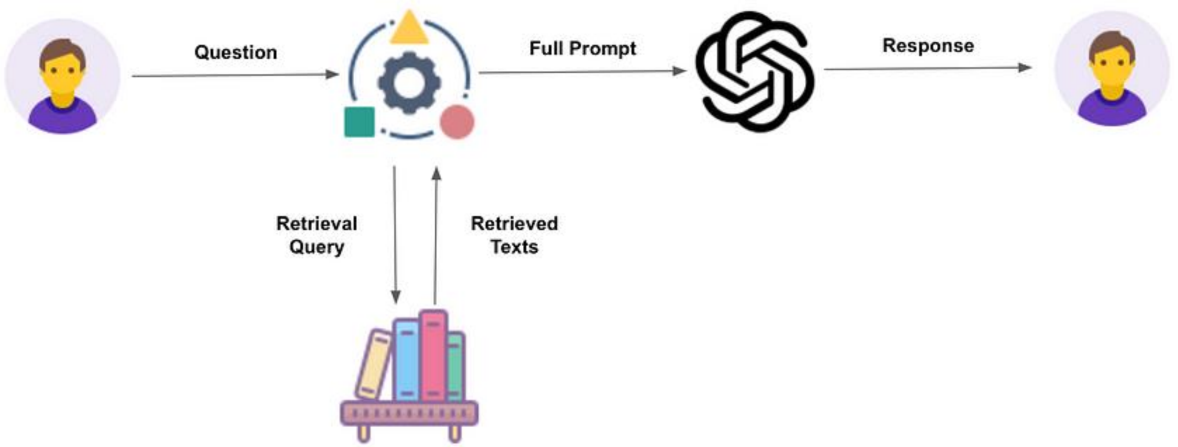
Introduction

RAG

What is **RAG**?

- **RAG, Retrieval-augmented generation**
  - RAG **retrieves** relevant external information.
  - RAG can **alleviate** LLM hallucinations by enriching the input context with **accurate information**
    - **E.g.,** ***Knowledge from RAG: Jamaica → language_spoken → English***

      Question: Which language do Jamaican people speak*?*



Introduction

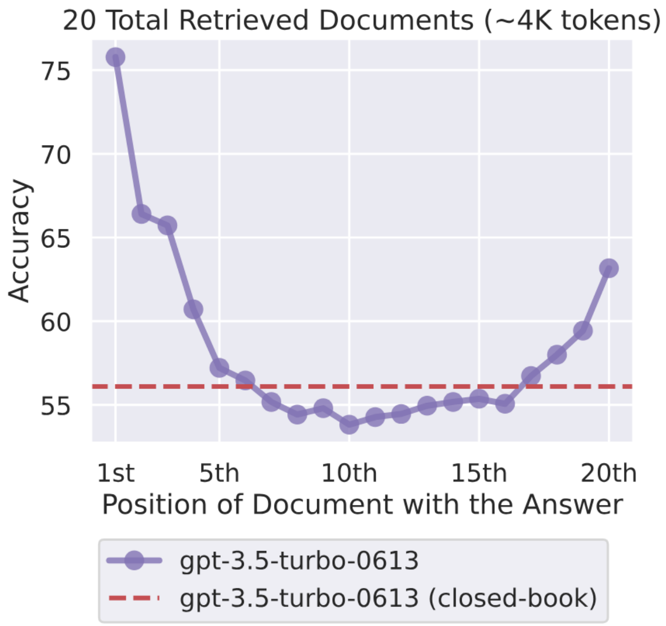
Existing Problems

- Existing **RAG** can **only** answer **local**-centered queries, **not global**-centered queries

  i.e., **Not suitable** for **global questions** and summarization tasks based on the **entire database**

  - **Can** answer ***'Who is the author of The Critique of Pure Reason?'***
  - **Can not** answer ***'What are the main themes in the dataset?'***

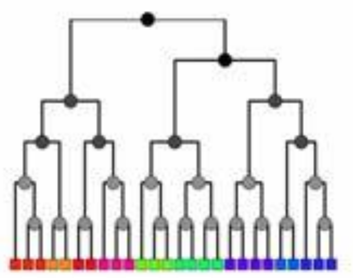
- **Why?**
  - The limitation of Context Window
  - Lost in the middle
    - Lost (disappear) the context in the middle of the long context window

