BENCHPRESS:

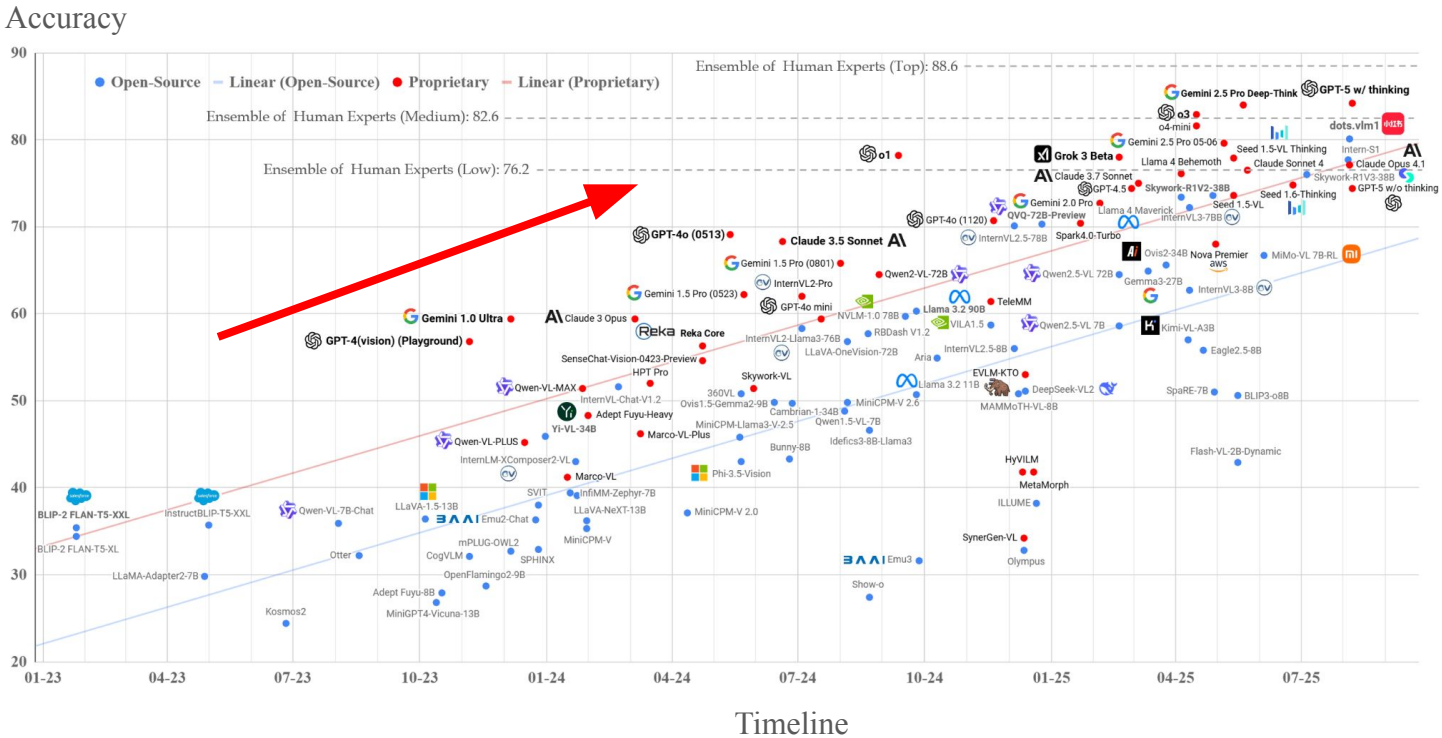
An Annotation System for Rapid Text-to-SQL Benchmark Curation

Fabian Wenz (TUM), Peter Baile Chen (MIT), Justin Choi (MIT), Cecil Gregg (MIT), Devin Yang (MIT), Omar Bouattour (TUM), NesimeTatbul (Intel Labs & MIT), Çağatay Demiralp (AWS AI Labs & MIT), Michael Stonebraker (MIT)

The *LLM Enterprise Group*

Making LLMs work for enterprise data tasks

LLMs Are Rapidly Improving



+LLMs are advancing rapidly

+Strong performance across many tasks:

- Question answering
- Code generation
- Data analysis

MMMU benchmark website (MMMU Benchmark Team, 2024), https://mmmu-benchmark.github.io, accessed Jan 18, 2026.

Text-to-SQL: A Critical Task for LLMs

Importance

- Databases remain the **source of truth**
- Most users cannot write SQL

LLMs enable Text-to-SQL:

Natural language access to data

Accessible data analysis

Reduced reliance on SQL experts

Public Benchmarks Fueled Rapid Progress

Existing Benchmarks

PUBLIC

95 **DBs**
Misc. **Domain**
1534 **Queries**

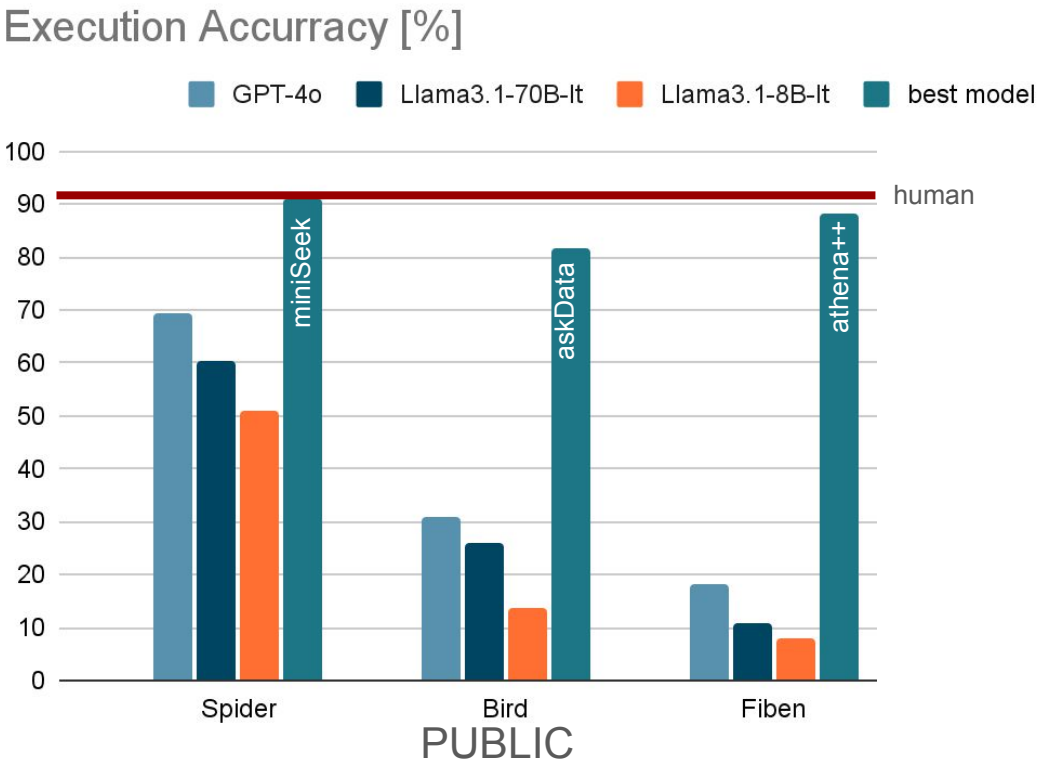
200 **DBs**
Misc. **Domain**
1034 **Queries**

