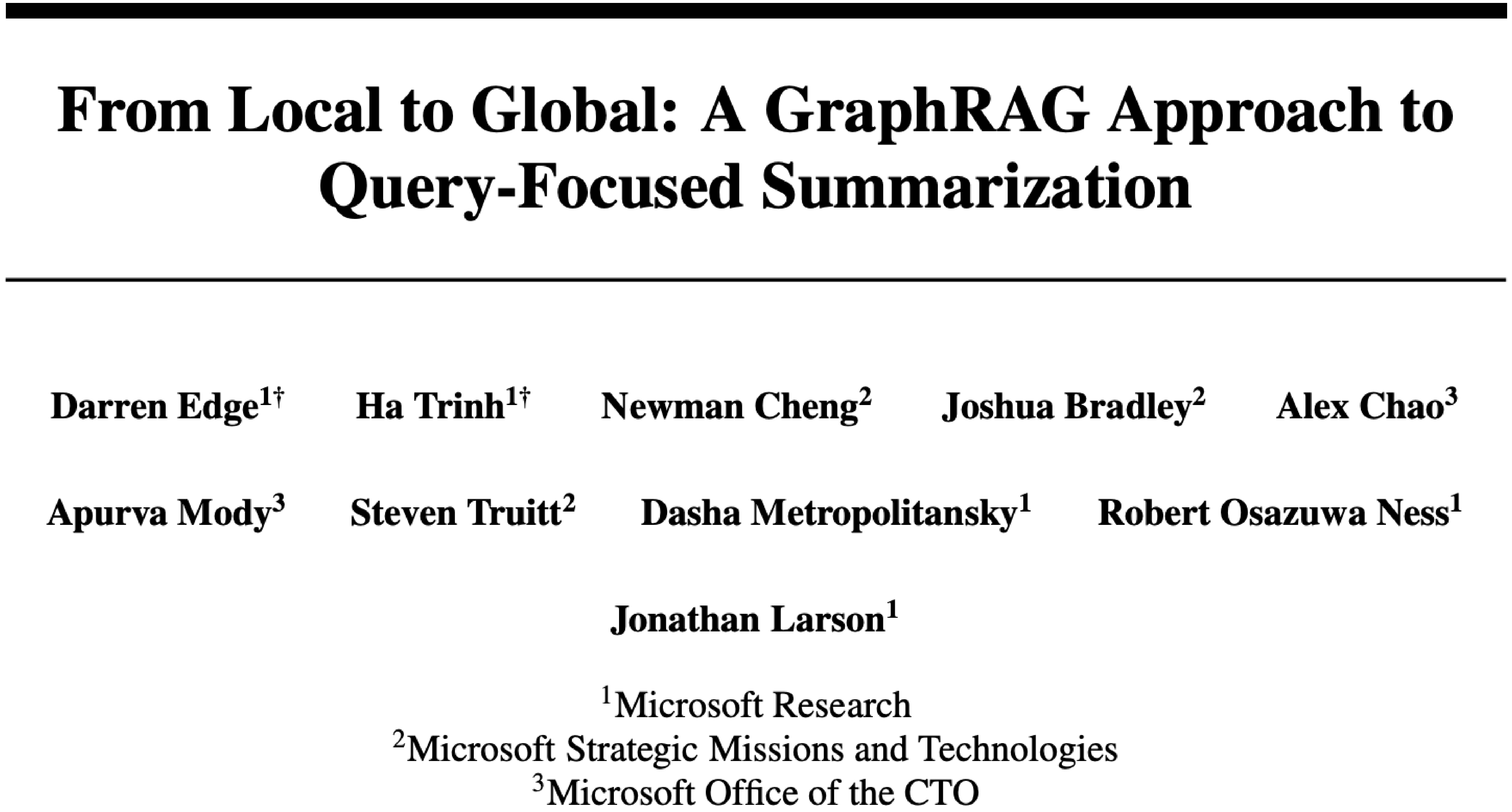


From Local to Global: A GraphRAG Approach to Query-Focused Summarization

Darren Edge[1†] Ha Trinh[1†] Newman Cheng[2] Joshua Bradley[2] Alex Chao[3]

Apurva Mody[3] Steven Truitt[2] Dasha Metropolitansky[1] Robert Osazuwa Ness[1]

Jonathan Larson[1]

