On Different Flavors hence Needs for Tabular Reasoning

Madelon Hulsebos

26 June 2026
SAP AI Research Retreat

Back to the Basics: *Semantic* Reasoning

What is this table about?

crop rotation : Tabelle

Nr	ID	seed rate	yield	crop	cultivar	pre crop	pre-pre crop	pre-pre-pre	soil type	precipita	tempera	comment
1	68		91	winter wheat		sugar beets	beans		sandy loam, loe	636	9,6	wb, sg,
2	68		100	winter wheat		sugar beets	rotation fallow		sandy loam, loe	636	9,6	cultivation
3	68		97	winter wheat		sugar beets	fallow land (5,5y)		sandy loam, loe	636	9,6	1993-1996
4	136		95	winter wheat		oats	sugar beets		sandy loam, loe	636	9,6	
5	136		96	winter wheat		potatos	sugar beets		sandy loam, loe	636	9,5	cultivation
6	136		107	winter wheat		sugar beets	maize		sandy loam, loe	636	9,5	1991-1994
7	136		107	winter wheat		sugar beetsn	summer wheat	maize	sandy loam, loe	636	9,5	
8	136		82	winter wheat		oats	sugar beets	sugar beets	sandy loam, loe	636	9,5	organic
9	136		77	winter wheat		potatos	sugar beets		sandy loam, loe	636	9,5	organic
10	136		85	winter wheat		sugar beets	maize	maize	sandy loam, loe	636	9,5	organic
11	136		84	winter wheat		sugar beets	summer wheat	sugar beets	sandy loam, loe	636	9,5	organic
12	57	371	98	winter wheat	Sperbe	sugar beets	winter barley	winter wheat	sandy loam, loe	635		wb, ww
13	57	365	98	winter wheat	Sperbe	potatos	sugar beets	summer barl	sandy loam, loe	635		cultivation, weed
14	57	365	105	winter wheat	Sperbe	sugar beets	maize	maize	sandy loam, loe	635		1987-1992
15	57	365	97	winter wheat	Sperbe	sugar beets	winter wheat	sugar beets	sandy loam, loe	635		
16	39	433	90	winter wheat	Okapi	summer barley			sandy loam, loe	690	8,5	oats, cultivation, weec
17	39	433	100	winter wheat	Okapi	oats			clay, silt	690	8,5	1982-1986
18	39	433	97	winter wheat	Okapi	winter wheat			clay, silt	690	8,5	

Pre crop?
Pre-pre crop?
Pre-pre-pre?

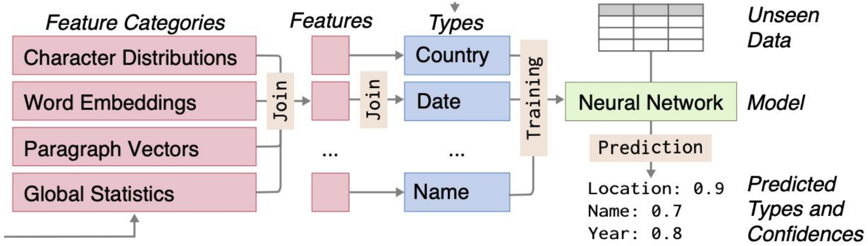
Semantics key in downstream tasks (data cleaning, integration, predictive ML)

Sherlock: A Deep Learning Approach to Semantic Data Type Detection by Hulsebos et al., KDD, 2019

Can we infer *column types* through embeddings?

Sherlock: neural model w/ thousands of params trained on column representations

78 semantic types (`name`, `address`, etc)



Performance (F1)

Sherlock	0.89
Human Consensus	0.32

🐘 **LLMs?** Sherlock (*tiny* model) outperformed (*huge*) GPT-3.5 with +6%

Sherlock: A Deep Learning Approach to Semantic Data Type Detection by Hulsebos et al., KDD, 2019

Why care about semantics: Sensitive Data Detection

Sherlock classifies PII types (e.g. `name`, `address`) with confidence!

