

# STaRK: Benchmarking LLM Retrieval on Textual and Relational Knowledge Bases

**Shirley Wu<sup>\*§</sup>, Shiyu Zhao<sup>\*§</sup>, Michihiro Yasunaga<sup>§</sup>, Kexin Huang<sup>§</sup>, Kaidi Cao<sup>§</sup>, Qian Huang<sup>§</sup>,  
Vassilis N. Ioannidis<sup>†</sup>, Karthik Subbian<sup>†</sup>, James Zou<sup>‡§</sup>, Jure Leskovec<sup>‡§</sup>**

<sup>§</sup>Department of Computer Science, Stanford University   <sup>†</sup>Amazon

Joohyung Yun  
2024.5.21

# Overview of This Seminar

- Introduction
- Background
- Problem Definition
- Methodology of Benchmark Generation
  - Generated Semi-structured Knowledge Bases
  - Query-Ground Truth Entity Set Generation
- Analysis of Benchmark
- Experiments

# Introduction

# Semi-Structure Retrieval Benchmark

**Description:** Generate a benchmark for the task of *retrieving nodes via natural language* on semi-structured knowledge base.

**Example use case:** Information retrieval on e-commerce web search

- People express complex concepts that need both structured and unstructured information.

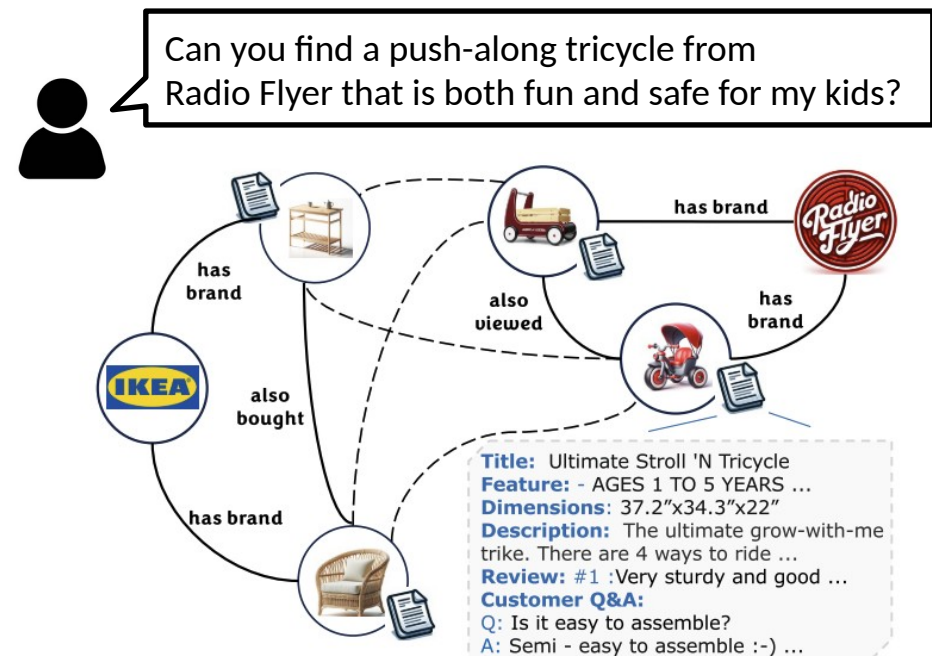


Figure 1. An example case of semi-structure retrieval

# Motivation of Research

## Problems of previous works

- Focused on either purely textual queries on unstructured knowledge or structured SQL or knowledge graph queries.
- Multi-hop reasoning benchmarks are either limited in the span of knowledge or lack in diverse textual properties and explicit relational information.
- LLM retrieval techniques remain an open question of how to effectively apply them to the specific challenge of retrieval from semi-structured knowledge bases.

# Contributions

- **Natural-sounding queries on semi-structured knowledge:** the queries are crafted to incorporate rich relational information and complex textual properties.
- **Context-specific reasoning:** The queries entail reasoning capabilities specific to the context.
- **Diverse domains:** wide range of content and domains, reflecting the diverse nature of real-world information needs.

# Background

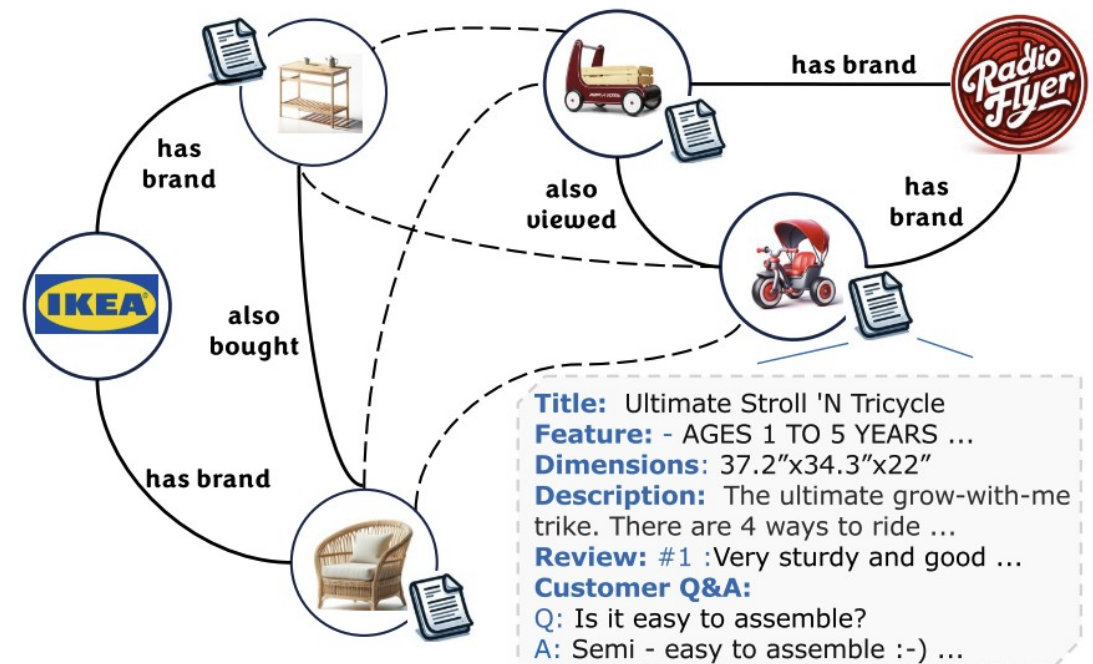
# Background

# Semi-Structured Knowledge Base (SKB)

A semi-structured knowledge base corresponds to a *labeled property graph*.

| Component in labeled property graph | Component in semi-structured knowledge base |
|-------------------------------------|---------------------------------------------|
| Node                                | Entity                                      |
| Node label                          | Entity type                                 |
| Node properties and values          | Textual information                         |
| Edge                                | Relation                                    |
| Edge label                          | Relation type                               |

**Table 1.** Relationship between the components of a labeled property graph and the components in SKB



**Amazon Semi-structured Knowledge Base**

**Figure 2.** An example SKB

# Background

## Semi-Structured Knowledge Bases (SKB)

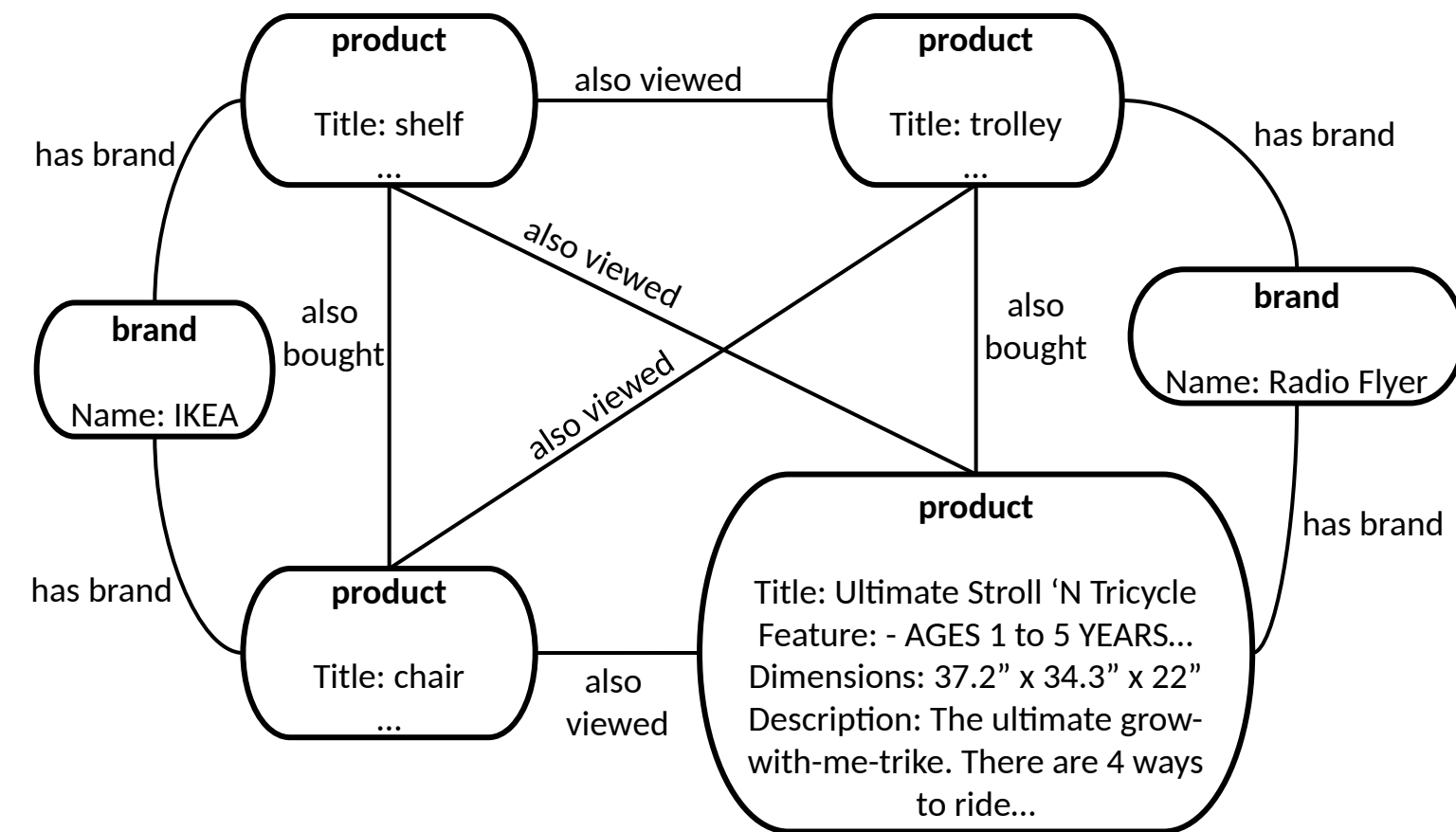
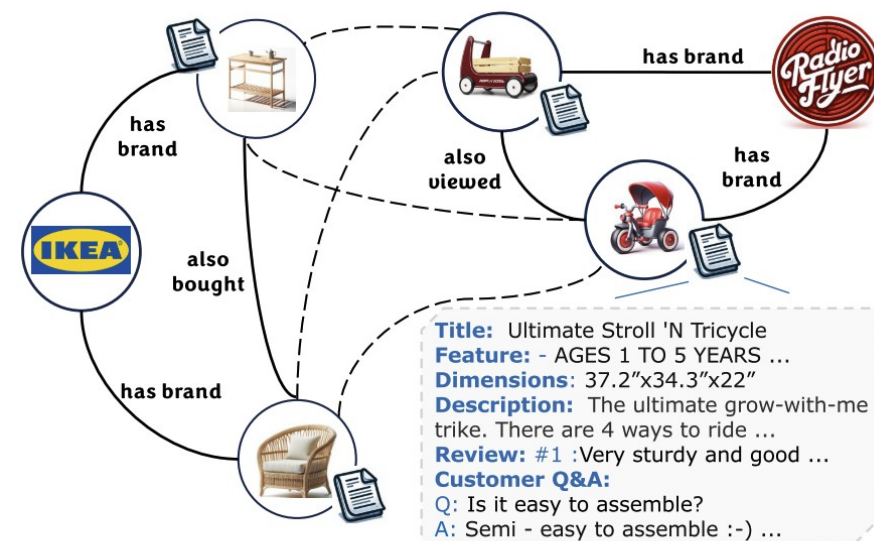


