

PUBLISHED: NOVEMBER 25TH · BY SHREYAS KALYANPAD

A confidence score is the number a product prints next to an AI answer to say how sure it is.

The problem: on most screens it does nothing. The answer looks the same at 62 percent as it does at 98 percent. You can copy it, send it or approve it either way. That is not transparency. It is a badge.

> *If the number cannot stop an action, it is decoration with a decimal point.*

Five ways the number fails



The needle moves. Nothing downstream of it does.

1. The score gates nothing

What teams ship: A percentage beside the output, and every button stays enabled.

Why it fails: In a study of 184 people solving logic puzzles, well calibrated confidence lifted decision accuracy by 20 percent. Miscalibrated confidence lifted it by 2 percent. It also made people accept wrong answers that were stated confidently.

Ship instead: Pick a threshold. Below it, Apply is disabled and the job routes to a person.

2. Nobody checked the number

What teams ship: Whatever the model says when you ask it how sure it is.

Why it fails: Instruction tuned models are badly calibrated. They also have an ownership bias. They rate their own answers up to 26 percent higher than the same answer typed in by a user.

Ship instead: Measure the score against real outcomes before you show it. If 90 percent is not right nine times in ten, do not print 90 percent.

3. Honesty with no action attached

What teams ship: A note saying the score may not be reliable.

Why it fails: Researchers tried this. Telling people the calibration level did help them spot a bad score, but it cut their trust so far that they rejected correct answers too. Decision quality did not improve.

Ship instead: Do not move the judgement onto the user. Change what the interface allows.

4. Fake precision

What teams ship: 98.7 percent, a gauge, or four coloured bars.

Why it fails: That decimal is usually a raw score, a rank or a similarity value wearing the costume of a probability. It reads as certainty the system cannot back.

Ship instead: Hide the raw number. Say low, medium or high, name what it covers, and say what happens next.

5. The score goes stale

What teams ship: A confidence value stored alongside the answer.

Why it fails: The user edits the input, the sources change, the model updates. The old number is still sitting there looking current.

Ship instead: Clear it on any change, and say so before the user acts.

What the number has to earn

DECORATION	A SIGNAL THAT WORKS
Percentage shown, all actions live	Below threshold, the action is held
The model's own self report	Checked against real outcomes
98.7 percent	Low, with the reason
Warning text	A route to a human
Stored with the answer	Cleared when anything changes

The audit

Open your product. Find the confidence score. Then answer three questions.

What changes on screen between the highest and the lowest value it can show? If the only thing that changes is the number, you have decoration.

Who checked that 80 percent means right four times in five? If nobody has, it is a guess with a percent sign on it.

Which action gets held back, and does the user know how to get unstuck? A dead end with good manners is still a dead end.

Checked against

Read on 25 November 2026.

- Too Sure for Our Own Good: A User Study on AI Confidence and Human Reliance. 184 participants; well calibrated confidence raised accuracy 20 percent, miscalibrated raised it 2 percent and increased automation bias
- Understanding the Effects of Miscalibrated AI Confidence on User Trust, Reliance, and Decision Efficacy. Miscalibration is hard for users to spot, and telling them about it caused under-reliance rather than better decisions
- Large Language Models Are Overconfident in Their Own Responses. Instruction tuned models show an ownership bias, rating their own answers up to 26 percent higher than identical user-supplied answers
- AI confidence shown as fake precision. UX Patterns Guide on treating uncalibrated exact scores as an anti-pattern, and gating actions instead

---