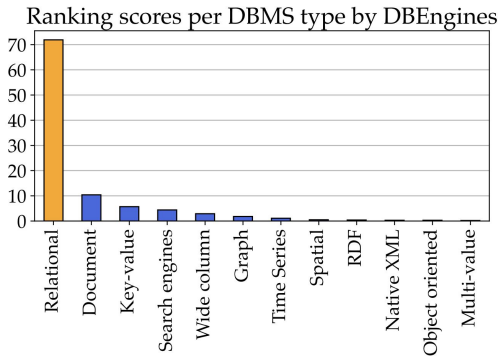
Towards Robust Open-Domain Querying over

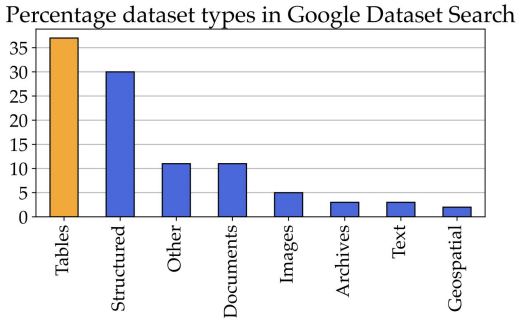
Madelon Hulsebos (CWI)
25 July 2025
SAP Business AI Retreat

Why tables?

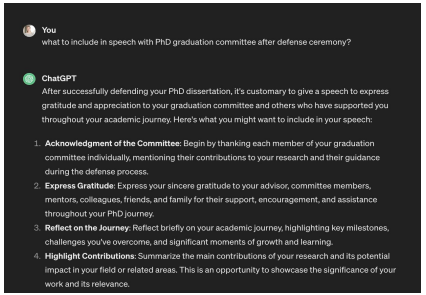
Tabular data:

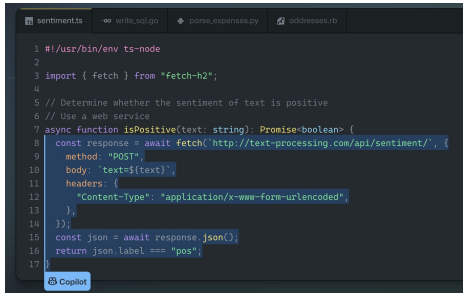
- **Dominant** in data landscape
- **High-value** decisions in enterprise, government, finance, healthcare..
- **Challenging** in structure, relations, size, heterogeneity...





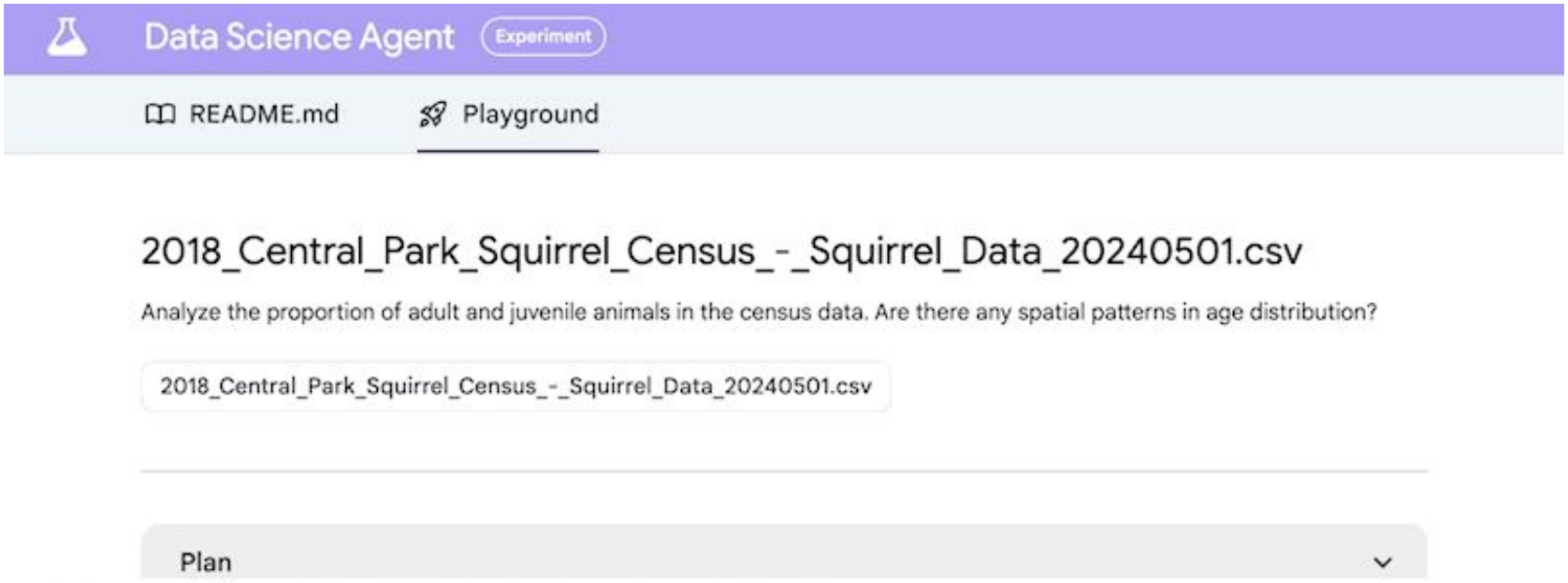
Yet... *neural models* mainly for **text**, **images**, and **code**.





