AtScale's BIRD-Interact Benchmark Results, Methodology and Analysis

What seven AI models scored on BIRD-Interact with and without a semantic layer, and what it'll take to close the last 7%.

Dave Mariani, CTO and Founder

Jay Schuren, President

Mark Palmer, CSMO

ATSCALE

An AI benchmark built on business questions.

> **“Only 5% to 10% of the current value of the technology” is actually being used.**
>
> Dario Amodei, CEO of Anthropic, on stage with Marc Benioff, 15 September 2026. Benioff called enterprise adoption uneven; Salesforce is spending $300 million with Anthropic this year.

Ask an AI model a business question and you’ll get a confident, well-written answer.

Sometimes it’s wrong, and you can’t tell by reading it.

In a chat window that’s an annoyance. In a forecast or a board deck, it’s a real cost.

BIRD-Interact measures how often it happens. It poses 600 business questions against 22 messy PostgreSQL databases, withholds the documentation that says which columns to use, and caps what each task can spend. We ran it twice per model, once with AtScale’s semantic layer between the model and the database and once without.

This brief covers how the benchmark works, what each model scored, and the 21 questions we still missed, since those show what 100% will take.

What's inside.

SECTION ONE · BIRD-INTERACT, EXPLAINED

How BIRD-Interact works 05

Finally, a real-world AI accuracy test 06

Eight real questions, and where each one goes wrong 07

The public BIRD-Interact leaderboard, September 16, 2026 08

SECTION TWO · THE TEST AND THE RESULTS

The architecture we ran 10

Three takeaways from BIRD-Interact 11

All 600 tasks, with and without AtScale 12

410 business questions, with and without AtScale 13

The follow-up question 14

Gemini 3.1 Pro: the best question-asker 15

The cost of a right answer 16

But 50% accuracy stinks 17

SECTION THREE · THE HARD PATH TO 100% AI ACCURACY

On the 317 answerable questions, AtScale doubles accuracy 19

21 lines of AI-generated SQL, or 3 lines of AtScale 20

What a semantic layer changes 21

The AtScale AI Accuracy Flywheel 22

Five misses, five fixes 23

The path to the pursuit of 100% AI accuracy 28

SECTION ONE

BIRD-Interact, explained.

BIRD-Interact tests AI the way a business uses it, with vague questions, messy data and a fixed budget. This section covers how it works and where models stand on the public leaderboard.

How BIRD-Interact works.

A simulated user asks a question and pushes back when the answer's off, the way a colleague would. A clarifying question costs a few BIRD COINs, the benchmark's budget unit, and a wrong query costs a lot more.

An incorrect interaction.

"Give me a ranked list of stars in the galaxy."

```
SELECT star_name FROM stars ORDER BY star_name;
```

"That's alphabetical. I asked for a ranking."

```
SELECT star_name FROM stars ORDER BY radius_km DESC;
```

"That's by size. Still not what I asked for."

FAIL — Two guesses used up the budget. The user wanted the stars ranked by brightness.

A correct interaction.

"Give me a ranked list of stars in the galaxy."

"Ranked by what? Three measures are defined:"

By size	`radius_km`
By brightness	`luminosity`
By distance	`distance from Earth`

"By brightness."

```
SELECT star_name, luminosity FROM stars
ORDER BY luminosity DESC;
```

"That's correct."

PASS — The agent asked which measure to rank by, then got the answer right.

Finally, a real-world AI accuracy test.

BIRD-Interact builds in four conditions you'll find in any production warehouse. Scoring is strict: an answer has to match the right number, and a close one scores zero.

01

Ambiguous prompts

People type short, vague requests. Someone who wants urgent care facilities ranked by average visit satisfaction types "I need urgent care."

02

Vague semantics

Nothing in the schema is called average visit satisfaction. The number has to be derived from 2,010 columns that never name it.

03

Real-world dirty data

NULL values sit in critical fields, column names disagree across tables, and numbers arrive mixed with their units.

04

Multi-pass metrics

Average visit satisfaction has to be computed over the right calendar and the right set of visits. No column holds it.

THE UNDERLYING DATA AND PROMPTS IN THE BIRD-INTERACT BENCHMARK

600
prompts, one per task

22
PostgreSQL databases, 2,010 columns

17.86
BIRD COINs, the budget for each task

With 17.86 BIRD COINs per task, a model can look at the schema a few times before the budget runs out. Exploring costs coins, so many models guess early, and the scores show it.

Eight real questions, and where each one goes wrong.

These eight come from the public question set. Each depends on a definition, calculation or value the database doesn't spell out.

FAILURE TYPE	THE PROMPT	WHY IT BREAKS
NEEDS A SEMANTIC LAYER	**"How many of our plants are real lemons, both running inefficiently and breaking down all the time?"** solar_panel_5	The database never defines a lemon.
NEEDS A SEMANTIC LAYER	**"For the machines that are always breaking down, what's the average price tag on a single repair?"** solar_panel_17	Easy SQL averages per machine. The question asks per repair.
NEEDS A SEMANTIC LAYER	**"Freeze all those super-risky accounts, marking them suspended unless they already are."** fake_account_m_1	Super-risky needs a certified threshold to act on.
METRIC CALCULATION	**"What's the actual juice each plant is sending to the grid right now?"** solar_panel_10	Net power is a computed value, gross minus loss.
METRIC CALCULATION	**"Build a materialized view containing all ultra-credible accounts."** fake_account_m_4	A weighted composite of three scores, defined in a document.
DIRTY DATA	**"Which of our plants are a headache for warranty claims? I need a list of sites."** solar_panel_13	Warranty state is stored in mixed case.
DIRTY DATA	**"Make a helper that converts percentage strings to numeric."** fake_account_m_6	Percentages are stored as text with a sign attached.
VAGUE PROMPT	**"Add a rule to the loyalty table so the trust value stays within a specific range."** virtual_idol_m_10	The range is never stated.

