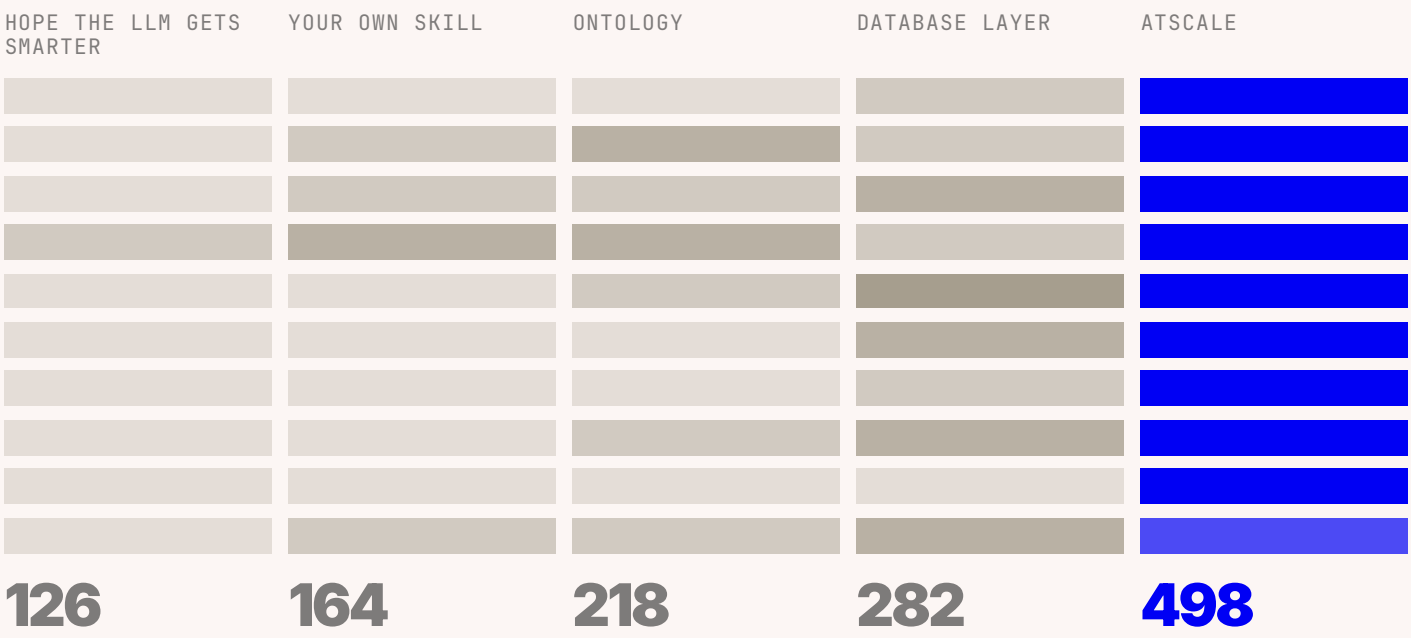
THE COMPLETE BUYER'S GUIDE TO A SEMANTIC LAYER FOR AI.



ROWS FOLLOW THE TEN EVALUATION CRITERIA, 01 AT THE TOP, SHADED BY SCORE (DARKER IS STRONGER). TOTALS ARE WEIGHTED SCORECARD POINTS, OUT OF 510.

CONTENTS

YOUR AI STRATEGY STARTS WITH THE SEMANTIC LAYER.

Every AI strategy rests on two questions: can you trust the numbers, and can you afford them?

01	**Can you trust the numbers?**	03
02	**Can you afford the answers?**	05
03	**The path to 100%**	06
04	**Five signs you need one**	07
05	**Ten criteria to evaluate**	08
06	**How AtScale fits the criteria**	19
07	**Run a 30-day proof of value**	20
08	**The scorecard**	21
09	**Download the scorecard**	22

CAN YOU TRUST THE NUMBERS?

Ask an AI model a business question against raw tables and it's left guessing what each word means and where the data lives. A semantic layer swaps that guess for a certified definition and computes the answer. On BIRD-Interact, a public benchmark, that alone doubled accuracy.

46%

AI, left on its own to guess at SQL, gets enterprise questions right 46% of the time.*

93%

once an AtScale semantic layer is generated automatically, with no tuning.

