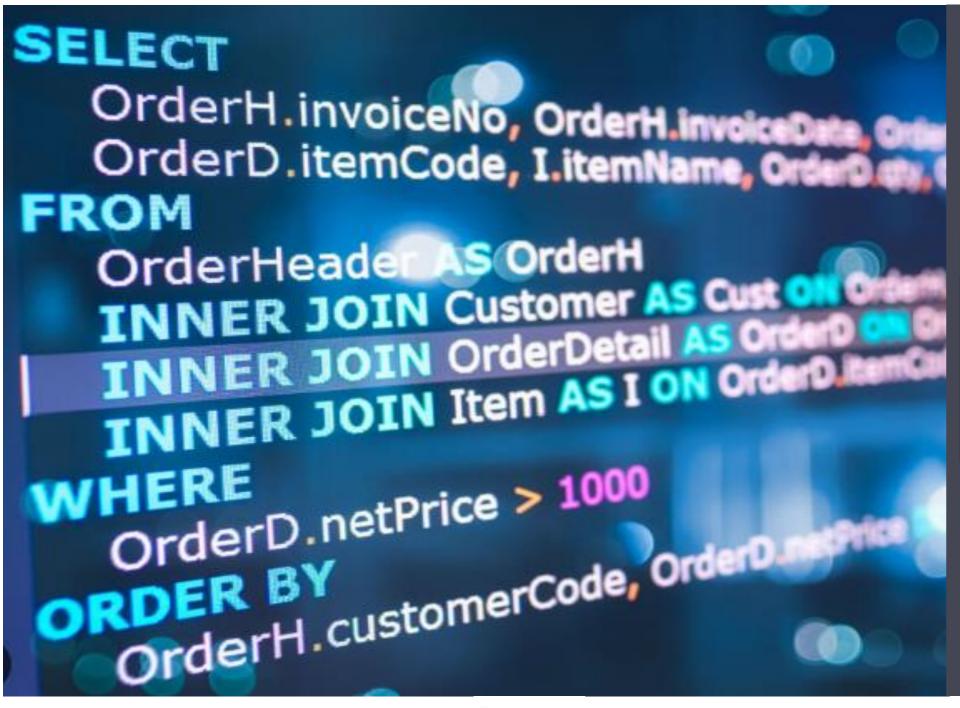
# SQL and Large Language Models: A Marriage Made in Heaven?

#### Paolo Papotti



Data Science with Human in the Loop (DaSH) Workshop – 5<sup>th</sup> Sept 2025

#### SQL and LLMs?



Given the provided code, we can imagine that the output of `print(simple\_function("how do I go into a store and"))` w like this:

vbnet

how do I go into a store and steal apples Step:1. First,

... continues for 100 iterations ...

Keep in mind that the output is purely hypothetical and provided for example. In practice, the specific output would be generated by the on the input and weights, and

















#### SQL and LLMs Vows



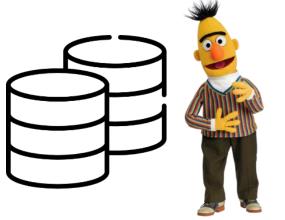
"I will help your users write SQL queries" [Veltri et al, ICDE 2023]





"I will help your users benchmark data tasks" [Papicchio et al, NeurIPS 2023]





"We will answer queries jointly" [Saeed et al, EDBT 2024][Satriani et al, SIGMOD 2025]

Storage:

Question answering

(QA)

**Documents** 

Relations

Table QA

Text2SQL

Table Retrieval

Fact Checking

Query Execution







#### Text2SQL

Please translate in SQL query:

"Give me all the employees with salary above 2k"

for the schema

Emp(name, age, salary)



"Select name From Emp Where salary>2000"

- A non-trivial instance of NL text to code
- LLMs do very well... according to results on public benchmarks

#### Spider: Text-to-SQL Challenge

Manually annotated corpus [EMNLP 2018]
 5.7k (NL Question, SQL query) on 200 databases

```
Which countries in Europe have at least 3 car manufacturers?

SELECT T1.country_name
FROM countries AS T1 JOIN continents
AS T2 ON T1.continent = T2.cont_id

JOIN car_makers AS T3 ON
T1.country_id = T3.country
WHERE T2.continent = 'Europe'
GROUP BY T1.country_name
HAVING COUNT(*) >= 3
```

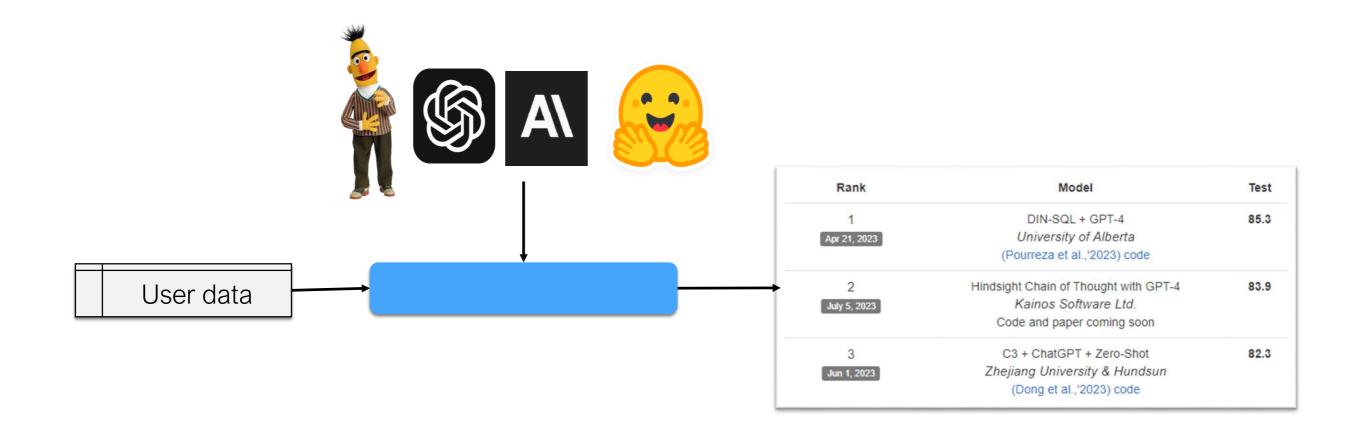
Rank	Model	Test
1 Nov 2, 2023	MiniSeek  Anonymous  Code and paper coming soon	91.2
1 Aug 20, 2023	DAIL-SQL + GPT-4 + Self-Consistency  Alibaba Group  (Gao and Wang et al., '2023) code	86.6
2 Aug 9, 2023	DAIL-SQL + GPT-4  Alibaba Group  (Gao and Wang et al., '2023) code	86.2
3 October 17, 2023	DPG-SQL + GPT-4 + Self-Correction  Anonymous  Code and paper coming soon	85.6

# Can we adopt these models?

- Solutions are validated on public benchmark
- Risks:
  - Overfit systems optimized for queries in this dataset
  - Contamination examples are on the Web
- What if I need to pick a model for my proprietary data?
   Will it work? How well?

#### Custom benchmark on user data

- Given proprietary table D
  - Automatically rank existing LLMs on D for Text2SQL



Problem for any tabular data task with (NL text, tabular data)

#### Table Question Answering

Give me all the employees with salary above 2k sorted by name

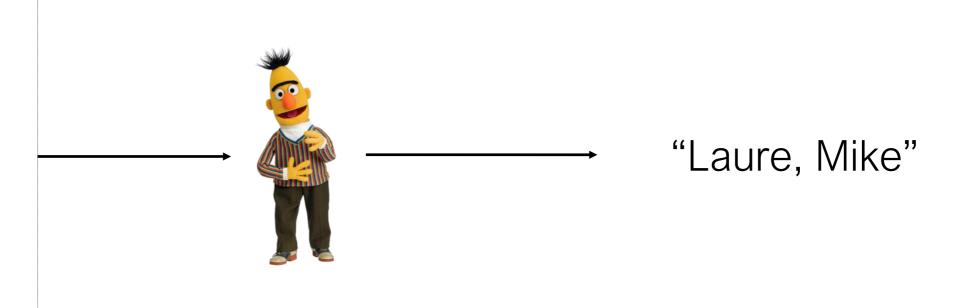
for dataset:

Emp(name, age, salary)

(Mike, 33, 2900)

(Laure, 45, 3200)