🤔 Q1: are **all** PII typed columns, sensitive? (current practice)

personal `address` of human aid workers

Name	Role	Address	Region
Lina Haddad	Field Coordinator	14 Olive Grove Rd, Sector 3, Al Noor District	Northern Syria
David Mensah	Medical Logistics Officer	House 8, Kivu Relief Compound, East Ridge	Eastern DRC
Sara Petrov	Emergency Specialist	22 Cedar Lane, Block B, River Quarter	South Sudan

organizational `address` of professors

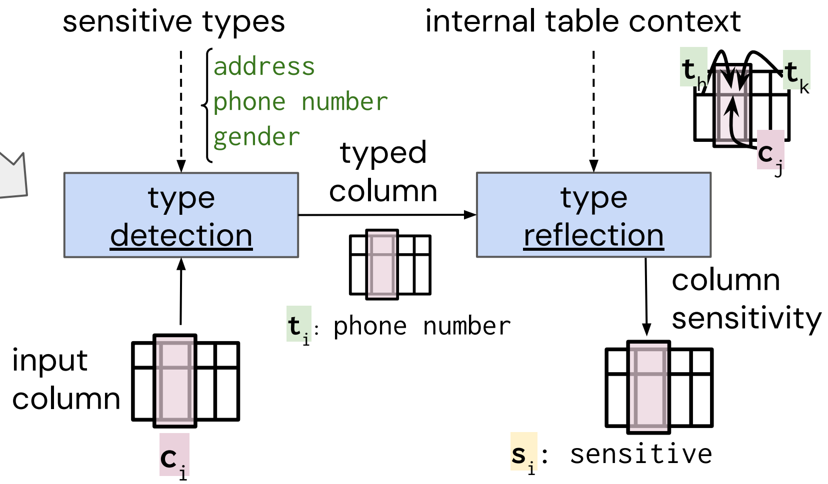
Name	Department	Address	Institution
Dr. Elena Vos	Sociology	Room 4.12, Social Sciences Building, 120 University Ave	Utrecht University
Prof. Michael Chen	Physics	Office 2-318, Science Hall, 85 Academic Way	TU Delft
Dr. Amina Rahman	History	Room 5B, Humanities Centre, 44 College Street	Leiden University

Towards Contextual Sensitive Data Detection, Telkamp & Hulsebos, Findings of EMNLP, 2026.

CWI

Mechanism: *type-contextualization*

w/o PII types:
recall **-5%**



✨**integrated** in UN's data platform

🏆**Amsterdam AI Thesis award**

94% sensitive PII detected
vs 63% with Google's DLP!

Towards Contextual Sensitive Data Detection, Telkamp & Hulsebos, Findings of EMNLP, 2026.

External Semantics: dataset context matters

🤔 Q2: are **only** PII typed columns, sensitive?

`geo-coordinates` from hospitals in **Gaza**

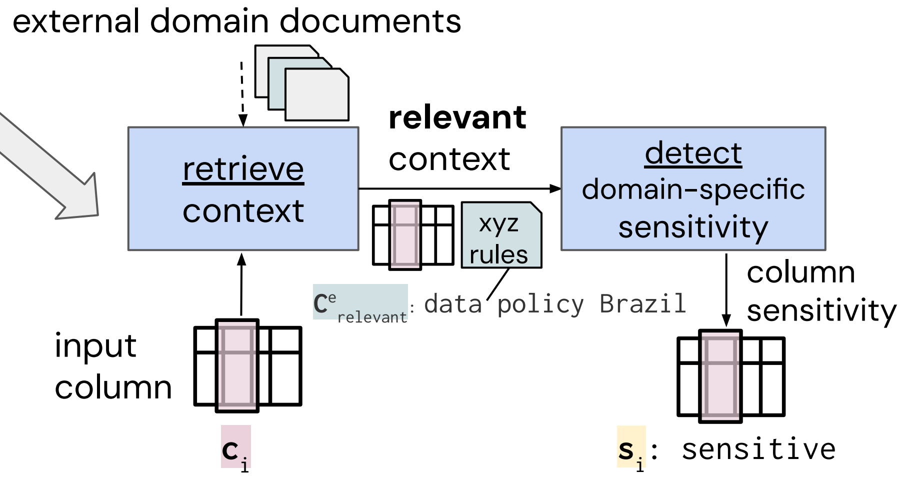
Facility Name	Lat	Long	Region
Al Amal Field Hospital	31.4872	34.4386	Central Gaza
Coastal Emergency Clinic	31.5129	34.4551	Northern Gaza
Rafah Community Hospital	31.2964	34.2478	Southern Gaza

`geo-coordinates` from hospitals in **Germany**

Facility Name	Lat	Long	Region
St. Marien Medical Center	52.5218	13.4097	Berlin
Rhine General Hospital	50.9386	6.9603	Cologne
Bavaria University Hospital	48.1397	11.5654	Munich

Mechanism: *domain-contextualization*

w/o policy rules:
precision -8%



✨**integrated** in UN's data platform
🏆**Amsterdam AI Thesis award**

100% sensitive non-PII data detected but yields some FPs (acceptable amount)

Towards Contextual Sensitive Data Detection, Telkamp & Hulsebos, Findings of EMNLP, 2026.