Prompts from the public bird-interact-full dataset, lightly shortened. Labels are the benchmark's own task IDs. The failure type on each row is our classification of where the meaning is missing.

The public BIRD-Interact leaderboard, September 16, 2026.

Anyone can run BIRD-Interact, and the leaderboard lists only verified results. We compare our numbers with the leaderboard as published on September 16, 2026. The top entry scores 38.0%.

BIRD-INTERACT Leaderboard - Stress Mode ⚡

Verified

Mode: Stress Mode | Dataset: Full (600 tasks) | Interaction Mode: a-Interact | User Simulator: All | Performance: Success Rate

Priority Questions | Follow Ups

Rank	Model	Institute	Link	Date	User Simulator	Efficiency	Success Rate (%)
	MERIT + GPT-5.4			2026-05-22	GPT-4o	17.86	New Sub 38.00
	Claude-Opus-4.6			2026-02-17	Claude-Haiku-4-5	17.86	33.00
	MERIT + Claude-Opus-4.6			2026-05-22	Claude-Haiku-4-5	17.86	New Sub 32.83
4	GPT-5			2025-08-22	GPT-4o	17.86	29.17
5	Kimi-2.5			2026-02-17	Claude-Haiku-4-5	17.86	28.17
6	Claude-Sonnet-4			2025-08-22	GPT-4o	17.86	27.83
7	GLM-4.7			2026-02-17	Claude-Haiku-4-5	17.86	25.83
8	MiniMax-M2.1			2026-02-17	Claude-Haiku-4-5	17.86	24.83
9	Claude-3.7-Sonnet			2025-08-22	GPT-4o	17.86	21.00
10	GPT-4o			2025-08-22	GPT-4o	17.86	21.00
11	Gemini-2-5-Pro			2025-08-22	GPT-4o	17.86	20.33
12	O3-Mini			2025-08-22	GPT-4o	17.86	19.83
13	Qwen-3-235B			2025-08-22	GPT-4o	17.86	19.17
14	Gemini-2.0-Flash			2025-08-22	GPT-4o	17.86	18.17
15	Qwen3-Coder-30B			2026-02-02	Claude-Haiku-4-5	17.86	18.17
16	DeepSeek-Chat-V3.1			2025-08-22	GPT-4o	17.86	17.17
17	Llama-3.1-70B			2026-02-02	Claude-Haiku-4-5	17.86	15.17
18	Qwen3-8B			2026-02-02	Claude-Haiku-4-5	17.86	14.50
19	Qwen3-14B			2026-02-02	Claude-Haiku-4-5	17.86	13.83
20	Qwen-3-Coder-480B			2025-08-22	GPT-4o	17.86	13.33
21	Qwen3-32B			2026-02-02	Claude-Haiku-4-5	17.86	11.17
22	Llama-3.1-8B			2026-02-02	Claude-Haiku-4-5	17.86	2.94

