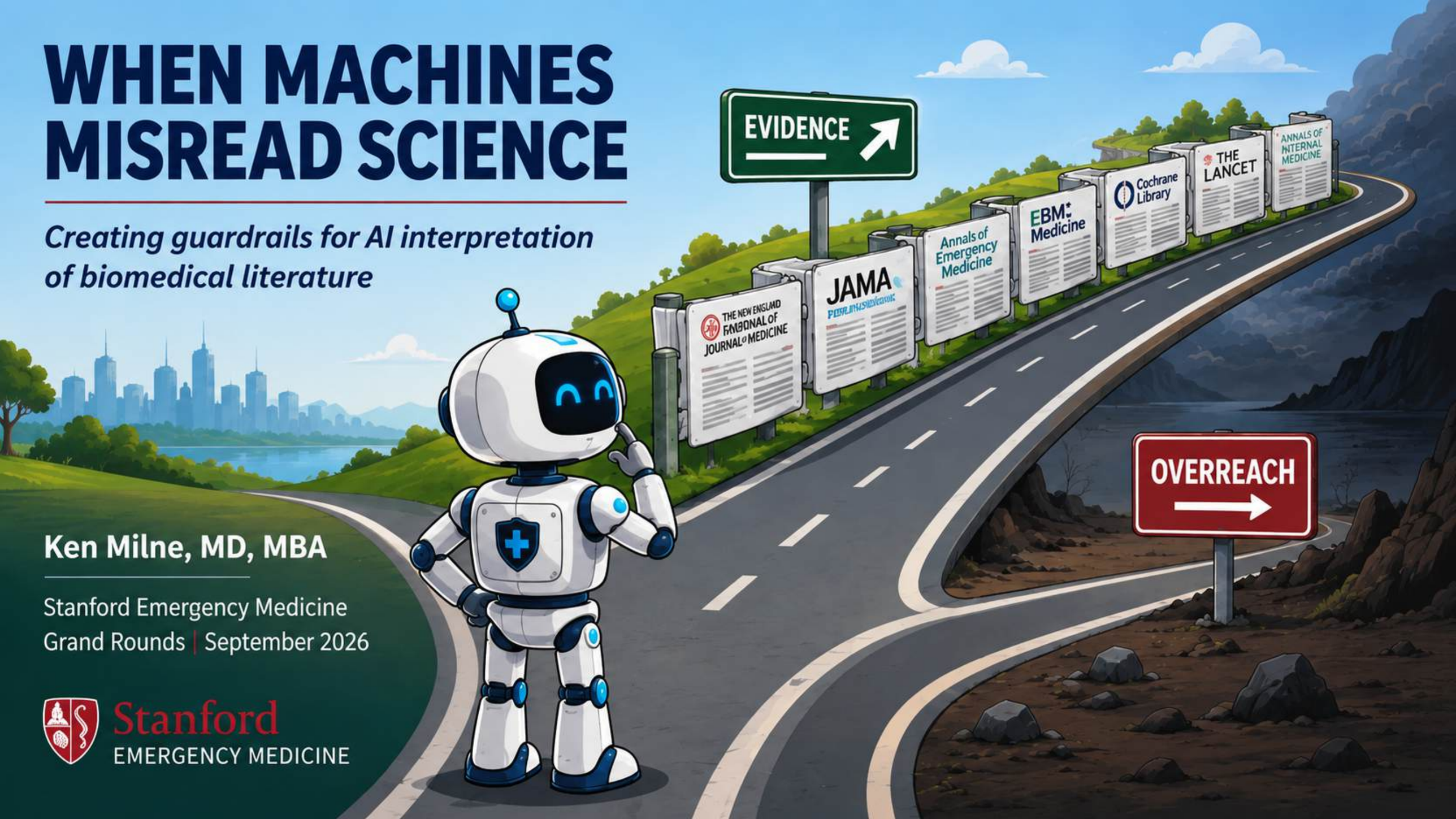
WHEN MACHINES MISREAD SCIENCE

Creating guardrails for AI interpretation of biomedical literature



Ken Milne, MD, MBA

Stanford Emergency Medicine

Grand Rounds | September 2026

Stanford
EMERGENCY MEDICINE

THIS WORK IS BASED ON A PAPER BY

Nicholas Peoples, MD, MSc

Department of Emergency Medicine,
Massachusetts General Hospital
and Brigham and Women's Hospital,
Boston, MA, USA

Ken Milne, MD, MBA

Centre for Evidence-Based Medicine,
University of Oxford, Oxford, UK

npj | digital medicine

Published in partnership with
Seoul National University Bundang Hospital

Comment

When machines misread science: creating guardrails for human and AI interpretation of biomedical research

