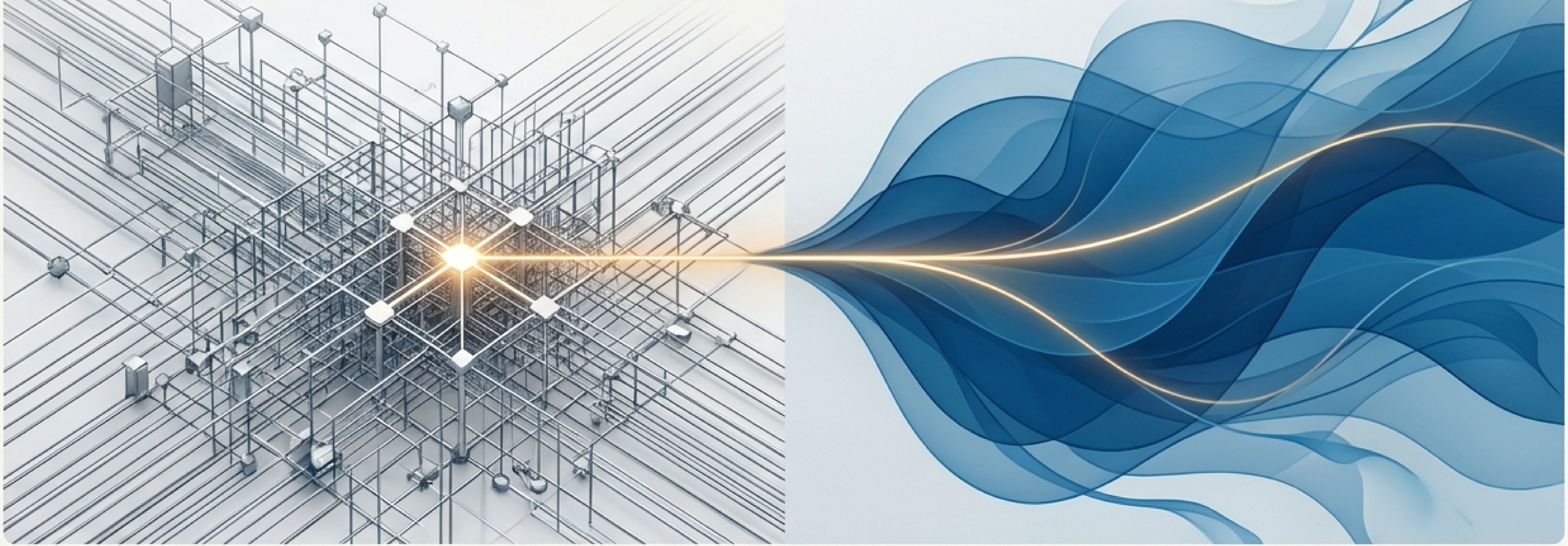
To Unlock Trustworthy AI for Enterprise Data, We Must Solve the Context Problem.

A benchmark analysis of Large Language Model accuracy for question answering on enterprise SQL databases, and the critical role of Knowledge Graphs



