

Towards Robust Open-Domain Querying over

tabular	data	✨

Madelon Hulsebos (CWI)

25 July 2025

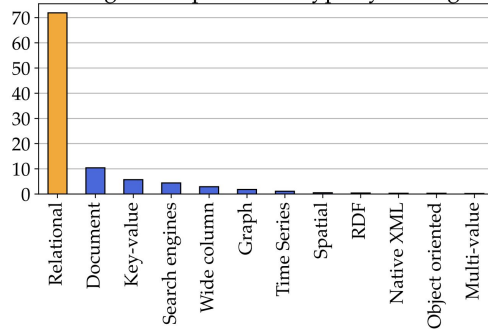
SAP Business AI Retreat

Why tables?

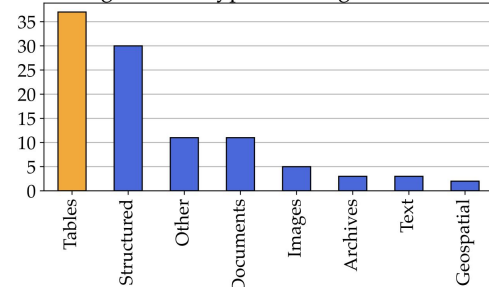
Tabular data:

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

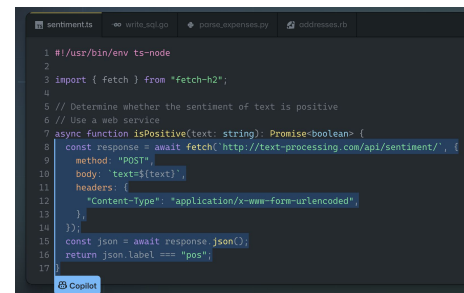
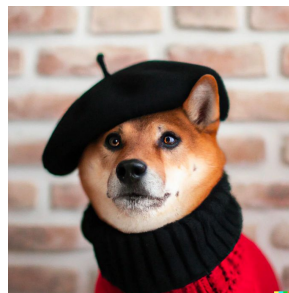
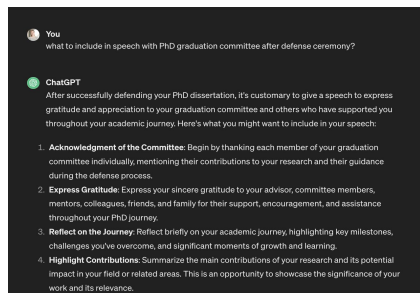
Ranking scores per DBMS type by DBEngines