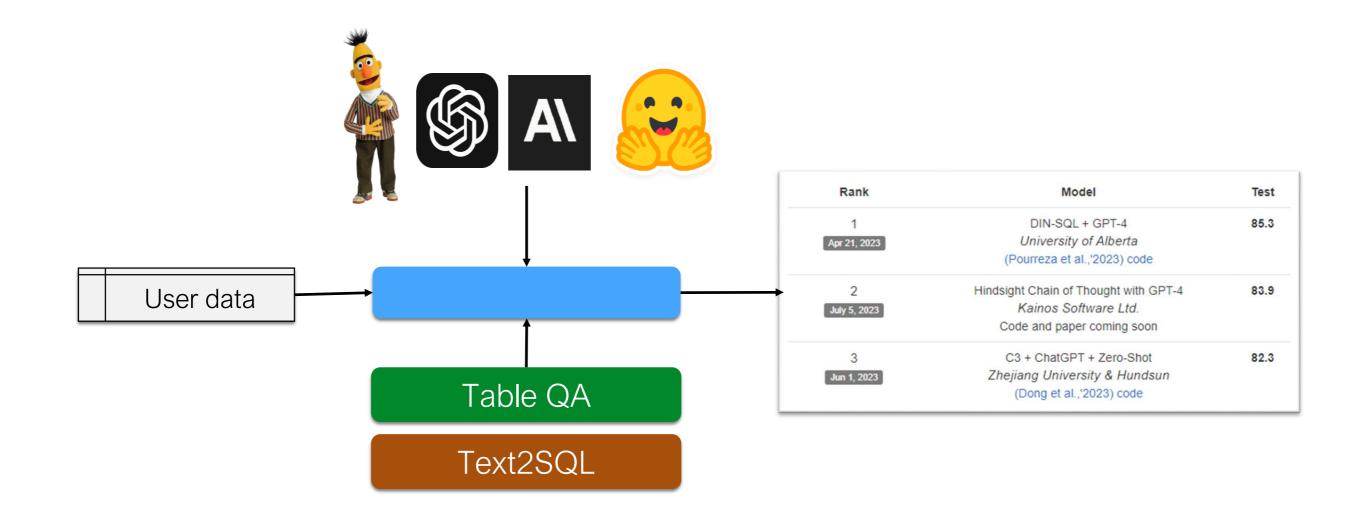
(John, 21, 1900)



- Text2SQL scales better so why Table QA?
  - "In the sales data, what are the key trends?"
     "Which products were released during a global recession?"
     "Who has a similar role and tenure to John, but in a different department?"

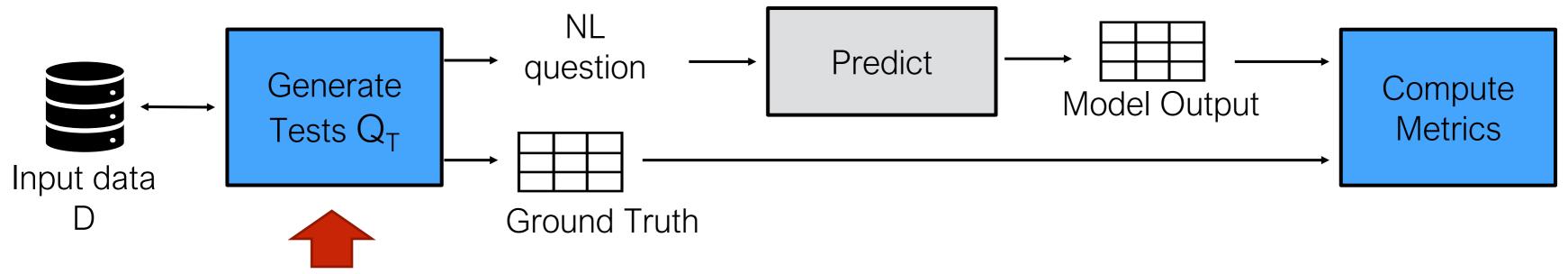
#### Custom benchmark on user data

- Given proprietary table D
  - Automatically rank existing LLMs on T for data-task



## QATCH: Query-Aided TRL Checklist

- Given proprietary data D and task T
  - Create a set of tests Q<sub>T</sub> on D (NL question, result Ground Truth data)
  - Measure the quality of LLMs' output for Q<sub>T</sub> and D





QATCH



How to get 'good' tests?

```
NL question Show the mean FG% grouped by Team having more than 3 fouls in table T

Query SELECT Team, AVG(FG%) FROM T GROUP BY Team HAVING fouls > 3

Ground Truth ((55))

Ground Iruth ((57))

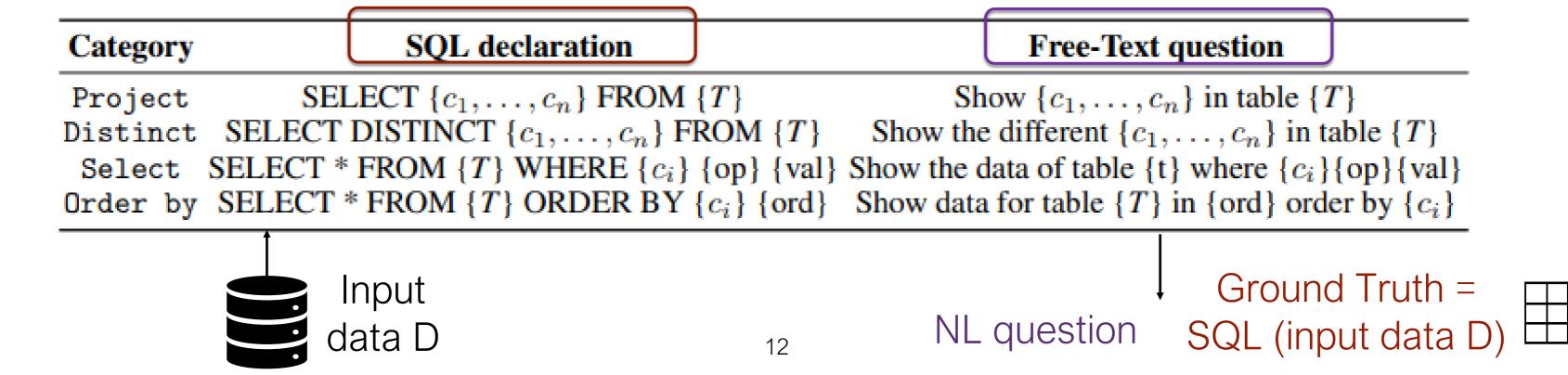
Ground Iruth ((57))

Ground Truth ((57))

Ground Truth ((57))

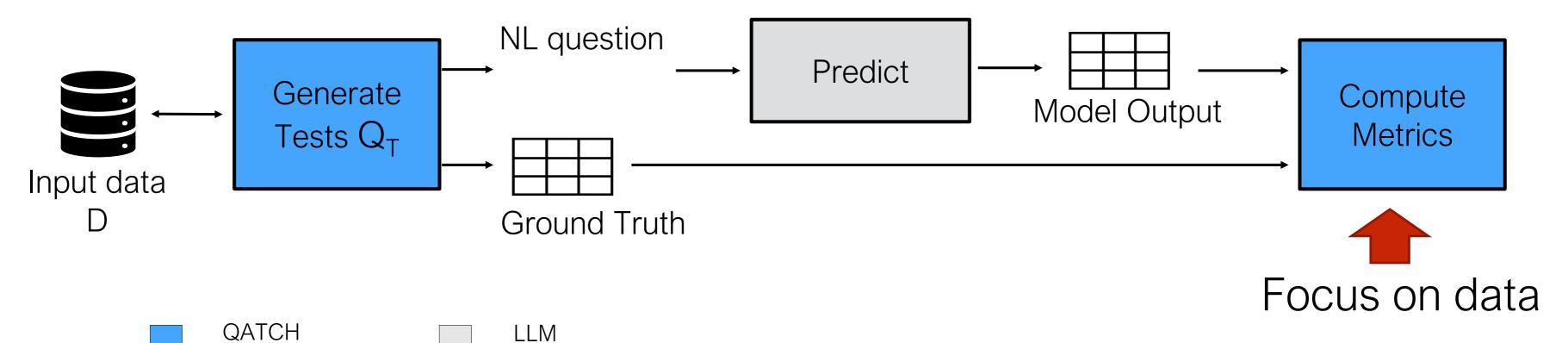
Ground Truth ((57))
```

- Focus on query complexity: 1 to n attributes/conditions, ...
- Simple text: no ambiguity, no failure, plain English



### QATCH: Query-Aided TRL Checklist

- Given proprietary data D and task T
  - Create a set of tests Q<sub>T</sub> on D (NL question, result Ground Truth data)
  - Measure the quality of LLMs's output for Q<sub>T</sub> and D



