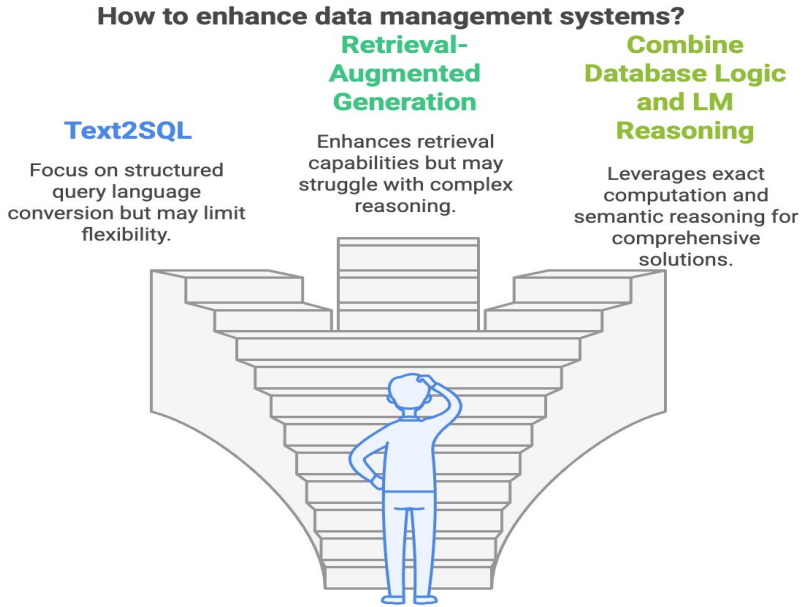
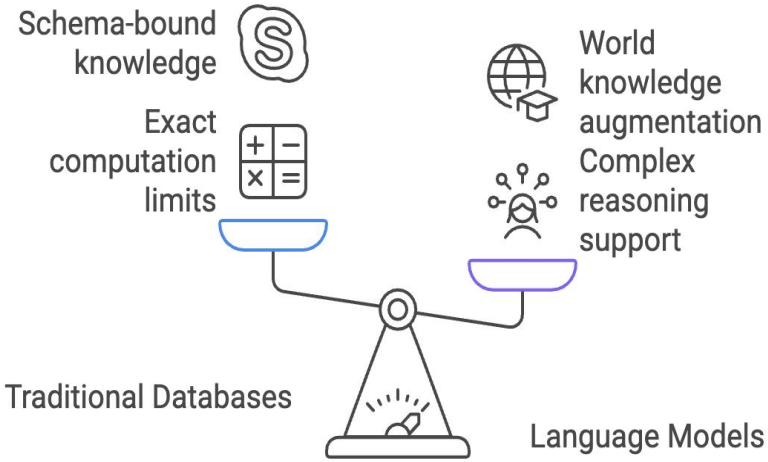
Text2SQL is Not Enough: Unifying AI and Databases with TAG

Asim Biswal, Liana Patel, Siddharth Jha, Amog Kamsetty, Shu Liu, Joseph E. Gonzalez, Carlos Guestrin, Matei Zaharia

UC Berkeley, Stanford University

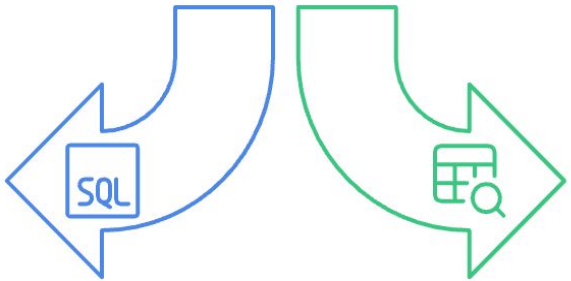
Presented by: Rudra Ranajee Saha





Comparing Data Processing Approaches

Which method best handles natural language queries?

