Trinetro Labs™

FREE EBOOK · AI AND TRUST

Numbers You Can Defend

A practical guide to using AI for finance and reporting, with seven tests and a 30-minute bake-off to tell an answer you can sign off from one you can only hope is right.

7	30	1	12
tests for any AI answer	minute bake-off	printable scorecard	questions for vendors

TRINETRO LABS · BANGALORE

OCTOBER 2026

WELCOME

Read this first

Anyone can now ask a business question in plain English and get a number back in seconds. The hard part has moved. It is no longer getting an answer; it is deciding whether you can put your name under it.

This guide is for the people who do put their name under numbers: CFOs, finance controllers, heads of data, and the founders who play both roles. It does not ask you to pick a side in the AI debate. It gives you a way to test any tool, ours included, against the standard your auditors already hold you to.

~18 min
to read it all. Chapters 3 and 4 are the ones to keep open.

30 min
to run the bake-off in Chapter 4 on your own data.

0 setup
to use the tests. All you need is a question you already know the answer to.

1 **The sign-off problem** 3
Why "usually right" is the wrong standard for money

2 **Two ways a machine answers a question** 4
Generating an answer versus computing one, and the one question that separates them

3 **The seven tests** 5
Repeat, Trace, Define, Scope, Reconcile, Doubt, Recover

4 **The 30-minute bake-off** 9
A fair comparison of any two tools, with a scorecard to print

5 **Nine places Indian finance data goes quietly wrong** 11
Financial year, GST, credit notes, lakh and crore, and more

6 **Getting your team ready** 12
Definitions, owners, access, and a runbook for when two numbers disagree

7 **Twelve questions for any vendor** 13
Plus where Trinetro fits, in one honest page

01

CHAPTER 1

The sign-off problem

Every finance leader knows this Monday. A number appears in a pack, someone asks where it came from, and the room goes quiet while a person is sent to find out.

The cost of that silence is not the hour it takes to answer. It is the next thought in everyone's head: if this one cannot be explained, what about the others?

AI assistants change the shape of the problem. They make the silence **shorter**, because anyone can ask a question and get a figure in seconds. They also make it **riskier**, because a fluent answer looks finished. A wrong number in a clear sentence is harder to catch than a wrong number in a messy spreadsheet.

> A number is ready for sign-off when you could hand it to an auditor and explain, without improvising, where it came from and why it would come out the same tomorrow.

That is the standard this guide uses, and it is deliberately not "is the AI usually right?". Usually right is a fine standard for drafting an email. It is a poor one for a figure that moves money, sets a bonus or goes into a filing.

What sign-off asks of an answer

- **It can be repeated.** Ask again tomorrow, or ask a colleague, and the figure is the same.
- **It can be explained.** Someone who was not in the room can follow the steps from source to total.
- **It can be defended.** The definitions used (what counts as revenue, which date, which entity) were agreed in advance, not guessed.
- **It can be limited.** It only used data the asker was allowed to see.

ILLUSTRATIVE EXAMPLE

You ask a tool, "What was revenue last quarter?" On Monday it answers **₹4.82 crore**. On Wednesday, asked in slightly different words, it answers **₹4.61 crore**. The gap is ₹21 lakh. It turns out one run netted off credit notes and the other did not.

This is an invented example to show the pattern, not a report of a real incident. Neither answer was "wrong" in the tool's eyes. The problem is that nobody chose between them on purpose.