Figure 3. Representation of SKB in Figure 2. in a labeled knowledge graph



Amazon Semi-structured Knowledge Base

Figure 2. An example SKB

# Problem Definition

# Problem Definition

## Retrieval Task on SKB

**Description:** Retrieve node entities over an SKB given a natural language query.

- Input: natural language query
- Input: SKB
- Output: a set of nodes within the SKB

|              | Example query                                                                                                                       | Title of ground truth node(s)                                                         |
|--------------|-------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
| STARK-AMAZON | <i>Looking for durable Dart World brand dart flights that resist easy tearing. Any recommendations?</i>                             | <Amazon Standard Flights><br><Dart World Broken Glass Flight> (12 more)               |
|              | <i>What are recommended scuba diving weights for experienced divers that would fit well with my Gorilla PRO XL waterproof bag?</i>  | <Sea Pearls Vinyl Coated Lace Thru Weight>                                            |
| STARK-MAG    | <i>Search publications by Hao-Sheng Zeng on non-Markovian dynamics.</i>                                                             | <Distribution of non-Markovian intervals...><br><Comparison between non-Markovian...> |
|              | <i>What are some nanofluid heat transfer research papers published by scholars from Philadelphia University?</i>                    | <A Numerical Study on Convection Around A Square Cylinder using AL2O3-H2O Nanofluid>  |
| STARK-PRIME  | <i>Could you provide a list of investigational drugs that interact with genes or proteins active in the epididymal region?</i>      | <(S)-3-phenyllactic Acid>,<br><Anisomycin>, <Puromycin>                               |
|              | <i>Search for diseases without known treatments and induce pruritus in pregnant women, potentially associated with Autoimmune.</i>  | <Intrahepatic Cholestasis>                                                            |
|              | <i>Please find pathways involving the POLR3D gene within nucleoplasm.</i>                                                           | <RNA Polymerase III Chain Elongation>                                                 |
|              | <i>Which gene or protein associated with lichen amyloidosis can bind interleukin-31 to activate the PI3K/AKT and MAPK pathways?</i> | <OSMR>, <IL31RA>                                                                      |

**Table 2.** Example pairs of query and answer node set for SKB retrieval task

## Problem Definition

# Benchmark for Retrieval Task on SKB

A benchmark for retrieval task on SKB is comprised as

- A semi-structured knowledge base
- A set of (query, ground-truth entity set) pairs

# Benchmark Generation

# Generated SKBs

- STaRK has generated 3 SKBs
- The statistics for each SKB are as follows:

|              | #entity<br>types | #relation<br>types | avg.<br>degree | #entities | #relations | #tokens     |
|--------------|------------------|--------------------|----------------|-----------|------------|-------------|
| STARK-AMAZON | 2                | 3                  | 3.0            | 1,032,407 | 3,886,603  | 592,067,882 |
| STARK-MAG    | 4                | 4                  | 10.6           | 1,872,968 | 19,919,698 | 212,602,571 |
| STARK-PRIME  | 10               | 18                 | 62.6           | 129,375   | 8,100,498  | 31,844,769  |

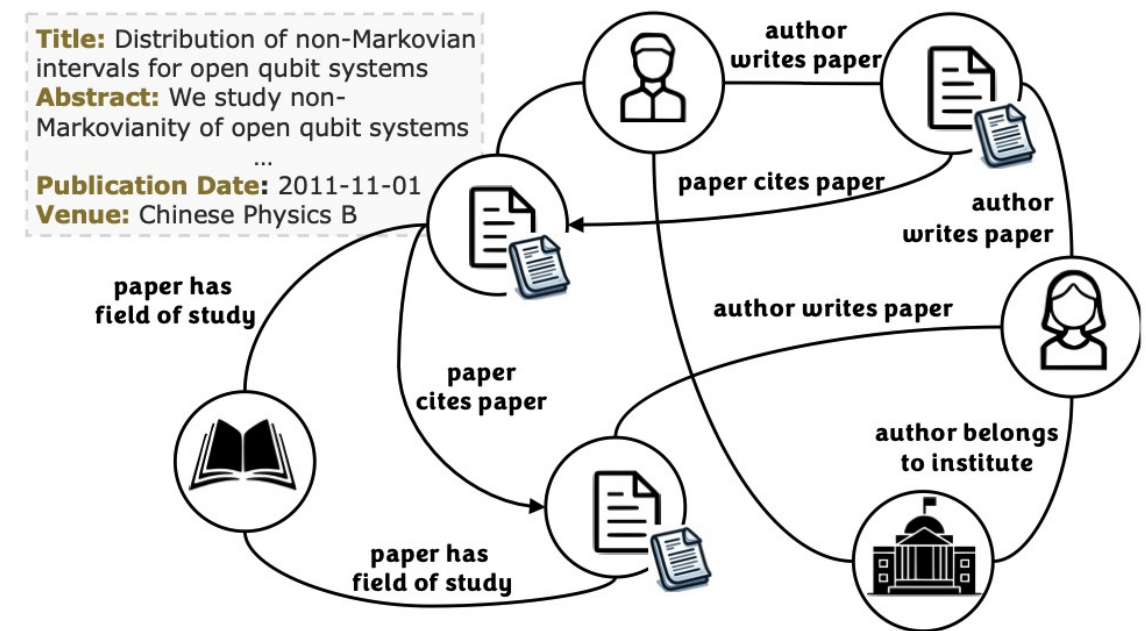
**Table 3.** Statistics for the 3 semi-structured knowledge bases of STaRK

# Generated SKB : Example

## MAG Semi-structured Knowledge Base

### Generation steps:

1. Select papers shared between *obgn-MAG* and *obgn-papers100M*.
2. Extract subgraph of *obgn-MAG* based on the selected papers to earn relational structure.
3. Extract textual information such as titles and abstracts for papers from *obgn-papers100M*.
4. Augment the textual information by integrating details (e.g., venue, author and institution names) from the *Microsoft Academic Graph database*.



**MAG Semi-structured Knowledge Base**

Figure 5. The MAG semi-structured knowledge base

# Query-Ground Truth Entity Set Generation

## Overall Idea

The process of generating a (query, ground truth entity set) pair from a SKB can be interpreted as

- 1. Query generation:** Picking a subgraph of SKB
  1. Determine the nodes, node labels, relations types of a subgraph.
  2. Determine the properties of a subgraph.
- 2. Ground truth entity set generation:** Filtering entities that match the subgraph of (1)