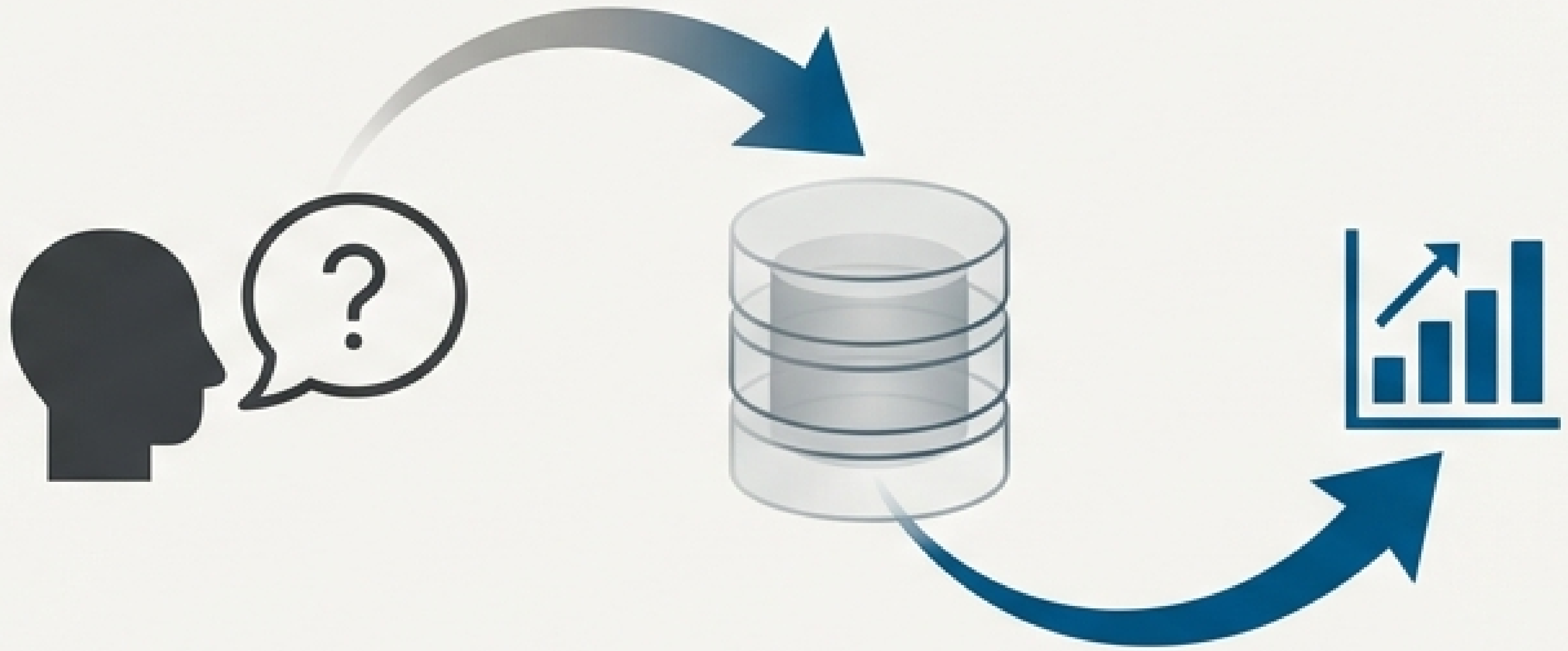
NotebookLM

The Universal Goal is to "Chat with Your Data" for Faster, Better Decisions

The promise of Generative AI is transformative: enabling any user, from an executive to an analyst, to ask complex questions in natural language and receive accurate, data-backed answers instantly.

This capability has the potential to fundamentally change how data-driven decisions are made.

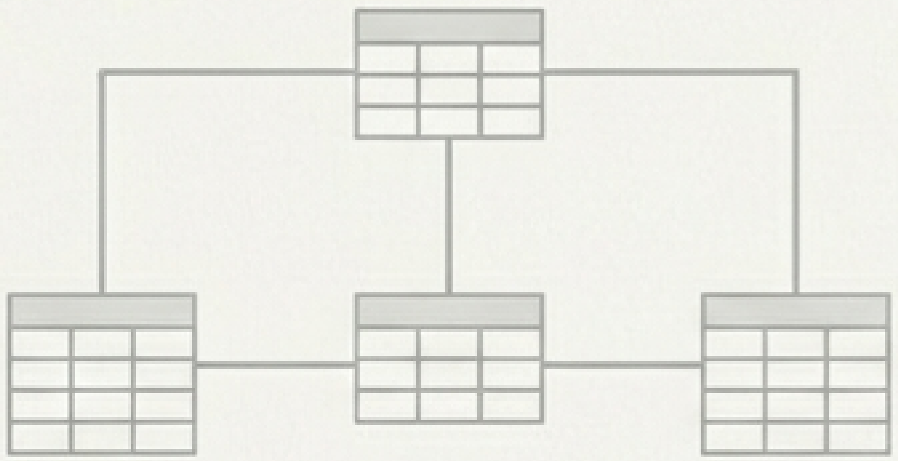
- Democratize data access
- Accelerate operational and strategic planning
- Unlock insights hidden in complex databases



