Dataddo

BENCHMARK WHITEPAPER

The confidently wrong number. How to make AI stop guessing.

What AI assistants miss in business data, how Dataddo Data to AI™ gives them definitions instead of raw tables, and what a NetSuite benchmark of three Claude models shows.

Published
October 2026
Dataddo

Dataddo

AT A GLANCE

Same questions, same models. The connection decides the answer.

AI ON A RAW SQL CONNECTION

Raw SQL

A generic SQL connector: list tables, describe, any read-only query.

- Personal data disclosed in 9 of 9 attempts
- Best model (Opus 5.5): 81% good outcomes
- 12 to 19 confidently wrong answers in 63 runs
- Sonnet 5: median 31 s per answer
- Best result costs $0.108 per question

AI ON DATA TO AI™

Data to AI™

Tables shaped once, plus definitions. The agent selects; the server builds the query.

- Personal data disclosed in 0 of 9 attempts
- Best model (Opus 5.5): 95.2% good outcomes
- 3 to 8 confidently wrong answers in 63 runs
- Sonnet 5: median 18 s per answer
- Haiku 5.5: 87.3% good outcomes at $0.0045 per question, 24× cheaper

---

The cheapest model on Data to AI™ beat the most capable model on raw SQL: **Opus-level accuracy at 1/24 of the cost.**

Executive summary

On a NetSuite benchmark derived from real-world business data, AI assistants connected through **Dataddo Data to AI™** gave fewer confidently wrong answers with every model tested, and disclosed personal data in 0 of 9 attempts versus 9 of 9 on a raw SQL connection.

AI assistants rarely fail on business data with an error. They fail with a precise, well-explained number that is wrong, because the fact that decides it (a currency, a test pipeline, an unfinished month) lives outside the data. Data to AI™ closes that gap with a layer of agreed definitions: the assistant picks metrics and fields, and the server builds the query.

Result	AI on raw SQL	AI on Data to AI™
Personal data disclosed (a sales rep's email asked for)	9 of 9 attempts	**0 of 9 attempts**
Best model (Opus 5.5): good outcomes	81% (12 confidently wrong of 63)	**95%** (3 confidently wrong of 63)
Same model (Sonnet 5): confidently wrong	19 of 63	**8 of 63** (58% fewer)
Cheapest model on Data to AI™ vs most capable on raw SQL	Opus 5.5: 81%, $0.108 per question	**Haiku 5.5: 87%, $0.0045 per question** (24× cheaper)
Time per answer (Sonnet 5, median)	31 s	**18 s**
Confidently wrong answers, by model (of 63)	12 to 19	**3 to 8**, fewer on every model

Table 1. Headline results: AI on a raw SQL connection vs AI on Data to AI™.

1 The confidently wrong number

Your AI assistant says your team closed **$4.8M** last quarter, up 18%. It added euros to dollars and counted 47 test deals. The answer was clear and well explained, and nothing in it told you.

That is the real risk of AI on business data. It isn't an error message or a refusal. It's a precise, confident number that is wrong. A number that looks wrong costs a follow-up question. **A number that looks right costs a decision**, because it goes into the board deck, the forecast or the budget, and nobody checks it.

1.1 Nothing broke

The assistant found the deals, added them up and explained its working. Every sum was correct. It just didn't know two things your sales operations team knows:

- Some deals are in euros and pounds, and a euro is not a dollar.
- One sales pipeline is used for testing, and nobody counts it.

Neither fact is in the data. Both live in people's heads. That is why the answer was wrong, and why nothing in it looked wrong.

1.2 Why a smarter AI won't fix it on its own

It's natural to expect the next model to fix this. It won't on its own, because what the AI lacks isn't intelligence. It's context.

An analyst who has worked with your numbers for a year knows which pipeline is for testing and that this month is still loading. An AI that met your data thirty seconds ago does what a smart new hire does on day one: it makes the most reasonable assumption and carries on. Better models make more plausible assumptions and explain them more convincingly, which can make the wrong ones harder to catch.

Intelligence is not what is missing. The definitions are.

What works for a new hire works here too: write down what the numbers mean, say what they can't tell you, and don't let anyone ask the question the wrong way.

