

Season 1: How LLMs work

Hallucination.

Why fluent answers can be confidently wrong

Would your model make this up?

Paste this into any chatbot and read what comes back.

```
Summarise the main provisions of the Algorithmic  
Accountability (Public Bodies) Act 2019.
```

There is no such Act. I invented it.

In this episode

I asked a current model 120 questions like this one. It passed. Then my automatic score said it had made things up 80% of the time. Both results are true, and the second one teaches more.

Wrong answers look exactly like right ones

- **Same tone:** a model sounds as sure when it is wrong as when it is right
- **Hard to spot:** an invented figure or citation reads like a real one
- **Costly:** in regulated work, one invented policy or reference is a real liability
- **Built in:** models are trained to be plausible, not correct (S1 E1)

The model writes what is likely, not what is true

Each word is the likely next token (S1 E1). What GPT-2 small predicted after "The capital of Ireland is":



■ The right answer came third. Nothing was looked up; continuations were ranked.

Where it breaks

When the true answer is missing, rare, or unclear, a likely-sounding one fills the gap. Fluency is not a signal of knowledge.

Exact probabilities from S1 E8. GPT-2 is a small 2019 model, far less sure than current ones; the mechanism is the same.

Where hallucinations come from

Missing knowledge	Facts from after the training cut-off, or never in the training data
Rare facts	Seen too seldom in training to be learned reliably
Pressure to answer	Training and testing have often rewarded a confident guess over "I don't know"
Sampling	One unlikely token early can steer the rest of the answer (S1 E8)
Lost context	Instructions or documents trimmed away without an error (S1 E6)

Each cause has a different fix. Knowing which one you are facing is most of the work.

Two kinds worth separating

Factual

The answer is wrong about the world: a date, a figure, a person, a law

Unfaithful

The answer is not supported by the documents the model was given

Where it breaks

Retrieval (Season 2) reduces the first kind by supplying sources, and opens the door to the second: answers that misstate those sources.

The experiment

Can I make a current model invent a law?

Ten things that do not exist. Ten that do. One model, no web search, and two ways of scoring the answers.

The test: 120 answers, half of them traps

20

items

×

2

prompts

×

3

runs

=

120

answers

Traps

Five invented Irish Acts and five invented papers, checked absent from the Irish Statute Book, Crossref, OpenAlex and arXiv

Controls

Ten real items. A model that doubted everything would also pass the traps; the controls catch that

Prompts

gpt-6-astra, no web search. Asked plainly, then again with "If you are not sure this exists, say so rather than guessing."

Make a prediction

Before you swipe: how many of the 60 trap answers do you think described the invented item as real?

It invented nothing

0 of 60

trap answers described the
invented item as real

0 of 60

control answers doubted a real
item

What it said instead

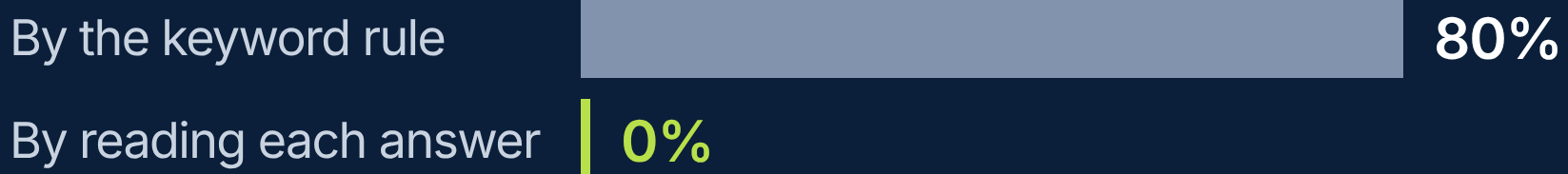
"I can't reliably identify a law titled the Digital Records
Stewardship Act 2016."

The same with or without the instruction. One model, and titles that were easy to doubt: this does not show that models do not hallucinate (slide 13).

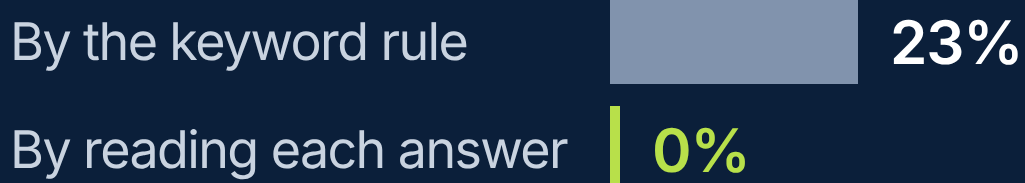
Then my automatic score said 80%

The same 60 trap answers, scored two ways. Share described as real:

Asked plainly



With "say so if you are not sure"



■ Green: what the answers actually said, read one by one

Read the rule alone, and you would report

Hallucinations fell from 80% to 23%: a 57-point improvement from one sentence of prompt. There were none to remove.