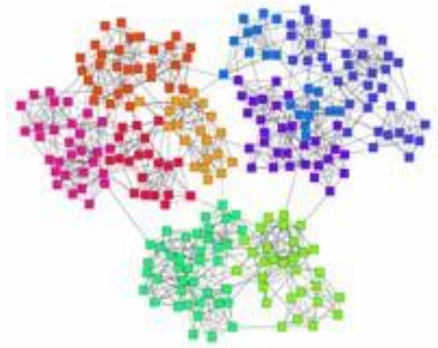


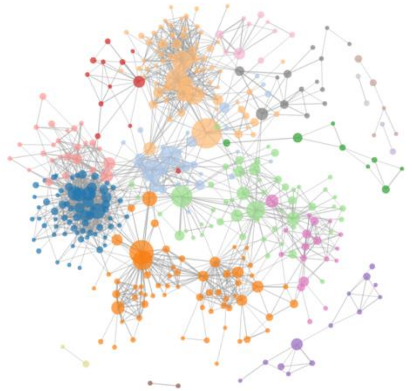
Introduction

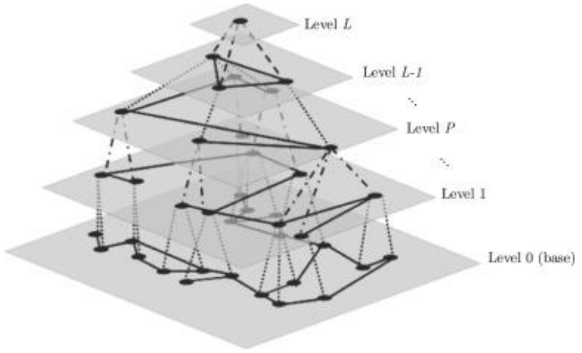
Motivation

- ❑ Incorporating **QFS** with **RAG**
  - ▪ **QFS** can summarize the documents into several sentences
  - ▪ However, QFS does not leverage the **relation** of documents
    - • Without relation, there will be just million of sentences
  - ▪ i.e., Not suitable for **global questions**
- ❑ Need to present a new RAG approach, **Graph RAG**
  - ▪ To leverage the **modularity** and **hierarchical** property of Graph