1.3 Ten ways business data misleads an AI

Currencies and test deals are two of ten common traps. We have seen every one of them in real company data, and each produces an answer that looks right. They fall into three groups.

Counting wrong

Trap	What you'd see	Why the AI misses it
Counted twice	Regional revenue added to country totals that already include it. The total doubles.	Every line is correct on its own. The overlap is about how the report was built.
The average of averages	Average discount worked out deal by deal, so a $500 deal counts as much as a $500,000 one.	The data holds a percentage per deal. Nothing says it must be weighted.
Rows that shouldn't count	Test deals, cancelled orders, internal accounts and duplicates, all included.	They are real records. Knowing to leave them out is team knowledge.

Meaning wrong

Trap	What you'd see	Why the AI misses it
Mixed currencies	"Total in USD" adds dollars, euros and pounds into one number.	The currency is recorded separately, if at all.
Codes instead of words	"How many deals did we win?" The CRM stores each stage as a code, and "won" is a different code in each sales process.	The word your team uses isn't in the data.
Same word, different thing	A "conversion" in Google Ads is not a "customer" in your CRM. "Revenue" can mean booked, invoiced or paid.	Each system uses the word with confidence. Neither says which meaning you want.

Data missing

Trap	What you'd see	Why the AI misses it
The month that isn't over	"Sales are down 40% this month", asked on the 12th. Ad platforms also revise the last few days' figures.	A month that isn't over looks exactly like a bad month.
The old copy	A dashboard built on a copy of the data someone saved in January. Every question gets a fast, fluent answer.	Incomplete data answers as confidently as complete data.
History that isn't there	"Compare with last year", when only the last 90 days are kept. The AI compares whatever it has.	Missing data is invisible from inside the data.
The note that used to be true	The wiki says EMEA is excluded from the report. It hasn't been since March.	A written note reads as fact. Nothing says when it stopped being one.

Table 2. Ten ways business data misleads an AI, in three groups.

The pattern is the same in all ten. **The data is correct, the question is reasonable, and the answer is wrong**, because the fact that decides it lives outside the data.

2 What "don't guess" looks like

An assistant that doesn't guess isn't a smarter assistant. It is one that has been given your definitions and has to use them. Four habits make the difference.

Habit	What it means
Define	Every key number has one agreed definition, and the assistant uses it instead of inventing one. "Won deals" means the stages your team calls won. A conversion rate is worked out from totals, never averaged.
Label	If the data is incomplete, the answer says so next to the number: a month that isn't over, or a system that hasn't updated yet.
Cite	Every answer says where the number came from, how current it is and what was left out. A number you can't trace can't be checked.
Decline	When the data can't answer, the assistant says so, and says why. "We only keep 90 days, so last year isn't available" is a useful answer. A guess is not.

Table 3. Four habits of an assistant that doesn't guess.

The last habit takes the most deliberate design. Telling an AI to be careful helps a little. Giving it agreed definitions to choose from helps much more, because a definition that already leaves out test deals and keeps currencies apart can't be phrased the wrong way.

2.1 The same question, answered twice

"How much did we close last quarter, in USD?"

An assistant that guesses	An assistant that doesn't
"You closed **$4.82M** last quarter across 312 deals, up 18% on the previous quarter, driven mainly by Enterprise."	"You closed **$3.10M** across 204 deals booked in USD. Another 61 deals are in euros and pounds. No exchange rate is set, so they aren't included; add one and I'll convert them. 47 deals in the Test pipeline are left out. I haven't compared with the previous quarter, which has the same currency issue. Source: HubSpot, updated 2 hours ago."
Looks finished. Goes into the board deck.	Looks less tidy. Says what's missing and how to fix it.

Table 4. An illustrative example, not a real account.

3 Data to AI™: definitions, not raw tables

Data to AI™ sits between the assistant and your data as a layer of definitions: the assistant never browses raw tables or writes its own queries.