100%

is within reach once people, process and tools catch and fix every miss.

*** About these numbers.** BIRD-Interact is an independent benchmark from the BIRD team and Google Cloud: 600 tasks over 22 messy PostgreSQL databases, with vague questions and a simulated user who holds the definitions. We ran Claude Sonnet 5 on its 317 answerable business questions, setting aside 93 whose answer key is itself wrong, once against the raw database and once through an automatically generated AtScale semantic layer. Both arms used the same grader, with the same six grading corrections. Without the layer it answered 46% correctly; with it, 93% (296 of 317). Results are preliminary, October 2026, and not yet verified by the BIRD team. Read the white paper at atscale.com/bird.

2x

Simply adding a semantic layer makes AI two times more accurate.

On BIRD-Interact's answerable business questions, Claude Sonnet 5 went from 46% to 93% once an AtScale semantic layer was generated automatically, with no tuning. Preliminary results, October 2026.

CAN YOU AFFORD THE ANSWERS?

Right answers cost less. An AI that guesses keeps spending tokens and compute on attempts, while one that calls a computed metric gets it in one pass.

AI SPEND

18%

Cheaper right answers.

On average, a right answer cost 18% less in API fees with AtScale, from 7% to 37% depending on the model.*

AGENT EFFORT

2.3x

Less work per answer.

Agents call certified metrics by name instead of rebuilding joins and formulas. Fewer calls, fewer retries.**

WAREHOUSE COMPUTE

10x

Lower query cost.

Aggregates and caching answer repeat questions without rescanning raw tables.***

Guessing keeps burning compute. AtScale gets the answer right the first time, so you pay for answers, not attempts.

* The simple average of API-cost savings per correct answer across all five agent models in AtScale's BIRD-Interact-Full run (410 query tasks, leaderboard rules, list API prices, best run of three): Sonnet 5 7%, Kimi K2.5 13%, Opus 4.6 16%, Grok 4.3 18%, GPT-5 37%.
** BIRD-coins spent per correct answer, Claude Sonnet 5 on the corrected harness: 19.9 with AtScale vs 46.4 without.
*** AtScale TPC-DS benchmark on Google BigQuery, on-demand pricing.
BIRD-Interact results are preliminary, October 2026, and not yet verified by the BIRD team.

THE PATH TO 100%.

Doubling accuracy is the start. Four products work in one loop to close the rest of the gap, and the ten criteria that follow show what to look for in each.

1 · RESOLVE

AVA

Asks which metric a vague question means. Agents call certified metrics by name through MCP.

2 · COMPUTE

ACE

Computes every metric the same way, inside your data platform, and returns the SQL.

3 · MODEL

COMODEL

Drafts model changes in SML from your schemas and docs, for a person to approve.

4 · ATTEST

CANARY

Scores answers on your questions with and without AtScale and sorts every miss by cause.

UNDERNEATH THE LOOP: THE UNIVERSAL SEMANTIC LAYER. CANARY AND COMODEL ARRIVE THIS FALL.

FIVE SIGNS YOU NEED A SEMANTIC LAYER FOR AI.

1

Anecdotally, AI gives many wrong answers.

On BIRD-Interact, accuracy runs from 20% to above 90%. Below 90%, your skills or semantics need work.

2

Metric logic is trapped in your database.

Definitions sit in views and platform semantic views tied to one engine, so every other tool rebuilds them.

3

Agents are about to act on numbers.

A person reads a chat answer first. An agent places the order or books the accrual with nobody checking.

4

AI queries are driving up warehouse cost.

Every question can become a big query, and an agent that retries failed SQL pays for every attempt.

5

You don't measure AI accuracy or cost.

With no test set of your own questions, nobody can say how often the AI is right or what each answer costs.

TEN CRITERIA TO EVALUATE.

Ten tests for the semantic layer behind your AI strategy. Tap any criterion to jump to it.

01 AI metric computation →

02 Prompt disambiguation →

03 Semantic layer foundation →

04 Sovereign semantics →

05 Semantic governance →

06 SQL optimization →

07 Performance optimization →

08 AI cost containment →

09 Accuracy attestation →

10 AI-assisted semantic modeling →

AI METRIC COMPUTATION

Wrong numbers come from SQL the model makes up. Don't let it write any.