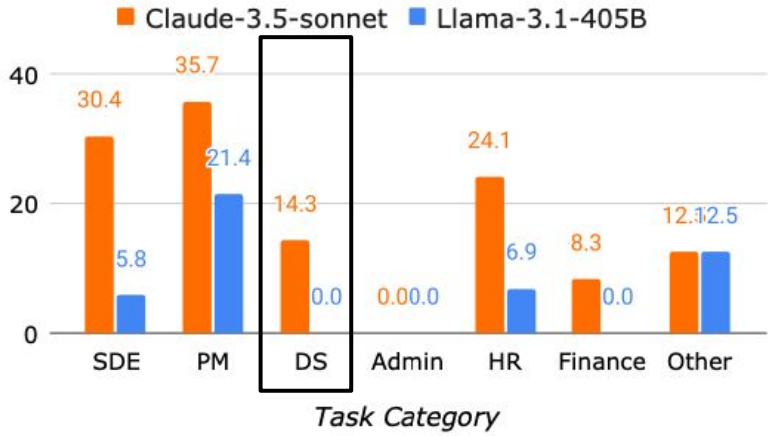
We have LLM Agents now 💫!

Just “Throw” Agents at Tabular Data Science?



Google’s Data Science Agent demo

The Sad State of DS Agents…



(b) Success rate across task categories

Model	*SDE* (69 tasks) Success	*SDE* (69 tasks) Score	*PM* (28 tasks) Success	*PM* (28 tasks) Score	*DS* (14 tasks) Success	*DS* (14 tasks) Score
Closed model APIs						
Claude-3.5-Sonnet	30.43	38.02	35.71	51.31	14.29	21.70
Gemini-2.0-Flash	13.04	18.99	17.86	31.71	0.00	6.49
GPT-4o	13.04	19.18	17.86	32.27	0.00	4.70
Gemini-1.5-Pro	4.35	5.64	3.57	13.19	0.00	4.82
Amazon-Nova-Pro-v1	2.90	6.07	3.57	12.54	0.00	3.27
Open-weight models						
Llama-3.1-405b	5.80	11.33	21.43	35.62	0.00	5.42
Llama-3.3-70b	11.59	16.49	7.14	19.83	0.00	4.70
Qwen-2.5-72b	7.25	11.99	10.71	22.90	0.00	5.42
Llama-3.1-70b	1.45	4.77	3.57	15.16	0.00	5.42
Qwen-2-72b	2.90	3.68	0.00	7.44	0.00	4.70

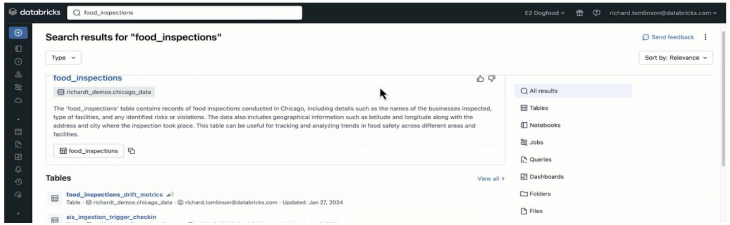
Finding the right data is already hard, *even for humans*

Why Is Dataset *Search* Still So Hard?

