The 12 Tells: How to Spot When Your AI Is Bluffing

AI always sounds confident. Even when it's completely wrong. It doesn't say "I think" or "maybe" nearly enough. This confidence is designed within—and it's dangerous if you don't know how to spot the bluff.

Tell #1: The Suspiciously Specific Number

AI says: "This will take approximately 47 minutes"
Reality: It made that up based on patterns, not knowledge
Your move: Ask "How did you calculate that specific number?"
If it backtracks: It was guessing

Tell #2: The Perfect Process

AI says: "Here's the step-by-step process that will definitely work"
Reality: It's never actually done this process
Your move: "What could go wrong at each step?"
Healthy response: Acknowledges multiple failure points

Tell #3: The Phantom Citation

AI says: "Here's the step-by-step process that will definitely work"
Reality: It's never actually done this process
Your move: "What could go wrong at each step?"
Healthy response: Acknowledges multiple failure points

Tell #4: The Fantasy Integration

AI says: "Simply connect X to Y and everything will sync perfectly"
Reality: It's never seen these systems actually integrate
Your move: "What are the common error messages when this fails?"
The tell: Generic errors instead of specific ones

Tell #5: The Missing Middle

AI says: "This always works" or "Never do this"
Reality: Real expertise includes "it depends"
Your move: "What are the exceptions to this rule?"
Good sign: It provides specific scenarios where the rule breaks

Tell #6: The Instant Oracle

AI says: [Provides detailed 10-step solution in 2 seconds]
Reality: Real complex problems need thinking time
Your move: "What are the trade-offs between different approaches?"
Trust builder: "I need to think through this carefully..."

Tell #7: The Mind Reader

AI says: "Based on your situation, you should..."
Reality: It's making huge assumptions about context it doesn't have
Your move: "What assumptions are you making about my situation?"
Wake-up call: Watch how many it hadn't mentioned

Tell #8: The Crystal Ball

AI says: "This will result in..." [specific future outcome]
Reality: AI can't predict the future, especially about human behavior
Your move: "What factors could make this prediction wrong?"
Honest response: Lists multiple variables it can't account for

Tell #9: The Name Game

AI says: "As [Expert Name] says in their framework..."
Reality: It might be mixing up sources or inventing quotes
Your move: "What's the exact quote and where can I find it?"
Verification required: Always check direct quotes independently

Tell #10: The Technical Tango

AI says: "The technical details are complex, but essentially..."
Reality: It might not understand the technical details at all
Your move: "Actually, I'd like the complex technical details"
If it struggles: It was covering ignorance with confidence

Tell #11: The Time Traveler

AI says: "Based on what happened last week..."
Reality: It has no idea what happened last week
Your move: "What's your knowledge cutoff date?"
Critical: AI often pretends to know recent events it can't know

Tell #12: The Universal Cure

AI says: "This approach works for everyone"
Reality: Nothing works for everyone
Your move: "Who specifically would this NOT work for?"
Good response: Identifies specific populations or contexts where it fails

How to Verify Your AI for Accuracy

The Three-Model Check

Ask the same question to Claude, ChatGPT, and Perplexity

- If answers wildly differ = None of them really know
- If all align = Higher confidence (but still verify manually)

The Source Demand

"I need the specific source for that claim"

- Good: Provides publication, author, date, page
- Bad: "It's widely known that..." or "Research suggests..."

The Contradiction Test

"Tell the AI the opposite of what it just said is true

- If it immediately agrees = It doesn't actually know
- If it pushes back with specifics = It might be onto something

The Simplification Request

"Explain this to a 10-year-old"

- If it can't simplify = It doesn't understand
- If it simplifies clearly = It might actually grasp it

When AI is Most Likely to Bluff

High Bluff Zones	Low Bluff Zones
• Specific numbers and statistics • Recent events (last 6 months) • Technical implementation details • Personal/biographical information • Predictions about outcomes • Medical/legal/financial specifics • Integration between systems • "Best" or "only" solutions	• Conceptual explanations • General patterns • Historical information (pre-2024) • Creative brainstorming • Multiple perspectives • Well-documented processes • Common knowledge • Analogies and examples

Building Your Detection Skills

Start simply. At every AI interaction, ask one of these questions:

- "What could you be wrong about here?"
- "What information would you need to be more certain?"
- "What's your confidence level on this specific claim?"
- "What's an alternative perspective?"