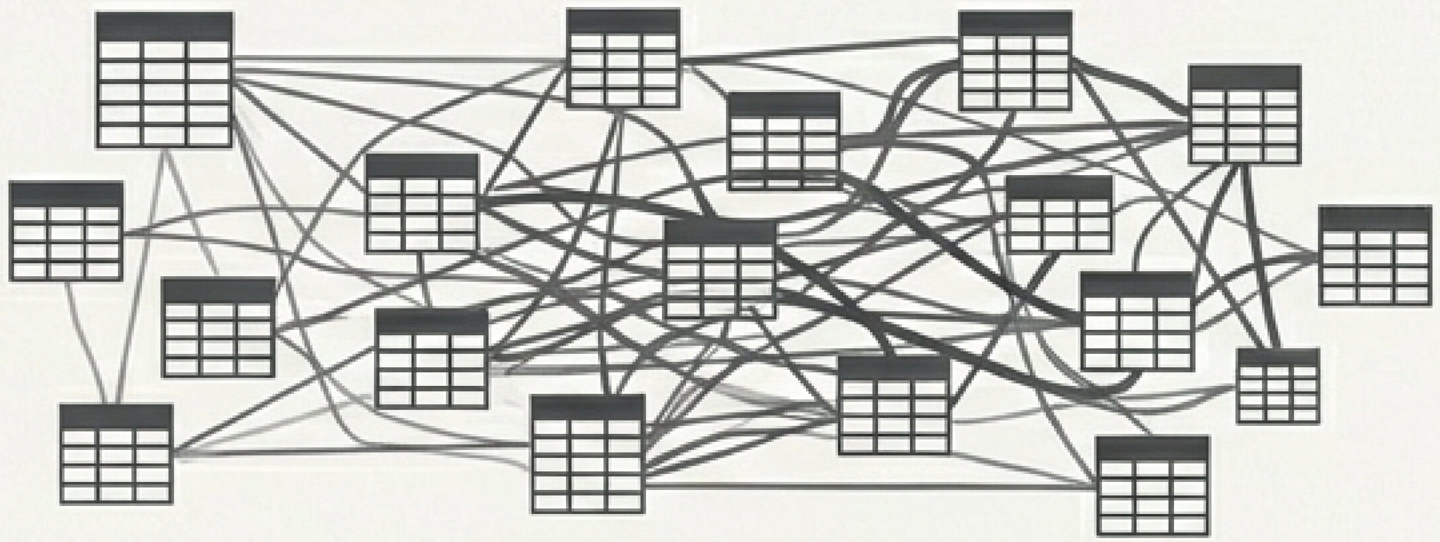
NotebookLM

Existing AI Benchmarks Don’t Reflect the Reality of Enterprise Data.

While LLMs show remarkable performance on public Text-to-SQL benchmarks like Spider or WikiSQL, these are misaligned with typical enterprise settings. This disconnect creates a false sense of security and leads to poor outcomes in production.



Typical Benchmark



Enterprise Reality

1. Schema Complexity

Benchmarks overlook complex database schemas that often comprise hundreds of tables.

2. Question Complexity

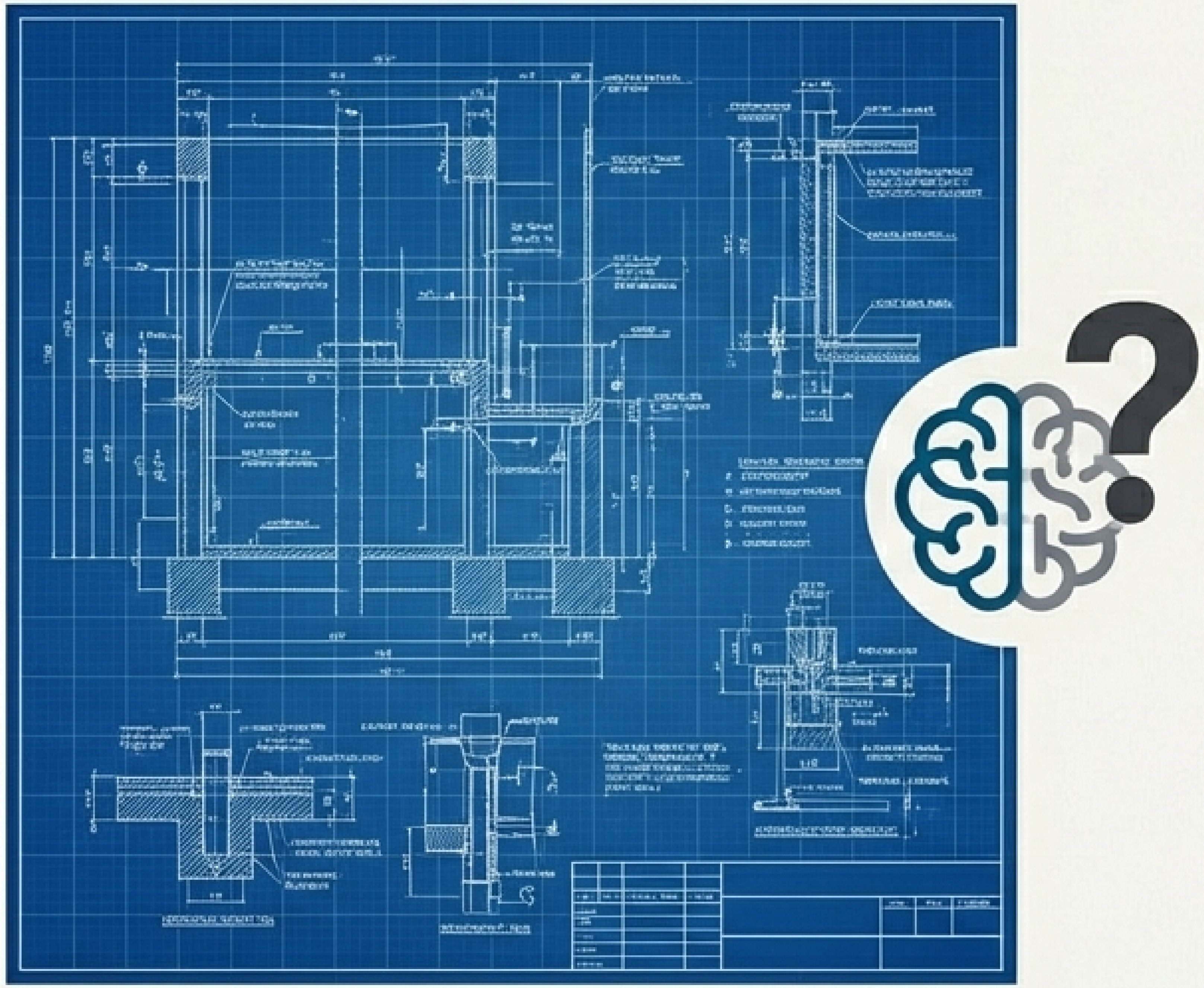
They disregard crucial business questions related to reporting, metrics, and KPIs.