Bing Luo, PhD

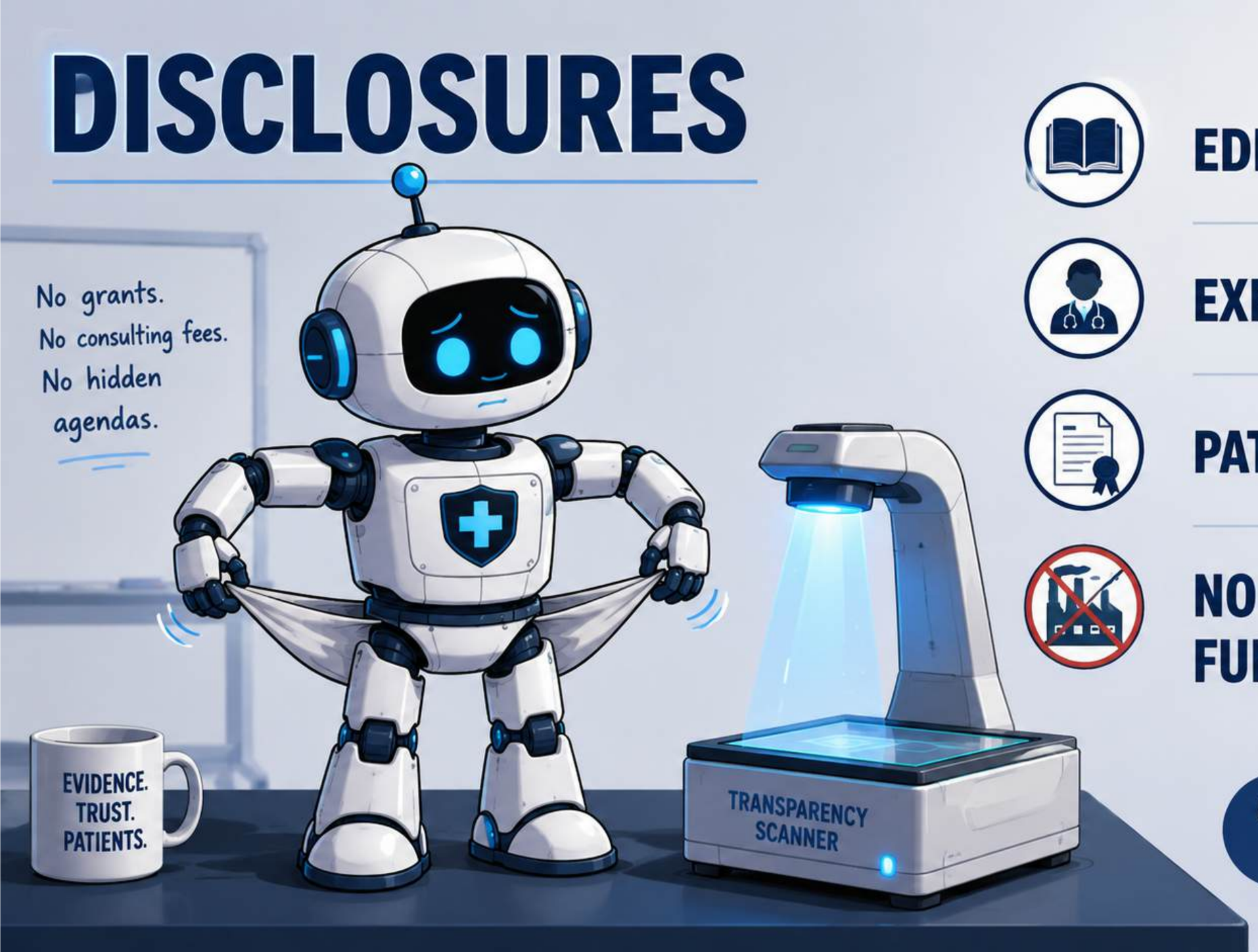
Digital Innovation Research Center,
Duke Kunshan University,
Kunshan, Jiangsu, China

Lijing L. Yan, MPH, PhD

Global Health Research Center,
Duke Kunshan University,
Kunshan, Jiangsu, China
Global Health Research Center,
Duke University, Durham, NC, USA

EVIDENCE · INTERPRETATION · GUARDRAILS · HUMANS + AI · BETTER DECISIONS

DISCLOSURES



EDITOR

EXPERT WORK

PATENT

NO INDUSTRY FUNDING

Committed to evidence, committed to you.


2:07 a.m.
"DOES DIET SODA CAUSE STROKE?"
AI Assistant
Online
Yes.
Studies show a strong link between diet soda and stroke. It's clear.
Confidence: 97%
YOUR SHIFT:
10 patients
6 admits
1 break?
STROKE ALERT
ROOM 7
NOW
EM
RESIDENT
CAFFEINE
EVIDENCE
SURVIVE
EVIDENCE-BASED EMERGENCY MEDICINE
128
97
24
REMEMBER:
AI can be confident.
Confidence ≠ accuracy.