# Query-Ground Truth Entity Set Generation Overview

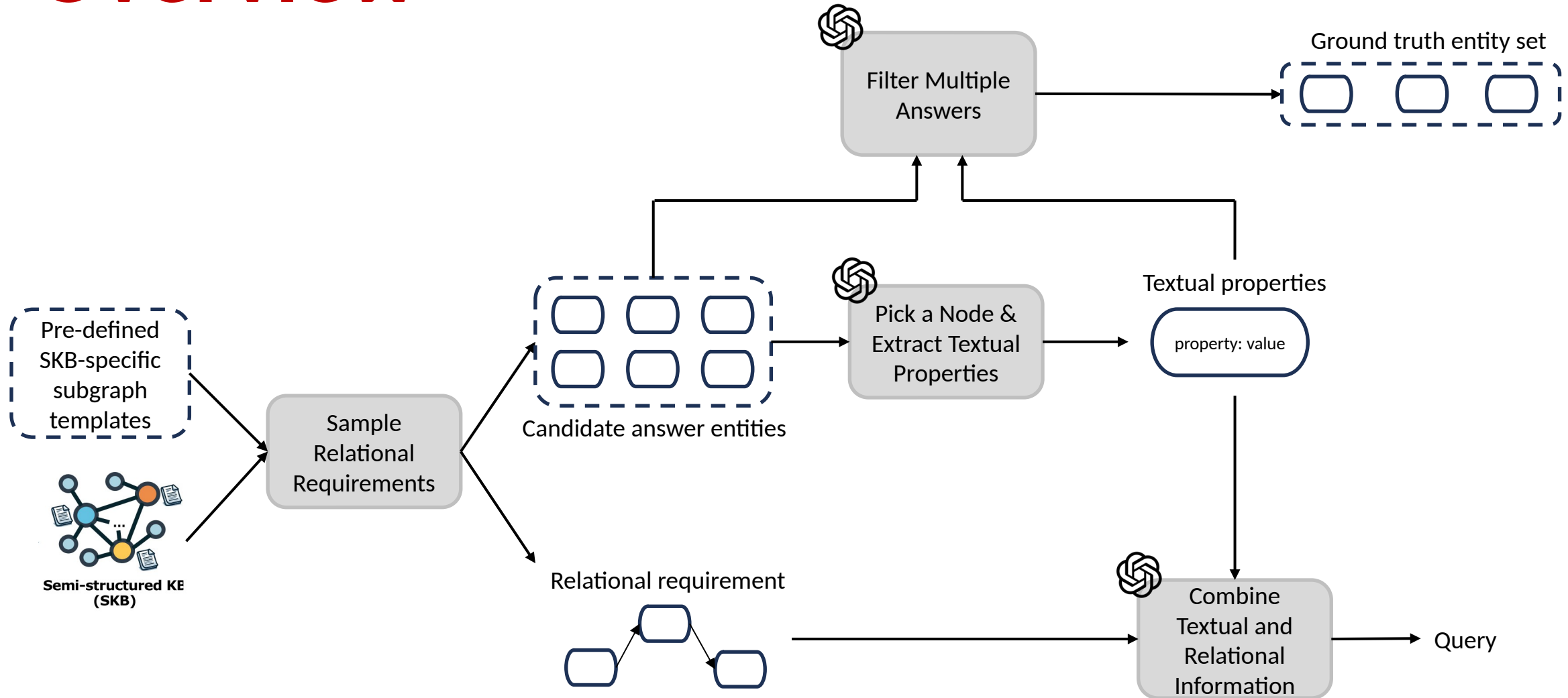
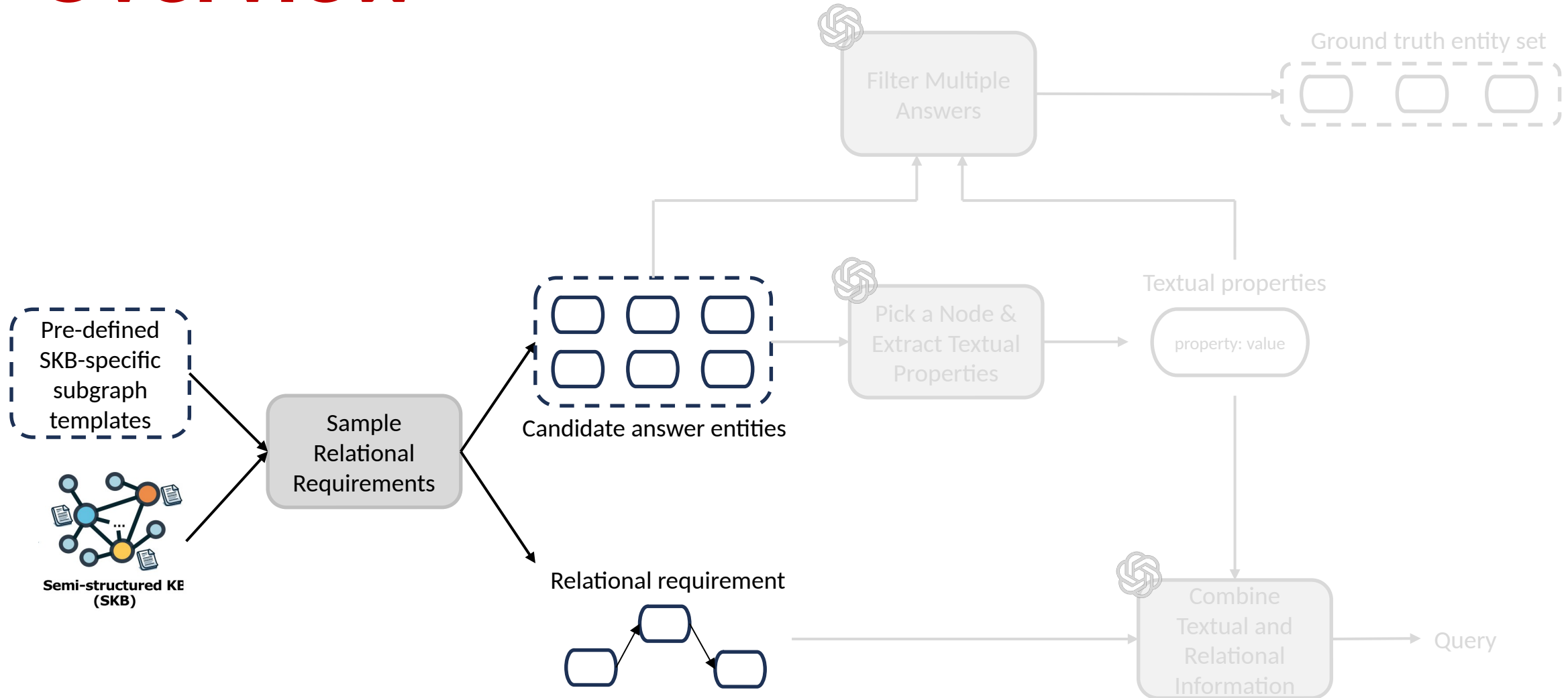
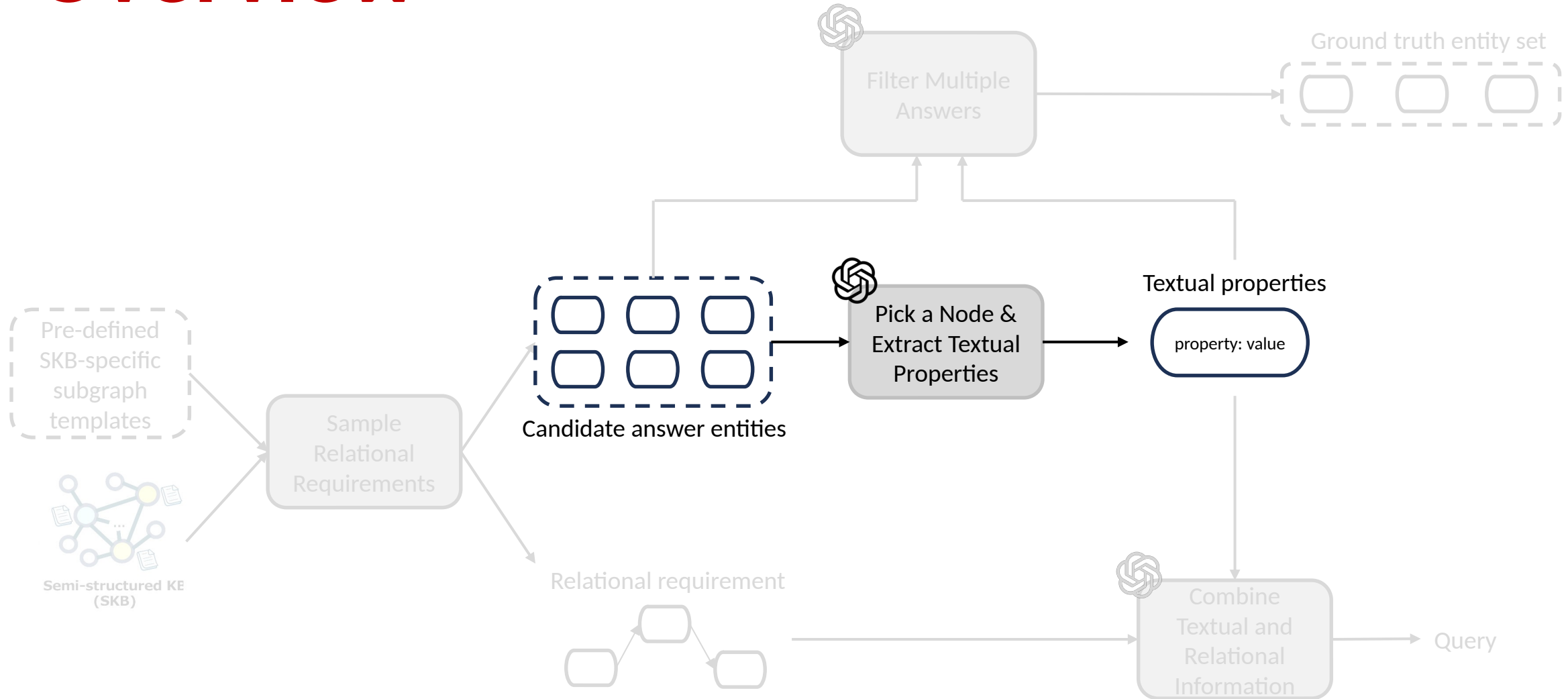


**Figure 4.** Overall flow of a single query - ground truth entity set pair generation.

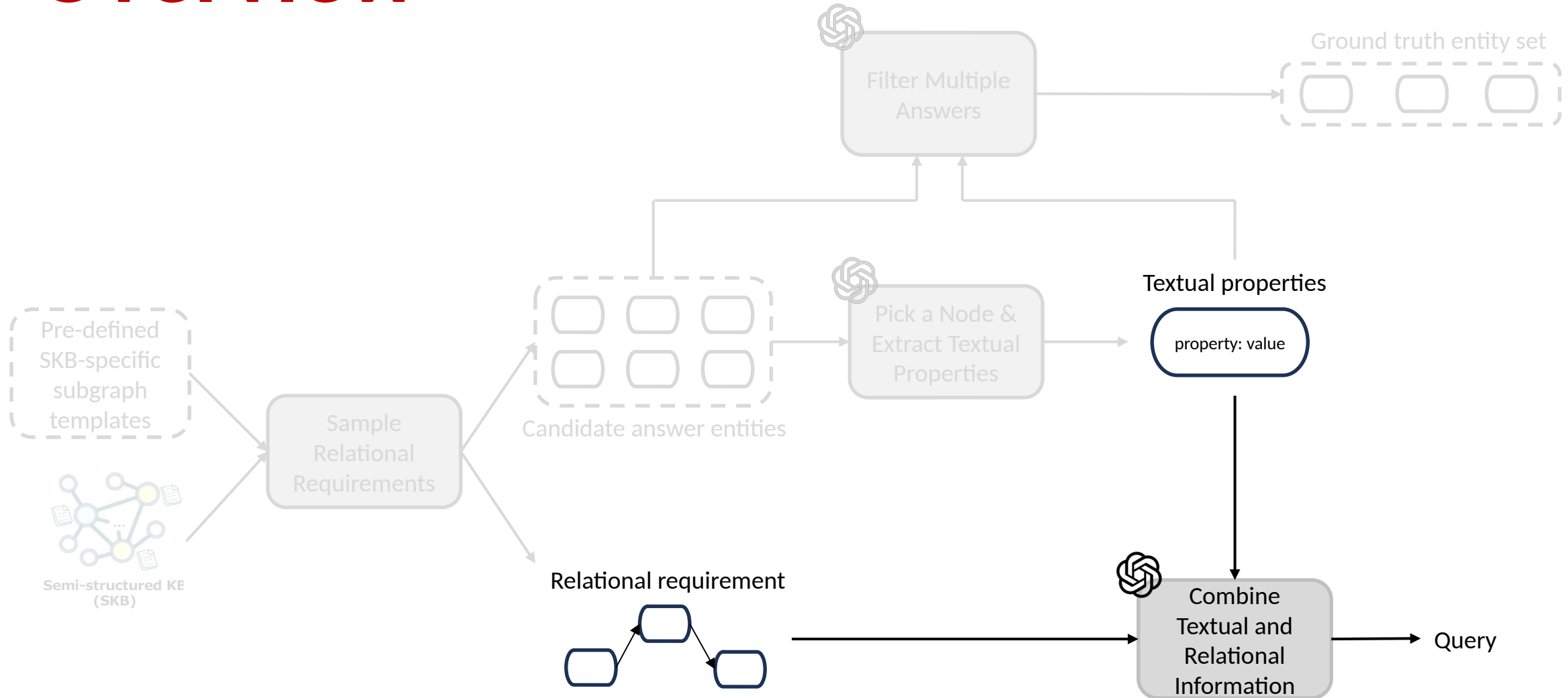
# Query-Ground Truth Entity Set Generation Overview