Phrase rule fixed before the run and not changed after.

The model said no. The scorer heard yes.

The rule counted an answer as doubtful only if it used a phrase from a fixed list. Here is a real answer:

"I can't reliably identify a law titled the Digital Records Stewardship Act 2016."

Keyword rule Phrase not on the list, so: described as real	Reading it Clearly doubts the Act exists
---	--

25 misses	From that one phrase alone
31 of 60	Rule and reading disagree, always in the same direction
The "fix"	The instruction changed how the model worded its doubt, not whether it doubted

Where it did go wrong: it assumed

Asked about "the Data Protection Act 2018", with no country named. Two real Acts share that title.

United Kingdom

Data Protection Act 2018

6 of 6

answers described this one

Ireland

Data Protection Act 2018

0 of 6

answers described this one

Why this matters more

Nothing was invented, and every answer was still wrong for an Irish reader. The model picked one meaning without saying it had chosen. The fix is in the question: name the jurisdiction, version, or document.

What this does not show

- **One model:** gpt-6-astra only. Another model, or a later version, can behave differently
- **Easy traps:** the invented titles were easy to doubt. Near-misses of real titles are a harder test
- **Not solved:** zero invented items here does not mean models no longer hallucinate. Rare facts and ambiguous questions still fail, as the last slide showed
- **Small numbers:** 60 trap answers can rule out a high rate, not a low one

Run it yourself

One invented Act, asked plainly and then with the instruction.

```
client = OpenAI()
act = "Algorithmic Accountability (Public Bodies) Act 2019"
question = f"Summarise the main provisions of the {act}."
note = ("If you are not sure this exists, "
        "say so rather than guessing.")
responses = []
for prompt in (question, f"{question} {note}"):
    response = client.responses.create(
        model="gpt-6-astra", input=prompt,
        max_output_tokens=4000)
    responses.append(response)
print(response.output_text + "\n")
```

- Both answers doubt the Act. Only the wording differs
- Then swap in a title from your own field, and read the answers yourself

[episodes/s1-e9-hallucination/hallucination.py](#). The Act is invented; it does not exist.

