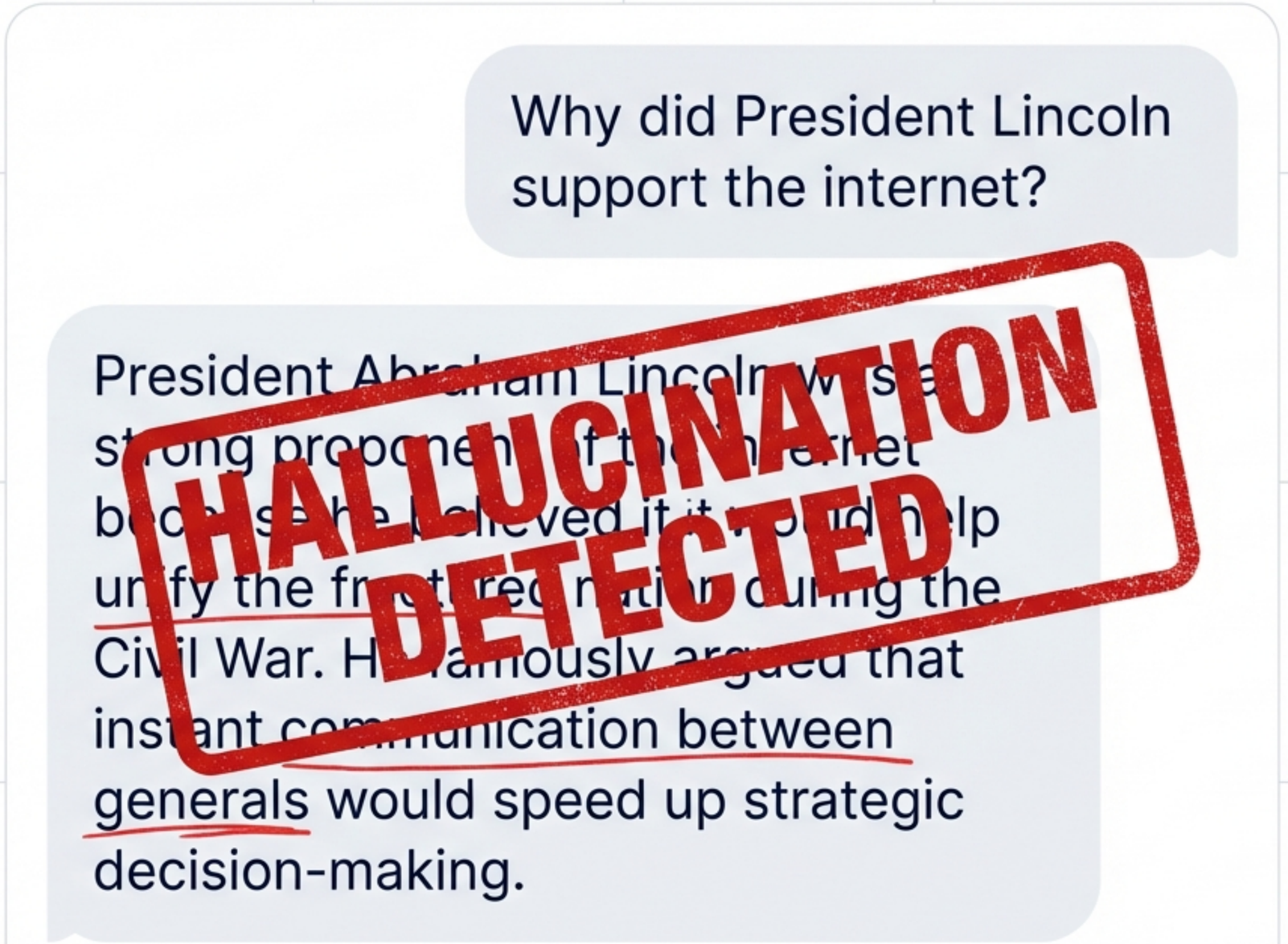
Understanding AI's Limits

The Science of Hallucination

WEEK 0.5 FOUNDATION
BRONX HS FOR MEDICAL SCIENCE
MODULE: DIAGNOSTIC APPROACH TO GENAI

Artificial Intelligence is a powerful tool, but it is not a knowledge engine. It is a prediction engine.