Percentage dataset types in Google Dataset Search

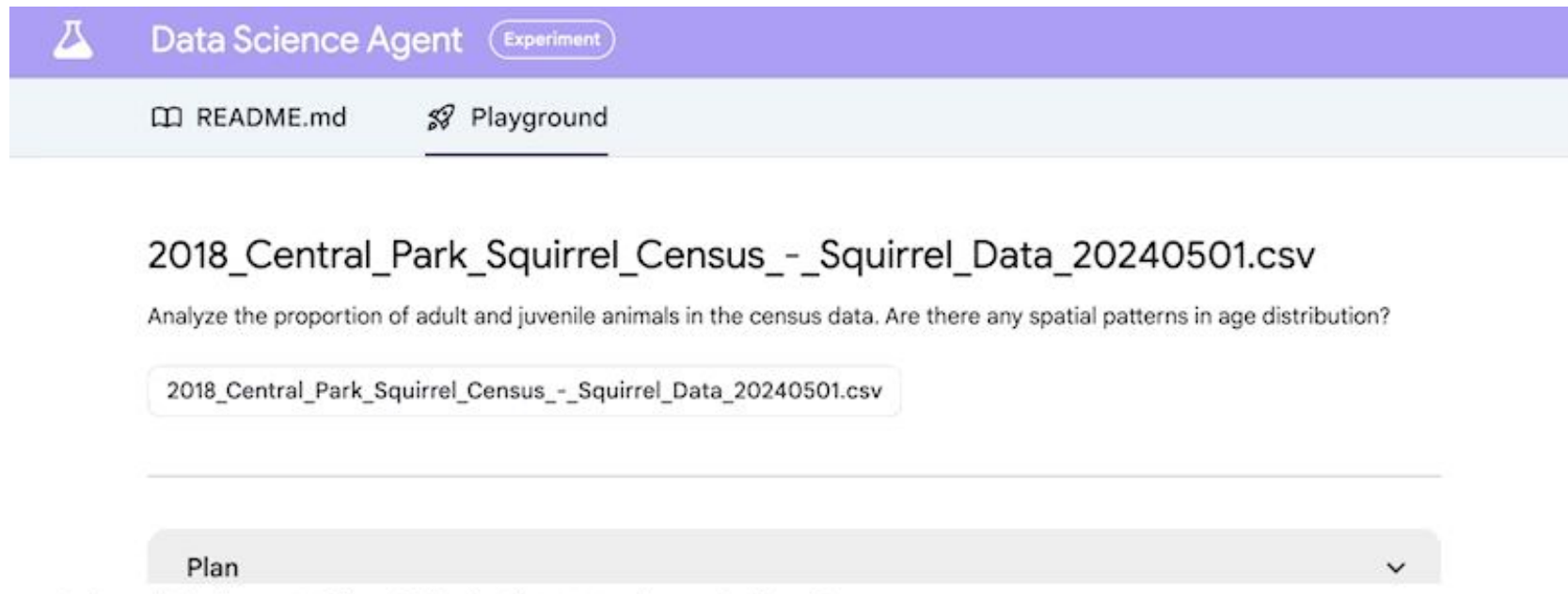


Yet... *neural models* mainly for **text**, **images**, and **code**.



We have LLM Agents now 🌟!

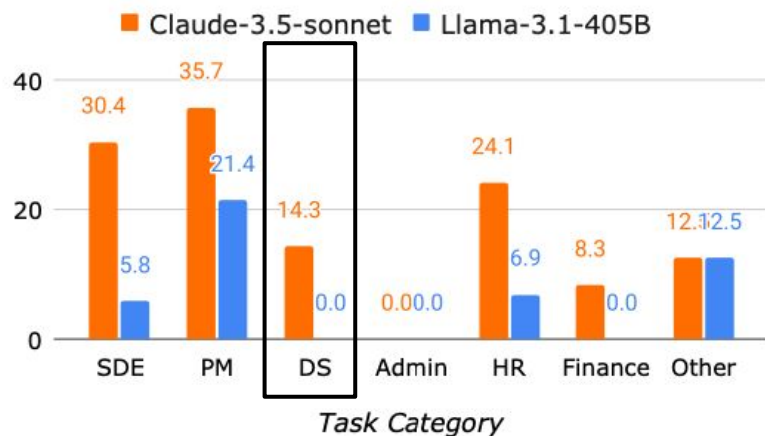
Just “Throw” Agents at Tabular Data Science?



The screenshot displays the Google Data Science Agent interface. At the top, there is a purple header with a flask icon, the text "Data Science Agent", and a button labeled "Experiment". Below this is a light blue navigation bar with a book icon and "README.md" on the left, and a rocket icon and "Playground" on the right, which is underlined. The main area shows a task title "2018_Central_Park_Squirrel_Census_-_Squirrel_Data_20240501.csv" followed by a description: "Analyze the proportion of adult and juvenile animals in the census data. Are there any spatial patterns in age distribution?". Below the description is a text input field containing the same filename. At the bottom, there is a grey bar with the word "Plan" and a downward arrow.

Google's Data Science Agent demo

The Sad State of DS Agents...



(b) Success rate across task categories

Model	SDE (69 tasks)		PM (28 tasks)		DS (14 tasks)	
	Success	Score	Success	Score	Success	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

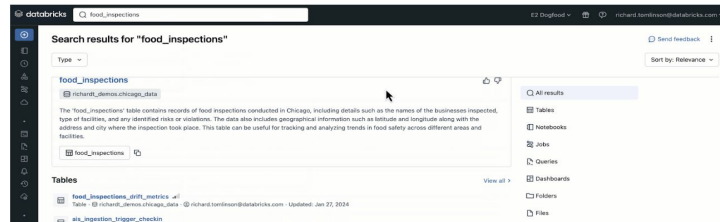
Finding the right data is already hard, *even for humans*

Why Is Dataset Search Still So Hard?

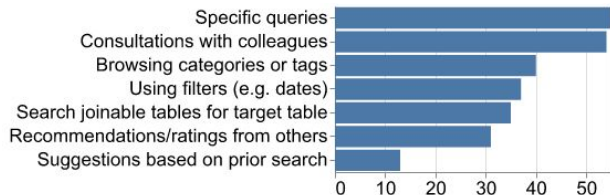
How systems facilitate dataset search

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 look for others."

