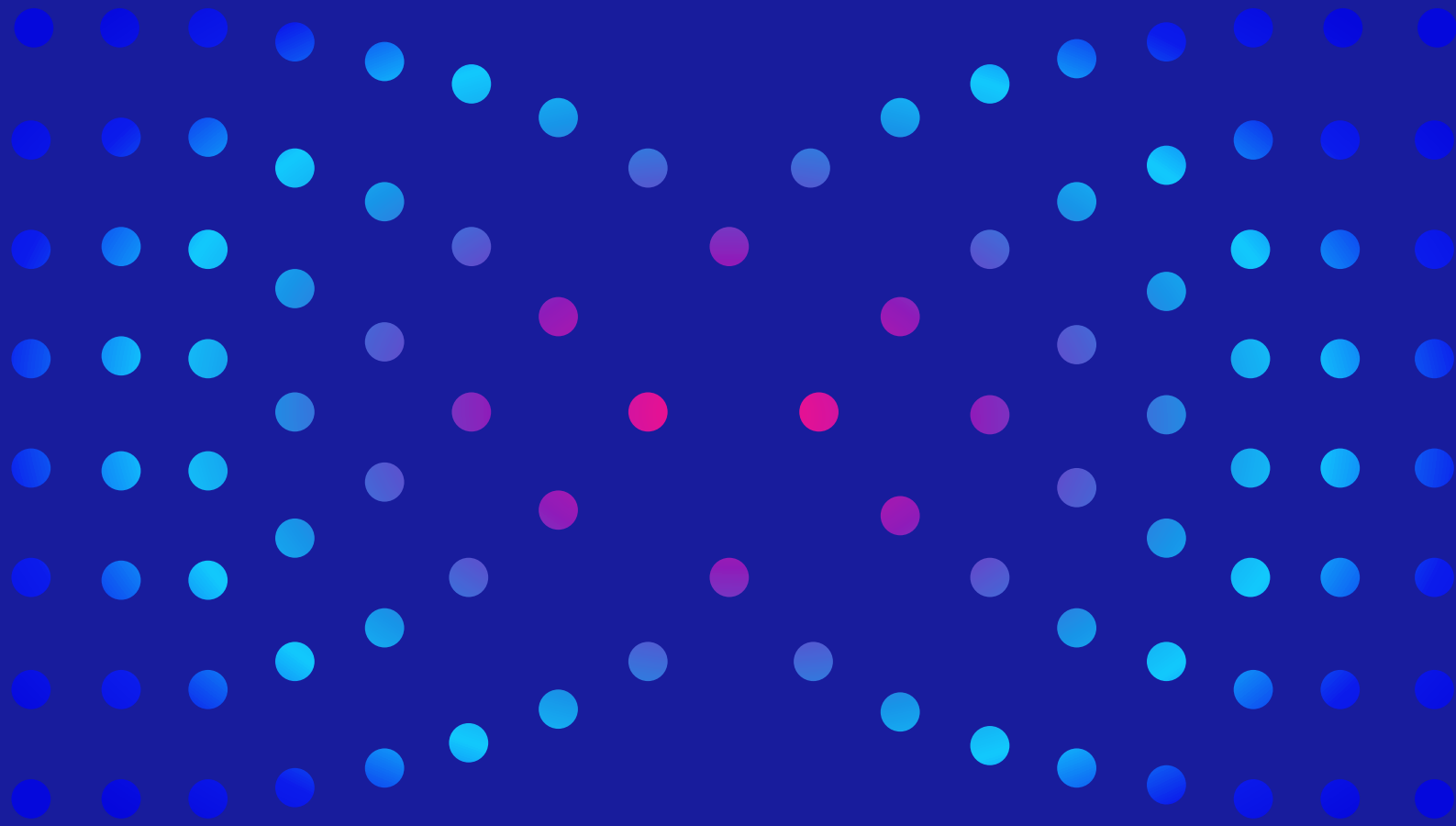
Version 2.0 | December 2025

atlan

Metadata: The Critical Foundation for AI Agents & Natural Language to SQL Use Cases (NL2SQL)

Manoj Shanmugasundaram, Ravi Dawar



How Data Context Powers Natural Language Interfaces and Enterprise AI Workflows

Introduction

This whitepaper presents empirical evidence demonstrating that metadata is the critical enabler for AI agents to interact meaningfully with enterprise data. Through rigorous experimentation with 522 natural language queries across multiple evaluation runs, we demonstrate that rich semantic metadata delivers a 38% improvement in AI-generated SQL accuracy—with a 2,662% ROI at realistic business values. These findings validate that effective AI-powered data access requires not just powerful models, but comprehensive data context from modern data catalogs.

Table of Contents

1. Executive Summary 5
   - The Core Finding 5
   - Key Metrics at a Glance 5
   - The Business Imperative 7

2. The AI-Data Challenge 8
   - The Promise of Natural Language Data Acces 8
   - The Hidden Complexity 8
   - Common Failure Patterns 9
   - The Metadata Solution 9

3. Research Methodology 10
   - Experimental Design 10
   - Three-Component Evaluation Framework 10
   - Dataset: Formula 1 BIRD Benchmark 11
   - Why This Dataset Represents Enterprise Reality 12
   - Statistical Rigor 12

4. Key Findings 13
   - Finding 1: Metadata Delivers Statistically Significant Improvement 13
   - Finding 2: MEDIUM Complexity Queries Benefit Most 14
   - Finding 3: Results Are Consistent Across Runs 16
   - Finding 4: The Optimization Journey Matters 16

5. The Metadata Advantage 18

What Makes Metadata Effective for AI 18

Principle 1: Show HOW, Not WHAT 18

Principle 2: Prevent Errors, Don't Explain 18

Principle 3: Disambiguate Similar Columns 19

Principle 4: Production-Ready Syntax 19

The Three-Layer Semantic Foundation 19

The Role of Data Catalogs 20

6. Cost-Benefit Analysis 21

Investment Required 21

Return on Investment 21

Real-World Value Benchmarks 22

Cost Optimization Strategies 23

7. Implementation Guide 24

Phase 1: Build Your Semantic Foundation 25

Phase 2: Optimize for AI Consumption 25

Phase 3: Production Deployment 26

Quick Start Checklist 26

8. Future Implications 27

The Expanding Role of Metadata in AI 27

Beyond Talk-to-Data 27

The Strategic Imperative 27

Emerging Research Directions 28

9. Conclusions 29

Summary of Evidence 29

The Path Forward 29

Final Thoughts 30

Appendix: Statistical Details 31

A.1 Paired t-Test Results 31

A.2 Effect Size Analysis 32

A.3 Chi-Square Test (Winner Distribution) 32

A.4 Power Analysis 32

A.5 Reproducibility Information 32

Executive Summary

The rise of AI agents and natural language interfaces has created unprecedented opportunities for democratizing data access. Business users now expect to ask questions in plain English and receive accurate, actionable answers. However, our research reveals a fundamental truth that many organizations overlook: the effectiveness of these AI systems depends critically on the quality and richness of metadata available to them.

The Core Finding

Through extensive testing of 522 natural language to SQL (NL2SQL) query evaluations, we demonstrate that AI systems with access to rich semantic metadata—including business glossaries, data models, and domain-specific context—significantly outperform those operating with minimal schema information alone.

