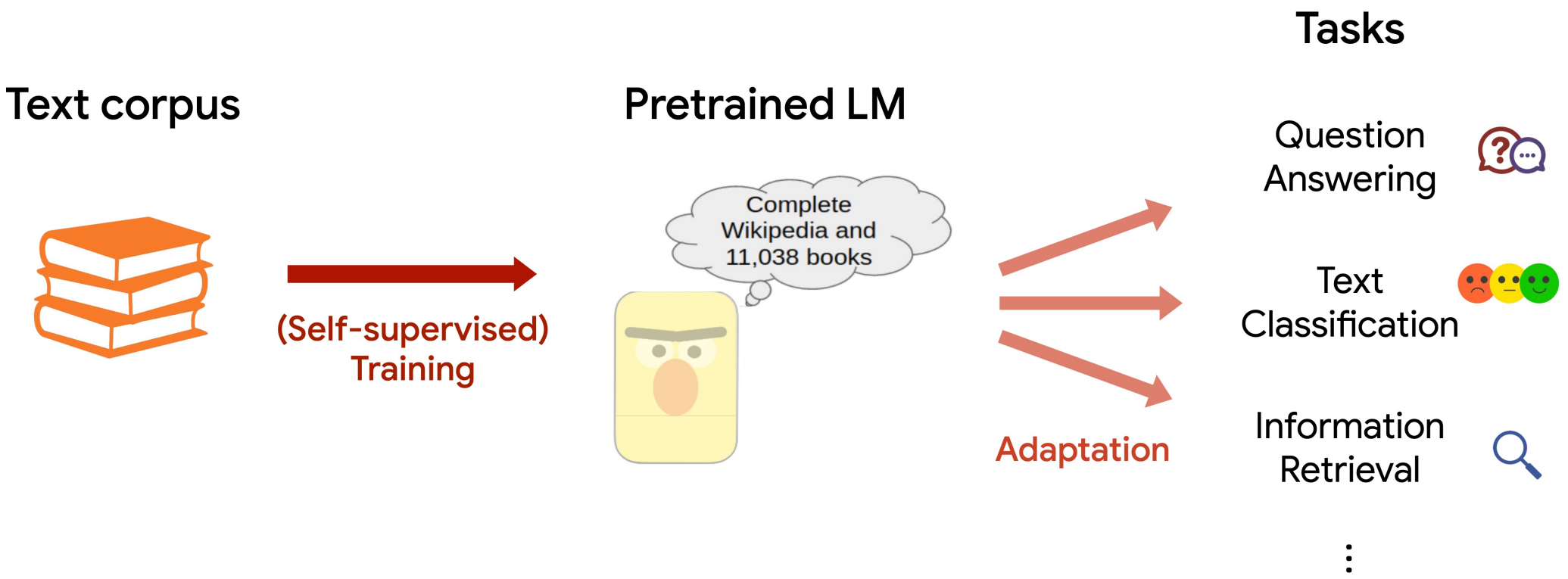
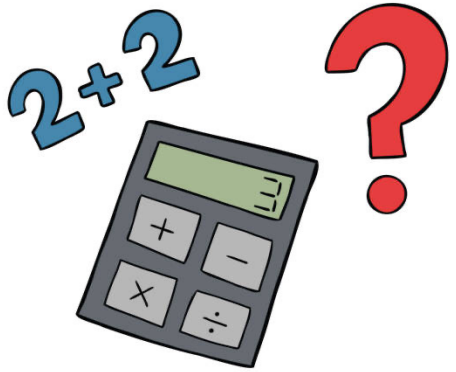
Pre-Training



Limitations of LLMs

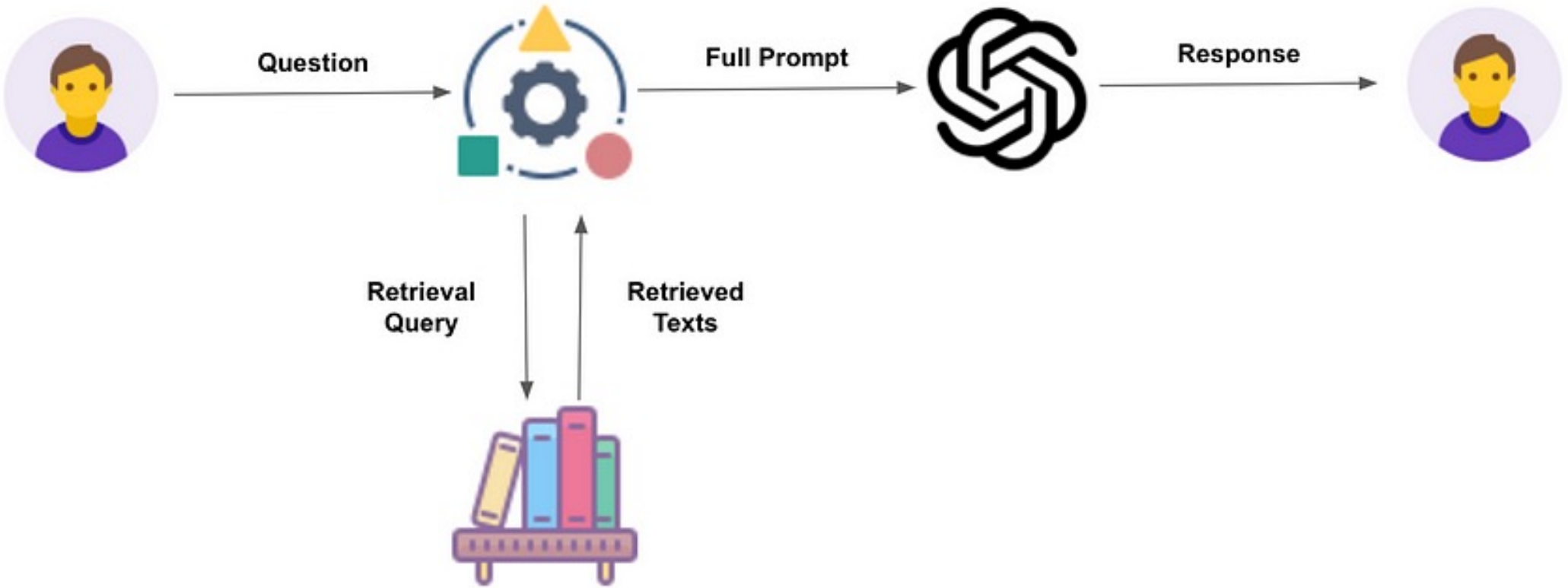


Retrieval-Augmented Generation

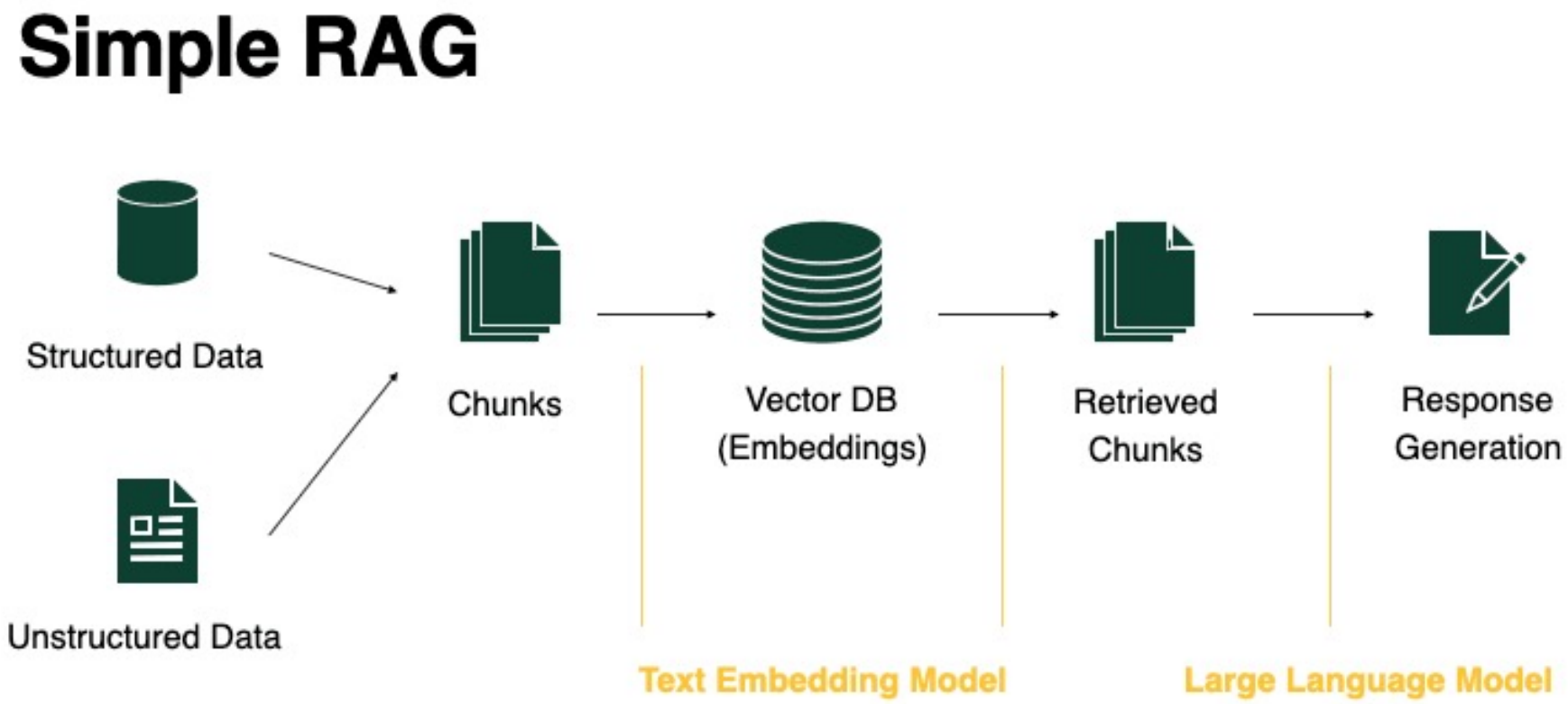
Private Knowledge

→

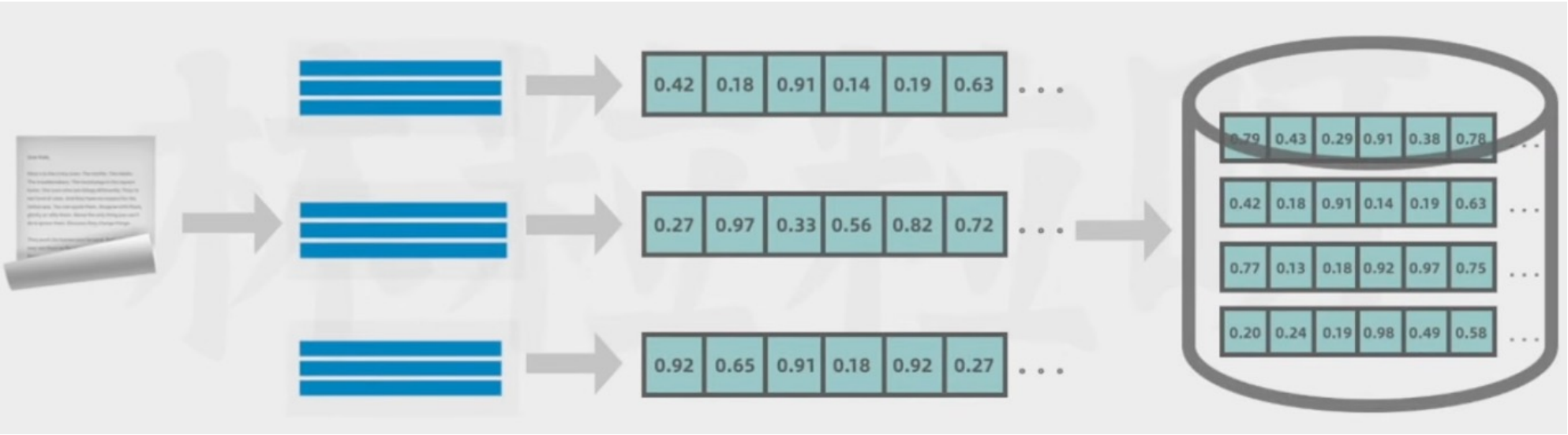
Retrieval-Augmented Generation



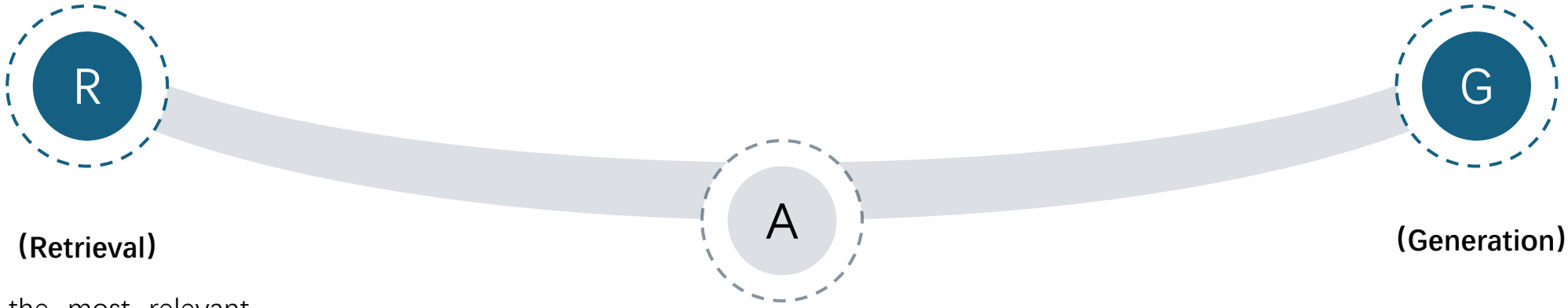
Retrieval-Augmented Generation (RAG)



Retrieval-Augmented Generation (RAG)



RAG (Retrieval-Augmented Generation)



(Retrieval)

Finds the most relevant information from external knowledge bases or document collections. By vectorizing and storing text data, the system efficiently retrieves the best matches using query vectors from user input.