Colab case-study: UN humanitarian data platform

Reported location on NFIs suppliers							
nfi_supplier_	nfi_supplier_loc- value -from_Iran	nfi_supplier_	nfi_supplier_	nfi_supplier_	nfi_supplier_	nfi_supplier_	nfi_supplier_
NA	0%	0%	0%	100%	0%	0%	0%
NA	0%	0%	0%	33%	0%	67%	0%
NA	0%	0%	0%	100%	0%	0%	0%

Context-grounded LLM output

1. interpretation of data

2. assessment of sensitivity

3. context-grounded explanation

1. The column "nfi_supplier_loc- value -from_Iran" indicates the origin of NFIs suppliers.

2. This information, even with country-level disaggregation, doesn't reveal personal or sensitive data. It relates to the location of suppliers, which is generally not considered sensitive.

3. It aligns with the ISP's guideline on "Facility data (e.g., health, education, water points) at national or regional level, unless explicitly restricted by clusters." Since this column is about locations of suppliers and not about individuals or specific operations, it is categorized as non-sensitive.

✨ UN Data Responsibility Lead: *"I noticed the LLM explanation<s> and have to say it already looks great."*

Towards Contextual Sensitive Data Detection, Telkamp & Hulsebos, Findings of EMNLP, 2026.

CWI

From *Semantic* to *Analytical* Reasoning

Can we also just use LLMs to query tables…?

*What is the **average budget** of Drama movies with a **greater or equal to 8.0** rating?*

MovieID	Title	MPAA Rating	Budget	Gross	Release Date	Genre	Runtime	Rating	Rating Count
121	Dangerous Minds	R	23 000 000	179 519 401	1995-08-11	Crime	99	6.5	45 479
377	Slumdog Millionaire	R	15 000 000	377 910 544	2008-05-12	Drama	120	8.0	758 215
386	Kung Fu Panda	PG	130 000 000	631 744 560	2008-06-04	Animation	90	7.5	390 478

ActorID	Name	Year of birth	Birth City	Birth Country	Height	Biography	Gender	Ethnicity	NetWorth
410	Dustin Hoffman	1937-08-08	LA	USA	65.0	“”	Male	-	$100M
553	George Dzundza	1945-07-19	Rosenheim	Germany	69.0	“”	Male	White	$2M
2099	Saurabh Shukla	1963-03-05	-	India	-	“”	-	-	-

MovieID	ActorID	Character Name	creditOrder
121	553	Hal Griffith	2
377	2099	Sergeant Srinivas	2
386	410	Shifu	2

DeepSeek reasoning trace:

“… Wait no—the data doesn’t show that…
… Wait I’m getting confused…”

How well do LLMs reason over tabular data, really? by Wolff & Hulsebos, TRL@ACL, 2025

CWI

The expected sad state of tabular QA with LLMs

LLM-as-a-judge on *averaging* questions over multiple tables in TQA-Bench

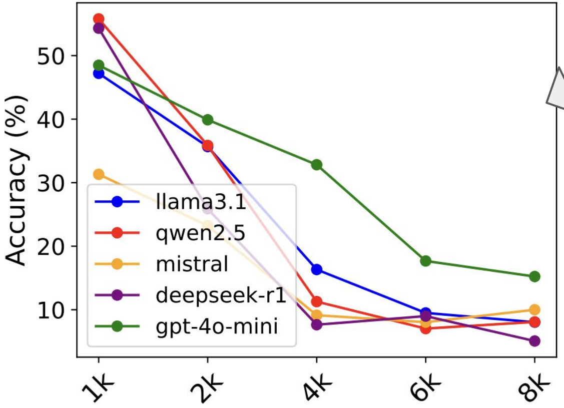


table context size

don't let yourself get fooled by MPC evals for LLMs

How well do LLMs reason over tabular data, really? by Wolff & Hulsebos, TRL@ACL, 2025