The BIRD-Interact leaderboard

SECTION TWO

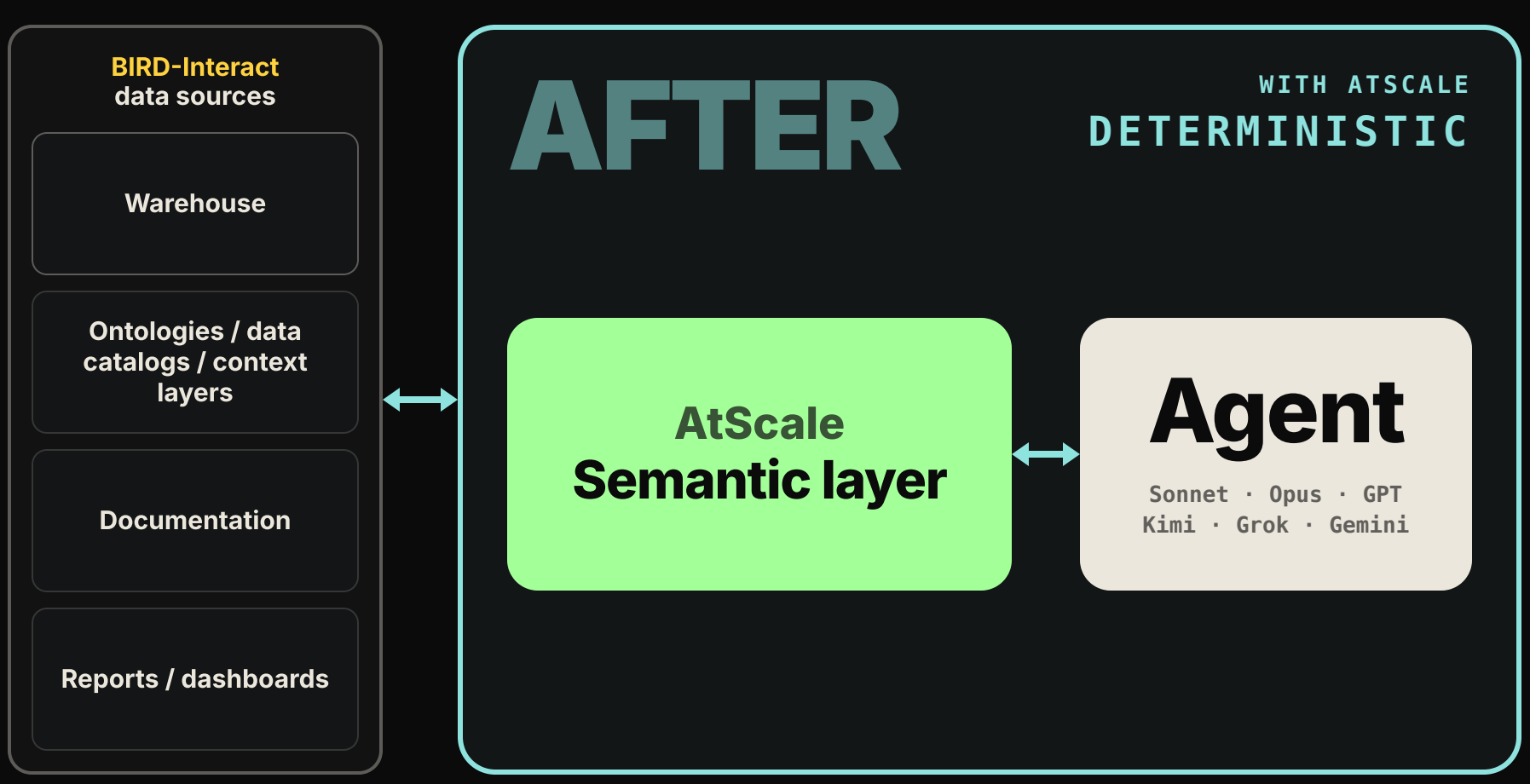
The test and the results.

The architecture we ran, seven models’ scores with and without AtScale, why Gemini 3.1 Pro gained the most on the follow-up question, the cost of a right answer, and the scores on the 317 questions that have one.

The architecture we ran.

Both lanes used the same agents, questions and data. The second lane adds a semantic layer between the agent and the warehouse.





Without AtScale, the agent reads raw tables and writes its own SQL, and each answer is its best guess. With AtScale, the agent asks for a metric by name and the semantic layer computes it the same way every time.

Every model we tested ran in both lanes. Page 12 lists the parts we switched off to stay inside the benchmark rules.

Three takeaways from BIRD-Interact.

We ran seven models from six vendors on all 600 tasks, with and without AtScale. These three results hold on the first, cold run, and the gaps are widest on the 410 business questions, the ones that ask for a number.

1

Every model improved with a semantic layer.

410 BUSINESS QUESTIONS, COLD RUN

+55% to +108%

Best of three: +72% to +137%

ALL 600 TASKS, BEST OF THREE **+38% to +76%**

2

Every model with AtScale beat every model without it.

410 BUSINESS QUESTIONS, COLD RUN

39% vs 36%

Grok 4.3 with AtScale against Gemini 3.1 Pro alone

ALL 600 TASKS, COLD RUN **40.7% vs 40.3%**

Nemotron 3 Ultra with AtScale against Gemini alone

3

Open-weight models rival frontier accuracy with a semantic layer.

410 BUSINESS QUESTIONS, BEST OF THREE

49% vs 51%

Open-weight Kimi K2.5 against frontier GPT-5. Open-weight Nemotron 3 Ultra beat Grok 4.3 on the 410 questions.

ALL 600 TASKS, BEST OF THREE **46% vs 48%**

410 business questions: first-question pass rate. All 600 tasks: success rate. With AtScale: cold first run, semantic memory empty, unless noted as best run of three. Without: one run per model. Seven models from six vendors. Preliminary results from AtScale's October 2026 runs; subject to change.

All 600 tasks, with and without AtScale.

The public leaderboard scores all 600 tasks, and so does this table. With AtScale, Gemini 3.1 Pro scored 57.3%, 19 points above the leaderboard's top entry (38.0%); every model with AtScale clears it, cold run included.

MODEL	WITHOUT ATSCALE	WITH ATSCALE	LIFT
Gemini 3.1 Pro GOOGLE, FRONTIER	40.3%	57.3%	+42%
Sonnet 5 ANTHROPIC, FRONTIER	31.8%	53.8%	+69%
Opus 4.6 ANTHROPIC, FRONTIER	28.2%	49.5%	+76%
GPT-5 OPENAI, FRONTIER	27.7%	47.7%	+72%
Kimi K2.5 MOONSHOT AI, OPEN WEIGHT	28.0%	46.2%	+65%
Nemotron 3 Ultra NVIDIA, OPEN WEIGHT	29.0%	44.5%	+53%
Grok 4.3 xAI, FRONTIER	32.3%	44.5%	+38%

AtScale lifted every model, by 38% to 76%, or 12 to 22 points. The lift is smaller than on the business questions because 190 of the 600 tasks are management operations, such as creating a table or a view. Those never touch a semantic layer and ran straight to the database in both arms.

