Text2SQL is Not Enough: Unifying AI and Databases with TAG

Presented by Liana Patel

Joint work with Asim Biswal, Sid Jha, Amog Kamsetty, Shu Liu, Joseph Gonzalez, Carlos Guestrin, and Matei Zaharia

HPTS Gong Show

How should we serve NL questions over large amounts of data?

How should we serve NL questions over large amounts of data?

Text2SQL

RAG

Point Solutions Fit Perfectly in Special Cases

Text2SQL

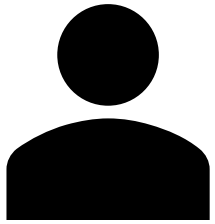
RAG

But in the general case...

Text2SQL

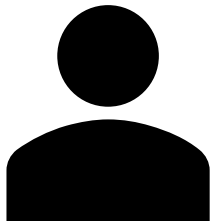
RAG

How should we serve NL questions over large amounts of data?



Conference Name	URL	Agenda	Chairs	Date	City	Country
			...			

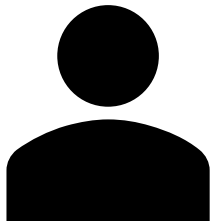
How should we serve NL questions over large amounts of data?



Which paper about vector databases makes the most outrageous claim?

Title	Arxiv URL	Abstract	Authors	H-Index	Date Published	Categories
...						

How should we serve NL questions over large amounts of data?



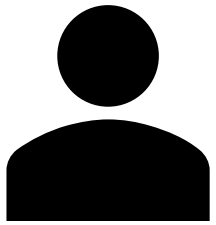
Which paper about vector databases makes the most outrageous claim?

Text2SQL

RAG

Title	Arxiv URL	Abstract	Authors	H-Index	Date Published	Categories
			...			

How should we serve NL questions over large amounts of data?



Which paper about vector databases makes the most outrageous claim?

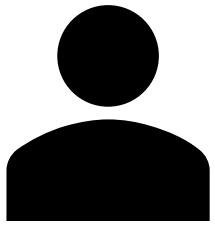
Title	Arxiv URL	Abstract	Authors	H-Index	Date Published	Categories
			...			

Text2SQL

Limited to NL queries that have relational equivalents

RAG

How should we serve NL questions over large amounts of data?



Which paper about vector databases makes the most outrageous claim?

Title	Arxiv URL	Abstract	Authors	H-Index	Date Published	Categories
...						

LM

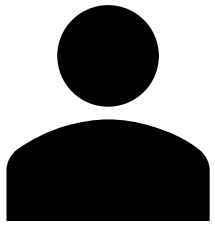
Text2SQL

Limited to NL queries that have relational equivalents

RAG

Good for point lookups, not bulk processing

How should we ask questions over large amounts of data?



How many papers claim to achieve SOTA on BIRD?

Summarize recent papers on vector databases

Which paper about vector databases has the funniest title?

Which papers obtains the highest MMLU score?

Which papers contradict claims made by other papers?

Which papers contradict the claim that LMs can efficiently use long context?

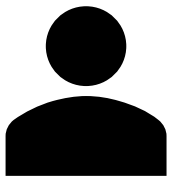
Title	Arxiv URL	Abstract	Authors	H-Index	Date Published	Categories

TAG: Toward a Unified Paradigm…

Table Augmented Generation

The TAG Model: 3 Key Steps

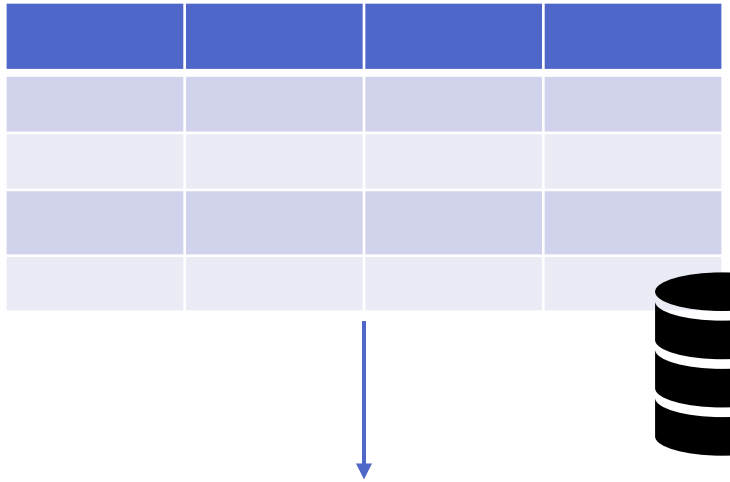
1 Query Synthesis

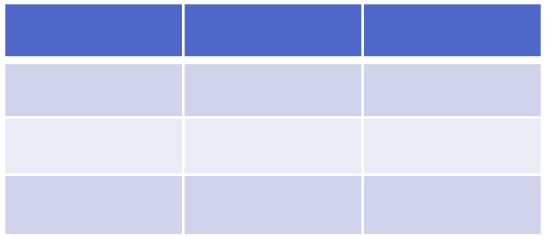


Which paper about vector databases makes the most outrageous claim

SELECT abstract, title **FROM** papers **WHERE SEM-FILTER**("{the {abstract} is about vector databases}

2 Query Execution



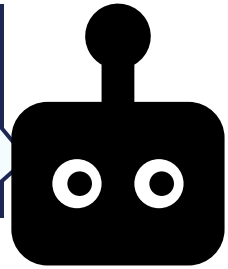


3 Answer Generation

"Which paper about vector databases has the funniest title?'."

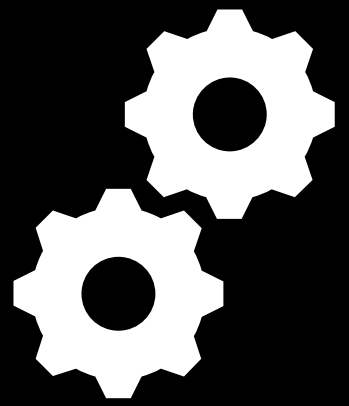
"{title: Efficient and robust approximate nearest neighbor search using HNSW graphs, abstract: We present a new ...}"

"The paper about vector databases with the most outrageous is..."



The TAG Design Space

1. Query Synthesis
2. Query Execution
3. Answer Generation



Many possible:

- Question types
- Data models
- Execution engines & APIs
- LM generation patterns

Measuring Performance with TAG-Bench

Text2SQL

RAG

Handwritten TAG

Programmed with the LOTUS[1] query engine

< 20 % Accuracy

55% Accuracy

[1] https://github.com/TAG-Research/lotus

Summary

- There's **a huge design space** for serving NL questions over data, much more than Text2SQL or RAG
- **TAG is a unified paradigm** that captures under-studied interactions between the LLM and databases
- There's tremendous promise for new TAG implementations!

Try it out! https://github.com/TAG-Research/

- **LOTUS**: declarative programming interface and query engine for AI-based data processing `pip install lotus-ai`
- **TAG-Bench**: an end-to-end benchmark of complex NL questions
- **ACORN**: a SoTA index for supporting vector search with arbitrary relational predicates

Please reach out!

@lianapatel_

lianapat@stanford.edu

Thank you!