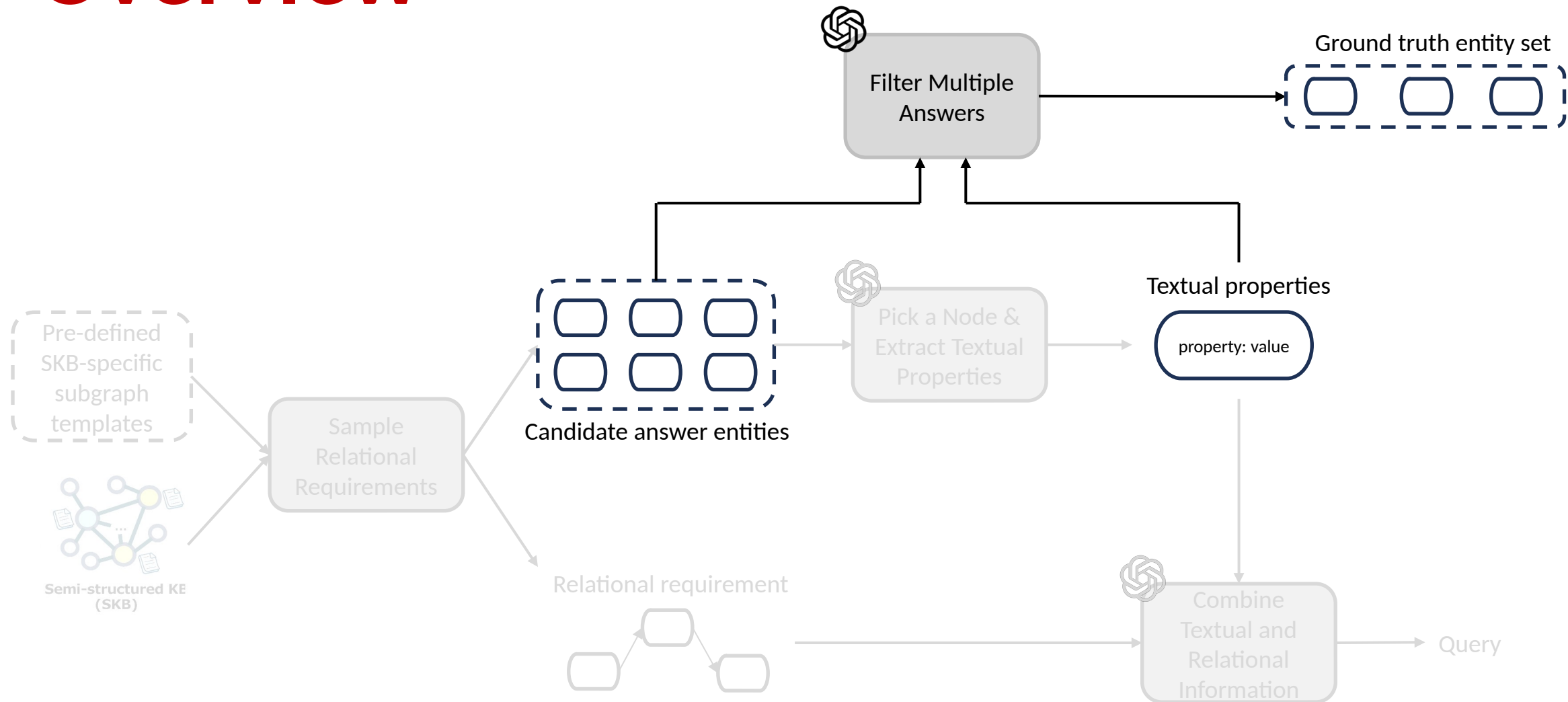
# Query-Ground Truth Entity Set Generation Overview



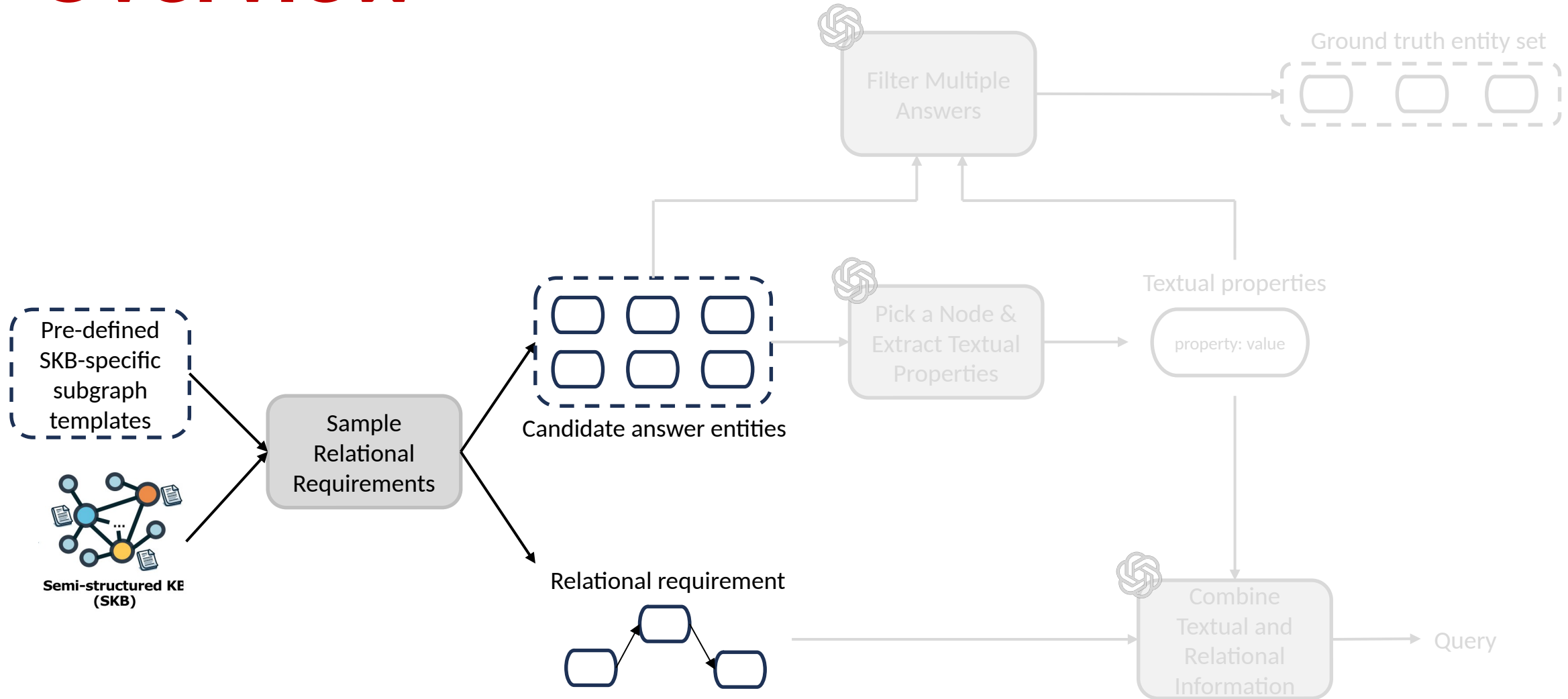
# Query-Ground Truth Entity Set Generation Overview



# Query-Ground Truth Entity Set Generation Overview



# Query-Ground Truth Entity Set Generation Overview



# Query-Ground Truth Entity Set Generation

## Step 1: Sample Relational Requirements

Initiate a relational requirement by sampling a relation template.

- **Relational requirement:** Subgraph expressed in string format
  - Nodes, node labels, relations, and one property-value is expressed.
- **Relation template:** A template for generating relational requirement

### Global example

**query:** Search publications by [Hao-Sheng Zeng](#) on non-Markovian dynamics.

**ground truth entity set:**

Distribution of non-Markovian intervals...

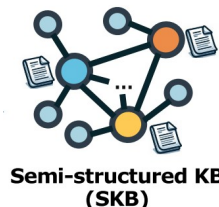
Comparison between non-Markovian...

### Relational templates

- Can you list the papers authored by <author>?
- Which papers have been cited by the paper <paper>?
- Can you provide a list of papers in the field of <field\_of\_study>?
- What papers have been published by researchers from <institution>?

Sample

Can you list the papers authored by <author>?



Pick a node to instantiate template

### Relational requirement

**author**

name: [Hao-Sheng Zeng](#)

**paper**

Can you list the papers authored by Hao-Sheng Zeng?

### Candidate ground truth entity set

**paper**

title: Equivalence of the measures of ...

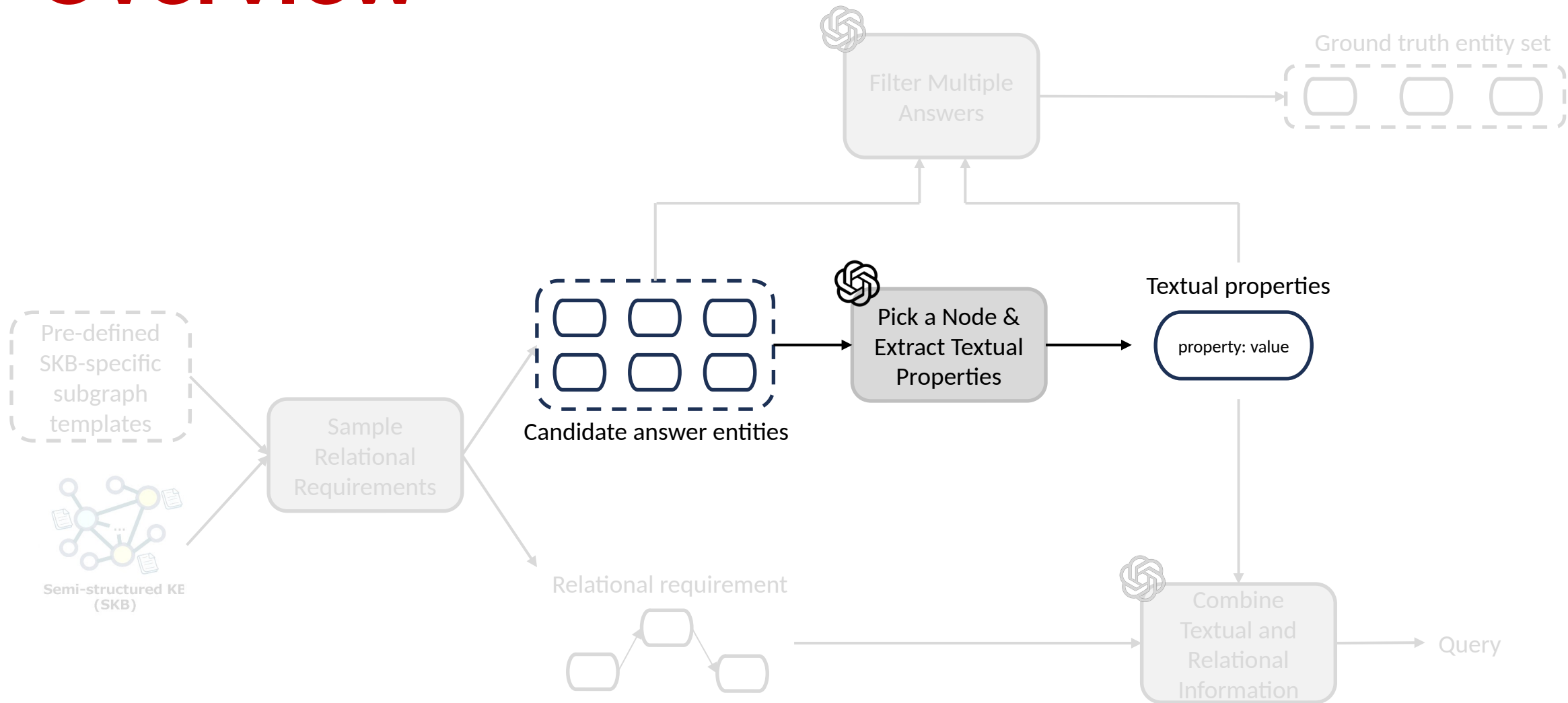
**paper**

title: Distribution of non-Markovian intervals...

**paper**

title: Comparison between non-Markovian

# Query-Ground Truth Entity Set Generation Overview



# Query-Ground Truth Entity Set Generation

## Step 2: Extract Textual Properties

1. Pick gold answer from candidate entity set.
2. Extract textual property from gold answer using LLM.
  - **Textual property:** Property-value constraint of gold answer expressed in string format.