How we want to search

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

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

Proactive Assistance for Dataset Search is Needed

Task-driven
HySE query

Getting started?

Answer a few questions to help you get started and brainstorm ideas for your task.

1. Do you have a specific task in mind, or are you exploring available options?

[I have a specific task](#) [I am exploring](#)

What is the primary goal of your task?

Train a classifier [Add a supervised feature](#) [Supervised learning](#) [Unsupervised learning](#)

Visualization [LLM prompting](#) [LLM fine-tuning](#) [Question-Answering](#) [Text sum-ppt](#)

2. What do you specifically want to do? Provide keywords or a sentence on the task you're interested in.

datasets indicating quality of life before, during, and after the COVID-19 pandemic

[Get Started](#)

Dataset Search Query

Task Specifications

task datasets indicating quality of life before, during, and after the COVID-19 pandemic

Suggestions to Refine your Search Query:

- Analyze cost of living trends adjusted for inflation pre and post-pandemic.
- Analyze temporal shifts in social media engagement during the Covid-19 pandemic
- Analyze the impact of the pandemic on remote work and work-life balance focuses on quality of life impact [11]

Top Dataset Results

Showing 7 to 100 of 100 datasets

- Countries' quality of life index, 2020 year 8 cols · 115 rows · 3.8 kB · 947
- Statewise Quality of Life Index 2024 7 cols · 520 rows · 11.1 kB · 516
- Stress level | Life satisfaction | Healthcare data
- Quality of Life for Each Country 19 cols · 230 rows · 9.2 kB · 536
- Lifestyle_and_Wellbeing_Data 24 cols · 10972 rows · 204.5 kB · 936

Result-driven
query suggestions

Task-driven
dataset relevance

COVID-19 on Working Professionals

[covid_impact_on_working_profs.csv](#)

quality score: 100% 10 cols · 10000 rows · 255.2 kB · 2.2k [banking](#) [health](#) [finance](#)

data classification: [Business](#) [Science](#) [Health Level Granularity](#) [Country Level Granularity](#)

Why is this dataset relevant for your task?

Utility: This dataset includes relevant attributes such as increased work hours and work from home, which can help in evaluating the effects of remote work on quality of life.

Limitation: The dataset does not specify time periods throughout the pandemic (no time-range provided in the description or data preview).

Description

This dataset offers a detailed look into the effects of the COVID-19 pandemic on work patterns across various sectors, comprising 10,000 unique data points that facilitate in-depth analysis. Each row corresponds to an individual and contains 10 columns detailing aspects such as increased work hours, remote work, productivity variations, and stress levels.

[Show More](#)

Dataset Preview

Stress_Level	Sector	Increased_Work_Hours	Work_From_Home	Hours_Worked_Per_Day	Meetings_Per_Day
Low	Retail	1	1	6.202.393.839.805.820	26.845.944.014.688.700
Low	IT	1	1	6.171.963.637.907.660	33.392.345.834.802.800

Dataset Search Query

Task Specifications

task Analyze the impact of the pandemic on remote work and work-life balance

Suggestions to Refine your Search Query:

- Analyze remote work's influence on employment quality during the pandemic
- Analyze the impact of remote work preferences on post-pandemic quality of life

Filters (0)