[1]Microsoft Research
[2]Microsoft Strategic Missions and Technologies
[3]Microsoft Office of the CTO

presented by: Yan-he(Evian) Chen

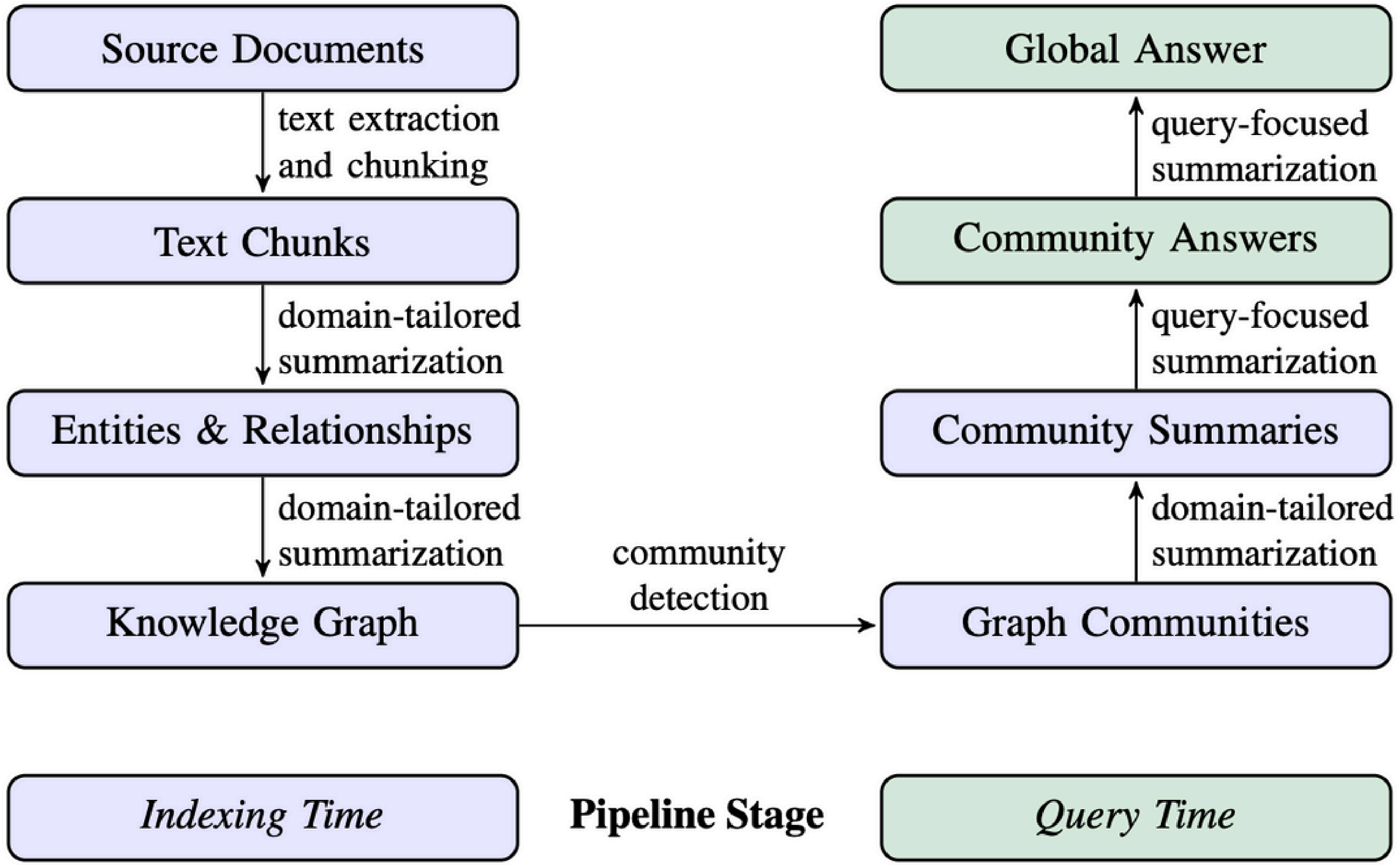
Outline

- Introduction
- Methodology

Introduction

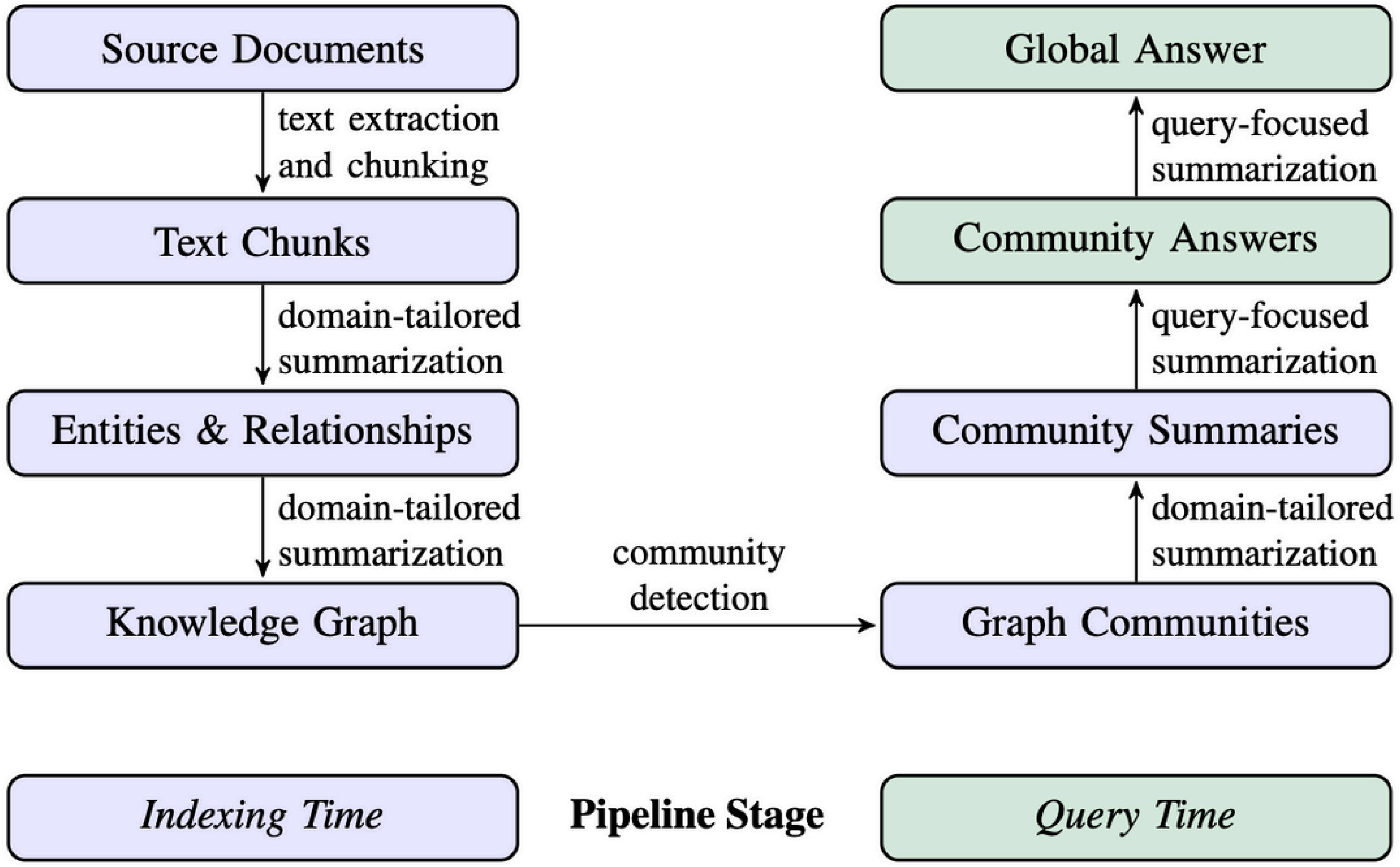
- Vector RAG approach does not support sensemaking queries (queries that require global understanding of the entire dataset)
- GraphRAG (graph-based RAG) enables sensemaking over a large text corpus
- GraphRAG uses community summaries from different levels to generate the final global answer
- In the experiments, there’s no gound truth. The authors developed LLM-as-a-judge technique suitable for questions targeting broad issues and themes