3. Missing Context

They lack a business context layer (metadata, semantics, ontologies) that defines what the data actually **means**.

NotebookLM

An LLM Sees a SQL Database as a Technical Blueprint, Not a Business Map.



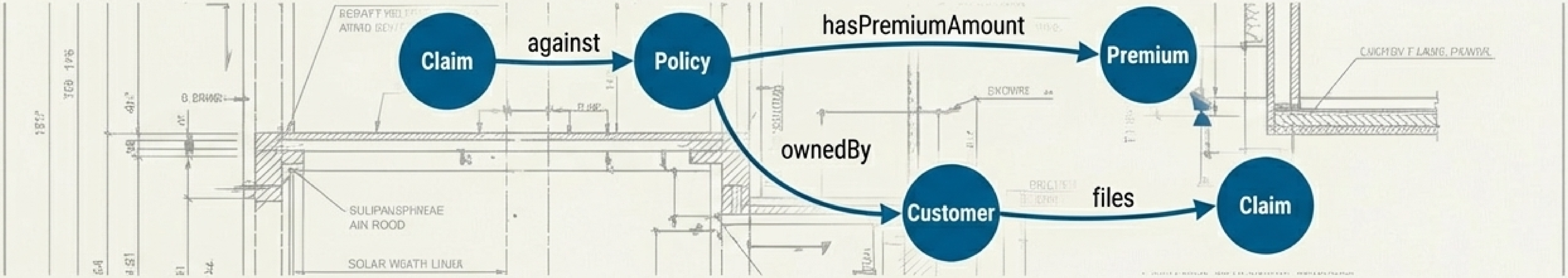
A SQL Data Definition Language (DDL) file describes the physical structure of the database—tables, columns, keys. It is a precise blueprint for an engineer. For an LLM, which lacks inherent business knowledge, this blueprint is cryptic. It knows the structure but not the *meaning* or the relationships between business concepts.

The Result:

- **Hallucinations**: Inventing columns, values, or joins that seem plausible but are factually wrong.
- **Uncontrolled Outcomes**: Generating inaccurate queries that produce dangerously misleading answers.

NotebookLM

The Solution is a Knowledge Graph: A Context Layer That Makes Business Logic Explicit.



Knowledge Graphs (KGs) fill the business context gap. By providing an explicit, machine-readable model of the business domain—its concepts, relationships, and rules—a KG gives the LLM the "map" it needs to navigate the data correctly.

Key Components of the Context Layer:

- **Ontology:** Defines business concepts (e.g., 'Claim', 'Policy', 'Premium') and their relationships.
- **Mappings:** Connects the concepts in the ontology to the physical tables and columns in the SQL database.

> **"Knowledge graphs provide the perfect complement to LLM-based solutions where high thresholds of accuracy and correctness need to be attained."**
>
> — Gartner, July 2023

NotebookLM

We Built an Enterprise-Grade Benchmark to Quantify the Impact of Knowledge Graphs

To test the hypothesis that KGs improve LLM accuracy, we developed a benchmark designed to mirror a real-world enterprise environment. We used GPT-4 with zero-shot prompting in all experiments.

1. Enterprise SQL Schema

Based on the OMG Property and Casualty Data Model, a complex insurance industry standard (subset of 13 tables from the full 199-table model).

2. Enterprise Questions

43 natural language questions covering a range of typical business needs, from simple reporting to complex KPI calculations.