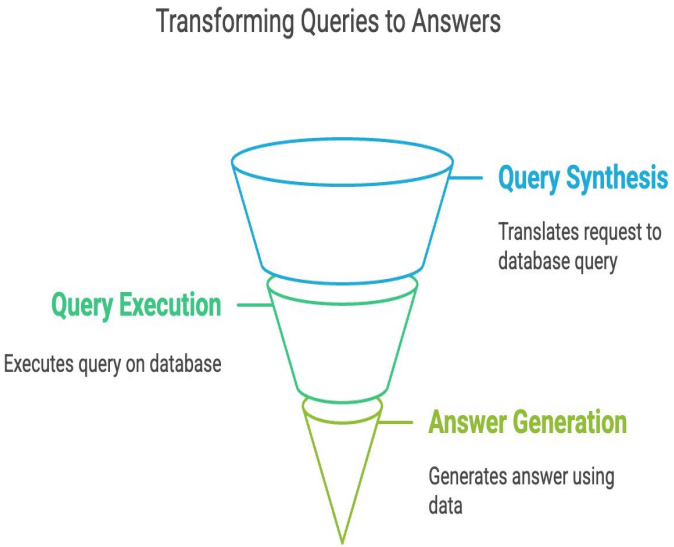


Text2SQL

Suitable for queries with direct relational equivalents but lacks semantic reasoning.

RAG Model

Limited to simple relevance-based lookups, failing to leverage full database capabilities.



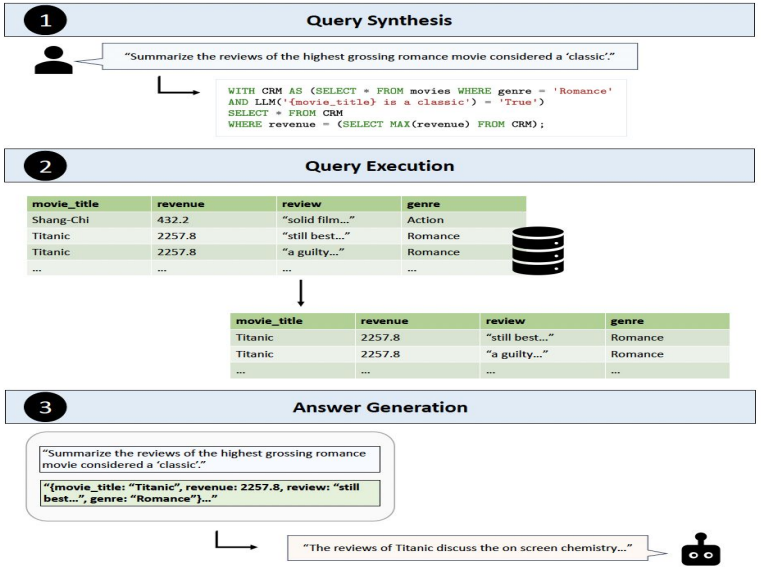
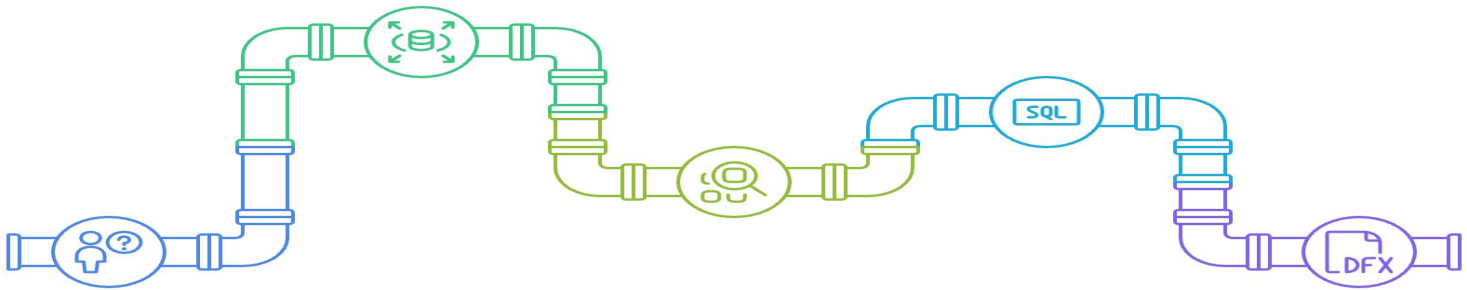


Figure 1: An example TAG implementation for answering the user's natural language question over a table about movies. The TAG pipeline proceeds in three stages: query synthesis, query execution, and answer generation

Query Synthesis Process



01

User Request Identification

Input: the user's natural language request

02

Data Relevance Deduction

Determining which data is relevant to the request

03

Semantic Parsing

Translating the user request into a query

04

SQL Query Generation

Creating an executable SQL query for the database

05

Execution of LM UDFs

Applying language model user-defined functions to execute the query

Query Execution Process

01

Initiate Query Execution

The exec function begins processing the query.