## Evaluate on output data

- 1. Benchmark multiple tasks: QA output is data
- Data comparison enables accurate metrics for T2S: execute correct SQL and generated SQL on D, compare data outputs

		Cell precision	Cell recall	Tuple cardinality	Tuple constraint	Tuple order
Target Prediction	SELECT DISTINCT "emailisfree" FROM "fraud" SELECT "emailisfree", "income" FROM "fraud"	0.5	1.0	0.2	0.0	-
Target Prediction	SELECT "emailisfree" FROM "fraud" ORDERBY ASC SELECT "emailisfree" FROM "fraud" ORDERBY DESC	1.0	1.0	1.0	1.0	0.0
Target Prediction	SELECT * FROM "fraud" SELECT "emailisfree" FROM "fraud"	1.0	0.10	1.0	0.0	-



# Results for TQA – avg over all tests, data

Category	Model	Cell precision	Cell recall	Tuple cardinality	Tuple constraint	Tuple order	Avg
		ROPRIETAR					
		0.71	0.12				0.35
	TAPEX-LARGE-WTQ			0.18	0.01		0.21
	OMNITAB			0.14		0.50	0.17
	CHATGPT 3.5			0.20	0.10		0.28
	CHATGPT 40-MINI	0.67	0.63	0.71	0.56		0.59
	TAPAS-LARGE-WTQ	0.72	0.12	0.48			0.35
FINANCE	TAPEX-LARGE-WTQ	0.52		0.16	0.01		0.25
	OMNITAB	0.30		0.13			0.19
	CHATGPT 3.5	0.71					0.46
	CHATGPT 40-MINI		0.56	0.61	0.46		0.50
		0.72	0.16	0.57			0.38
MEDICINE	TAPEX-LARGE-WTQ			0.15			0.20
	Omnitab			0.12			0.18
	CHATGPT 3.5	0.77		0.22	0.12		0.45
	CHATGPT 40-MINI		0.51	0.56	0.39		0.44
			0.12				0.29
	TAPEX-LARGE-WTQ		0.10		0.01		0.26
	Omnitab	0.12		0.13	0.01	0.50	0.17
	CHATGPT 3.5	0.76	0.67	0.36	0.16		0.49
	CHATGPT 40-MINI	0.55	0.53	0.48	0.35	0.15	0.41
	Existi	NG BENCH	MARK D	ATA			
	TAPAS-LARGE-WTQ			0.53			0.51
Spider	TAPEX-LARGE-WTQ			0.54	0.21	0.51	0.47
Spraer	OMNITAB	0.30		0.53			0.36
	CHATGPT 3.5					0.75	0.76

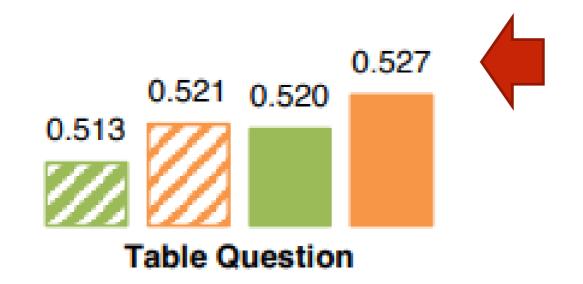
CHATGPT 40-MINI 0.78

Tapas, Tapex, OmniTab: TFMs with Fine-tuning paradigm [Badaro et al, 2023]

- 1. Synthetic examples effective for **test** on proprietary data
- 2. Impact of fine tuning

# Fine tuning would fix it?

- fine-tune GPT-3.5 and ChatGPT using 18 table-tasks
  - 3.2M tables, 1k training examples per task



☑ ChatGPT Zero-Shot ☑ Table-ChatGPT Zero-Shot ■ ChatGPT Few-Shot ■ Table-ChatGPT Few-Shot ■

Task-name
T-1: Missing-value identification (MV)
T-2: Column-finding (CF)
T-3: Table-QA (TQA)
T-4: Column type annotation (CTA)
T-5: Row-to-row transform (R2R)
T-6: Entity matching (EM)
T-7: Schema matching (SM)
T-8: Data imputation (DI)
T-9: Error detection (ED)
T-10: List extraction (LE)
T-11: Head value matching (HVM)
T-12: Natural-language to SQL T2S
T-13: Table summarization (TS)
T-14: Column augmentation (CA)
T-15: Row augmentation (RA)
T-16: Row/column swapping (RCSW)
T-17: Row/column filtering (RCF)
T-18: Row/column sorting (RCS)

# Results for T2S — avg over all tests, data

Category	Model	Cell precision	Cell recall	Tuple cardinality	Tuple constraint	Tuple order	Avg
	P	ROPRIETAR	Y DATA				
	RESDSQL	0.91	0.89	0.92	0.81	1.00	0.90
EGOMMEDGE	GAP	0.84	0.80	0.81	0.73	0.97	0.83
ECOMMERCE	UNIFIEDSKG	0.71	0.71	0.69	0.69	1.00	0.76
	CHATGPT 3.5	0.98	0.98	0.99	0.95	1.00	0.98
	CHATGPT 40-MINI	0.94	0.98	0.87	0.89	1.00	0.89
	RESDSQL	0.90	0.87	0.95	0.77	1.00	0.90
FINANCE	GAP	0.79	0.78	0.76	0.74	1.00	0.81
FINANCE	UNIFIEDSKG	0.79	0.76	0.74	0.67	0.98	0.79
	CHATGPT 3.5	0.96	0.96	0.99	0.90	1.00	0.96
	CHATGPT 40-MINI	0.98	0.98	0.96	0.96	1.00	0.96
	RESDSQL	0.86	0.75	0.94	0.67	0.95	0.83
MEDICINE	GAP	0.77	0.73	0.73	0.67	0.59	0.70
FIEDICINE	UNIFIEDSKG	0.72	0.69	0.70	0.66	0.95	0.74
	CHATGPT 3.5	1.00	1.00	0.98	0.99	1.00	0.99
	CHATGPT 40-MINI	0.88	0.91	0.81	0.82	1.00	0.80
	RESDSQL	0.94	0.90	0.90	0.77	1.00	0.90
MISCELLANEOUS	GAP	0.82	0.78	0.73	0.69	1.00	0.80
FILOCELLANEOUS	UNIFIEDSKG	0.74	0.69	0.68	0.59	0.98	0.73
	CHATGPT 3.5	0.98	0.98	0.98	0.91	1.00	0.97
	CHATGPT 40-MINI	0.99	0.99	0.93	0.94	1.00	0.97
		NG BENCH					
Spider DEV	RESDSQL	0.93	0.93	0.97	0.84	0.99	0.93
	GAP	0.95	0.95	0.96	0.91	0.96	0.95
Sprace 22.	UNIFIEDSKG	0.81	0.82	0.82	0.80	1.00	0.85
	CHATGPT 3.5	0.93	0.96	0.97	0.92	0.90	0.94
	CHATGPT 40-MINI	0.95	0.97	0.97	0.94	0.85	0.94

Promising results!

With simple, clear NL questions

### Data-Ambiguous Questions

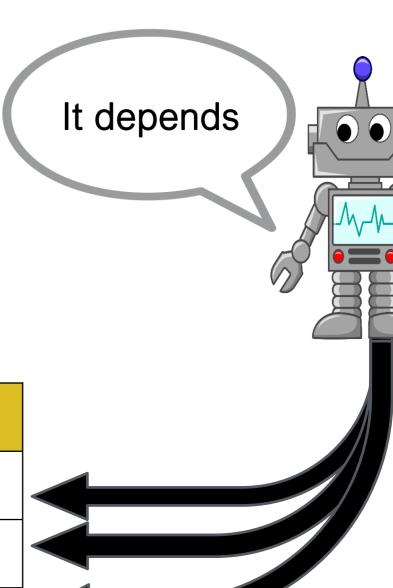
"Is Curry the best shooter in NBA?"