Key criteria

Computed, not generated
The layer computes the metric, so the model never writes the SQL.

Correct at every level
Ratios, distinct counts and semi-additive balances roll up correctly.

Time-aware
Fiscal calendars, period comparisons and hierarchies are built in.

Pushed down
Computation runs inside your data platform, where the data lives.

Questions for vendors

Ask the vendor

When an agent asks for a ratio by region, what generates the SQL, and how are ratios, distinct counts and semi-additive measures handled at each level?

Test it yourself

Ask for average order value by region and by quarter, and inventory on hand across weeks. Check them against numbers you computed by hand.

How AtScale does it

ACE computes each metric inside your data platform and rebuilds ratios from numerator and denominator at every level.

Common eval mistake: Ontologies and catalogs

Expecting an ontology, catalog or context layer to fix accuracy. They hand the model hints, and the model still writes the SQL. As Andrew Brust puts it, hints don't lead to AI accuracy.

PROMPT DISAMBIGUATION

Many wrong answers start with a vague term that maps to more than one metric.

Key criteria

Knows the candidates
Holds every certified metric a vague term could map to.

Asks first
Stops and asks which one the person means before anything runs.

Remembers the answer
Stores each answer as a rule or synonym, so it doesn't ask twice.

Shows its source
Every answer names the metric and definition behind it.

Questions for vendors

Ask the vendor

What happens when a question could mean more than one metric?

Test it yourself

Ask "What was our deal size last quarter?" and see whether it picks a metric silently or asks which one.

How AtScale does it

Ava reads the semantic model, asks which metric a vague question means, and answers with its source.

Common eval mistake: Custom skill files

Testing only clear questions. Catalogs, ontologies and skill files pass the model context and let it pick a metric on its own. Put a vague question in your test set, like "What was deal size last quarter?", and watch whether the tool asks.

SEMANTIC LAYER FOUNDATION

Every AI answer rests on the semantic layer, so it has to be open, proven, governed and still moving forward.

Key criteria

Open
Definitions in an open format, kept in your own Git repository.

Proven
Years in production at enterprise scale, with references you can call.

Governed
One certified definition per metric, and a person approves every change.

Innovative
Measures and attests accuracy, so the model gets better over time.

Questions for vendors

Ask the vendor

How long has the layer run in production, at what scale, how is each definition governed, and how do you measure accuracy?

Test it yourself

Call two references running AI on the layer at scale. Then export a model to your Git repository and push one change through approval.

How AtScale does it

AtScale has run in production for more than a decade at companies such as TELUS, Vodafone and Wayfair. Models are written in SML, an open-source format kept in your Git repository, and every change is reviewed. Canary and CoModel, arriving this fall, add accuracy attestation and AI-assisted modeling.

Common eval mistake: Build-your-own layers

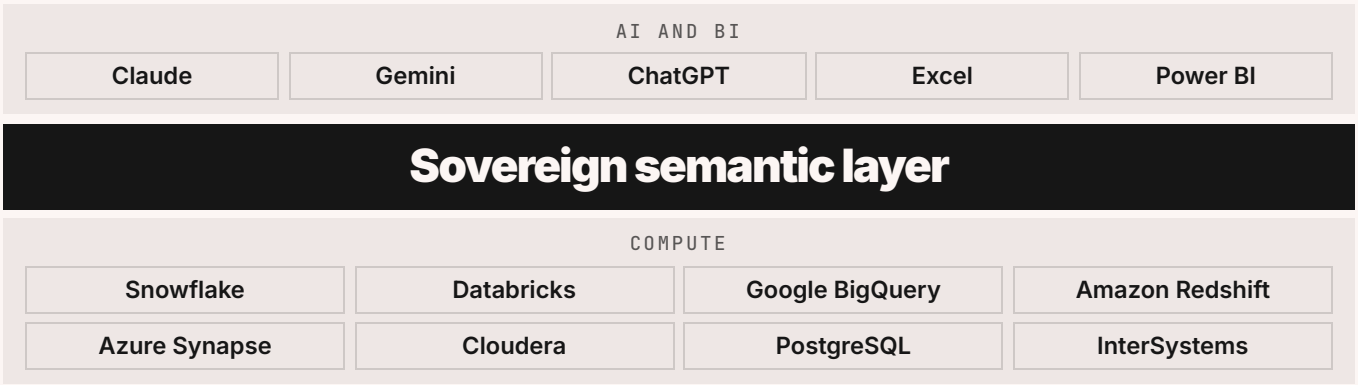
Choosing a layer on a demo alone. A young tool or a skill file can look complete in a pilot. Ask how long it's been in production, at what scale, and which customers will take your call.