300 **DBs**
Finance **Domain**
300 **Queries**

Enable:

- Rapid iteration
- Fair model comparison
- Reproducible research

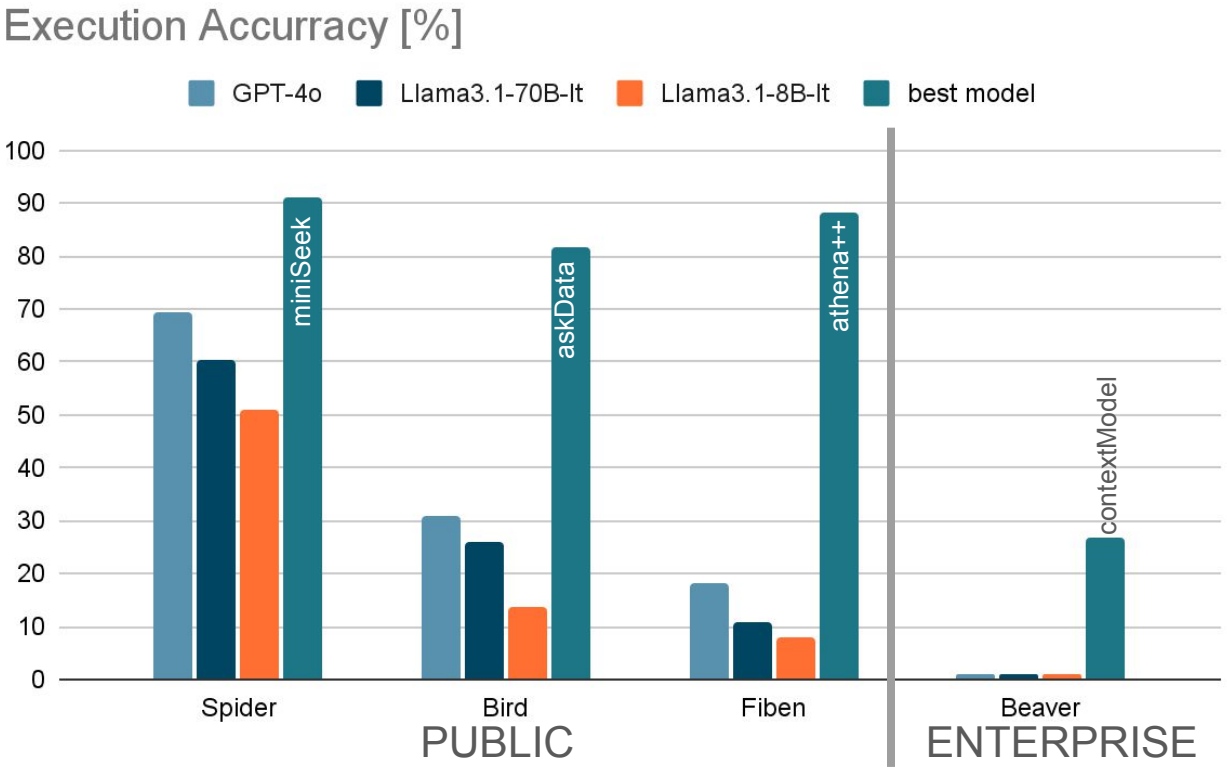
The Illusion of “Solved” Text-to-SQL



- **Near-perfect** results on public datasets
- Text-to-SQL appears ***“solved”***

But Public Benchmarks Are Reaching Their Limits

Models struggle with Enterprise

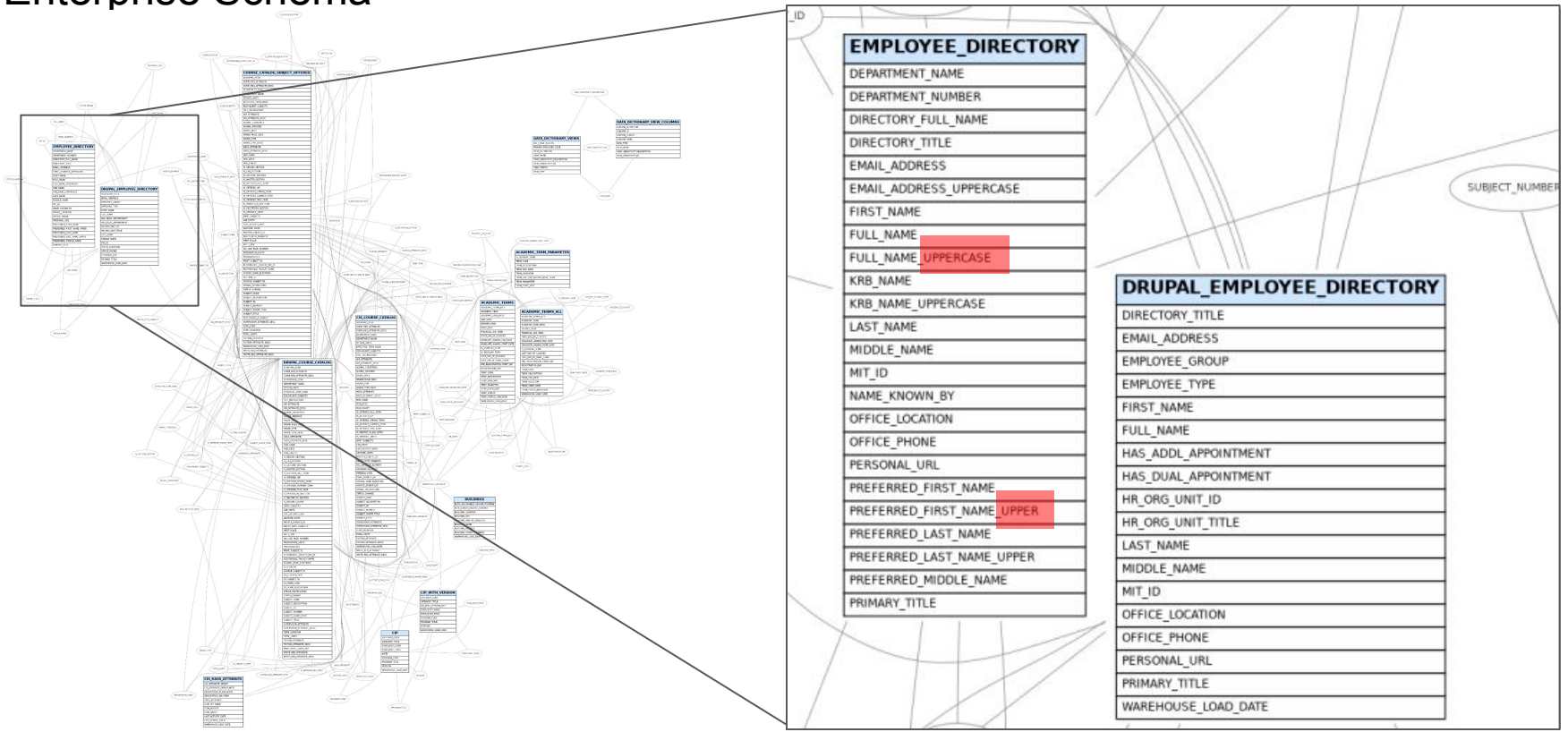


But this raises a question:

“Solved for whom?”