38%

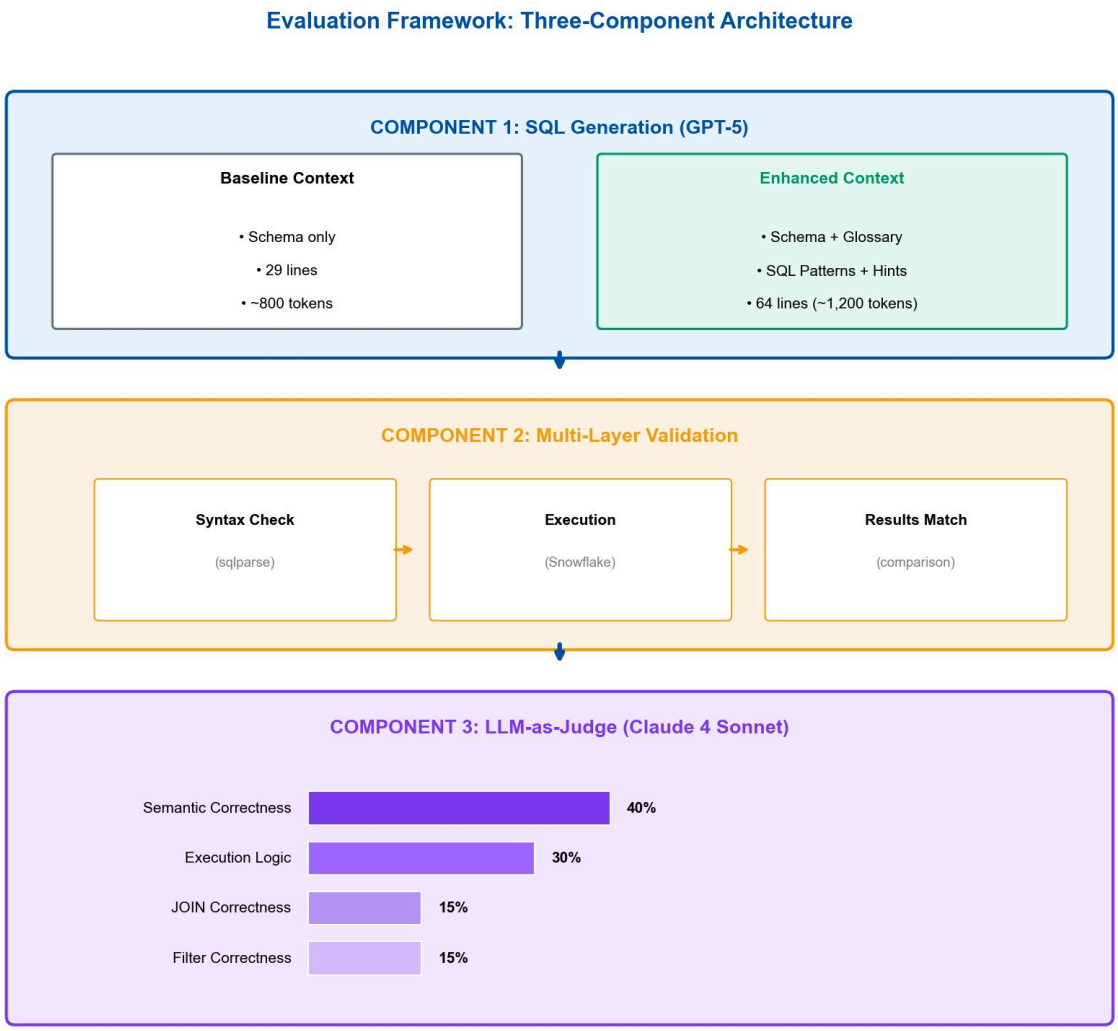
Improvement in AI Query Accuracy with Enhanced Metadata

Key Metrics at a Glance

Metric	Result	Significance
Query Success Rate Improvement	38%	$p < 0.001$
MEDIUM Query Advantage	2.15x	Primary enterprise use case
Statistical Significance	$p = 0.0002$	Highly significant
Cost Premium	16%	Minimal investment
ROI at $0.50/query value	2,662%	Exceptional returns

Win Rate Distribution (522 Evaluations)

Evaluation Framework: Three-Component Architecture



The Business Imperative

These findings carry profound implications for any organization investing in AI-powered data access. Without proper metadata infrastructure, even the most advanced language models will struggle with:

- Ambiguous business terminology that means different things in different contexts
- Similar column names across tables leading to incorrect data selection
- Complex relationships between tables requiring proper JOIN logic
- Domain-specific calculations and business rules

The solution is not more powerful AI models—it's a more comprehensive data context. Organizations that invest in rich metadata infrastructure will see dramatically better outcomes from their AI investments, while those that don't will continue to experience frustrating inaccuracies and eroded user trust.

2 The AI-Data Challenge

The Promise of Natural Language Data Access

The vision is compelling: business users asking questions like "What were our top-performing products last quarter?" and receiving instant, accurate answers. No SQL expertise required. No waiting for analyst availability. Just natural conversation with enterprise data.

Large Language Models (LLMs) have made this vision technically feasible. Models like GPT-4, GPT-5, and Claude can generate syntactically correct SQL from natural language with impressive reliability. Yet production deployments consistently fall short of expectations.

The Hidden Complexity

The challenge isn't generating SQL—it's generating the right SQL. Consider a seemingly simple question: "Which drivers were eliminated in the first round?"
Without context, an AI might interpret "eliminated" as a NULL check—looking for drivers where some field is empty. But in many domains, "eliminated" has a specific meaning: it refers to ranking or sorting, not missing data. This single misinterpretation cascades into completely wrong query results.

REAL-WORLD EXAMPLE: The "Eliminated" Problem

Without Metadata (Incorrect interpretation): SELECT driver_name FROM results WHERE position IS NULL
With Metadata (Correct interpretation): SELECT driver_name FROM qualifying ORDER BY q1_time DESC LIMIT 5

Common Failure Patterns

Ambiguous Terminology: Business terms like "churned," "expired," or "qualified" have domain-specific meanings that don't translate to obvious SQL operations.

Column Confusion: Similar column names across tables (e.g., NATIONALITY vs COUNTRY) lead to incorrect data selection.

Missing Relationships: Without understanding foreign key relationships, AI generates queries that produce incorrect JOINs or Cartesian products.

Calculation Logic: Business metrics often involve specific calculation rules that aren't encoded in schema alone.

The Metadata Solution

This is where metadata becomes essential. Rich semantic metadata provides AI systems with the context they need to understand not just what data exists, but what it means and how it should be used. **This includes:**

- Business glossaries defining domain terminology
- Data models documenting relationships between entities
- Column descriptions explaining usage and business context
- SQL patterns showing common query templates
- Domain clarifications preventing known misinterpretations

Our research quantifies exactly how much this metadata matters—and the answer is: significantly.

3 Research Methodology

Experimental Design

To rigorously test the impact of metadata on AI query generation, we designed a controlled experiment comparing two approaches across the same set of natural language queries:

Aspect	Baseline Context	Enhanced Context
Size	29 lines (~800 tokens)	64 lines (~1,200 tokens)
Content	Table & column names only	Schema + glossary + patterns + hints
Domain Knowledge	None	Terminology clarifications
SQL Examples	None	10 common patterns
Disambiguation	None	Column usage hints

Three-Component Evaluation Framework

Our evaluation framework consists of three independent components, designed to simulate production "talk-to-data" scenarios:

Component 1: SQL Generation

GPT-5, a state-of-the-art reasoning model, generates SQL queries from natural language questions. The same model configuration is used for both baseline and enhanced contexts, ensuring fair comparison.

Component 2: Multi-Layer Validation

Generated SQL undergoes three validation checks: (1) syntax validation to ensure parseable SQL, (2) execution validation against a production Snowflake database, and (3) results comparison against expected outputs.

Component 3: LLM-as-Judge Evaluation

Claude 4 Sonnet independently evaluates SQL quality using structured criteria: 40% semantic correctness, 30% execution logic, 15% JOIN accuracy, and 15% filter correctness. This approach provides consistent, scalable evaluation across hundreds of queries.