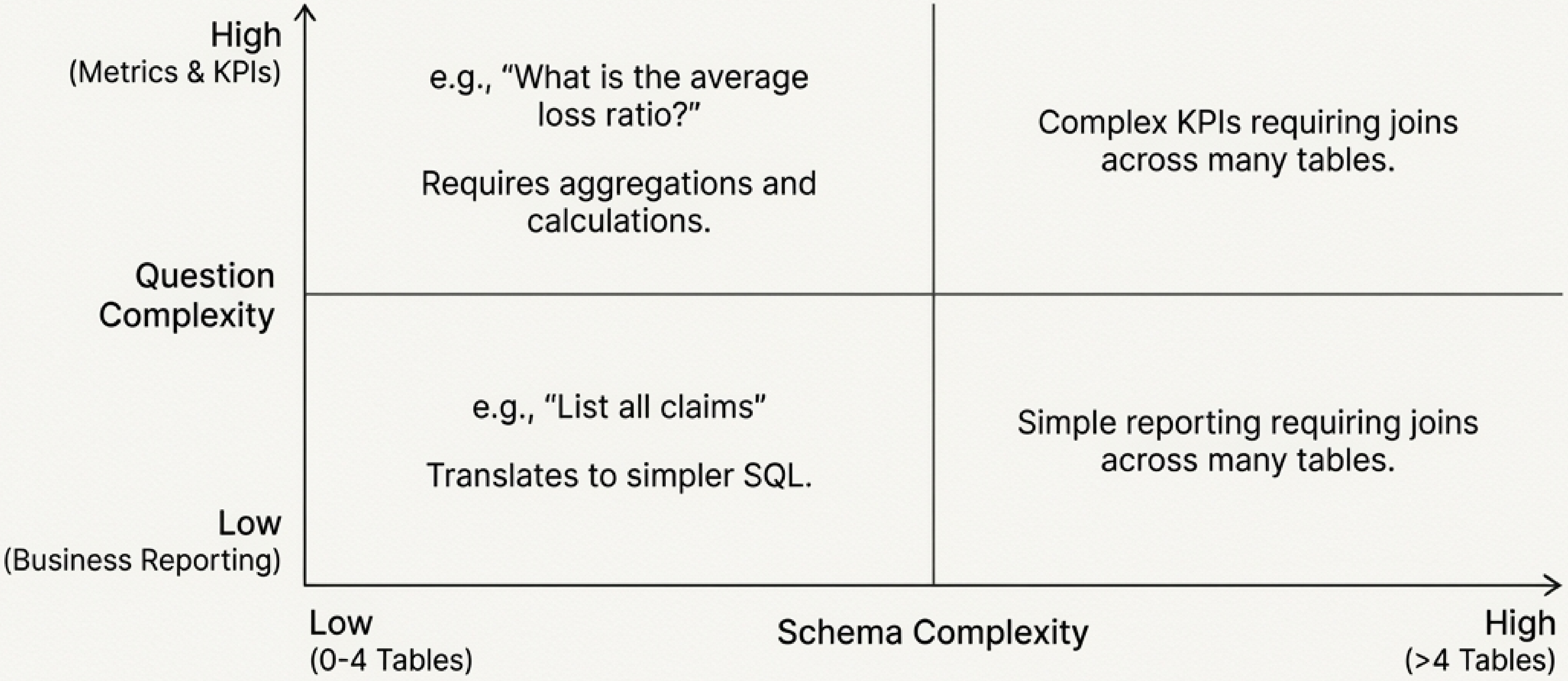
3. Context Layer

A business ontology (in OWL) describing the insurance domain, with mappings (in R2RML) to the SQL schema, creating the Knowledge Graph representation.