Better use SQL…

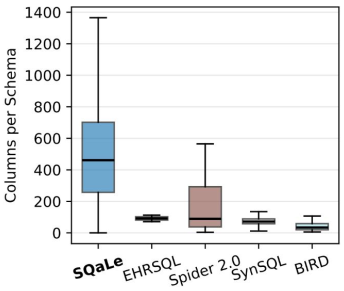
✨ **SQaLe v1 downloaded 6K+ times**

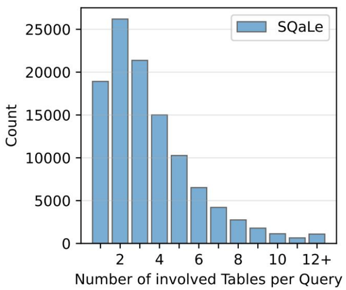
Text-to-*SQL* is not entirely solved though:

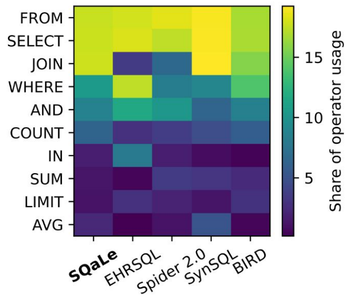
- SOTA: throwing massive LLM pipelines at it; that's it? fit for the agent-era?
- Complex queries/data: text-to-SQL needs *data-grounding*

We need **SQaLe** to specialize.

- First **large** text-to-SQL dataset
- 500K+ (question, sql, schema)
- Queries over 23K *real* schemas from SchemaPile (GitHub)







SQALE: A Large Dataset for Specialized Text-to-SQL by Wolff, Gomm, and Hulsebos, AITD@EurIPS
-> full paper coming soon

Complex queries, you said?

An example question in **SQaLe**:

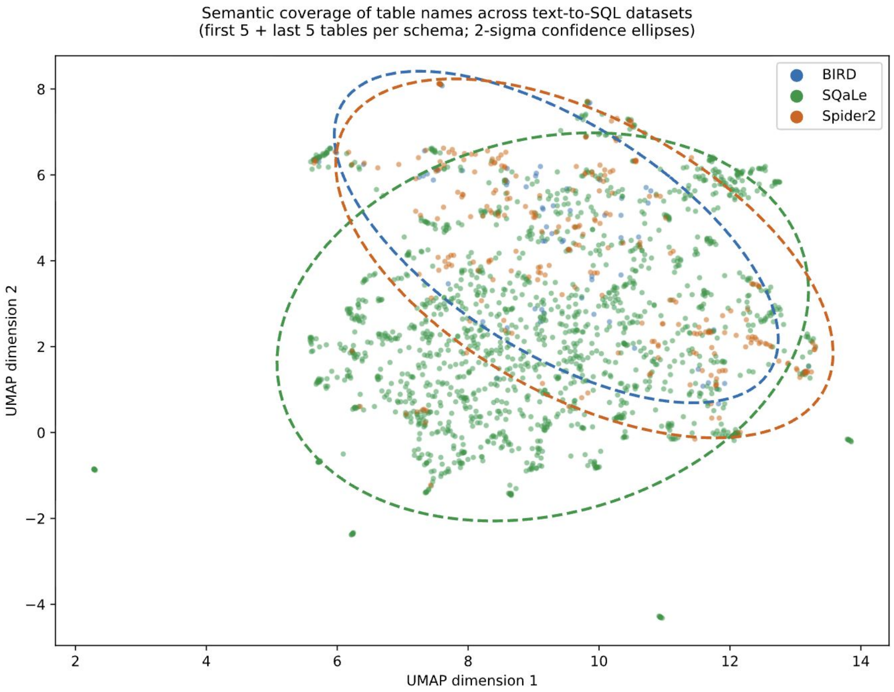
"For each customer in the gold or platinum loyalty tier who has placed at least two delivered orders in the last six months, return their full name, email, city and state, the total number of delivered orders, the total amount they paid based only on captured payments, their most frequently ordered product category in that period, and whether any of the products they ordered currently have low inventory in any warehouse (defined as quantity on hand minus quantity reserved less than 10), sorted by total captured payment amount from highest to lowest."