(Augmentation)

Uses retrieved information as additional context for the language model, enhancing its understanding and improving answer accuracy.

(Generation)

Powered by LLMs, it generates coherent, context-aware, high-quality answers by combining user queries, retrieved information, and its language generation capabilities.

Naive RAG Limitation

1

Relies on flat data, limiting complex relationships.

Naive RAG uses text chunks, ignoring entity relations and complex cross-document dependencies.

2

Lacks context awareness, leading to incoherent answers.

Retrieval feeds generation linearly, lacking global context integration.

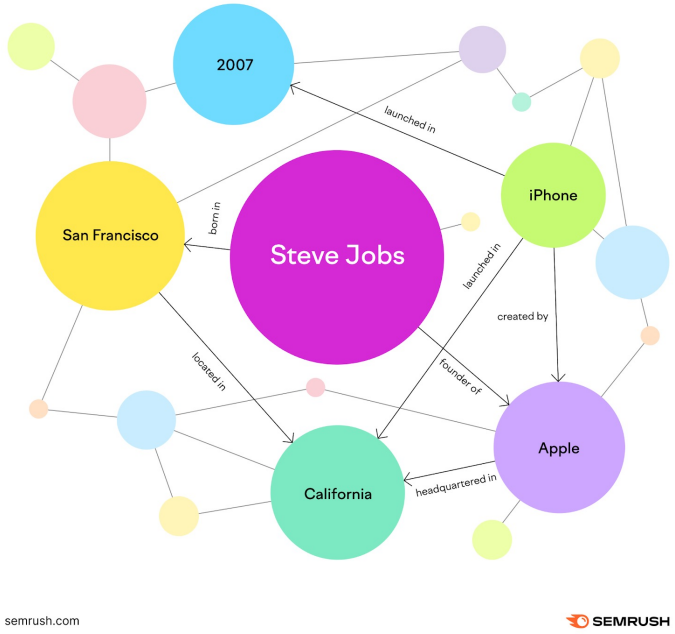
3

Text blocks are redundant with irrelevant content.

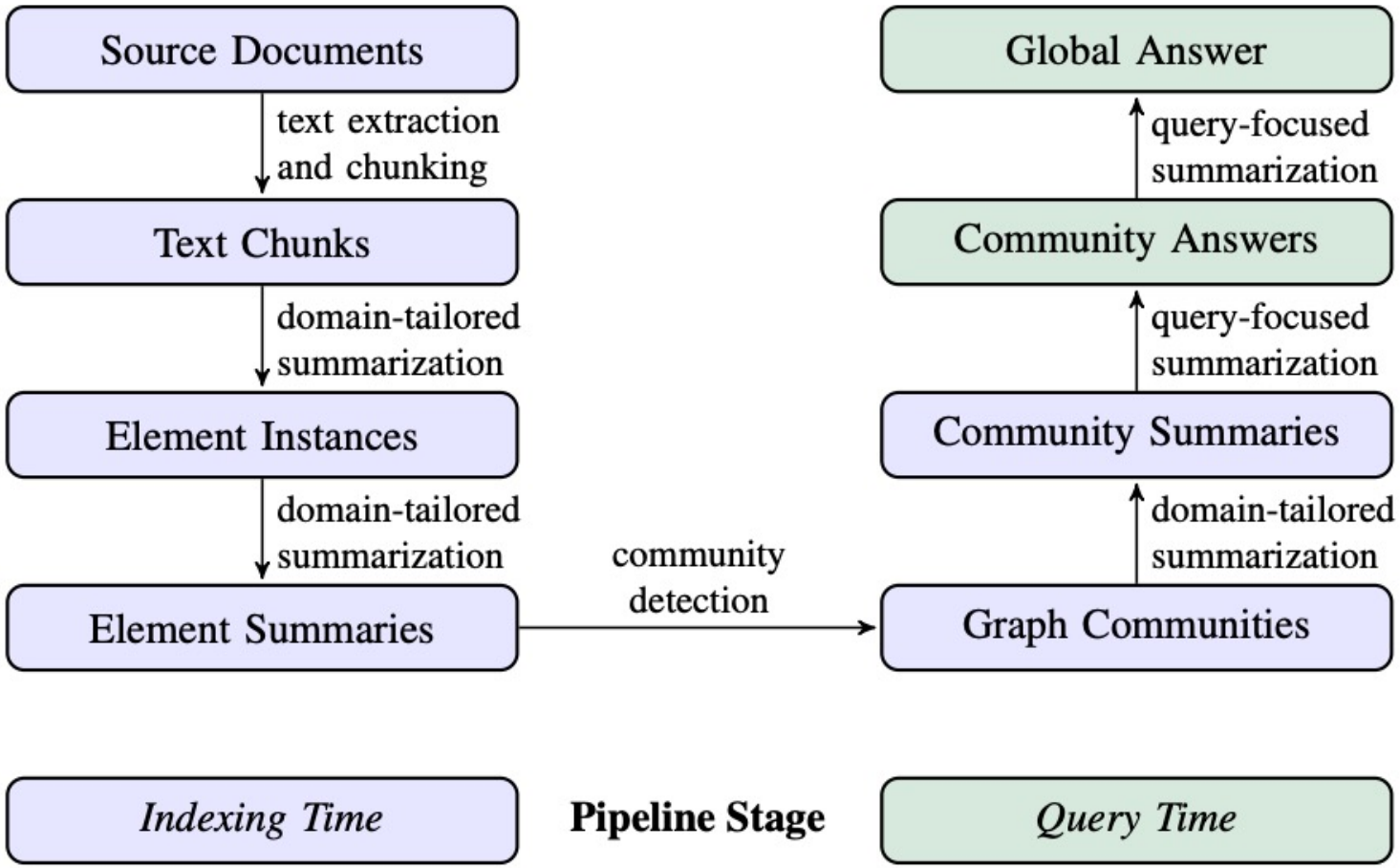
Naive RAG retrieval often includes irrelevant information, increasing computational load and reducing answer accuracy.

Graph RAG

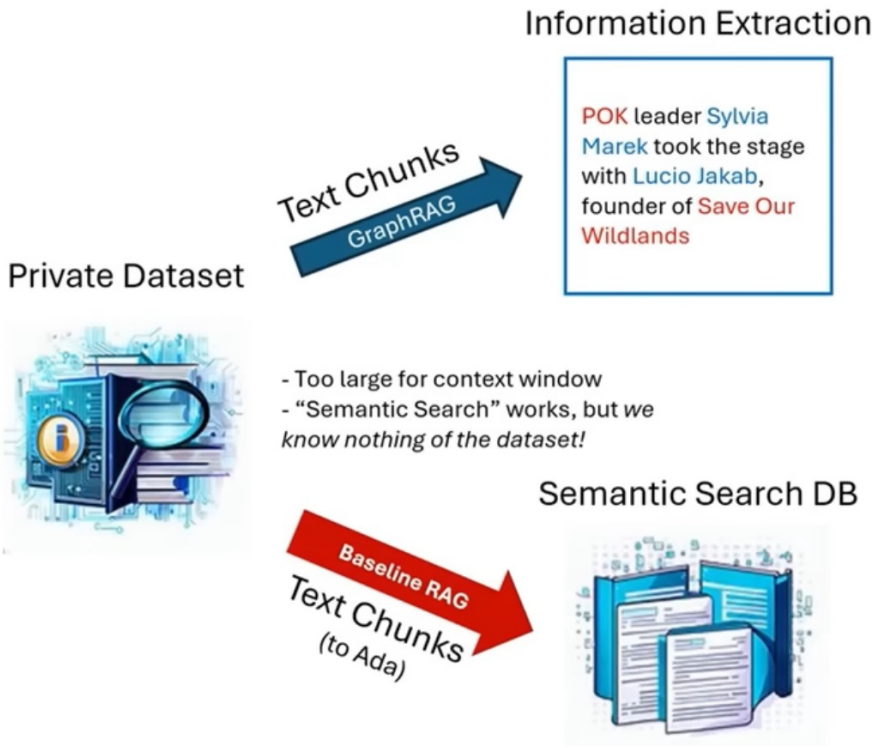
Unlike a baseline RAG that uses a vector database to retrieve semantically similar text, GraphRAG enhances RAG by incorporating knowledge graphs (KGs). Knowledge graphs are data structures that store and link related or unrelated data based on their relationships.

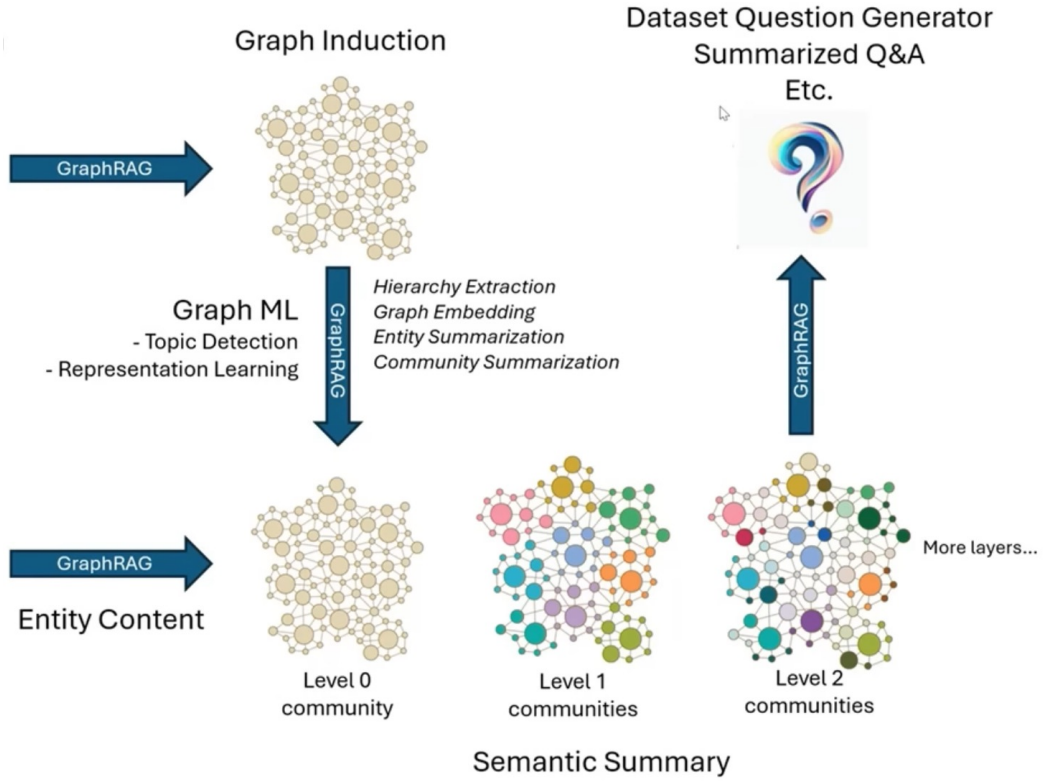


Graph RAG


Source Documents
text extraction and chunking
Text Chunks
domain-tailored summarization
Element Instances
domain-tailored summarization
Element Summaries
community detection
Graph Communities
domain-tailored summarization
Community Summaries
query-focused summarization
Community Answers
query-focused summarization
Global Answer
Indexing Time
Pipeline Stage
Query Time

Graph RAG

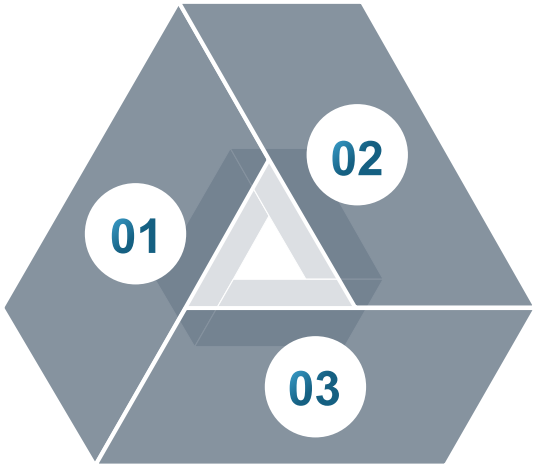

Information Extraction
Text Chunks
GraphRAG
POK leader Sylvia Marek took the stage with Lucio Jakab, founder of Save Our Wildlands
Private Dataset
- Too large for context window
- "Semantic Search" works, but we know nothing of the dataset!
Semantic Search DB
Baseline RAG
Text Chunks
(to Ada)


Graph Induction
Dataset Question Generator
Summarized Q&A
Etc.
GraphRAG
Graph ML
- Topic Detection
- Representation Learning
GraphRAG
Hierarchy Extraction
Graph Embedding
Entity Summarization
Community Summarization
GraphRAG
GraphRAG
Entity Content
Level 0 community
Level 1 communities
Level 2 communities
More layers...
Semantic Summary

GraphRAG—Limitation

High Cost

Generating and traversing community reports for retrieval and query stages leads to extremely high costs.



Inefficiency

Both generating community reports and traversing the community require significant time, which greatly reduces system efficiency.

Poor Scalability

Merging new data requires rebuilding communities, raising costs and limiting scalability.

LightRAG

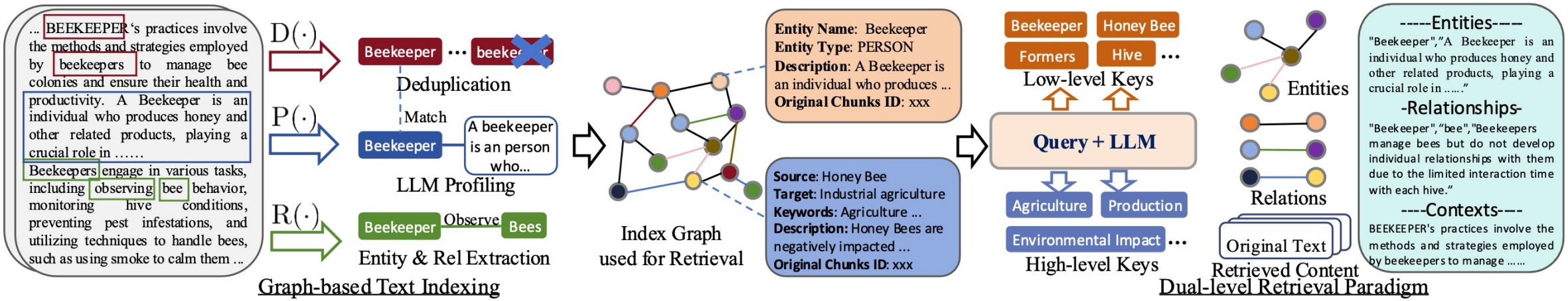


Figure 1: Overall architecture of the proposed LightRAG framework.