02

Leverage Database Query Engine

The query engine efficiently executes the query on data.

03

Utilize Database API

The API facilitates query execution across systems.

04

Apply LM-based Operators

LM-based operators enhance query execution with additional capabilities.

05

Execute SQL Query

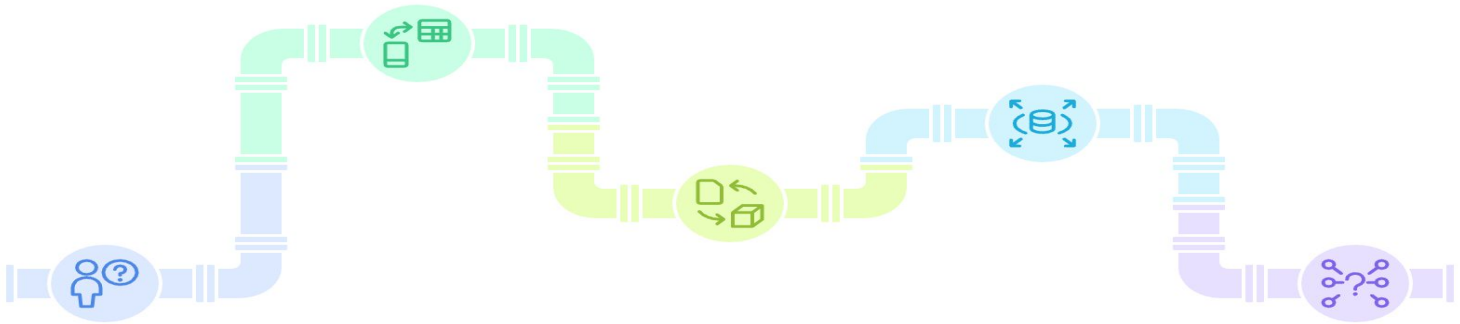
The SQL query is executed to retrieve relevant data.

06

Select and Rank Data

The query selects and ranks data based on specific criteria.

Answer Generation Process in TAG



01

User Request

Input: user's natural language request.

02

Data Encoding

Relevant data is encoded into a string format for processing.

03

Model Input Preparation

The encoded data and user request are prepared as inputs for the language model.

04

Semantic Reasoning Application

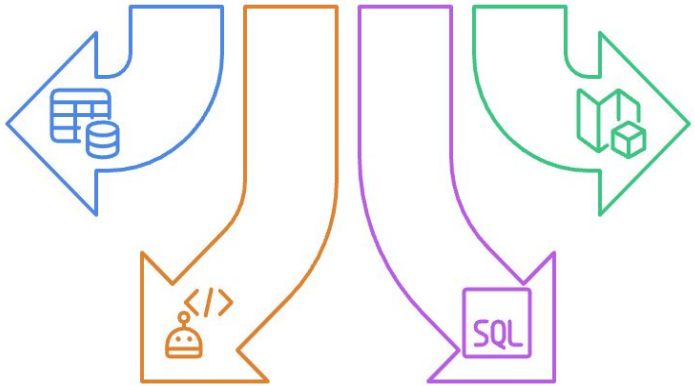
The language model applies semantic reasoning to the inputs.

05

Answer Generation

An answer is generated by the model based on its reasoning.

Which database execution engine and API setting to implement for TAG systems?



SQL-Based Query Engine

Leverages relational data and table schema for data retrieval.

Vector Embedding Retrieval

Transforms queries into embeddings for similarity-based retrieval.