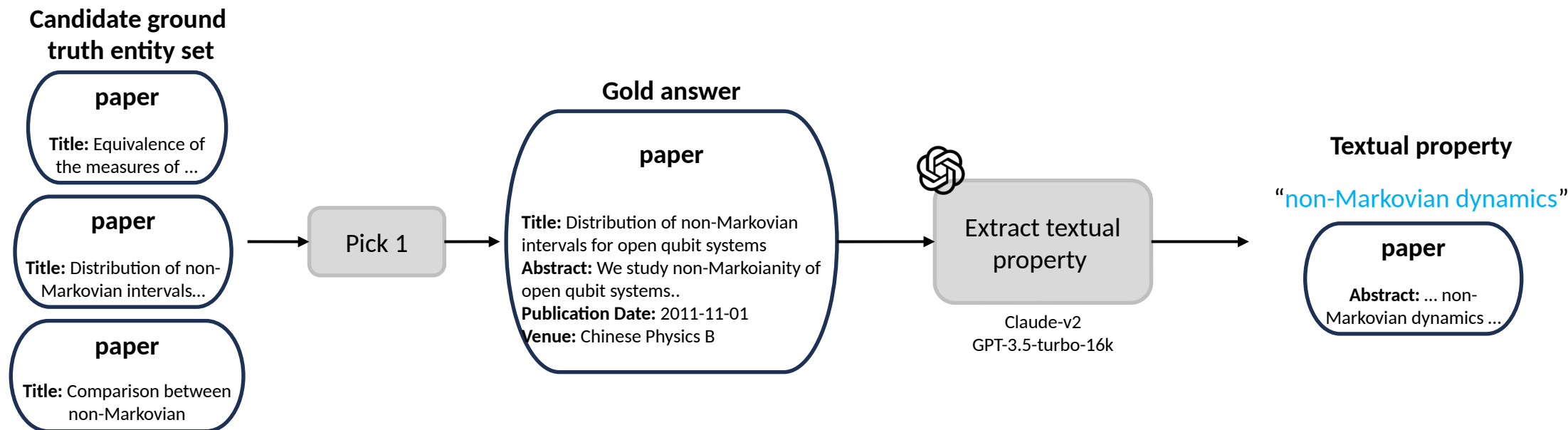
### Global example

**query:** Search publications by Hao-Sheng Zeng on **non-Markovian dynamics**.

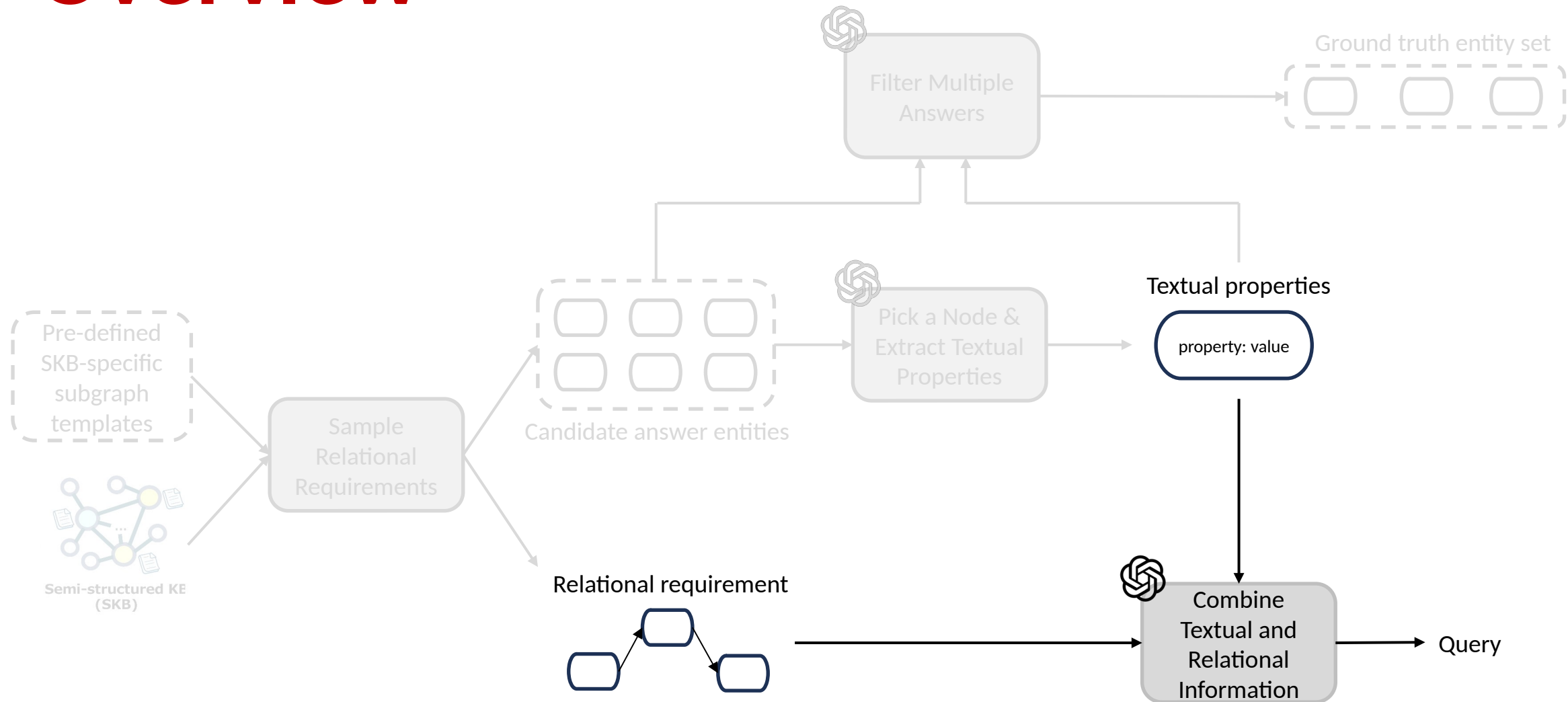
**ground truth entity set:**

Distribution of non-Markovian intervals...

Comparison between non-Markovian...



# Query-Ground Truth Entity Set Generation Overview



# Query-Ground Truth Entity Set Generation

## Step 3: Query Generation

Fuse relational requirement and textual property with LLMs.

1. Ensure natural-sounding query
2. Enrich context, rephrase the language

Global example

query: Search publications by Hao-Sheng Zeng  
on non-Markovian dynamics.

ground truth entity set:

Distribution of non-  
Markovian intervals...

Comparison between  
non-Markovian...

Relational requirement

author

name: Hao-Sheng Zeng

paper

Can you list the papers  
authored by Hao-Sheng Zeng?

Textual property

"non-Markovian dynamics"

paper

Abstract: ... non-  
Markovian dynamics ...



Fusion: stage 1

Claude-v2



Middle result  
sentence

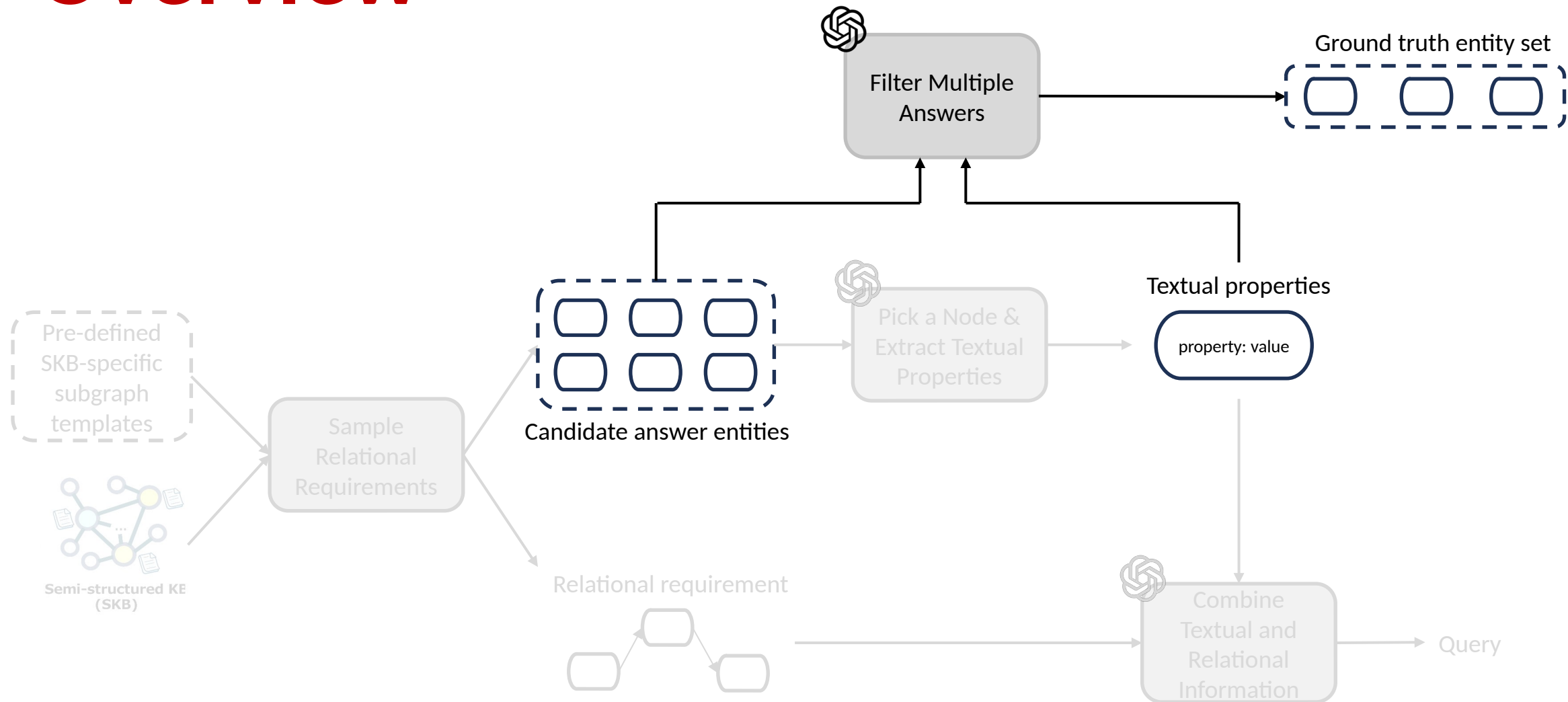


Fusion: stage 2

GPT-3.5-turbo-16k

Search publications by  
Hao-Sheng Zeng on non-  
Markovian dynamics.

# Query-Ground Truth Entity Set Generation Overview



# Query-Ground Truth Entity Set Generation

## Step 4: Ground Truth Entity Set Generation

Assess whether each remaining candidate meets the additional textual requirement.

- Candidates that pass the verification across 3 Claude models are included in the final ground truth answer set.

Global example

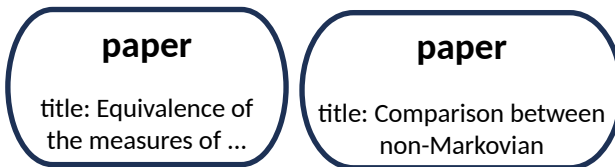
**query:** Search publications by Hao-Sheng Zeng on non-Markovian dynamics.

**ground truth entity set:**

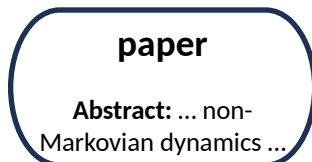
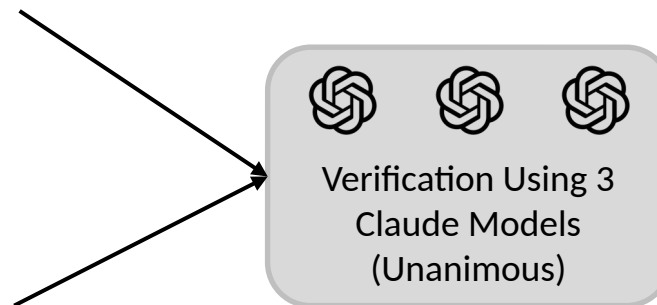
Distribution of non-Markovian intervals...

Comparison between non-Markovian...

