FIELD GUIDE

ATSCALE FOR AGENTIC AI

AtScale is the metric calculator for agentic AI. Your agents get certified numbers over MCP today. Canary, coming this fall, will test every answer against the BIRD-Interact test suite before automation acts on it.

ATSCALE

Most enterprises still use AI only for chat.

McKinsey, Deloitte and Gartner found that few companies let agents act on business numbers yet.

10%

of companies, at most, scale AI agents in any single business function.

MCKINSEY, STATE OF AI 2025

21%

of companies have mature governance for autonomous agents.

DELOITTE, STATE OF AI IN THE ENTERPRISE 2026

60%

of MCP-only agentic analytics projects will fail by 2028 without a semantic layer.

GARTNER, MARKET GUIDE FOR AGENTIC ANALYTICS, FEBRUARY 2026

It's easy to miss the two errors AI made on lines 209 and 214.

Tomorrow's reorder list

LINES 208 TO 215 OF 600

LINE	SKU	CORRECT ORDER	AGENT'S ORDER
⋮ LINES 1 TO 207			
208	40-110	620	620
209	**40-112**	**1,240**	430 −65%
210	40-115	860	860
211	40-118	1,240	1,240
212	40-122	2,480	2,480
213	40-130	980	980
214	**40-131**	**1,560**	2,340 +50%
215	40-134	740	740
⋮ LINES 216 TO 600			

Take a retailer whose agent builds tomorrow's reorder list overnight. It adds inventory across weeks and orders 810 units short on line 209, then counts a transfer twice and orders 780 too many on line 214. The buyer checks 600 lines in the morning and approves them. Both errors ship.

A better model won't fix this. An agent that writes its own SQL has to guess what the numbers mean, and it guesses differently each run.

Every model got more accurate with a semantic layer.

AtScale ran five agent models from four vendors on the 410 business questions in the public BIRD-Interact benchmark. Each model answered once against the raw tables and once through a semantic layer AtScale generated automatically. Every model improved, by 79% to 137%. The weakest model with AtScale, Grok 4.3 at 44.6%, beat the strongest model on its own, Claude Sonnet 5 at 25.9%.

MODEL	NO SEMANTIC LAYER	WITH ATSCALE	LIFT
Claude Sonnet 5 ANTHROPIC · FRONTIER	25.9%	56.6%	+119%
Claude Opus 4.6 ANTHROPIC · FRONTIER	22.7%	53.7%	+137%
GPT-5 OPENAI · FRONTIER	21.5%	50.5%	+135%
Kimi K2.5 MOONSHOT AI · OPEN WEIGHT	23.4%	49.3%	+111%
Grok 4.3 XAI · FRONTIER	24.9%	44.6%	+79%

On questions with a correct answer, 46% becomes 93%.

Some of BIRD-Interact's own answer keys are wrong, so we also graded the 317 business questions that have a correct answer. Both arms got the same six grader fixes for dates, letter case, column order, ties and rounding, and we set aside 93 questions where the benchmark's gold SQL is wrong.

CLAUDE SONNET 5 ALONE

46% → **93%**

WITH ATSCALE

PRELIMINARY RESULTS FROM ATSCALE'S OCTOBER 2026 RUNS, SUBJECT TO CHANGE / 410 BUSINESS QUESTIONS, BEST RUN OF THREE / 317-QUESTION SET: BEST RUNS 46% AND 93%, MEANS OF THREE RUNS 45.8% AND 90.6%, NOT COMPARABLE TO THE LEADERBOARD'S SUCCESS RATE / ATSCALE.COM/BIRD

The follow-up question is where the layer pays off.

BIRD-Interact's second phase asks a follow-up, and a task only counts when the first answer and the follow-up both pass. That's how agents work, since each step builds on the last one. On their own, no model got past 16.1%. With AtScale, every model gained 11 to 35 points.

MODEL	NO SEMANTIC LAYER	WITH ATSCALE	LIFT
Gemini 3.1 Pro GOOGLE · FRONTIER	10.2%	45.6%	+345%
Claude Sonnet 5 ANTHROPIC · FRONTIER	15.4%	42.7%	+178%
Claude Opus 4.6 ANTHROPIC · FRONTIER	16.1%	40.2%	+150%
GPT-5 OPENAI · FRONTIER	13.4%	36.3%	+171%
Kimi K2.5 MOONSHOT AI · OPEN WEIGHT	12.0%	30.5%	+155%
Grok 4.3 XAI · FRONTIER	14.4%	25.6%	+78%

Gemini 3.1 Pro went from last to first.

Gemini asks more clarifying questions than any other model. Without a layer, each question spends coins from BIRD-Interact's fixed budget, and Gemini runs out before it finishes the follow-up. With AtScale, the semantic model answers those questions, so Gemini's habit of asking turns into accuracy instead of cost.