🔍 Comprehensive Information Retrieval with Complex Interdependencies

⚙️ Efficient Information Retrieval through a Dual-Level Retrieval Paradigm

⚡ Rapid Adaptability to Dynamic Data Changes

Table 4: Statistical information of the datasets.

Statistics	Agriculture	CS	Legal	Mix
Total Documents	12	10	94	61
Total Tokens	2,017,886	2,306,535	5,081,069	619,009

Evaluation Datasets. To conduct a comprehensive analysis of LightRAG, we selected four datasets from the UltraDomain benchmark (Qian et al., 2024). The UltraDomain data is sourced from 428 college textbooks and encompasses 18 distinct domains, including agriculture, social sciences, and humanities. From these, we chose the Agriculture, CS, Legal, and Mix datasets. Each dataset contains between 600,000 and 5,000,000 tokens, with detailed information provided in Table 4. Below is a specific introduction to the four domains utilized in our experiments:

i) **Comprehensiveness**: How thoroughly does the answer address all aspects and details of the question? ii) **Diversity**: How varied and rich is the answer in offering different perspectives and insights related to the question? iii) **Empowerment**: How effectively does the answer enable the reader to understand the topic and make informed judgments? iv) **Overall**: This dimension assesses the cumulative performance across the three preceding criteria to identify the best overall answer.

Table 1: Win rates (%) of baselines v.s. LightRAG across four datasets and four evaluation dimensions.

	Agriculture		**CS**		**Legal**		**Mix**	
	NaiveRAG	**LightRAG**	NaiveRAG	**LightRAG**	NaiveRAG	**LightRAG**	NaiveRAG	**LightRAG**
Comprehensiveness	32.4%	67.6%	38.4%	61.6%	16.4%	83.6%	38.8%	61.2%
Diversity	23.6%	76.4%	38.0%	62.0%	13.6%	86.4%	32.4%	67.6%
Empowerment	32.4%	67.6%	38.8%	61.2%	16.4%	83.6%	42.8%	57.2%
Overall	32.4%	67.6%	38.8%	61.2%	15.2%	84.8%	40.0%	60.0%
	RQ-RAG	**LightRAG**	RQ-RAG	**LightRAG**	RQ-RAG	**LightRAG**	RQ-RAG	**LightRAG**
Comprehensiveness	31.6%	68.4%	38.8%	61.2%	15.2%	84.8%	39.2%	60.8%
Diversity	29.2%	70.8%	39.2%	60.8%	11.6%	88.4%	30.8%	69.2%
Empowerment	31.6%	68.4%	36.4%	63.6%	15.2%	84.8%	42.4%	57.6%
Overall	32.4%	67.6%	38.0%	62.0%	14.4%	85.6%	40.0%	60.0%
	HyDE	**LightRAG**	HyDE	**LightRAG**	HyDE	**LightRAG**	HyDE	**LightRAG**
Comprehensiveness	26.0%	74.0%	41.6%	58.4%	26.8%	73.2%	40.4%	59.6%
Diversity	24.0%	76.0%	38.8%	61.2%	20.0%	80.0%	32.4%	67.6%
Empowerment	25.2%	74.8%	40.8%	59.2%	26.0%	74.0%	46.0%	54.0%
Overall	24.8%	75.2%	41.6%	58.4%	26.4%	73.6%	42.4%	57.6%
	GraphRAG	**LightRAG**	GraphRAG	**LightRAG**	GraphRAG	**LightRAG**	GraphRAG	**LightRAG**
Comprehensiveness	45.6%	54.4%	48.4%	51.6%	48.4%	51.6%	50.4%	49.6%
Diversity	22.8%	77.2%	40.8%	59.2%	26.4%	73.6%	36.0%	64.0%
Empowerment	41.2%	58.8%	45.2%	54.8%	43.6%	56.4%	50.8%	49.2%
Overall	45.2%	54.8%	48.0%	52.0%	47.2%	52.8%	50.4%	49.6%

Table 2: Performance of ablated versions of LightRAG, using NaiveRAG as reference.

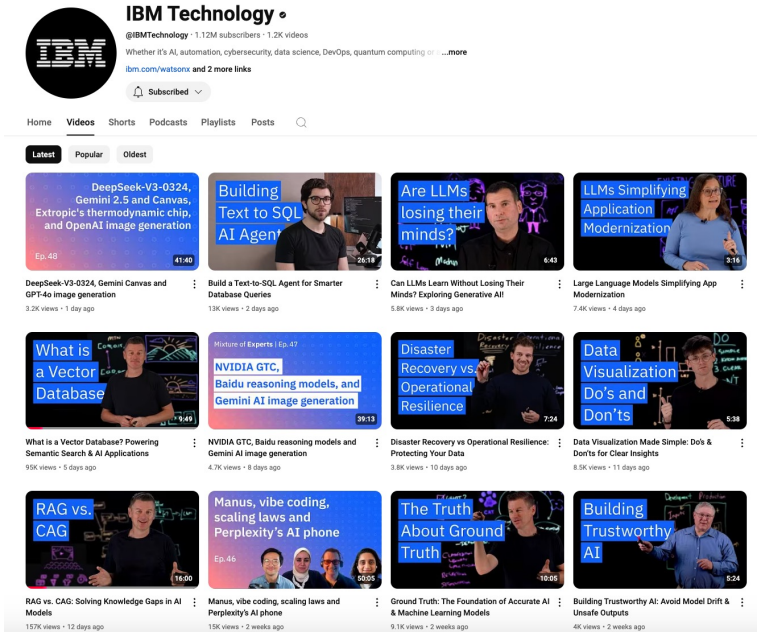
	Agriculture		**CS**		**Legal**		**Mix**	
	NaiveRAG	**LightRAG**	NaiveRAG	**LightRAG**	NaiveRAG	**LightRAG**	NaiveRAG	**LightRAG**
Comprehensiveness	32.4%	67.6%	38.4%	61.6%	16.4%	83.6%	38.8%	61.2%
Diversity	23.6%	76.4%	38.0%	62.0%	13.6%	86.4%	32.4%	67.6%
Empowerment	32.4%	67.6%	38.8%	61.2%	16.4%	83.6%	42.8%	57.2%
Overall	32.4%	67.6%	38.8%	61.2%	15.2%	84.8%	40.0%	60.0%
	NaiveRAG	**-High**	NaiveRAG	**-High**	NaiveRAG	**-High**	NaiveRAG	**-High**
Comprehensiveness	34.8%	65.2%	42.8%	57.2%	23.6%	76.4%	40.4%	59.6%
Diversity	27.2%	72.8%	36.8%	63.2%	16.8%	83.2%	36.0%	64.0%
Empowerment	36.0%	64.0%	42.4%	57.6%	22.8%	77.2%	47.6%	52.4%
Overall	35.2%	64.8%	44.0%	56.0%	22.0%	78.0%	42.4%	57.6%
	NaiveRAG	**-Low**	NaiveRAG	**-Low**	NaiveRAG	**-Low**	NaiveRAG	**-Low**
Comprehensiveness	36.0%	64.0%	43.2%	56.8%	19.2%	80.8%	36.0%	64.0%
Diversity	28.0%	72.0%	39.6%	60.4%	13.6%	86.4%	33.2%	66.8%
Empowerment	34.8%	65.2%	42.8%	57.2%	16.4%	83.6%	35.2%	64.8%
Overall	34.8%	65.2%	43.6%	56.4%	18.8%	81.2%	35.2%	64.8%
	NaiveRAG	**-Origin**	NaiveRAG	**-Origin**	NaiveRAG	**-Origin**	NaiveRAG	**-Origin**
Comprehensiveness	24.8%	75.2%	39.2%	60.8%	16.4%	83.6%	44.4%	55.6%
Diversity	26.4%	73.6%	44.8%	55.2%	14.4%	85.6%	25.6%	74.4%
Empowerment	32.0%	68.0%	43.2%	56.8%	17.2%	82.8%	45.2%	54.8%
Overall	25.6%	74.4%	39.2%	60.8%	15.6%	84.4%	44.4%	55.6%

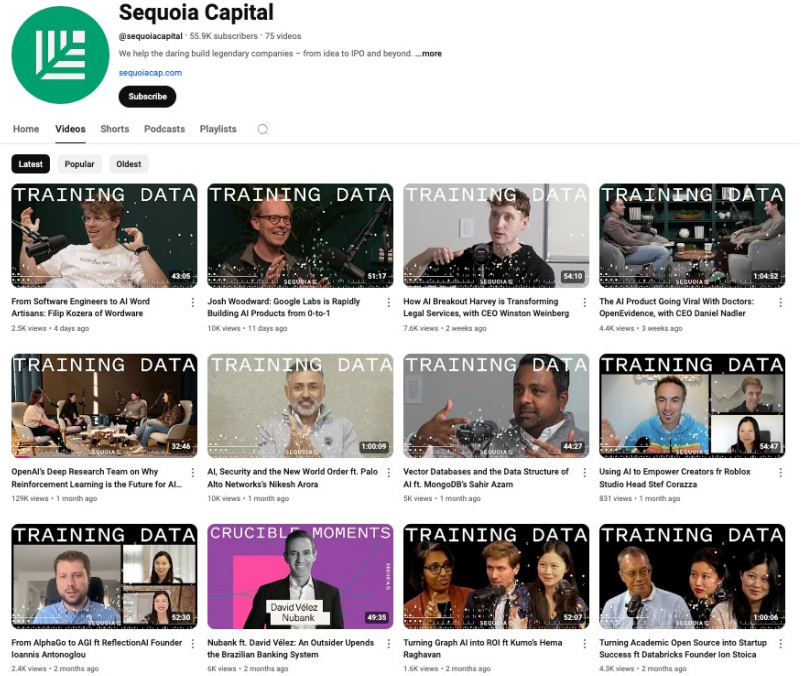
On-Device RAG

Data Privacy

Resource Limited

Multi-Modal RAG



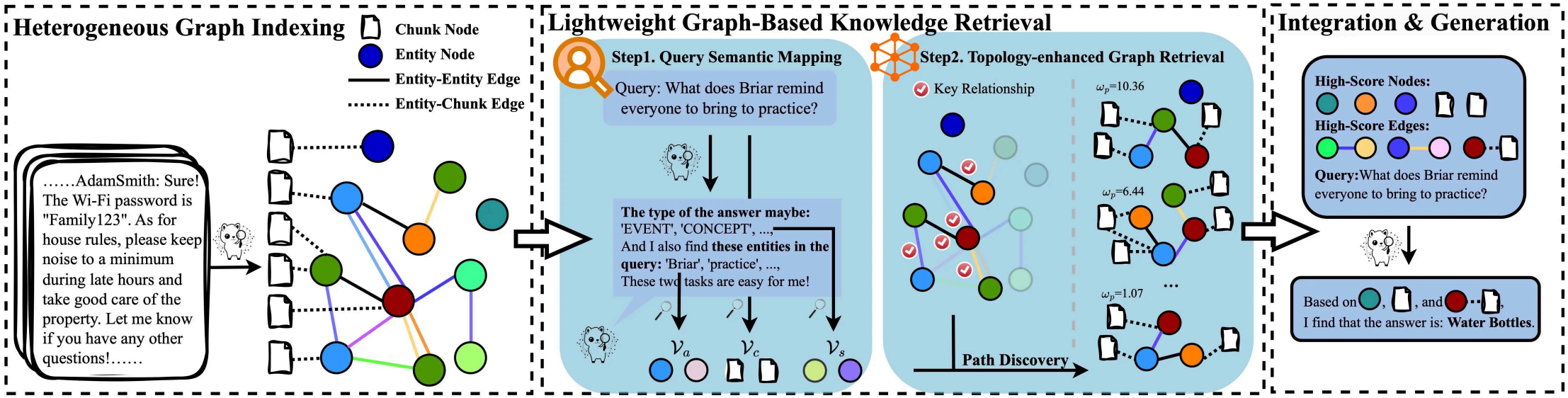


2025/8/20

MiniRAG


MiniRAG


Making new friends
Morning and joys
Li Hua
WhatsApp
Worries and joys
Worries and joys


Heterogeneous Graph Indexing
Chunk Node
Entity Node
Entity-Entity Edge
Entity-Chunk Edge
……AdamSmith: Sure! The Wi-Fi password is "Family123". As for house rules, please keep noise to a minimum during late hours and take good care of the property. Let me know if you have any other questions!……
Lightweight Graph-Based Knowledge Retrieval
Step1. Query Semantic Mapping
Query: What does Briar remind everyone to bring to practice?
The type of the answer maybe: 'EVENT', 'CONCEPT', ...,
And I also find these entities in the query: 'Briar', 'practice', ...,
These two tasks are easy for me!
$\mathcal{V}_a$
$\mathcal{V}_c$
$\mathcal{V}_s$
Step2. Topology-enhanced Graph Retrieval
Key Relationship
Path Discovery
ω_p=10.36
ω_p=6.44
ω_p=1.07
Integration & Generation
High-Score Nodes:
High-Score Edges:
Query:What does Briar remind everyone to bring to practice?
Based on , , and ,
I find that the answer is: Water Bottles.

What is MiniRAG?

MiniRAG is a **free**, **lightweight**, yet **powerful** RAG framework.