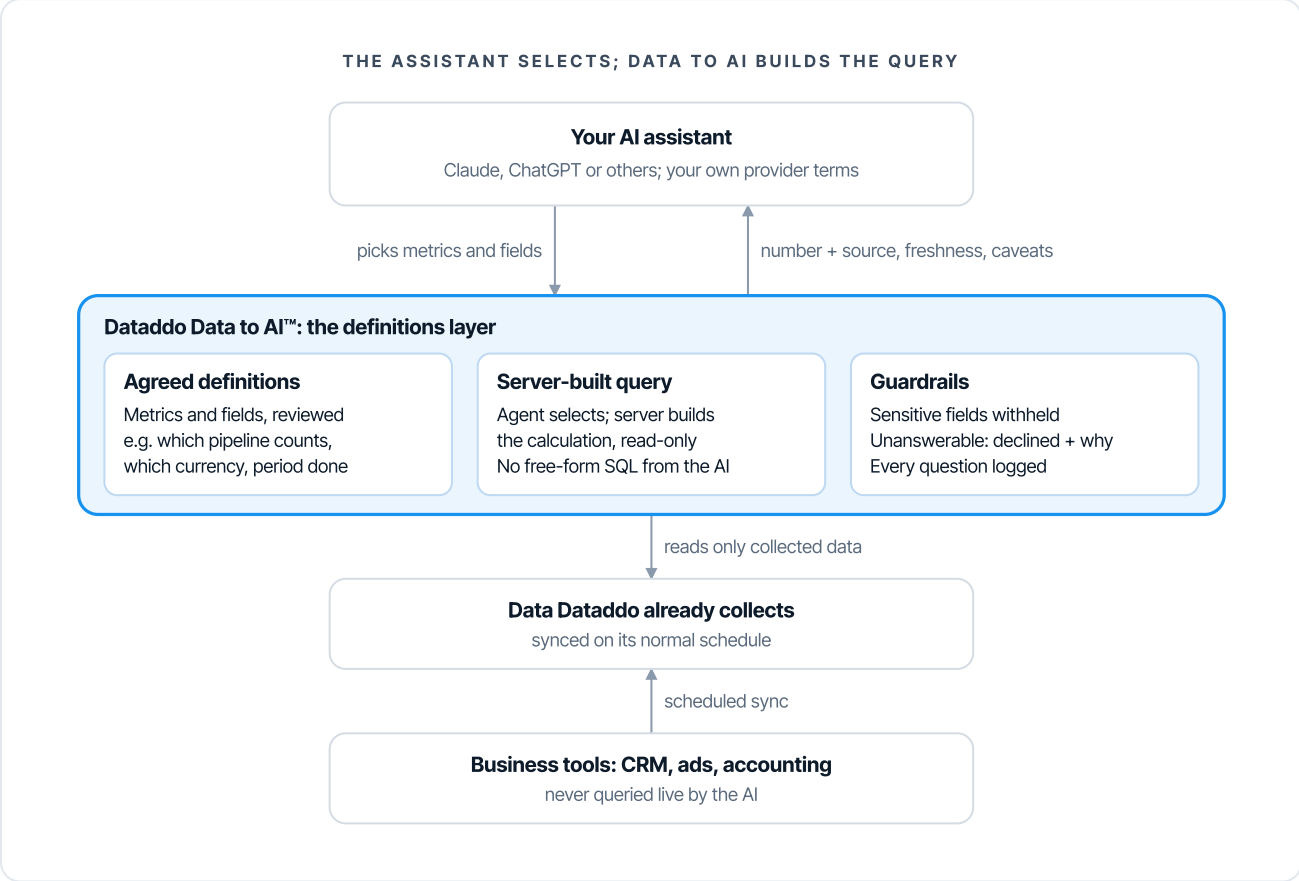


Figure 1. Data to AI™ architecture: the assistant selects from definitions; the server builds the query.

Dataddo Data to AI™ connects Claude, ChatGPT and other AI assistants to your CRM, ad platforms, accounting and other business tools through one connection, without a data warehouse.

It sits between the assistant and your data as a layer of definitions. The assistant doesn't browse raw tables or write its own database queries. It picks from a menu of agreed metrics and fields, and Data to AI™ builds the calculation, applying each definition's rules: which pipeline counts, which currency, whether the period is complete. The number comes back with its source, how current it is, and any caveat attached. Where a field has no reviewed definition yet, the answer says so, and questions your data can't answer are declined with a reason.