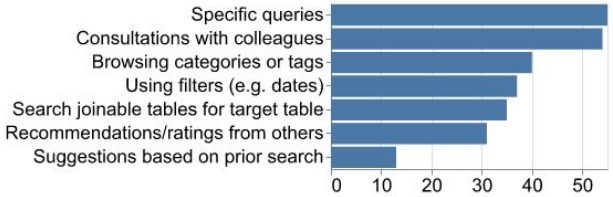
How systems *facilitate* dataset search

`k1, k2, k3`

Keywords that ***perfectly align*** with the dataset ***needed but is unknown***.



How we *actually* search (survey insights)



"*Identify the **problem, and the data for the problem**, … then specific keyword or tag search. Also, identify **people** who have worked on **similar problems**…*"

"*Having so many tables, I ask more **experienced colleagues which ones are most inherent** to the analysis I need to do. I then navigate through the categories and tags to looks for others.*"

How we *want* to search

"*Dataset to <**solve issue of …**> with columns **<1,2,3,...>** on <**granularity desired**>*"

Berkeley UNIVERSITY OF CALIFORNIA CWI

Hulsebos, M., Lin, W., et al. *"'It took longer than I was expecting': Why is dataset search still so hard?"* HILDA@SIGMOD 2024.

Task-driven *Search* Reduces Domain-Expert Reliance

Task query Q, instead of keywords:

"Dataset to <train an **ML model** to **forecast demand** for **drug types** across **suppliers**>, …"

Search w/ Hypothetical Schema Embeddings (hyse):

1) generate `hypothetical schema` for task Q:

```
medication table: medication id, medication name, …
sales table: medication id, supplier id, date, quantity, …
```

2) embed `hypothetical schema`

3) retrieve relevant tables from retrieval corpus based on embedding similarity

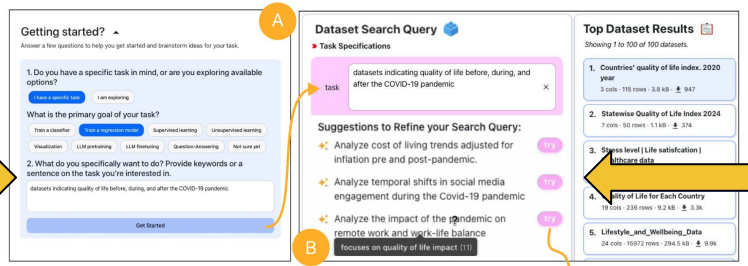
Berkeley UNIVERSITY OF CALIFORNIA CWI

Preliminary results are promising!

Lin, W., Parameswaran, A. Hulsebos, M., "*Task-driven Dataset Search through Hypothetical Schema Embeddings*", WIP, 2025.

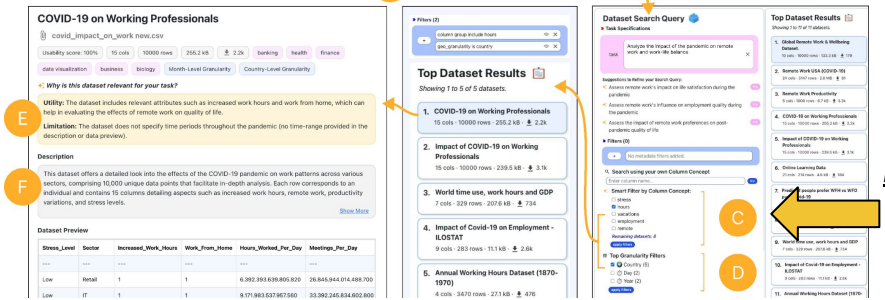
Proactive Assistance for Dataset *Search* is Needed

Task-driven HySE query



Result-driven query suggestions

Task-driven dataset relevance



Data-driven column suggestions

Proactive assistance makes it **easier** to find **more relevant** results, with **higher success rate**.

Condition	Ease-of-use	Relevance	# Successes
(A) Kaggle	μ=3.08; σ=0.51	μ=3.25; σ=1.05	7 *of* 12
(B) Semantic Baseline	μ=3.75; σ=0.45	μ=3.25; σ=0.86	6 *of* 12
(C) DataScout	μ=4.75; σ=0.45	μ=3.67; σ=0.78	10 *of* 12

Berkeley UNIVERSITY OF CALIFORNIA CWI

Lin, R., Chopra, B., Lin, W., Shankar, S., Hulsebos, M., Parameswaran, A. "*Rethinking Dataset Discovery with DataScout*" UIST 2025.