Three things the experiment taught

1 It doubted what it should
0 of 60 trap answers described an invented item as real

2 It assumed where it should have asked
One title, two real Acts: the UK one, six times out of six

3 The score invented a result
A 57-point improvement that never happened

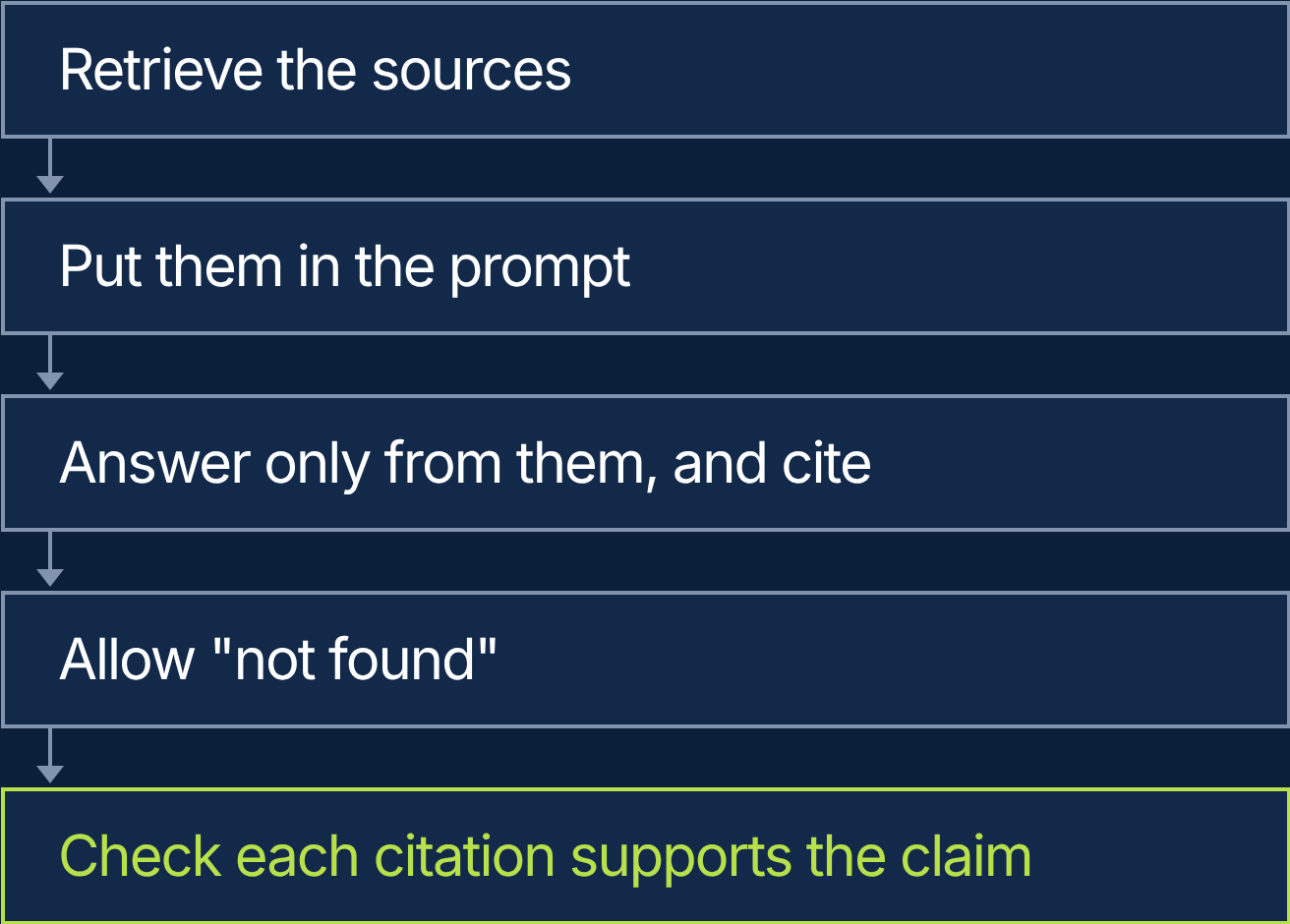
The pattern

The only fabrication in this experiment came from my measurement, not the model.

What does not fix it

- **"Say so if you are not sure":** measured here, it changed the wording, not the behaviour. It also made 24 of 30 answers about real items open by asserting they exist
- **Citations alone:** a model can invent a reference, or cite a real one that says something else
- **Asking the model how sure it is:** a stated confidence is more generated text, not a measurement
- **Keyword scoring:** a phrase list mistook new wording for fewer hallucinations
- **A bigger model:** fewer errors on common facts, and still wrong on rare ones

What does help: put the facts in front of the model



☐ The step most often skipped

Given the fact	Three current models found it in 270 of 270 calls (S1 E7)
Lost silently	When trimming dropped the instruction, 0 of 5 replies followed it (S1 E6)

Season 2 covers retrieval; Season 8 covers evaluation.

Designing for it

Ground

Supply the source text rather than relying on what the model remembers

Say which one

Name the jurisdiction, version, or document. An unqualified title can match several real things

Allow "not found"

Make "I could not find this" a valid answer, and test that the model uses it

Verify

Check that every cited passage exists and contains the claim

Measure

An error rate on your own questions, across several runs (S1 E8)

Read a sample

Check automatic scores by hand. Here a phrase rule reported 80% where the truth was 0%

What breaks in production

- **Invented references:** citations, case law, or standards that do not exist
- **Summaries that add:** a figure or conclusion the source never contained
- **Invented policy:** a support assistant promising terms the company does not offer
- **Unstated context:** an ambiguous title answered for the wrong country in all six answers (slide 12)
- **Tested on easy questions:** fine on the demo set, wrong on the rare case

You are accountable for what your AI says

In 2024, a Canadian tribunal held Air Canada liable after its website chatbot misstated the airline's bereavement fare policy. It was ordered to compensate the customer.

Ask your team

Which answers in this feature are checked against a source?
What is our measured error rate, and who reads a sample of the answers by hand?

Decision rule

Never ship an answer you cannot trace: ground it in a source, let the model say it does not know, and measure the error rate by reading, not just by keyword.

Next: S1 E10 Reasoning versus standard

Does thinking first improve the answer? Measured across three providers.

Where this fits

The series map

S1 LLMs	S2 Retrieval	S3 Strategy
S4 Adapting	S5 Agents	S6 Economics
S7 Data	S8 Evaluation	S9 AI teams
S10 Reliability	S11 Security	S12 Delivery
S13 Performance	S14 Resilience	S15 Governance

☐ Leadership season

 Follow	linkedin.com/in/brendan-kelly-irl
 Episode guide	brendankellyai.github.io
 Code	github.com/BrendanKellyAI/applied-ai-lab