TARGET: Benchmarking Table Retrieval for Generative Tasks

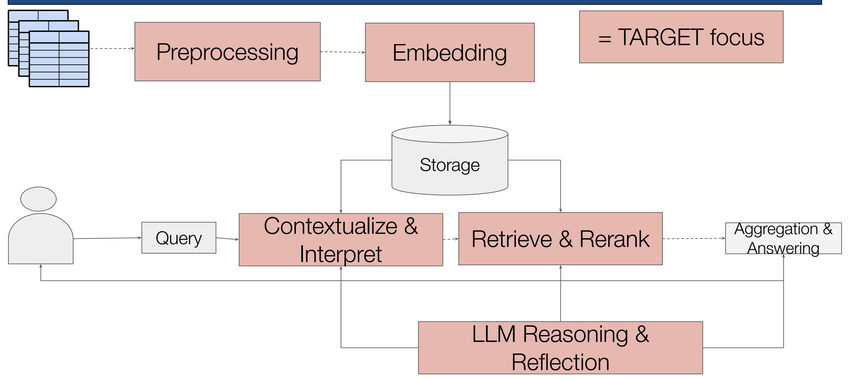
Madelon Hulsebos, Carl Ji, Rachel Xin, Aditya Parameswaran

Evaluating Structured Data Retrieval in RAG Pipelines

Why table retrieval?

- People need quality RAG! 🤖
  - LLMs allow ***reasoning*** over ***large datasets***.
  - RAG techniques for ***structured data*** requires further exploration & benchmarking.
- Focus on Table Retrieval 📊
  - Current benchmarks for IR evaluates ***end result*** (ie, fact verification, table QA, etc.).
  - Capabilities of RAG tools to ***retrieve the correct tables*** heavily influences downstream task generation quality.

Common RAG Pipeline



Why TARGET?

- Problems 🧐
  - Different tools ***vary significantly*** in how structured data is preprocessed and embedded!
  - Different tools make diff. assumptions about the data, focus on different tasks, etc.
- **Questions ❓**
  - How to evaluate the effectiveness of table retrieval?
  - Can we build a benchmark that easily adapts to these different RAG tools?

TARGET Evaluation Workflow

Datasets with various types of downstream tasks

What kind of tasks? 🏗️

- Text-to-SQL/Pandas
- Table Question Answering
- Fact Verification
- Table Grounded Dialogue Generation
- Data Analysis (reasoning)

Queries

"What were the voting results of the 1975 North Vietnamese legislative election with regards to seats?"

User Interface

Custom Table Retrieval Tool

TARGET Interface 💻

A simple interface for users to plug in their custom model. Only need to specify how to embed corpus & retrieve.

Fixed Generators ⚙️

Standardize downstream task completion to isolate retrieval capabilities.

Fixed Downstream Task Generators

Retrieved Tables

Party	Votes	%	Seats
...	...	...	...
Vietnamese Fatherland Front	10,561,314	100	424
Invalid/blank votes	105,807	-	-
Total	10,667,121	100	424
...	...	...	...

Generated Answers

In the 1975 North Vietnamese legislative election, the Vietnamese Fatherland Front won all 424 seats.

Evaluate **Table Retrieval** Results

Evaluate **Downstream** Results

Task Evaluation 📋

- **Accuracy**: accuracy of retrieval?
- **Efficiency**: performance v resource usage trade-offs?
- **Query intent**: How well do retrieval methods perform across different types of tasks?
- **Generalizability**: How well do retrieval methods generalize across datasets per task?
- **Metrics**: precision@, recall@, index storage, query latency, etc.