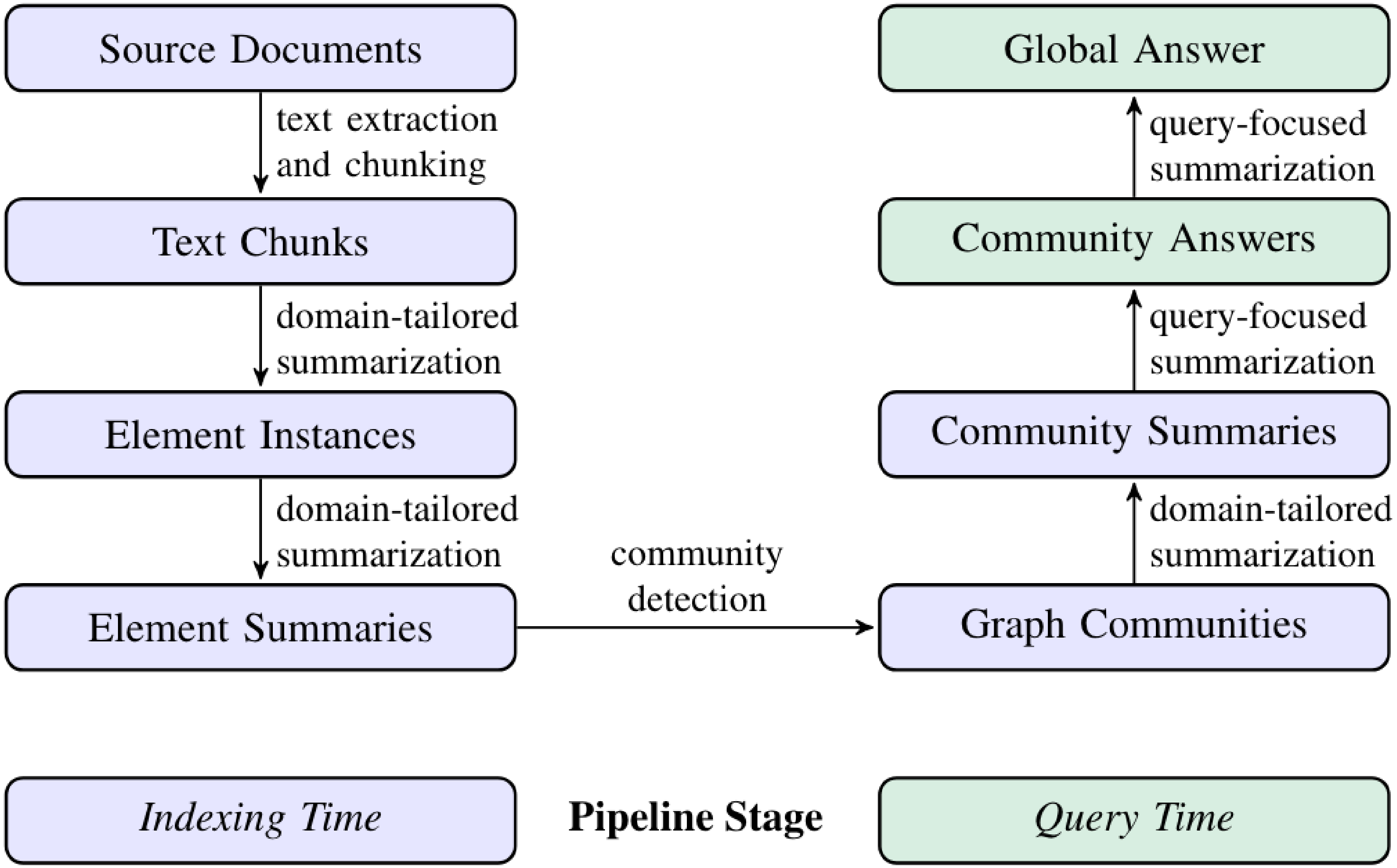


2. Methodology

- Overview
- Indexing Time
- Query Time

Methodology

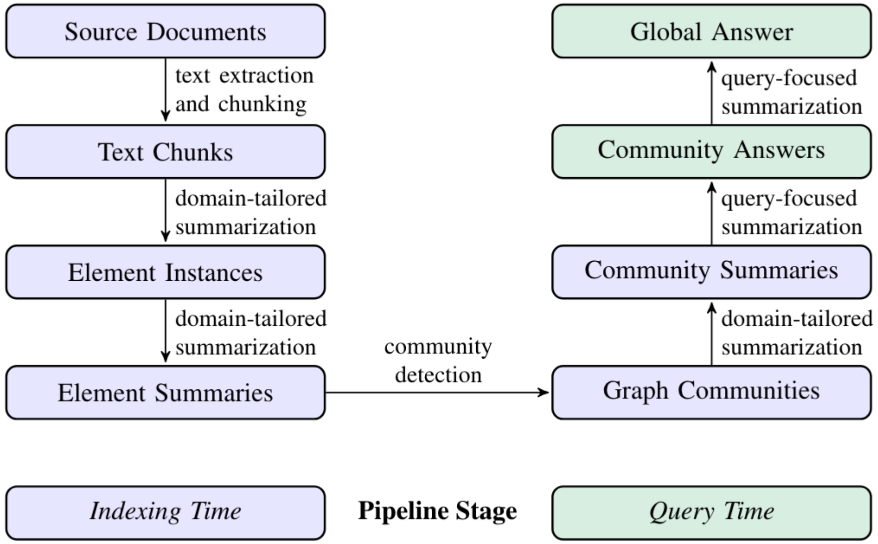
Overview



- **Indexing Time**
  1. Generate **KG** from Documents
  2. Detect communities from the KG
  3. Make **community summaries**
- **Query Time**
  - Given a user **query**, generate a **final answer**

Methodology

Indexing Time

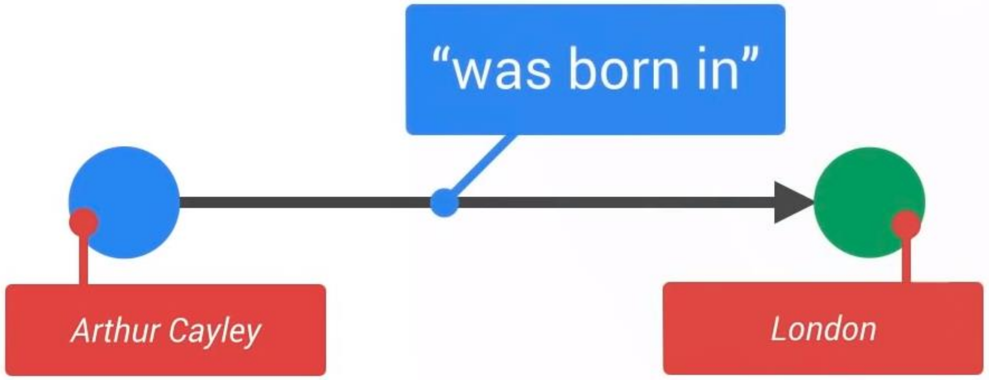


❑ Source Documents → Text Chunks

- **Split** the documents into text chunks

❑ Text Chunks → Element Instances

- Text Chunks are passed to a set of **LLM** prompts **to extract the various elements** of a graph index
- Multipart LLM prompt
  - Identify the **entities**, including their name, type, and description, as well as their **relations**
  - Identify any additional covariates called **claims** including subject, type, description, source text span, and dates

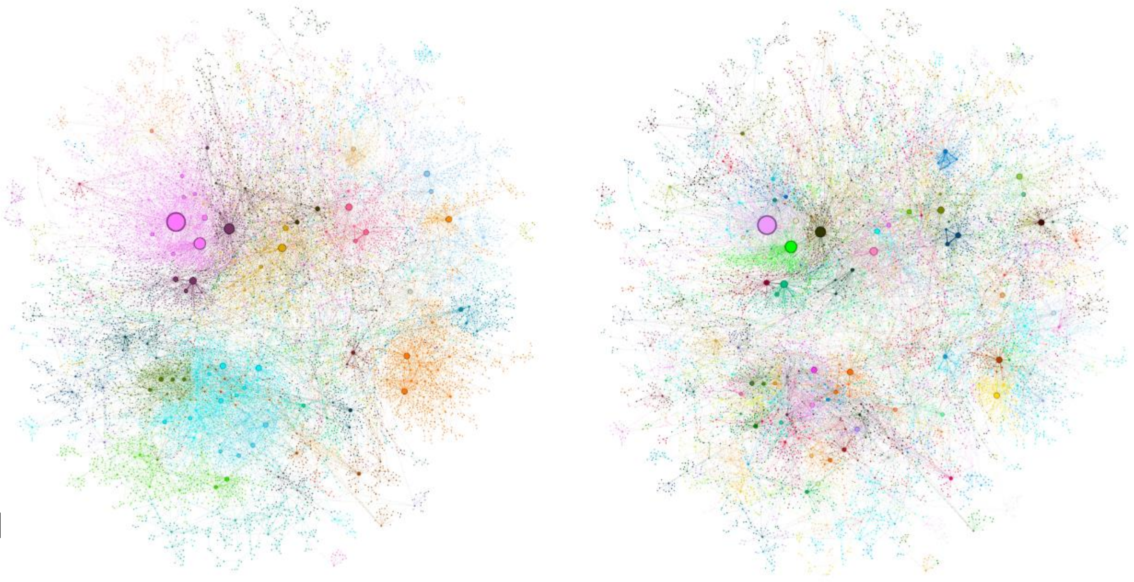


Subject (주체)	Type (유형)	Description (설명)	Source Text Span (출처)	Dates (날짜)
Cayley	Birth Record	Cayley was born in London, England.	"Cayley, a renowned mathematician, was born in London in 1821."	1821
Cayley	Nationality	As he was born in London, he held British nationality.	"Born in London, Cayley was a British mathematician."	-
Cayley	Historical Context	London in the early 19th century was a hub of scientific advancements.	"During the 19th century, London was a major center for scientific research."	1800s
Cayley	Birth Location Details	Cayley was born in Richmond, London, which was a suburban area at the time.	"Cayley was born in Richmond, a district in London, in 1821."	1821