NotebookLM

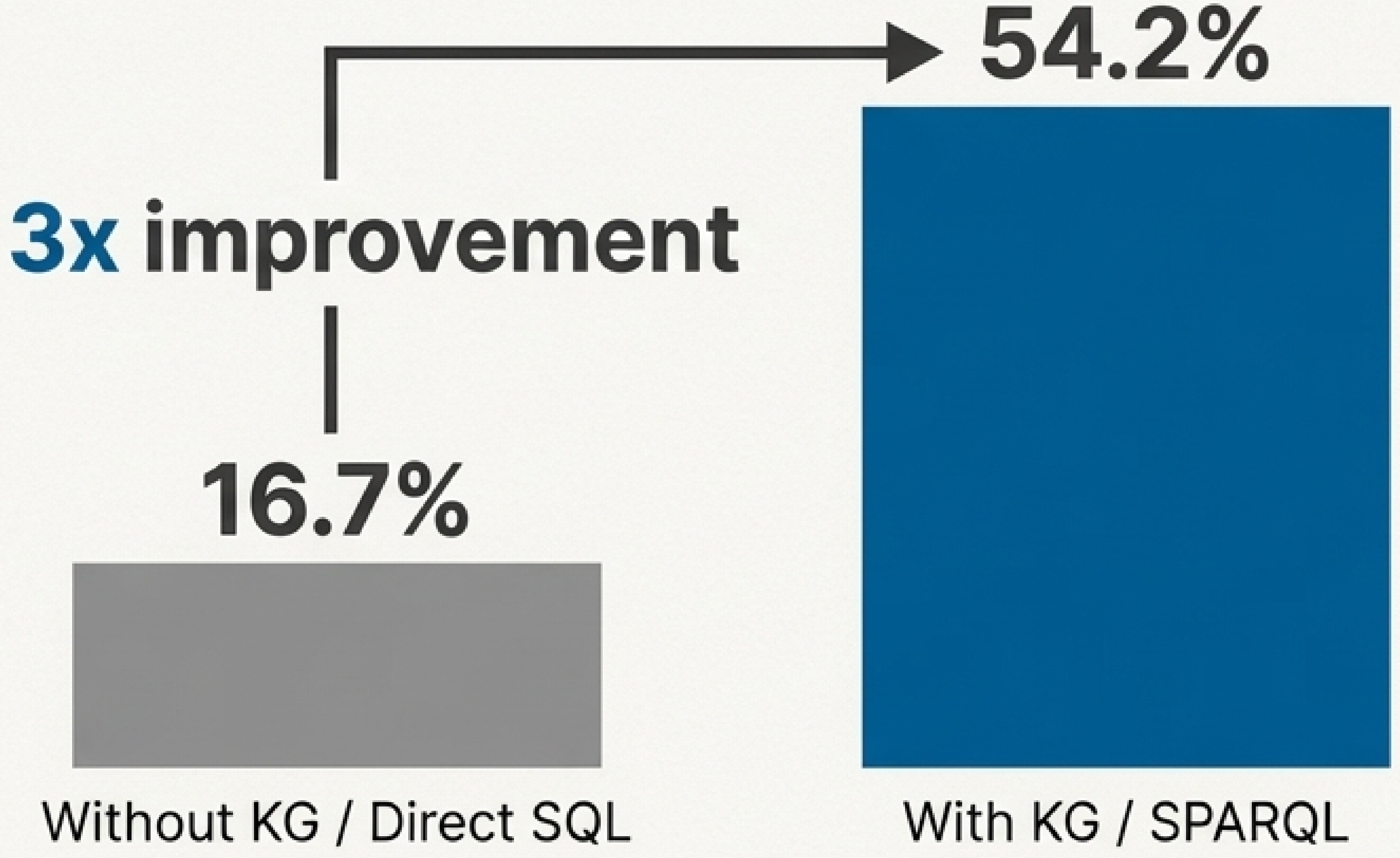
Questions Were Classified by Business and Technical Complexity

Not all questions are equal. We classified our 43 questions along two axes of complexity to understand performance under different conditions.