The weakest models with AtScale, Grok 4.3 and Nemotron 3 Ultra at 44.5%, still beat the strongest model on its own, Gemini 3.1 Pro at 40.3%.

Success rate over all 600 tasks, the leaderboard's own column. AtScale figures are best of three runs; each control is one run. To stay inside the benchmark rules, aggregates were switched off, so every query the warehouse ran touched only base tables, and Semantic Memory was cleared before run 1. Runs 2026-10-03 to 10-07.

410 business questions, with and without AtScale.

The business questions are the part of BIRD-Interact a semantic layer touches. Each model's first answer was scored with and without AtScale.

MODEL	WITHOUT ATSCALE	ATSCALE COLD RUN	ATSCALE BEST OF 3	COLD LIFT
Gemini 3.1 Pro GOOGLE, FRONTIER	36.1%	55.9%	62.2%	+55% BEST +72%
Sonnet 5 ANTHROPIC, FRONTIER	25.9%	50.2%	56.6%	+94% BEST +119%
Opus 4.6 ANTHROPIC, FRONTIER	22.7%	45.4%	53.7%	+100% BEST +137%
GPT-5 OPENAI, FRONTIER	21.5%	44.6%	50.5%	+108% BEST +135%
Kimi K2.5 MOONSHOT AI, OPEN WEIGHT	23.4%	41.7%	49.3%	+78% BEST +110%
Nemotron 3 Ultra NVIDIA, OPEN WEIGHT	23.2%	41.2%	45.9%	+78% BEST +98%
Grok 4.3 xAI, FRONTIER	24.9%	39.0%	44.6%	+57% BEST +79%

The layer added 14 to 24 points on the cold run (20 to 31 best of three). With the layer on, the seven models sit within 18 points of each other, so the layer moves the score further than the choice of model does.

On the cold run, the weakest model with AtScale, Grok 4.3 at 39.0%, beat the strongest model on its own, Gemini 3.1 Pro at 36.1%.

First-question pass rate on the 410 query tasks, public leaderboard protocol. With AtScale: cold run is run 1 with semantic memory empty; best of three is the best of runs 1 to 3; lift is cold-run lift, best-of-three lift below it. Without: one run on the same harness. GPT-5 used GPT-4o as the user simulator; the other six used Claude Haiku 4.5. Runs 2026-10-03 to 10-07.

The follow-up question: every model gains 78%+.

Each BIRD-Interact task has two parts: a first question and, if the agent gets that right, a follow-up.

PHASE 1

The first question

“Give me a ranked list of stars in the galaxy.”

The agent may ask clarifying questions, then submits one answer. The 410-question tables score this pass.

→

PHASE 2

The follow-up

“Now just the ten brightest.”

Asked only if phase 1 was right, and it builds on that answer. The task counts here only when both answers are right.

Getting two answers right in a row is harder, and that’s where AtScale’s lift was largest.

MODEL	WITHOUT ATSCALE	WITH ATSCALE	LIFT
Gemini 3.1 Pro GOOGLE, FRONTIER	10.2%	45.6%	+345%
Sonnet 5 ANTHROPIC, FRONTIER	15.4%	42.7%	+178%
Opus 4.6 ANTHROPIC, FRONTIER	16.1%	40.2%	+150%
GPT-5 OPENAI, FRONTIER	13.4%	36.3%	+171%
Nemotron 3 Ultra NVIDIA, OPEN WEIGHT	13.7%	30.7%	+124%
Kimi K2.5 MOONSHOT AI, OPEN WEIGHT	12.0%	30.5%	+155%
Grok 4.3 xAI, FRONTIER	14.4%	25.6%	+78%

Phase 2 pass rate on the 410 business questions. Follow-up example is illustrative. With AtScale: best run of three. Without: one run per model, so each lift carries the sampling error of a single run. Runs 2026-10-03 to 10-07.

GEMINI 3.1 PRO · THE FOLLOW-UP QUESTION

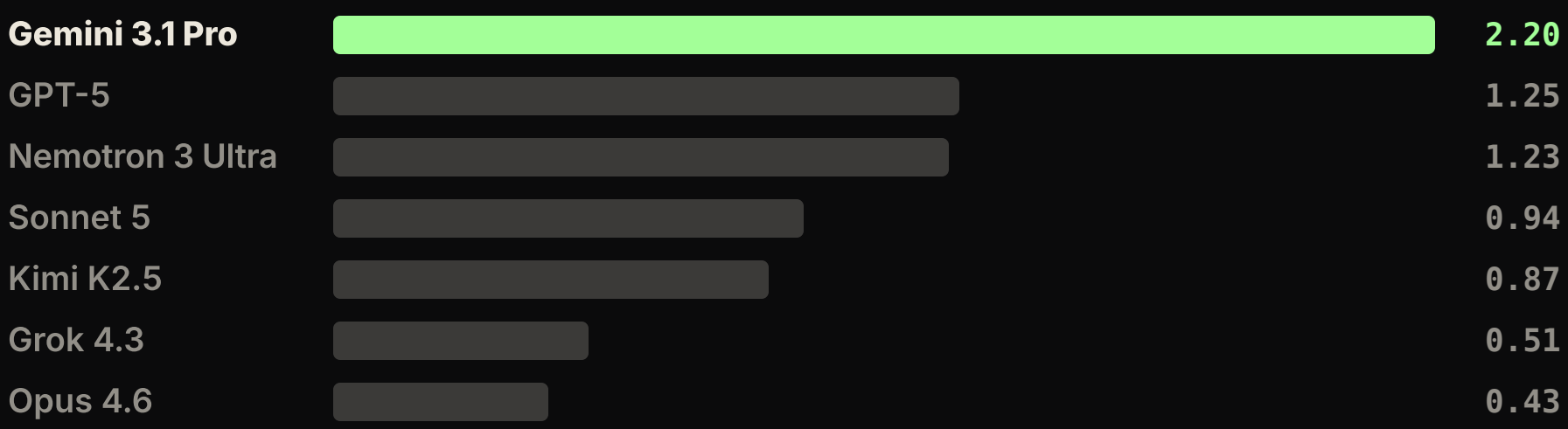
Gemini is the best question-asker.