Methodology



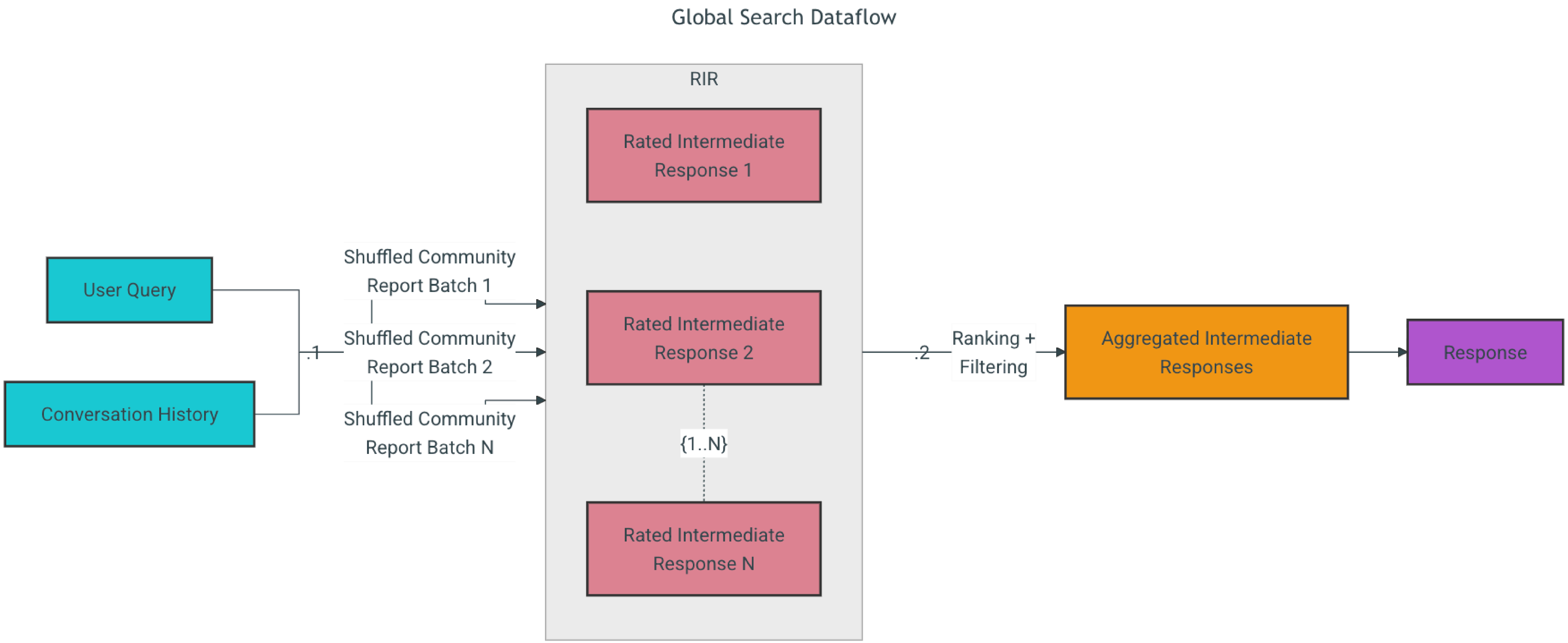
- The documents are split into text chucks. The size of chunks is a hyperparameter
- LLM is prompted to extract important entities and the relationships between them from a given chuck. Claims can also be extracted using prompts
- Entities and relationships become nodes and edges in the graph
- Using Leiden community detection to partition the community
- Generate report-like summaries od each community in the hierarchy