SOVEREIGN SEMANTICS

Pick your LLM and your data platform on their merits, and switch either one without rebuilding a single metric.

Key criteria

A sovereign semantic layer connects every AI and BI tool to every data platform you run. Define a metric once, and it's the same answer everywhere.



Questions for vendors

Ask the vendor

If I define my metrics in your tool, can I use them in a competitor's? If I define them in Snowflake, can I use the same definitions in Databricks?

Test it yourself

Define one metric, then query it from Claude and Excel on two different data platforms. You should get the same number without rebuilding anything.

How AtScale does it

One model serves Claude, Gemini, ChatGPT and your agents over MCP, Excel and Power BI over XMLA, and SQL tools over JDBC and ODBC. ACE computes on Snowflake, Databricks, Google BigQuery, Amazon Redshift, Azure Synapse, Cloudera, PostgreSQL and InterSystems, and the data stays where it is.

Common eval mistake: Warehouse semantic views

Building your semantics into the data warehouse alone. Platform-native semantic views are useful, but they tie your definitions to one engine. Sovereign semantics stay independent of the platform underneath.

SEMANTIC GOVERNANCE

A metric definition should ship the way production code does: versioned, tested, approved and traceable.

Key criteria

Versioned
Every change to a metric is a commit with an author, a timestamp and a diff. A rollback is a revert.

Shipped through CI/CD
A change opens a pull request, tests run, a named reviewer approves, then it moves from dev to QA to production.

Authenticated
Every person and agent signs in with its own identity and sees only what its policies allow.

Traceable
Any number traces back to the definition version and the SQL that produced it.

Questions for vendors

Ask the vendor

How does a change to a metric reach production, who approves it, and can you show the definition version behind any past answer?

Test it yourself

Change one metric in a branch, push it through review to production, then roll it back. Query it as two users and check who saw what.

How AtScale does it

Every SML file lives in Git. A change opens a pull request, must compile with access rules unchanged, and needs a named reviewer before it's promoted to production. People and agents sign in through Okta, Microsoft Entra ID or Active Directory.

Common eval mistake: Data catalogs

Not governing semantics at all, or assuming a governed catalog covers it. A catalog, ontology or context layer governs descriptions and table access. The calculation the AI runs still needs a version, a review and an approval.

SQL OPTIMIZATION

AI-written SQL is often long, slow and different every time. You need to see the SQL behind every answer and know it's tuned for your platform.

Key criteria

Visible
The SQL behind every answer is available by query ID.

Platform-tuned
SQL is written for each platform's dialect and optimizer.

Lean
Modeled joins and formulas replace long AI-written queries.

Consistent
The same question produces the same SQL every time.

Questions for vendors

Ask the vendor

Can we see the SQL behind any answer, and is it optimized for our data platform?

Test it yourself

Pull the SQL behind a few answers and compare it with what your best engineer would write.

How AtScale does it

ACE writes SQL tuned for each platform, and the MCP server returns the SQL behind any past query. On one BIRD-Interact task, 21 lines of AI-written SQL became 3 lines through AtScale, and both returned the right number.

Common eval mistake: Text-to-SQL agents

Accepting an answer with no trail. Skill files and text-to-SQL agents write fresh SQL each time and rarely keep it. Ask for the SQL behind an answer from yesterday.

PERFORMANCE OPTIMIZATION

AI sends far more queries than dashboards did. Rescanning raw tables for each one runs up the bill.

Key criteria

Aggregate-aware
Queries route to the smallest structure that can answer them.

Self-tuning
Aggregates are built and retired from real query patterns.

Cached
Repeat questions return without rescanning raw tables.

Built for AI volume
Response stays fast as agents multiply the query load.