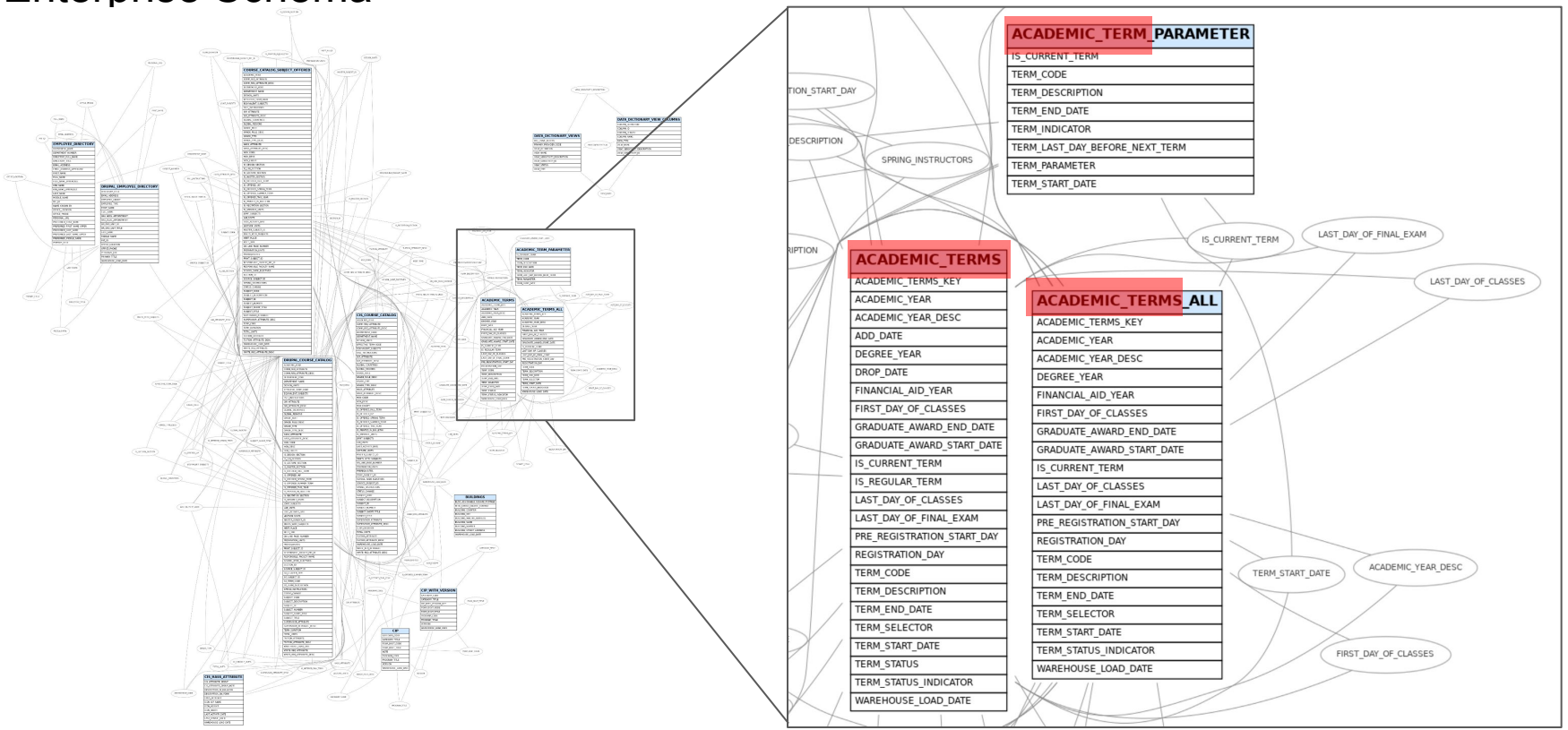
Enterprise Reality Is Fundamentally Different

Enterprise Schema



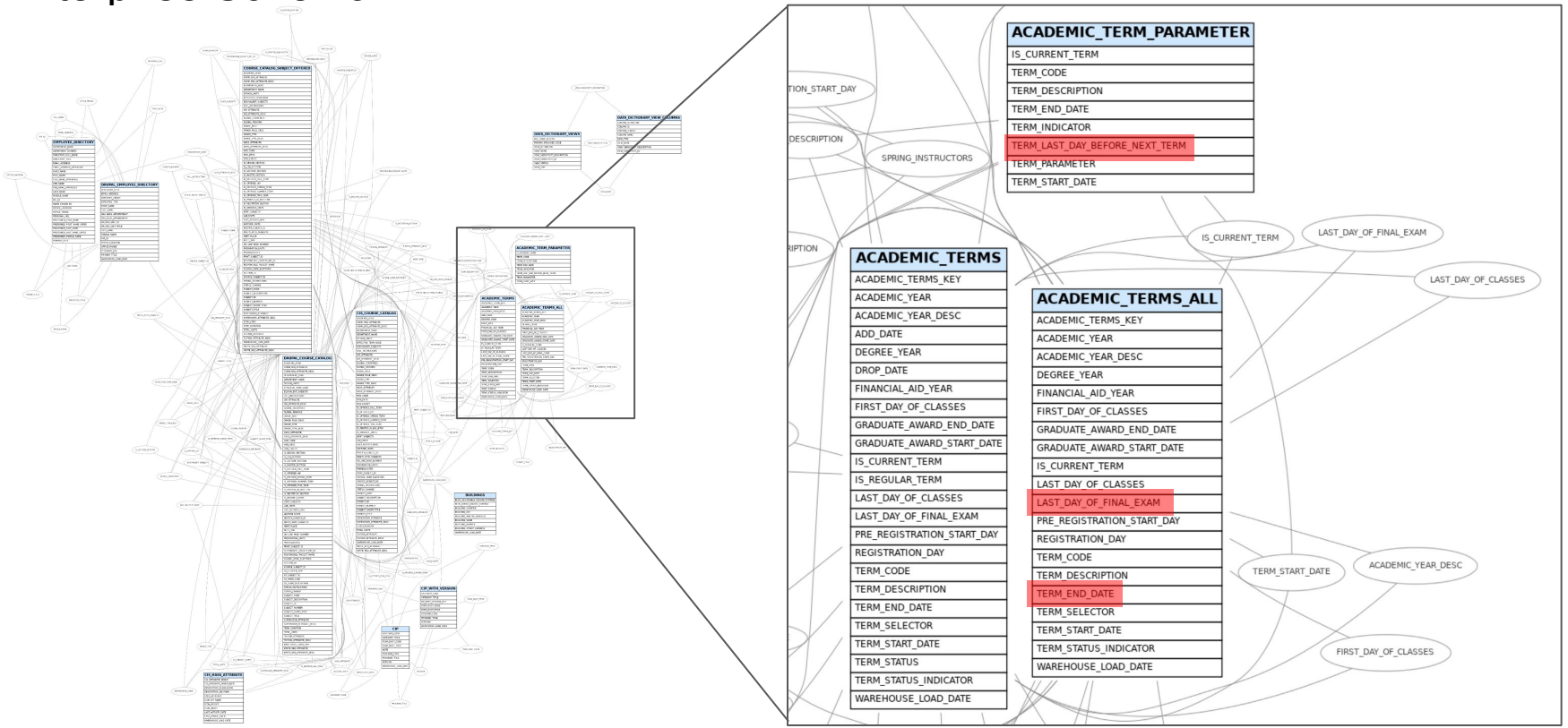
Enterprise Reality Is Fundamentally Different