Methodology

Indexing Time

- **Element Instances → Elements Summaries**
  - **Convert** all such **instance** into **single text summary** for each graph element
  - One summary corresponds to one element (an entity node, a relationship edge, or a claim)
- **Element Summaries → Graph Communities**
  - Use **community detection** algorithms, ***Leiden***
  - ***Leiden*** can represent hierarchical community structure
- **Graph Communities→ Community Summaries**
  - Create summaries of each community by using **LLM**
  - **Leaf-level communities**: element level summaries.
  - **Higher-level communities**: summaries including seve



(a) Root communities at level 0

(b) Sub-communities at level 1

Methodology

Query Time

- **Community Summaries → Community Answers → Global Answer**
  - Generate a final answer in a multi-stage process **Map-Reduce Method**
  - All Community summaries are divided into chunks (why?)
    - E.g., '***AI industry***' community summary is divided into 5 sentences.
    - E.g., First Context window contain 'Query, AI industry chunk 3, displaced situation chunk 4, and others'
  - Generate **intermediate answers** in parallel by using **LLM** with given query and divided summaries
    - E.g., 'Query, AI industry chunk 1, global warming chunk3'
  - Request the **LLM** to generate a **helpfulness score (0-100)** for the answer about query
  - Intermediated answers are sorted and iteratively added into a new context window until the token limit
  - **LLM** return **final answer**

3. Experiments

- Setup
- Evaluation Metrics
- Results

Experiments

Setup

Datasets

- Podcast transcripts & News articles: 1M token range

Queries

- No existing high level summarization **queries** datasets
- Generated by **LLM**
  - Given a short description of dataset, identify potential users and tasks per user

Experiments

Setup

- Given a short description of dataset,

  1) **identify potential users** and **2) tasks per user** and **3) generate the queries**

Dataset	Example activity framing and generation of global sensemaking questions
Podcast transcripts	*User*: A tech journalist looking for insights and trends in the tech industry *Task*: Understanding how tech leaders view the role of policy and regulation *Questions*: 1. Which episodes deal primarily with tech policy and government regulation? 2. How do guests perceive the impact of privacy laws on technology development? 3. Do any guests discuss the balance between innovation and ethical considerations? 4. What are the suggested changes to current policies mentioned by the guests? 5. Are collaborations between tech companies and governments discussed and how?
News articles	*User*: Educator incorporating current affairs into curricula *Task*: Teaching about health and wellness *Questions*: 1. What current topics in health can be integrated into health education curricula? 2. How do news articles address the concepts of preventive medicine and wellness? 3. Are there examples of health articles that contradict each other, and if so, why? 4. What insights can be gleaned about public health priorities based on news coverage? 5. How can educators use the dataset to highlight the importance of health literacy?

Experiments

Evaluation Metrics

Head-to-head comparison

- Due to **the lack of gold standard answers**, use **LLM evaluator** to compare the metrics
- Metrics
  - **Comprehensiveness**: Does the answer cover all aspects of the question in detail?
  - **Diversity**: Does the answer provide varied perspectives and insights?
  - **Empowerment**: Does the answer help the reader understand?
  - **Directness**: Does the answer clearly and specifically address the question?
- **How to evaluate?**
  - The **LLM** is provided with the question, target metric, and a pair of answers
  - Then, the **LLM** is asked to answer which answer is better according to the metric

Experiments

Evaluation Metrics

Question	Which public figures are repeatedly mentioned across various entertainment articles?
Graph RAG	**Overview of Prominent Public Figures in Entertainment** The entertainment industry is vast and diverse, encompassing film, television, music, sports, and digital media. Certain public figures stand out due to their significant contributions and influence across these sectors. The following summary highlights key individuals who are repeatedly mentioned in various entertainment articles, reflecting their impact and presence within the industry. **Actors and Directors** [...] **Public Figures in Controversy** [...] **Musicians and Executives** [...] **Athletes and Coaches** [...] **Influencers and Entrepreneurs** [...] The repeated mention of these figures in entertainment articles signifies their ongoing relevance and the public's interest in their work. Their influence spans across various aspects of entertainment, from shaping cultural narratives in film and television to driving trends in music and digital media. These individuals not only contribute to their respective fields but also influence the broader cultural landscape, often becoming central figures in social discussions and public discourse.
Naïve RAG	Public figures who are repeatedly mentioned across various entertainment articles include Taylor Swift, Travis Kelce, Britney Spears, and Justin Timberlake. These individuals have been highlighted for various reasons, ranging from their professional achievements to their personal lives. **Taylor Swift** [...] **Travis Kelce** [...] **Britney Spears** [...] **Justin Timberlake** [...] These figures are frequently covered due to their high-profile status and the public's interest in their careers and personal lives. Their activities, whether in music, sports, or personal relationships, have significant cultural and economic impacts, as evidenced by the media coverage and public reactions.