Free: No need for API-based LLMs; local SLMs can also handle RAG tasks.
Lightweight: Requires only 25% of the storage space compared to leading baselines.
Strong: MiniRAG+3B model > GraphRAG+gpt-4o-mini

Observation of SLMs on RAG tasks

Limitation of Indexing

Original Text:
LiHua: Hey Adam! Hope you're doing well. Just wanted to ask for the Wi-Fi password when you get a chance. Also, are there any house rules I should know about? Thanks!
AdamSmith: Sure! The Wi-Fi password is "Family123". As for house rules, please keep noise to a minimum during late hours and take good care of the property. Let me know if you have any other questions!
LiHua: Got it, thanks! I'll make sure to keep it down at night. Just one more thing, is it okay to have friends over occasionally?
AdamSmith: You're welcome! Yes, having friends over occasionally is fine, just try to keep the gatherings reasonable. Enjoy!
LiHua: Awesome, appreciate it! Looking forward to settling in.
gpt-4o-mini response:
Entity: <HOUSE RULES>
Description: <House Rules refer to the guidelines provided by Adam Smith regarding noise and care of the property, aimed at maintaining a peaceful living environment.>
Phi-3.5-mini response:
Entity: <HOUSE RULES>
Description: <House rules are guidelines provided by AdamSmith for LiHua to follow while living in the house.>

Limitation of Answering

Input Prompt:
You are an assistant designed to answer questions. Please respond to the user's query based on the information provided by the user. If there is insufficient information, please reply with "I don't have enough information."
Here is the information:
Entities: <HOUSE RULES> ... and 49 other entities.
Relations: <LiHua, AdamSmith> ... and 49 other relationships.
Chunk: 10 chunks
Query: What are the House Rules mentioned by Adam Smith?

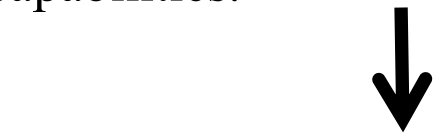
gpt-4o-mini response:
Based on <HOUSE RULES> and <LiHua, AdamSmith>, I know that the House Rules mentioned by Adam Smith is about...
Phi-3.5-mini response:
I'm sorry, I don't have enough information.

Left: SLMs generate notably lower-quality descriptions than LLMs. **Right:** When processing identical inputs, SLMs struggle to locate relevant information in large contexts, while LLMs perform this task effectively.

Observation 1: while SLMs struggle with sophisticated semantic understanding, they excel at pattern matching and localized text processing
Observation 2: explicit structural information can effectively compensate for limited semantic capabilities
Observation 3: decomposing complex RAG operations into simpler, well-defined steps can maintain robustness without requiring advanced reasoning capabilities.

Motivation of MiniRAG

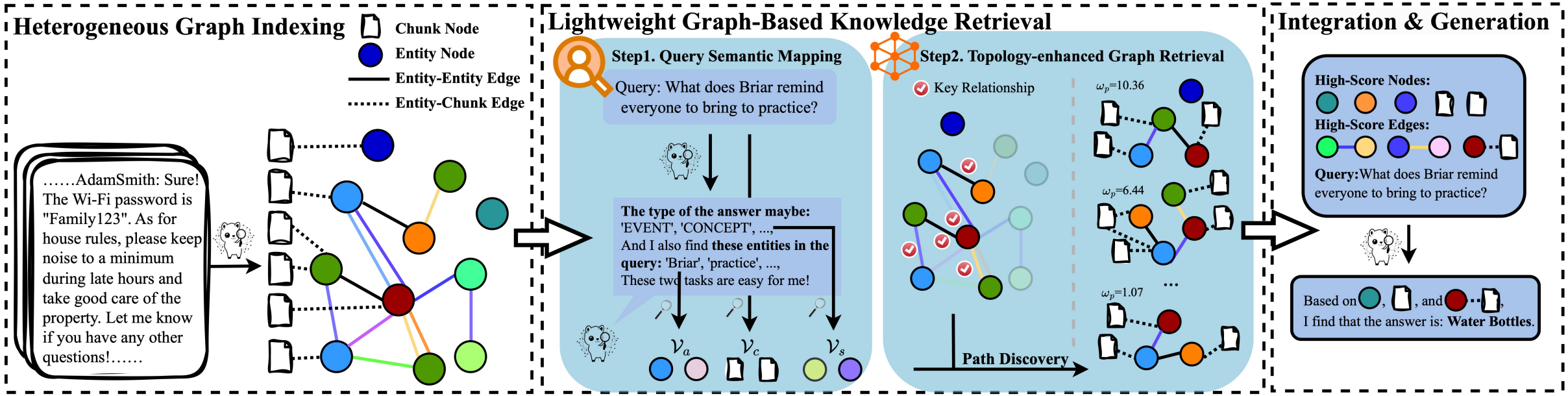
Observation 1: while SLMs struggle with sophisticated semantic understanding, they excel at pattern matching and localized text processing
Observation 2: explicit structural information can effectively compensate for limited semantic capabilities
Observation 3: decomposing complex RAG operations into simpler, well-defined steps can maintain robustness without requiring advanced reasoning capabilities.



To **alleviate the high demands on model capabilities** imposed by the RAG framework, We should use
(1) a new indexing mechanism to reduce reliance on complex semantic understanding
(2) a new retrieval approach for efficient knowledge discovery.

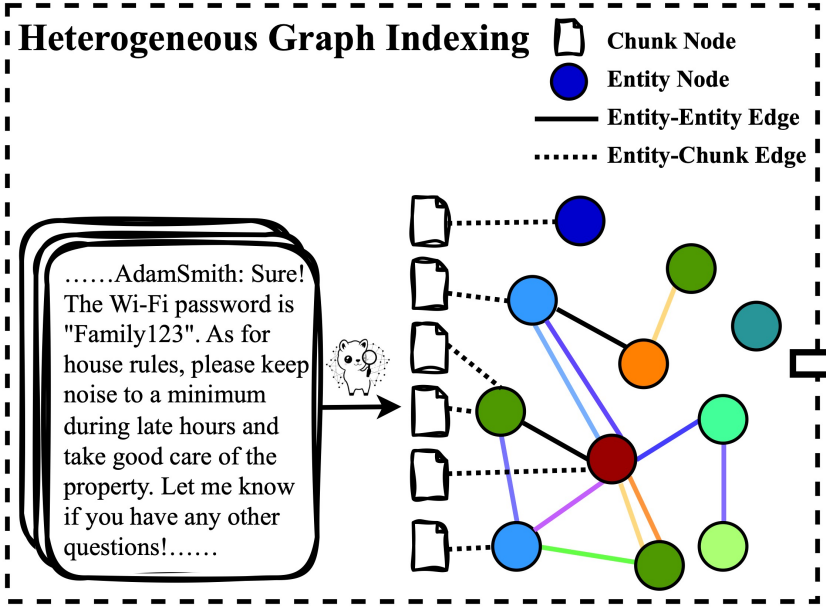
The MiniRAG framework

(1) a new indexing mechanism to reduce reliance on complex semantic understanding
(2) a new retrieval approach for efficient knowledge discovery.



The MiniRAG framework employs a streamlined workflow built on the key components: **(1) heterogeneous graph indexing** and **(2) lightweight graph-based knowledge retrieval**.

Details of MiniRAG: Heterogeneous Graph Indexing



Target:

(1) The indexing mechanism should **extract the key relationships and contextual connections within the data**.
(2) The indexing approach should **condense the retrieved content to its most query-relevant elements**.

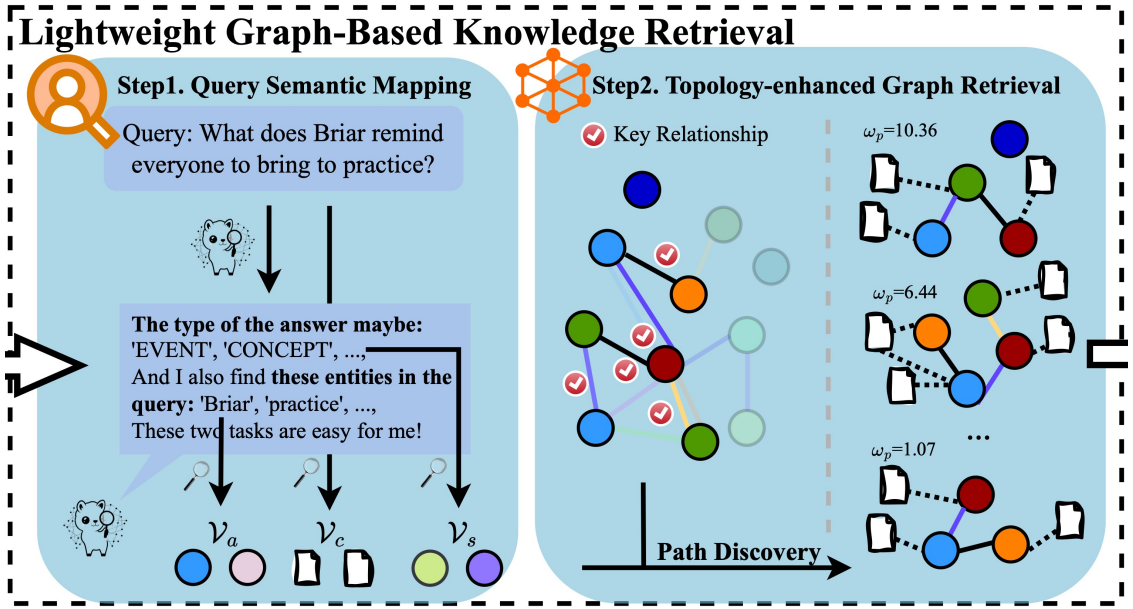
Key components:

- Text Chunk Node v_c
- Entity Node v_e
- Entity-Entity Connection $\mathcal{E}_\alpha$
- Entity-Chunk Connection $\mathcal{E}_\beta$

Advantages:

1. It systematically incorporates both text chunks and named entities, **creating a rich semantic network**
2. It enables data chunks to directly participate in retrieval and effectively **mitigates information distortion**
3. It **avoids challenging tasks** (e.g., sophisticated semantic understanding) that could cause SLMs to fails.

Details of MiniRAG: **lightweight graph-based knowledge retrieval**.



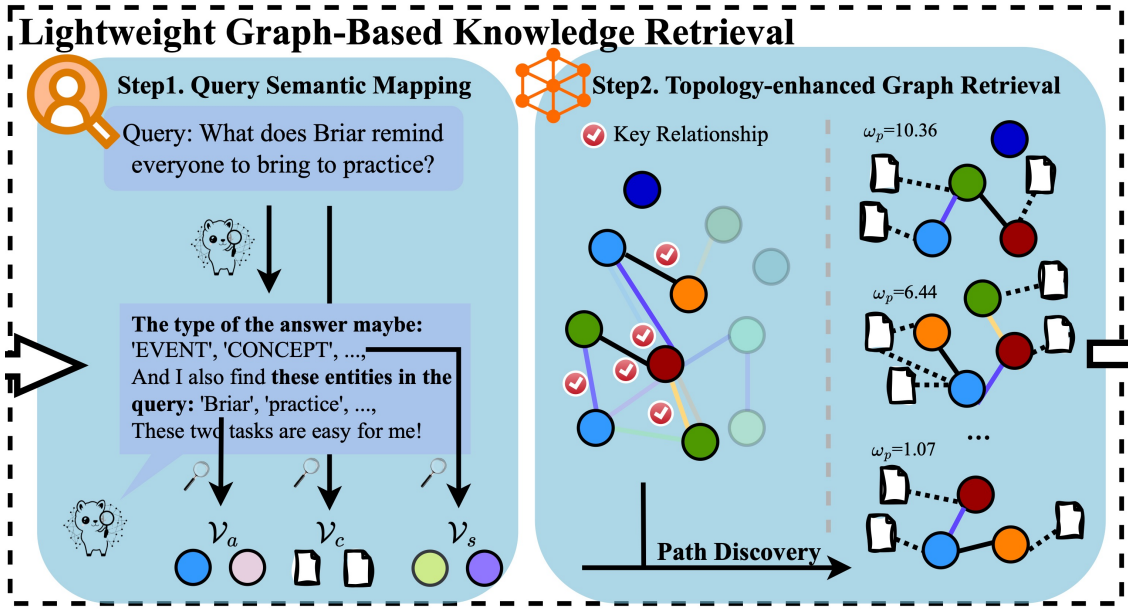
Target:

(1) **Reduce the complexity of input content.** For generation, ensuring that semantic information is clear and concise;
(2) **Shorten the length of input content.** For SLMs, facilitating improved comprehension and retrieval accuracy.

Step 1: Query Semantic Mapping

(1) Extract relevant entities v_q and predict potential types of query
(2) Query-driven reasoning path discovery:
- Initial Entity Identification $\hat{v}_s$
- Answer-Aware Entity Selection $\hat{v}_a$
- Context-Rich Path Formation $\hat{v}_c$

Details of MiniRAG: **lightweight graph-based knowledge retrieval**.



Target:

(1) **Reduce the complexity of input content.** For generation, ensuring that semantic information is clear and concise; ✓

(2) **Shorten the length of input content.** For SLMs, facilitating improved comprehension and retrieval accuracy.