Enterprise Schema



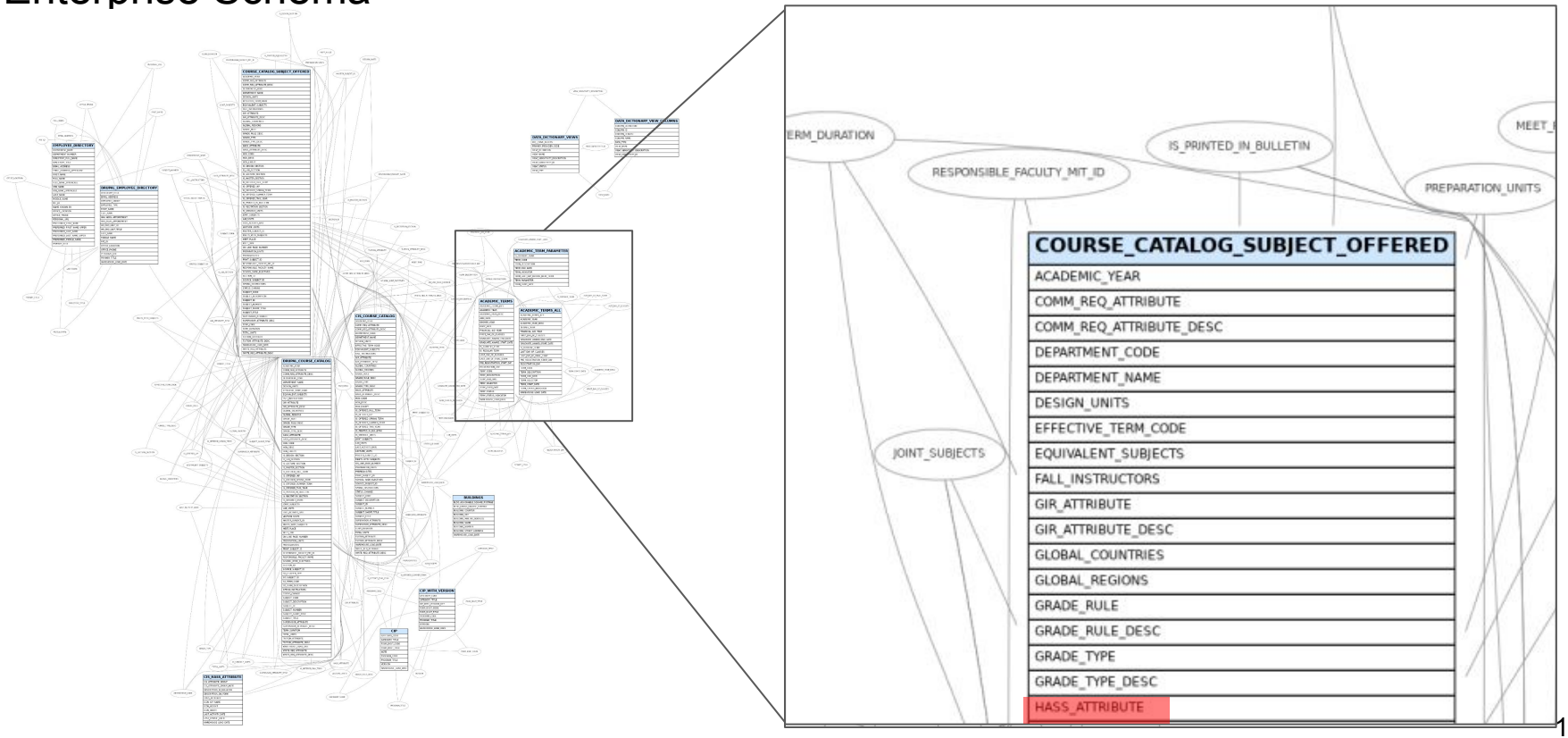
Enterprise Reality Is Fundamentally Different

Enterprise Schema



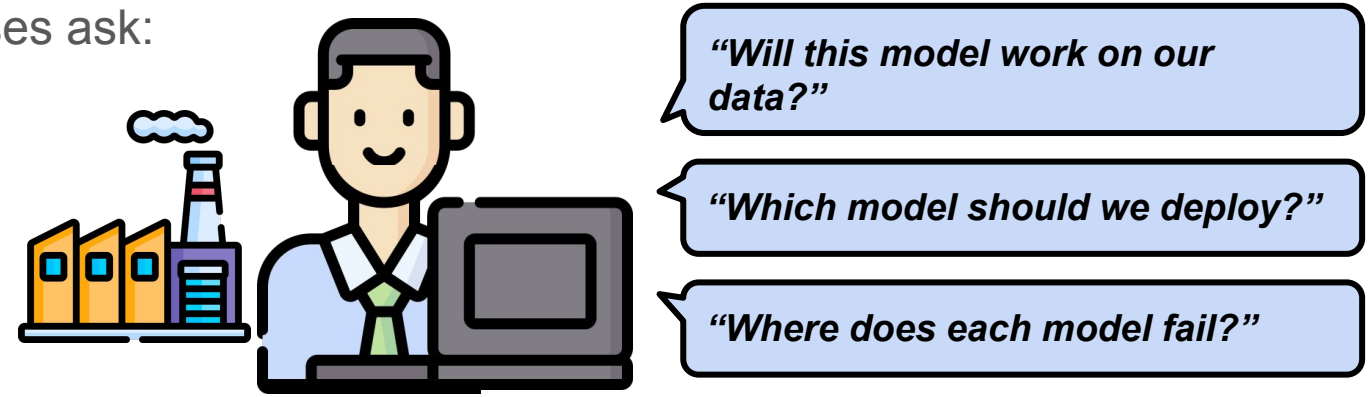
Enterprise Reality Is Fundamentally Different

