When Your AI Says 70 Percent, Is It Right 70 Percent of the Time?

Choosing the right shape of AI for business forecasts, from a live, public benchmark on the 2026 World Cup

inocta.io · The Intelligence of Operations.

Executive summary

Businesses are starting to let AI put a number on the future: the probability a deal closes, a customer churns, an invoice gets paid on time. The moment that number drives a decision, one question decides whether it helps or hurts: **when the AI says 70 percent, is it right 70 percent of the time?**

We built a live, public benchmark to answer it. Three different ways of setting up an AI forecaster called every match of the 2026 World Cup, in real time, with the rules locked before the first game. We measured not who guessed the most winners, but whose stated confidence was **honest**. We call that the **trust score**.

Three findings an operator can act on:

1. **A council of AI models earned the best trust score of any setup we tested.** When a confidence number drives a decision that matters (a credit line, a capacity commitment, guidance to a board), the council is the setup to reach for.
2. **More expensive did not mean more trustworthy.** The most elaborate setups cost twenty times the cheapest and bought no better trust score. Spend and trustworthiness are two different things, and most organizations pay for orchestration they never validated.
3. **The finding matters more inside your company than it did on the pitch.** Football has a betting market that already prices every game. Your Q3 pipeline has no such market. So a trustworthy AI forecast is not competing with a reference. It **is** the reference you would otherwise lack.

If you read nothing else: before you trust an AI's confidence in a real decision, measure its trust score. On this evidence, a council is the shape to reach for, and elaboration alone is not worth paying for.

---

1. The problem: an un-audited confidence number is a liability

Take a weighted sales pipeline. Deal size times close-probability, summed. Standard revenue arithmetic. But it is only arithmetic if the probabilities are honest. A book of deals marked "70 percent" must actually close about 70 percent of the time, or the weighted total is fiction with decimal places.

When an AI starts supplying those probabilities, that honesty question becomes a finance question. An operator who adopts AI-generated confidence without measuring whether it is honest has not automated forecasting. They have automated the production of unaudited numbers, and every downstream decision inherits the error.

The natural objection: *we already check our forecasts against last quarter's actuals*. True, and where you have enough closed history, run that backtest. But it cannot give you what a trust score can. A trust score is available **before** the real results come in, when a new segment or a young product has too few closed deals to backtest. And it is a **per-forecast** signal: your actuals tell you last quarter's total missed, while a trust score tells you **which** confidence levels are the ones that lie.

---

2. The three AI setups, in plain terms

We tested three ways to arrange an AI forecaster. A manager choosing between them needs to know what each one **is** and when to reach for it. The technical build details are in the scientific paper. Here is what each does and what it costs.

Single call

One AI model gets the information and returns a forecast with a confidence number. The simplest thing you can deploy. We ran it two ways: with a full pack of context (stats, news,

history) and with almost none. **Reach for it** when the decision is high-volume and low-stakes and a rough probability is enough. It is the cheapest option by a wide margin.

Pipeline

A chain of specialists. One AI step reads the numbers, a second reads the narrative and context, a third writes the final forecast. On the hard cases where its own confidence is low, it escalates to a heavier method. More moving parts, more cost. **Reach for it** when you want a staged process with a built-in escalation on the difficult forecasts, and you are willing to pay for the orchestration.