Step 2: Topology-Enhanced Graph Retrieval

(1) Key Relationship Identification

$$\omega_e(e) = \sum_{\hat{v}_s \in \hat{\mathcal{V}}_s} \text{count}(\hat{v}_s, \hat{\mathcal{G}}_{e,k}) + \sum_{\hat{v}_a \in \hat{\mathcal{V}}_a} \text{count}(\hat{v}_a, \hat{\mathcal{G}}_{e,k}),$$

(2) Query-guided path discovery

$$\omega_p(p \mid v_q) = \omega_v(\hat{v}_s \mid v_q) \cdot (1 + \sum_{v \in (p \wedge \hat{\mathcal{V}}_a)} \text{count}(v, p) + \sum_{e \in (p \wedge \hat{\mathcal{E}}_\alpha)} \omega_e(e)).$$

(3) Retrieval of Query-Relevant Text Chunks

①Candidate Filtering $\hat{v}_c \wedge v_c^q$ ②Similarity Computation ③Ranking&Selection

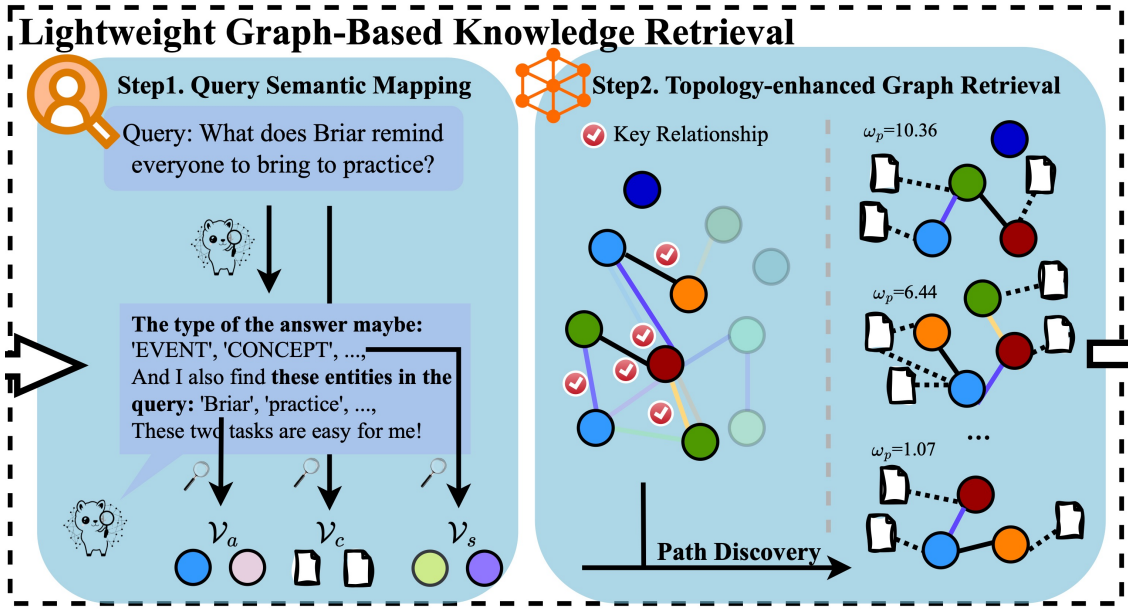
(4) Integration for Augmented Generation

Details of MiniRAG: **lightweight graph-based knowledge retrieval**.

Target:

(1) **Reduce the complexity of input content.** For generation, ensuring that semantic information is clear and concise; ✓

(2) **Shorten the length of input content.** For SLMs, facilitating improved comprehension and retrieval accuracy.✓



Advantages:

1. It effectively **mitigates the issue of significant noise** introduced into the retrieval process due to the limited semantic understanding capabilities of SLMs.
2. It **avoids challenging tasks** (e.g., predicting the high-level information of the query) that could cause SLMs to fails.

New Datasets: LiHua-World

[Design for on-device RAG] The LiHua-World dataset is specifically designed for local RAG scenarios, containing **one year of chat records**.
[Diverse types of questions] The dataset includes single-hop, multi-hop, and summarization questions, each accompanied by manually annotated answers and supporting documents.
[Wide variety of events] The content covers various aspects of daily life, such as social interactions, fitness training, recreational activities, and personal affairs.

Experiment

RQ1: Comparative Performance. How does MiniRAG perform against state-of-the-art alternatives in terms of retrieval accuracy and efficiency?
RQ2: Component Analysis. What is the contribution of key components to MiniRAG's overall effectiveness?
RQ3: Case Studies. How effectively does MiniRAG handle complex, multi-step reasoning tasks with small language models, as demonstrated through practical case studies?

Table 1: Performance evaluation using accuracy (acc) and error (err) rates, measured as percentages (%). Higher accuracy and lower error rates indicate better RAG performance. Results compare various baseline methods against our MiniRAG across multiple datasets. Bold values indicate best performance, while "/" denotes cases where methods failed to generate effective responses.

LiHuaWorld	NaiveRAG		GraphRAG		LightRAG		MiniRAG	
	acc↑	*err*↓	*acc*↑	*err*↓	*acc*↑	*err*↓	*acc*↑	*err*↓
Phi-3.5-mini-instruct	41.22%	23.20%	/	/	39.81%	25.39%	**53.29%**	23.35%
GLM-Edge-1.5B-Chat	42.79%	24.76%	/	/	35.74%	25.86%	**52.51%**	25.71%
Qwen2.5-3B-Instruct	43.73%	24.14%	/	/	39.18%	28.68%	**48.75%**	26.02%
MiniCPM3-4B	43.42%	17.08%	/	/	35.42%	21.94%	**51.25%**	21.79%
gpt-4o-mini	46.55%	19.12%	35.27%	37.77%	**56.90%**	20.85%	54.08%	19.44%

MultiHop-RAG	NaiveRAG		GraphRAG		LightRAG		MiniRAG	
	acc↑	*err*↓	*acc*↑	*err*↓	*acc*↑	*err*↓	*acc*↑	*err*↓
Phi-3.5-mini-instruct	42.72%	31.34%	/	/	27.03%	11.78%	**49.96%**	28.44%
GLM-Edge-1.5B-Chat	44.44%	24.26%	/	/	/	/	**51.41%**	23.44%
Qwen2.5-3B-Instruct	39.48%	31.69%	/	/	21.91%	13.73%	**48.55%**	33.10%
MiniCPM3-4B	39.24%	31.42%	/	/	19.48%	10.41%	**47.77%**	26.88%
gpt-4o-mini	53.60%	27.19%	60.92%	16.86%	64.91%	19.37%	**68.43%**	19.41%

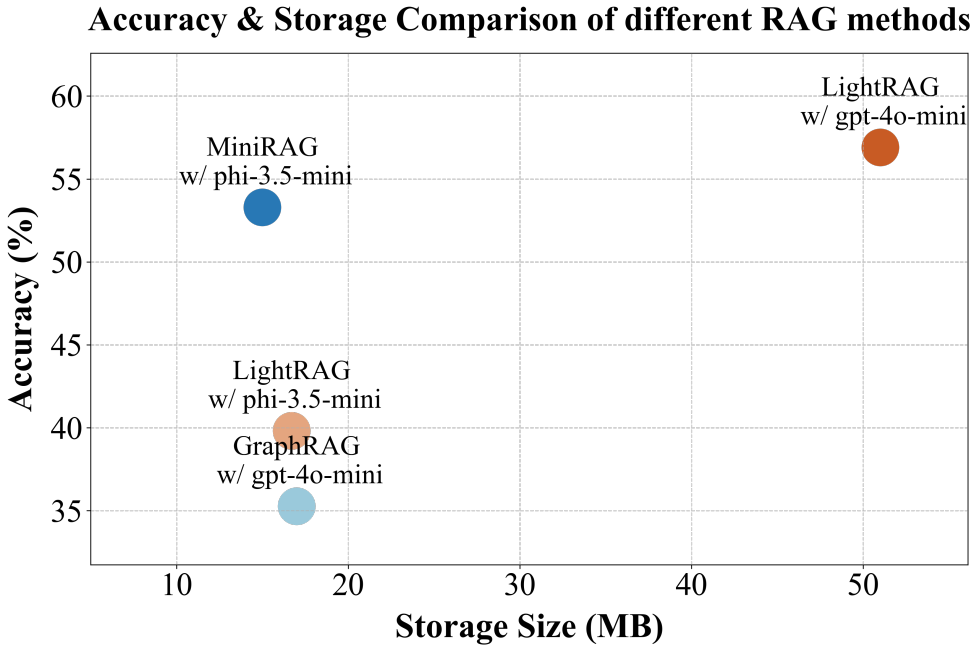
Experiment

RQ1: Comparative Performance. How does MiniRAG perform against state-of-the-art alternatives in terms of retrieval accuracy and efficiency?

Table 1: Performance evaluation using accuracy (acc) and error (err) rates, measured as percentages (%). Higher accuracy and lower error rates indicate better RAG performance. Results compare various baseline methods against our MiniRAG across multiple datasets. Bold values indicate best performance, while "/" denotes cases where methods failed to generate effective responses.

LiHuaWorld	NaiveRAG		GraphRAG		LightRAG		MiniRAG	
	acc↑	*err*↓	*acc*↑	*err*↓	*acc*↑	*err*↓	*acc*↑	*err*↓
Phi-3.5-mini-instruct	41.22%	23.20%	/	/	39.81%	25.39%	**53.29%**	23.35%
GLM-Edge-1.5B-Chat	42.79%	24.76%	/	/	35.74%	25.86%	**52.51%**	25.71%
Qwen2.5-3B-Instruct	43.73%	24.14%	/	/	39.18%	28.68%	**48.75%**	26.02%
MiniCPM3-4B	43.42%	17.08%	/	/	35.42%	21.94%	**51.25%**	21.79%
gpt-4o-mini	46.55%	19.12%	35.27%	37.77%	**56.90%**	20.85%	54.08%	19.44%

MultiHop-RAG	NaiveRAG		GraphRAG		LightRAG		MiniRAG	
	acc↑	*err*↓	*acc*↑	*err*↓	*acc*↑	*err*↓	*acc*↑	*err*↓
Phi-3.5-mini-instruct	42.72%	31.34%	/	/	27.03%	11.78%	**49.96%**	28.44%
GLM-Edge-1.5B-Chat	44.44%	24.26%	/	/	/	/	**51.41%**	23.44%
Qwen2.5-3B-Instruct	39.48%	31.69%	/	/	21.91%	13.73%	**48.55%**	33.10%
MiniCPM3-4B	39.24%	31.42%	/	/	19.48%	10.41%	**47.77%**	26.88%
gpt-4o-mini	53.60%	27.19%	60.92%	16.86%	64.91%	19.37%	**68.43%**	19.41%



Observation:

1. Current RAG systems **face critical challenges** when operating with SLMs.
2. MiniRAG **maintain strong performance** even with SLMs.
3. MiniRAG demonstrates exceptional **storage efficiency** while preserving high accuracy levels.

Experiment

RQ2: Component Analysis. What is the contribution of key components to MiniRAG's overall effectiveness?

Table 2: Ablation study results comparing accuracy (acc, ↑) and error rate (err, ↓) (%) across architectural variants: baseline MiniRAG versus variants with (i) semantic-driven indexing replacement (-$\mathcal{I}$), (ii) edge information removal (-$\mathcal{R}_{edge}$), and (iii) chunk nodes removal (-$\mathcal{R}_{chunk}$). Results validate SLM limitations and the effectiveness of query-guided reasoning path components.

LiHuaWorld	**MiniRAG**		-$\mathcal{I}$		-$\mathcal{R}_{chunk}$		-$\mathcal{R}_{edge}$	
	acc↑	err↓	acc↑	err↓	acc↑	err↓	acc↑	err↓
Phi-3.5-mini-instruct	53.29%	23.35%	26.02%	19.12%	48.90%	17.40%	50.47%	15.36%
GLM-Edge-1.5B-Chat	52.51%	25.71%	25.08%	31.50%	46.24%	16.77%	47.81%	20.53%
Qwen2.5-3B-Instruct	48.75%	26.02%	24.14%	15.67%	40.91%	16.14%	48.43%	18.65%
MiniCPM3-4B	51.25%	21.79%	26.18%	15.52%	46.39%	15.83%	48.59%	19.44%

Observation:

1. SLM Limitations. Replacing indexing with text semantic-driven techniques (-*I*) causes significant performance drop, **validating SLMs' constraints in semantic understanding**, impacting knowledge graph generation and text descriptions.

2. Effectiveness of Query-guided Reasoning Path Discovery. Removing edge info (-R_{edge}) or chunk nodes (-R_{chunk}) harms performance. These components enable query matching and **compensate for SLMs' limitations during indexing.**

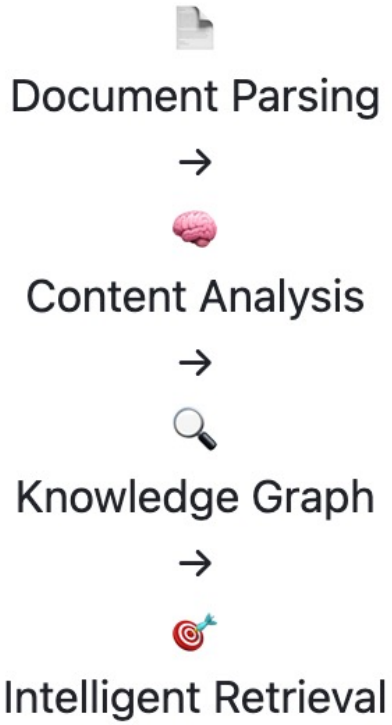
🚀 RAG-Anything: All-in-One RAG System

🎯 Key Features

- 🔄 **End-to-End Multimodal Pipeline** - Complete workflow from document ingestion and parsing to intelligent multimodal query answering
- 📄 **Universal Document Support** - Seamless processing of PDFs, Office documents, images, and diverse file formats
- 🧠 **Specialized Content Analysis** - Dedicated processors for images, tables, mathematical equations, and heterogeneous content types
- 🔗 **Multimodal Knowledge Graph** - Automatic entity extraction and cross-modal relationship discovery for enhanced understanding
- ⚡ **Adaptive Processing Modes** - Flexible MinerU-based parsing or direct multimodal content injection workflows
- 📋 **Direct Content List Insertion** - Bypass document parsing by directly inserting pre-parsed content lists from external sources
- 🎯 **Hybrid Intelligent Retrieval** - Advanced search capabilities spanning textual and multimodal content with contextual understanding

Core Algorithm

RAG-Anything implements an effective **multi-stage multimodal pipeline** that fundamentally extends traditional RAG architectures to seamlessly handle diverse content modalities through intelligent orchestration and cross-modal understanding.



1. Document Parsing Stage

The system provides high-fidelity document extraction through adaptive content decomposition. It intelligently segments heterogeneous elements while preserving contextual relationships. Universal format compatibility is achieved via specialized optimized parsers.

Key Components:

- ⚙️ **MinerU Integration**: Leverages MinerU for high-fidelity document structure extraction and semantic preservation across complex layouts.
- 🧩 **Adaptive Content Decomposition**: Automatically segments documents into coherent text blocks, visual elements, structured tables, mathematical equations, and specialized content types while preserving contextual relationships.
- 📂 **Universal Format Support**: Provides comprehensive handling of PDFs, Office documents (DOC/DOCX/PPT/PPTX/XLS/XLSX), images, and emerging formats through specialized parsers with format-specific optimization.