Enterprise Schema



Enterprises Have a Benchmarking Problem

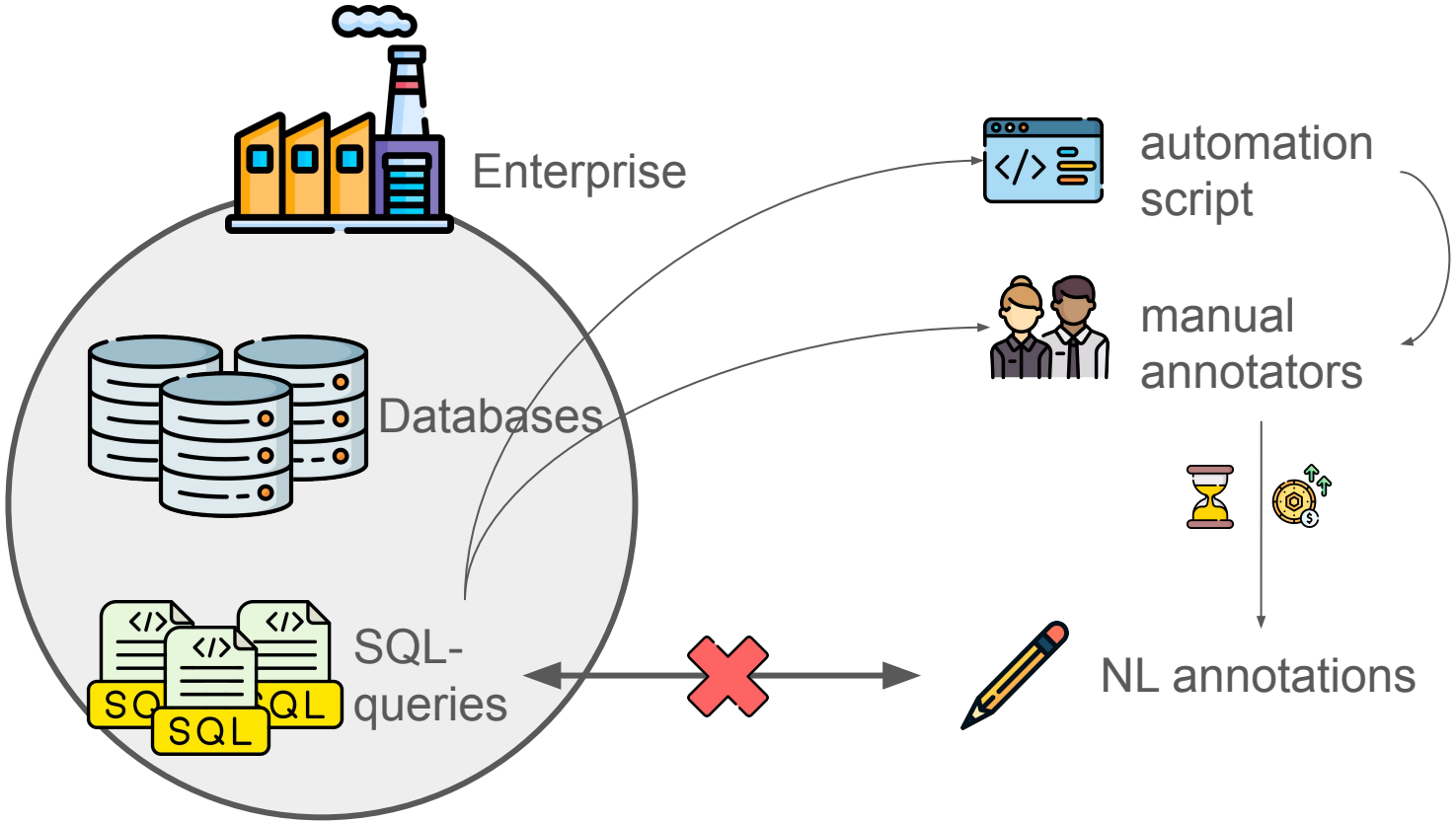
Enterprises ask:



But they lack

- Domain-specific benchmarks
- Realistic evaluation workloads

Why Enterprise Benchmarks Are Hard to Create



This Is the Missing Tooling Gap

Key Concepts Behind BenchPress

Existing tools focus on:

NL → SQL(model inference)

Enterprises need:

SQL → **NL**(benchmark creation)

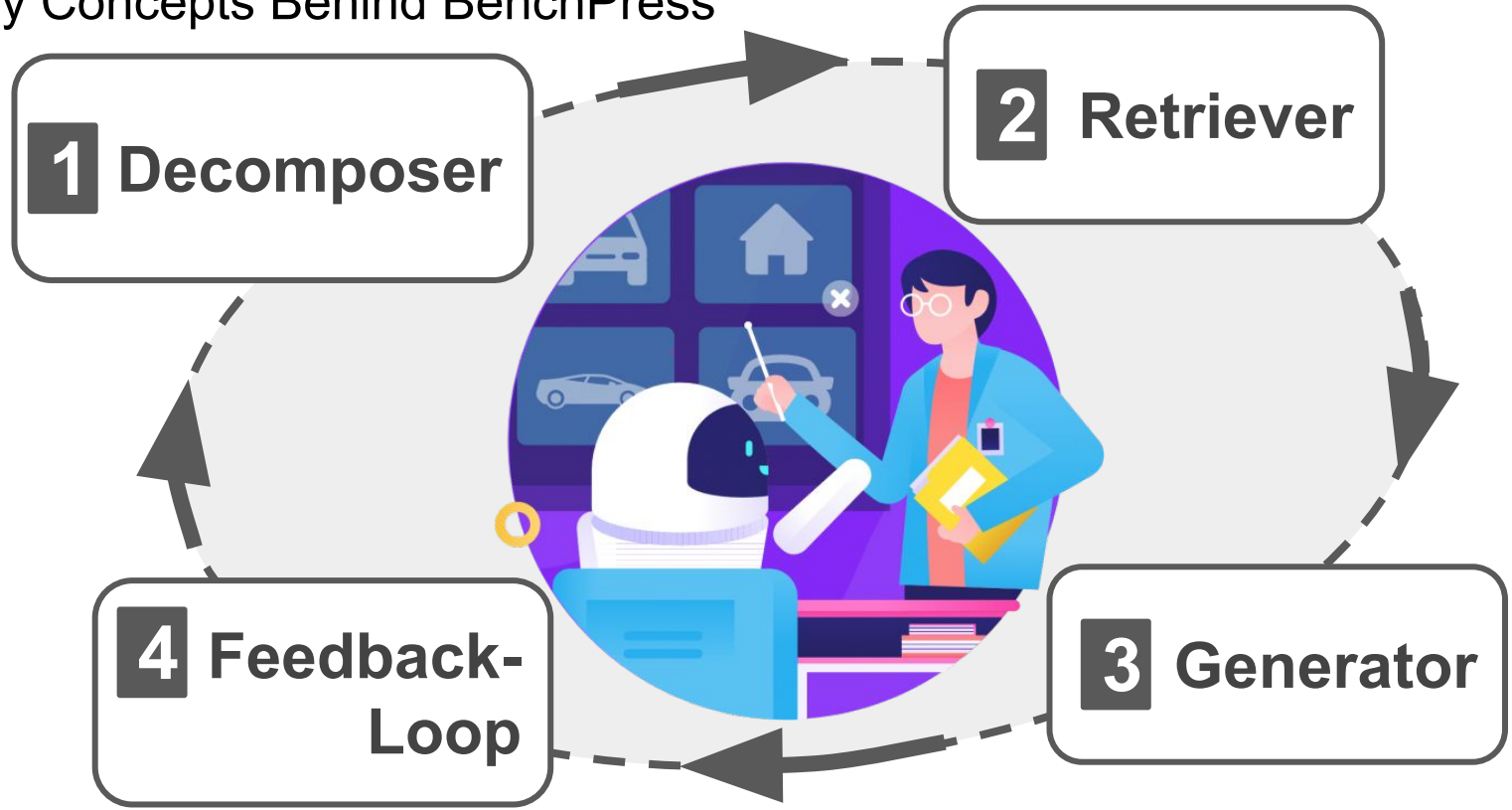
BENCHPRESS

Key insight:

- Benchmark creation, not modeling, is the bottleneck

Introducing BenchPress

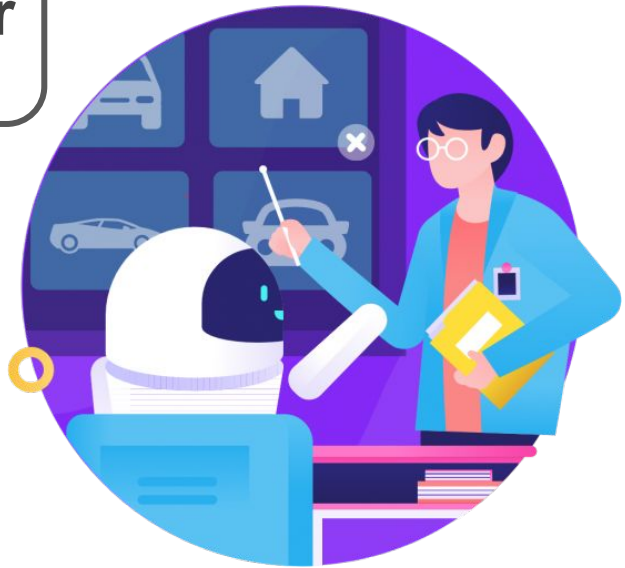
Key Concepts Behind BenchPress



Introducing BenchPress

Key Concepts Behind BenchPress

1 Decomposer

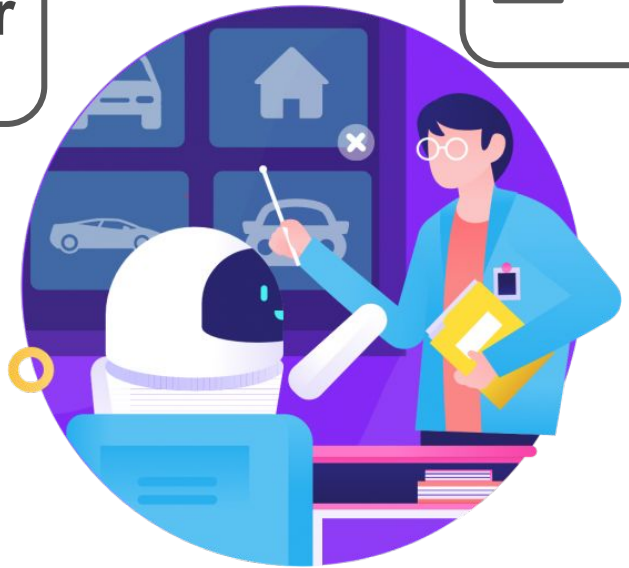


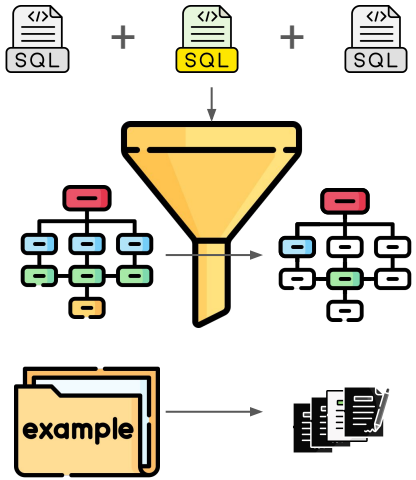
Introducing BenchPress

Key Concepts Behind BenchPress

1 Decomposer

2 Retriever





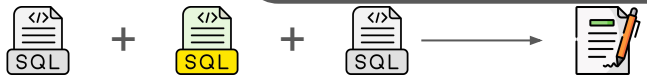
Introducing BenchPress

Key Concepts Behind BenchPress

1 Decomposer

2 Retriever

3 Generator



Introducing BenchPress

Key Concepts Behind BenchPress

1 Decomposer

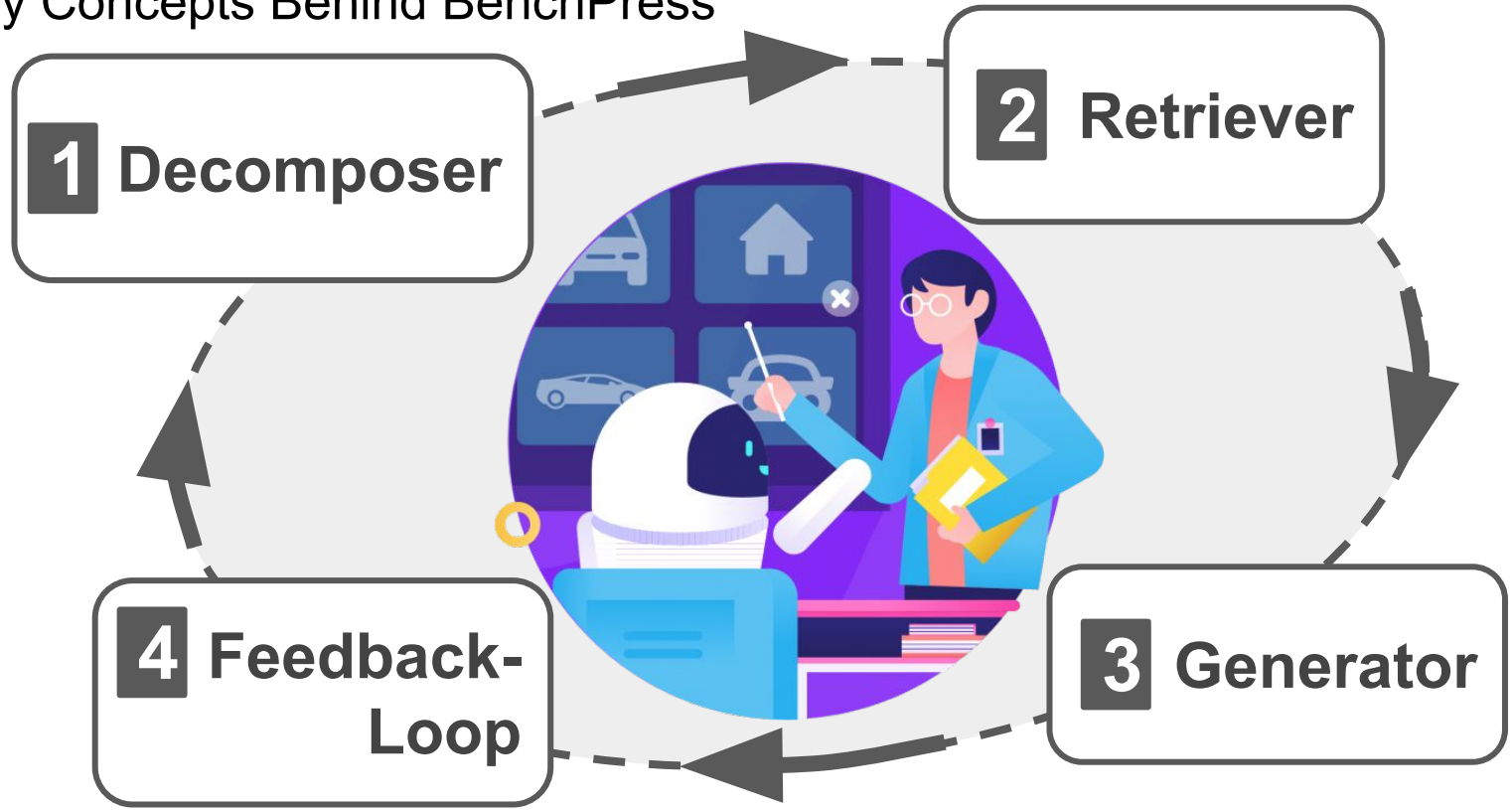
2 Retriever

3 Generator

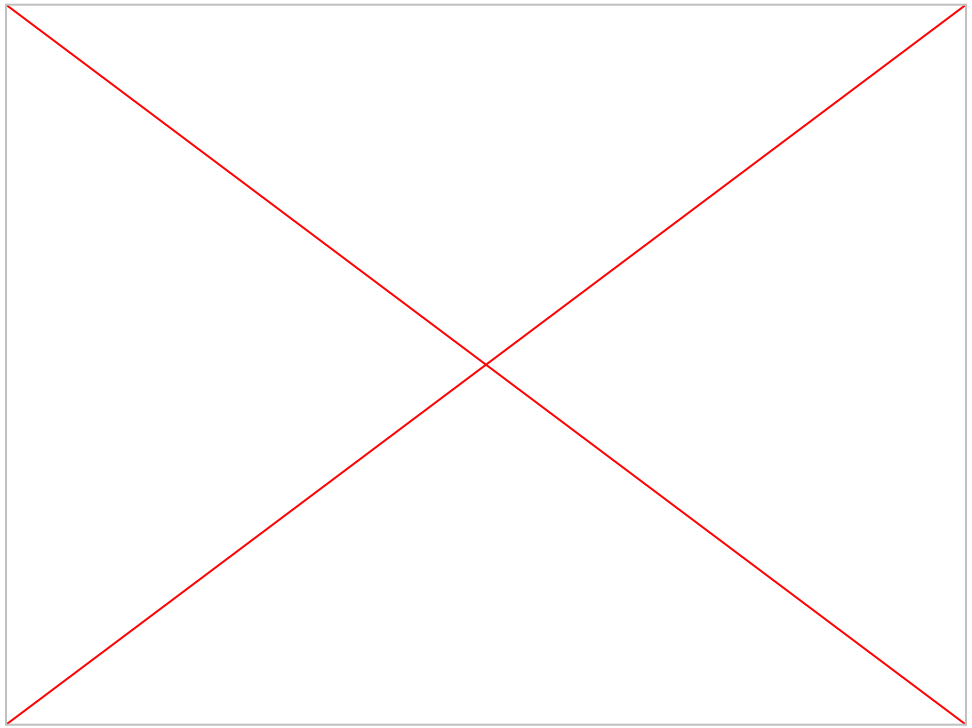
4 Feedback-Loop

Introducing BenchPress

Key Concepts Behind BenchPress

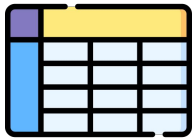


Short Demo