Methodology

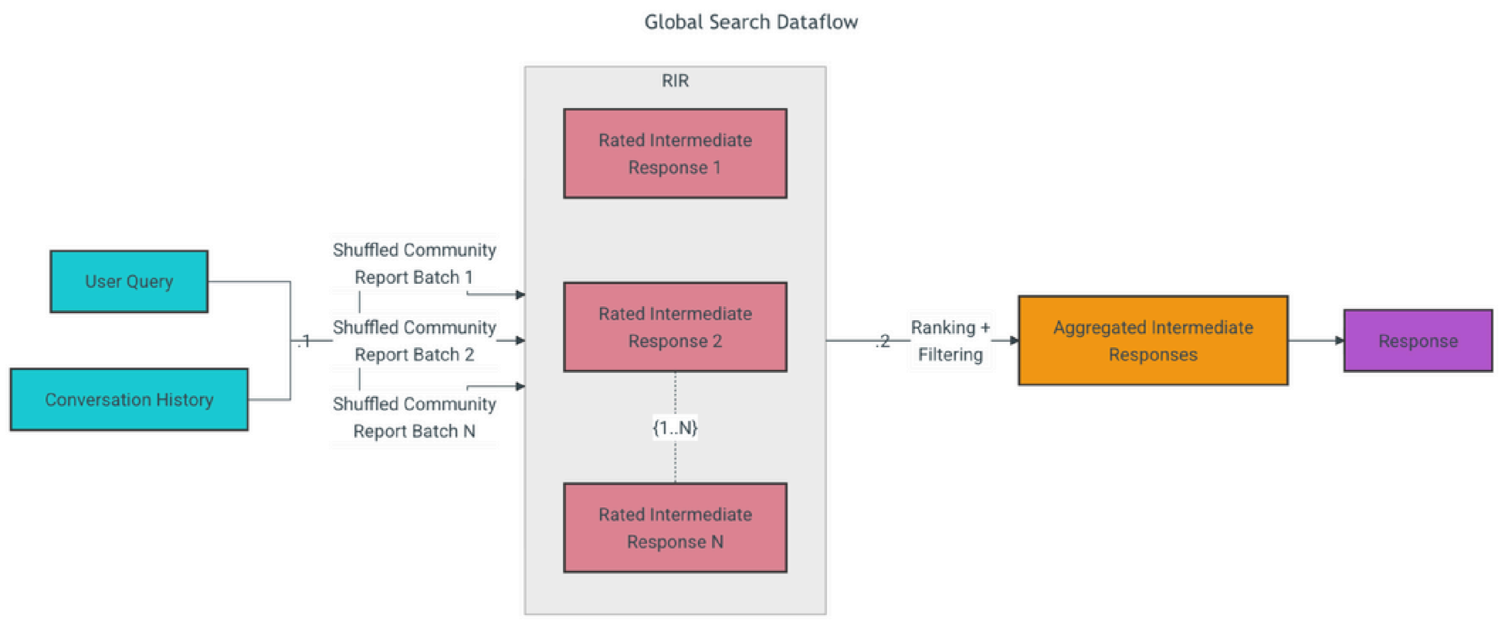


- Questions can be answered using the community summaries from different levels
- Community summaries are shuffuled to ensure relevant is distributed across chunks
- Every chunk will be fed into LLM with user query, LLM will generate score 0-100 indicating how helpful it is
- Intermediate answers are sorted in descending order and fed into LLM until token limit