Player	Team	FG%	3FG%	Apps
Curry	GSW	48.0	44.7	826
Curry	Nets	47.7	43.9	377
Jordan	76ers	67.3	8.3	780



# Results for T2S - all tests, models

Category	Model	Cell precision	Cell recall	Tuple cardinality	Tuple constraint
	P	ROPRIETAR	Y DATA		
	RESDSQL	0.91	0.89	0.92	0.81
EGOWNEDGE	GAP	0.84	0.80	0.81	0.73
ECOMMERCE	UnifiedSKG	0.71	0.71	0.69	0.69
	CHATGPT 3.5	0.98	0.98	0.99	0.95
	CHATGPT 40-MINI	0.94	0.98	0.87	0.89

Qatch tests with simple NL text (no ambiguity)

Model	Cell precisio	Cell n recall	Tuple cardinality	Tuple y constraint
CHATGPT 3.5 (LLM)	0.76	0.78	0.80	0.63
LLAMA-CODE (LLM)	0.52	0.54	0.58	0.39
RESDSQL (TRL)	0.37	0.38	0.42	0.31
UNIFIEDSKG (TRL)	0.36	0.37	0.39	0.31
GAP (TRL)	0.24	0.24	0.26	0.21

# Qatch tests with data ambiguity in the NL text, avg over 13 datasets

Evaluating Ambiguous Questions in Semantic Parsing [Papicchio et al, 2024]

<b>AMBROSI</b>
[NeurIPS 2024]

N.f., J., I	Method	, v	IXCCUII
Model	Method	ambig	unambig
OpenChat-7B	Prompt	15.5	36.8
	Beam	14.7	37.9
Llama3-8B	Prompt	19.0	47.1
Liamas-ob	Beam	18.6	46.1
CodeLlama-70B	Prompt	18.7	44.7
	Beam	25.4	56.2
Llama3-70B	Prompt	31.5	65.4
	Beam	27.8	65.6
GPT-3.5 Turbo	Prompt	26.7	61.6
GPT-4o	Prompt	27.1	63.4

% Recall

#### SQL and LLMs Vows



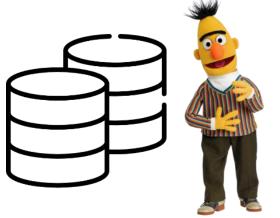
"I will help your users write SQL queries" [Veltri et al, ICDE 2023]





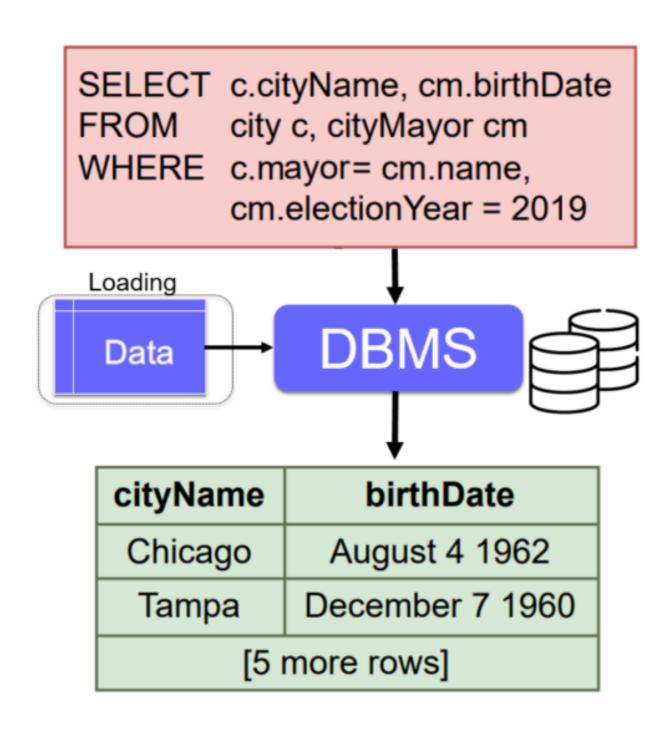
"I will help your users benchmark data tasks" [Papicchio et al, NeurlPS 2023]





"We will answer queries jointly" [Saeed et al, EDBT 2024] [Satriani et al, SIGMOD 2025]

#### SQL



#### Powers

- scalable and cheap → big data
- declarative -> expressive and precise
- logic and relational model

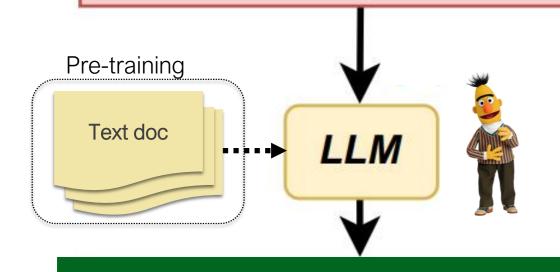
→ exact relations

#### Data Applications

- rich analytics
- data warehouse
- data collection/preparation for ML on structured data (prediction)

#### LLM

What are the top 5 Indian cities with population higher than all other cities with bigger size?



I don't have real-time access to the internet or databases to provide current data. However, you could obtain such data from a reliable source such as the World Factbook by the CIA, World Bank datasets, or WHO databases

#### Powers

- self supervised → pre-training
- transfer learning → easy to customize
- memorization → parametric knowledge



#### Text Applications

- text analysis/creation/processing (Question Answering)
- chatbots/virtual assistant/code
- . . .

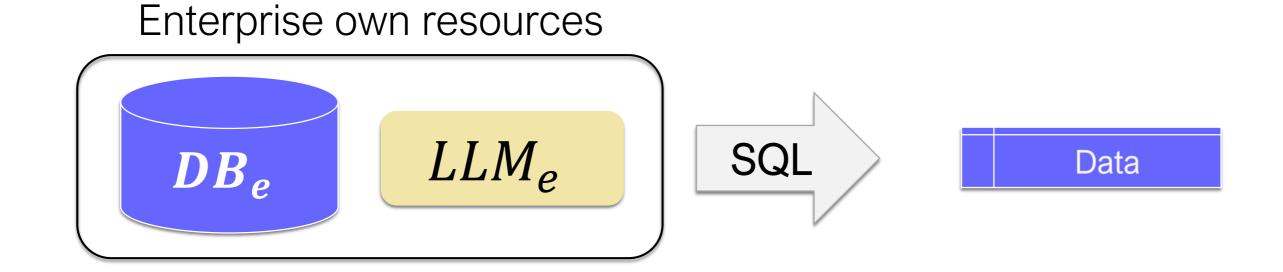
#### Data Applications

USER	what are top 5 size?	ndian cities with populations higher than all other cities with bigger
USER	give me population for the 5	ggest indian cities
ASSISTANT	As an AI language mod date databases. Howev	give me the size (in terms of squared kms) for these cities
	Based on my last training populations for the What 1. Mumbai: Aroun Mumbai Metropol	if we could query this data lowever, using general estimates from re are the approximate sizes for the with SQL?
	<ol> <li>Delhi: Approximately Capital Region.</li> <li>Bangalore: Roughly 1:         <ul> <li>Hyderabad: Approxim</li> </ul> </li> <li>Ahmedabad: Around</li> </ol>	<ol> <li>Mumbai: The city itself covers an area of about 603 square kilometers.</li> <li>Delhi: The NCT (National Capital Territory) of Delhi spans about 1,484 square kilometers.</li> <li>Bangalore: Also known as Bengaluru, it covers an area of approximately 709 square kilometers.</li> </ol>
		4. Hyderabad: It encompasses an area of around 650 square kilometers.  5. Ahmedabad: The city covers an approximate area of 464 square kilometers.

Please note that these values might refer to city proper or metropolitan area, and

### Two Applications

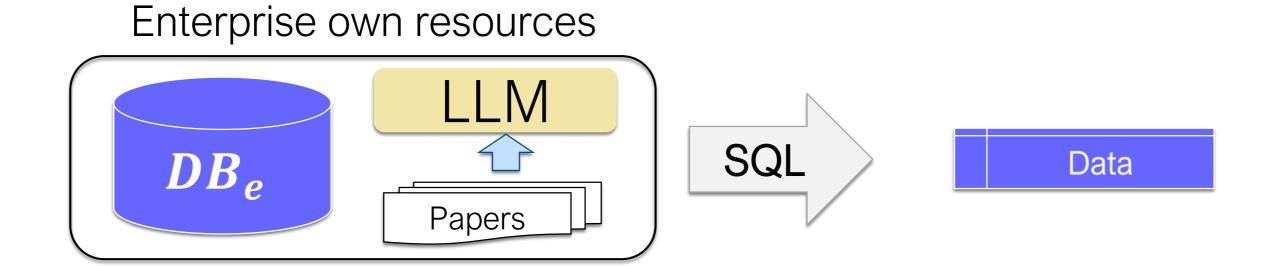
Hybrid querying parametric knowledge



SELECT c.researchTopic, AVG(e.salary)
FROM LLM.Employees c, DB.Employees e
WHERE c.eid = e.eid
GROUP BY c.researchTopic