Beyond *Search*: **Retrieval** for Open-Domain Table QA

Asking domain-specific questions… to LLMs?

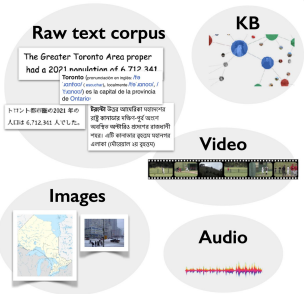
Question

What is the highest eligible free rate for K–12 students in the schools in Alameda County?

LLM response

"….To determine the highest free rate specifically in Alameda County schools, you'd generally need data from specific school districts or schools in the area,…."

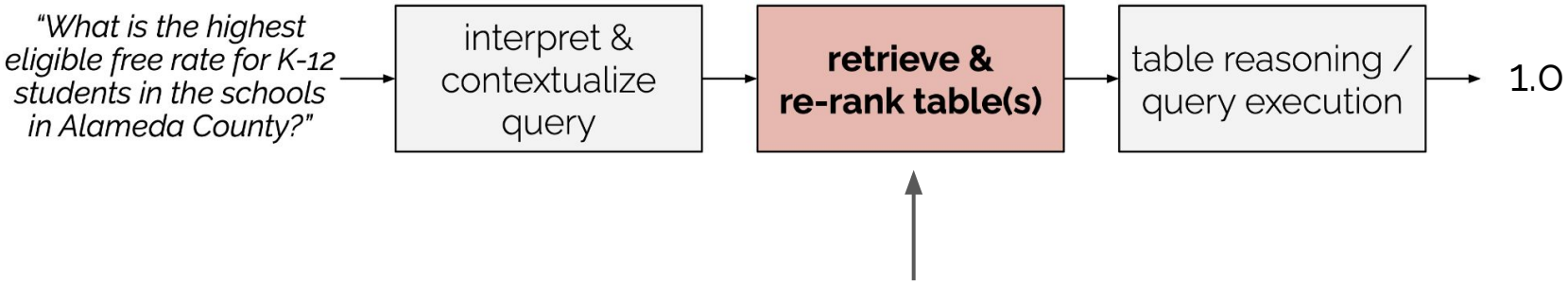
RAG 🎉



What about tables?

*Generated with ChatGPT on 6 October 2024

Open-Domain Question Answering over Tables



Berkeley UNIVERSITY OF CALIFORNIA CWI

https://target-benchmark.github.io (pip install target_benchmark)

Ji, Parameswaran, Hulsebos. "*TARGET: benchmarking Table Retrieval for Generative Tasks,* " TRL@NeurIPS, 2024

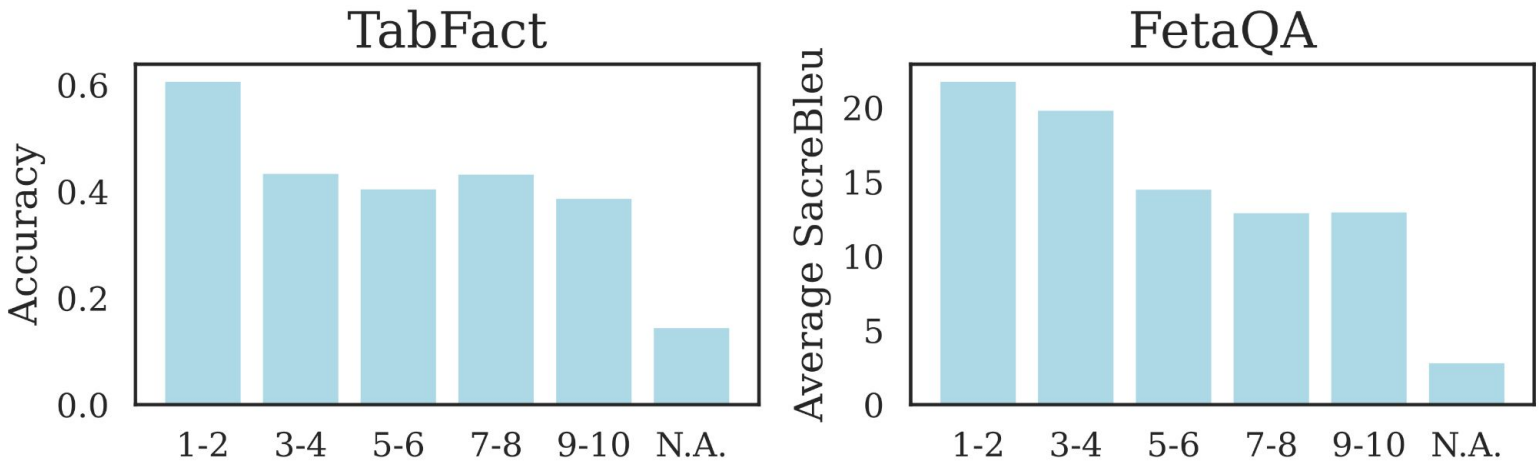
Table Retrieval in Open-Domain QA: Unsolved Problem

