

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

Paolo Papotti



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

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]



User Input:

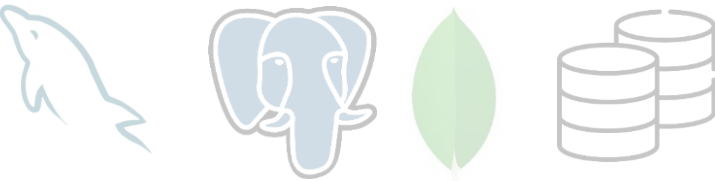
NL Prompt

SQL Query

Storage:

Documents

Relations



Question answering
(QA)

- 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”

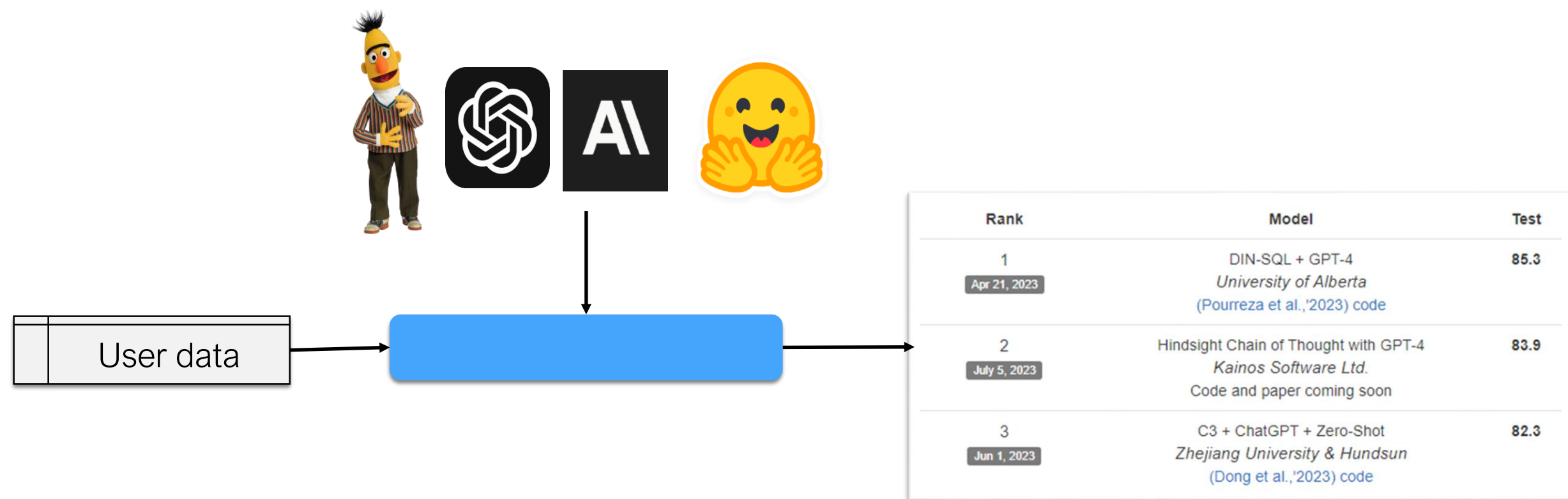
- A non-trivial instance of *NL text to code*
- LLMs do very well... according to results on public benchmarks

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)