LLM Decision	**Comprehensiveness: Winner=1 (Graph RAG)** Answer 1 is better because it provides a more comprehensive and detailed list of public figures from a wider range of entertainment sectors, including film, television, music, sports, gaming, and digital media. It also includes specific examples of their contributions and the impact they have on their respective fields, as well as mentions of controversies and their implications. Answer 2, while detailed in its coverage of a few individuals, is limited to a smaller number of public figures and focuses primarily on their personal lives and relationships rather than a broad spectrum of their professional influence across the entertainment industry. **Diversity: Winner=1 (Graph RAG)** Answer 1 is better because it provides a more varied and rich response by covering a wide range of public figures from different sectors of the entertainment industry, including film, television, music, sports, gaming, and digital media. It offers insights into the contributions and influence of these figures, as well as controversies and their impact on public discourse. The answer also cites specific data sources for each mentioned figure, indicating a diverse range of evidence to support the claims. In contrast, Answer 2 focuses on a smaller group of public figures, primarily from the music industry and sports, and relies heavily on a single source for data, which makes it less diverse in perspectives and insights. **Empowerment: Winner=1 (Graph RAG)** Answer 1 is better because it provides a comprehensive and structured overview of public figures across various sectors of the entertainment industry, including film, television, music, sports, and digital media. It lists multiple individuals, providing specific examples of their contributions and the context in which they are mentioned in entertainment articles, along with references to data reports for each claim. This approach helps the reader understand the breadth of the topic and make informed judgments without being misled. In contrast, Answer 2 focuses on a smaller group of public figures and primarily discusses their personal lives and relationships, which may not provide as broad an understanding of the topic. While Answer 2 also cites sources, it does not match the depth and variety of Answer 1. **Directness: Winner=2 (Naïve RAG)** Answer 2 is better because it directly lists specific public figures who are repeatedly mentioned across various entertainment articles, such as Taylor Swift, Travis Kelce, Britney Spears, and Justin Timberlake, and provides concise explanations for their frequent mentions. Answer 1, while comprehensive, includes a lot of detailed information about various figures in different sectors of entertainment, which, while informative, does not directly answer the question with the same level of conciseness and specificity as Answer 2.

Experiments

Result

Podcast transcripts

Comprehensiveness

	SS	TS	C0	C1	C2	C3
SS	50	17	28	25	22	21
TS	83	50	50	48	43	44
C0	72	50	50	53	50	49
C1	75	52	47	50	52	50
C2	**78**	**57**	**50**	**48**	**50**	**52**
C3	79	56	51	50	48	50

Diversity

	SS	TS	C0	C1	C2	C3
SS	50	18	23	25	19	19
TS	82	50	50	50	43	46
C0	77	50	50	50	46	44
C1	75	50	50	50	44	45
C2	81	57	54	56	50	48
C3	**81**	**54**	**56**	**55**	**52**	**50**

Empowerment

	SS	TS	C0	C1	C2	C3
SS	50	42	57	52	49	51
TS	**58**	**50**	**59**	**55**	**52**	**51**
C0	43	41	50	49	47	48
C1	48	45	51	50	49	50
C2	51	48	53	51	50	51
C3	49	49	52	50	49	50

Directness

	SS	TS	C0	C1	C2	C3
SS	**50**	**56**	**65**	**60**	**60**	**60**
TS	44	50	55	52	51	52
C0	35	45	50	47	48	48
C1	40	48	53	50	50	50
C2	40	49	52	50	50	50
C3	40	48	52	50	50	50

News articles

Comprehensiveness

	SS	TS	C0	C1	C2	C3
SS	50	20	28	25	21	21
TS	80	50	44	41	38	36
C0	72	56	50	52	54	52
C1	**75**	**59**	**48**	**50**	**58**	**55**
C2	79	62	46	42	50	59
C3	79	64	48	45	41	50

Diversity

	SS	TS	C0	C1	C2	C3
SS	50	33	38	35	29	31
TS	67	50	53	45	44	40
C0	62	47	50	40	41	41
C1	65	55	60	50	50	50
C2	**71**	**56**	**59**	**50**	**50**	**51**
C3	69	60	59	50	49	50

Empowerment

	SS	TS	C0	C1	C2	C3
SS	50	47	57	49	50	50
TS	53	50	58	50	50	48
C0	43	42	50	42	45	44
C1	**51**	**50**	**58**	**50**	**52**	**51**
C2	50	50	55	48	50	50
C3	50	52	56	49	50	50

Directness

	SS	TS	C0	C1	C2	C3
SS	**50**	**54**	**59**	**55**	**55**	**54**
TS	46	50	55	53	52	52
C0	41	45	50	48	48	47
C1	45	47	52	50	49	49
C2	45	48	52	51	50	49
C3	46	48	53	51	51	50

Baseline methods

- C0: root level community
- C1: high level community
- C2: intermediate level community
- C3: low level community
- TS: Only use map reduce without KG
  (Use pure text not KG)
- SS: naïve semantic RAG
  (Do not specify.
  Probably semantic search)

4. Conclusion

- **Contribution**

Conclusion

Contribution

Contributions

- ❑ Show effectiveness on **global question** answering on large datasets
- ❑ Leverage the power of LLMs for constructing, summarizing, and evaluating
- ❑ Use the natural modularity of graphs to partition data for global summarization