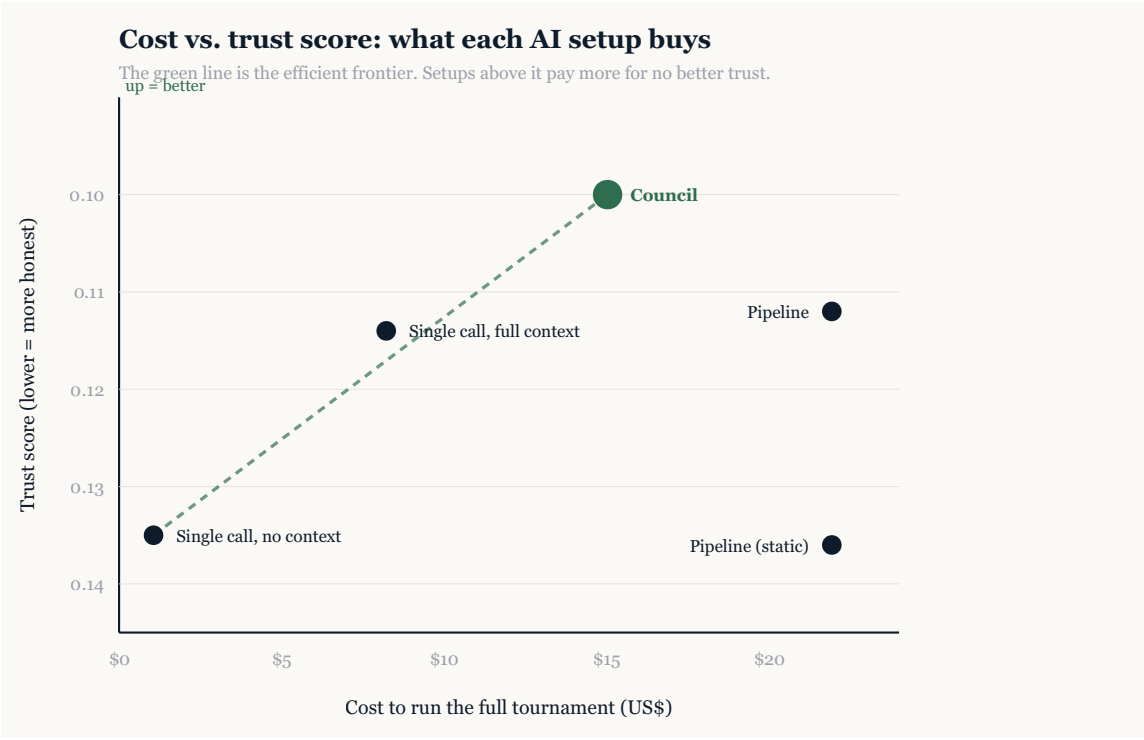
Council

Three independent AI models (from different families) each produce a forecast in parallel, and a final step weighs their agreement and disagreement into one answer. When the three agree, confidence is high. When they diverge, confidence drops to reflect the genuine uncertainty. That disagreement-aware behaviour is what earned it the best trust score. **Reach for it** when the decision is calibration-critical, when a confidence number drives an expensive or irreversible commitment.

---

3. What the benchmark found

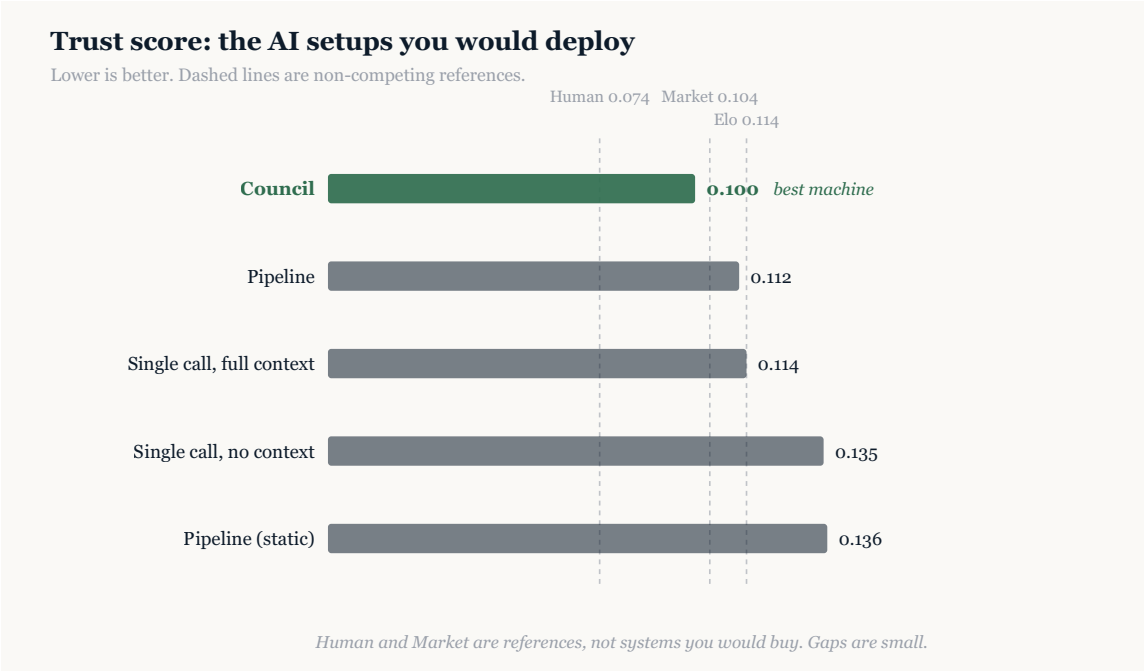
Across all 104 matches, each setup earned a trust score (lower is more honest) at a known cost. The cost figures below are the total spend to forecast all 104 matches, start to finish.



Cost vs. trust score frontier

The chart above is the buy decision in one image. Two setups sit on the efficient frontier (the green line): the cheap single-call at one end (lowest cost) and the council at the other (best trust score). The pipelines sit off the frontier, up and to the right, meaning they cost far

more without buying a better trust score. They are the two most expensive systems in the study, and neither earns its premium.



Ranked head to head, the council posts the best trust score of any machine. We report the margins honestly, two ways. First, when we re-ran the ranking thousands of times on resampled matches, the council held the top or near-top spot far more often than not, but its lead over the next-best setup was not wide enough to call a certainty on a single 104-match tournament. Second, and this is the harder test: on the final stretch of matches we held back and scored blind, the council did not keep its lead. On that thin, held-back set a cheaper setup edged ahead. So read this as a directional answer, not a photo finish: across the full tournament the council is the setup to reach for, and a second tournament would sharpen the margin, but we will not tell you the clean held-back test crowned it, because it did not. A directional best, disclosed as directional, is still the answer an operator asked for, and it is a more honest one than a false certainty.

