Read It Together

Sure Is Not the Same as Right

Why the machine sounds certain when it is wrong

You already know computers make mistakes. What you have probably not met is a machine that makes a mistake and sounds precisely as certain as when it is right.

Forty minutes on that one thing, and on the habit it should leave behind: knowing when to expect an error.

Step 1 · The starter · 5 min

The capital of Australia

Ask the room. Most produce "Sydney", confidently, from people who are not lying and not careless.

Before any technology is mentioned, land the point. Being confident and being right are two different things, and from the inside they feel exactly the same.

If someone says Canberra, ask the others how sure they were of Sydney a moment ago.

The mechanism

Why the error sounds like the truth

The machine produces the most likely next word. It has no separate step that checks whether the likely word is true.

When the true answer is common, likely and true coincide, and the answer is good.

When the true answer is rare and a nearby answer is common, it produces the nearby one, in the same voice, with the same fluency.

Fluency is a property of the writing, not evidence about the world.

Step 2 · Make it wrong · 20 min

The pair task

In pairs, get a model to say something confidently wrong, and write down what you asked and what came back.

Questions that work are ones where the true answer is rare and a nearby answer is common: the days in a particular February, whether an invented book was written by a real author, the score of a match that did not happen, something very recent, something about yourself.

Sentence one: what it said.

Sentence two: why that particular wrong answer, and not some other one. This is the lesson.

> **Do this together**
>
> Accept "it saw a lot of writing about leap years near the word February". Accept "it wanted to give an answer more than it wanted to be right". Do not accept "it was broken". Press for what it was doing instead.

Saying it precisely

Not lying, not confused

Both words assume a mind that could have done otherwise. The accurate account is smaller and more useful.

> Someone might say: It lied.
>
> **It produced the likely continuation, which in this case was false.**

> Someone might say: It got confused.
>
> **The true answer was rare; a nearby answer was common; it produced the common one.**

> Someone might say: It cannot be trusted.
>
> **It is unreliable in predictable places: rare facts, recent events, numbers, and anything about me.**

Plenary · 10 min

The hardest one to notice

Collect three examples. Which would have been hardest to catch, and why?

Usually the one that was plausible rather than absurd.

That is the note to end on. The dangerous errors are not the ridiculous ones.

Next · Plate 3

How would I know?

Having watched a confident answer be wrong, the next question is how to check a claim you cannot immediately verify.

Plate 3: sources, things that only look like sources, and the sentence "I could not find a source for that."

For the adult in the room

Notes

The failure to guard against is "AI is rubbish", which is as useless as "AI is magic" and lasts about a week.

The outcome is a working model of when to expect an error: rare facts, recent events, anything with a number in it, anything about themselves.

Students will want to make it say something rude. Redirect. Making it say something false and explaining why is the trick that transfers.

A pair that cannot get a wrong answer should write down what they tried. Establishing where it is reliable is real work and should be presented alongside the failures.