2. Multi-Modal Content Understanding & Processing

The system automatically categorizes and routes content through optimized channels. It uses concurrent pipelines for parallel text and multimodal processing. Document hierarchy and relationships are preserved during transformation.

Key Components:

- 🎯 **Autonomous Content Categorization and Routing**: Automatically identify, categorize, and route different content types through optimized execution channels.
- ⚡ **Concurrent Multi-Pipeline Architecture**: Implements concurrent execution of textual and multimodal content through dedicated processing pipelines. This approach maximizes throughput efficiency while preserving content integrity.
- 🏗️ **Document Hierarchy Extraction**: Extracts and preserves original document hierarchy and inter-element relationships during content transformation.

3. Multimodal Analysis Engine

The system deploys modality-aware processing units for heterogeneous data modalities:

Specialized Analyzers:

- 🔍 **Visual Content Analyzer**:
  - Integrate vision model for image analysis.
  - Generates context-aware descriptive captions based on visual semantics.
  - Extracts spatial relationships and hierarchical structures between visual elements.
- 📊 **Structured Data Interpreter**:
  - Performs systematic interpretation of tabular and structured data formats.
  - Implements statistical pattern recognition algorithms for data trend analysis.
  - Identifies semantic relationships and dependencies across multiple tabular datasets.
- 📐 **Mathematical Expression Parser**:
  - Parses complex mathematical expressions and formulas with high accuracy.
  - Provides native LaTeX format support for seamless integration with academic workflows.
  - Establishes conceptual mappings between mathematical equations and domain-specific knowledge bases.

4. Multimodal Knowledge Graph Index

The multi-modal knowledge graph construction module transforms document content into structured semantic representations. It extracts multimodal entities, establishes cross-modal relationships, and preserves hierarchical organization. The system applies weighted relevance scoring for optimized knowledge retrieval.

Core Functions:

- 🔍 **Multi-Modal Entity Extraction**: Transforms significant multimodal elements into structured knowledge graph entities. The process includes semantic annotations and metadata preservation.
- 🔗 **Cross-Modal Relationship Mapping**: Establishes semantic connections and dependencies between textual entities and multimodal components. This is achieved through automated relationship inference algorithms.
- 🏗️ **Hierarchical Structure Preservation**: Maintains original document organization through "belongs_to" relationship chains. These chains preserve logical content hierarchy and sectional dependencies.
- ⚖️ **Weighted Relationship Scoring**: Assigns quantitative relevance scores to relationship types. Scoring is based on semantic proximity and contextual significance within the document structure.

5. Modality-Aware Retrieval

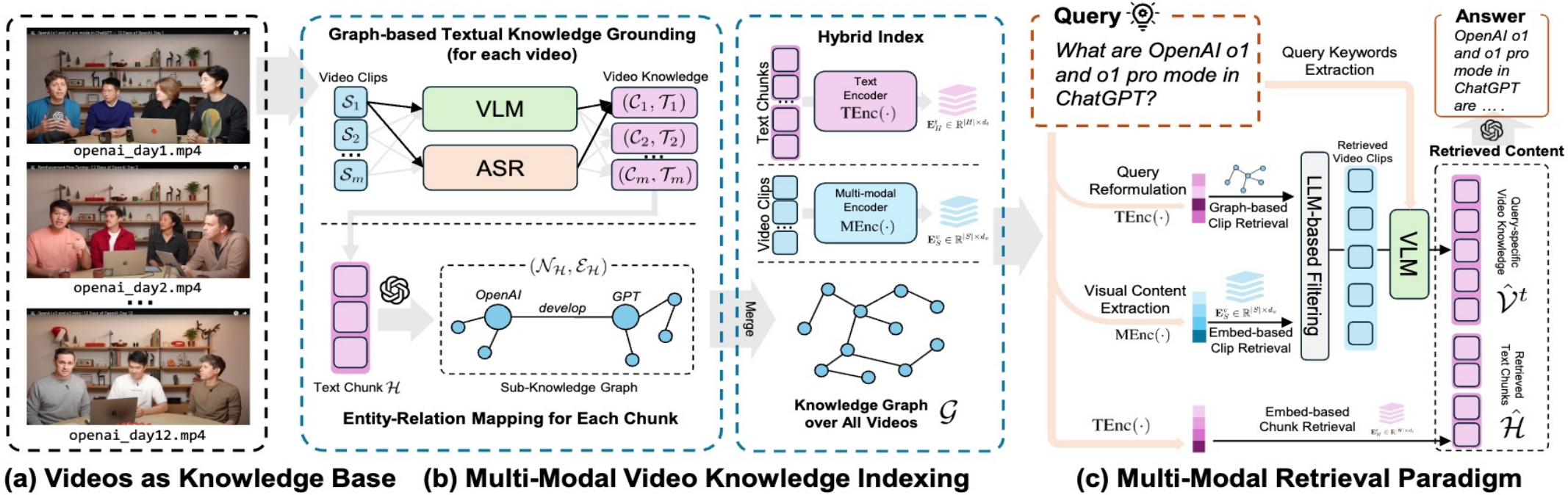
The hybrid retrieval system combines vector similarity search with graph traversal algorithms for comprehensive content retrieval. It implements modality-aware ranking mechanisms and maintains relational coherence between retrieved elements to ensure contextually integrated information delivery.

Retrieval Mechanisms:

- 🔀 **Vector-Graph Fusion**: Integrates vector similarity search with graph traversal algorithms. This approach leverages both semantic embeddings and structural relationships for comprehensive content retrieval.
- 📊 **Modality-Aware Ranking**: Implements adaptive scoring mechanisms that weight retrieval results based on content type relevance. The system adjusts rankings according to query-specific modality preferences.
- 🔗 **Relational Coherence Maintenance**: Maintains semantic and structural relationships between retrieved elements. This ensures coherent information delivery and contextual integrity.

VideoRAG

Video Type	#video list	#video	#query	#avg. queries per list	#overall duration
Lecture	12	135	376	31.3	~ 64.3 hours
Documentary	5	12	114	22.8	~ 28.5 hours
Entertainment	5	17	112	22.4	~ 41.9 hours
All	22	164	602	27.4	~ 134.6 hours



Jerry from Uganda: An Open Learner's Story

Lotfullah from Afghani An Open Learner's Stor

OUR OWN PDF gChain

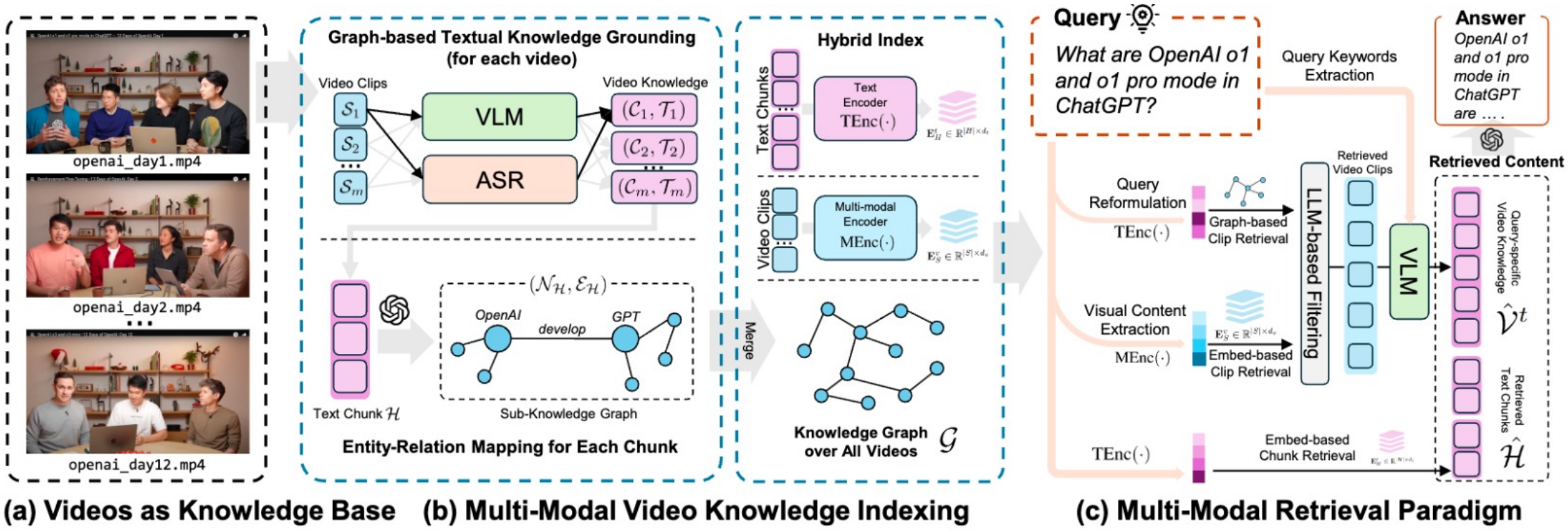
Langchair (GUI) | Ch - 2 | urse...

Cash flow for Beg Equity Research |

Videos on the Internet contain a wealth of knowledge.

Can AI watch all the videos and learn from them?

We present VideoRAG



An unlimited AI framework for watching long-context videos on a single RTX 3090.

What is Retrieval-Augmented Generation (RAG)?

Knowledge Data Base (e.g., Documents, Videos) $\mathcal{D}$

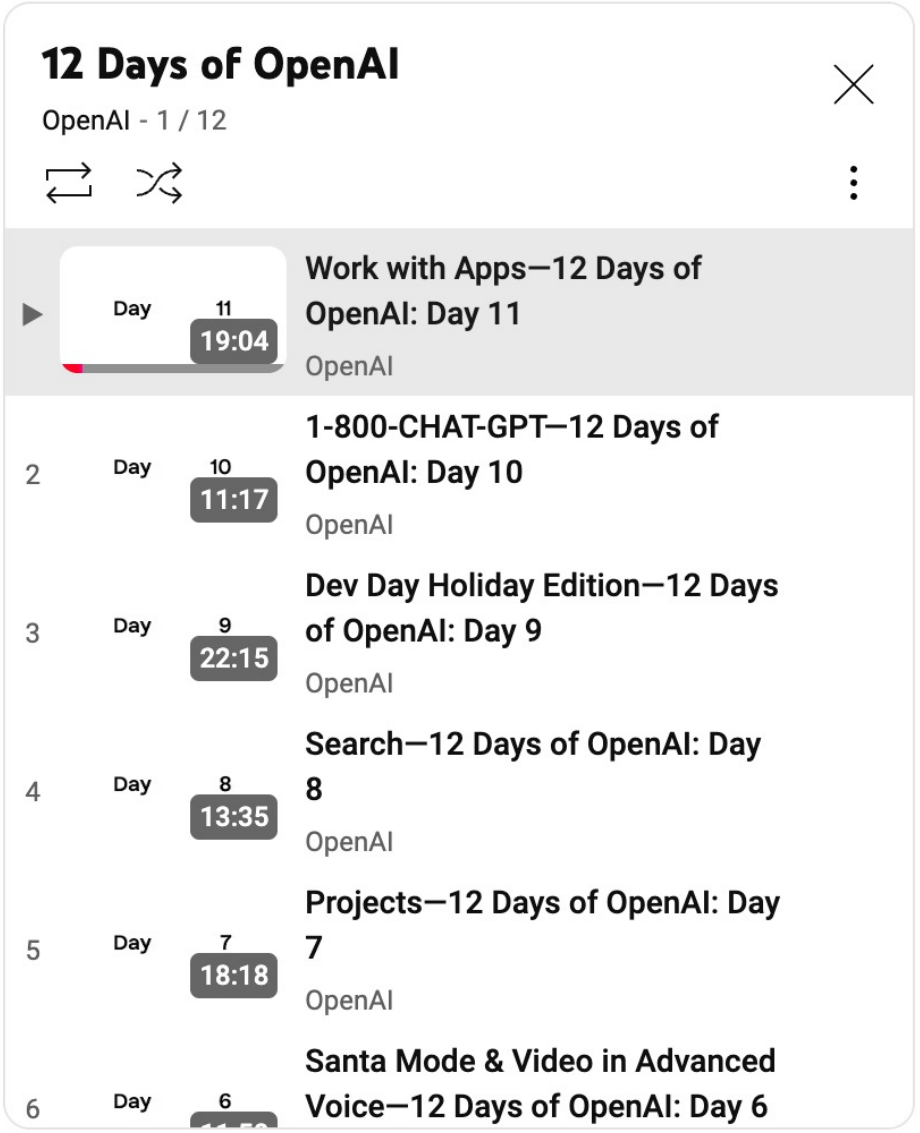
Indexing Process $\hat{\mathcal{D}} = \varphi(\mathcal{D})$

Retrieval Knowledge for Input Query $\psi(q, \hat{\mathcal{D}})$

Answer Generation $\mathbf{LLM}(q, \psi(q, \hat{\mathcal{D}}))$

VideoRAG Knowledge Base

- We are considering leveraging **video list** as the knowledge base for the VideoRAG framework, such as OpenAI's 12 Days Show.
- There are no limitations on the duration of each video or the number of videos.