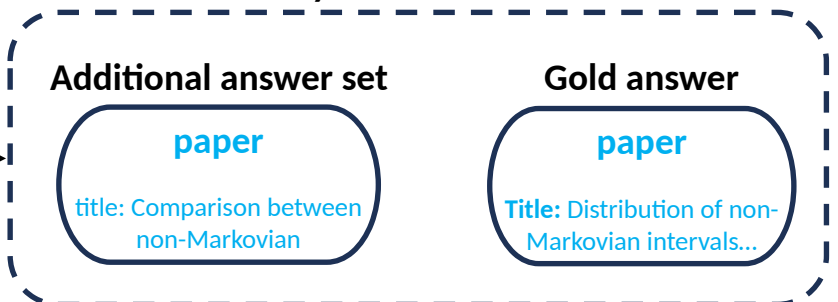
Candidate ground truth entity set  
(excluding gold answer)



Textual property  
"non-Markovian dynamics"



Ground truth entity set



# Analysis

# Query and Answer Length

- **Query length:** Number of tokens in a query
  - Reflects the amount of context information provided by user.
- **Answer length:** Number of ground truth answer nodes for a query
  - Indicates the level of inclusiveness or ambiguity of queries within a specific SKB.

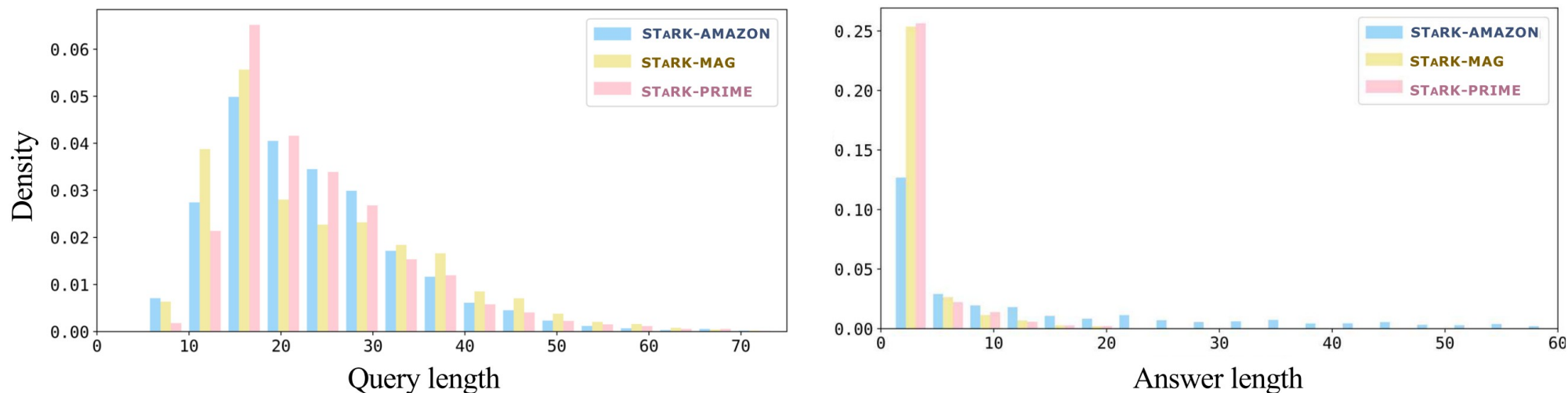


**Figure 5.** Distribution of query and answer lengths on STaRK datasets.

# Query Diversity

The diversity of query is measured with the diversity of tokens (token distribution).

- Entropy
  - $H(x) = -\sum_{x \in Vocab} p(x) \log p(x)$
- Type-Token Ratio
  - $TTR = \frac{\# \text{ of unique tokens}}{\text{Total number of tokens}}$

|                        | Shannon<br>Entropy | Type-Token<br>Ratio |
|------------------------|--------------------|---------------------|
| STARK-AMAZON           | 10.39              | 0.179               |
| STARK-MAG              | 10.25              | 0.180               |
| STARK-PRIME            | 9.63               | 0.143               |
| Reference <sup>1</sup> | 10.44              | 0.261               |

**Table 4.** Query diversity measurement on STARK.

# Ratio of Relational vs Textual Information

Analyze the proportionality between relational and textual information.

- Calculate the ratio of the length of relational requirements to the length of textual properties.

| Example query |                                                                                                                                                    |
|---------------|----------------------------------------------------------------------------------------------------------------------------------------------------|
| STARK-AMAZON  | Looking for <i>durable Dart World brand</i> dart flights that resist <i>easy tearing</i> . Any recommendations?                                    |
|               | What are recommended scuba diving weights for <i>experienced divers</i> that would fit well with my <i>Gorilla PRO XL waterproof bag</i> ?         |
| STARK-MAG     | Search publications by <i>Hao-Sheng Zeng</i> on <i>non-Markovian dynamics</i> .                                                                    |
|               | What are some <i>nanofluid heat transfer</i> research papers published by scholars from <i>Philadelphia University</i> ?                           |
| STARK-PRIME   | Could you provide a list of <i>investigational drugs</i> that interact with <i>genes or proteins active in the epididymal region</i> ?             |
|               | Search for diseases <i>without known treatments</i> and induce <i>pruritus in pregnant women</i> , potentially associated with <i>Autoimmune</i> . |
|               | Please find pathways <i>involving the POLR3D gene</i> within <i>nucleoplasm</i> .                                                                  |
|               | Which gene or protein associated with <i>lichen amyloidosis</i> can bind <i>interleukin-31</i> to activate the <i>PI3K/AKT and MAPK pathways</i> ? |

**Table 5.** Example pairs of query and answer node set for SKB retrieval task  
(**Relational requirements are expressed in blue**,  
and **textual properties are expressed in green**)

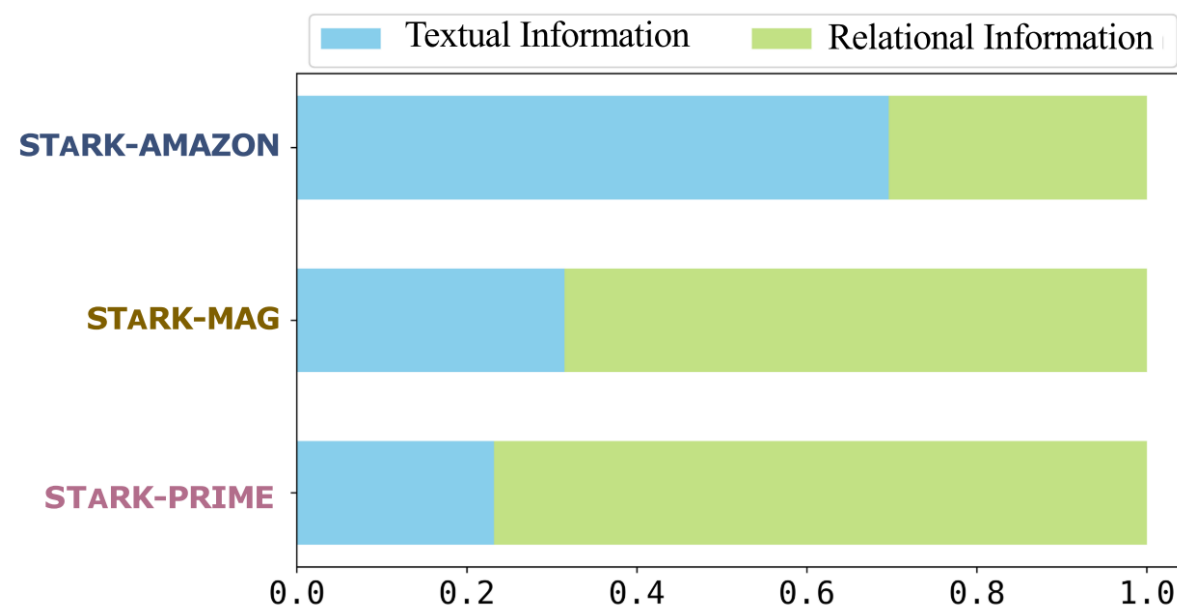


**Figure 6.** Average relative composition of relational requirement vs. textual property

# Human Evaluation

63 participants evaluated STaRK with 3 aspects in 1 ~ 5 points:

- Natural-sounding (resembling natural conversation)
- Diverse (covering a wide range of question structures and complexity levels)
- Practical (relevant and useful in real-life situations)

|              | Naturalness | Diversity   | Practicality |
|--------------|-------------|-------------|--------------|
| STARK-AMAZON | 73.6 / 89.5 | 68.4 / 89.5 | 89.5 / 94.7  |
| STARK-MAG    | 94.7 / 100  | 73.7 / 84.2 | 68.4 / 84.2  |
| STARK-PRIME  | 67.8 / 92.8 | 71.4 / 82.1 | 71.4 / 89.3  |
| Average      | 78.7 / 94.1 | 71.0 / 85.3 | 76.4 / 89.4  |

**Table 6.** Positive (4~) / Non-negative (3~) rates (%) of responses in each aspect of human evaluation.

# Experiment

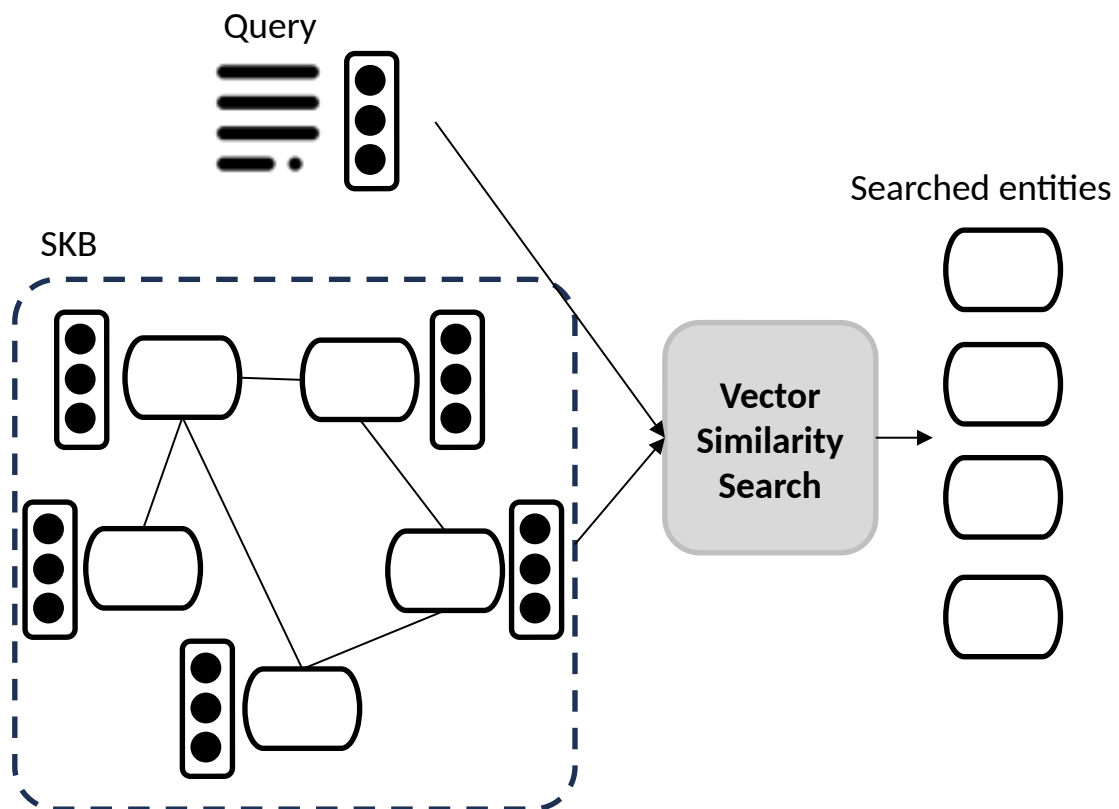
# Experiment Goals and Settings

- **Goals**
  - How well do current retrieval approaches perform on our benchmark datasets?
  - What challenges are revealed by the experiments, and what are the potential future directions for improvement?
  