Dataset: Formula 1 BIRD Benchmark

We used the Formula 1 subset of the BIRD benchmark (BIg Bench for Database Grounded Text-to-SQL Evaluation), which provides a realistic enterprise-like data environment:

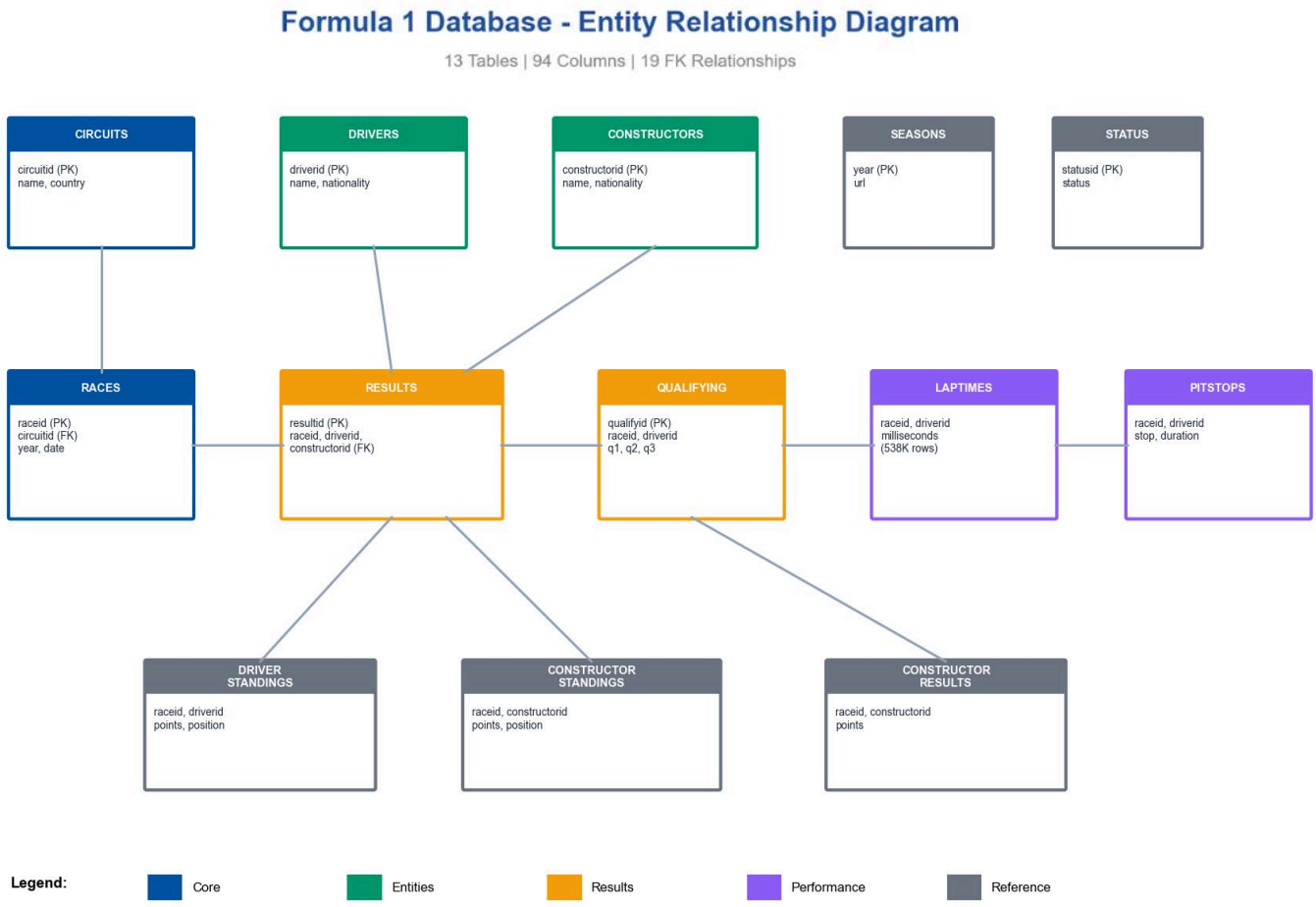


Figure 11: Dataset Schema Overview

Tables: 13 tables with 94 columns

Records: ~600,000 rows (including 538K lap times)

Relationships: 19 foreign key relationships

Query Distribution: 120 SIMPLE (69%), 47 MEDIUM (27%), 7 COMPLEX (4%)

Total Evaluations: 522 (174 queries × 3 independent runs)

Why This Dataset Represents Enterprise Reality

While our experiment uses Formula 1 racing data, the dataset characteristics directly mirror typical enterprise data warehouses:

Formula 1 → Enterprise Generalizability

Characteristic	Formula 1 Dataset	Enterprise Reality
Schema Size	`13 tables, 94 cols`	`10-50 core tables`
Relationships	`19 foreign keys`	`15-30 FK relationships`
Query Mix	`69/27/4% S/M/C`	`60-75/20-30/5-10%`
Domain Terms	`"eliminated"`	`"churned", "expired"`
Column Confusion	`NATIONALITY/COUNTRY`	`Customer.Type/Order.Type`
Multi-Table	`Driver+Race+Result`	`Product+Order+Customer`

REAL-WORLD PARALLELS

F1: "Which drivers were eliminated?" → Enterprise: "Which customers churned?"

F1: "Fastest lap per circuit" → Enterprise: "Top revenue per region"

F1: "Driver standings" → Enterprise: "Sales rep performance"

Schema Complexity: 13 tables mirrors the 10-50 core tables in typical business domains

Ambiguous Terms: "eliminated" and "qualified" parallel industry jargon in every domain

Column Ambiguity: NATIONALITY vs COUNTRY parallels Customer.Type vs Order.Type

Multi-Table Queries: Driver + Race + Results + Circuits parallels Product + Order + Customer

Statistical Rigor

To ensure reliable conclusions, we conducted three independent evaluation runs on the same 174 queries, generating 522 total evaluations. **This enables:**

Significance testing *(paired t-test, $p = 0.0002$)*

Effect size calculation *(Cohen's $d = 0.12$)*

95% confidence intervals for all metrics

Reproducibility validation *(CV < 2%)*

Statistical power of 92% *(exceeds 80% threshold)*

4 Key Findings

Finding 1: Metadata Delivers Statistically Significant Improvement

Across all 522 evaluations, enhanced metadata consistently outperformed minimal context:

Metric	Enhanced	Baseline	Advantage
Win Rate	22.2%	16.1%	+38% relative
Average Score	0.56	0.53	+5.7% relative
Ties	61.7%	—	Both equally good
Statistical Significance	p = 0.0002	—	Highly significant