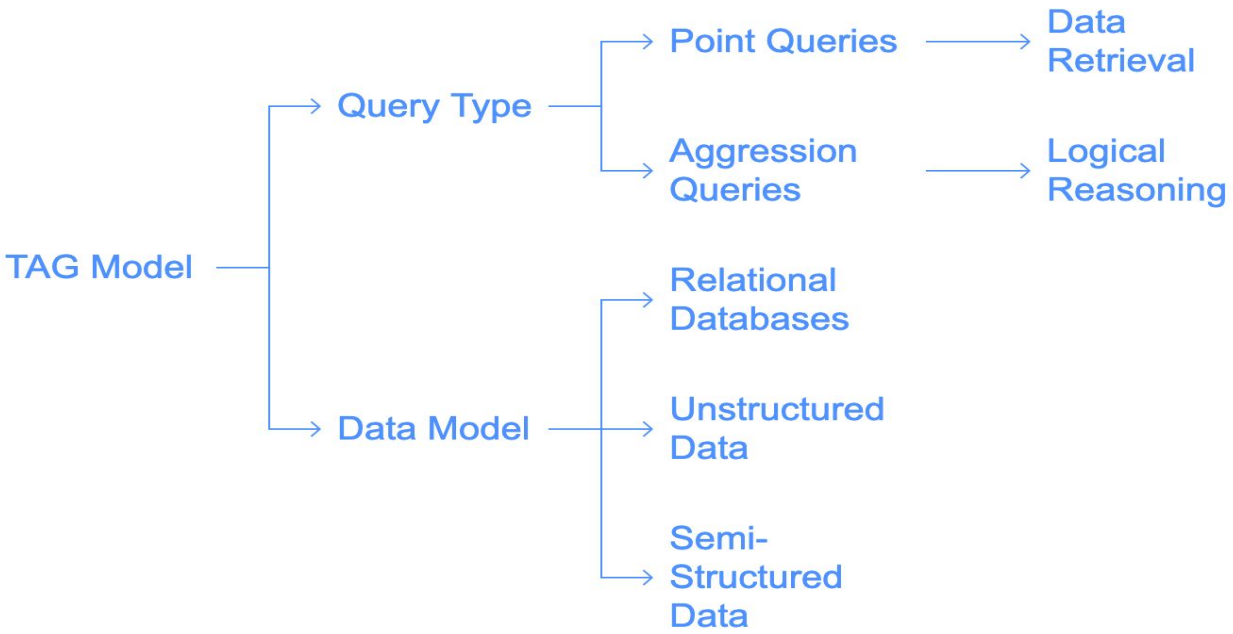
Semantic Operators

Enhances relational models with AI-based operators for dynamic querying.