NotebookLM

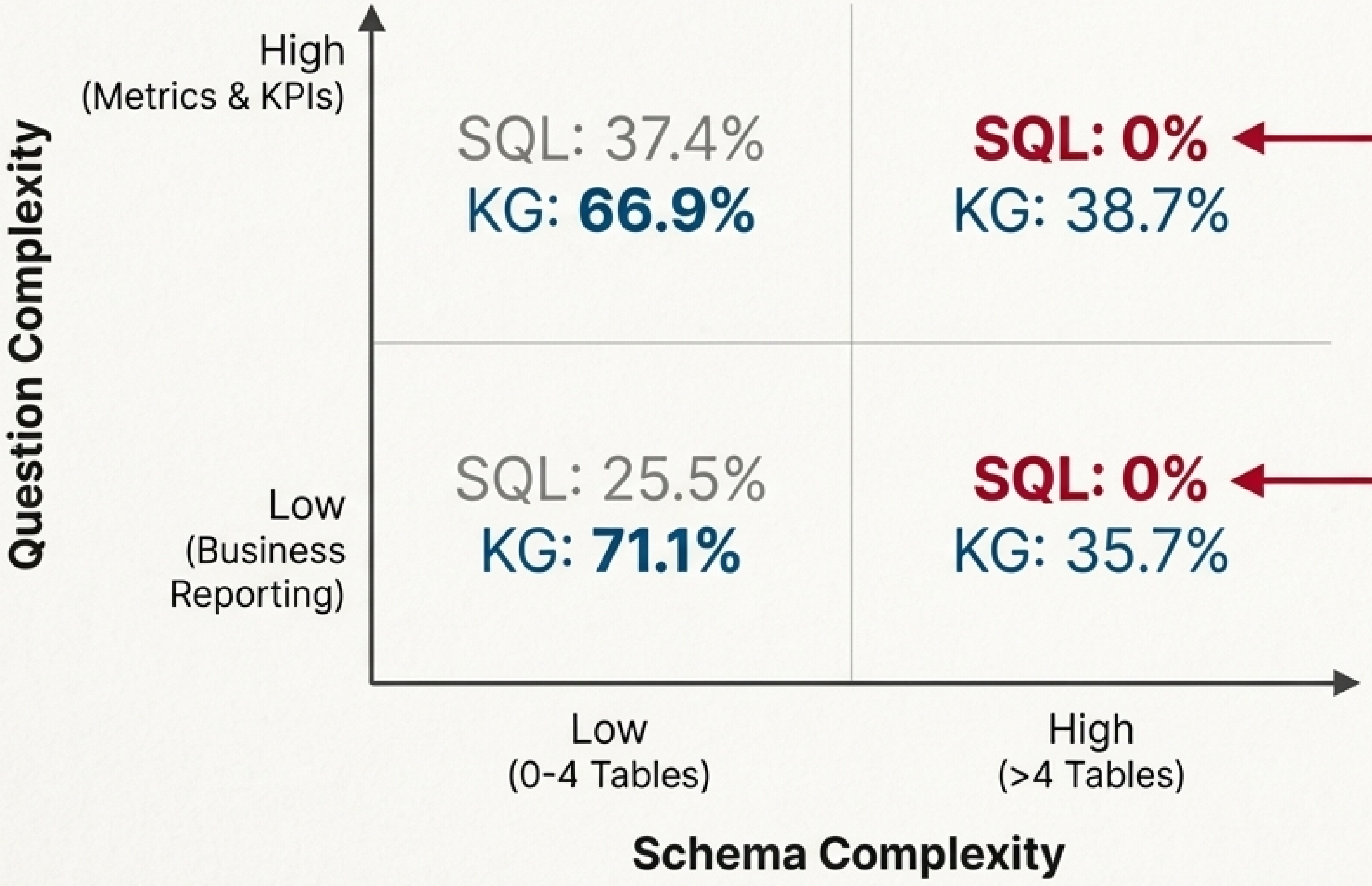
Knowledge Graphs Triple the Accuracy of LLM-Powered Answers



Average Overall Execution Accuracy (AOEA) across all 43 enterprise questions using GPT-4.
The Knowledge Graph provides a **3x improvement** in accuracy.

NotebookLM

Direct SQL Accuracy Collapses to 0% When Answering Strategically Important Questions.



For questions requiring joins across **more than 4 tables, the direct-to-SQL** approach failed every single time.

NotebookLM

The Anatomy of Failure: Direct SQL Prompts Lead to Dangerous Hallucinations.

Without the business context provided by a Knowledge Graph, the LLM is forced to invent details to bridge gaps in its understanding. We observed three primary types of SQL inaccuracies:

1. Column Name Hallucinations

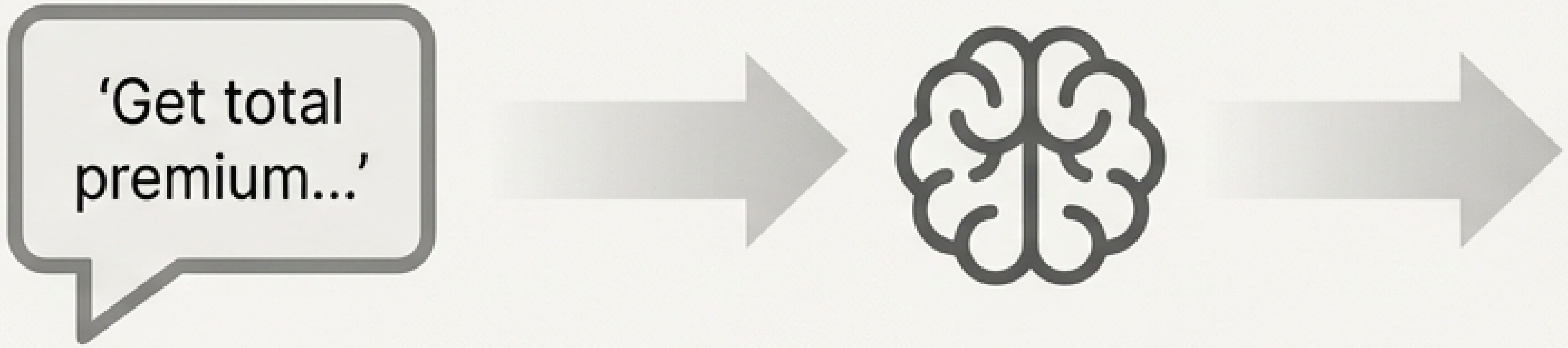
Generating queries with column names that do not exist in the database (e.g., asking for total_premium_paid when the column is named policy_amount).