*SQaLe has different verbosity levels :)

```sql
WITH gold_platinum_customers AS
  (SELECT c.id,
          c.first_name,
          c.last_name,
          c.email,
          c.city,
          c.state
   FROM customers c
   JOIN orders o ON c.id = o.customer_id
   WHERE c.loyalty_tier IN ('gold',
                            'platinum')
     AND o.status = 'delivered'
     AND o.delivered_at >= date('now', '-6 months')
   GROUP BY c.id
   HAVING COUNT(*) >= 2),
     order_details AS
  (SELECT gpc.id,
          gpc.first_name,
          gpc.last_name,
          gpc.email,
          gpc.city,
          gpc.state,
          o.order_number,
          SUM(p.amount) AS total_paid,
          p.status AS payment_status
   FROM gold_platinum_customers gpc
   JOIN orders o ON gpc.id = o.customer_id
   LEFT JOIN payments p ON o.id = p.order_id
   AND p.status = 'captured'
   WHERE o.delivered_at >= date('now', '-6 months')
     AND o.status IN ('delivered',
                      'completed')
   GROUP BY gpc.id,
            o.id),
     product_category AS
  (SELECT od.id,
          od.first_name,
          od.last_name,
          od.email,
          od.city,
          od.state,
          p.category,
          ROW_NUMBER() OVER (PARTITION BY od.id
                             ORDER BY COUNT(*) DESC) AS rn
   FROM order_details od
   JOIN order_items oi ON od.order_number = CAST(oi.order_id AS TEXT)
   JOIN products p ON oi.product_id = p.id
   GROUP BY od.id)
SELECT gpc.first_name,
       gpc.last_name,
       gpc.email,
       gpc.city,
       gpc.state,
       COUNT(od.order_number) AS total_orders,
       SUM(od.total_paid) AS total_captured_payment_amount
FROM gold_platinum_customers gpc
JOIN order_details od ON gpc.id = od.id
AND od.payment_status = 'captured'
LEFT JOIN product_category pc ON od.id = pc.id
AND pc.rn = 1
WHERE pc.category IS NOT NULL
GROUP BY gpc.id
ORDER BY total_captured_payment_amount DESC;
```

SQALE: A Large Dataset for Specialized Text-to-SQL by Wolff, Gomm, and Hulsebos, AITD@EurIPS
-> full paper coming soon

Diversity, really?

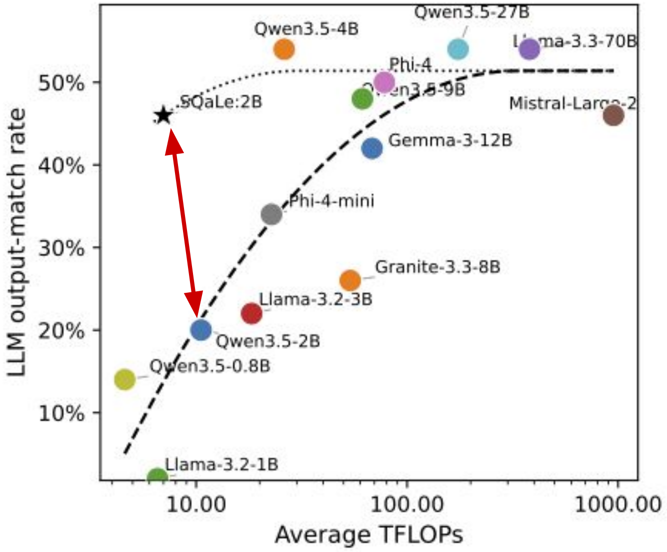


SQALE: A Large Dataset for Specialized Text-to-SQL by Wolff, Gomm, and Hulsebos, AITD@EurIPS
-> full paper coming soon

CWI

SQaLe enables the vision of *specialized small* text-to-SQL

Initial Qwen2B model fine-tuned on **SQaLe** (33% of v2) for text-to-SQL



SQALE: A Large Dataset for Specialized Text-to-SQL by Wolff, Gomm, and Hulsebos, AITD@EurIPS
-> full paper coming soon

SQaLe utility for *end-to-end* text-to-SQL: retrieval

Problem: semantic gap question <> tables

Idea: align questions & tables within *generative table retrieval model* (**GenTR**)

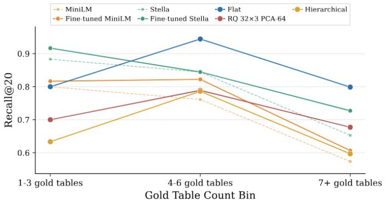
question → **GenTR** → Table IDs

Question: by pre-training at scale (on SQaLe), can GenTR adapt to **new corpora**?