Questions for vendors

Ask the vendor

How do you keep warehouse cost down as AI query volume grows?

Test it yourself

Replay a day of expected AI questions and measure warehouse credits and latency, with the layer and without.

How AtScale does it

ACE builds and manages aggregates from query patterns, so repeat questions skip the raw tables.

In a production benchmark at a Tier 1 bank, five everyday AI questions billed **21,903x fewer bytes** through AtScale than when the LLM queried BigQuery directly. The cost fell from $17.93 to less than a tenth of a cent.

Common eval mistake: Raw-table queries

Benchmarking on a small sample. Raw-table scans look cheap on demo data and expensive at production scale. Test against full-size tables at a realistic query volume.

AI COST CONTAINMENT

An agent that crawls schemas and retries failed SQL pays for every attempt, and most teams can't see what a right answer costs.

Key criteria

Fewer tokens
Agents call metrics by name instead of crawling schemas.

Fewer retries
Computed answers come back right the first time more often.

Cost per right answer
Spend is measured per correct answer, not per call.

Model choice
Cheaper models become viable once the layer does the hard work.

Questions for vendors

Ask the vendor

How many tokens and tool calls does an agent spend per correct answer, and can you report spend per right answer, not just per query?

Test it yourself

Run your test set with the layer and without. Log tokens and retries, then divide total spend by the answers that came back right.

How AtScale does it

An agent calls a certified metric by name and skips the schema crawling.

On BIRD-Interact, a right answer cost **18% less** in API fees with AtScale, averaged across five models. Results are preliminary.

Common eval mistake: The LLM on its own

Comparing price per token, or budgeting by query count. A layer can spend more per call and still cost less per right answer. On BIRD-Interact, spend per run rose while cost per right answer fell 18% on average (preliminary).

ACCURACY ATTESTATION

A vendor's demo questions won't tell you much about your data. Yours will.

Key criteria

Your questions
Test sets come from what your people actually ask.

Gold answers
Each test has a validated result and gold SQL.

Diagnosed misses
Every wrong answer gets a cause: prompt, model, computation or data.

Before and after
Each model change shows whether it fixed what it was meant to.

Questions for vendors

Ask the vendor

How do you measure answer accuracy, and can we see the result for each question?

Test it yourself

Run 20 of your own questions per domain, each with a validated answer, through the AI with the semantic layer and without it.

How AtScale does it

Canary, arriving this fall, runs your tests with and without AtScale and sorts every miss into prompt, model, computation or data.

Common eval mistake: Vendor demos

Trusting the vendor's demo. Demo questions are picked to pass. Bring 20 of your own, with answers you know are right, and score every one.

AI-ASSISTED SEMANTIC MODELING

Agents ask about more of the business than analysts ever did. A model that takes months to build holds back every AI project.

Key criteria

Drafted by AI
Agents draft the model from schemas, docs and pipeline code.

Approved by people
A named reviewer approves every change.

Tested before it ships
Each draft runs against your test cases first.

Kept current
Fixes from failed tests flow back into the model.

Questions for vendors

Ask the vendor

How does the model get built from our schemas, docs and pipeline code, and who approves each change?

Test it yourself

Point the tool at a subject area you haven't modeled and time how long it takes to reach a tested, approved model.

How AtScale does it

CoModel, arriving this fall, drafts the model in SML from your schemas, docs and existing metadata, tests it through Ava, and waits for a person to approve each change.

Common eval mistake: Unreviewed AI modeling

Treating an AI-drafted model as finished. A model nobody reviews can encode the same guesses the LLM would make. Ask who approves each change and how it's tested before it ships.

THE AI ACCURACY FLYWHEEL.

Four products work in one loop, and each turn fixes what the last one missed.

1 · RESOLVE

AVA

Asks which metric a vague question means. Agents call certified metrics by name through MCP.

2 · COMPUTE

ACE

Computes every metric the same way, inside your data platform, and returns the SQL.

3 · MODEL

COMODEL

Drafts model changes in SML from your schemas and docs, for a person to approve.

4 · ATTEST

CANARY

Scores answers on your questions with and without AtScale and sorts every miss by cause.

UNDERNEATH THE LOOP: THE UNIVERSAL SEMANTIC LAYER. CANARY AND COMODEL ARRIVE THIS FALL.