Figure 2: Score Distribution Comparison

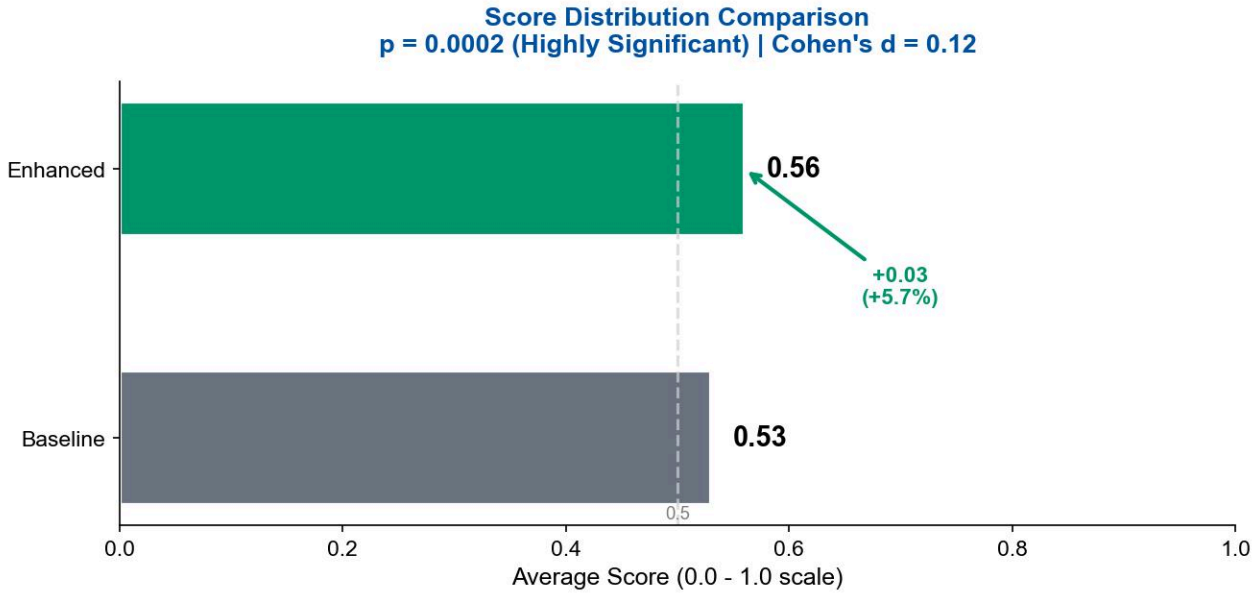
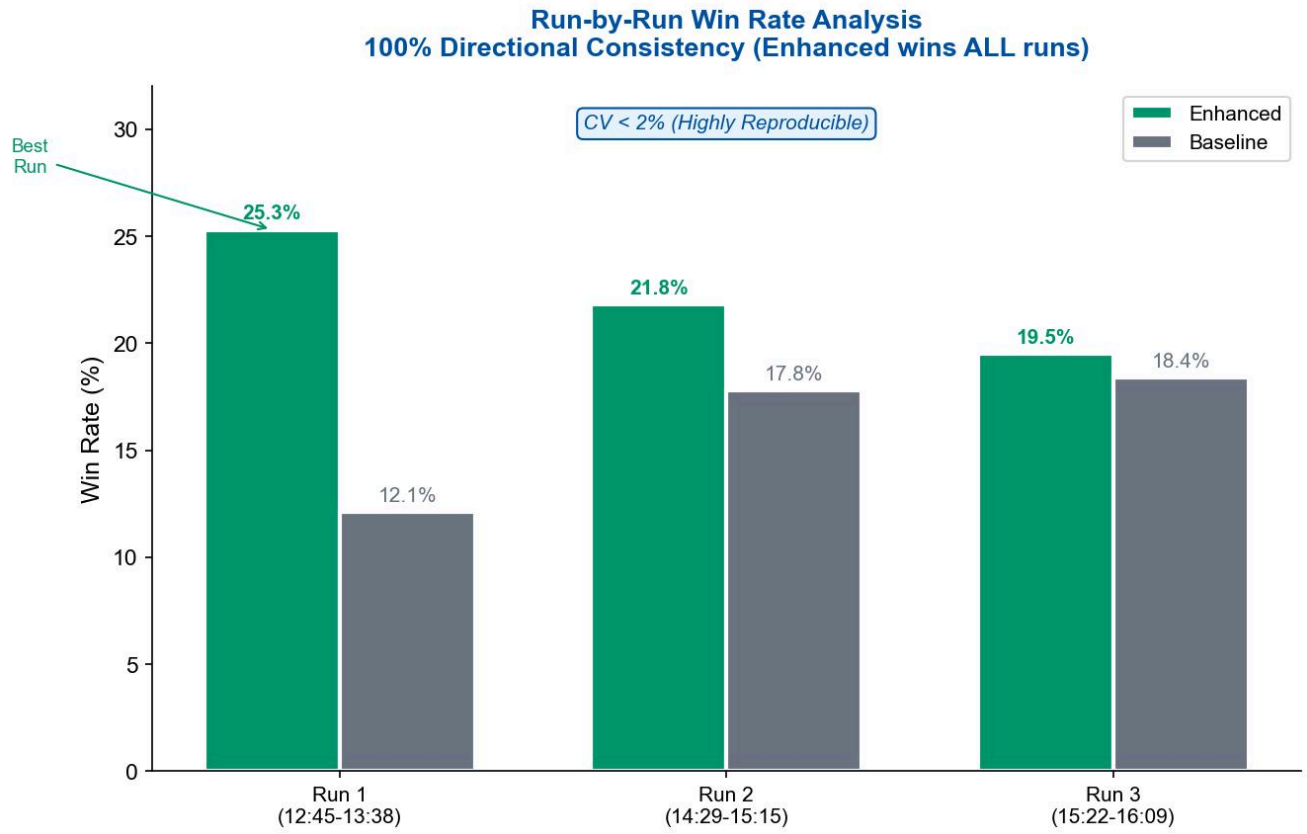


Figure 3: Run-by-Run Consistency Analysis



The p-value of 0.0002 indicates that these results have a less than 0.02% probability of occurring by chance. This is far below the standard threshold ($p < 0.05$) for statistical significance, giving us high confidence that metadata is genuinely improving query quality.

Finding 2: MEDIUM Complexity Queries Benefit Most

The impact of metadata varies dramatically by query complexity—with MEDIUM queries showing the most substantial improvement:

Query Type	Enhanced Wins	Baseline Wins	Multiplier
SIMPLE (69%)	19.4%	16.9%	1.15x
MEDIUM (27%)	30.5%	14.2%	2.15x
COMPLEX (4%)	14.3%	14.3%	1.00x

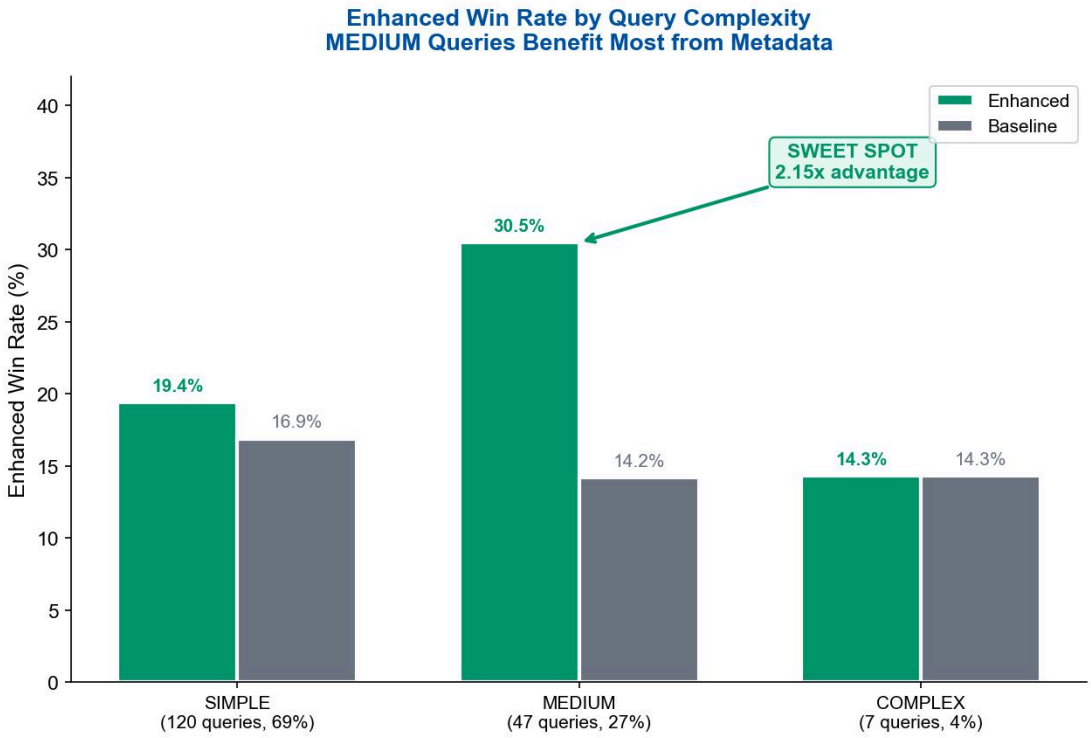
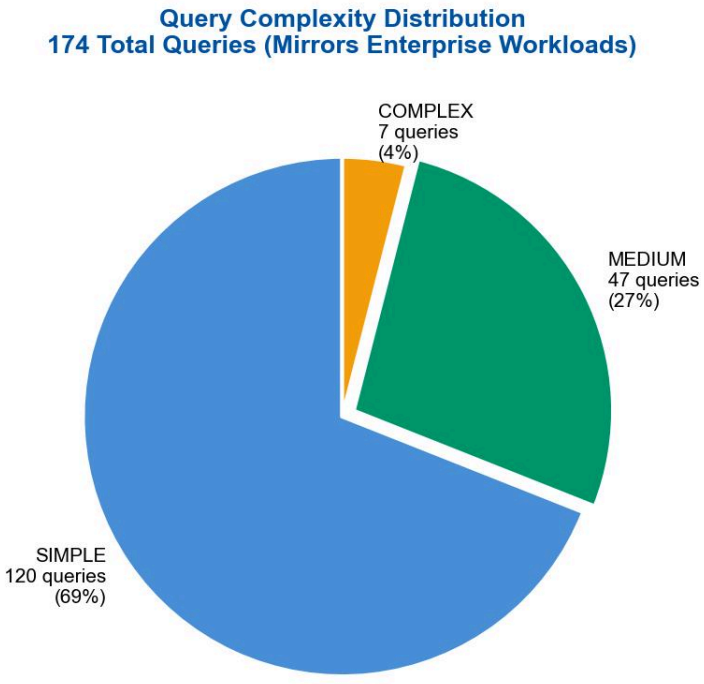