Global Search

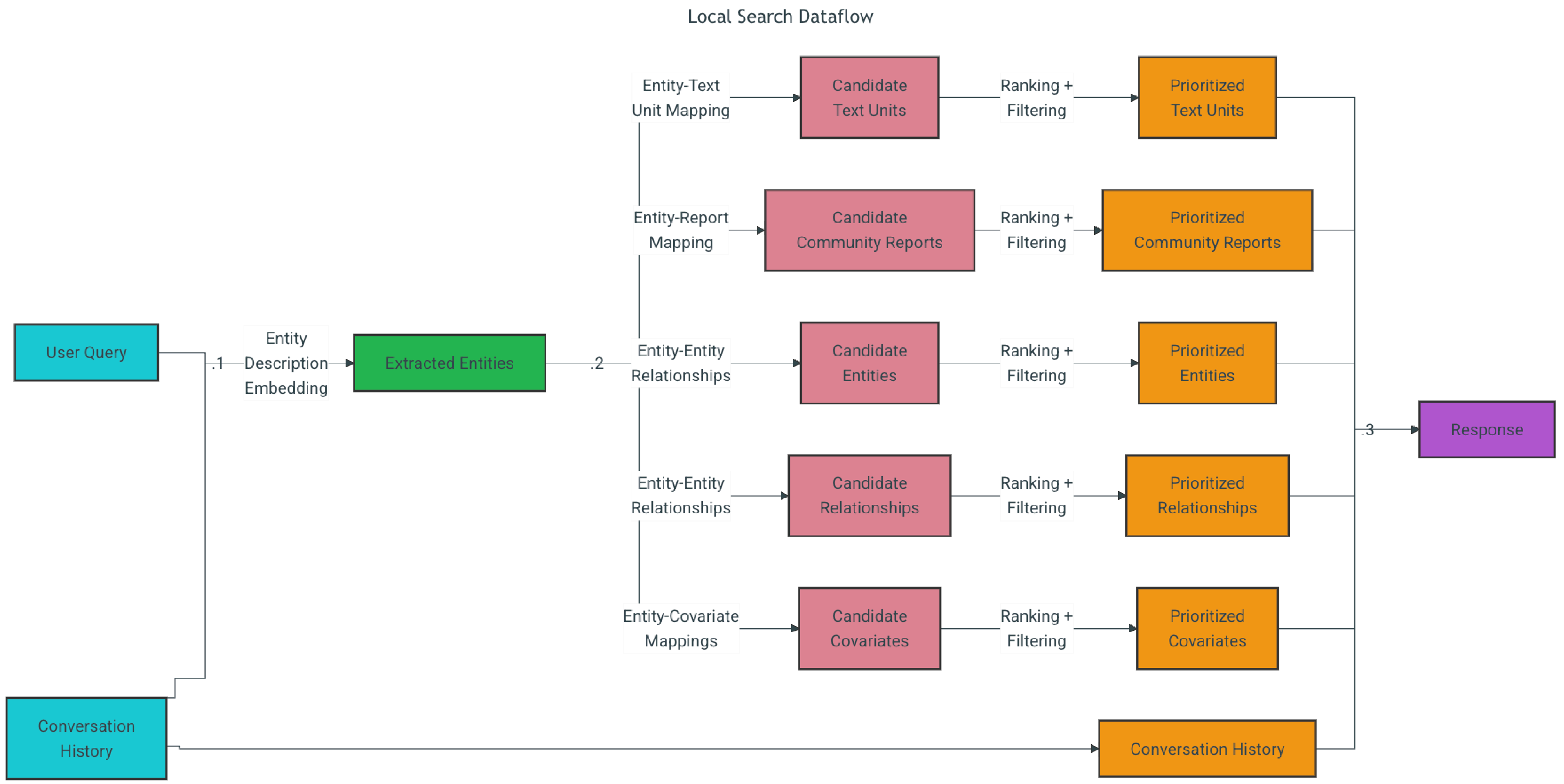

Global Search Dataflow
RIR
Rated Intermediate Response 1
User Query
Conversation History
.1
Shuffled Community Report Batch 1
Shuffled Community Report Batch 2
Shuffled Community Report Batch N
Rated Intermediate Response 2
{1..N}
Rated Intermediate Response N
.2
Ranking + Filtering
Aggregated Intermediate Responses
Response

Global Search

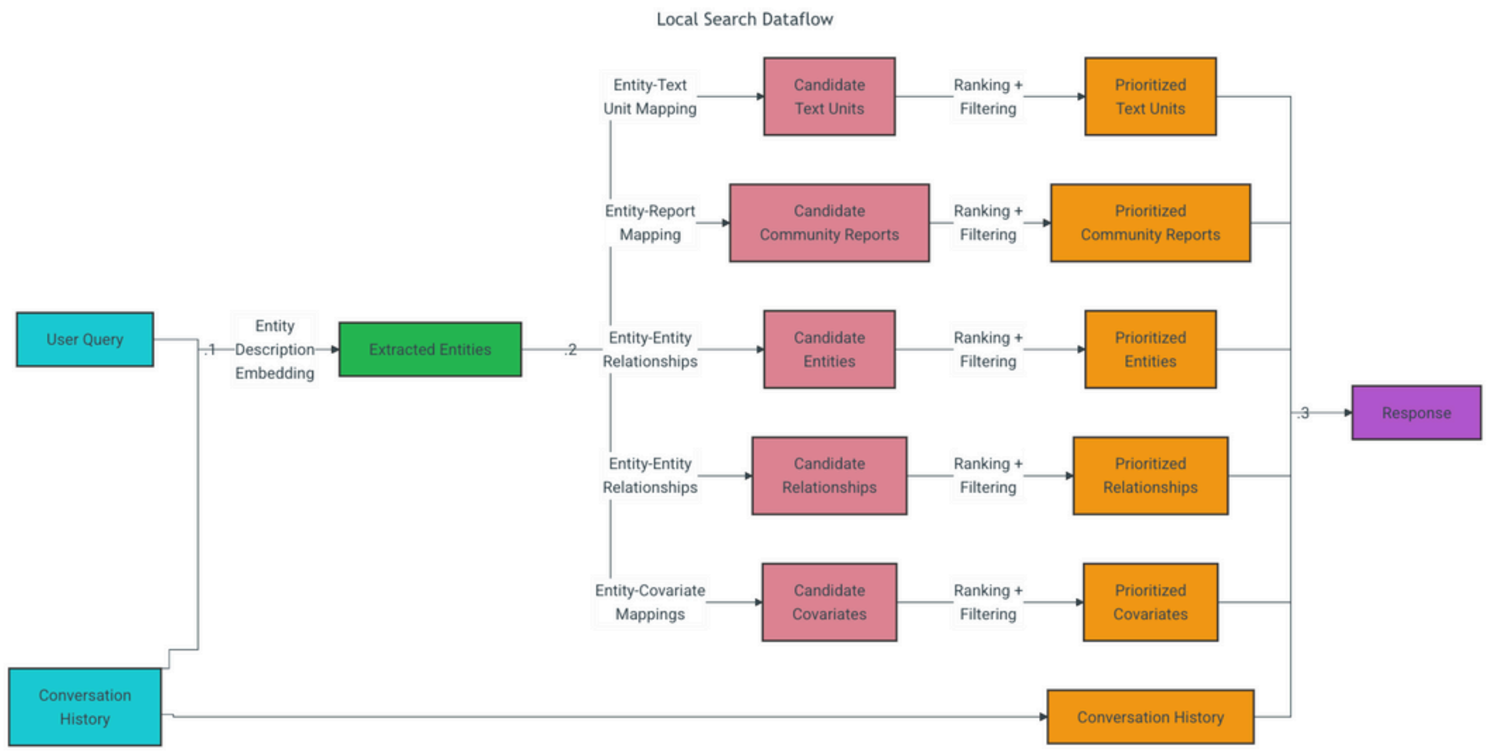

Global Search Dataflow
RIR
Rated Intermediate Response 1
User Query
Conversation History
1
Shuffled Community Report Batch 1
Shuffled Community Report Batch 2
Shuffled Community Report Batch N
Rated Intermediate Response 2
{1..N}
Rated Intermediate Response N
2
Ranking + Filtering
Aggregated Intermediate Responses
Response