+345%

LIFT ON THE FOLLOW-UP

Without a semantic layer, Gemini 3.1 Pro finished last on the follow-up. With AtScale, it finished first: 10.2% to 45.6%.

CLARIFYING QUESTIONS ASKED PER TASK, WITH ATSCALE



Each clarifying question and test query spends BIRD COINs from a fixed budget. Without a semantic layer, Gemini spent almost all of it and left 4 tasks unanswered. With AtScale, its questions got answers from defined metrics, and it finished every task with coins to spare.

WITHOUT ATSCALE	
10.2%	
BIRD COINs spent	**17.73 of 17.86**
Tasks left with no answer	**4**

WITH ATSCALE	
45.6%	
BIRD COINs left over	**2.10**
Tasks left with no answer	**0**
Tool calls	**30% fewer**

Gemini 3.1 Pro on the 410 business questions: best AtScale run (run 3) against one run without. Budget 17.86 BIRD COINs per task. Runs 2026-10-03 to 10-07.

ATSCALE

The cost of a right answer fell for every model.

A model writing its own SQL explores the schema, tests queries and retries, and every call costs money. With AtScale, each model got more answers right per dollar.

21%

CHEAPER RIGHT ANSWERS

About 21% on average across seven models, from 7% to 44%.

2.3x

LESS WORK PER ANSWER

BIRD COINs per correct answer, Sonnet 5: 19.9 vs 46.4.

10x

LOWER QUERY COST

Warehouse compute, AtScale TPC-DS benchmark on BigQuery.

COST PER RIGHT ANSWER, SEVEN MODELS

MODEL	WITHOUT ATSCALE	WITH ATSCALE	SAVING
Nemotron 3 Ultra NVIDIA, OPEN WEIGHT	$0.54	$0.30*	−44%
GPT-5 OPENAI, FRONTIER	$0.79	$0.50	−36%
Grok 4.3 xAI, FRONTIER	$0.17	$0.14	−18%
Opus 4.6 ANTHROPIC, FRONTIER	$0.87	$0.73	−16%
Gemini 3.1 Pro GOOGLE, FRONTIER	$0.74	$0.63	−14%
Kimi K2.5 MOONSHOT AI, OPEN WEIGHT	$0.56	$0.49	−13%
Sonnet 5 ANTHROPIC, FRONTIER	$0.56	$0.52	−7%

Spend per run rose for six models, by 6% (GPT-5) to 49% (Sonnet 5). Right answers rose faster.

Cost per right answer: mean API spend per run at list prices, divided by first-question passes over all 600 tasks; three AtScale runs against one run without. *Nemotron 3 Ultra: runs 2 and 3 only; its run-1 cost record is incomplete. BIRD COINs per correct answer: Sonnet 5 on the corrected harness. 10x: AtScale's TPC-DS benchmark on Google BigQuery at on-demand pricing, a warehouse compute figure, not LLM spend.

But 50% accuracy

STINKS.

With AtScale, the seven models averaged 49.1% on all 600 tasks. The best, Gemini 3.1 Pro, reached 57.3%.

SECTION THREE

The hard path to 100% AI accuracy.

On the 317 questions with a right answer, AtScale reached 93%. This section covers the 21 it still missed and what it takes to get to 100%.

On the questions with a right answer, AtScale doubles accuracy.

To see how close a semantic layer can get to 100%, we set aside the business questions that can't be answered correctly and scored the 317 that can.

LLM ALONE

46%

Claude Sonnet 5, writing raw SQL

→ 2X

WITH ATSCALE

93%

Claude Sonnet 5, through the semantic layer

WHY 93 OF THE 410 BUSINESS QUESTIONS WERE SET ASIDE

1 **The answer key contradicts the instructions**
The benchmark's own document defines a business term one way, and its answer key computes it another. Follow the document and you're marked wrong.

2 **The answer key reads a field that isn't there**
It looks up a field the database doesn't contain, so no query can return it.

3 **The answer key's row order is random**
It never says how to order the rows, so the same right answer passes on one run and fails on the next.

Both arms got the same six grading fixes: dates, row order only when asked, letter case, column order, ties and numeric tolerance.

A guard stops the simulated user misstating the answer key's numbers, in both arms. Only the AtScale agent was told not to round.