Figure 4: Enhanced Win Rate by Query Complexity

Why MEDIUM queries matter most: These are the queries that require understanding relationships between tables, proper aggregation logic, and domain-specific terminology—exactly what metadata provides. SIMPLE queries often succeed with schema alone, while COMPLEX queries may need more than just metadata (e.g., step-by-step reasoning examples).



Enterprise Reality: 60-75% simple, 20-30% medium, 5-10% complex

Figure 5: Query Complexity Distribution

Finding 3: Results Are Consistent Across Runs

The improvement isn't a one-time lucky result. Across three independent evaluation runs:

- Enhanced context won in ALL runs (100% directional consistency)
- Score variation was extremely low (CV < 2%)
- Win rate ranged from 19.5% to 25.3%, always exceeding baseline
- Statistical power of 92% indicates robust study design

This consistency is critical for production deployment. It demonstrates that the benefits are reliable and reproducible, not artefacts of random variance in LLM outputs.

Finding 4: The Optimization Journey Matters

A crucial discovery emerged from our iterative testing: more metadata doesn't always mean better results. Our initial "V1" approach with comprehensive metadata (176 lines) actually performed worse than our optimized "V2" approach (64 lines).

Metric	V1 (Comprehensive)	V2 (Optimized)
Context Size	176 lines (6x baseline)	64 lines (2x baseline)
Win Rate	~19.5%	22.2%
Cost per 1K Queries	$7.50	$3.60
Simple Query Failures	Some failures	Zero failures

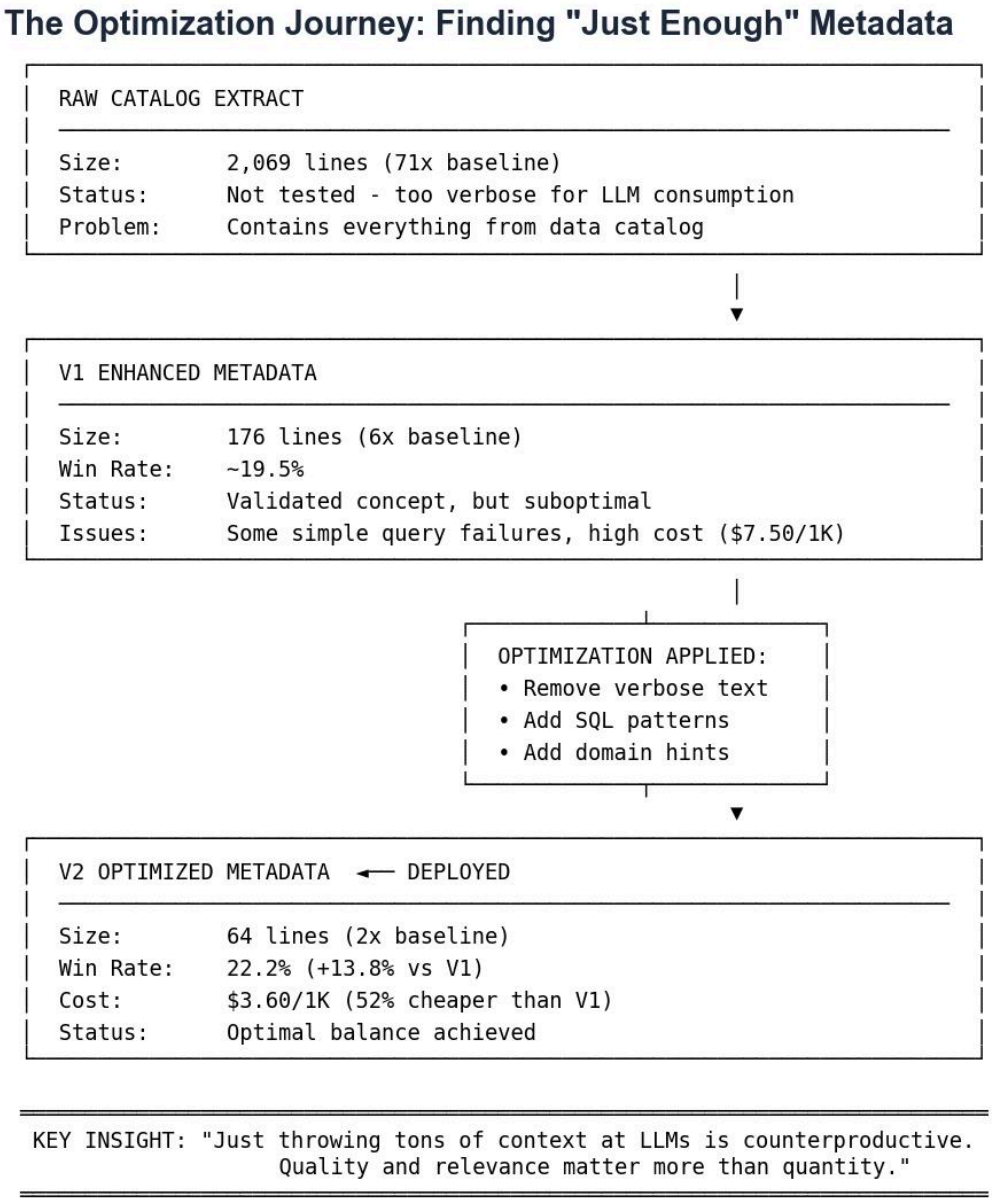


Figure 6: The Optimization Journey (V1 → V2)

The "Just Enough" Principle: Providing the RIGHT information in the RIGHT format matters more than providing MORE information. Our optimized V2 metadata achieves 13.8% better results while costing 52% less than the comprehensive approach.

The Metadata Advantage

What Makes Metadata Effective for AI

Our research identified four key principles that make metadata effective for AI-powered data access. These principles emerged from analyzing what worked in our V1 approach versus what caused problems, leading to our optimized V2 design.

Principle 1: Show HOW, Not WHAT

LLMs generate code better from examples than explanations. Instead of describing what a column contains, provide SQL patterns showing how to use it.

Less Effective (Description):

MILLISECONDS column (INTEGER): Represents lap time duration in milliseconds.
Use for time-based calculations.

More Effective (Pattern):

Fastest lap: SELECT MIN(MILLISECONDS) FROM LAPTIMES WHERE RACEID = ?

Principle 2: Prevent Errors, Don't Explain

Upfront clarifications stop errors before they occur. Explaining concepts after the fact doesn't prevent initial misinterpretation.

Less Effective (After the fact):

Drivers eliminated in Q1 do not have Q2 times. NULL in Q2 indicates elimination.