What I want to provide

- ❑ The power of LLM in the research
- ❑ How to leverage Graph RAG with LLMs

Contents Beyond Relevant Documents

Beyond Relevant Documents

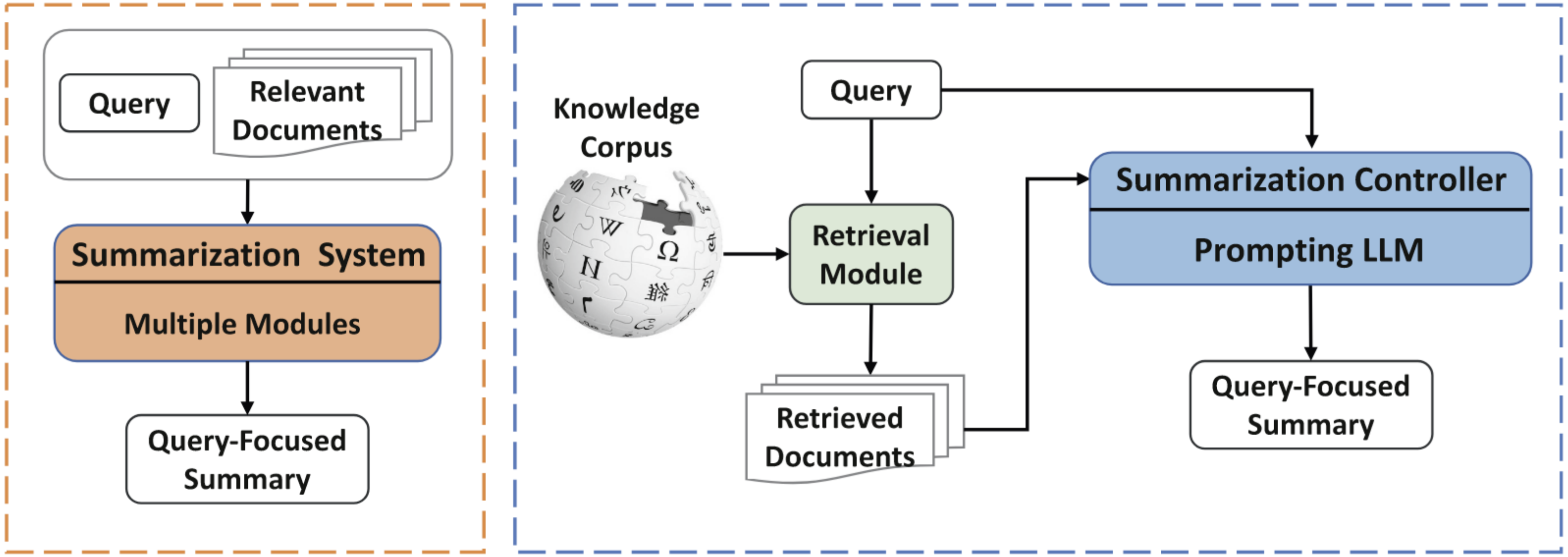
1. Introduction

- Existing Problems
- Motivation

Introduction

Existing Problems & Motivation

- Existing QFS methods **assume the availability of a set of relevant documents**
- However, this assumption is **impractical**
- Therefore, need to reframe QFS as a **knowledge-intensive (KI) task setup**
  - **Knowledge intensive task**: An environment where a **very large** external knowledge corpus are utilized



(a) Conventional Approach (b) Our Knowledge-Intensive Approach

Beyond Relevant Documents

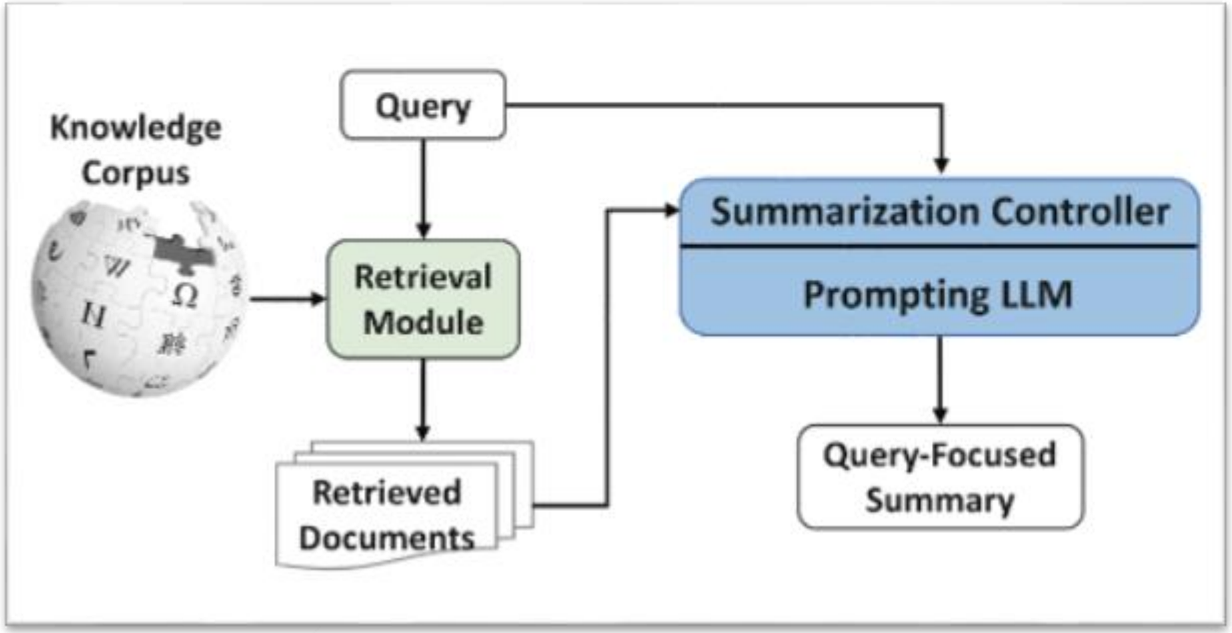
2. Methodology

- Retrieval Module
- Summarization Controller
- Datasets setup

Methodology

Retrieval Module

- Estimates the **relevance Score Rel(q, d_i)** and then **ranks** all documents
- Two models
  - Sparse model, **BM25**[1]: Based on using weighted counts of overlapping terms
  - Dense model, **DPR**[2]: Based on embedding vector similarity