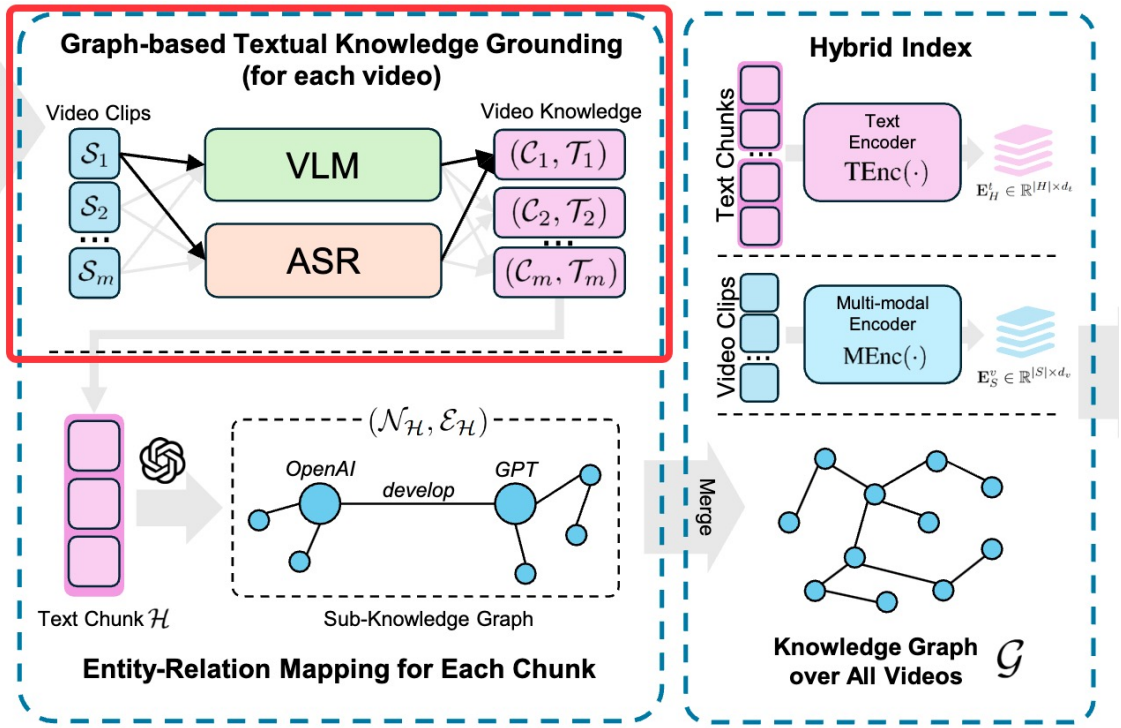
VideoRAG
Multi-Modal Video Knowledge Indexing

- For each video, we split it into sub-clips and use VLMs and ASR models to extract both visual and audio information.

$$\mathcal{C}_j = \mathbf{VLM}(\mathcal{T}_j, \{\mathbf{F}_1, \ldots, \mathbf{F}_k\} \mid \mathbf{F} \in \mathcal{S}_j)$$

$$\mathcal{T}_j = \mathbf{ASR}(\mathcal{S}_j)$$

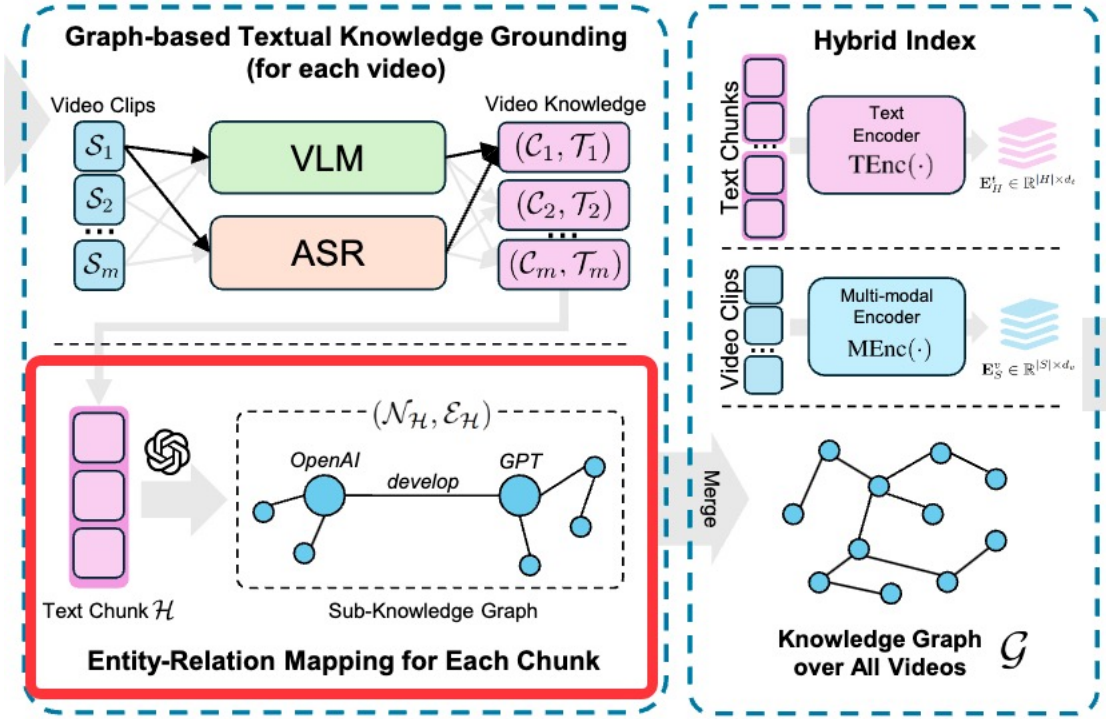
$$\mathcal{V}^t = \{(\mathcal{C}_l, \mathcal{T}_l) \mid l \in [1, m]\}$$



VideoRAG
Multi-Modal Video Knowledge Indexing

- Next, we construct a knowledge graph from the extracted textual knowledge of the sub clips to link information across multiple videos.

$$\mathcal{G} = (\mathcal{N}, \mathcal{E}) = \bigcup_{\mathcal{H} \in \{\mathcal{V}_1^t, \ldots, \mathcal{V}_n^t\}} (\mathcal{N}_\mathcal{H}, \mathcal{E}_\mathcal{H}),$$



VideoRAG
Multi-Modal Video Knowledge Indexing

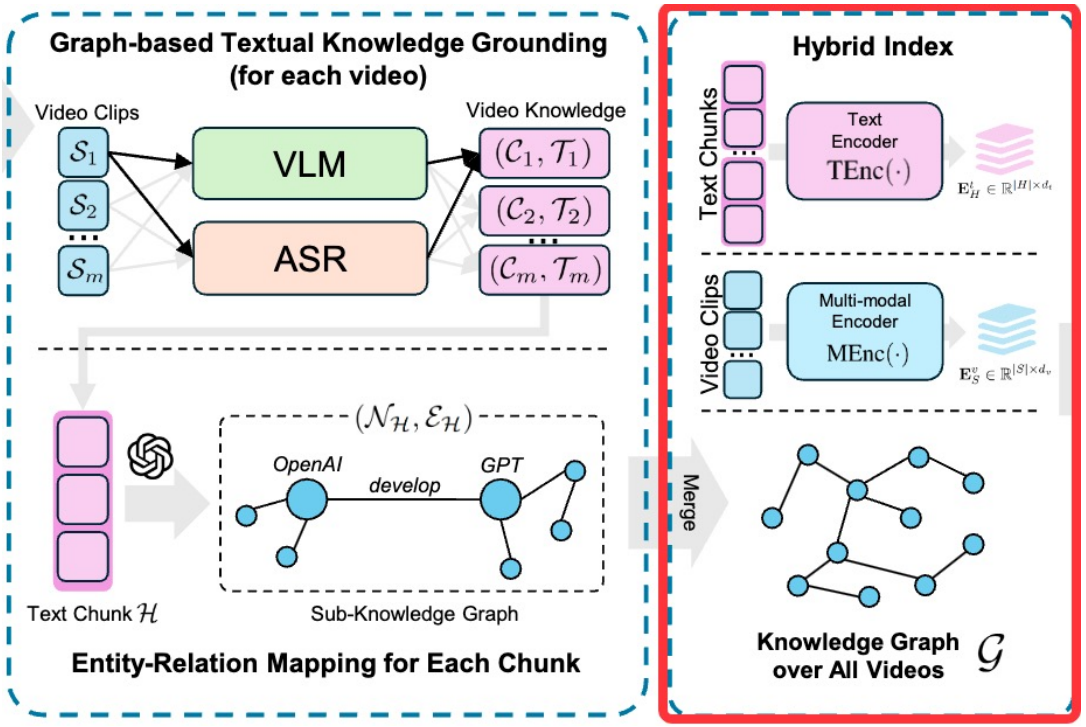
- Meanwhile, We encoded multi-modal embeddings for textual chunks and video clips separately with multi-modal embedders.

$$\mathbf{E}_H^t \in \mathbb{R}^{|H| \times d_t}$$

$$\mathbf{E}_S^v \in \mathbb{R}^{|S| \times d_v} \quad w.r.t. \quad \mathbf{e}_S^v = \mathbf{MEnc}(S).$$

- The final indexing results in a hybrid system that incorporates both multi-modal embeddings and knowledge graphs (KGs).

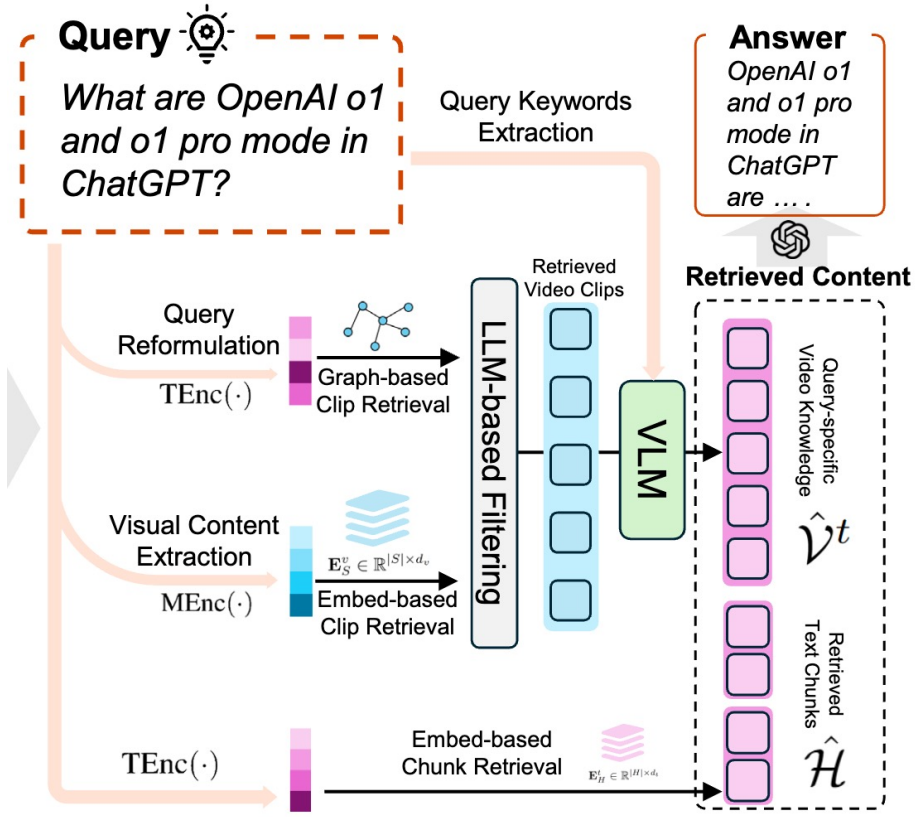
$$\hat{\mathcal{D}} = \varphi(\mathcal{D}) = (\mathcal{G}, \mathbf{E}_H^t, \mathbf{E}_S^v).$$



VideoRAG
Multi-Modal Retrieval Paradigm

- We perform video clip-level retrieval using both textual semantic matching and visual retrieval techniques.
- For the retrieved video clips, we then filter them by assessing their relevance to the input query.
- Finally, the VLM is employed once more to extract fine-grained, query-specific knowledge from the clips.

$$\hat{\mathcal{C}} = \mathbf{VLM}(\mathcal{K}_q, \hat{\mathcal{T}}, \{\mathbf{F}_1, \ldots, \mathbf{F}_{\hat{k}}\} \mid \mathbf{F} \in \hat{\mathcal{S}})$$

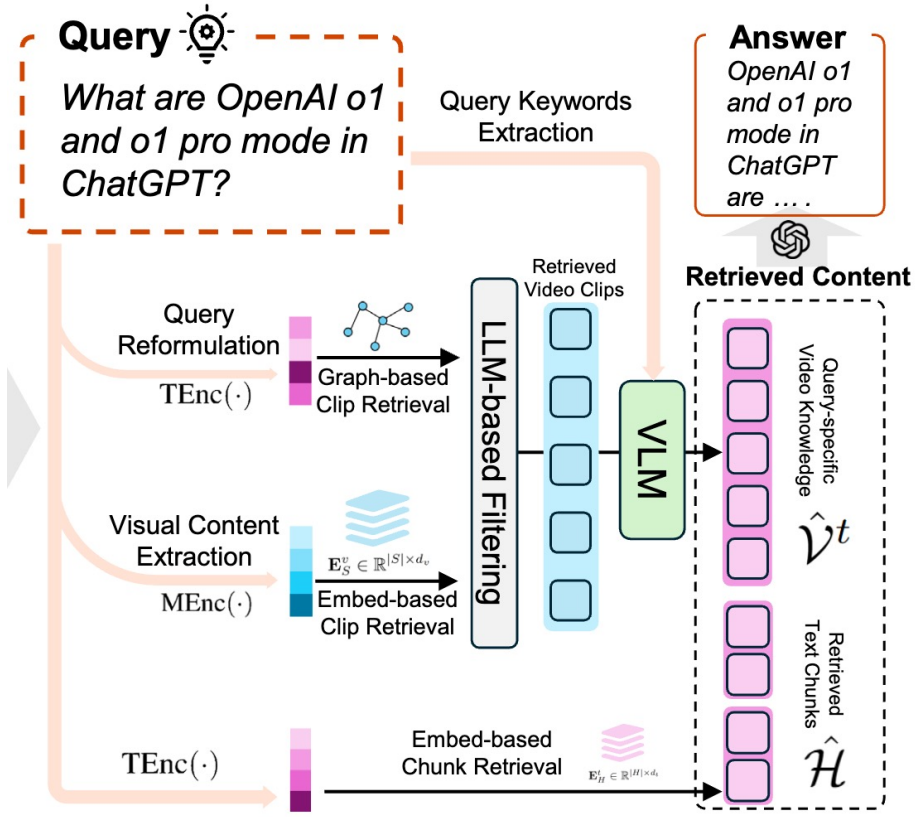


VideoRAG
Response Generation

- After retrieving the query-specific knowledge from the videos, we concatenate it with the extracted query-relevant text chunks for the LLMs to generate responses.

$$\psi(q, \hat{\mathcal{D}}) = (\hat{\mathcal{V}}^t, \hat{\mathcal{H}})$$

$$\mathbf{LLM}(q, \psi(q, \hat{\mathcal{D}}))$$



Benchmark (LongerVideos)

We constructed the *LongerVideos* benchmark to evaluate the model's performance in comprehending multiple long-context videos and answering open-ended queries.