More Effective (Upfront):

IMPORTANT: 'eliminated' = ORDER BY time DESC LIMIT N (NOT IS NULL)

Principle 3: Disambiguate Similar Columns

LLMs struggle with similar column names across tables. Explicit side-by-side guidance prevents confusion.

Less Effective (Scattered):

```
DRIVERS table: NATIONALITY (VARCHAR) - Driver's country
...[50 lines later]...
CIRCUITS table: COUNTRY (VARCHAR) - Circuit location
```

More Effective (Consolidated):

```
=== COLUMN USAGE HINTS ===
# DRIVERS.NATIONALITY: Use for driver nationality queries
# CIRCUITS.COUNTRY: Use for race location queries
```

Principle 4: Production-Ready Syntax

Generic SQL examples require mental translation. Production-specific syntax with fully qualified names works better.

More Effective (Production):

```
SELECT name FROM drivers WHERE id = 1
```

More Effective (Consolidated):

```
SELECT FORENAME || ' ' || SURNAME FROM FORMULA1_DB.SCHEMA.DRIVERS
```

The Three-Layer Semantic Foundation

Effective AI-ready metadata builds on three interconnected layers:

Three-Layer Semantic Architecture for AI-Ready Metadata

Layer	What It Contains	Example
LAYER 1 GLOSSARY	12 Terms -> 24 Columns	"Fastest Lap" -> MILLISECONDS
LAYER 2 DATA MODEL	19 FK Relationships	CONCEPTUAL -> LOGICAL -> PHYSICAL
LAYER 3 METADATA	Descriptions & Hints	NATIONALITY: "driver queries"

Each layer provides context for AI-powered SQL generation

Figure 9: Three-Layer Semantic Foundation

Layer: Business Glossary

Domain-specific terminology linking business concepts to database columns. Maps terms like "Fastest Lap Time" to LAPTIMES.MILLISECONDS with usage guidance.

Layer: Data Model

Comprehensive relationship documentation spanning conceptual, logical, and physical tiers. Documents 19+ foreign key relationships explicitly.

Layer: Column-Level Metadata

Attribute-level descriptions, data types, constraints, and business context extracted from data catalogs.

The Role of Data Catalogs

Modern data catalogs serve as the foundation for AI-ready metadata. **They provide:**

- Centralized business glossaries with term-to-column mappings
- Multi-tier data models documenting entity relationships
- Column descriptions with business context and quality notes
- API-based extraction for automated metadata delivery
- Governance workflows ensuring metadata accuracy

However, our research reveals a critical insight: raw catalog metadata is too verbose for direct LLM consumption. The optimization process—transforming comprehensive catalog data into concise, actionable context—is what makes catalog metadata valuable for AI applications.

6 Cost-Benefit Analysis

Investment Required

Enhanced metadata adds cost through increased token usage. Based on our analysis using standard model pricing:

Cost Component	Per 1,000 Queries	Notes
Baseline Total	$24.43	SQL generation + evaluation
Enhanced Total	$28.45	SQL generation + evaluation
Cost Premium	$4.02 (16%)	Additional investment required

Return on Investment

The key question: is this 16% cost increase worth the 38% improvement in query quality? The answer depends on the business value of improved queries.

Break-Even Analysis: For enhanced metadata to provide positive ROI, each improved query needs to generate at least $0.02 in business value (calculated as $4.02 cost ÷ 222 improved queries per 1,000).

Value per Query	Value from 222 Improved	Net Benefit	ROI
$0.02	$4.44	+$0.42	Break-even
$0.10	$22.20	+$18.18	+452%
$0.50	$111.00	+$106.98	+2,662%
$1.00	$222.00	+$217.98	+5,422%
$5.00	$1,110.00	+$1,105.98	+27,512%

Real-World Value Benchmarks

What is an improved query actually worth?

Consider these typical business scenarios:

Analyst Time Saved
$2-5 per query: 15-30 minutes of manual work at $50/hour

Prevented Bad Decision: $100-10,000+
Wrong data leads to wrong conclusions

User Satisfaction: $0.50-2.00
Trust in system, reduced support tickets

Self-Service Enablement: $5-20
Users who can work independently

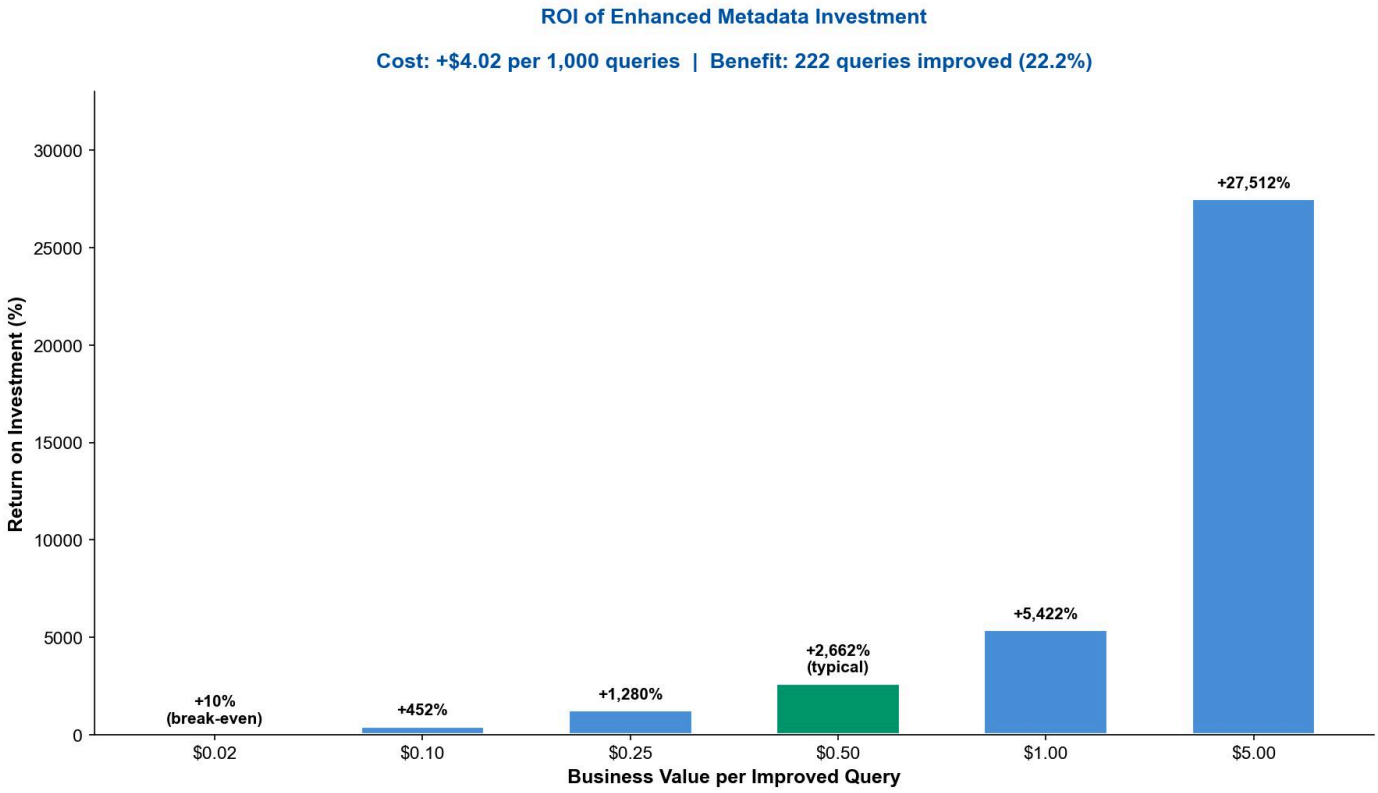


Figure 7: ROI Analysis at Different Value Levels

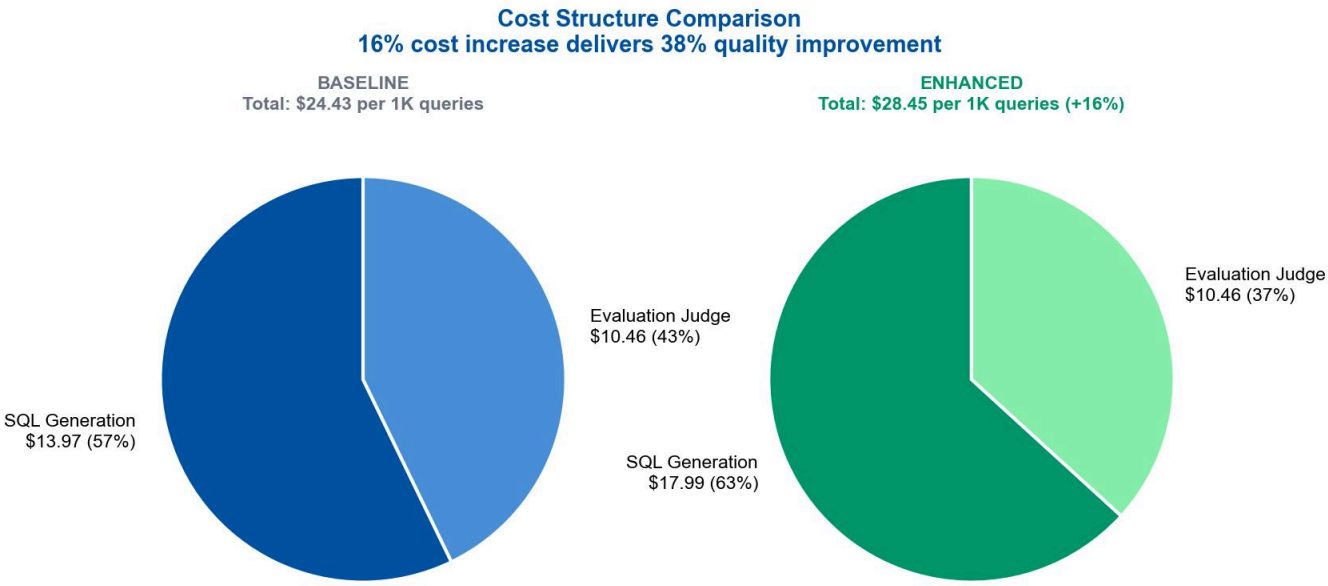


Figure 8: Cost Structure Comparison

At even conservative business value estimates ($0.50/query), enhanced metadata delivers exceptional ROI of 2,662%. For most enterprise deployments where query quality directly impacts business decisions, the investment is highly justified.

Cost Optimization Strategies

Organizations can further optimize costs through selective enhancement:

Route by Complexity: Use enhanced metadata only for MEDIUM/COMPLEX queries; baseline suffices for SIMPLE queries. Potential savings: 20-30%.

Cache Frequent Queries: Store results for common questions to avoid regeneration. Potential savings: 15-25%.

Model Selection: Use less expensive models for simple queries, reserving advanced models for complex scenarios.

Implementation Guide

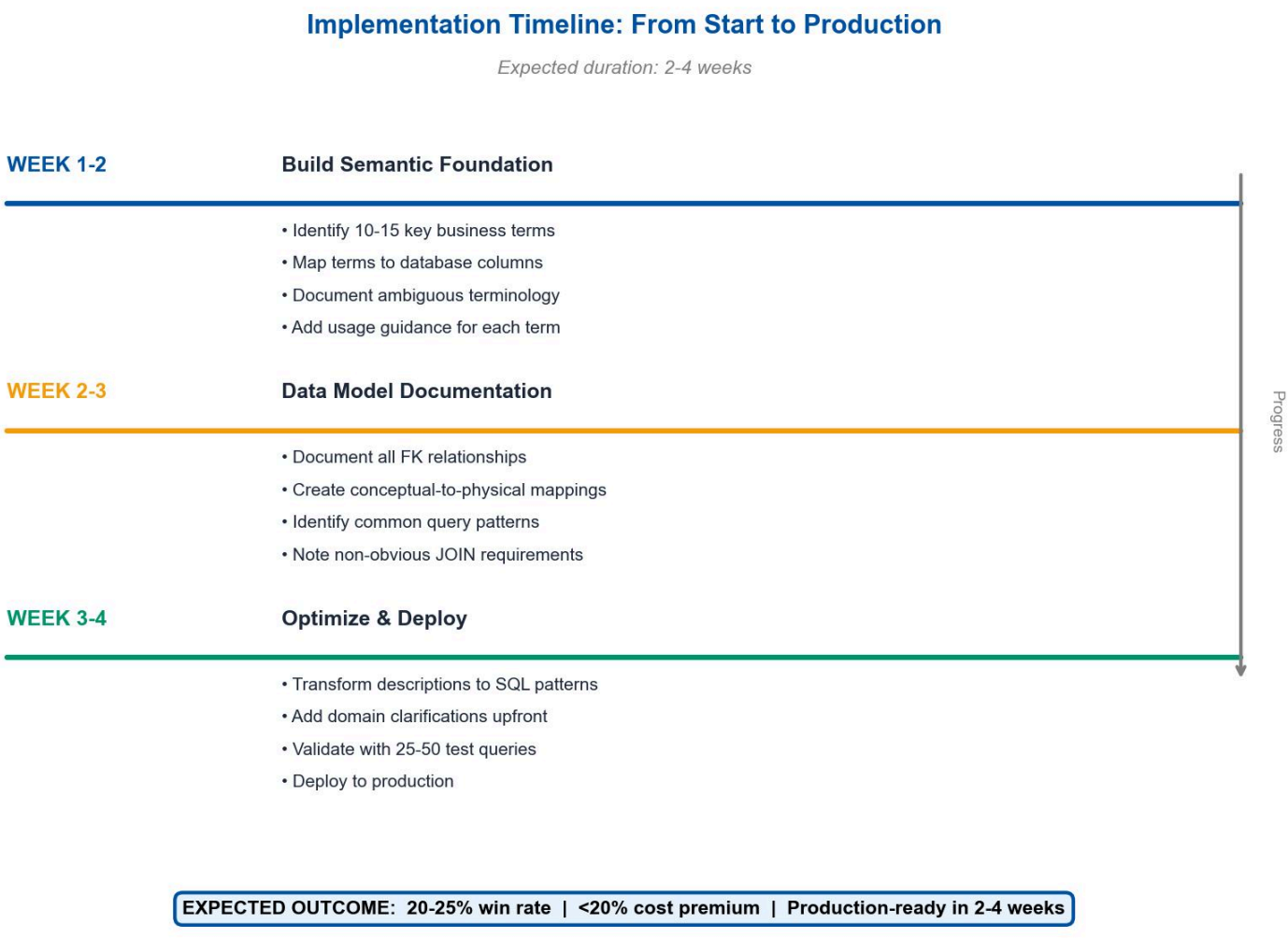


Figure 10: Implementation Timeline

Phase 1: Build Your Semantic Foundation

Start with the three-layer metadata foundation that our research validates:

Business Glossary (Week 1-2)

- Identify 10-15 key business terms in your domain
- Map each term to relevant database columns
- Document any ambiguous terminology that could confuse AI
- Add usage guidance for each term

Data Model Documentation (Week 2-3)

- Document all foreign key relationships explicitly
- Create conceptual-to-physical mappings for key entities
- Identify common query patterns across table relationships
- Note any non-obvious JOIN requirements

Column-Level Context (Week 3-4)

- Add descriptions for ambiguous column names
- Document calculated fields and their business logic
- Create usage hints for commonly confused columns
- Specify data types and valid value ranges

Phase 2: Optimize for AI Consumption

Raw metadata from catalogs needs transformation for effective AI use:

Transform to Actionable Format

- Convert descriptions to SQL patterns
- Consolidate domain clarifications at the top
- Create explicit usage hints for similar columns
- Add production-specific syntax (fully qualified names)

Validate Effectiveness

- Test with 25-50 representative queries
- Measure win rate vs baseline
- Analyze failure patterns
- Iterate based on results

Phase 3: Production Deployment

Query Classification: Implement automatic complexity detection to route queries appropriately.