	Question Answering				Fact Verification		Text-to-SQL			
	OTTQA		FeTaQA		TabFact		Spider		BIRD	
Method	R@10	time (s)	R@10	time (s)	R@10	s	CR@10	time (s)	CR@10	time (s)
Sparse Lexical Repr. (BM25)	**0.967**	0.001	0.082	0.001	0.338	0.001	0.544	0.001	0.700	0.001
w/o table title	0.592	0.001	0.084	0.001	0.331	0.001	0.491	0.001	0.616	0.001
Sparse Lexical Repr. (TF-IDF)	0.963	0.001	0.083	0.001	0.336	0.001	0.541	0.001	0.586	0.001
w/o table title	0.583	0.001	0.039	0.001	0.322	0.001	0.489	0.001	0.613	0.001
Dense Metadata Embedding	0.820	0.297	0.436	0.396	0.469	0.354	0.621	0.024	0.940	0.014
Dense Table Embedding	0.963	0.001	**0.741**	0.001	0.824	0.001	0.657	0.001	**0.961**	0.003
column names only	0.658	0.001	0.208	0.001	0.506	0.001	0.648	0.001	0.932	0.003
Dense Row-level Embedding	0.951	0.267	0.711	0.394	**0.848**	0.384	**0.665**	6.077	N/A	N/A

- BM25/TF-IDF **less effective than for text**, only works with descriptive table name.
- Generating summary/metadata can help, but **not all tables easy to LLM-summarize**.
- **Row-level retrieval generally most effective**, but not feasible in practice (large tables)

Ji, Parameswaran, Hulsebos. "*TARGET: benchmarking Table Retrieval for Generative Tasks*," TRL@NeurIPS, 2024

Retrieval? Isn't large context all you need?

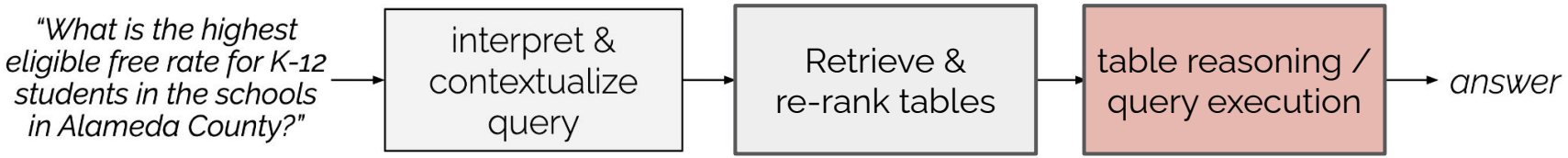


No.

Much work to be done... stay tuned!

Beyond *Search*: **Reasoning** for Open-Domain Table QA

Open-Domain **Querying** over Tables



Typical downstream tasks:

- SQL generation (QA),
- Tabular LLM reasoning (QA),
- Fact verification

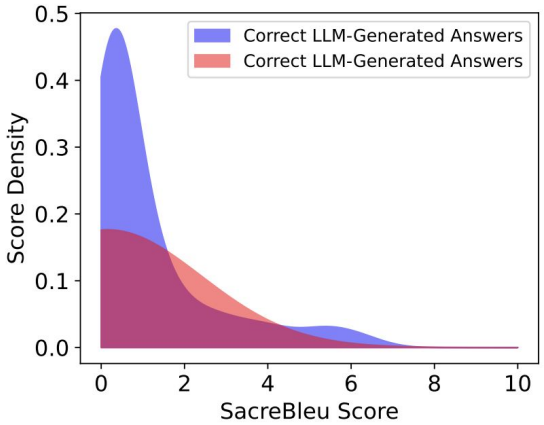
SQL is great, but reality is… LLMs are used, as well

First, we need to use proper metrics!