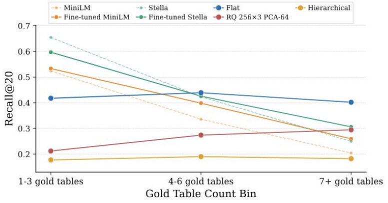
Generalization performance

Benchmark		Method	Source	R@5	R@10	R@20	MRR@20
FIBEN	Baselines	BM25	–	0.049	0.049	0.059	0.173
		MiniLM	–	0.140	0.234	0.390	0.319
		Stella	–	0.318	0.431	0.541	0.593
	Fine-tuned	MiniLM	–	0.372	0.643	0.881	0.458
		Stella	–	0.759	0.963	0.994	0.930
	GenTR	LM-Head Tuning	10k	**0.971**	**0.999**	**1.000**	**0.983**

Robust in complex cases



(a) SQaLe2-300. Dense baselines degrade with increasing gold-table count. Flat remains stable and leads on queries with 4+ gold tables.

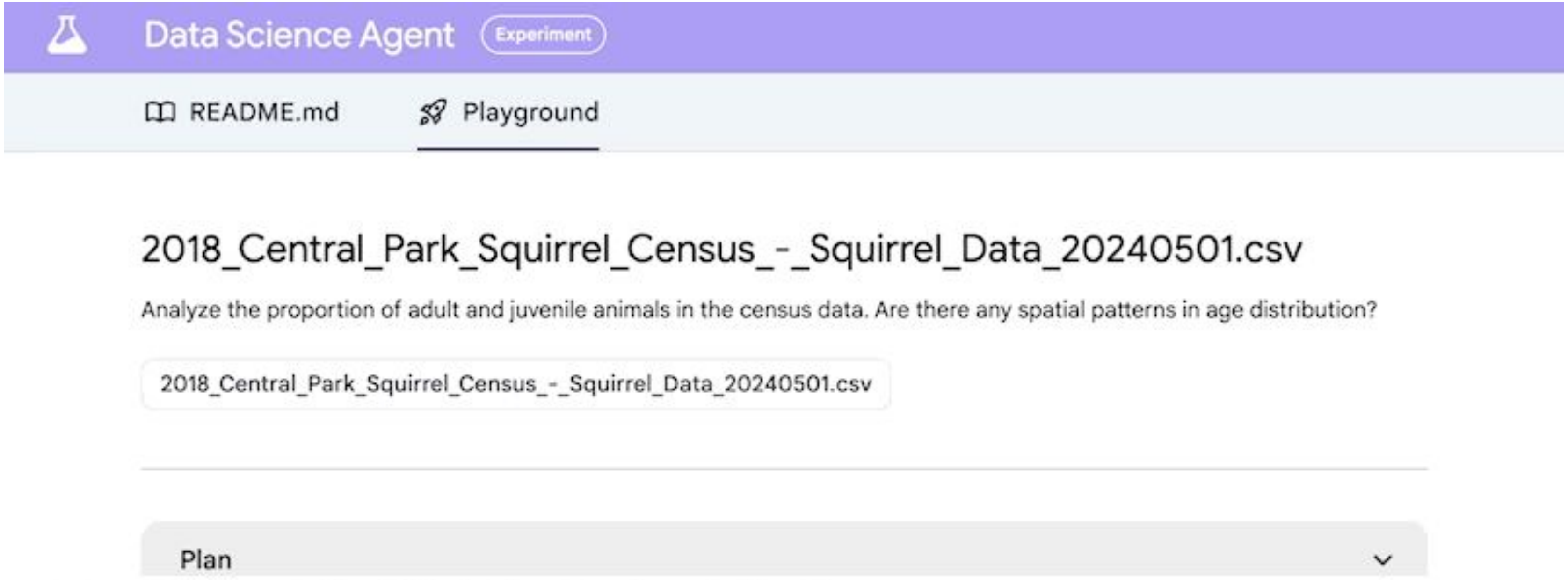


(b) SQaLe2-10k. Dense baselines degrade while generative methods remain approximately flat. RQ surpasses zero-shot Stella and approaches fine-tuned Stella on queries with 7+ gold tables.

Generative Table Retrieval by Kosiuk, W., 2026.

From *Analytical* to *<u>Statistical</u>* Reasoning

Recall: flashy demos of Data Science Agents



Problem Solved?