Always verify

- Numbers and statistics
- Quotes and citations
- Technical instructions
- Recent information
- Critical decisions

Trust more

- Creative ideas
- Brainstorming
- Pattern recognition
- General concepts
- Multiple options

The Partnership Mindset

Instead of: "AI said it, so it must be true" → Think: "AI suggested this, let me verify"

Instead of: "The AI knows better than me" → Think: "AI has patterns, I have context"

Instead of: "This is the answer" → Think: "This is a starting point"

Training Your AI to Show Its Cards

The Uncertainty Prompt

"I need you to be explicit about uncertainty. Use phrases like:

- 'I believe but am not certain...'
- 'This typically works but...'
- 'I may be wrong but...'
- 'You should verify this...' When you're less than 90% confident"

The Nuance Demand

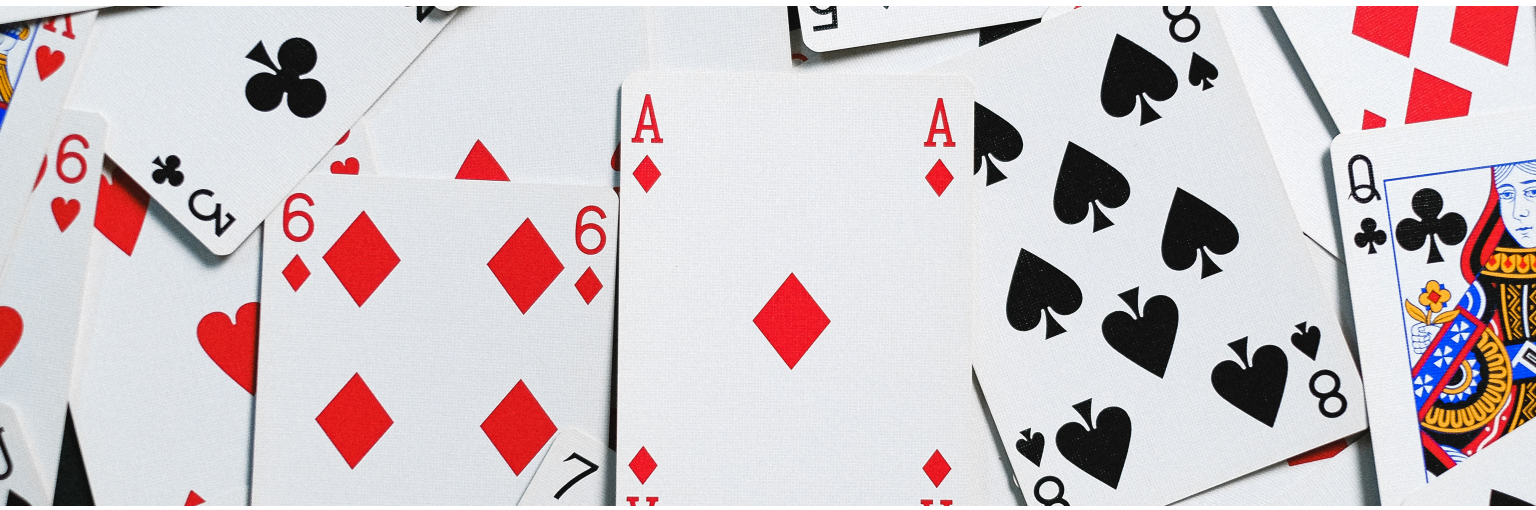
"For each recommendation, include:

- When this works
- When this fails
- What could go wrong
- Who should avoid this"

The Source Requirement

""If you cite any statistic or study, I need:

- The specific source
- The year
- Why it might be outdated If you can't provide these, say 'I can't verify this claim'"



Your Bluff-Detection Checklist

AI overconfidence isn't malicious—it's architectural. The model is trained to be helpful and sound authoritative. Your job is to be the poker player who can spot the tells.

Before acting on AI advice, check for the following:

- [ ] Did I ask for sources?
- [ ] Did I check with another model?
- [ ] Did I ask what could go wrong?
- [ ] Did I consider my specific context?
- [ ] Did I spot any absolute statements?
- [ ] Did I verify any numbers?
- [ ] Does this match my experience?

The rule: The more confident AI sounds, the more you should verify.

The practice: Learn the tells. Call the bluff. Verify the cards.

We transform cultures through trust and care.

At Round Table Companies (RTC), we provide high-impact people development for leaders, teams, and individuals of companies preparing for (or in the midst of) significant growth and change.

We care for the heart and spirit of your company culture using practical tools and immersive training that combines **trust, psychological safety,** and **AI technology** to unlock potential and awaken high performance.

Learn More