02

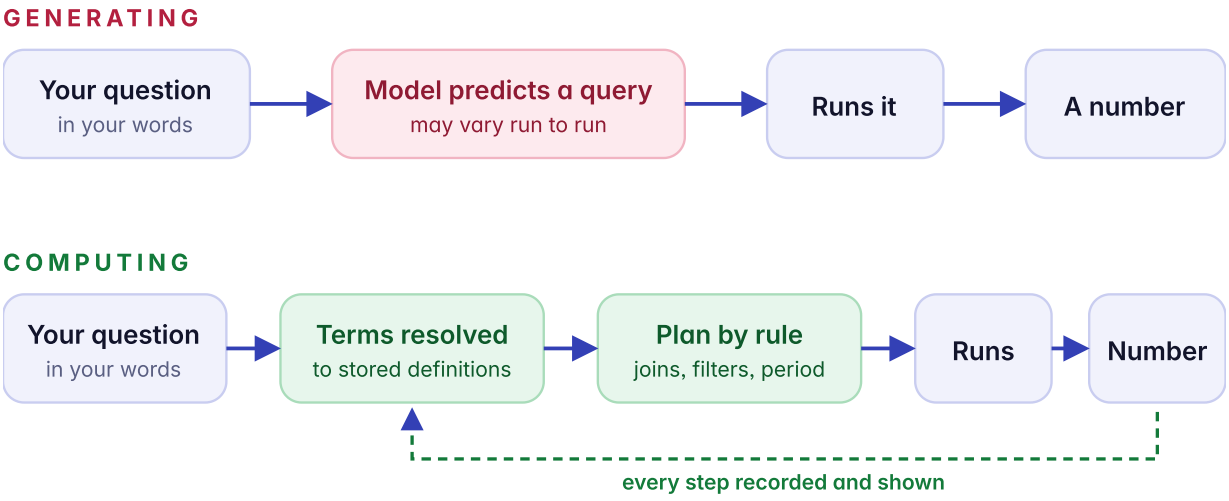
CHAPTER 2

Two ways a machine answers a question

Almost every "chat with your data" tool does one of two things between your question and the number. Knowing which one is the single most useful thing you can learn about it.

Generating means a language model reads your question and writes a plausible query or answer from patterns it has seen. It is flexible, quick to start and often right. But it is a prediction, so it can differ from one run to the next, and it cannot show you why it chose what it chose.

Computing means the system maps each term in your question to a defined field or metric, plans the joins by rule, runs the query and records every step. The same question on the same data gives the same answer, and the steps are the explanation.



The one question that separates them

> Which part of the system decides the number: a model's prediction, or a rule you can read?

Many good tools combine both. A language model may read the question, and a rules layer checks and runs it. That is a sensible design. What matters is the answer to the question above, because it tells you where an error could come from and whether you could find it.

03

CHAPTER 3

The seven tests

Each test takes about ten minutes and needs nothing more than a question you already know the answer to. Run them in order on any tool. Score each 0, 1 or 2 on the scorecard in Chapter 4.

1 REPEAT · 2 TRACE · 3 DEFINE · 4 SCOPE · 5 RECONCILE · 6 DOUBT · 7 RECOVER

1 Repeat

Do I get the same number every time?

If a figure changes when nothing underneath it has, you cannot tell which version is right, and neither can your auditor. It is the cheapest test, and the one that most often ends a trial early.

RUN IT IN TEN MINUTES

Ask one question with a known answer three times in a row. Then ask it once in different words. Ask a colleague to ask it too, from their own login.

PASS LOOKS LIKE

Identical figures every time. When the wording changes, the tool shows it understood the question the same way.

RED FLAG

The figures differ and nobody can say why, or the tool says "results may vary".

2 Trace

Can I see how it got there?

An explanation you cannot inspect is a claim. A trace is a record of the tables, filters, joins and definitions used, so a second person can verify the answer without trusting the first.

RUN IT IN TEN MINUTES

Take any answer and ask to see how it was produced. Give the trace to a colleague who was not present and ask them to explain it back to you.

PASS LOOKS LIKE

Plain steps from source to total, naming the data, the filters and the definitions used. It can be saved or exported with the answer.

RED FLAG

"The AI determined this", a confident paragraph with no steps, or a query you cannot map to your own data.

3 Define

Did it use a definition I chose?

Most disputed numbers are not arithmetic errors. They are two people meaning different things by the same word. A tool that quietly picks a meaning for "revenue" or "active customer" is making a finance policy decision on your behalf.

RUN IT IN TEN MINUTES

Ask for revenue, then ask what it counted. Ask the same for "active customer" and "outstanding receivables". Find out where the definition lives and who can change it.

