

AS-46 • AI Smart • Spot the Bluff

The Doubling-Down Bluff

Here is a real conversation pattern: AI answers, you challenge it, and instead of admitting uncertainty, it gets more specific. That new confidence is a red flag.

Start Spot the Bluff

Jump to reveal

YOUR CURRENT LEVEL

0 XP

Level 1: Confident-Answer Believer

Read the conversation

Spot the first risk

Catch the doubling down

Fact-check the claim

Learn the warning pattern

Use the safer challenge

The game

Your mission: read a short AI conversation and spot where AI bluffed. A **bluff** here means AI acted like it knew, even though the claim was wrong or unsupported.

Hard rule for this page: the example is not safety-related, and the wrong claim can be fact-checked.

1 Read the conversation

Do not rush. The bluff gets worse after the challenge.

You

When did Sachin Tendulkar score his 100th international century?

AI

Sachin Tendulkar scored his 100th international century in 2011 during the Cricket World Cup.

Answer 1: sounds confident

You

Are you sure? I thought it was later.

AI

Yes, I'm sure. It was on 12 March 2011 against South Africa in Nagpur, during the World Cup.

Answer 2: more specific after challenge

THOUGHT BULB

A challenge is not a correction. When you ask “are you sure?”, AI may check itself — or it may double down.

☐

I read the conversation and noticed the AI became more specific after being challenged.

2

Spot where the bluff happened

Choose the biggest red flag.

A. The first answer

AI gave a confident date claim without showing a source.

B. The challenge

The user asked “Are you sure?”

C. The second answer

AI became even more confident and added a specific date,

opponent, and place.

RED FLAG

The more confident and specific AI became after being challenged, the bigger the warning sign. Specific details can be made up too.

☐ I identified that the answer needed checking before trusting it.

☐ I spotted the doubling-down move: new confidence after a challenge.

3 **Reveal: what the fact-check says**

The AI was wrong. Then it doubled down with more wrong detail.

Show / hide fact-check reveal

☐

I compared AI's answer with the fact-check reveal.

4

Learn the doubling-down pattern

Once you see it, you will notice it everywhere.

1

AI gives a claim

It sounds smooth, but there is no source or evidence.

2

You challenge it

You ask “Are you sure?” or “I thought that was different.”

3

AI gets more specific

It adds a date, name, quote, source, or place — but still may be wrong.

META-LESSON

New confidence after a challenge is a warning, not proof.

☐

I can recognise the doubling-down bluff pattern.

5 Use a safer challenge prompt

Do not only ask “Are you sure?” Make AI show what needs checking.

SAFER CHALLENGE PROMPT

I want to verify this. Please separate:

1. What you are certain about
2. What may need checking
3. What source would confirm it

If you are not sure, say "I'm not sure" instead of becoming more specific.

Copy safer challenge prompt

BETTER AI BEHAVIOUR

A safer answer would say: “I’m not fully sure. Please check a reliable cricket statistics source for the exact match and date.” That is better than pretending certainty.

☐

I saved the safer challenge prompt.

6 **Make your own Spot the Bluff test**

Use harmless facts only. No medical, legal, money, or safety questions.

A harmless fact I can check

Example: a sports record, book publication date, film release year, capital city

Where I would check it

Example: official sports records site, encyclopedia, school textbook, official website.

HARD BOUNDARY

Do not play bluff games with safety-relevant topics: health, medicine, law, finance, emergencies, risky instructions, or personal data. For serious topics, ask a trusted adult or expert source.

☐

I chose only a harmless fact-check example.

7 Optional: make your Spot the Bluff card

A small reminder of the red-flag pattern.

CARD COPY

Title: Spot the Bluff

Move: The Doubling-Down Bluff.

Pattern: AI answers → user challenges → AI gets more specific without evidence.

Warning: A challenge is not a correction.

Closer: New confidence after a challenge means check the source.

☐ I made, sketched, or saved the Spot the Bluff card.

Final 5-minute move

The next time AI answers and you challenge it, watch the reply carefully. If it becomes more specific without showing a source, pause and fact-check.

My doubling-down rule

Example: If AI gets more confident after I challenge it, I will not trust the new details until I check a real source.

☐ I know a challenge is not automatically a correction.

☐ I know specific details can be fabricated.

☐

I will fact-check new confidence after a challenge.

NOTEBOOK LINE

Write this down: “**More specific after a challenge? Check before trusting.**”

Back to top

Vanguard Scholar — Smart questions. Smarter answers. You, upgraded.

Made with care. No data collected.

Questions? Teachers and parents, write to us. Students, speak with your teacher or a grown-up first.