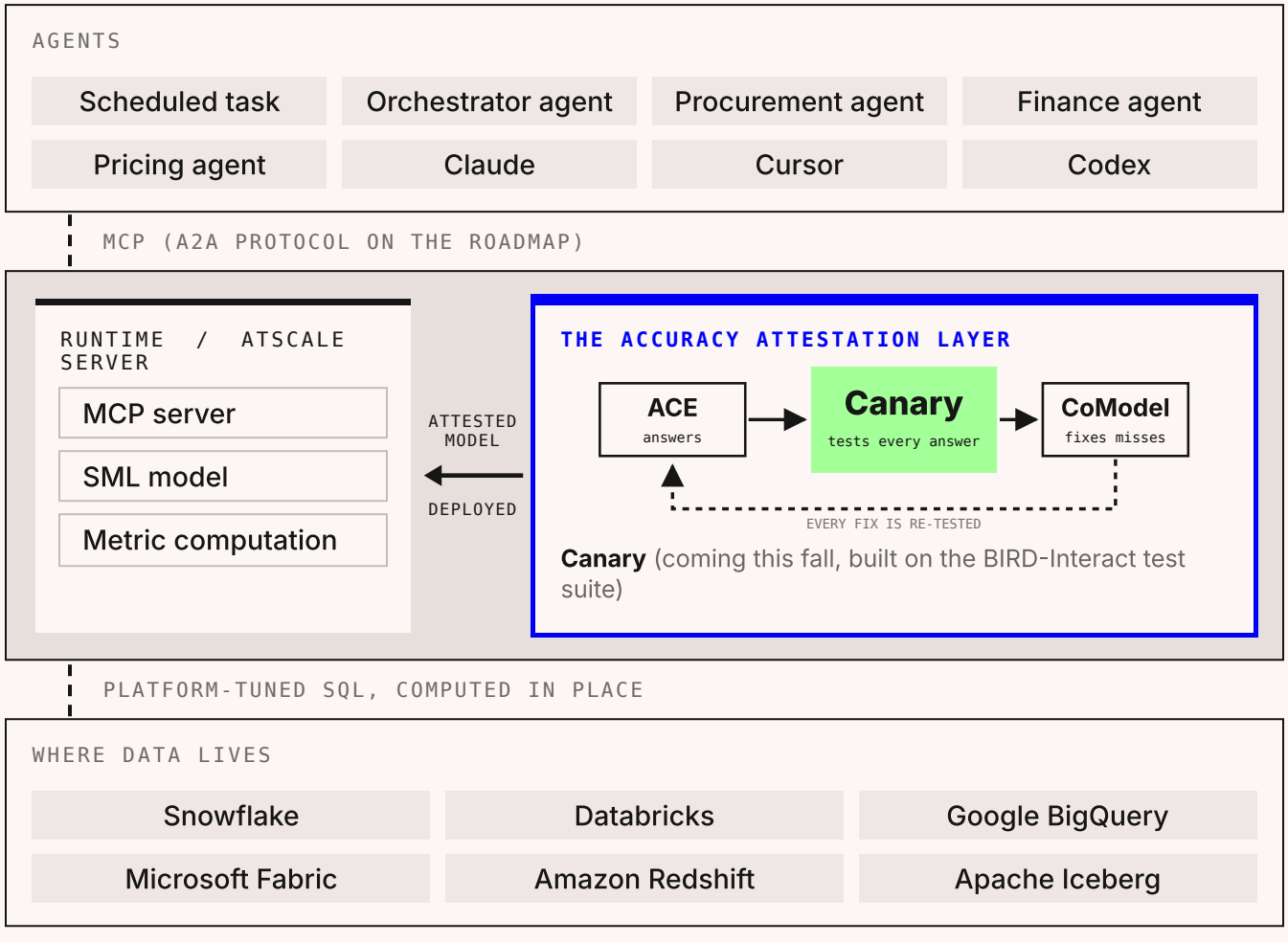
37% chance a 100-step agent workflow gets every step right, even when each step is 99% accurate. Small misses compound, so every point of accuracy counts.

PHASE 2 COUNTS A TASK ONLY IF ITS FIRST QUESTION AND ITS FOLLOW-UP BOTH PASS / ATSCALE: BEST OF THREE RUNS / NO LAYER: ONE RUN PER MODEL / RUNS OCTOBER 3 TO 7, 2026 / PRELIMINARY, SUBJECT TO CHANGE / ODDS = 0.99 RAISED TO THE NUMBER OF STEPS, ASSUMING INDEPENDENT ERRORS

An accuracy attestation layer for AI agents.

Say a procurement agent needs projected stockout days by SKU family. It calls AtScale's MCP server, and ACE computes the number from its certified definition. Nothing gets guessed. Once Canary arrives this fall, it'll test questions like that against known answers first and flag any ambiguous term, so only metrics that pass on your own questions reach agents.

FIG. 01 / WHERE ATSCALE SITS IN AN AGENTIC SYSTEM



FOR HOW THE RUNTIME AND DEVELOPMENT LOOPS WORK, READ THE ATSCALE TECHNICAL OVERVIEW 2026.

Four MCP tool calls between an agent and the right number.

The major agent frameworks all call MCP tools, so every agent reaches AtScale the same way. The answer doesn’t change.

Step 01

Find the metric

MCP TOOL CALL

`list_models`
List attested models

`explore_columns`
So the agent can explore

The agent looks up certified metrics by their business descriptions, so it doesn't have to guess from raw table names.