Experimental Design Process

- 

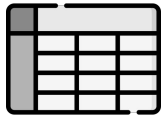
SQL query +  SQL output

Acc on Beaver:

Experimental Design Process

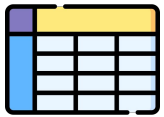
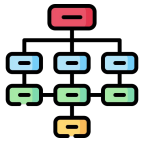
Acc on Beaver:

- 

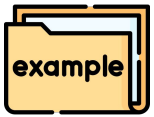
SQL query +  SQL output

⬆️28.9

- 

query +  output +  schema

⬆️41.7

+ 

examples

Experimental Design Process

Acc on Beaver:

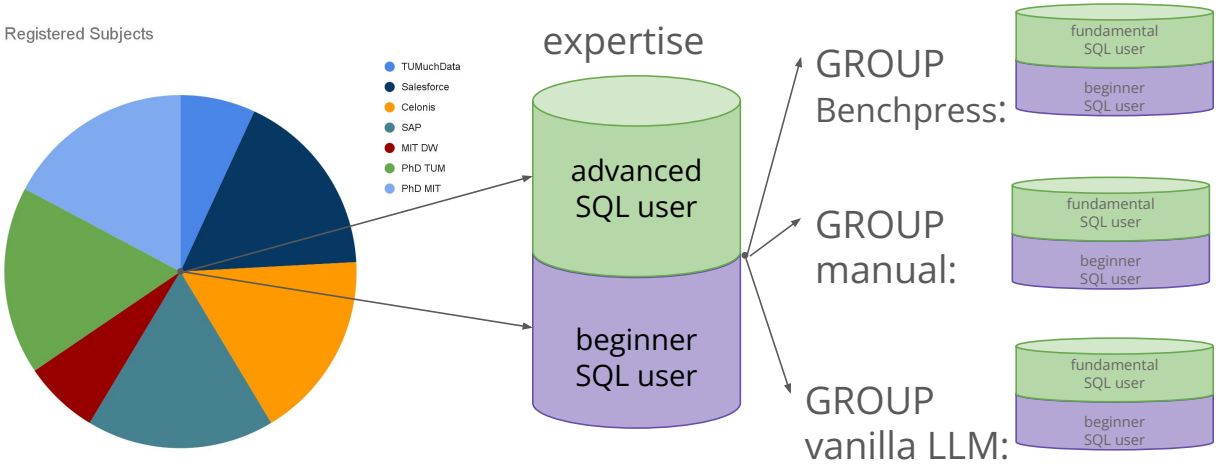
- SQL query + SQL output — ⬆28.9
- query + output + schemas + examples — ⬆41.7
- decompose:
  - query1 + output1 + schemas1 + examples1
  - …

⬆68.0

User-Study: SQL-to-Text

Overview

- **Goal**: Evaluate the effectiveness of BenchPress for generating high-quality natural language annotations for SQL queries.
- **Task**: Write a natural language description that explains what the given SQL query does.