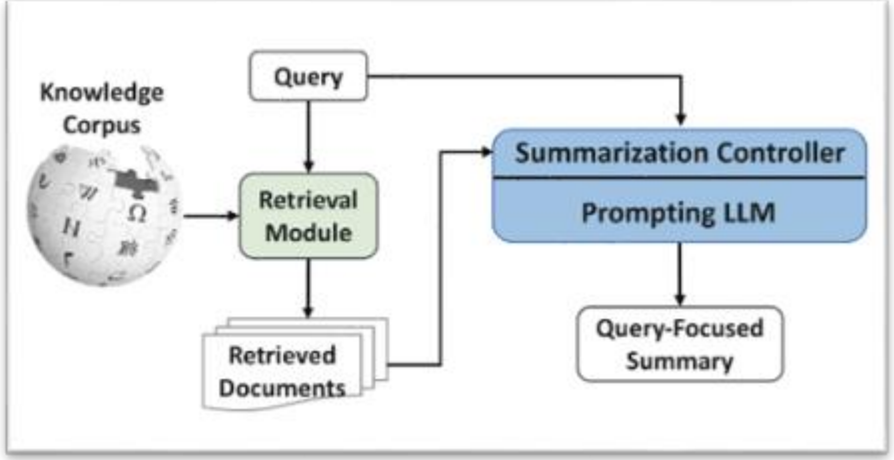


[1] The probabilistic relevance framework: BM25 and beyond. Found. Trends Inf. Retr. '09
[2] Dense passage retrieval for open-domain question answering. EMNLP '20

Methodology

Summarization Controller

- Feed the query and the documents into **LLM** with specific prompt to generate a summary
  - Prompts contain **identifying relevant information** and **generating a summary**



Prompt 3.1: Summarization Controller

Instruction: You will be given a query and a set of documents. Your task is to generate an informative, fluent, and accurate query-focused summary. To do so, you should obtain a query-focused summary step by step.

Step 1: Query-Relevant Information Identification
In this step, you will be given a query and a set of documents. Your task is to find and identify query-relevant information from each document. This relevant information can be at any level, such as phrases, sentences, or paragraphs.

Step 2: Controllable Summarization
In this step, you should take the query and query-relevant information obtained from Step 1 as the inputs. Your task is to summarize this information. The summary should be concise, include only non-redundant, query-relevant evidence, and be approximately 250 words long.

Demonstrations:
Few-shot human-written demonstrations.

Query: {*Input query*}
Documents: {*Retrieved documents*}

Methodology

Datasets setup

- **Variation of datasets**
  - **$\mathbf{CORPUS_{Int}}$**: Combine all of the reference documents of DUC 2005-2007 (**32K** documents)
  - **$\mathbf{CORPUS_{Ext}}$**: Use Wikipedia dump datasets as external (not reference) documents (**21M** documents)
  - **$\mathbf{CORPUS_{Avg}}$**: Combine $CORPUS_{Int}$ and $CORPUS_{Ext}$
- **Relevance Annotation**
  - Collect **human annotations** to evaluate **the relevancy score** between query and documents
  - Collect 200 candidate documents for each query by using each Retrieval Module
  - Ask workers to label whether a document contains either part of the summary or key to a query

Beyond Relevant Documents

3. Experiment

- Setup & Metrics
- Result

Experiment

Setup & Metrics

- **Backbone Models**
  - **Retrieval Module**: BM25 or DPR
  - **Summarization controller**: GPT-3.5
  - **Number or retrieved documents**: 50
- **Evaluation Metrics**
  - **Lexical Metrics**
    - **ROUGE**: lexical overlap based on n-gram matching and recall
  - **Semantic Metrics**
    - **BERTScore**: using **contextual** embeddings from a pre-trained BERT model
    - **BARTScore**: how likely a reference summary is the generated summary by capturing fluency and coherence (Predict next sentence)

Experiment

Reuslt2

Beyond Relevant Documents

Corpus	Model	P@10	P@50	R@10	R@50
CORPUS$_{Int}$	BM25	**16.0***	**12.0**	**8.2***	**30.2***
	DPR	11.6	11.5	6.0	27.8
CORPUS$_{Ext}$	BM25	12.7	8.6	**7.0**	22.2
	DPR	**13.8***	**12.6***	6.9	**29.7***
CORPUS$_{Aug}$	BM25	12.2	8.9	6.8	23.2
	DPR	**14.4***	**12.5***	**7.4**	**30.3***

	CORPUS$_{Int}$					CORPUS$_{Ext}$					CORPUS$_{Aug}$				
Model	R1	R2	RS	BE	BA	R1	R2	RS	BE	BA	R1	R2	RS	BE	BA
Weakly Supervised															
QUERYSUM [34]	36.1	7.5	12.7	8.5	32.3	31.1	4.5	10.2	2.0	28.9	32.6	5.5	11.1	3.0	29.8
MARGESUM [35]	38.0	9.1	14.3	11.5	32.8	34.4	6.5	12.2	5.9	30.3	36.7	8.1	13.5	8.7	32.1
Supervised															
RAG-Sequence [19]	28.9	5.7	10.1	12.6	4.1	32.3	5.2	10.8	8.0	3.9	27.1	4.6	9.0	8.3	3.9
FiD [12]															
- BM25	42.4	11.3	16.5	21.4	38.1	38.8	8.4	14.2	15.7	35.0	41.4	10.8	16.1	20.0	36.8
- DPR	41.5	10.7	15.9	21.4	39.0	38.6	8.0	14.1	15.5	34.1	40.0	9.4	15.1	18.0	36.7
Zero-Shot Prompted															
NaiveRAG [9]															
- BM25	36.4	10.5	14.4	26.8	39.5	31.3	7.0	11.4	19.6	34.9	32.8	8.1	12.4	23.4	37.2
- DPR	37.1	10.4	14.7	27.3	39.5	31.7	7.1	11.4	18.3	34.8	33.7	8.4	12.6	21.8	37.1
Ours															
- BM25	43.1	11.0	16.5	25.9	40.7	37.9	7.3	13.1	19.9	34.5	39.4	8.8	14.3	22.1	37.2
- DPR	42.8	11.0	16.3	25.5	40.6	36.7	7.0	12.5	17.7	33.8	39.2	8.6	14.0	21.5	37.2
Few-Shot Prompted															
NaiveRAG [9]															
- BM25	41.7	11.9	16.5	24.4	41.7	35.2	7.8	12.8	17.1	35.5	37.5	9.1	14.2	20.2	37.7
- DPR	42.3	11.7	16.5	25.1	41.6	34.3	7.5	12.3	16.7	35.1	36.9	9.1	13.9	19.1	37.9
Ours															
- BM25	**45.8***	**13.4***	**18.6***	**29.2***	44.7	**41.7***	**9.6**	**15.6***	22.5	**37.3***	**43.6***	**11.3**	**16.7**	**26.1***	40.8
- DPR	45.1	12.9	18.3	28.6	**44.8***	41.1	9.4	15.2	**22.7***	36.8	42.9	10.7	16.2	24.7	**41.7***