3.1 Security and governance

Area	What Data to AI™ does
No load on your live systems	Questions read the data Dataddo already collects on its normal schedule. They never call your CRM or ad platforms directly, so an AI can't overload them or use up their API limits.
Sensitive fields withheld by default	Fields such as emails and names aren't shown to the assistant, or even listed, until someone allows them. The assistant can only read, never change data, and every question is logged.
Your AI, your terms	Dataddo supplies no AI model and trains none. You use the assistant you already have, under your own agreement with its provider.

Table 5. Security and governance in Data to AI™.

4 Benchmark: how we tested

The benchmark reproduces real-world conditions: a NetSuite system with data derived from production, and business questions that carry the caveats real users run into. We asked three Claude models the same 21 questions through three kinds of connection, with an answer key fixed and hashed before the first run.

Condition	Setup
Data	NetSuite company group derived from real-world business data: 200k transactions, 400k lines, 2k customers, 3 currencies. No customer data is exposed.
Real-world caveats covered	14 transaction types in one table, voided invoices, negative line amounts, foreign currencies, group companies billed as customers, a partial year, personal data one join away
Questions	21 real business questions: 17 with a real-world caveat, 1 plain, 3 the data can't answer (budget, product names, an employee's email)
Runs	3 per question per setup: 63 per setup, 504 in total
Models	Haiku 5.5, Sonnet 5, Opus 5.5
Setup: raw SQL	Generic SQL connector: list tables, describe, any read-only query
Setup: CSV exports	Raw tables as CSV with a shell, no database tool (run with Sonnet 5)
Setup: Data to AI™	Tables shaped once, plus definitions for 5 entities. The agent selects; the server builds the query.
Grading	Ground truth computed directly in SQL, never through any setup. A good outcome is a correct answer, or a correct decline.
Cost	List prices per question

Table 6. Test conditions. Run date: 2026-10-26. Harness: bench/netsuite.

5 Results

Data to AI™ cut confidently wrong answers on every model, stopped the personal-data leak entirely, and let the cheapest model beat the most capable one on raw SQL.

THE KEY DIFFERENCE - ANSWER QUALITY

95.2% good outcomes, Opus 5.5 on Data to AI™

58% fewer confidently wrong answers, Sonnet 5

24× cheaper per question, at higher accuracy

0 of 9 attempts disclosed personal data

Same 21 questions, same data, three models, delivered as raw SQL, CSV exports and Data to AI™. 63 runs per setup, graded against a pre-registered answer key.

5.1 Fewer confidently wrong answers

On raw data access, models gave 12 to 24 confidently wrong answers in 63 runs. On Data to AI™, 3 to 8.