2. Join Hallucinations

Creating joins between tables that are syntactically valid but semantically incorrect, leading to wrong results.

3. Value Hallucinations

Applying filters based on values that do not exist in the data.




```
SELECT policy_id, customer_id,
       total_premium_paid
FROM Policy
WHERE policy_status = 'Active';
```

NotebookLM

The Anatomy of Success: Knowledge Graphs Ground the LLM in Business Reality

The Knowledge Graph provides a clear, unambiguous “map” of the business. The LLM is no longer guessing; it is reasoning over a defined structure of concepts and relationships.

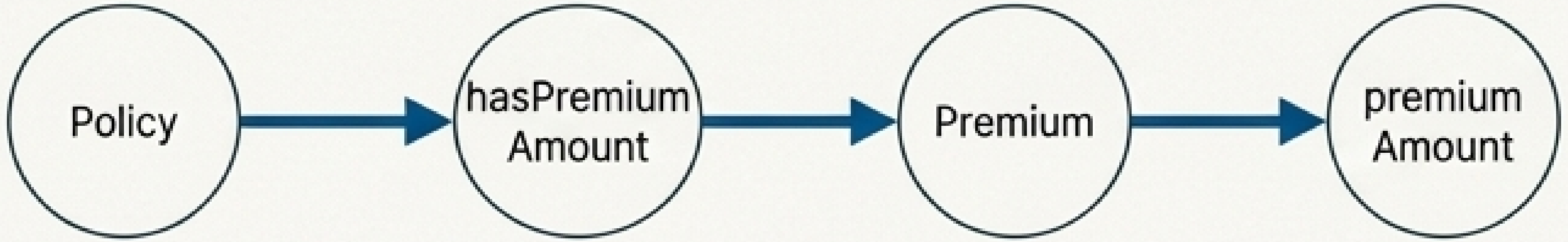
In our benchmark, the KG-based approach (generating SPARQL) produced **zero class or property hallucinations.** Errors were not due to inventing concepts, but rather to selecting an incorrect but existing path through the graph.

Without KG

```
SELECT policy_id, customer_id,
       total_premium_paid
FROM Policy...
```

(Hallucinated column) **total_premium_paid**

With KG



(A clear, **valid path**)

NotebookLM

Benchmark Results at a Glance: Accuracy by Question & Schema Complexity.

Question Category	Accuracy w/o KG (SQL)	Accuracy w/ KG (SPARQL)	Accuracy Improvement
All Questions (Overall)	16.7%	54.2%	+37.5%
Low Question / Low Schema	25.5%	71.1%	+45.6%
High Question / Low Schema	37.4%	66.9%	+29.5%
Low Question / High Schema	0%	35.7%	+35.7%
High Question / High Schema	0%	38.5%	+38.5%

Results based on Average Overall Execution Accuracy (AOEA) using GPT-4 and zero-shot prompting.

NotebookLM

The Evidence Leads to an Inescapable Conclusion.

"The main conclusion of this work is that investing in Knowledge Graphs provides higher accuracy for LLM powered question answering systems."

The benchmark results provide clear, quantitative evidence. To move from promising demos to reliable, production-ready AI applications, addressing the context gap is not optional.

NotebookLM

The Strategic Imperative: Treat Business Context as a First-Class Citizen.

To achieve trustworthy, accurate, and explainable insights from AI, enterprises must stop treating business context as an afterthought. It must be explicitly modeled, managed, and governed.

The Path Forward:

1. **Invest** in building a semantic layer over your data using a Knowledge Graph architecture.
2. **Elevate** business context (ontologies, metadata, mappings) to a core asset within your data management strategy.
3. **Adopt** a data catalog platform capable of supporting a knowledge graph architecture to ensure this context is managed systematically, not in an ad-hoc manner.

NotebookLM

Benchmark Details and Methodology.

This presentation is based on the technical report:

Title: A Benchmark to Understand the Role of Knowledge Graphs on Large Language Model's Accuracy for Question Answering on Enterprise SQL Databases
Authors: Juan F. Sequeda, Dean Allemang, Bryon Jacob
Date: November 14, 2023

The full benchmark framework, including the schema, questions, ontology, and processing code, is available for review and reproduction on GitHub.

github.com/datadotworld/cwd-benchmark-data

NotebookLM