PASS LOOKS LIKE

One stored definition per term, visible with the answer. It says what is in and out (credit notes, GST, which date) and who owns it.

RED FLAG

The tool picks silently, gives two different meanings on two days, or lets anyone redefine a term without a record.

4 Scope

Does it only use what the asker may see?

A tool that answers every question for everyone is a data leak with a friendly face. Access has to be enforced where the data is read, not hidden in the screen. The safest starting position is that a new person sees nothing until someone grants it.

RUN IT IN TEN MINUTES

Create a test user with limited access. Ask for something outside their scope, by name and then indirectly ("what did we pay the top supplier?"). Check what a brand-new member can see before anyone grants anything.

PASS LOOKS LIKE

A new member sees nothing by default. Limits are set per data source, down to named tables or sheets, and apply to every answer.

RED FLAG

Everyone sees everything until restricted, or the restriction is only a hidden menu that a clever question walks around.

WHY THIS ORDERING

Tests 1 and 2 tell you whether the tool is *capable* of an auditable answer. Tests 3 and 4 tell you whether it is *governed*. A tool can pass the first pair and fail the second, which is why a good demo alone never settles it.

5 Reconcile

Does it tie to the books?

Your ledger has already been checked by people with something at stake. If a tool cannot reproduce totals you already trust, it is not ready to produce totals you do not yet know.

RUN IT IN TEN MINUTES

Pick five closed figures, such as last month's sales, receivables at month end or payroll cost, from your accounting system. Ask the tool for each. Compare to the rupee, then ask it to explain any gap.

PASS LOOKS LIKE

Totals tie exactly, or the difference is explained by a stated rule such as a timing cut-off or a rounding line.

RED FLAG

"Close enough", a difference nobody can locate, or a tool that has never been compared to a ledger.

6 Doubt

Does it ask when the question is unclear?

Real questions are ambiguous. "Revenue by region" could mean billing region, shipping region or sales territory. A tool that always answers confidently is guessing some of the time, and you will not know which.

RUN IT IN TEN MINUTES

Ask three deliberately ambiguous questions: a term with two meanings, a period with no year, and a measure the data does not contain. Note what it does each time.

PASS LOOKS LIKE

It asks one precise question, or states its assumption plainly and lets you change it. It says so when the data cannot answer.

RED FLAG

A fluent, confident answer to every question, including the ones that cannot be answered.

> An answer that says "I need to know which you mean" is worth more than three that guess.

7 Recover

Can I see what changed, and go back?

Definitions change, and they should. A policy decision in March must not silently rewrite the numbers you reported in February. Recovery means every change has an author and a date, and old answers can be reproduced.

RUN IT IN TEN MINUTES

Change a definition on a copy of your setup. Ask what changed and who changed it. Then ask for last quarter's figure the way it was reported at the time.

PASS LOOKS LIKE

Versioned definitions with a visible history. An answer can be traced to the version that produced it.

RED FLAG

Edits overwrite the past, or there is no record of who changed a definition.

The seven at a glance

TEST	THE QUESTION	WHAT YOU WILL SEE IF IT PASSES
1 Repeat	Same number every time?	Identical across runs, wordings and users
2 Trace	Can I see how?	Plain, exportable steps from source to total
3 Define	Whose definition?	One stored, owned, visible meaning per term
4 Scope	Who may see it?	Default deny, limits enforced on every answer
5 Reconcile	Does it tie out?	Matches the ledger, or explains the gap
6 Doubt	Does it ask?	Clarifies or states assumptions; admits limits
7 Recover	Can I go back?	Versioned history; old answers reproducible

HOW TO SCORE

0 means it fails or cannot be tested. **1** means it passes with manual effort or only in some cases. **2** means it passes cleanly and you could show someone else.

Set your bar before you start. A sensible one: **no zeros on tests 1, 2, 4 or 5** for any tool that will touch money, and a total of at least 11 out of 14.

04

CHAPTER 4

The 30-minute bake-off

A demo is a performance. A bake-off is a measurement. This one runs on your questions, against your numbers, and treats every tool the same way.