- **Setting**
  - GPU setting: NVIDIA A100-SXM4-80GB \* 1

# Proposed Baseline Retrieval Methods

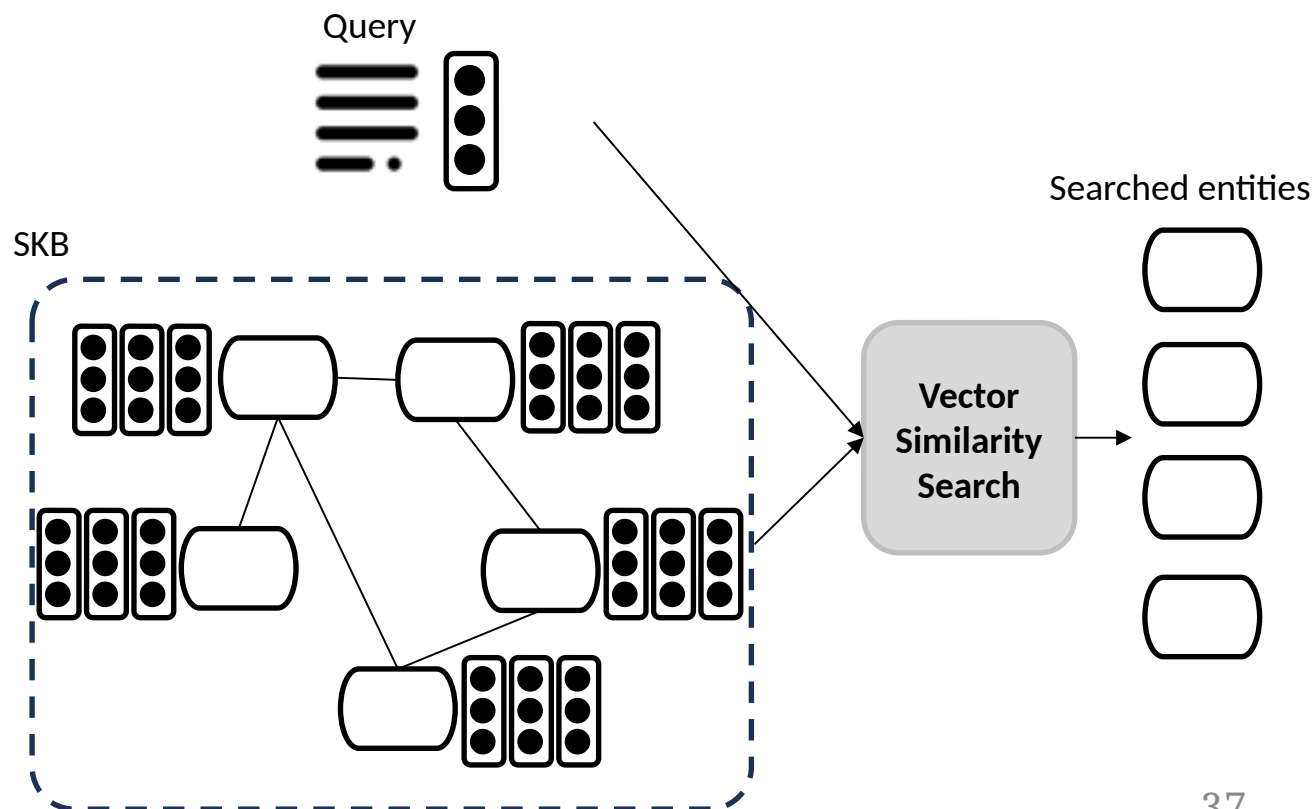
## Vector Similarity Search

encoder = ada-002



## Multi-Vector Similarity Search

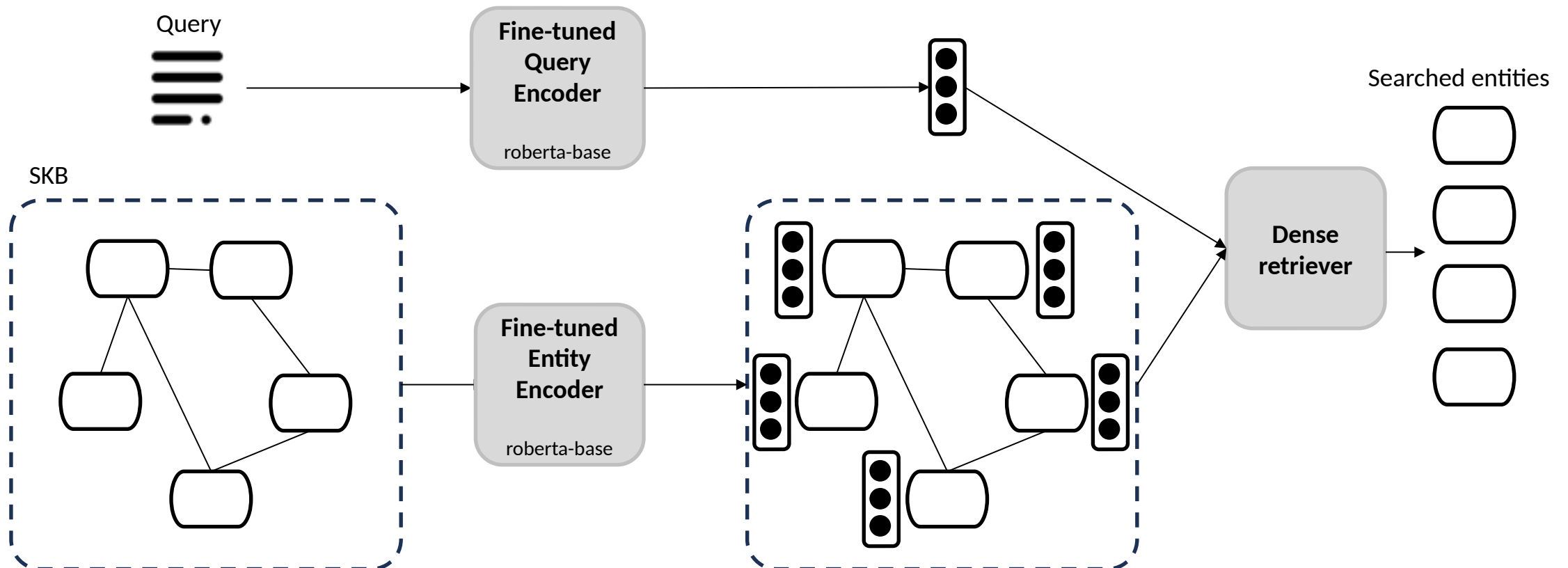
encoder = ada-002



# Proposed Baseline Retrieval Methods

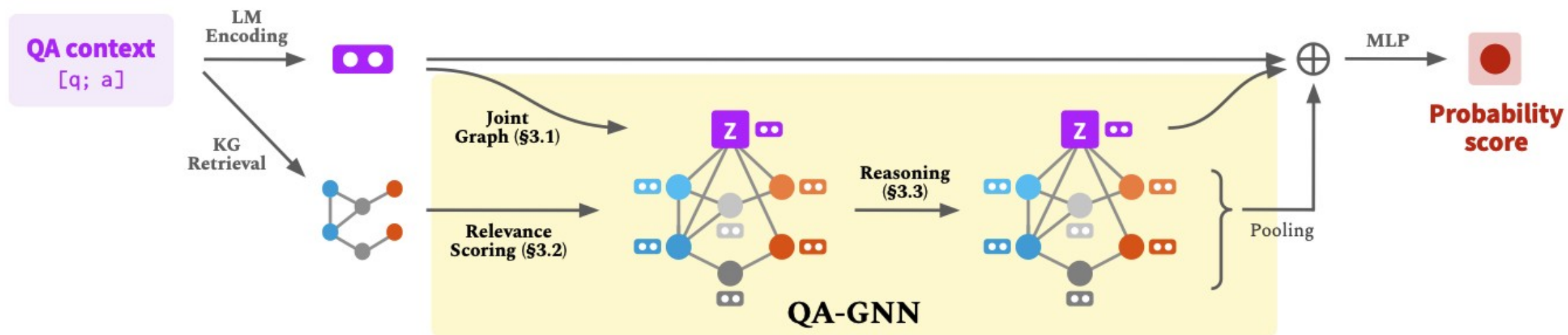
## Dense Retriever

Encoder fine-tuning is done via contrastive learning, with 20 hard negatives per positive sample