Step 02

Learn the rules

MCP TOOL CALL

`get_sml_skills`
The agent discovers skills

It reads the instructions for querying the semantic layer, including what to do with an ambiguous term.

Step 03

Get the answer

MCP TOOL CALL

`run_query`
Run the calculation

ACE computes the result, and the agent gets the rows back with a query ID it'll need in step four.

Step 04

Keep the record

MCP TOOL CALL

`get_outbound_queries`
Log the interaction

The workflow pulls the exact SQL ACE sent to the warehouse and logs what it acted on.

The MCP server signs callers in through OAuth, so an agent sees the same rows and columns as the person it's signed in as.

Ask a procurement agent to compute inventory.

ACE computes each metric from one certified definition. Take inventory, which is semi-additive. Ten units in Boston plus ten in Dallas is twenty. Ten units on Monday and the same ten on Tuesday is still ten. Exchange rates and customers who move regions trip agents up the same way. ACE gets all of them right.

AN AGENT WRITING ITS OWN SQL	AN AGENT CALLING ACE
✗ Adds inventory across weeks, sees stock that isn't there and orders too little.	✓ Reads the balance on the date asked, since the model defines inventory once.
✗ Books last year's sales to each customer's current region and converts every euro at today's rate.	✓ Uses each customer's region at the time of the sale, and that day's exchange rate.
✗ Tomorrow it may calculate the same number a different way.	✓ Tomorrow's number matches today's, and you can see the SQL that produced it.

AtScale for agentic computing.

ATSCALE MCP

One interface for every agent.

Agents reach AtScale through its MCP server, the open standard most agents and agent frameworks already speak. OAuth sign-in carries each caller's permissions through to the data.

ATSCALE ACE

The calculator agents call.

ACE holds the rules for every metric and applies them the same way on every query. It pushes platform-tuned SQL to the warehouse, and the data stays where it lives.

Test the math before agents act on it.

Coming this fall. Canary will turn the questions a workflow asks into test cases, score them with AtScale and without it, and sort every miss by cause. When you switch models, it'll run the suite again.

ATSCALE CANARY

A model that keeps up with agents.

Agents ask about more of the business than any analyst team ever did. Coming this fall, CoModel will draft each model change in SML for a person to approve, and the change will ship through Git like any other code.

ATSCALE COMODEL

What makes AtScale agent-ready today.

An MCP server
Six tools any MCP client can call, from finding a metric to pulling the SQL behind an answer.

Certified metrics in SML
Each metric is defined once in an open-source language and reviewed in Git like any other code.

The same math every time
ACE handles semi-additive metrics and currency conversion the same way on every query.

A record of every query
get_outbound_queries returns the exact SQL behind an answer, so a workflow can log what it acted on.

Point-in-time answers
See which model version and SQL produced an answer an agent gave last month.

Agent sign-in
OAuth through Okta or Entra. Row and column rules follow the person the agent works for.

Rules for each agent
Each agent signs in as itself, with its own access to models and data.

Data stays in place
Platform-tuned SQL runs on Snowflake, Databricks, BigQuery, Fabric, Redshift and Iceberg.

ON THE WAY

Canary
Coming this fall, it'll test computations against known answers.

CoModel
Coming this fall, it'll draft model changes for a person to approve.

A2A support
On the roadmap, so agents on other platforms can reach AtScale over Agent2Agent.

Nine agents to build on AtScale, one for each industry.

Each one acts on a metric that's easy to get wrong. ACE computes it the same way every time.

INDUSTRY AND AGENT		METRIC	THE TRAP
RETAIL	**Store-planning agent** Sets labor and stock by store.	**Same-store sales**	Which stores count as comparable, on a 4-5-4 calendar
FINANCE	**Treasury agent** Flags products losing margin.	**Net interest margin**	Average balances don't sum across time
HEALTHCARE	**Care-transition agent** Calls patients likely to return.	**30-day readmission rate**	A rolling window from each discharge, minus exclusions
INSURANCE	**Reserving agent** Sets reserves as claims reopen.	**Loss ratio**	Accident year versus calendar year
MEDIA	**Licensing agent** Ranks which titles to renew.	**Views and subscribers**	Three seconds or a full episode, one household or four
MANUFACTURING	**Maintenance agent** Fixes the lines losing hours.	**Availability**	Machine, ERP and sensor clocks at different grains
SUPPLY CHAIN	**Carrier agent** Moves freight off late carriers.	**On-time delivery**	Promised versus requested date, and partial deliveries
TELECOM	**Retention agent** Sends offers where churn rises.	**Churn**	Monthly churn doesn't add up into a quarter
ADTECH	**Budget-pacing agent** Shifts spend to what converts.	**Conversions**	Which action counts, over which lookback window

See how ACE computes each metric in the AtScale Industry Guide to AI Metric Computation. READ THE GUIDE →

MAKE OUR DAY. START BUILDING AGENTS WITH ATSCALE.

We're building agentic systems with partners right now. Book an AtScale agentic workshop and start on yours.

Book a workshop →

ATSCALE