THE LAST 7%

46% → 93% → 100%

AtScale still got 21 of the 317 wrong. Each one points to work a model can't do on its own: a definition someone has to write down, a question someone has to clarify, a test someone has to keep. The rest of this section works through them.

Claude Sonnet 5. Best runs: 46% without AtScale, 93% (296 of 317) with it. Means of three runs: 45.8% and 90.6%. Unadjusted, leaderboard rules: 53.8% vs 25.9% (3-run mean). Adjusted rates can't be compared with the leaderboard's success rate. Preliminary results from AtScale's October 2026 runs.

21 lines of AI-generated SQL.

Or 3 lines of AtScale.

3 LINES · ATSCALE SEMANTIC LAYER

```
SELECT "Average Order Modification Intensity (OMI)"
FROM   "Insider Trading Surveillance"
WHERE  "Volatile Event Speculator" = 'Yes';
```

No AI-generated SQL.

The 4 subqueries, 5 tables and 6 joins live in the semantic model.

What a semantic layer changes.

Without a semantic layer, the model works out the joins, the grain and the metric math on every question. With one, it calls a certified metric by name and ACE writes the SQL the warehouse runs.

SQL Kimi generated **21** **SQL lines** **6** **joins** **All AI-generated.**

```
WITH event_driven_traders AS (SELECT DISTINCT tr.tr_anchor FROM trade_records tr
  JOIN corporate_events ce ON tr."REC_KEY" = ce."REC_EVT"),
aggressive_event_speculators AS (SELECT edt.tr_anchor FROM event_driven_traders edt
  JOIN traders t ON edt.tr_anchor = t."TR_KEY"
  WHERE t.trader_fin_data->>'risk_lvl' = 'Aggressive'),
volatile_event_speculators AS (SELECT aes.tr_anchor FROM aggressive_event_speculators aes
  JOIN trade_records tr ON aes.tr_anchor = tr.tr_anchor
  JOIN sentiment_analytics sa ON tr."REC_KEY" = sa."REC_SA"
  WHERE ABS(COALESCE((sa.sentiment_data->>'news_score')::real, 0)
          - COALESCE((sa.sentiment_data->>'soc_sent')::real, 0)) > 1.0),
omi_calculation AS (SELECT tr.tr_anchor, (ob.order_metrics->>'mod_freq')::real AS mod_freq,
  (ob.order_metrics->>'cancel_pct')::real AS cancel_pct,
  CASE WHEN ob.order_metrics->>'cancel_pct' IS NULL THEN NULL
       WHEN (ob.order_metrics->>'cancel_pct')::real = 1.0 THEN NULL
       ELSE (ob.order_metrics->>'mod_freq')::real
            / (1.0 - (ob.order_metrics->>'cancel_pct')::real)
  END AS omi FROM volatile_event_speculators ves
  JOIN trade_records tr ON ves.tr_anchor = tr.tr_anchor
  JOIN order_behaviour ob ON tr."REC_KEY" = ob."REC_NODE"
  WHERE ob.order_metrics->>'mod_freq' IS NOT NULL)
SELECT AVG(omi) AS average_omi FROM omi_calculation WHERE omi IS NOT NULL;
```

AtScale: **3** **lines** **Validated, gold SQL.**

```
SELECT "Average Order Modification Intensity (OMI)"
FROM   "Insider Trading Surveillance"
WHERE  "Volatile Event Speculator" = 'Yes';
```

No AI-generated SQL.

The 4 subqueries, 5 tables and 6 joins live in the semantic model.

Kimi K2.5 on a BIRD-Interact insider_trading question, as it wrote the SQL.

The AtScale AI Accuracy Flywheel.

Each step feeds the next, and the last feeds back into the first, so a fixed error stays fixed. CoModel codifies meaning, AVA refuses to guess when a question is ambiguous and captures the resolution as a rule, ACE computes from the certified definition, and Canary attests to the result and turns every caught error into a permanent regression case.

01 Semantic Modeling

ATSCALE COMODEL

`Average_Visit_Satisfaction`

02 Prompt Disambiguation

ATSCALE AVA

"I need urgent care."

03 AI Metric Computation

ATSCALE ACE

Multi-pass · Aggregations · Time series · Semantic memory

04 Accuracy Attestation

ATSCALE CANARY

Measure. Assess. Fix. Remember. Repeat.

EVERY TURN GETS CLOSER TO **100%**

Hard-to-resolve intent: an undefined business term

THE QUESTION · fake_account_9

The question asks for a ranking of accounts by VPN abuse. The benchmark withholds what qualifies as a VPN abuser until somebody asks.

01

Semantic Modeling

CODIFIES WHAT THE BUSINESS MEANS

Holds no certified definition of a VPN abuser, so there is nothing to resolve to.

02 MISS FIX

Prompt Disambiguation

RESOLVES THE QUESTION

"VPN abuser" is vague.

No column or value in the schema defines it, so the LLM invents a threshold. AVA asks the user what qualifies before anything computes.

03

AI Metric Computation

COMPUTES THE ANSWER

Computed the guess faithfully. It has nothing to flag.

04 RESOLVE

AI Accuracy Attestation

GRADES IT AND KEEPS THE FIX

Test "VPN abuser" disambiguation.