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

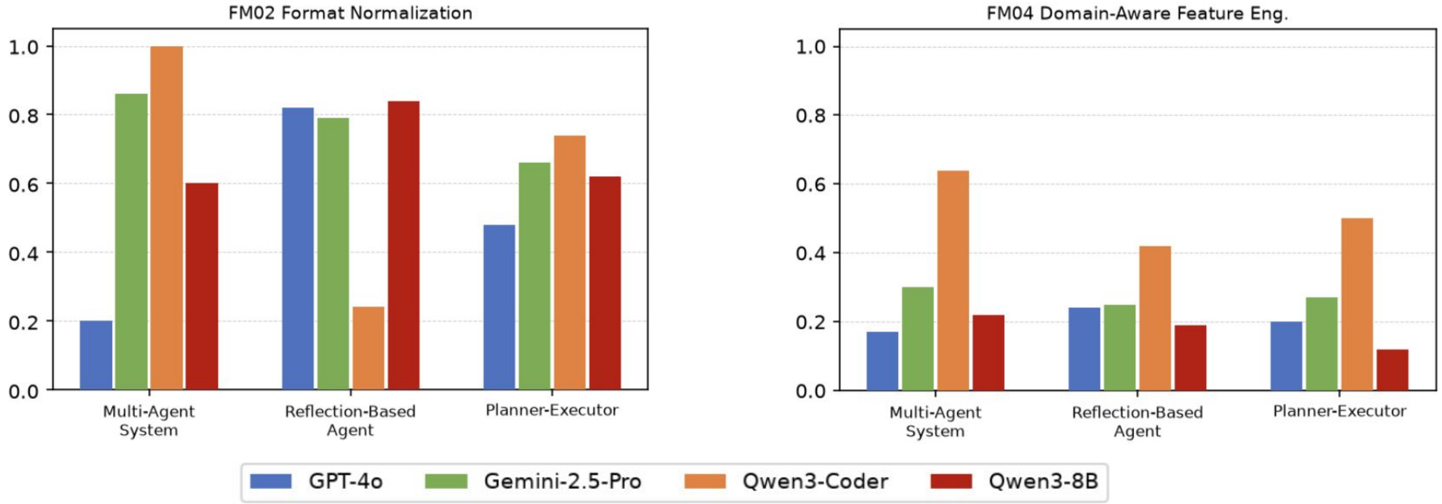
data science is
not just
code generation

From *TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks* by Xu et al., NeurIPS, 2025.

Failure modes in agentic data science

Data preparation... ✅

DS / domain knowledge... ❌



Why do Data Science Agents fail? by Xia, H., et al, 2026.
-> full paper coming soon

Wrapping up…

- ➢ Many tasks (e.g. sensitivity val) powered by table semantics, **contextualization** for LLMs key in improvement, interpretability, ánd consistency of outcomes.
- ➢ Tabular AI is multifaceted: analytical/statistical reasoning have different needs. Future for ***small specialized*** models, **SQaLe** powers this for text2sql, retrieval, …
- ➢ Agentic data science for predictive tasks needs go beyond contextualized LLMs/code gen: domain- and data science **knowledge** are key.
- ➢ Things we're working on:
  - **Table retrieval**: end-to-end QA, generative table retrieval, specialized table embeddings
  - Understanding **tabular insight needs**, and which tables address them **appropriately**
  - **Agentic data science**: from failure modes to knowledge-grounding guardrails
  - **Predictive tabular learning** for large datasets, and next-gen capabilities

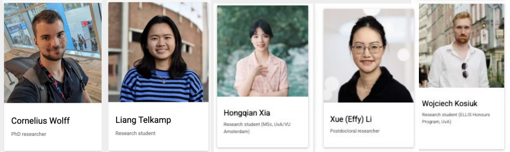
That's it!

TRL Lab

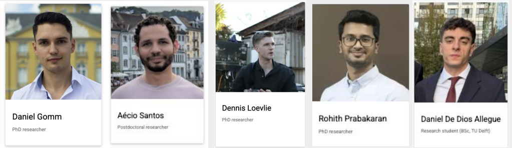
Madelon Hulsebos
madelonhulsebos.com
trl-lab.github.io

› **madelon@cwi.nl**

Leads of projects



Other members



Funding & Collaboration

CWI Centrum Wiskunde & Informatica
ellis UNIT AMSTERDAM
MIT
Berkeley UNIVERSITY OF CALIFORNIA
Universiteit van Amsterdam
SAP
NWO
Funded by the European Union
NextGenerationEU