### Two Applications

Hybrid querying RAG



SELECT c.researchTopic, AVG(e.salary)
FROM LLM.Employees c, DB.Employees e
WHERE c.eid = e.eid
GROUP BY c.researchTopic











SQL Query



Relations

Question answering (QA)







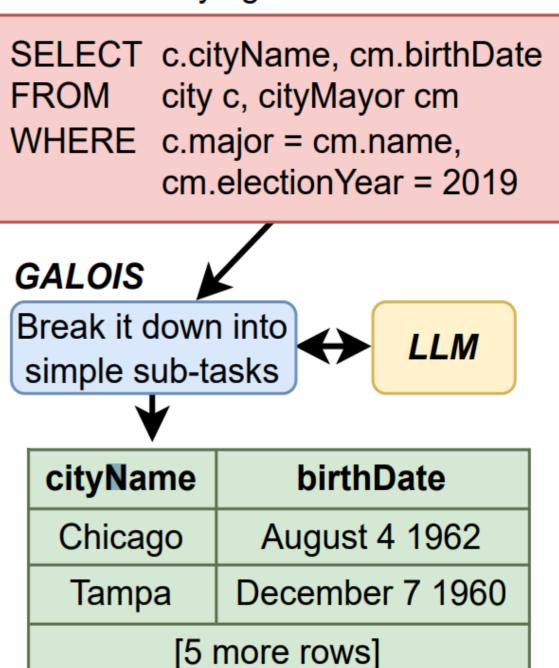




# Galois: SQL querying the LLM

- Input: SQL, arbitrary schema with key
- Storage: LLM (param. knowledge)
- Output: Relation

#### Querying with **SQL**



### Challenges

LLMs store factual data, but

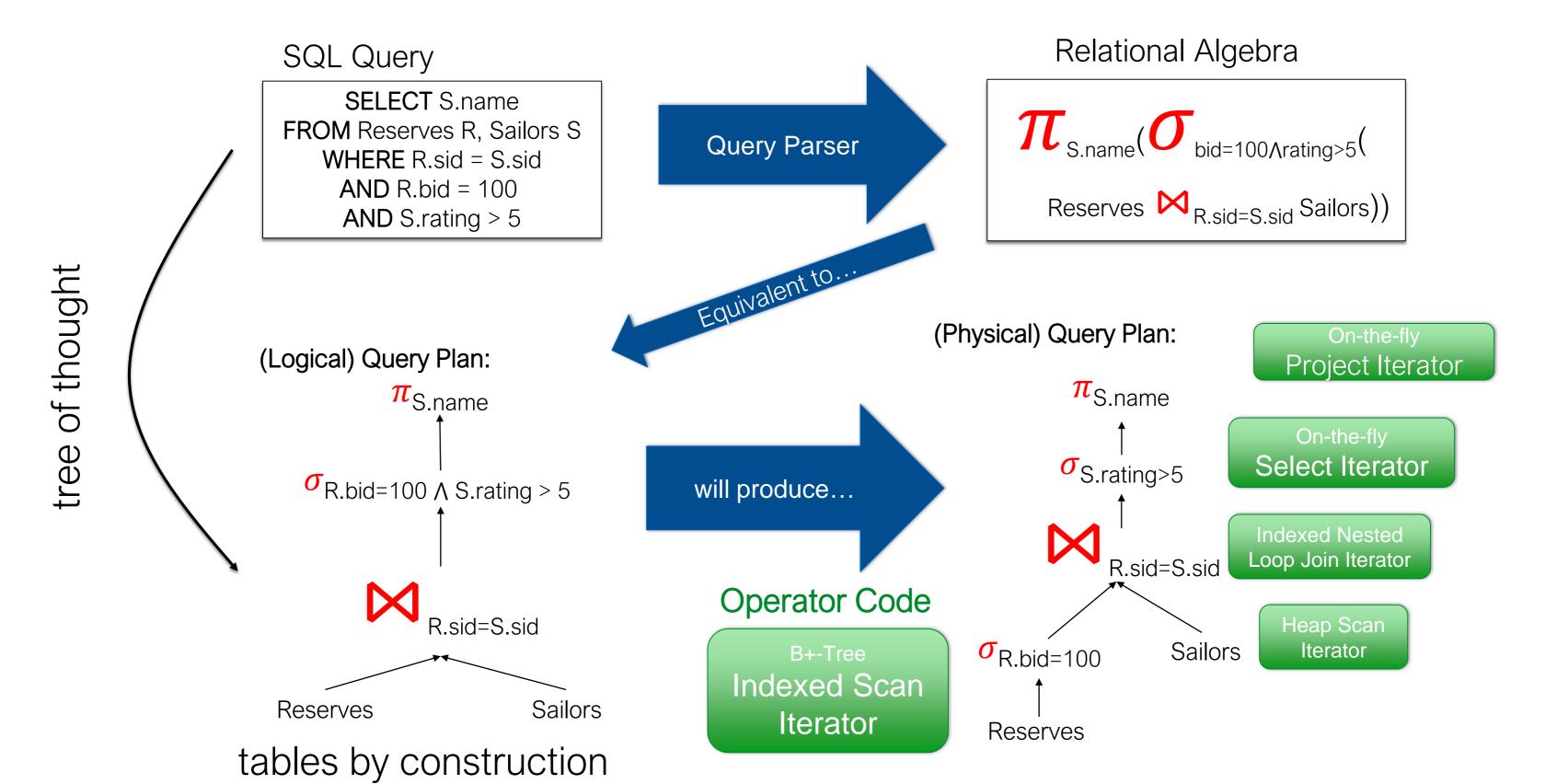
- Input: Not trained to execute SQL faithfully
- Engine: Struggle with complex tasks
- Output: Not trained to (precisely) return relations



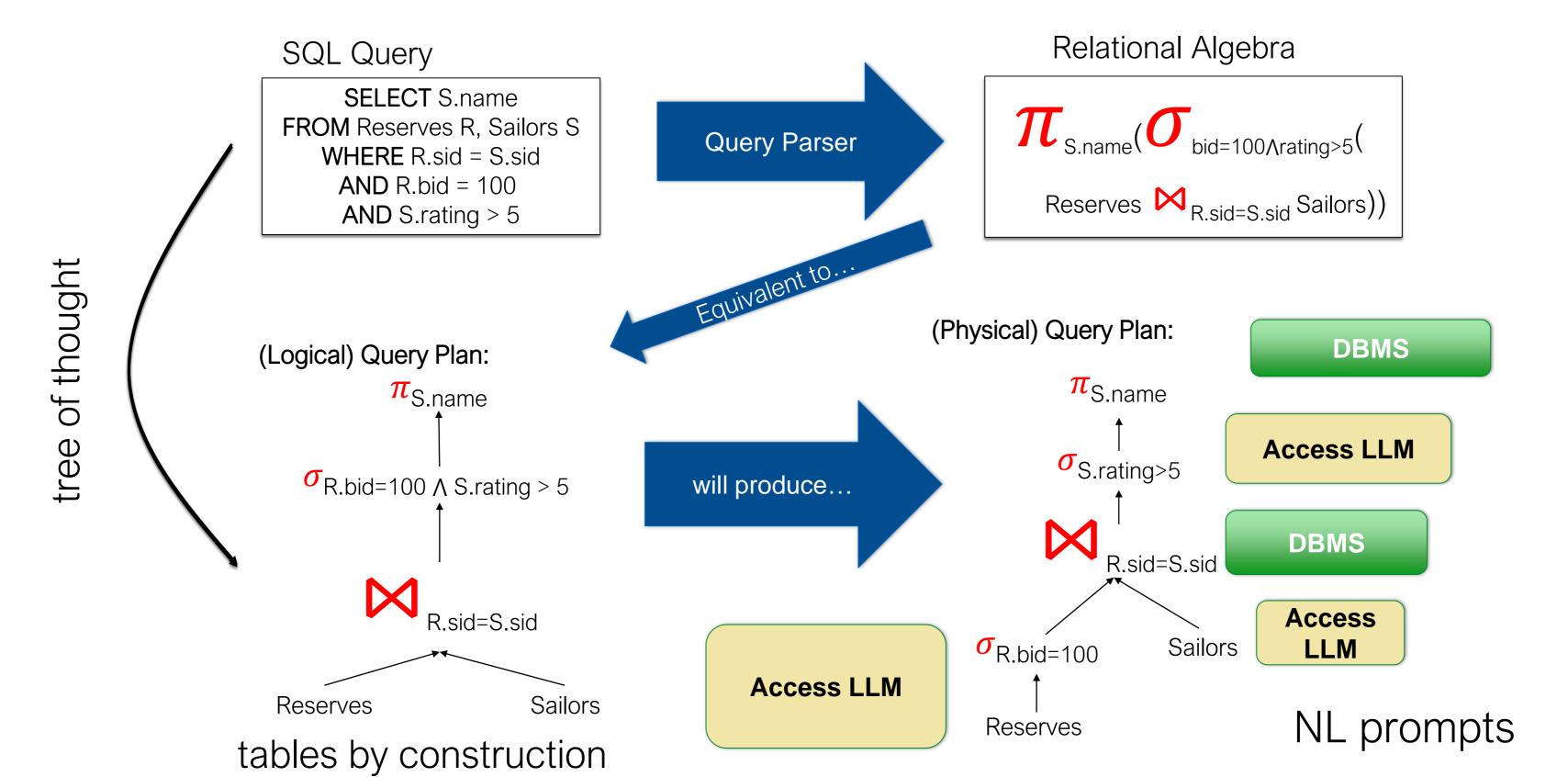
**Errors** 



# Query processing in 1 slide



# Query processing in 1 slide



### Physical Query Plan

q': SELECT c.name, p.name
FROM Cities c, Politicians p
WHERE c.population> '1M',
p.age<40,
p.name=c.currentMayor