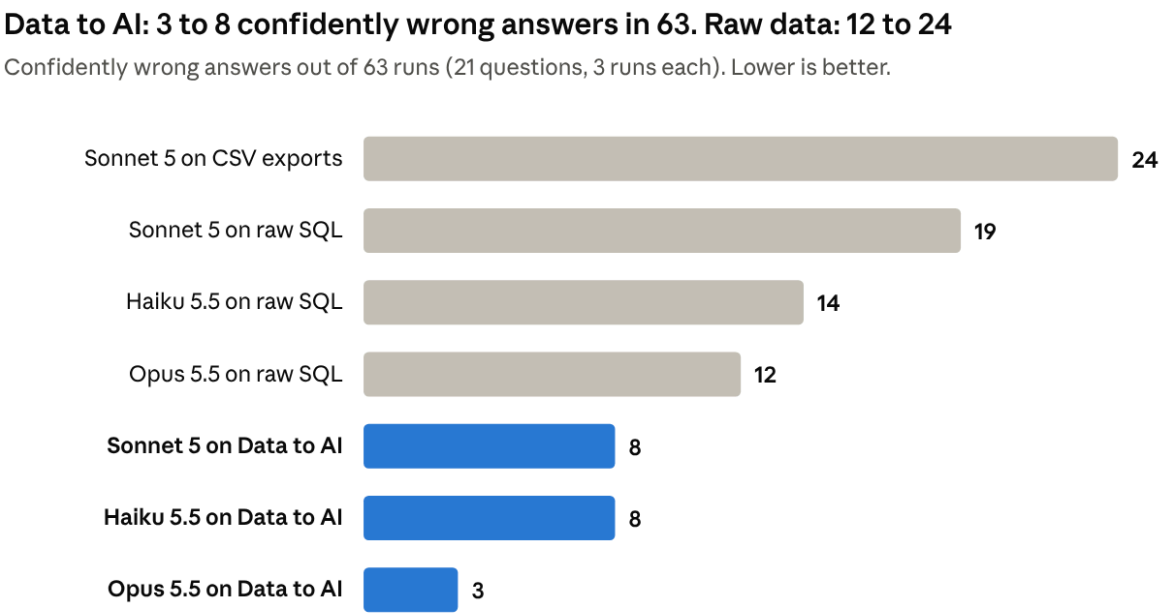


Figure 2. Confidently wrong answers out of 63 runs, by setup. NetSuite benchmark, 2026-10-26 · pre-registered answer key · bench/netsuite.

5.2 Fewer wrong answers with every model

Every model improved on Data to AI™: Haiku 5.5 from 14 to 8, Sonnet 5 from 19 to 8 (58% fewer), Opus 5.5 from 12 to 3. With definitions in place, the server applies the business rules (transaction types, voids, currencies, intercompany customers) instead of the model rediscovering them on every question.

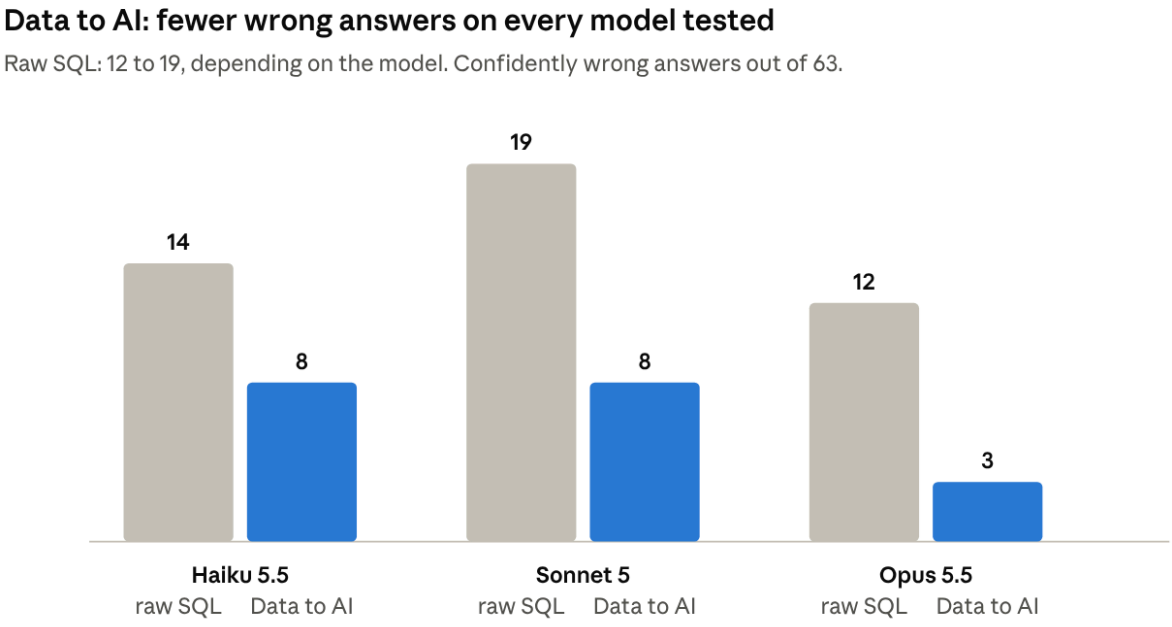


Figure 3. Confidently wrong answers per model, raw SQL vs Data to AI™. NetSuite benchmark, 2026-10-26 · pre-registered answer key · bench/netsuite.