EVERY TURN GETS CLOSER TO 100%

COMPARE VENDORS ON YOUR QUESTIONS BEFORE FEATURE LISTS.

Each step tests specific criteria. Together they cover all ten.

STEP 1

Pick 20 questions per domain.

Choose questions your team asks often, from one or two subject areas, with answers you know are right. Include a few vague ones and a few multi-pass calculations.

TESTS 09

STEP 2

Set the baseline.

Run them through your AI client against raw tables, with no semantic layer. Record accuracy, latency, tokens and warehouse cost.

TESTS 08

STEP 3

Build the model.

Model the subject area from your existing schemas and docs, and record how long it took.

TESTS 03 · 10

STEP 4

Rerun the questions.

Run the same questions with the semantic layer and compare each result with the baseline.

TESTS 01 · 02 · 04 · 07

STEP 5

Test one agentic workflow.

Have an agent retrieve a metric, act on it in a test system, and log the SQL behind the number.

TESTS 05 · 06

STEP 6

Review the misses.

For every wrong answer, find the cause (prompt, model, computation or data) and time the fix.

TESTS 09

AtScale runs this benchmark with you on your questions, and you keep the model it builds.

SCORE THEM YOURSELF.

Scores run 1 (doesn't do it) to 5 (does it fully). Weights run 1 to 5. Red marks a 1 or 2. The full scorecard adds database-specific layers and an LLM on its own.

CATEGORY	FEATURE	WEIGHT	BUILD YOUR OWN SKILL	ONTOLOGY LAYER	ATSCALE
01 AI METRIC COMPUTATION	Computes each metric and generates the SQL, so the model never writes its own	5	1	1	5
	Correct ratios, distinct counts and semi-additive balances at every level	5	1	1	5
	Time intelligence, fiscal calendars and hierarchies	4	1	2	5
02 PROMPT DISAMBIGUATION	Asks which metric a vague question means before anything runs	4	2	3	5
	Stores each answer as a rule or synonym	3	2	3	5
03 SEMANTIC LAYER FOUNDATION	Keeps one definition per metric, with no drift between copies	4	1	3	5
	Stores the model as open code in your Git repository, with CI/CD	3	4	2	5
	Supports Apache Ossie for exchanging definitions	3	1	2	5
04 SOVEREIGN SEMANTICS	Serves AI clients, SQL tools, Excel and Power BI from one model	4	1	2	5
	Runs on any data platform	5	4	4	5
	Works with any AI agent, AI client or LLM	5	3	4	5
	Fully supports MCP: discovery, queries and the SQL behind each answer	4	3	4	5
	Serves certified metrics to agents through MCP	5	3	2	5
05 SEMANTIC GOVERNANCE	Gives each agent its own identity, access and row policies	5	1	3	4
	Passes each user's identity through to the data platform	5	1	2	5
	Applies row- and column-level security the same way for every client	5	1	2	5
	Integrates with your single sign-on	3	1	3	5
06 SQL OPTIMIZATION	Returns the exact SQL behind any past answer	4	1	1	5
07 PERFORMANCE OPTIMIZATION	Builds and manages aggregates automatically	3	1	1	5
08 AI COST CONTAINMENT	Keeps token and warehouse cost per answer low	4	1	2	5
09 ACCURACY ATTESTATION	Measures accuracy on your own questions, one at a time	5	1	1	5*
	Compares accuracy with and without the semantic layer	4	1	1	5*
	Diagnoses why each wrong answer went wrong	4	1	1	4*
	Publishes reproducible results on a public benchmark	3	1	1	5
10 AI-ASSISTED SEMANTIC MODELING	Drafts the model with AI, with a person approving every change	3	2	2	4*
	Weighted total, out of 510		**164**	**218**	**498**
	Share of maximum		32%	43%	98%

* Arrives with CoModel and Canary this fall. Scores reflect a typical implementation, based on AtScale's assessment.

See all five approaches, change the weights and add your vendors.

OPEN THE SCORECARD →

DOWNLOAD THE SCORECARD. THEN GIVE US A CALL.

Score the vendors you're considering against the ten criteria. When you're ready to test them on your own questions, we'll run the benchmark with you.

Download the evaluation scorecard

Talk to AtScale