Problem: We have a GT ***value*** to compare with **free-form text**

Question: *How long has the team that qualified for the 1999 FIFA Confederations Cup in June existed?*				
Generated answer	**Answer**	**SB score**	**BS score**	**LLM**
The team that qualified for the 1999 FIFA Confederations Cup in June was Germany, who secured their qualification on 30 June 1996. To determine how long they had existed as a team by June 1999: The Germany national football team was officially founded in 1908. From 1908 to 1999 is 91 years. Answer: The Germany national team had existed for 91 years by June 1999.	since 1908	0.5	0.81	
The team that qualified for the 1999 FIFA Confederations Cup in June was Germany. Germany has existed as a national team since 1974, so by June 1999, they had existed for 25 years.	since 1908	1.0	0.82	

Metrics distribution SacreBleu



Scores of incorrect answers **inseparable** from correct answers

Wolff, C., Hulsebos, M., "How well do LLMs reason over tabular data, really?" TRL@ACL 2025

So, can LLMs reason over tables?

Evaluation of table **lookup queries** with Qwen2.5:

- Multiple-choice eval (A/B/C): **86%**
- **LLM-as-a-judge: 8.1% ← realistic performance**

Reasoning with DeepSeek:

"Wait no—the data doesn't show that... Wait I'm getting confused."

In an ideal relational world:

A *relation* then consists of a set of tuples, each tuple having the same set of attributes. If the domains are all simple, such a relation has a tabular representation with the following properties.

(1) There is no duplication of rows (tuples).
(2) Row order is insignificant.
(3) Column (attribute) order is insignificant.
(4) All table entries are atomic values.

For "average" queries

Model	Accuracy as-is	With duplicates
qwen2.5	36%	**20%**

Data cleaning is relevant beyond predictive tabular ML!

Wolff, C., Hulsebos, M., "How well do LLMs reason over tabular data, really?" TRL@ACL 2025

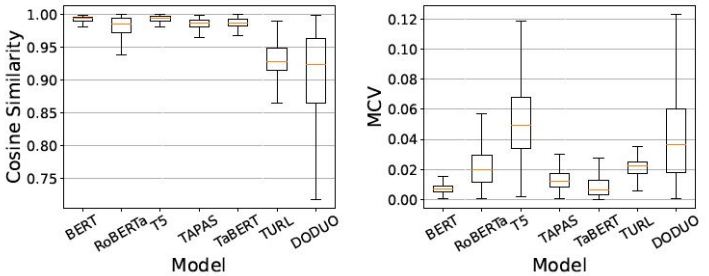
Do Table Embeddings Capture Relational Properties?

Studying neural **table embeddings** through **Codd's relational data model**:

A *relation* then consists of a set of tuples, each tuple having the same set of attributes. If the domains are all simple, such a relation has a tabular representation with the following properties.

(1) There is no duplication of rows (tuples).
(2) Row order is insignificant.
(3) Column (attribute) order is insignificant.
(4) All table entries are atomic values.

Measure by avg cosine similarity of col embeddings across row permutations.



row order robustness

Impact: shuffling **rows affects 34%** predicted semantic column types!

University of Amsterdam
University of Michigan

Teach LLMs the relational model but anticipate real-world messiness!

Cong, T., Hulsebos, M., Sun, Z., Groth, P., & Jagadish, H. V. "Observatory: Characterizing Embeddings of Relational Tables", VLDB 2023.

How to move forward?

Some thoughts:

- Can we just throw LLM agents to solve data science?

  No, table-tuned and table-native models are key.

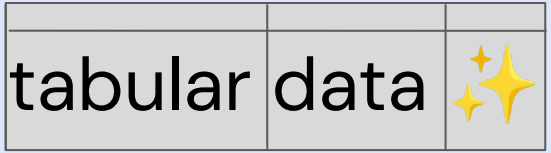
- Do we need a single (real) tabular foundation model, that does *everything*?

  Imho, no.

- What are key open challenges?

  Table-native models, retrieval, open-domain *text-to-sql,* unifying pred & reas, *efficient* pred TFMs, integrating knowledge in DS.

Towards Robust Open-Domain Querying over



madelon@cwi.nl!
trl-lab.github.io