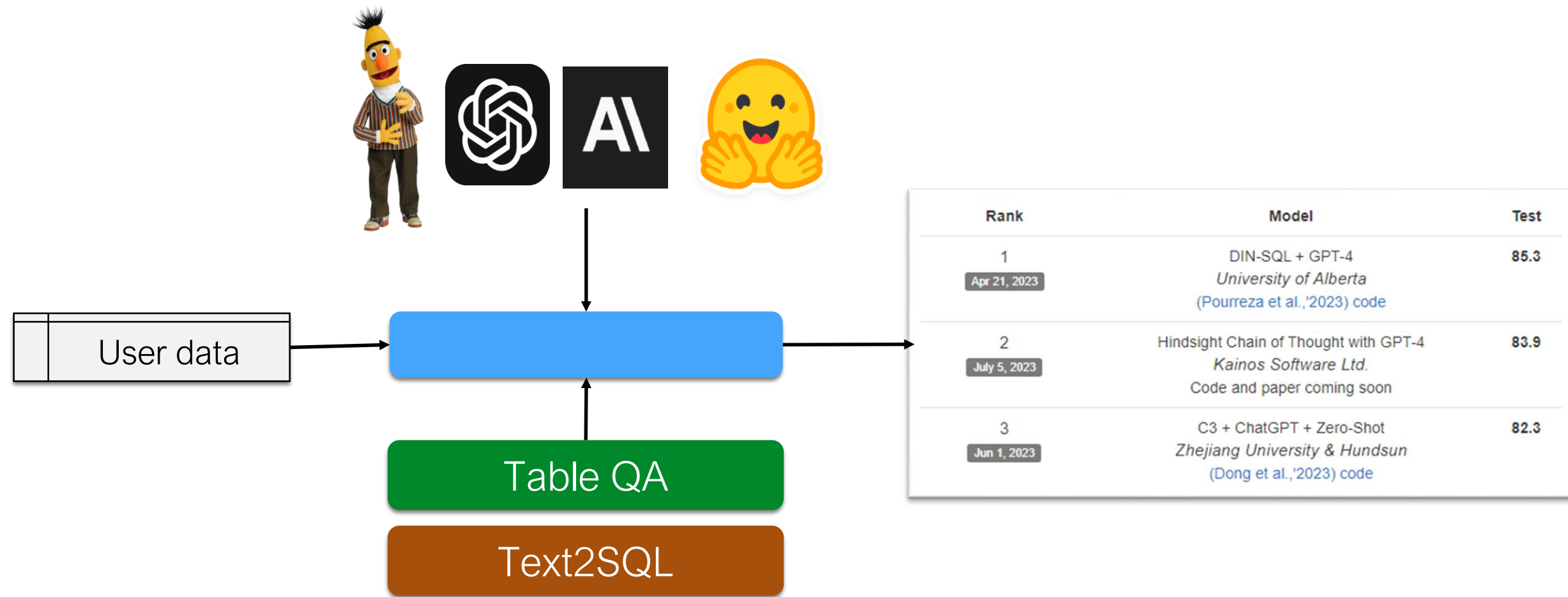


“Laure, Mike”

- 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?”

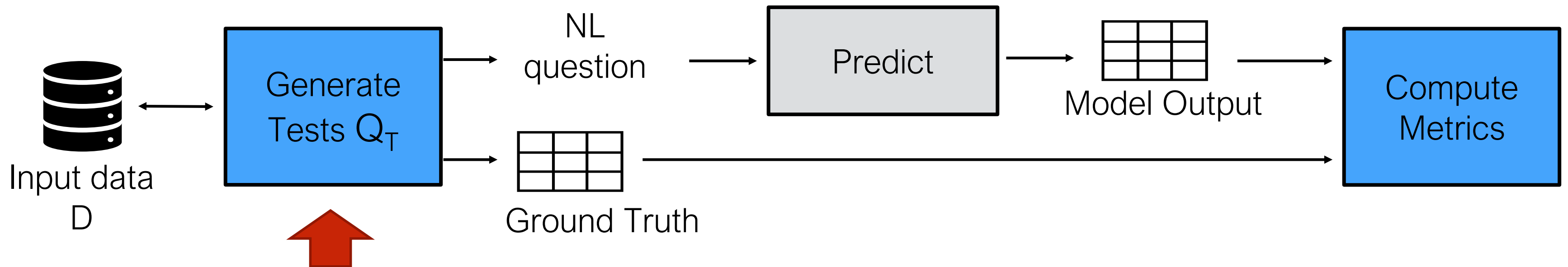
Custom benchmark on *user data*

- 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_T on D (NL question, result Ground Truth data)
 - Measure the quality of LLMs' output for Q_T and D



■ QATCH

■ LLM

How to get 'good' tests?

QATCH: Query-Aided TRL Checklist

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