The reasoning that resolves VPN abuser is added to the Canary run, to confirm it is disambiguated correctly every time.

Hard-to-resolve intent: an unstated ranking measure

THE QUESTION · planets_data_4

“Give me a ranked list of stars in the galaxy.”

The planets_data database is an astronomy catalog of stars and their planets. The question asks for a ranked list, but doesn’t say how to measure them, and an ordered comparison is graded row by row.

01
Semantic Modeling
CODIFIES WHAT THE BUSINESS MEANS

Every column the question needs is already modeled.

↓

02 FIX
Prompt Disambiguation
RESOLVES THE QUESTION

Ask what to rank the stars by.

The prompt is ambiguous: it asks for a ranking and names no measure. AVA asks the user which measure to rank by before anything is computed.

↓

03 MISS FIX
AI Metric Computation
COMPUTES THE ANSWER

The sort key is unstated.

The database is sorted alphabetically by name. There may be multiple ways to rank the stars: by size, by brightness, by distance from another planet. Each requires a different calculation.

↓

04 RESOLVE
AI Accuracy Attestation
GRADES IT AND KEEPS THE FIX

Continuously test AVA clarification logic.

Every subsequent run rechecks that a ranked request with no stated measure still triggers the clarification.

Semantic data gaps: a metric that lives in a document

THE QUESTION · polar_equipment_15

"How reliable is this equipment?"

The polar_equipment database in BIRD-Interact tracks field equipment. Composite Reliability Index is a formula in the knowledge base.

01 MISS FIX

Semantic Modeling

CODIFIES WHAT THE BUSINESS MEANS

Add reliability as a metric in the semantic layer.

A data engineer carefully specifies how to compute Composite Reliability Index, using the RELIAB_IDX and BSCSI columns.

02

Prompt Disambiguation

RESOLVES THE QUESTION

The question is clear. There is nothing to disambiguate. Tell the user we will be using the CRI metric.

03 FIX

AI Metric Computation

COMPUTES THE ANSWER

ACE computes CRI accurately.

ACE contains the SQL required to get CRI right every time.

04 RESOLVE

AI Accuracy Attestation

GRADES IT AND KEEPS THE FIX

Add CRI to the test suite.

Each subsequent Canary run checks that CRI is always calculated correctly.

Tough-to-compute metrics: a join that fans out

THE QUESTION · mental_health_4 and _7

“What share of each doctor’s patients take their medication as prescribed?”

The mental_health database in BIRD-Interact records patient visits, and the treatments given at each visit. The answer is one rate per doctor.

01 MISS FIX

Semantic Modeling

CODIFIES WHAT THE BUSINESS MEANS

Joining the two tables counts a visit twice.

The rate needs both tables, so they are joined. A visit with three treatments then appears three times and is counted three times. 123 of the 351 visits carry more than one treatment, and 42 of the 157 doctors came out wrong. Counting at visit level keeps each visit worth one.

02

Prompt Disambiguation

RESOLVES THE QUESTION

The question is clear. There is nothing to disambiguate.

03 FIX

AI Metric Computation

COMPUTES THE ANSWER

ACE counts each visit once.

ACE computes the rate at visit grain from the model’s definition, so a visit with three treatments still counts once.

04 RESOLVE

AI Accuracy Attestation

GRADES IT AND KEEPS THE FIX

Continuously test the visit count.

Each Canary run checks that a visit still counts once, whatever it carries.

Dirty or unnormalized data: one entity, many spellings

THE QUESTION · households

“How many households are in each region?”

The households database in BIRD-Interact stores a region name on every household, and the same region arrives spelled several ways. “Chicago”, “CHICAGO” and “chicago” with a stray tab on the end are three stored values for one real place, and a GROUP BY counts them as three regions.

01 MISS FIX

Semantic Modeling

CODIFIES WHAT THE BUSINESS MEANS

Reconcile spellings.

The semantic layer maps every variant of a region name, the case differences, the trailing tabs and the alternate spellings, onto one region. Every query that groups or filters on region then sees one region instead of three.

02

Prompt Disambiguation

RESOLVES THE QUESTION

The question is clear. There is nothing to disambiguate.

03

AI Metric Computation

COMPUTES THE ANSWER

Groups exactly what it is given, and returns three regions where there is one.