1. Search using your own Column Concept

2. Search Filter by Column Concept

3. Filter by

4. Filter by

5. Filter by

6. Filter by

7. Filter by

8. Filter by

9. Filter by

10. Filter by

11. Filter by

12. Filter by

13. Filter by

14. Filter by

15. Filter by

16. Filter by

17. Filter by

18. Filter by

19. Filter by

20. Filter by

21. Filter by

22. Filter by

23. Filter by

24. Filter by

25. Filter by

26. Filter by

27. Filter by

28. Filter by

29. Filter by

30. Filter by

31. Filter by

32. Filter by

33. Filter by

34. Filter by

35. Filter by

36. Filter by

37. Filter by

38. Filter by

39. Filter by

40. Filter by

41. Filter by

42. Filter by

43. Filter by

44. Filter by

45. Filter by

46. Filter by

47. Filter by

48. Filter by

49. Filter by

50. Filter by

51. Filter by

52. Filter by

53. Filter by

54. Filter by

55. Filter by

56. Filter by

57. Filter by

58. Filter by

59. Filter by

60. Filter by

61. Filter by

62. Filter by

63. Filter by

64. Filter by

65. Filter by

66. Filter by

67. Filter by

68. Filter by

69. Filter by

70. Filter by

71. Filter by

72. Filter by

73. Filter by

74. Filter by

75. Filter by

76. Filter by

77. Filter by

78. Filter by

79. Filter by

80. Filter by

81. Filter by

82. Filter by

83. Filter by

84. Filter by

85. Filter by

86. Filter by

87. Filter by

88. Filter by

89. Filter by

90. Filter by

91. Filter by

92. Filter by

93. Filter by

94. Filter by

95. Filter by

96. Filter by

97. Filter by

98. Filter by

99. Filter by

100. Filter by

Top Dataset Results

Showing 1 to 5 of 5 datasets

- COVID-19 on Working Professionals 10 cols · 10000 rows · 255.2 kB · 2.2k
- Impact of COVID-19 on Working Professionals 10 cols · 10000 rows · 238.8 kB · 5.1k
- World time use, work hours and GDP 7 cols · 329 rows · 207.6 kB · 734
- Impact of Covid-19 on Employment - ILOSTAT 9 cols · 283 rows · 11.1 kB · 2.6k
- Annual Working Hours Dataset (1870-1970) 4 cols · 3470 rows · 271.1 kB · 476

Data-driven
column suggestions

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

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

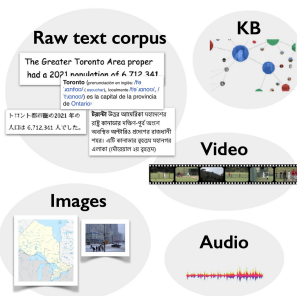
Asking domain-specific questions... to LLMs?

Question

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



RAG



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,..."

What about tables?

Open-Domain Question Answering over Tables

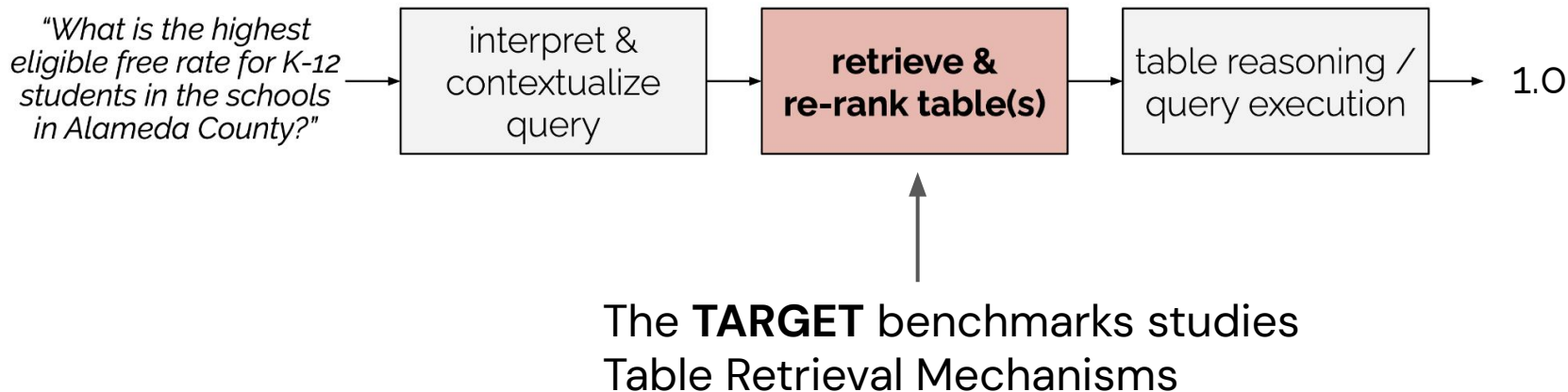
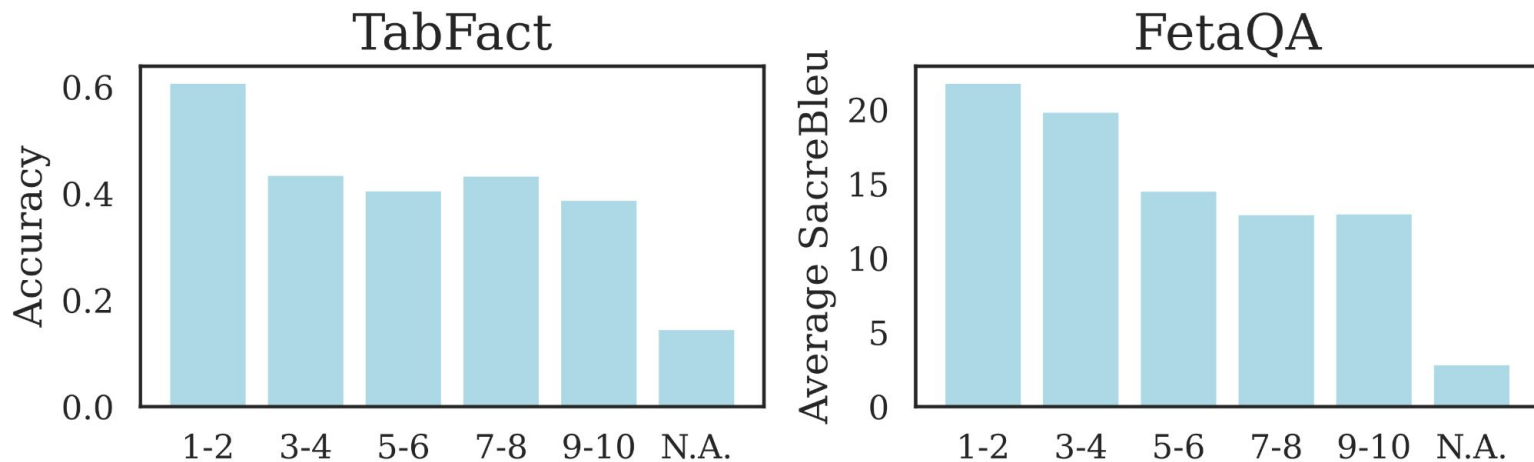