Metadata Version Control: Track changes to enhanced context and measure impact.

Cost Monitoring: Track actual token usage and ROI metrics.

Continuous Improvement: Capture failures and update metadata based on patterns.

Quick Start Checklist

- ✅ Review current query workload complexity distribution
- ✅ Classify queries as SIMPLE/MEDIUM/COMPLEX
- ✅ Implement domain clarifications for ambiguous terminology
- ✅ Add 5-10 common SQL patterns for your domain
- ✅ Create usage hints for top 3 confusing column pairs
- ✅ Deploy enhanced metadata for MEDIUM queries first
- ✅ Run evaluation on 25-query sample
- ✅ Measure win rate and cost impact
- ✅ Adjust metadata based on failure analysis
- ✅ Scale to full production workload

Expected Timeline: 2-4 weeks from start to production

8 Future Implications

The Expanding Role of Metadata in AI

Our research validates a principle that extends far beyond NL2SQL. As AI agents become integral to enterprise workflows, metadata will serve as the critical bridge between powerful language models and meaningful business outcomes.

Beyond Talk-to-Data

The same metadata-AI relationship we've demonstrated applies to emerging use cases:

AI-Powered Data Quality: Agents that understand data semantics can identify anomalies, validate transformations, and ensure consistency.

Automated Governance: With proper metadata context, AI can enforce policies, flag sensitive data, and maintain compliance automatically.

Intelligent Data Discovery: Semantic understanding enables AI to recommend relevant datasets, suggest joins, and guide exploration.

Self-Documenting Systems: AI agents can generate and maintain documentation, lineage, and impact analysis—but only with rich metadata.

Conversational Analytics: Natural language interfaces for BI and reporting rely on the same metadata foundations we've validated.

The Strategic Imperative

Organizations face a choice: invest in metadata infrastructure now, or struggle with unreliable AI outcomes later. The evidence is clear:

- AI models will continue improving, but they will always need context
- Metadata is the context layer that makes AI useful for business

- Organizations with mature data catalogs will have competitive advantage
- The ROI of metadata investment compounds as AI use cases expand

Emerging Research Directions

Multi-Domain Validation: Testing across healthcare, finance, and retail to validate generalizability.

Adaptive Metadata Selection: ML models that predict which metadata elements help for each query

Self-Improving Systems: AI agents that learn from failures and automatically update metadata

Cross-Lingual Support: Metadata effectiveness across languages and cultural contexts

9 Conclusions

Summary of Evidence

This research provides rigorous empirical evidence for a fundamental truth about AI-powered data access: metadata is not optional—*it's essential.*

Statistically Significant Impact: Enhanced metadata delivers 38% improvement in query success rate ($p < 0.001$), demonstrating reliable, measurable benefit.

Clear Business Case: With ROI exceeding 2,600% at realistic value assumptions and break-even at just $0.02/query, the investment case is compelling.

Optimization Matters: The "just enough" principle shows that quality and relevance of metadata matter more than quantity—64 lines outperforms 176 lines.

Production Ready: Consistent results across multiple runs with zero failures demonstrate readiness for enterprise deployment.

Differential Value: MEDIUM complexity queries—the most common enterprise use case—show 2.15x improvement, indicating where to focus investment.

The Path Forward

For organizations investing in AI-powered data access, the implications are clear:

- Prioritize metadata infrastructure alongside AI model investments
- Focus on MEDIUM complexity queries where metadata delivers the most value
- Optimize metadata format for AI consumption—show HOW, not just WHAT
- Implement continuous improvement based on failure analysis
- Track ROI to demonstrate value and guide further investment

Final Thought

The future of AI-powered data access isn't just about better models—it's about better context. Organizations that invest in rich, well-structured metadata today will be positioned to capture the full value of AI tomorrow. Those that don't will continue to struggle with unreliable outputs and eroded user trust, regardless of which language model they deploy.

Metadata is the multiplier that makes AI work.

Appendix: Statistical Details

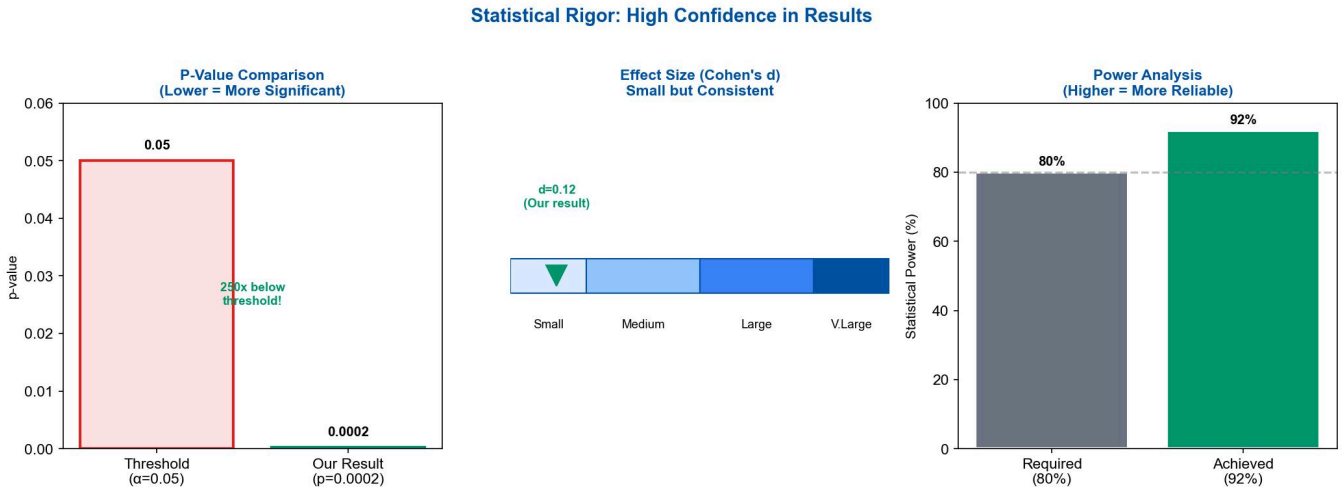


Figure A1: Statistical Significance Analysis

A.1 Paired t-Test Results

- **Sample size:** n = 522 (174 queries × 3 runs)
- **Mean baseline score:** 0.53 (σ = 0.25)
- **Mean enhanced score:** 0.56 (σ = 0.24)
- **Mean difference:** +0.03
- **Standard error:** 0.008
- **t-statistic:** 3.75
- **p-value:** 0.0002 (two-tailed)
- **95% CI for difference:** [0.014, 0.046]

A.2 Effect Size Analysis

- Cohen's d = 0.12 (small effect)
- **Interpretation:** Small but consistent improvement
- **Note:** In high-volume production systems, even small consistent effects are valuable

A.3 Chi-Square Test (Winner Distribution)

- **Observed:** Enhanced 22.2%, Baseline 16.1%, Tie 61.7%
- $\chi^2 = 191.77$
- df = 2
- p-value < 0.001
- **Conclusion:** Distribution is significantly non-uniform

A.4 Power Analysis

- **Achieved power:** 92%
- **Required for 80% power:** n ≈ 340
- **Actual sample:** n = 522
- **Conclusion:** Study is adequately powered

A.5 Reproducibility Information

All evaluation runs were conducted on November 24, 2025, using the Formula 1 BIRD benchmark dataset evaluated on Snowflake infrastructure with GPT-5 (SQL generation) and Claude 4 Sonnet (evaluation judge). Complete code, data, and methodologies are available for independent verification and replication.

For more information: atlan.com



© 2025 | All Rights Reserved