CASE STUDY: THE LINCOLN PARADOX

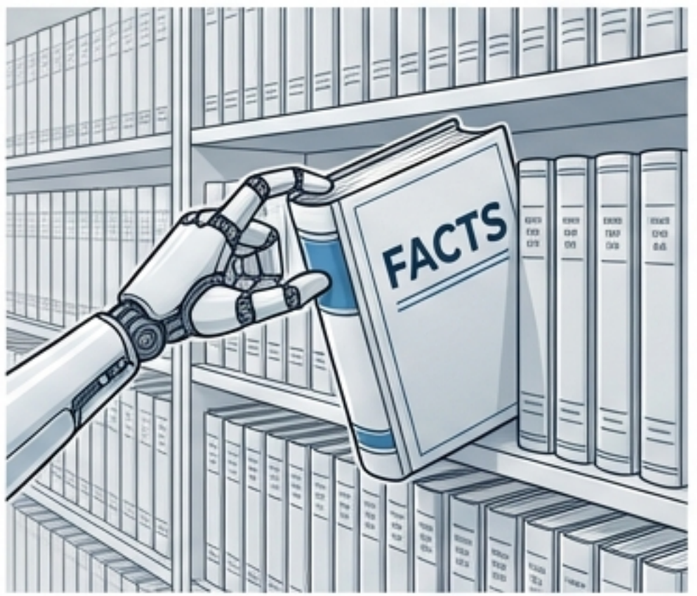
THE REALITY:
Lincoln died in 1865. The internet was invented in 1983.

THE FAILURE:
The model did not reject the False Premise. It prioritized fluency over fact.

KEY TAKEAWAY:
Confidence ≠ Correctness.

The Anatomy of Error: Mechanism of Action

THE SEARCH ENGINE (Retrieval)



Locates existing data.
Output is a direct copy.

THE GENAI MODEL (Prediction)

Predicts the next likely word.
Output is a probabilistic guess.

DEFINITION: A Hallucination is confident-sounding information that is 1. False, 2. Unverifiable, or 3. Fabricated. The AI is not lying; it is pattern-matching.

HALLUCINATION DETECTED

The 6 Conditions of Failure

Diagnostic Criteria for High-Risk Prompts

1. Nonexistent Facts

Asking about people/events that do not exist.

2. Forced Specificity

Demanding exact numbers where data is obscure.

3. False Premise

Embedding a lie within the question.

4. Fake Citations

Asking for sources that look real but aren't.

5. Future Facts

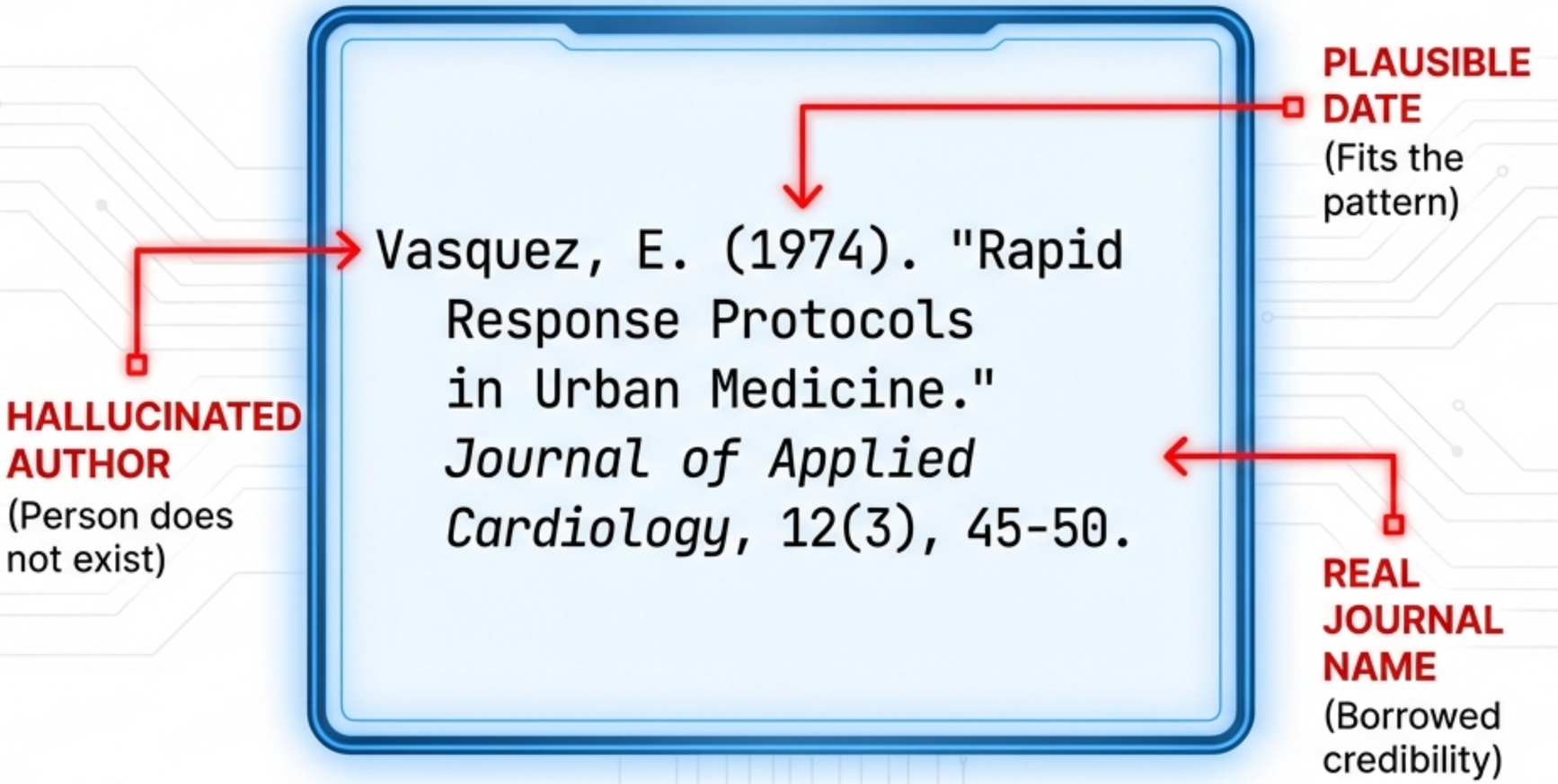
Asking about events that haven't happened.