Table Retrieval in Open-Domain QA: Unsolved Problem

Method	Question Answering				Fact Verification		Text-to-SQL			
	OTTQA		FeTaQA		TabFact		Spider		BIRD	
	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
<i>w/o table title</i>	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
<i>w/o table title</i>	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
<i>column names only</i>	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)

Retrieval? Isn't large context all you need?

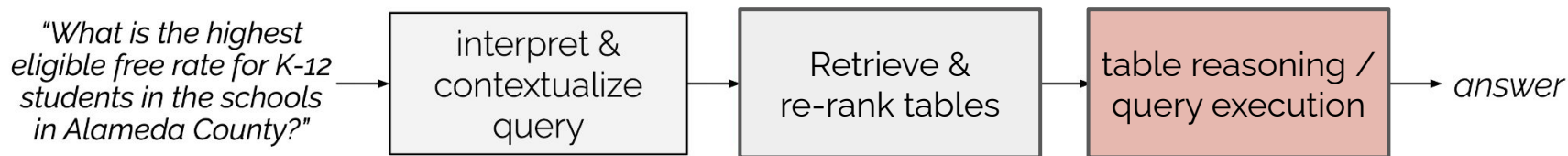


No.

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

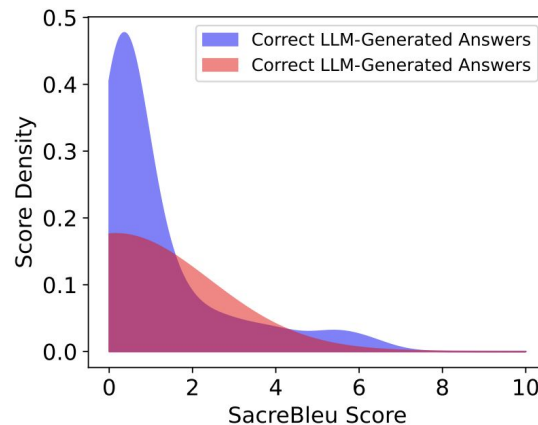
First, we need to use proper metrics!

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

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!

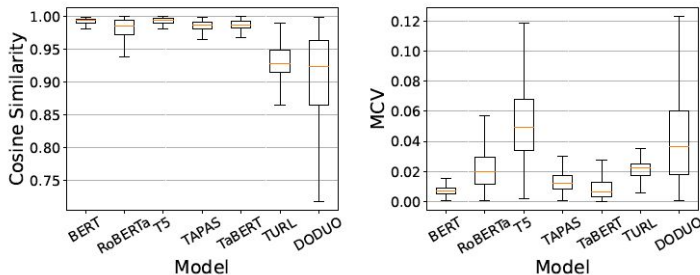
Do Table Embeddings Capture Relational Properties?

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



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

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

tabular	data	✨
---------	------	---

madelon@cw.nl!
[trl-lab.github.io](https://github.com/trl-lab)