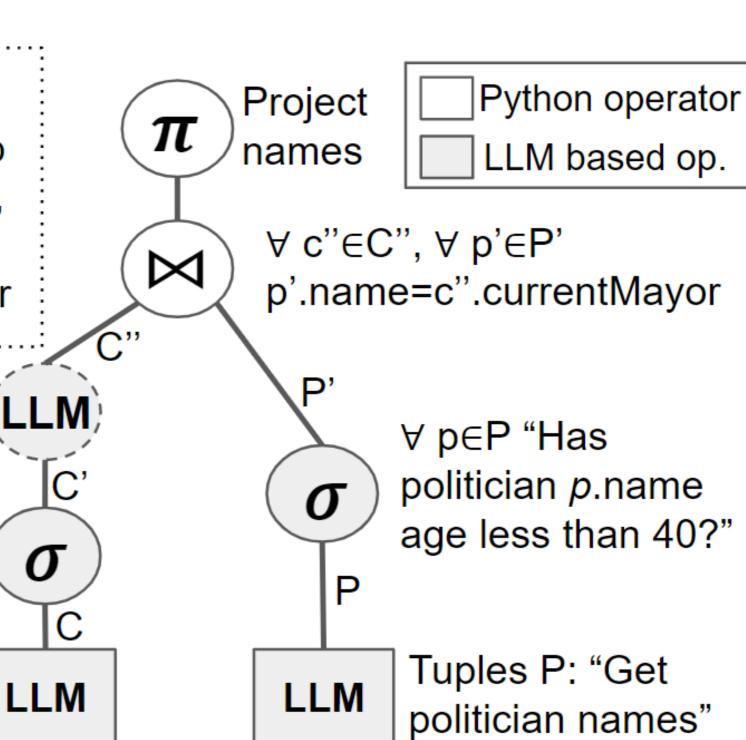
## Physical Query Plan

q': SELECT c.name, p.name FROM Cities c, Politicians p WHERE c.population> '1M', p.age<40, p.name=c.currentMayor

∀ c'∈C', c'.currentMayor = "Get current mayor of c'.name"

∀ c∈C, "Has city c.name more than 1M population?"

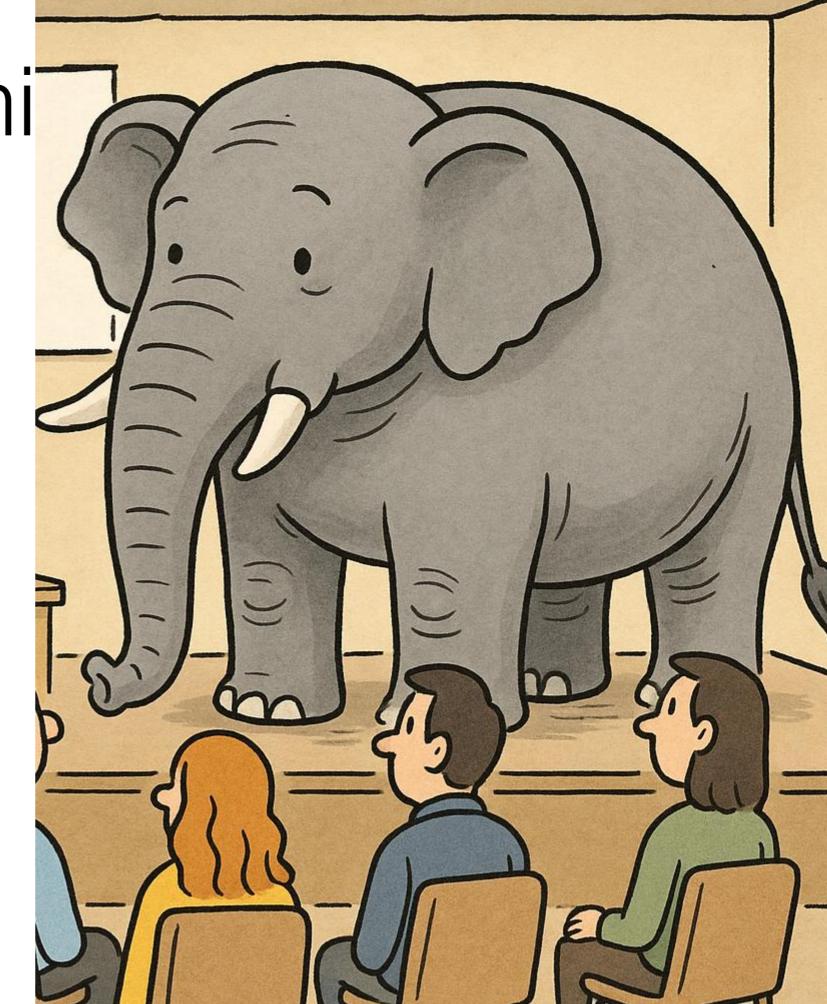
Tuples C: "Get city names"



σ

# Query optimi

- Physical: reduce errors
   → "get keys then values" vs "get tuples" scan operators
- Logical: Reduce LLM calls → push down selections ("get names of cities with > 1M population")
  - Optimize cost, quality.. building metadata/catalog from the LLM



### Factuality

- Decoder returns next token based on training data
- Such token may be based on either reliable acquired knowledge, or it may be a guess
  - → hallucinations
- + Models keep increasing the factuality of their answers\*
- + We observed the same pattern in Galois

#### Last updated on April 30th, 2024

Model	Hallucination Rate
GPT 4 Turbo	2.5 %
Snowflake Arctic	2.6 %
Intel Neural Chat 7B	2.8 %
GPT 4	3.0 %
Microsoft Orca-2-13b	3.2 %