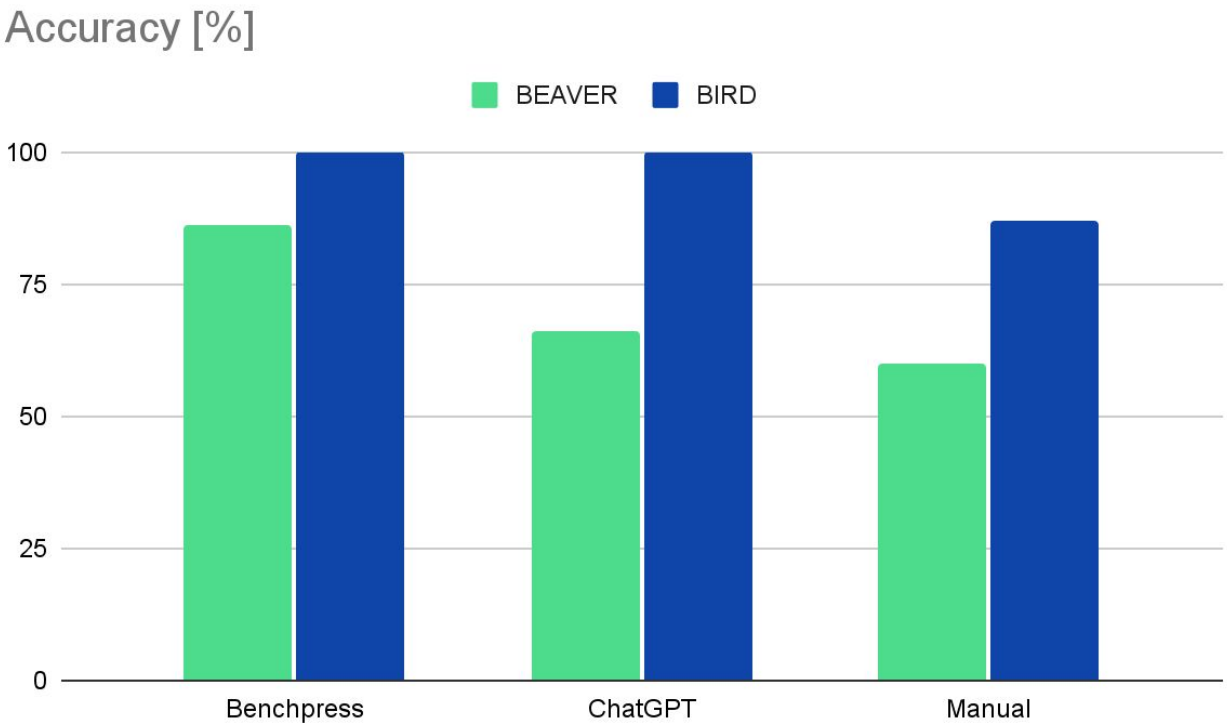


Study **measurements:**

- speed,
- accuracy,
- clarity

Results 🎯

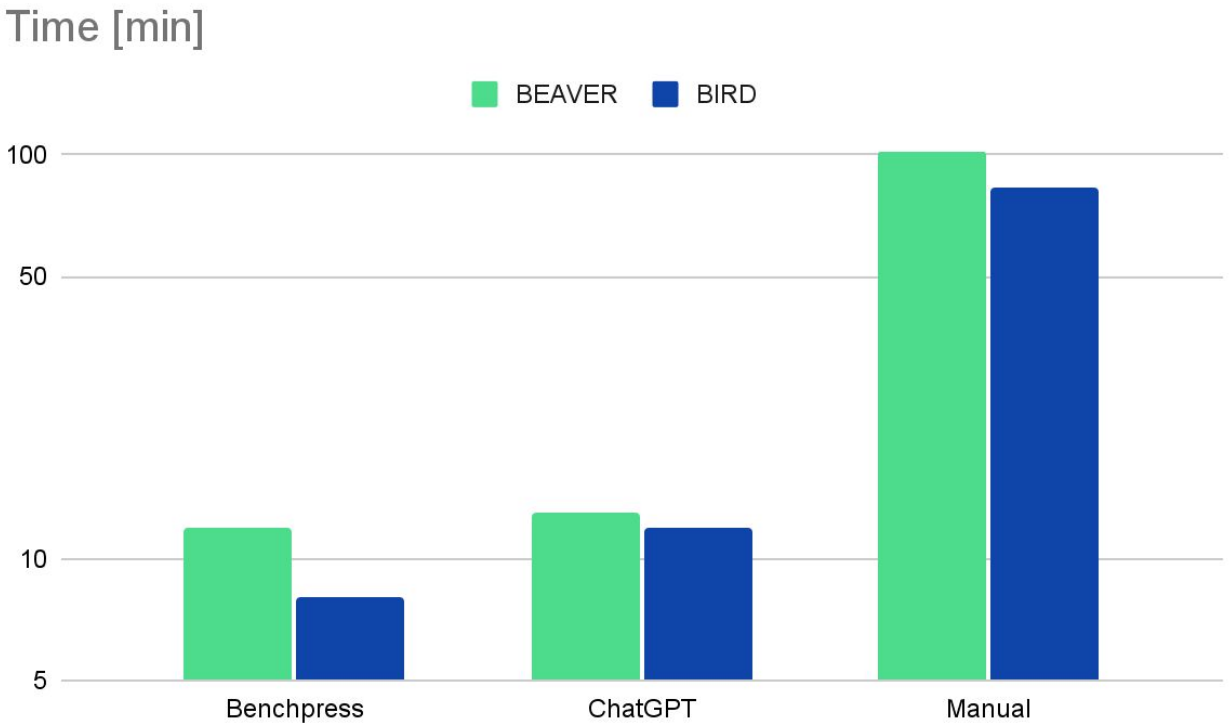
Accuracy: how often on average was the NL correct?



- **Public benchmark** accuracy is high across all methods.
- **Enterprise** (BEAVER) accuracy drops without BenchPress.

Results ⏱️

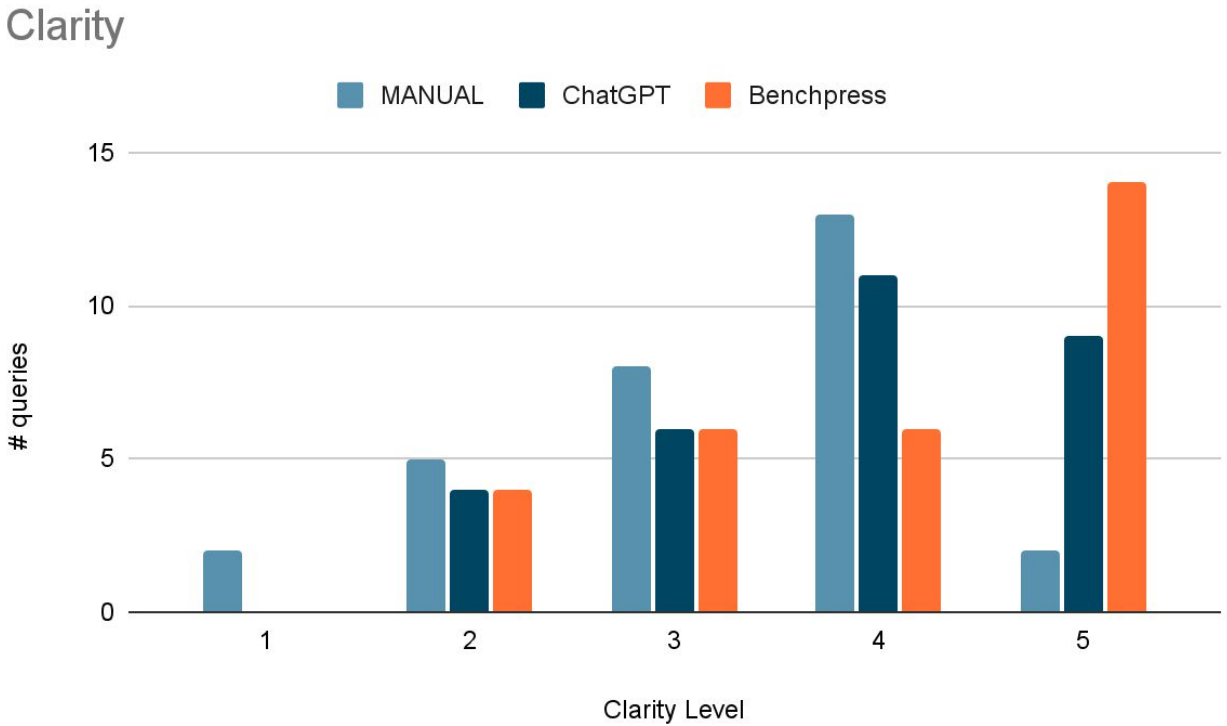
Time: how long on average to create all NL for all given SQL?



- **Public benchmark** are being annotated quicker across all methods.

Results

Clarity: how well did the backtranslation work?



Level 1: Doesn't run

Level 2: Runs but wrong join or wrong tables

Level 3: Runs but wrong columns/elements

Level 4: Runs but ordering issues or missing ORDER nuance

Level 5: Fully Correct

Conclusions

Thank you!

- **BenchPress** is a novel interactive system for Text-to-SQL benchmark generation
- **accelerates** enterprise benchmark creation while facilitating quality
- allows **robustness** checks for potentially flawed benchmarks
- enables scalable, realistic model **evaluation**

Thank You For your Attention!

References & Credits

- **Icons**: Figures use icons from Flaticon, created by Freepik and Muhammad Usman
- **Benchmark Progress**: MMMU Benchmark — tracking multimodal model improvements over time
- **Related CIDR Work**:
  - Jin et al., CIDR 2026 — Text-to-SQL Benchmarks are Broken: An In-Depth Analysis of Annotation Errors
  - Narasayya et al., CIDR 2026 — Leveraging Query Optimizers to Verify the Soundness of LLM-based Query Rewrites for Real-World Workloads, and More