1. **Write ten questions (10 min)**
   Four simple totals, three with a filter or period, two ambiguous ones, and one permission trap. Use questions from your last month-end pack.

2. **Write the expected answers first**
   From your ledger or a trusted report, before you open any tool. This is what makes the comparison fair and keeps the vendor from setting the bar.

3. **Run every question twice, by two people (10 min)**
   Use the same wording the first time, then different wording. Record each figure on the scorecard.

4. **Score the seven tests (5 min)**
   Use the 0, 1, 2 scale from Chapter 3. Put a note against anything you would not want to explain to an auditor.

5. **Sort the failures (5 min)**
   For each miss, decide whether it is a *setup* issue (a definition not yet entered, a table not yet connected) or a *fundamental* one (the figure changes, or there is no trace). Setup issues can be fixed. Fundamental ones are your answer.

CHOOSING THE TEN QUESTIONS

- **Four simple totals:** sales for a closed month, receivables at month end, expenses for a quarter, count of invoices.
- **Three with a twist:** a financial-year comparison, a split by branch or product, a figure net of returns.
- **Two ambiguous:** a term with two meanings, and a period with no year.
- **One permission trap:** a question a limited user should not be able to answer.

KEEP IT FAIR

Give each tool the same data, the same definitions and the same time. If you spend an hour tuning one, say so on the scorecard. A comparison that favours the tool you already like is not a comparison.

The scorecard is on the next page. Print it, or copy it into a spreadsheet.

SCORECARD

Bake-off scorecard

Tool A: _____ Tool B: _____ Date: _____ Run by: _____

Part 1. The seven tests (score 0, 1 or 2)

TEST	TOOL A	TOOL B	NOTES
1 Repeat			
2 Trace			
3 Define			
4 Scope			
5 Reconcile			
6 Doubt			
7 Recover			
Total (of 14)			

Part 2. The ten questions

#	QUESTION	EXPECTED	A: RUN 1	A: RUN 2	B: RUN 1	B: RUN 2
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						

Decision rule: no zero on tests 1, 2, 4 or 5, and a total of at least 11. Fundamental failures end the trial; setup issues get a date for a fix.

05

CHAPTER 5

Nine places Indian finance data goes quietly wrong

None of these is exotic. Each one produces a number that is internally consistent and still not the number you meant. Build them into your ten questions.

THE TRAP	WHAT GOES WRONG	HOW TO CHECK
Financial year versus calendar year	The Indian financial year runs April to March, so Q3 of FY 2026-27 is October to December 2026. A tool that assumes January to December puts "Q3" in July to September.	Ask for "Q3 this year" and check the dates it states.
GST inclusive or exclusive	GST collected is a pass-through, not revenue. Sales that include it overstate income by the tax.	Ask for sales, then ask whether tax is in the figure.
Credit notes and returns	Ignored, they inflate sales. Netted twice, they deflate them. This is the gap in the example in Chapter 1.	Pick a month with known returns and tie it out.
Lakh and crore	1 crore is 10 million and 1 lakh is 100,000. A figure shown "in lakhs" read as rupees is off by five zeros.	Always read the unit label. Ask for the full figure once.
Several GSTINs or entities	Branches and group companies registered separately are summed, or left out, without saying so.	Ask which entities are in the total.
Which date counts	Invoice date, voucher date, due date and payment date give different months for the same transaction.	Ask which date a "monthly" figure uses.
Intercompany transfers	Movements between your own entities counted as sales and as purchases.	Ask for a total with and without internal transfers.
Rounding	Summing rounded lines gives a different total from rounding the sum. Accounting systems often carry a rounding-off ledger.	Reconcile a total that has many small lines.
Foreign currency	Export invoices converted at the invoice-date rate, the month-end rate or the settlement rate give three different rupee figures.	Ask which rate and date it used.

Accounting treatments depend on your policies and on current rules. Use this chapter as a checklist of questions, and confirm treatment with your accountant.