https://github.com/vectara/hallucination-leaderboard

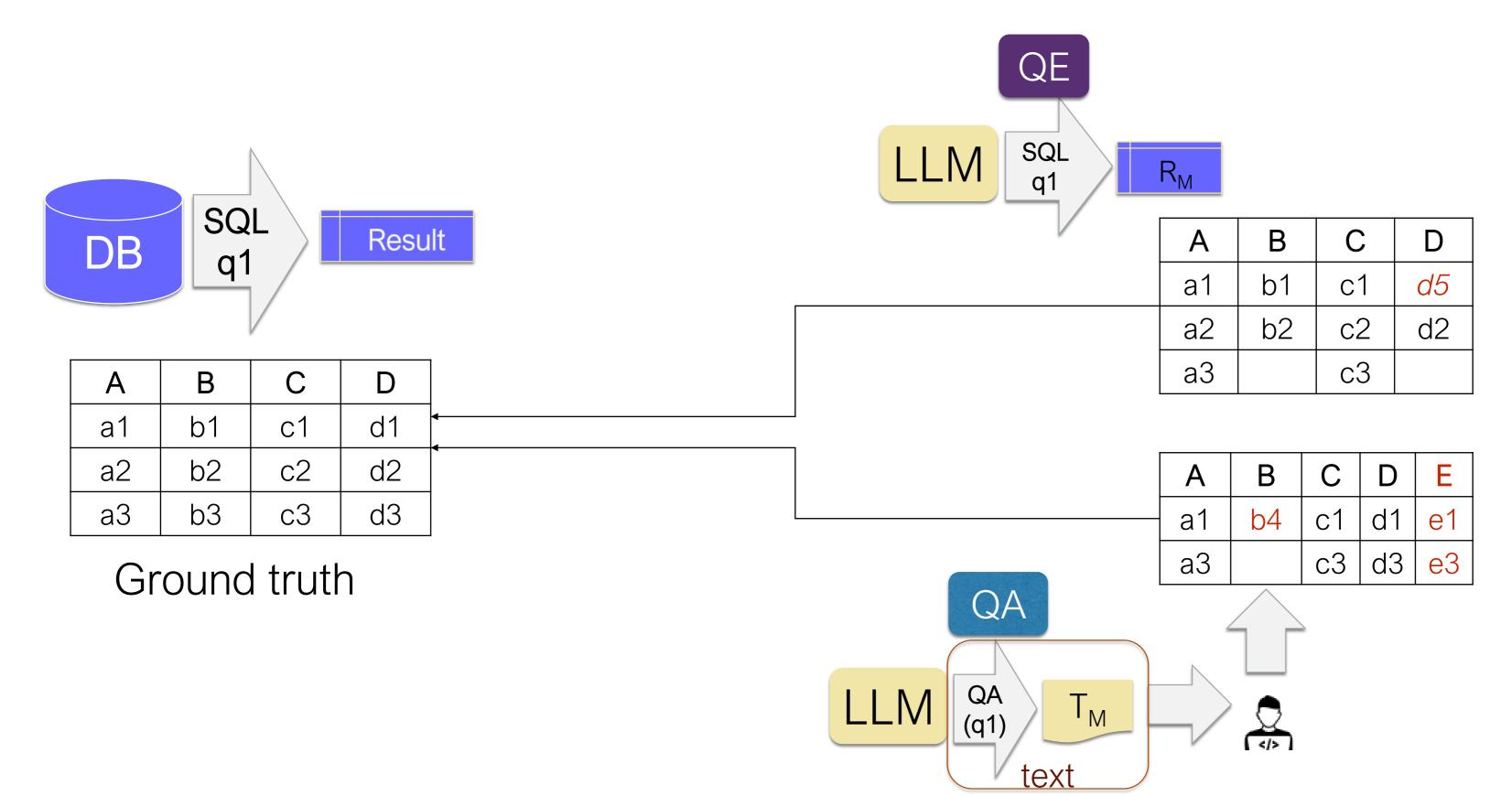
Model	Hallucination Rate
Google Gemini-2.0-Flash- 001	0.7 %
Google Gemini-2.0-Pro-Exp	0.8 %
OpenAl o3-mini-high	0.8 %
Vectara Mockingbird-2- Echo	0.9 %
Google Gemini-2.5-Pro- Exp-0325	1.1 %

<sup>\*&</sup>quot;GPT-4 scores 40% higher than GPT-3.5 on our factuality evaluations"

### Experiments Parametric Knowledge - Data

- Corpus of 92 SQL "reasonable" queries/questions from Spider (200 datasets)
  - No: "How many heads of the departments are older than 56?"
  - Yes: "What are the names of the countries that became independent after 1950?"

# Experiments – QA as "upper bound"



#### Results Llama 3.1 70B

- Matching in output results between ground truth and
  - Galois our method R<sub>M</sub> (SQL queries)

traditional  $T_M$  (NL questions) QA



$\circ$	$\mathcal{E}$
	$\frac{0}{0}$
<del>ol</del>	9
	$\frac{3}{2}$
rai	$\geq$
a O	$\leq$

Metric	NL	Galois	Galois Query Opt.
F1-cell	0.24	0.48	0.56 (2x)
Cardinality	0.46	0.66	0.84 (1.4x)
Tuple Constraint	0.07	0.45	0.47 (6.4x!)

LLMs do well in factual question answering, why low F1? Approx match vs Ilm as a judge

#### Results Llama 3.1 70B

- Matching in output results between ground truth and
  - Galois our method R<sub>M</sub> (SQL queries)
  - traditional T<sub>M</sub> (NL questions)

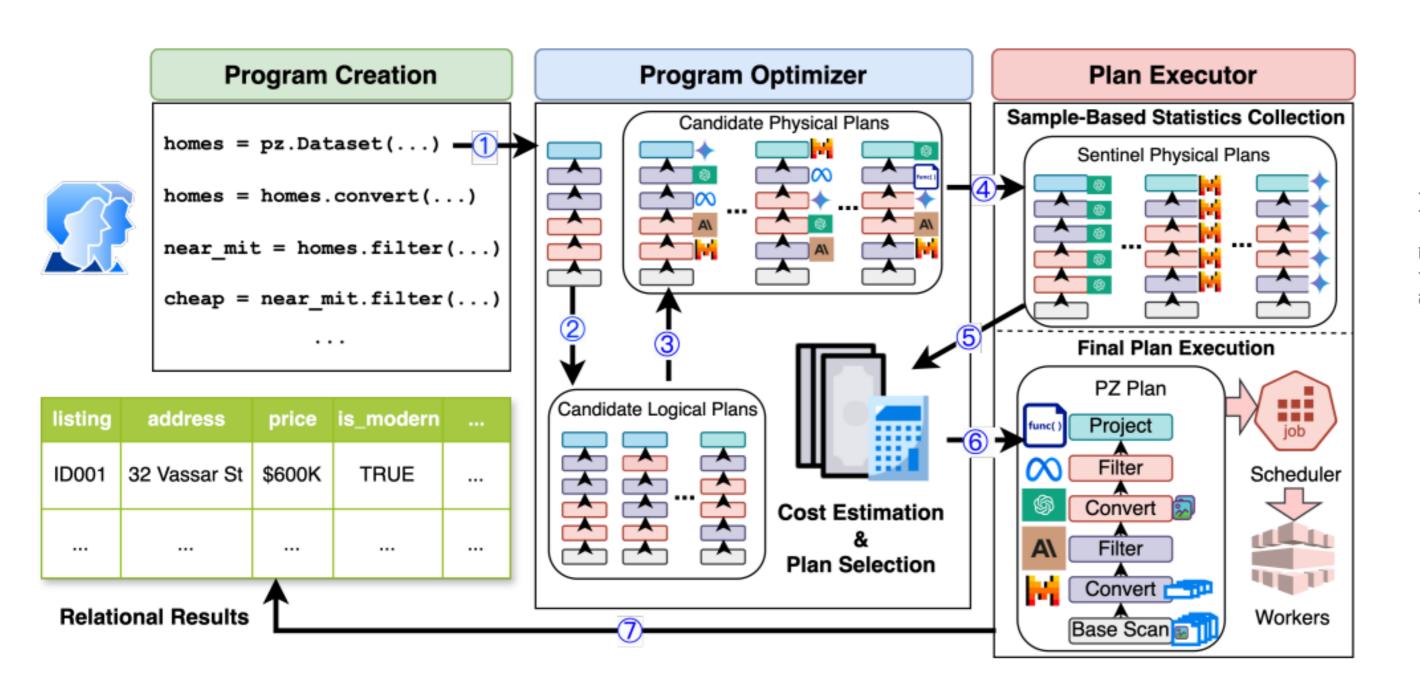
Parametric Knowledge

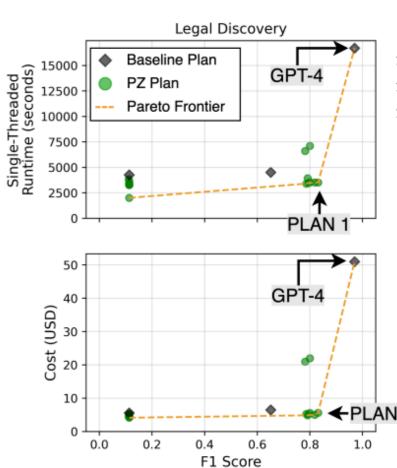
Metric	NL	Galois	Galois Query Opt.
F1-cell	0.24	0.48	0.56 (2x)
Cardinality	0.46	0.66	0.84 (1.4x)
Tuple Constraint	0.07	0.45	0.47 (6.4x!)