04 RESOLVE

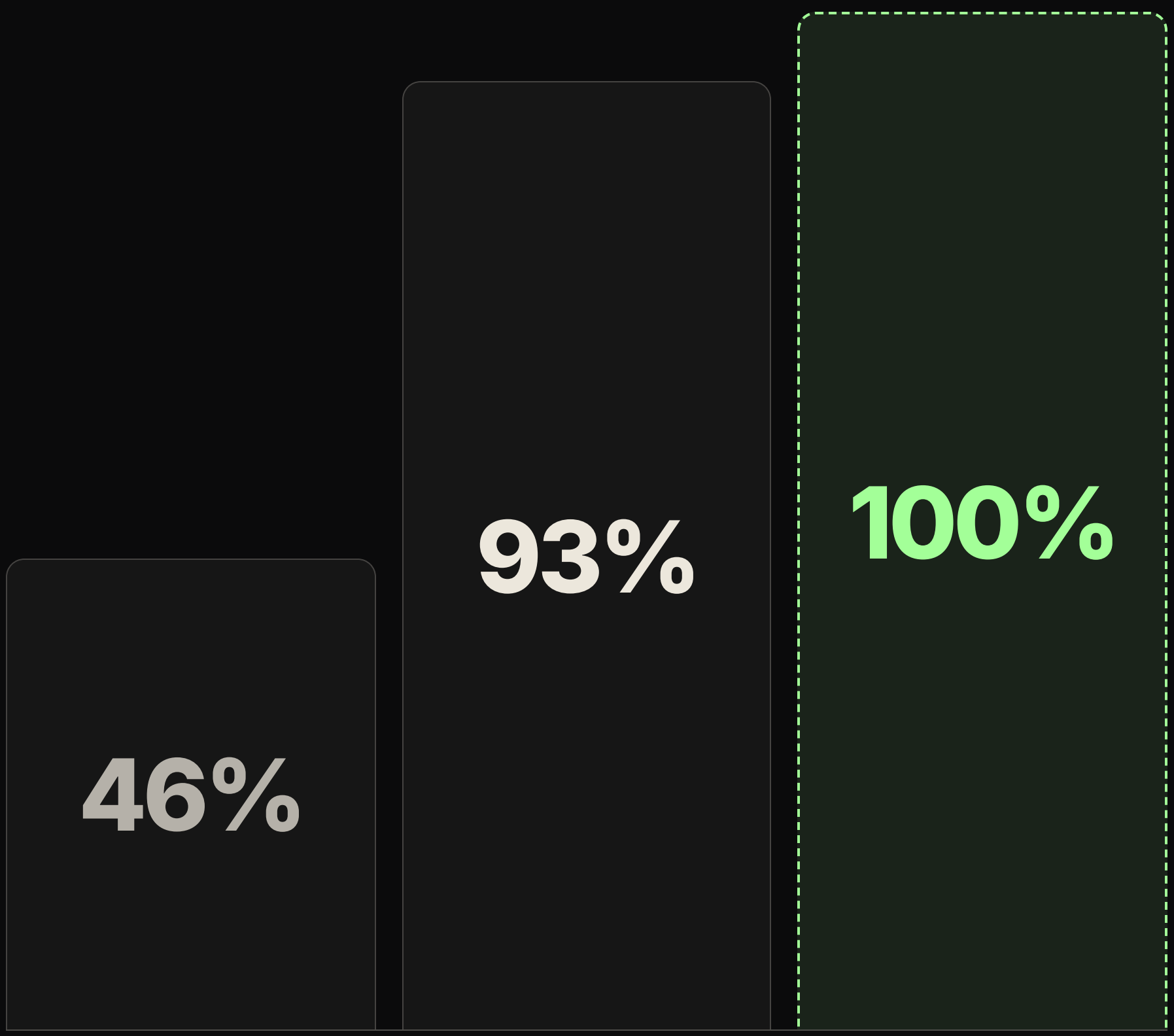
AI Accuracy Attestation

GRADES IT AND KEEPS THE FIX

Continuously test reconciliation.

Each Canary run checks the spellings in the data and the regions they map to, so a new spelling mistake is reconciled next time.

The path to the pursuit of 100% AI accuracy.



LLMs on their own, solvable questions

Sonnet 5 with no semantic layer, on the 317 questions that have a right answer to find.

AtScale auto-generated semantic layer

(generated by a CoModel preview, then tuned; see page 30)

Adding AtScale CoModel, AVA prompt disambiguation and ACE lifts accuracy to 93%.

The AI accuracy flywheel with analyst and data engineering assistance

The target. The five cases in this section show the fix for each kind of miss among the 21.

Join us in the pursuit of AI that never lies.

AtScale for AI.

Ava and ACE are available today, and agents reach ACE through the AtScale MCP server. CoModel and Canary arrive this fall.

1. **Modeling your semantics**
   atscale.com/comodel — ATSCALE COMODEL
2. **Disambiguating the prompt**
   atscale.com/ava — ATSCALE AVA
3. **Computing the metric**
   atscale.com/ace — ATSCALE ACE
4. **Measuring and attesting accuracy**
   atscale.com/canary — ATSCALE CANARY

Every number here can be checked.

All of it comes from runs on the public BIRD-Interact harness in stress mode, so anyone can reproduce it. We publish the data, the code, the per-run detail and our full results.

The 22 semantic models were generated by a CoModel preview from the same context the raw arm saw, then tuned on this benchmark; a September 23 audit removed every element traceable to the answer key, and all reported runs use the cleaned models. Results are preliminary and not yet verified by the BIRD team.

ATSCALE'S FULL FINDINGS

atscale.com/bird

Per-run detail, the database-level breakdown, the control runs, the semantic models we generated, and the list of all 93 excluded questions with the reason for each.

THE BENCHMARK ITSELF

github.com/bird-bench/BIRD-Interact

The harness, the 600 tasks, the 22 databases, and the scoring code.

THE LIVE LEADERBOARD

bird-interact.github.io

Every verified submission, updated as new entries are scored.

THE PAPER

arxiv.org/abs/2510.05318

How the BIRD team built the tasks and why they scored them this way.

FOR AGENTIC WORKFLOWS

AtScale for Agentic AI

Why agents that act on numbers need certified ones.

ATSCALE