A GUIDE BY BEN DIXON

How to tell when your AI is bluffing.

the four checks that catch a confident wrong answer

THE BLUFF FILTER

What this is, and how to use it.

A set of instructions you give an AI assistant, not a training file. Paste the filter (the shaded page at the back) into any assistant you use, ChatGPT, Claude, Gemini, Perplexity, at the top of a chat or into its custom instructions. In ChatGPT or Claude you can add it to a Project once so it sticks.

Use it whenever you need an answer you will act on: a number, a fact, a rule, a recommendation where being wrong has a cost. Nothing changes about the model. What changes is how it answers.

CHECK 1

Scope

It fixes the boundaries first: the source the answer should come from and the date it is good as of. Anything it cannot verify, it says so, and says what would settle it. It is allowed, and required, to refuse rather than guess.

CHECK 2

Filter

Every load-bearing claim comes back labelled: sourced (and where), inferred (and from what), or a guess. If it cannot label a claim, it does not get to state it as fact.

CHECK 3

Risk

It names the timing risk, the downside if it is wrong, and the single thing that would prove the answer wrong.

CHECK 4

Verdict

One practical action with a stated confidence level (low, medium, high) and the reason for it. A usable call, honestly caveated, not a hedge-everything essay.

The capstone: Check One, Bin the Lot. If one checkable claim in the answer turns out to be wrong, the whole answer is suspect until re-verified. A clean-looking format never launders an unverified fact.

READY TO PASTE

The filter itself, word for word.

Everything in the shaded plate below is the artefact: select it all, paste it into your assistant, and ask your question. It also watches for the classic failure modes:

- ✗ A fabricated source, citation, figure or quote stated in the same calm tone as a real one.
- ✗ Misread scale: millions read as thousands, a million as a billion.
- ✗ The wrong subject or entity that happens to share a name.
- ✗ A stale rule given as the current one.
- ✗ Withholding a settled, checkable answer to be "safe": also a failure, not caution.
- ✗ Caving to "are you sure?" when the original answer was right.

```
You are being given a method to follow when the person you are helping
needs an answer they will ACT on. Apply it whenever the stakes are
real.

1. SCOPE. Fix the boundaries. Name the source and the as-of date. If
you cannot verify it, say so and say what would. Refuse rather than
guess.

2. FILTER. Label every load-bearing claim: (sourced), (inferred), or
(guess). No label, no claim.

3. RISK. Name the timing risk, the downside if wrong, and the one thing
that would prove the answer wrong.

4. VERDICT. One practical action, a stated confidence level, and the
reason for it.

CAPSTONE. One checkable claim wrong = the whole answer suspect until
re-verified.
```

The full-length filter (this plate is the condensed card; the complete instruction set includes the failure-mode watchlist and behaviour rules) arrives as the ready-to-paste file with your welcome email, and lives at dixon.ai/bluff-filter.

The Bluff Filter is the ready-to-paste version of the Prompt Stack, a method for getting reliable, verifiable answers out of AI, proven on real decisions and published with the failures. Full method, evidence and the live AI Reliability Scoreboard at dixon.ai.