- Input user query and conversation history (optional)
- LLM generates reports of communities, segements reports into text chunks
- The shuffuled chunks will be used to generate intermediate response
- The filtered (by numerical rating indicating the importance) set of points are aggregated to generate final response

Local Search



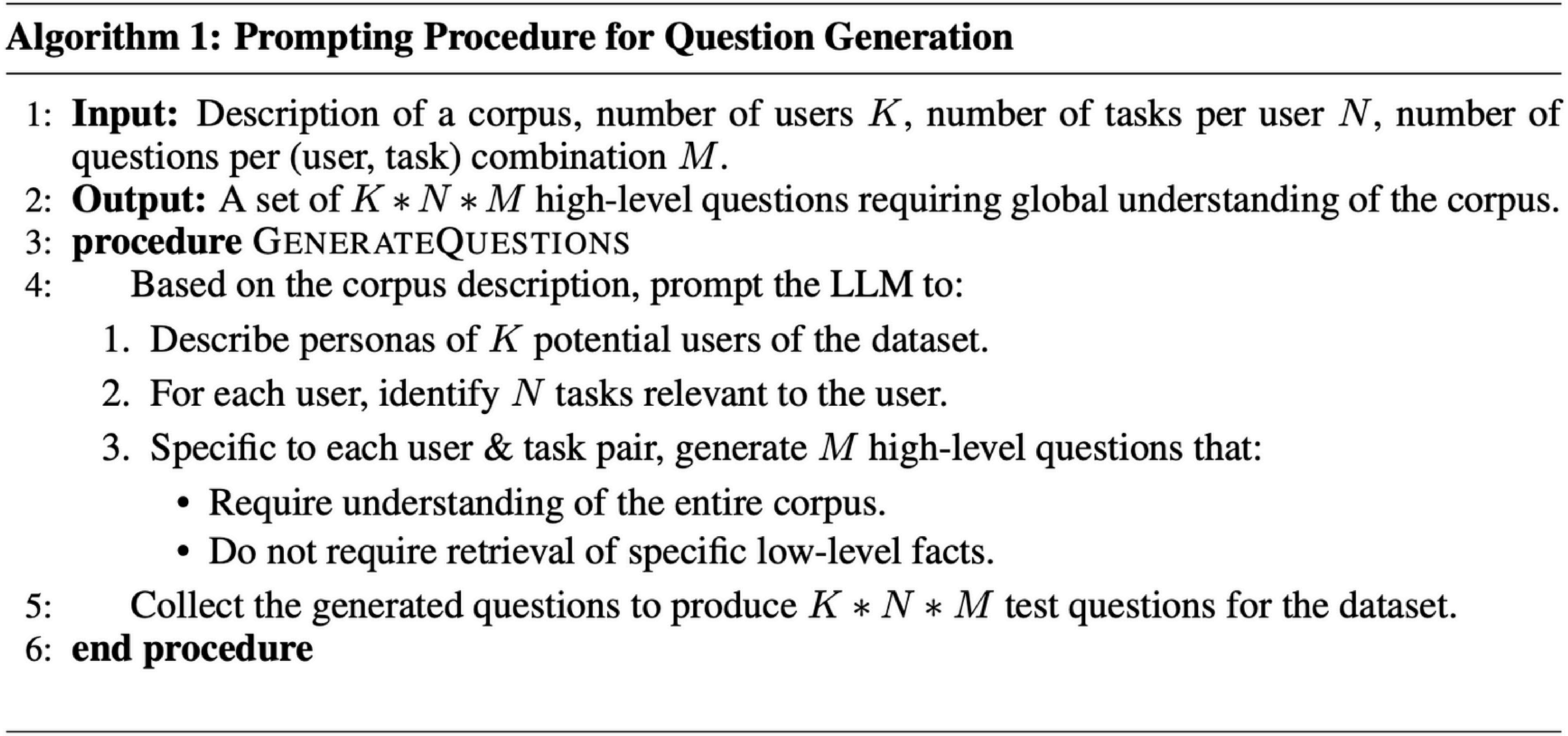
Local Search



- Input user query and conversation history (optional)
- LLM identifies the entities (access points into knowledge graph) to extract further relevant details
- All the data will be filtered and fit within a single context window to generate the final response