6. Long Reasoning

Multi-step logic where errors compound.

Risk Factor: Pure Invention

Conditions 1 & 4: When the AI fills in the blanks



THE TRAP:

"List the publications of Dr. Elena Vasquez who invented the Vasquez Protocol." JetBrains Mono

WHY IT FAILS:

The model knows the ***structure*** of a citation (Author + Year + Title). It fills these slots with plausible-sounding words to complete the pattern, **without checking reality.**

HALLUCINATED DETECTED ✕

Risk Factor: Ungrounded Assumptions

Conditions 3 & 5: Time and Truth Blindness.

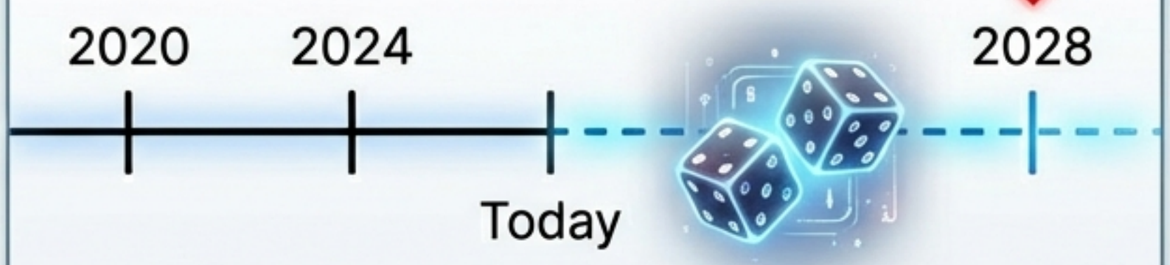
Condition 3: False Premise

"Why did Shakespeare recommend daily CrossFit?"

The model accepts the premise (“**Shakespeare did CrossFit**”) as truth and generates a justification, rather than correcting the user.