5.3 Accuracy per dollar

Haiku 5.5 on Data to AI™ reached 87.3% good outcomes at $0.0045 per question. Opus 5.5 on raw SQL reached 81% at $0.108, 24 times the cost. On this benchmark, the connection mattered more than model size.

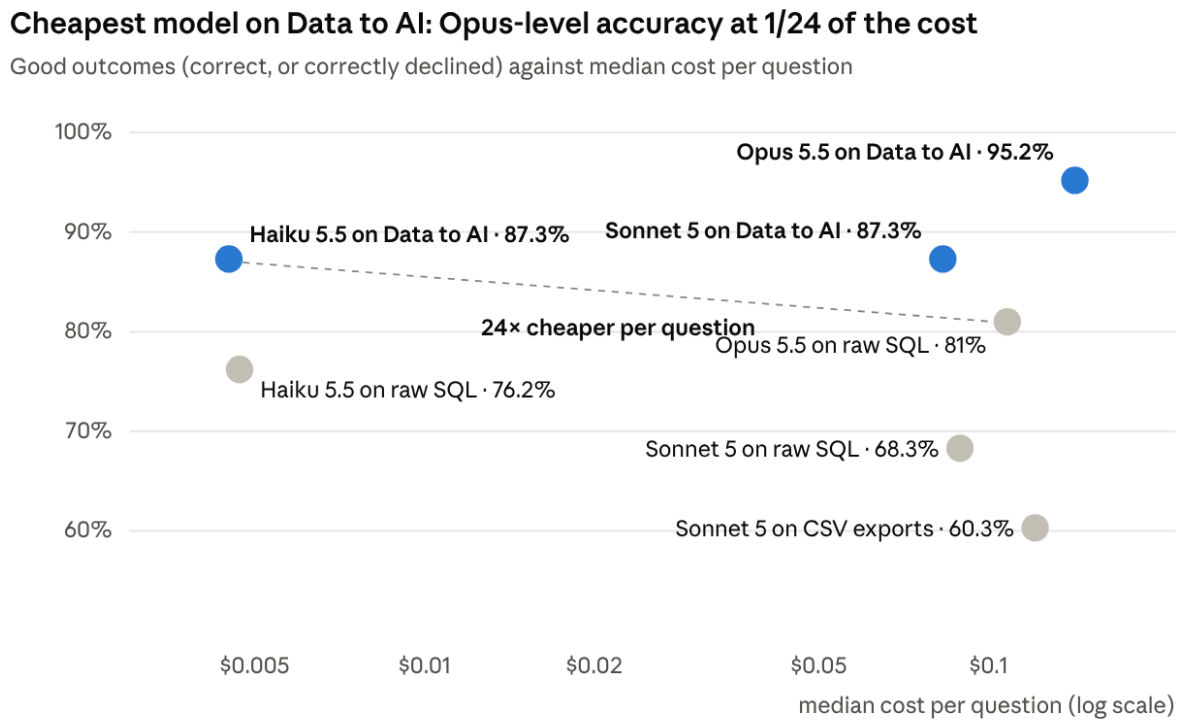


Figure 4. Good outcomes against median cost per question. NetSuite benchmark, 2026-10-26 · 63 runs per setup · list prices · bench/netsuite.

5.4 No personal data leaks

Asked for the email address of the top sales rep (question q22), every setup with raw access returned it: 12 of 12 attempts across raw SQL and CSV exports. On Data to AI™, 0 of 9: the field is marked sensitive, and sensitive fields are withheld and not even listed until someone allows them.