Question Generation

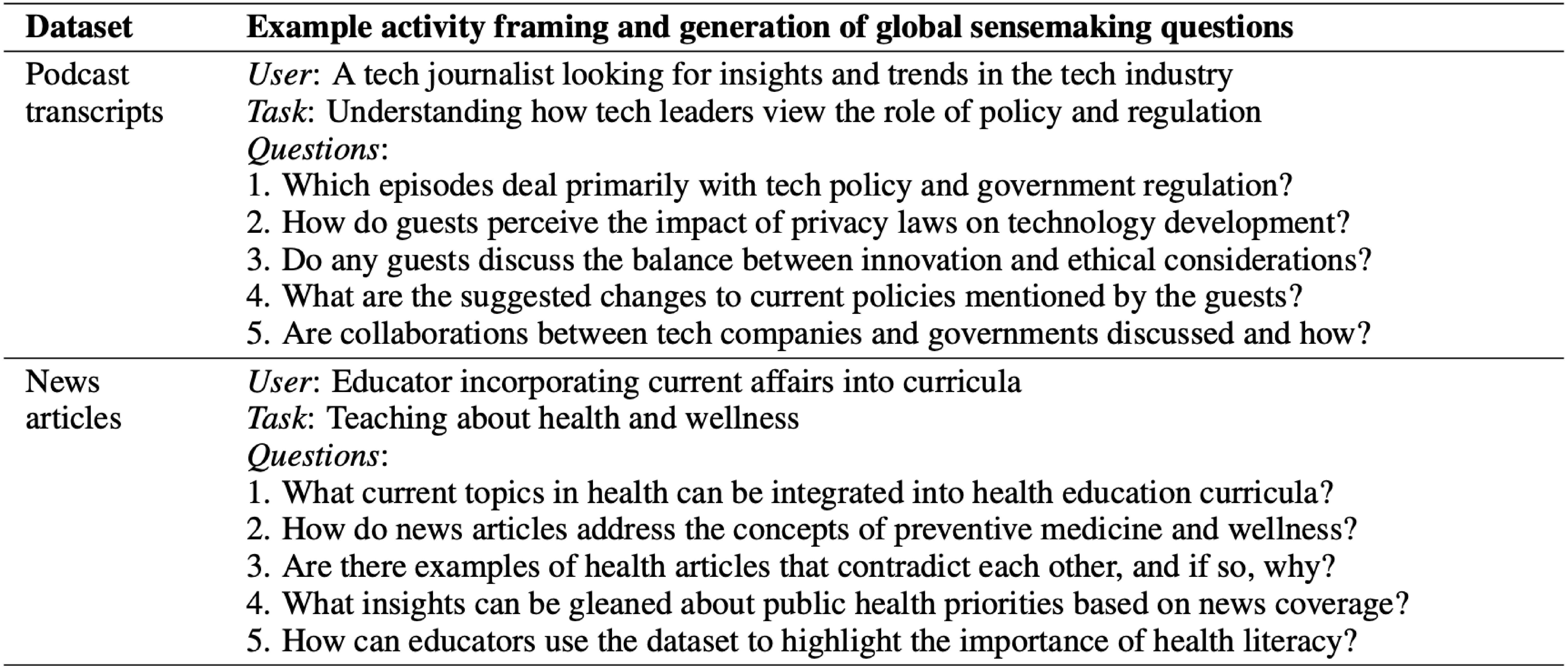
- To evaluate the RAG system for global sensemaking tasks, authors use LLM to generate corpus-specific questions



Algorithm 1: Prompting Procedure for Question Generation

```
Input: Description of a corpus, number of users $K$, number of tasks per user $N$, number of questions per (user, task) combination $M$.
Output: A set of $K * N * M$ high-level questions requiring global understanding of the corpus.
procedure GenerateQuestions
    Based on the corpus description, prompt the LLM to:
      1. Describe personas of $K$ potential users of the dataset.
      2. For each user, identify $N$ tasks relevant to the user.
      3. Specific to each user & task pair, generate $M$ high-level questions that:
           • Require understanding of the entire corpus.
           • Do not require retrieval of specific low-level facts.
    Collect the generated questions to produce $K * N * M$ test questions for the dataset.
end procedure
```

Question Generation - Example



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?

Reference

- GraphRAG paper: https://arxiv.org/pdf/2404.16130
- Global search: https://microsoft.github.io/graphrag/query/global_search/
- Local search: https://microsoft.github.io/graphrag/query/local_search/