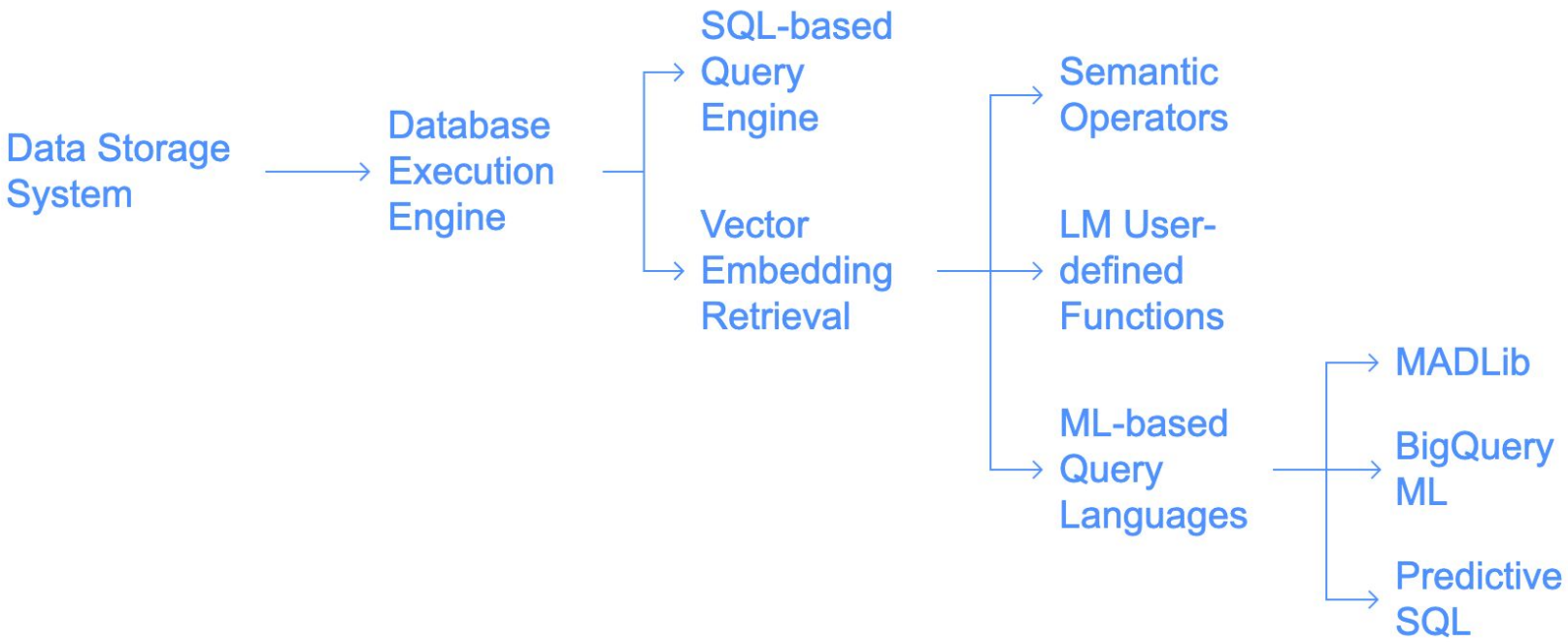
ML-Augmented SQL

Integrates machine learning functions with SQL for advanced data processing.

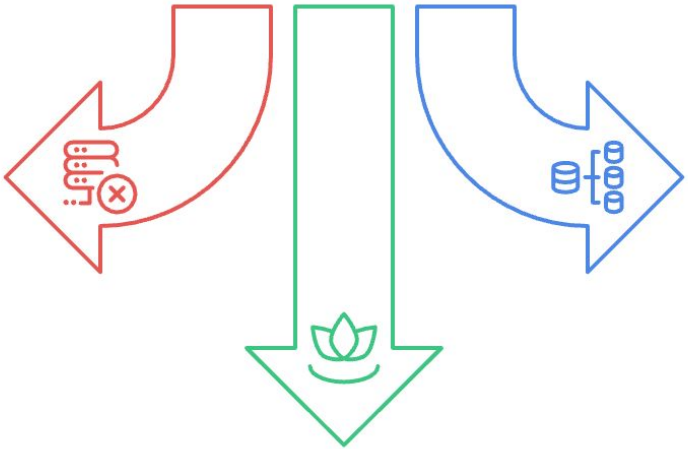
TAG Model Query Processing



TAG Model Database Execution Options



Which LM generation pattern to implement for query response?



Text2SQL

Stops after query execution, lacking final answer generation.

RAG Pipelines

Uses a single LM-call for generation, feeding data in context.

LOTUS

Composes iterative patterns for reasoning-based transformations.

How to evaluate table question answering methods?