Beyond Relevant Documents

4. Conclusion

- **Contribution**

Conclusion

Contribution

Contributions

- ❑ Propose brand-new QFS method in the knowledge-intensive (KI) task setup
- ❑ Construct new benchmark measuring document retrieval performance by using human annotations

What I want to provide

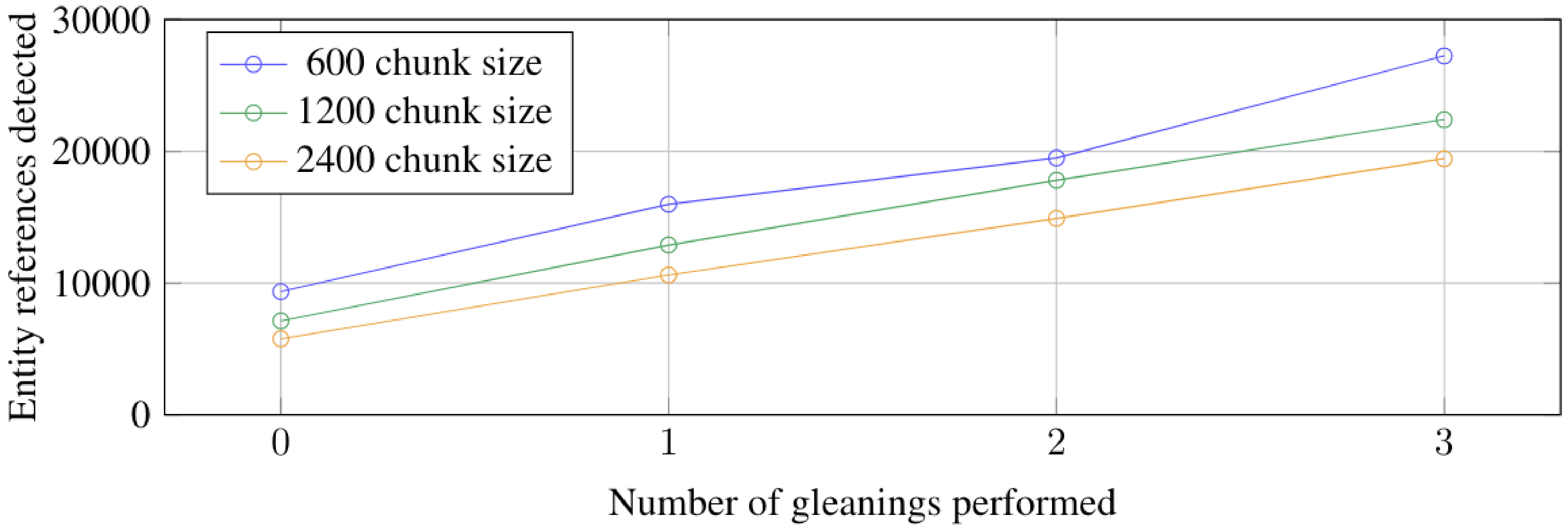
- ❑ The overview of QFS using LLM

A. Appendix

Appendix

- Granularity
  - **Longer text** chunks require **fewer LLM calls** for such extraction

    However, could suffer from the recall degradation of longer LLM context windows



Appendix

Text Chunks → Element Instances

- To **identify** and **extract** instances of graph nodes and edges from each chunk
  - Using **a multipart LLM prompt**
    - **First,** **identifies all entities** in the text, including name, type, and description

      and **clearly-related entities**, including the source and target entities and a relationship description
    - **Second,** **extract any additional covariates** that we would like to associate with the extracted nodes.

      Aims to extract ***claims*** linked to detected entities
      - Subject, object, type, description, source text span, and start and end dates
  - Use multiple rounds of '**gleanings**' up to a specified maximum,

    to encourage the LLM to detect any additional entities

Appendix

Element Instances → Elements Summaries

- ❑ Convert all instance-level summaries into single blocks of descriptive text for **each graph element**

  → **Convert into Knowledge Graph**

  - Requires a further round of LLM summarization over matching groups of instances
  - **Potential concern**: the LLM may inconsistently extract the same entity (e.g., MS and Microsoft)
    - All closely-related ‘communities’ of entities will be detected and summarized in the following step
    - LLM can understand the common entity behind multiple name variations

    → Our overall approach is resilient to such variations