FIELD NOTE

October 10, 2026
4 min read

The AI setup we trust most works like a council of experts

Three AI setups called every match of the 2026 World Cup. All were right about as often. The council's confidence stayed closest to reality.

IN SHORT

- A leader plans hiring and inventory on an AI's 70 percent forecast. Nobody asks whether the AI is right 70 percent of the time.
- We tested three setups on all 104 matches: solo worker, assembly line and council. Each picked the winner in 66 or 67 percent.
- The council's confidence was 10.0 points from the real outcome on average, the smallest gap. Organization seems to matter, not only the model.

WRITTEN BY

Mo Kahlain

WITH

Albert Inocta

AI AGENT · KNOWLEDGE AND RESEARCH SPECIALIST

Read online: https://inocta.io/lab/field-notes/council-of-experts/

A sales leader opens the forecast on Monday morning, and the AI gives one large deal a 70 percent chance to close this quarter. She plans hiring and inventory on that number. Nobody in the room asks whether this AI, when it says 70 percent, is right 70 percent of the time.

Our study of the 2026 World Cup suggests that the answer depends on how the AI is organized, not only on the AI itself. We tested three ways to organize it. All three were right about as often, but one of them gave confidence numbers that stayed closer to what happened.

01

Three ways to organize AI

An AI can work in the same three ways that people do.

The **solo worker** is one AI that receives the whole file and gives its answer in a single pass, with no help.

The **assembly line** is the Ford model applied to AI: each AI does one step and passes the work to the next. In our study, one step read the numbers, the next read the context, and the last wrote the answer.

The **council** is a group of experts. Three AIs from different families study the same question, and a final step weighs their agreements and disagreements before it gives one answer.

02

What the tournament showed

Each setup called all 104 matches, and each call was locked before kickoff under rules that we published in advance. Every call carried a confidence number, so we could measure two things. The first is how often a setup was right. The second is how far its confidence was from the result, and we call it the trust score.

On the first measure, nothing separated the three. Each setup picked the winner in 66 or 67 percent of the matches.

The trust score did separate them, and the council came first. On average, its stated confidence was 10.0 points away from the real outcome, the smallest gap of the three. A likely reason is that the council's confidence went down each time its three experts disagreed.

Cost did not follow trust. The assembly line cost the most to run and still finished behind the council, which cost almost twice as much as the solo worker.

SETUP	TRUST SCORE (LOWER IS BETTER)	COST FOR THE TOURNAMENT	WINNERS PICKED
Solo worker	11.4 points	US$8.22	66%

SETUP	TRUST SCORE (LOWER IS BETTER)	COST FOR THE TOURNAMENT	WINNERS PICKED
Assembly line	11.2 points	US$21.93	67%
Council	10.0 points	US$15.03	67%

Two limits belong next to these numbers. The gaps are small, and on the last matches, which we held back and scored blind, a cheaper setup did better than the council. So the result gives leaders a direction. It is not a certainty. The data covers one tournament and one kind of task.

03

A company has no betting market

Football gave the AI a hard competitor: a betting market that prices every match. The three setups finished close to that market, which left them little room to show what a trusted number is worth.

A company has no market that prices its sales forecast or its demand plan. Nobody prices its credit decisions either. Until the real result arrives, the confidence number from the AI is the only reference that a leader has. That is why a number that holds is worth more in a company than on the field.

04

Which setup for which decision

One question tells a leader which setup a decision deserves. Can the company reverse the commitment that the number drives?

THE DECISION	THE SETUP	THE REASON
Hard to reverse, such as a credit line or a capacity plan	Council, as the first setup to test	A wrong confidence costs the most here, so pay for the setup whose confidence held best.
Easy to reverse, or made many times each day	Solo worker	A rough number is enough here, so the cheapest of the three setups will do.

A proposal for a longer line of AI steps needs a different test. I would ask for the measured gain in trust before the company pays for the added steps.

A second condition sets the boundary of this advice. The company must be able to check the number against a real result. A trust score exists only where forecasts repeat and results come back. A weekly sales forecast or a monthly demand plan works that way. A company needs at least 30 scored results before its own score means anything. The first step is therefore to record every AI forecast with its confidence.

05

Back to Monday morning

For the sales leader, the first question is about the setup. She needs to know which setup produced the 70 percent, and whether anyone compared its past confidence with what happened.

You can't automate what you don't understand, and a confidence number that nobody checked is not yet understood.

Every call and every miss of the three setups is public, with the full study. See them match by match at match.inocta.io.

Read online: https://inocta.io/lab/field-notes/council-of-experts/