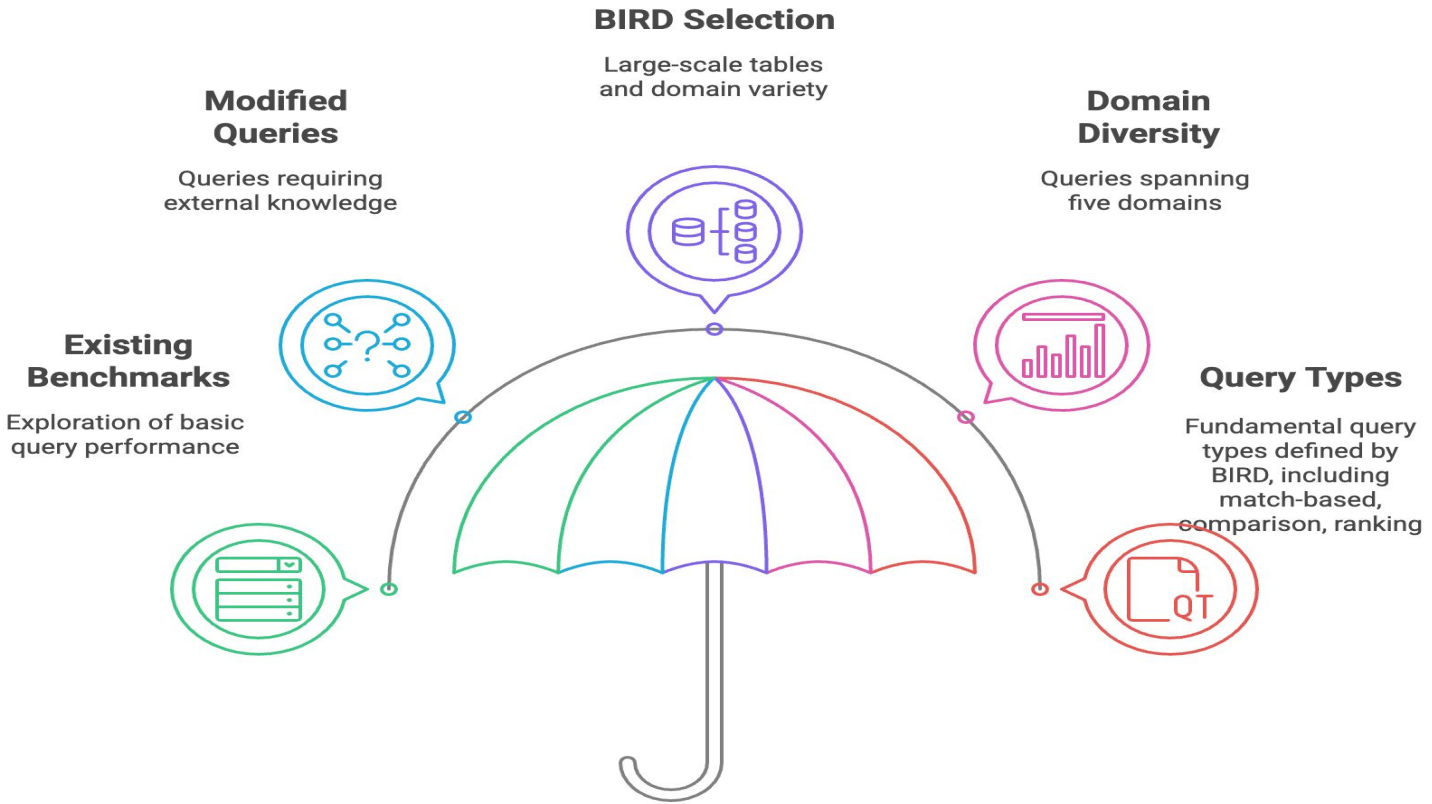
Existing Methods Performance

Assess how well current methods handle semantic reasoning and world knowledge.