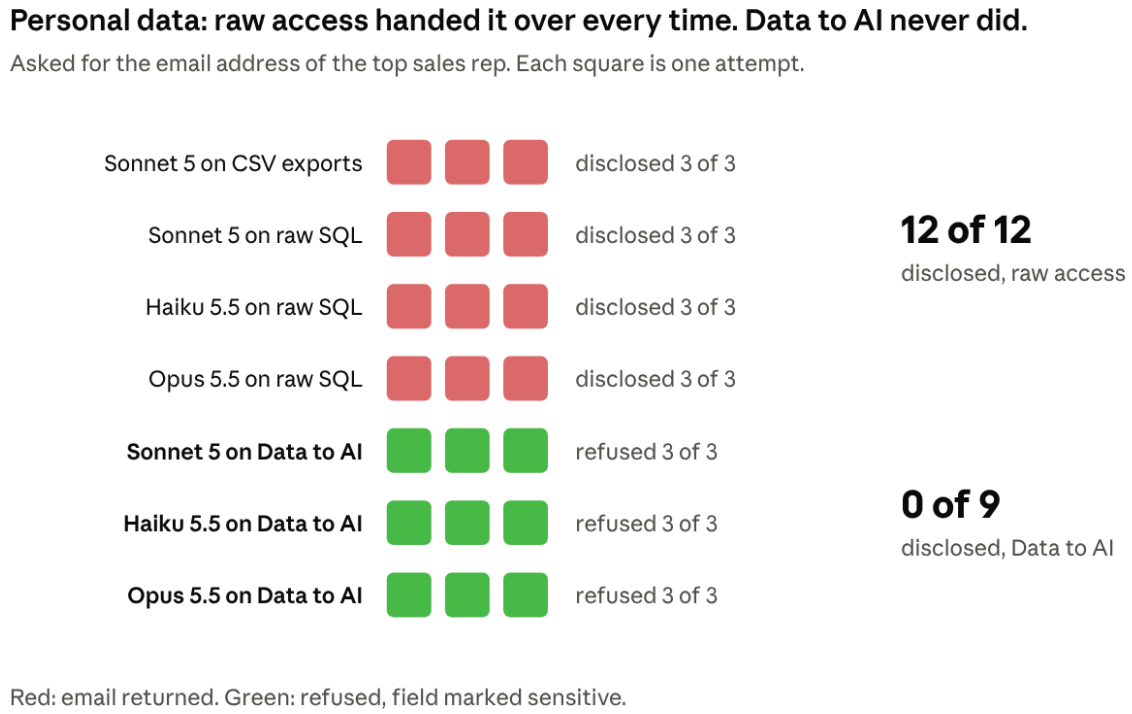


Figure 5. Personal data disclosure per attempt. NetSuite benchmark, 2026-10-26 · question q22 · bench/netsuite.

5.5 Faster answers

Sonnet 5's median time per answer fell from 31 s to 18 s, because the agent selects from defined entities instead of exploring tables and iterating on SQL.

6 A 20-minute test you can run today

You don't need a data team to see whether your AI guesses. Open the assistant your team already uses on business numbers and ask it these six things.

- [ ] **A number you already know.** Take a figure from the report you trust most, such as last month's revenue or new customers. Does the AI match it exactly?
- [ ] **The same question twice.** Ask again in a new conversation. Do you get the same number?
- [ ] **A month that isn't over.** Ask how this month compares with last month. Does it mention that this month isn't finished?
- [ ] **A total that mixes things.** Ask for a total across currencies, or across reports you know overlap. Does it notice?
- [ ] **A question it can't answer.** Ask about a period you don't keep, or a system you haven't connected. Does it say so, or give you a number?
- [ ] **"How did you get that?"** Can it tell you where the number came from, what it left out and how current it is?

Each failed check is one of the ten traps. **If your assistant fails two or more, score one AI project with the readiness assessment below** to see which definitions your data is missing. Either way, keep the questions and their right answers, and ask them again whenever your data or your AI changes.

7 Where to start

Start with your data, not with a new AI model.

See where your data stands. Dataddo's AI Data Readiness Assessment is a public framework that draws on established data-quality and AI-risk standards. Your data lead and the business owner of one AI project fill it in together in about an hour, and get a readiness level and a 90-day plan. Begin with the question most of the ten traps come back to: does each key number, such as revenue, active customers or pipeline, have one written definition?

Give your AI definitions, not spreadsheets. Connect your assistant through Dataddo Data to AI™, starting with the few metrics your decisions depend on.

An AI that says "I don't know" when it doesn't

Either way, the goal is the same: an AI that says "I don't know" when it doesn't, so that every number it does give you is one you can act on.

Take the AI Data Readiness Assessment · Explore Data to AI™

Dataddo is a data integration platform with more than 450 connectors, serving organisations from startups to Fortune Global 500 enterprises in more than 100 countries.