06

CHAPTER 6

Getting your team ready

Whatever tool you choose, the human side decides whether the numbers hold. Four habits do most of the work.

1. Write down ten definitions

For each term, settle the questions on the right and name one owner.

TERM	DECIDE BEFORE ANYONE ASKS
Revenue	Before or after GST? Credit notes netted? Recognised on invoice or on delivery?
Gross margin	Which costs sit in cost of sales? Freight in or out?
Active customer	Bought in the last 90 days, or has an open contract? Which entities count?
New customer	First invoice, first payment, or first signed order?
Churn	Cancelled, lapsed or downgraded? Measured monthly or annually?
Receivables	Gross or net of provisions? Aged from invoice date or due date?
Working capital	Which balance-sheet lines? Is cash included?
Run-rate	Latest month times twelve, or a trailing average?
Period	Financial year, calendar year, or rolling twelve months?

2. Start from "nobody can see anything"

Grant access per person and per data source, down to the tables they need. See **trinetrolabs.com/blogs/data-access-control-checklist**.

3. Keep a rhythm

Weekly: re-run your ten questions and compare. **Monthly:** review definition changes. **Quarterly:** review who has access to what.

4. WHEN TWO NUMBERS DISAGREE

Freeze both figures and who asked. **Trace** each one to its steps. **Compare** the definitions, period, entity and date basis used. **Decide** which is right for this purpose, record why, and update the definition so it cannot recur.

07

CHAPTER 7

Twelve questions for any vendor

Put these in your next review. A good vendor answers each in a sentence or two, and is glad to be asked.

THE ANSWER

01 Which part of the system decides the number: a model or a rule?
See Chapter 2.

02 Will the same question on the same data always return the same answer?
Ask them to show it, three times.

03 Can I see every step behind an answer and export it?

THE DEFINITIONS

04 Where are definitions stored, who can change them, and is every change versioned?

05 What happens when a term in my question is ambiguous?

THE PEOPLE

06 What can a newly added member see before anyone grants access, and where is that enforced?

07 Is a second sign-in factor compulsory for every account, or optional?

THE DATA

08 How are my stored credentials and files encrypted, and are keys shared across customers?

09 Is my data ever used to train or test a model?

THE RELATIONSHIP

10 Where can I read your past incident reports and status history?

11 If I leave, how do I export my data, and when is it deleted?

12 Can I run a bake-off on my own data before I sign anything?

More on trinetrolabs.com/blogs: **how-to-read-a-cloud-incident-report** (question 10) and **data-access-control-checklist** (questions 6 and 7).

WHERE TRINETRO FITS

One honest page from the publisher

We wrote this guide to be useful whichever tool you choose, and we built Trinetro to pass these tests. Here is how, in plain terms. Please run the bake-off on us too, on your own data.

Repeat, Trace	A deterministic engine with no language model in the answer path. The same question on the same data returns the same answer, with a reasoning trace that shows the joins, the metrics and a confidence score.
Define	Governed team memory: personal definitions, plus company-wide defaults that an admin promotes and can revoke at any time.
Scope	Default-deny access. A new member sees nothing until an admin grants it, down to named tables, enforced on the server.
Doubt	When a term is genuinely ambiguous, it asks one precise question instead of guessing.
Sign-in, keys	An authenticator app for every customer account, and a separate encryption key for each company.

Where we are not the right fit, we will say so. A tool that suits a data team writing SQL all day may suit them better than it suits a finance team that wants answers.

Run the bake-off on your own data

Connect a database or spreadsheet, ask in plain English and see exactly how every number was produced. The Cat plan is free for 30 days, with no credit card.

trinetrolabs.com/get-started **trinetrolabs.com/try**

Repeat · Trace · Define · Scope · Reconcile · Doubt · Recover

Trinetro Labs, Bangalore, India · contact@trinetrolabs.in · trinetrolabs.com
This guide is general information, not accounting, tax or legal advice. Examples marked illustrative are invented. Tax and accounting rules change; confirm current treatment with your advisers. © 2026 Trinetro Labs. You may share this PDF unchanged and in full.