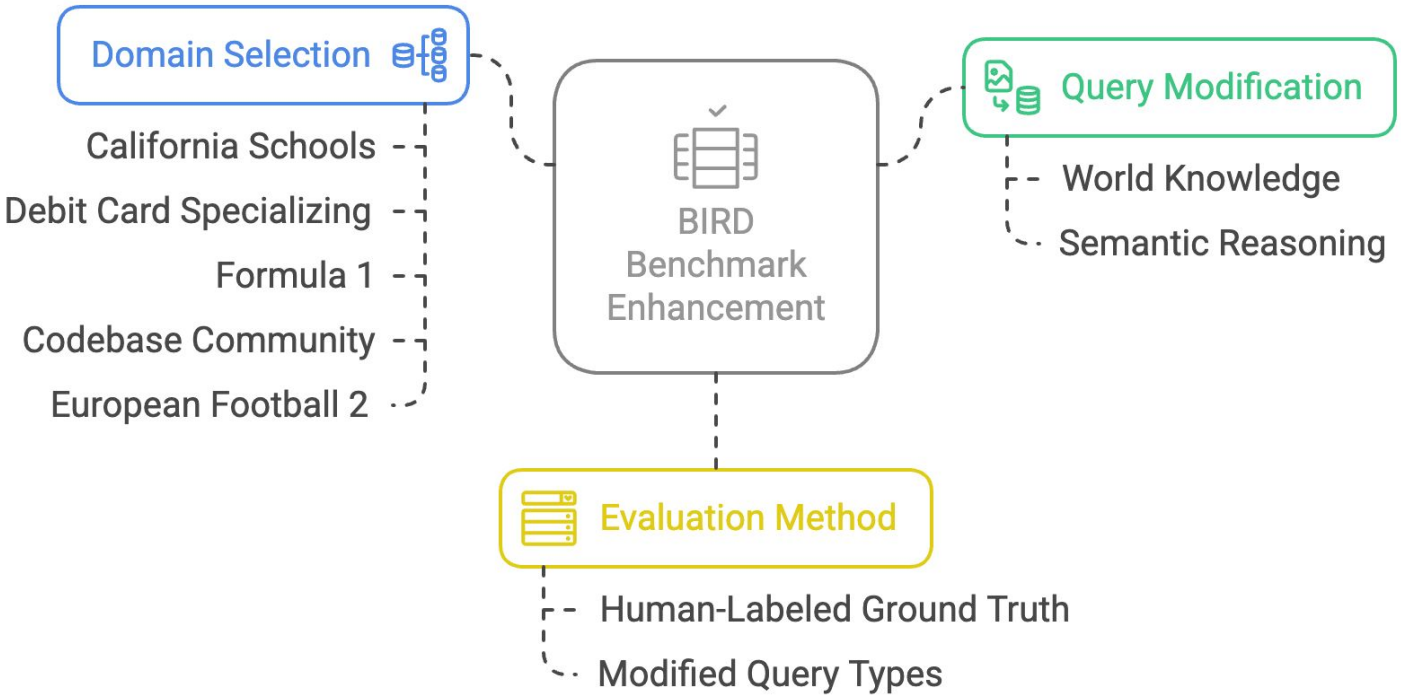
TAG Model Performance

Evaluate the effectiveness of the TAG model's approach to computational and reasoning tasks.

Enhancing Benchmark Queries



Enhancing BIRD Benchmark Queries with Knowledge and Reasoning



Evaluation and Setup Overview

Evaluation Metrics

Measures accuracy and execution time

Model Used

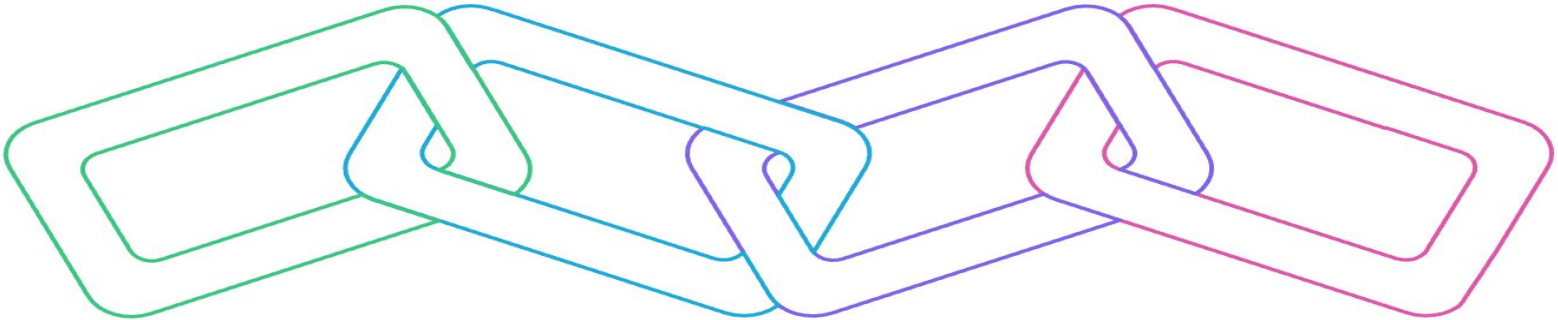
Meta's Llama-3.1 model

Database API

SQLite3 for SQL queries

Hardware

8 A100 80GB GPUs



Baseline Methods

1 Text2SQL

LM generates SQL for queries using table schemas; baseline evaluated by counting SQLite3 execution errors.

2 RAG

Enhancing query retrieval using row-level embeddings and vector similarity search to retrieve 10 relevant rows, with the NL question.

3 Retrieval + LM Rank

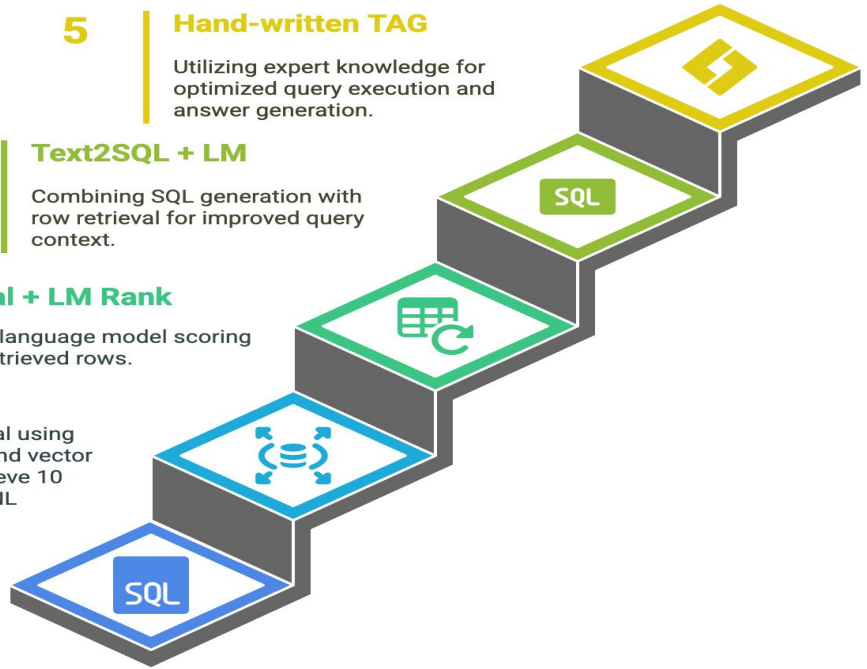
Integrating language model scoring to rerank retrieved rows.