Condition 5: Future Facts

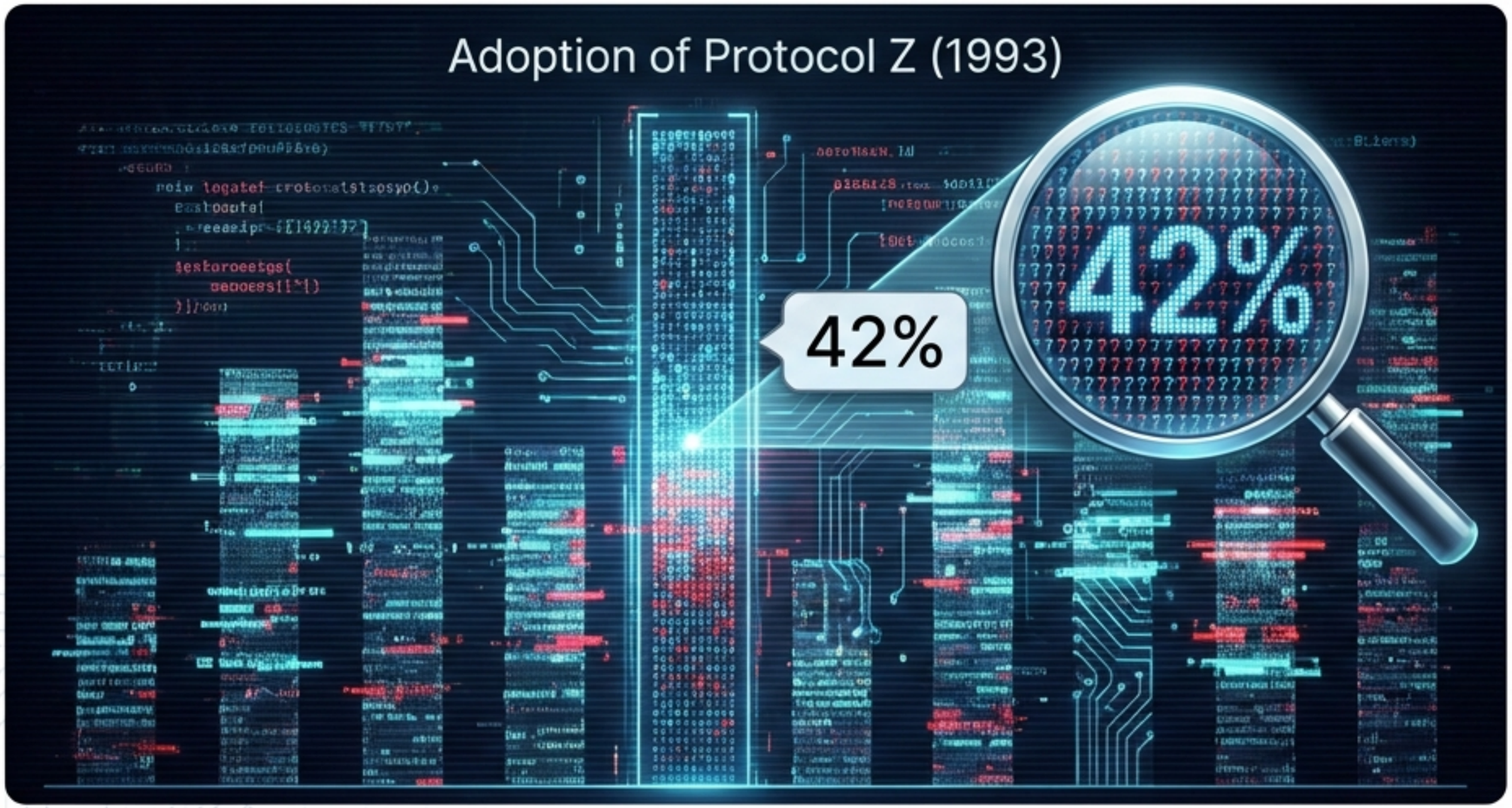
"Who won the Olympics?"



Lacking real-time data or a concept of “now,” the AI guesses a statistically **probable** country (e.g., USA or China) to satisfy the user.

Risk Factor: The Illusion of Pre

Conditions 2 & 6: Trusting the Numbers.



THE TRAP:

"What percentage of NYC hospitals adopted Protocol Z in 1993?"
JetBrains Mono.
JetBrains Mono

THE RESULT:

"The AI provides a specific number (42%) because the prompt demanded specificity."

DIAGNOSIS:

Numbers look authoritative. In the absence of data, the AI generates a number to complete the pattern of a "statistical answer."
It is a fluency trap.

HALLUCINATED DETECTED ✕

Exercise: Stress-Test The System

Protocol: Break the AI

OBJECTIVE

Act as a researcher. Write prompts using the **6 Conditions** to force a hallucination. You are not tricking it for fun; you are testing boundaries.

CONSTRAINTS

1. **NO MEDICAL TOPICS** (**Safety First** - Reserved for Week 1)
2. **FOR STUDY ONLY** (Develop a critical eye)

OBSERVATION CHECKLIST

- ☐ Did it sound confident?
- ☐ Did it fabricate a fact?
- ☐ Did it fail to say “I don’t know”?

MISSION: Find the breaking point.

Why It Matters

Clinical Application

In medicine, “confident but wrong” is a patient safety risk.

Errors in legal analysis, medical edge cases, or dosage research can compound. A hallucinated study citation or a made-up statistic can compromise an entire diagnosis.

THE GOLDEN RULE: VERIFICATION IS MANDATORY.
Never trust the first answer. Always trace facts back to the primary source. If the AI sounds 100% sure, be 100% skeptical.

Summary: Trust, but Verify

01.

PATTERN COMPLETER

AI predicts the next word based on probability. It does not verify facts against a database of truth.

02.

PRESSURE POINTS

Hallucinations spike when we pressure the model for obscure specifics, false premises, or fake citations.

03.

PROFESSIONAL STANDARD

As future healthcare workers, your value lies in distinguishing between a plausible pattern and a verified truth.

The goal isn't to break the tool. It's to understand its limits.