Table 1: Statistics of the experimental dataset *LongerVideos*.

Video Type	#video list	#video	#query	#avg. queries per list	#overall duration
Lecture	12	135	376	31.3	~ 64.3 hours
Documentary	5	12	114	22.8	~ 28.5 hours
Entertainment	5	17	112	22.4	~ 41.9 hours
All	22	164	602	27.4	~ 134.6 hours

Comparison with RAG baselines (Win-Rate)

Table 2: We analyze the performance of VideoRAG against RAG baselines on the LongerVideos dataset, presenting results both by individual video categories and across the complete dataset.

	Lecture		**Documentary**		**Entertainment**		**All**	
	NaiveRAG	**VideoRAG**	NaiveRAG	**VideoRAG**	NaiveRAG	**VideoRAG**	NaiveRAG	**VideoRAG**
Comprehensiveness	47.63%	52.37%	44.04%	55.96%	46.43%	53.57%	46.73%	53.27%
Empowerment	45.85%	54.15%	40.00%	60.00%	45.36%	54.64%	44.65%	55.35%
Trustworthiness	46.73%	53.27%	42.54%	57.46%	44.46%	55.54%	45.51%	54.49%
Depth	46.70%	53.30%	43.25%	56.75%	46.07%	53.93%	45.93%	54.07%
Density	46.73%	53.27%	44.21%	55.79%	44.29%	55.71%	45.80%	54.20%
Overall Winner	47.66%	52.34%	44.04%	55.96%	46.43%	53.57%	46.74%	53.26%
	GraphRAG-*l*	**VideoRAG**	GraphRAG-*l*	**VideoRAG**	GraphRAG-*l*	**VideoRAG**	GraphRAG-*l*	**VideoRAG**
Comprehensiveness	44.60%	55.40%	48.68%	51.32%	49.29%	50.71%	46.25%	53.75%
Empowerment	42.34%	57.66%	47.54%	52.46%	49.02%	50.98%	44.57%	55.43%
Trustworthiness	42.79%	57.21%	47.11%	52.89%	46.07%	53.93%	44.22%	55.78%
Depth	42.34%	57.66%	48.33%	51.67%	49.55%	50.45%	44.82%	55.18%
Density	39.26%	60.74%	45.26%	54.74%	46.52%	53.48%	41.74%	58.26%
Overall Winner	44.44%	55.56%	48.68%	51.32%	49.20%	50.80%	46.13%	53.87%
	GraphRAG-*g*	**VideoRAG**	GraphRAG-*g*	**VideoRAG**	GraphRAG-*g*	**VideoRAG**	GraphRAG-*g*	**VideoRAG**
Comprehensiveness	42.66%	57.34%	46.23%	53.77%	48.48%	51.52%	44.42%	55.58%
Empowerment	39.55%	60.45%	44.04%	55.96%	48.30%	51.70%	42.03%	57.97%
Trustworthiness	38.54%	61.46%	41.49%	58.51%	43.48%	56.52%	40.02%	59.98%
Depth	40.61%	59.39%	45.26%	54.74%	47.23%	52.77%	42.72%	57.28%
Density	37.55%	62.45%	46.93%	53.07%	48.04%	51.96%	41.28%	58.72%
Overall Winner	42.23%	57.77%	46.32%	53.68%	48.75%	51.25%	44.22%	55.78%
	LightRAG	**VideoRAG**	LightRAG	**VideoRAG**	LightRAG	**VideoRAG**	LightRAG	**VideoRAG**
Comprehensiveness	42.42%	57.58%	45.09%	54.91%	43.84%	56.16%	43.19%	56.81%
Empowerment	39.55%	60.45%	38.95%	61.05%	42.05%	57.95%	39.90%	60.10%
Trustworthiness	39.52%	60.48%	42.11%	57.89%	40.00%	60.00%	40.10%	59.90%
Depth	40.13%	59.87%	41.93%	58.07%	41.96%	58.04%	40.81%	59.19%
Density	39.57%	60.43%	42.37%	57.63%	41.61%	58.39%	40.48%	59.52%
Overall Winner	42.15%	57.85%	44.30%	55.70%	43.75%	56.25%	42.86%	57.14%

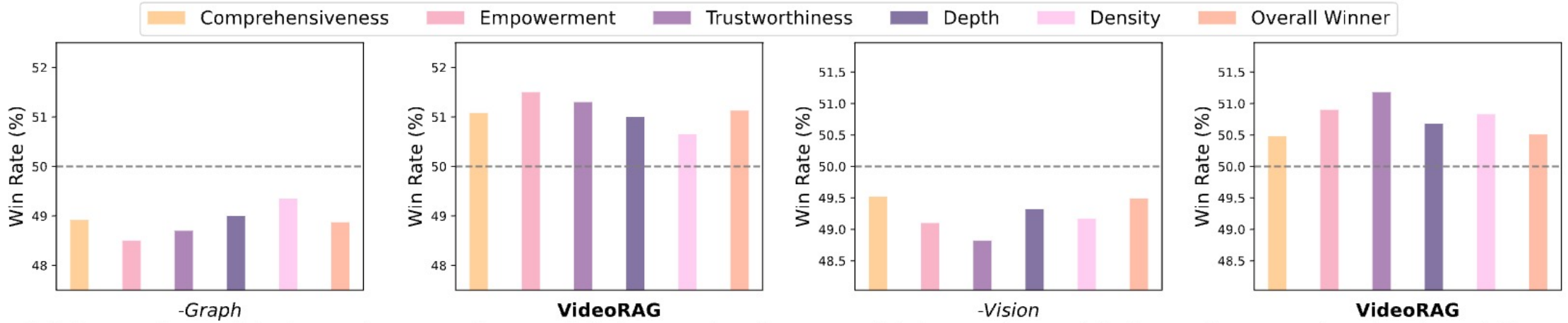
Comparison with Video Understanding baselines (Quantitative Comparison)

Quantitative Long-context Video Understanding Performance

Metric	LLaMA-VID				VideoAgent				NotebookLM				VideoRAG Agentic thinking			
	LEC	DOC	ENT	ALL	LEC	DOC	ENT	ALL	LEC	DOC	ENT	ALL	LEC	DOC	ENT	ALL
Comprehensiveness	2.36	2.62	2.54	**2.44**	2.02	1.99	1.80	**1.98**	3.53	3.20	2.96	**3.36**	4.48	4.51	4.44	4.48
Empowerment	2.79	3.03	2.86	**2.85**	2.42	2.37	2.10	**2.35**	3.88	3.62	3.29	**3.72**	4.51	4.55	4.45	4.51
Trustworthiness	3.15	3.30	3.35	**3.22**	2.83	2.73	2.65	**2.78**	3.95	3.80	3.71	**3.88**	4.50	4.54	4.48	4.50
Depth	2.01	2.06	2.00	**2.02**	1.79	1.75	1.62	**1.75**	3.14	2.89	2.55	**2.98**	4.34	4.42	4.31	4.35
Density	3.15	3.28	3.21	**3.18**	2.82	2.73	2.52	**2.75**	4.07	3.82	3.61	**3.94**	4.59	4.63	4.56	4.59
Overall Score	2.36	2.61	2.54	**2.44**	2.03	2.01	1.80	**1.98**	3.54	3.21	2.97	**3.37**	4.45	4.49	4.41	4.45

Ablation Study

Both variants that remove graph indexing (-Graph) or visual encoding (-Vision) result in decreased performance.

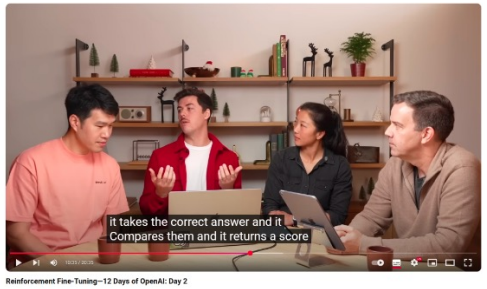


(a) Comparison with the variant w/o the graph index-retrieval. **(b) Comparison with the variant w/o the visual modality.**

Figure 2: Ablation on graph-based knowledge grounding and cross-modal retrieval components.

Case Study

VideoRAG can accurately retrieve four continuous video segments from OpenAI's 12-day show, specifically from 10:00 to 12:00 on Day 2 of the video titled "Reinforcement Fine-Tuning."



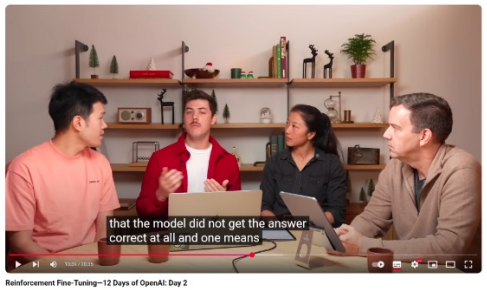


Table 4: A case study examining VideoRAG's response to a query about "*graders*" in OpenAI's 12-day edu cational series demonstrates its video retrieval capabilities and knowledge synthesis.

Query: Explain the purpose and functionality of 'graders' in the context of reinforcement fine-tuning.

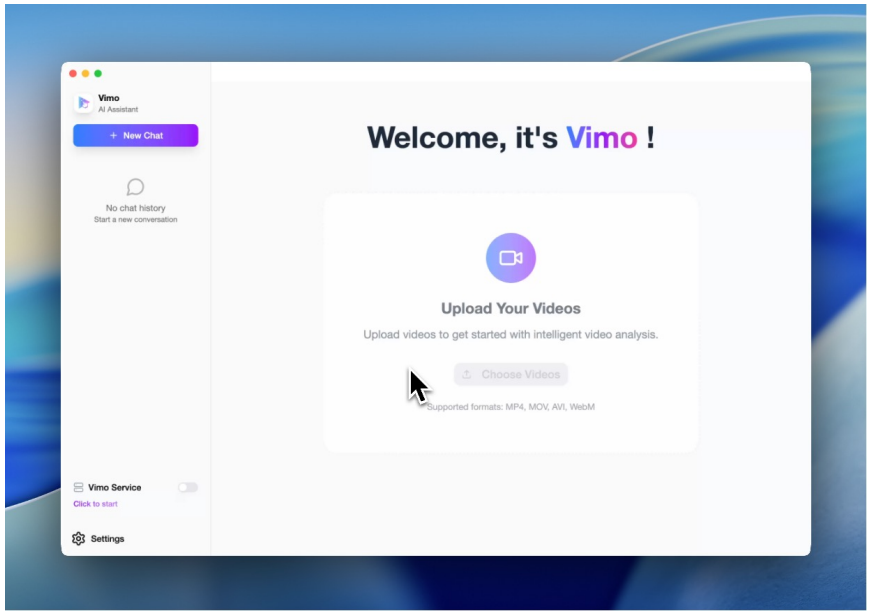
Input Videos: 12 Days of OpenAI (A total of 12 videos)

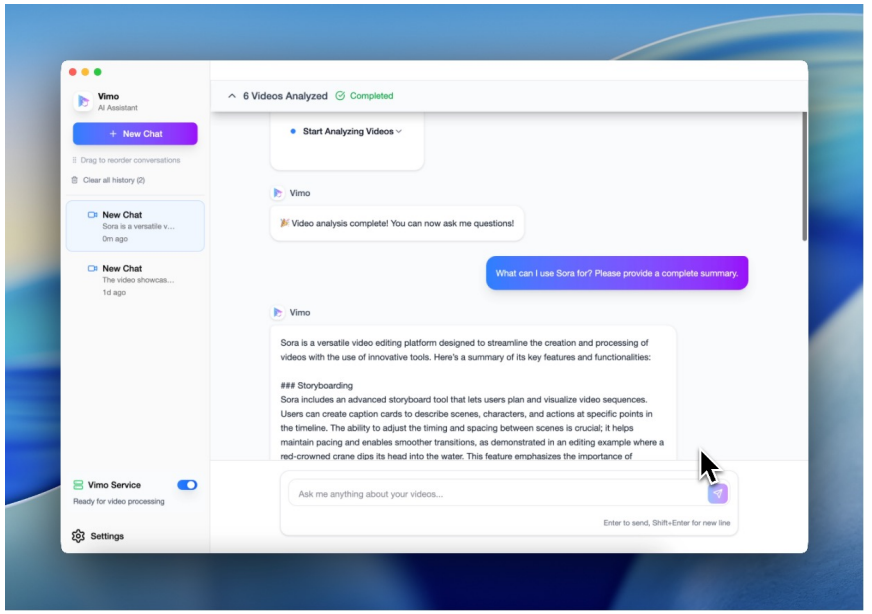
VideoRAG Response:

```
### Understanding Gradation in Reinforcement Fine-Tuning
**Purpose of Graders**
Graders are integral components in the process of reinforcement fine-tuning,
mainly utilized to evaluate and improve the performance of machine learning
models... (Omitted for brevity)
**Functionality of Graders**
The grading system operates by comparing a model's outputs against those of
verified correct responses. The mechanics of this assessment involve assigning
scores to the model's outputs based on their correctness. These scores range
from zero to one, where:
- A score of **0** indicates that the model's answer is completely incorrect.
- A score of **1** signals that the answer is fully correct.
- Scores can also be fractional, allowing for partial credit. For
instance, a score of **0.7** may denote that the model's suggestion was
correct but not the highest in a ranked list of possible answers.
```

Vimo

We deliver an open-source chat platform, **Vimo**, that can understand and analyze videos of unlimited length and answer your questions with remarkable accuracy. Now everyone can **chat with videos**.





Vimo

Vimo: Chat with Your Videos

✨Data Intelligence Lab@HKU✨

Chao Huang

Data Intelligence Lab

chaohuang75@gmail.com

https://github.com/HKUDS