4 Text2SQL + LM

Combining SQL generation with row retrieval for improved query context.

5 Hand-written TAG

Utilizing expert knowledge for optimized query execution and answer generation.



Performance Comparison

Table 1: Accuracy and execution time (ET) for TAG benchmark queries, averaged over all queries and each query type: TAG significantly improves answer quality while achieving the fastest or nearly fastest execution time.

Method	Overall		Match-based		Comparison		Ranking		Aggregation	
	Exact Match ↑[1]	ET (s) ↓	Exact Match ↑	ET (s) ↓	Exact Match ↑	ET (s) ↓	Exact Match ↑	ET (s)	Exact Match ↑	ET (s) ↓
Text2SQL	0.17	5.63	0.20	4.72	0.20	4.01	0.10	7.26	N/A	6.53
RAG	0.00	3.23	0.00	3.73	0.00	**2.29**	0.00	**2.01**	N/A	4.89
Retrieval + LM Rank	0.02	4.82	0.00	6.20	0.05	4.19	0.00	3.42	N/A	5.46
Text2SQL + LM	0.13	9.08	0.10	11.25	0.10	3.89	0.20	11.80	N/A	9.38
Hand-written TAG	**0.55**	**2.94**	**0.60**	**1.70**	**0.65**	5.05	**0.40**	2.50	N/A	**2.50**

[1] Excludes aggregation since accuracy is not measured.

Qualitative Analysis of Aggregation Queries

Query: Provide information about the races held on Sepang International Circuit		
RAG	**Text2SQL + LM**	**Hand-written TAG**
Based on the given data points, the following information is available about the races held on Sepang International Circuit: The races held on this circuit are the Malaysian Grand Prix. The location of the circuit is Kuala Lumpur, Malaysia. The following races were held on this circuit: 2001 Malaysian Grand Prix, 2003 Malaysian Grand Prix, 2002 Malaysian Grand Prix, 2000 Malaysian Grand Prix.	The data points provided do not contain specific information about the races held on Sepang International Circuit. However, based on general knowledge, the Sepang International Circuit is a racing circuit in Sepang, Malaysia, and it has hosted the Malaysian Grand Prix, a Formula One World Championship event, from 1999 to 2017.	The Sepang International Circuit in Kuala Lumpur, Malaysia, hosted the Malaysian Grand Prix from 1999 to 2017. The races were held on the following dates: 1999: October 17 (15th round) 2000: October 22 (2nd round), 2001: March 18 (2nd round), 2002: March 17 (2nd round), 2003: March 23 (2nd round), 2004: March 21 (2nd round), 2005: March 20 (2nd round), ..., 2016: October 2 (16th round), 2017: October 1 (15th round).

Figure 2: Example Aggregation Results: The RAG baseline provides an incomplete answer to the query while Text2SQL + LM fails to answer the question using any data from the DB. The Hand-written TAG baseline provides the most thorough answer, synthesizing data from the DB and its own world knowledge.

Performance for Knowledge/Reasoning based Queries

Table 2: TAG benchmark results averaged over queries requiring Knowledge or Reasoning: TAG performs consistently well with above 50% exact match accuracy on both Knowledge and Reasoning query types.

Method	Knowledge		Reasoning	
	Exact Match ↑	ET (s) ↓	Exact Match ↑	ET (s) ↓
Text2SQL	0.20	5.23	0.10	5.52
RAG	0.00	**2.73**	0.00	2.58
Retrieval + LM Rank	0.03	4.97	0.00	3.87
Text2SQL + LM	0.10	10.27	0.20	6.39
Hand-written TAG	**0.53**	3.50	**0.60**	**2.24**