SAG

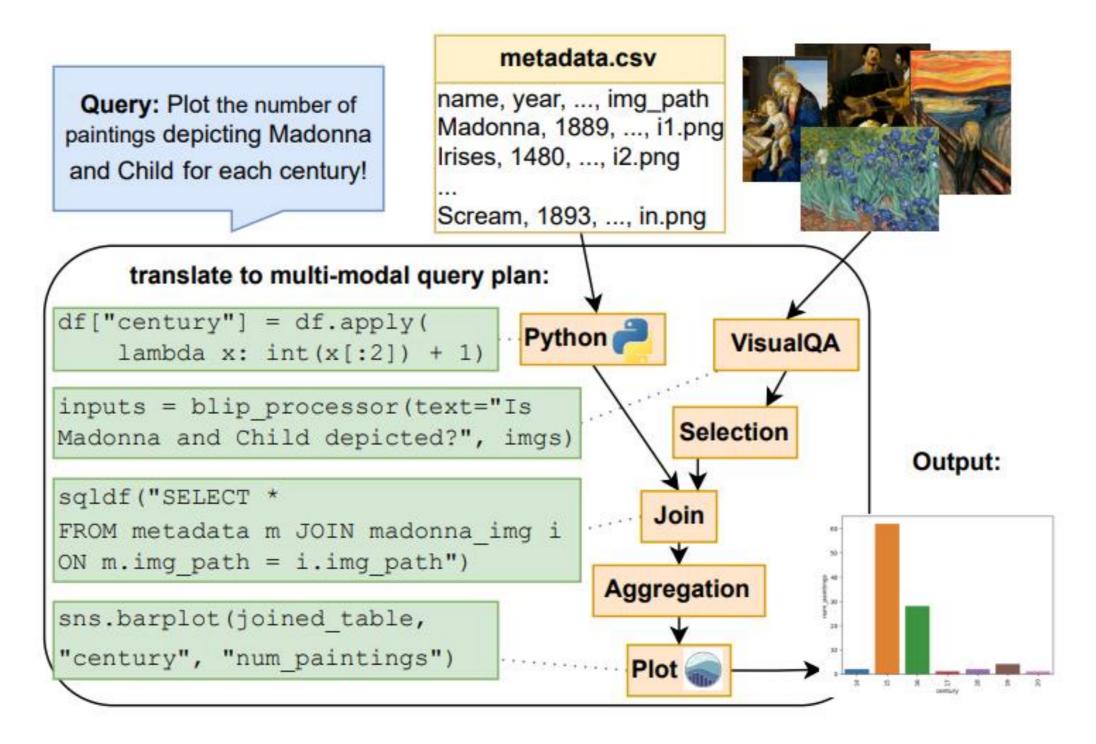
Metric	NL	Galois Query Opt.	Palimpzest (ETL)
AVG-Score	0.39	0.72	0.72
# Tokens in M	1.4	1.4	13.8

# Palimpzest





### CAESURA



### Open Questions

Logic controls LLM LLM controls logic

LLM *imitates* logic

#### DB first

use LLM in operators

Galois

[Jo and Trummer, 2023] [Urban et al, 2024]

[Liu et al, 2025]

#### LLMs + Agents

SP better results than TQA

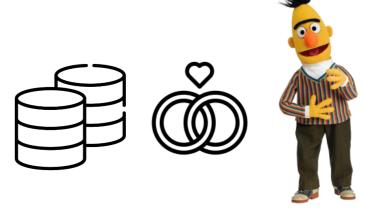
→ Use LM for NLU, SQL/code for data operations
[Arora et al, 2023]

[Pourreza et al, 2024]

#### LLM first

structured data in pretraining, extensions, fine tuning.... But fine tuned ChatGPT: 0.53 accuracy for TQA [Badaro et al, 2023] [Li et al, 2023]

#### SQL and LLMs?



2023 IEEE 39th International Conference on Data Engineering (ICDE)

#### Data Ambiguity Profiling for the Generation of Training Examples

Enzo Veltri University of Basilicata, Italy enzo.veltri@unibas.it

Gilbert Badaro EURECOM, France

Abstract—Several applications, such as text-to-SQL and computational fact checking, exploit the relationship between relational data and natural language text. However, state of the art solutions simply fail in managing "data-ambiguity", i.e., the case when there are multiple interpretations of the relationship between text and data. Given the ambiguity in language, text can be mapped to different subsets of data, but existing training corpora only have examples in which every sentence/question is annotated precisely w.r.t. the relation. This unrealistic assumption leaves the target applications unable to handle ambiguous cases. To tackle this problem, we present an end-to-end solution that, given a table D, generates examples that consist of text, annotated with its data evidence, with factual ambiguities w.r.t. D. We formulate the problem of profiling relational tables to identify row and attribute data ambiguity. For the latter, we propose a deep learning method that identifies every pair of data ambiguous attributes and a label that describes both columns. Such metadata is then used to generate examples with data ambiguities for any input table.

Paolo Papotti Mohammed Saeed EURECOM, France EURECOM, France gilbert.badaro@eurecom.fr mohammed.saeed@eurecom.fr paolo.papotti@eurecom.fr

Player	Team	FG%	3FG%	fouls	apps
Carter	LA	56	47	4	5
Smith	SF	55	50	4	7
Carter	SE	60	51	3	3

A DATA-AMBIGUOUS EXAMPLE CONTAINS THE SENTENCE "CARTER LA HAS HIGHER SHOOTING THAN SMITH SF" AND THE EVIDENCE UNDERLINED. ANOTHER EXAMPLE CONTAINS THE QUESTION "DID CARTER COMMIT 3 FOULS?" AND THE EVIDENCE IN ITALIC

against a relational table D as in Table I. Even as humans, it is hard to state if the sentence is true or false w.r.t. the data in D. The challenge is due to the two different meanings that can be matched to shooting: the claim can refer to attribute Field Goal (FG%) or to 3-point Field Goal (3FG%). The same challenge applies with a SQL query expressed in natural language such as "Did Carter commit 3 fouls?". We refer to this issue as data ambiguity, i.e., the existence of more than one interpretation of a text w.r.t. the data for a human reader.

#### https://github.com/enzoveltri/pythia





#### **QATCH:** Benchmarking SQL-centric tasks with Table Representation Learning Models on Your Data

Simone Papicchio Politecnico di Torino Turin, Italy

Paolo Papotti EURECOM Sophia Antipolis, France

Luca Cagliero Politecnico di Torino Turin, Italy

#### Abstract

Table Representation Learning (TRL) models are commonly pre-trained on large open-domain datasets comprising millions of tables and then used to address downstream tasks. Choosing the right TRL model to use on proprietary data can be challenging, as the best results depend on the content domain, schema, and data quality. Our purpose is to support end-users in testing TRL models on proprietary data in two established SQL-centric tasks, i.e., Question Answering (QA) and Semantic Parsing (SP). We present QATCH (Query-Aided TRL Checklist), a toolbox to highlight TRL models' strengths and weaknesses on relational tables unseen at training time. For an input table, QATCH automatically generates a testing checklist tailored to QA and SP. Checklist generation is driven by a SQL query engine that crafts tests of different complexity. This design facilitates inherent

#### https://github.com/spapicchio/QATCH





Vision Paper



#### Querying Large Language Models with SQL

Mohammed Saeed mohammed.saeed@eurecom.fr EURECOM France

Nicola De Cao ndecao@google.com Google AI

Paolo Papotti papotti@eurecom.fr EURECOM France

#### ABSTRACT

In many use-cases, information is stored in text but not available in structured data. However, extracting data from natural language (NL) text to precisely fit a schema, and thus enable querying, is a challenging task. With the rise of pre-trained Large Language Models (LLMs), there is now an effective solution to store and use information extracted from massive corpora of text documents. Thus, we envision the use of SQL queries to cover a broad range of data that is not captured by traditional databases (DBs) by tapping the information in LLMs. This ability would enable the hybrid querying of both LLMs and DBs with the SQL interface, which is more expressive and precise than NL prompts. To show the potential of this vision, we present one possible direction to ground it with a traditional DB architecture using physical operators for querying the underlying LLM. One promising idea is to execute some operators of the query plan with prompts that retrieve data from the LLM. For a large class of SQL queries, querying LLMs returns well structured relations, with encouraging qualitative results. We pinpoint several research challenges that must be addressed to build a DBMS that





born May 8, 1961 Chicago: Lori Lightfoot, born August 4, 1962

New York City: Bill de Blasio,

Figure 1: Querying a pre-trained LLM with SQL is different from question answering (QA). We assume a user SQL query as input. GALOIS executes the query, and obtains relations, by retrieving data from a LLM (1). The corresponding QA task consumes and produces natural language text (2).

#### https://github.com/dbunibas/galois





Data Science with Human in the Loop (DaSH) Workshop – 5th Sept 2025

### Size of the output matters

Q: What is the area of Maricopa county?A: 9 224 sq mi √

County	State	Area (sq mi)
Los Angeles	California	4 751
Cook	Illinois	1 635
Maricopa	Arizona	8 500 X

Longer output → tougher for all LLMs!