# Proposed Baseline Retrieval Methods

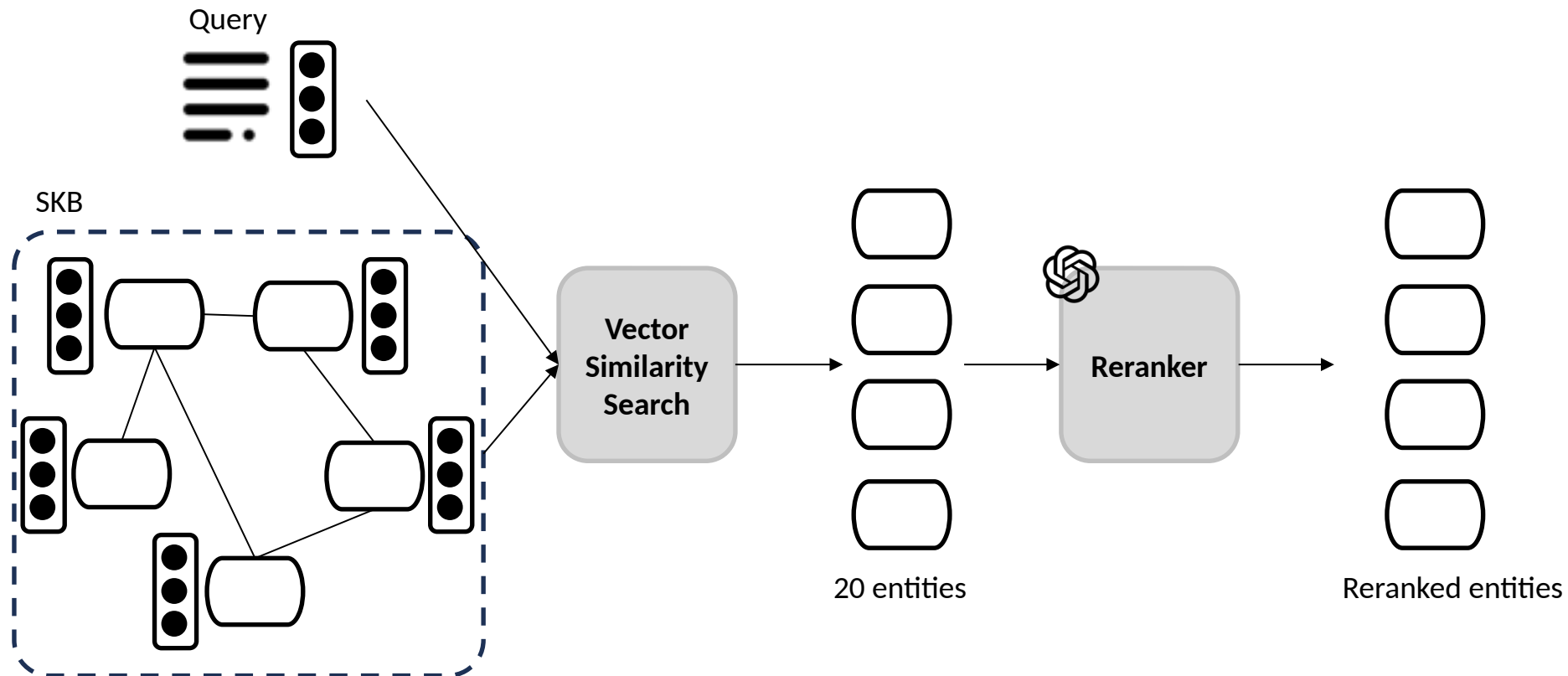
## QAGNN



# Proposed Baseline Retrieval Methods

## VSS + LLM Reranker

encoder = ada-002



# Metrics

## Effectivity

- Hit@k
  - Hit@1
  - Hit@5
- Recall@k
  - Recall@20
- Mean Reciprocal Rank

## Efficiency

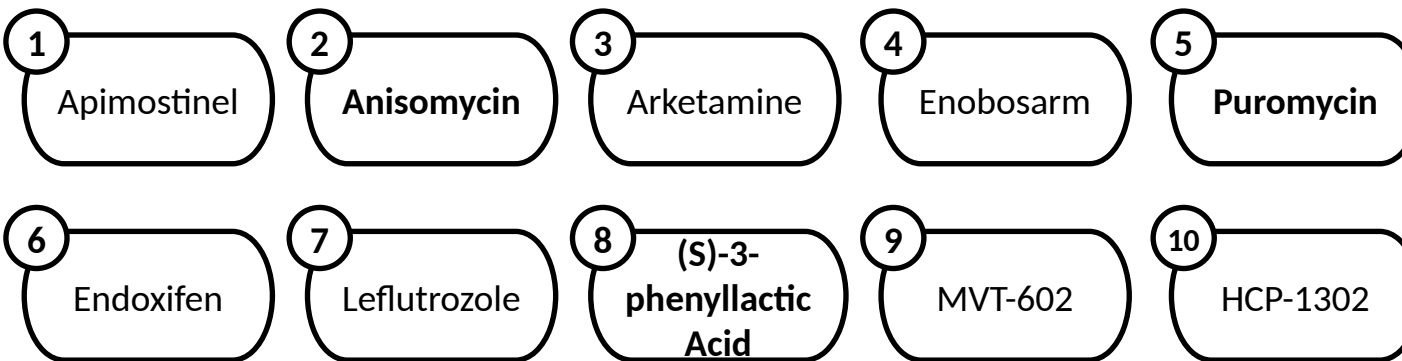
- Latency

**Query:** Could you provide a list of investigational drugs that interact with genes or proteins active in the epididymal region?

**Ground truth entities**



**Search results (sorted in descending similarity scores):**



| Hit@1 | Hit@5 | Recall@5 | MRR |
|-------|-------|----------|-----|
| 0     | 1     | 2/3      | 1/2 |

**Figure 7.** An example search result and its evaluation.

# Main Experimental Results

|              |                  | Dense Retriever<br>(roberta) | QAGNN<br>(roberta) | VSS<br>(ada-002) | Multi-VSS<br>(ada-002) | VSS+Claude2<br>Reranker | VSS+GPT4<br>Reranker |
|--------------|------------------|------------------------------|--------------------|------------------|------------------------|-------------------------|----------------------|
| STARK-AMAZON | <b>Hit@1</b>     | 0.1529                       | 0.2656             | 0.3916           | 0.4007                 | 0.4266                  | <b>0.4479</b>        |
|              | <b>Hit@5</b>     | 0.4793                       | 0.5001             | 0.6273           | 0.6498                 | 0.6746                  | <b>0.7117</b>        |
|              | <b>Recall@20</b> | 0.4449                       | 0.5205             | 0.5329           | 0.5512                 | 0.5376                  | <b>0.5535</b>        |
|              | <b>MRR</b>       | 0.3020                       | 0.3775             | 0.5035           | 0.5155                 | 0.5329                  | <b>0.5569</b>        |
| STARK-MAG    | <b>Hit@1</b>     | 0.1051                       | 0.1288             | 0.2908           | 0.2592                 | 0.3202                  | <b>0.4090</b>        |
|              | <b>Hit@5</b>     | 0.3523                       | 0.3901             | 0.4961           | 0.5043                 | 0.5334                  | <b>0.5818</b>        |
|              | <b>Recall@20</b> | 0.4211                       | 0.4697             | 0.4836           | <b>0.5080</b>          | 0.4834                  | 0.4860               |
|              | <b>MRR</b>       | 0.2134                       | 0.2912             | 0.3862           | 0.3694                 | 0.4129                  | <b>0.4900</b>        |
| STARK-PRIME  | <b>Hit@1</b>     | 0.0446                       | 0.0885             | 0.1263           | 0.1510                 | 0.1611                  | <b>0.1828</b>        |
|              | <b>Hit@5</b>     | 0.2185                       | 0.2135             | 0.3149           | 0.3356                 | 0.3582                  | <b>0.3728</b>        |
|              | <b>Recall@20</b> | 0.3013                       | 0.2963             | 0.3600           | <b>0.3805</b>          | 0.3598                  | 0.3405               |
|              | <b>MRR</b>       | 0.1238                       | 0.1473             | 0.2141           | 0.2349                 | 0.2466                  | <b>0.2655</b>        |

Table 7. Main experimental results.

# Efficiency Experiment Result

|                     | Dense Retriever | QAGNN | VSS  | Multi-VSS | VSS+Claude | VSS+GPT4 |
|---------------------|-----------------|-------|------|-----------|------------|----------|
| <b>STARK-AMAZON</b> | 2.34            | 2.32  | 5.71 | 4.87      | 27.24      | 24.76    |
| <b>STARK-MAG</b>    | 0.94            | 1.35  | 2.25 | 3.14      | 22.60      | 23.43    |
| <b>STARK-PRIME</b>  | 0.92            | 1.29  | 0.54 | 0.90      | 29.14      | 26.97    |
| <b>Average</b>      | 1.40            | 1.65  | 2.83 | 2.97      | 26.33      | 25.05    |

**Table 8.** Latency (s) of the retrieval systems on STaRK.

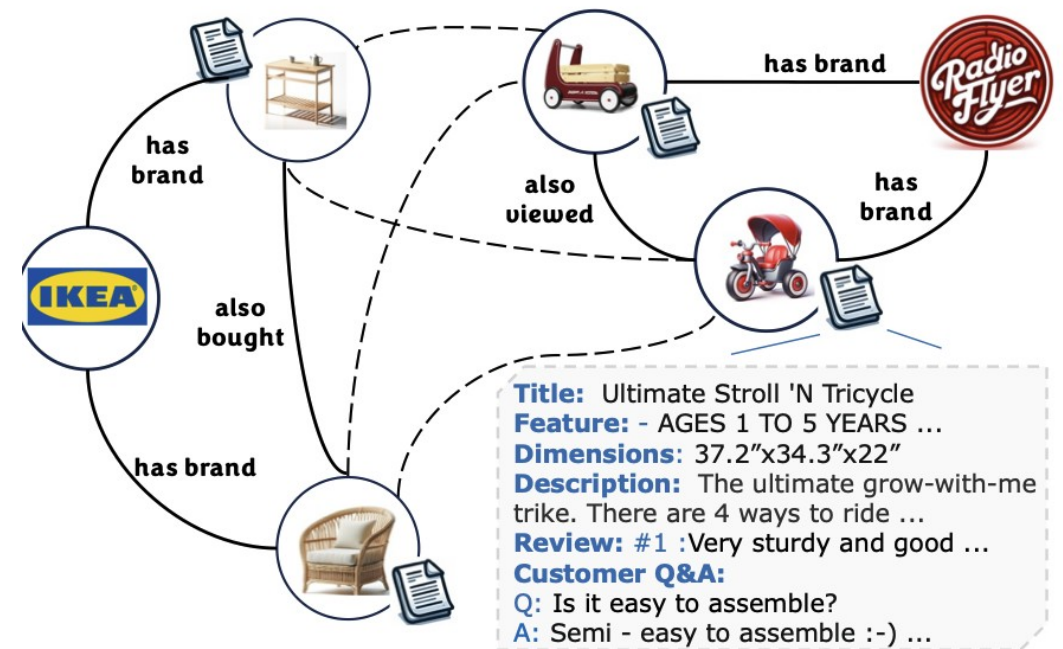
# APPENDIX

# Generated SKB 1

## Amazon Semi-structured Knowledge Base

### Generation method:

- Generation method of knowledge graph is not denoted.
- Textual information for products obtained by
  - the meta data from *Amazon Product Reviews*.
  - customer Q&A records from *Amazon Question and Answer Data*.



**Amazon Semi-structured Knowledge Base**

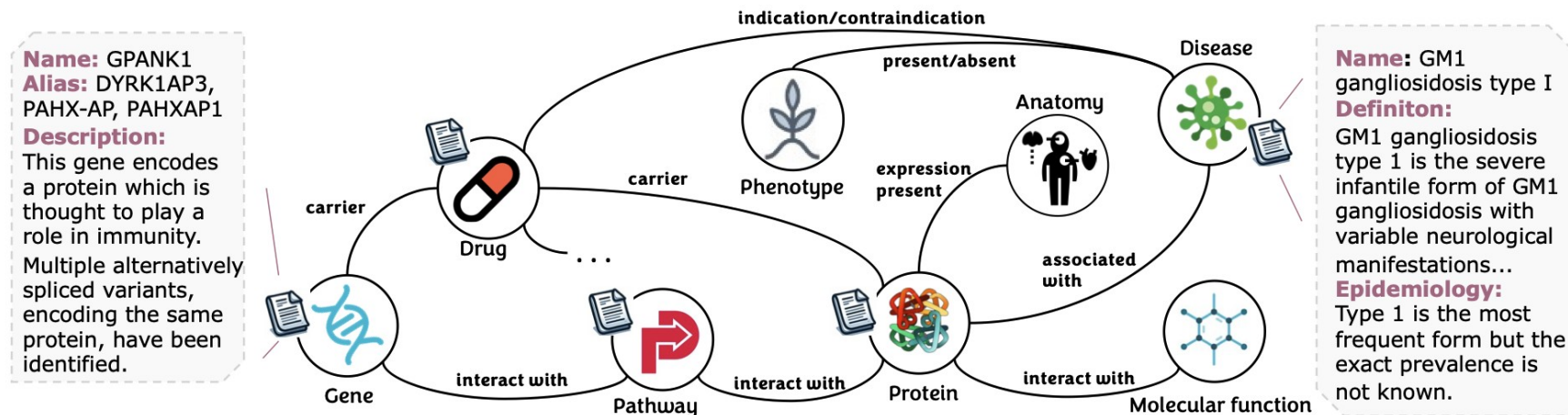
Figure 4. The Amazon semi-structured knowledge base

# Generated SKB 3

## Prime Semi-structured Knowledge Base

### Generation steps:

1. Leverage the existing knowledge graph *PrimeKG*
2. Integrate the textual details from multiple databases for gene/protein and pathway entities such as genomic position, gene activity summary and pathway orthologous event.



**Prime Semi-structured Knowledge Base**

Figure 6. The Prime semi-structured knowledge base