The AI setup	Trust score (lower better)	Cost to run	Hit rate
Single call, no context	0.135	$1.06	68%
Single call, full context	0.114	$8.22	66%
Council	**0.100 (best)**	$15.03	67%
Pipeline	0.112	$21.93	67%
Pipeline (static)	0.136	$21.93	67%

Hit rate is how often each setup picked the winner. Notice it barely moves across setups (66 to 68 percent): on a luck-heavy event, nobody guesses much better than anyone else, which is exactly why accuracy is the wrong thing to buy on and the trust score is the right one.

The technical name for the trust score is calibration. The scientific paper reports it as calibration error. Same measure, so a reader of either paper can map the two.

---

4. What it means for your business

Spend is not trust. The single clearest lesson: the most elaborate, most expensive setups were no more honest than far cheaper ones. Before you commit an integration budget to a rich AI orchestration, demand the trust-score difference it buys, measured, not assumed. On this task, it was zero.

The football result understates the business case. The AI setups landed close to the betting market because football **has** a deep public market pricing every match, a hard reference that leaves little room to separate. Your company has no such market. There is no closing line for your pipeline, your demand plan, your churn. So the trustworthy AI forecast is not squeezed against a reference. It is the reference you would otherwise lack, and that makes measuring its trust score more valuable inside a company, not less.

Where this applies, and where it does not. The finding transfers cleanly to a decision shaped like a football match: one you make on a schedule, with a committed deadline, and a real result you can score soon after. A weekly pipeline call, a monthly demand plan, a quarterly churn forecast. It transfers less to a one-off judgment with no clean outcome and no deadline. So point the trust score at your recurring, scored decisions first. That is where a confidence number is a promise you can hold the AI to.

---

5. Which setup for which decision

The tournament turns the setups into a priced menu. Match the setup to the decision:

	Low stakes (a rough number is fine)	**High stakes (a wrong confidence is costly)**
High volume (many forecasts)	**Single call, no context.** Cheapest, good enough.	**Council on the critical subset.** Reserve it for the decisions that can afford it.
Low volume (few, deliberate forecasts)	**Single call.** Simple, low cost.	**Council.** The best trust score, worth the premium.

The rule underneath the grid: reach for the council when a confidence number drives a commitment you cannot easily walk back. Use the cheap single call everywhere else, and put the money you save toward measuring the trust score itself.

---

6. Put it to work

Six questions to ask any AI forecasting vendor

Copy-paste these. A team that cannot answer them has not measured what it is selling.

1. “Show me the reliability curve for your system at 30-plus scored events. Not accuracy. The curve.”
2. “When your system says 70 percent, how often is it actually right? Give me that number, over how many events, on results you held back and did not train on.”
3. “What confidence threshold routes a forecast to a human, and where is that logged?”
4. “What did the context integration measurably add? Show me the no-context comparison.”
5. “Can you replay any past forecast exactly, with the inputs it saw at lock time? What prevents after-the-fact revision?”
6. “What does one forecast cost, and what does each point of trust-score improvement cost over our current process?”

How the trust score is computed, in plain terms

Group your forecasts by the confidence you assigned. Take every forecast you marked around 70 percent and check how many actually came true. If 70 percent of them did, that band is honest. If only half did, your 70 is really a 50, and the gap is the lie. Do that across every confidence band, average the gaps, and you have one number: how far your stated confidence sits from reality, on average. Lower is better. A perfectly honest forecaster scores zero.

That is the whole idea. The exact formula, and a copy-paste script that runs it on your own forecast log, are in the public OSF repository below, so you never have to build the calculation yourself.

Run it yourself in one quarter

Lock the outcome definition and the commit deadline in writing before the first forecast. Record every prediction with its confidence before the deadline, no revision.

What to run. Do not start with the council. Start with what you have plus a cheap floor: your current process, one simple historical average, and the cheapest single-call AI, all forecasting the same events in parallel. That tells you whether AI beats your status quo at all, and whether the expensive data integration earns its keep, before you spend on orchestration. Reserve the council for the critical subset of decisions once the cheap comparison has shown you the AI is worth trusting.

The one hard limit. A trust score needs at least 30 scored events to mean anything. A single quarter of a mid-market pipeline may close fewer than that. If it does, do not force a verdict on thin data: pool similar decisions together, or run two quarters, before you rank anything.

After enough scored events, a company of ordinary volume owns its first trust score. Which is to say, it knows whether its AI's 70 means 70.

The evidence behind this brief

Every claim here traces to a fully public, pre-registered study. Nothing is asserted that the data does not carry.

- **The full scientific paper, every prediction, every cost, the complete dataset:** OSF, DOI 10.17605/OSF.IO/R5SBJ (osf.io/r5sbj)
- **The live public dashboard:** match.inocta.io, where every prediction was published before kickoff
- **The locked pre-registration:** osf.io/ke35t, frozen before the first scored match

You can't automate what you don't understand, and a confidence number you have not audited is not something you understand yet.

inocta.io · The Intelligence of Operations.

inocta.io · The Intelligence of Operations.