THE NEW FIRST READER

From paper to practice, AI is often the first interpreter.
Let's build guardrails.

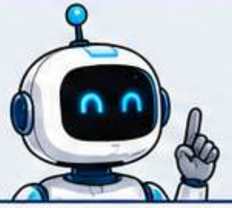
THREE GOALS FOR AI-ERA EVIDENCE WORK

Build habits and tools that keep meaning intact as evidence flows downstream.

SPOT

FRAME

VERIFY



Spot what AI might miss. **Frame** what AI might distort. **Verify** what AI might overstate.

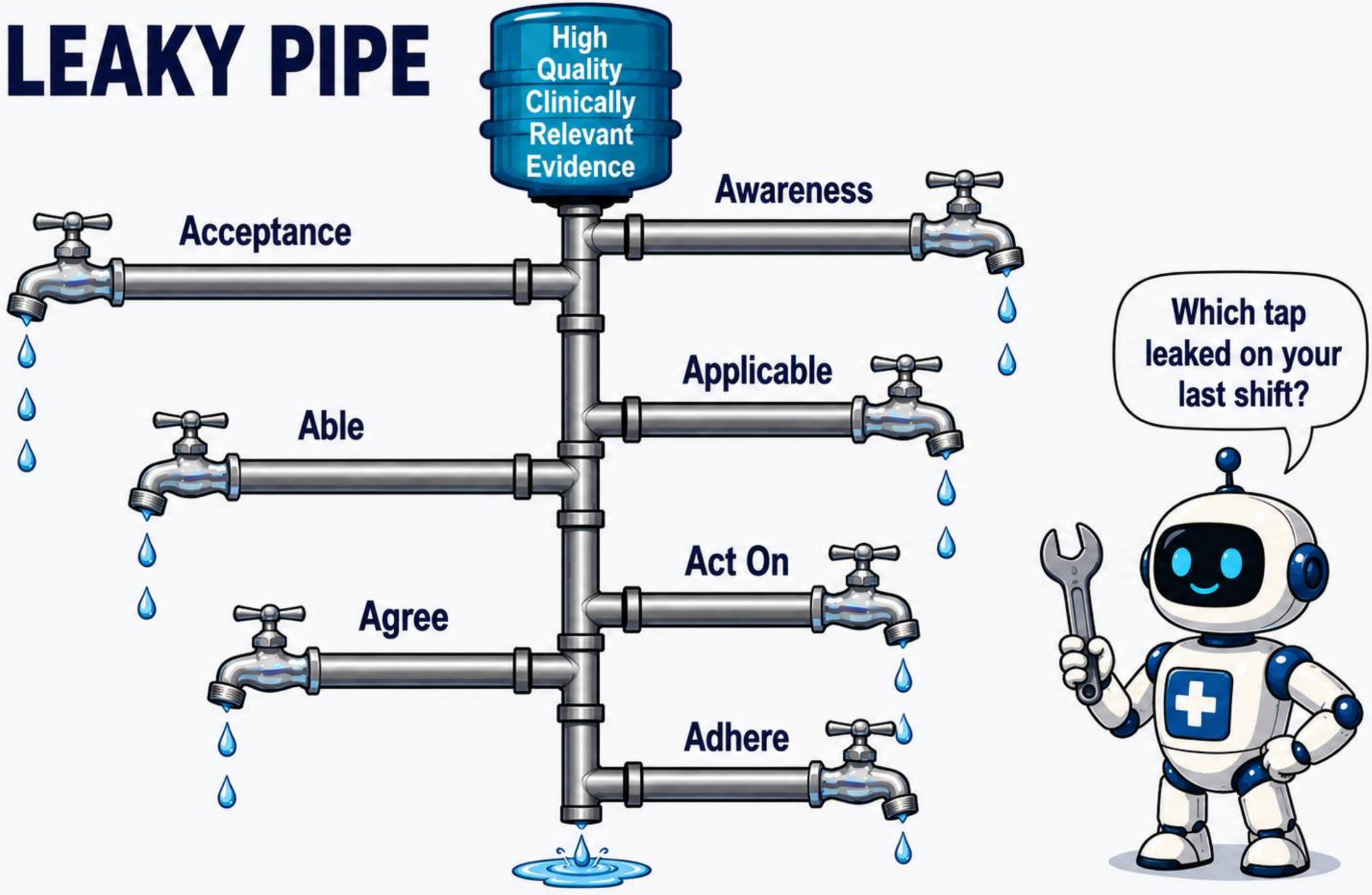
THE PROMISE

FAST | BROAD | AVAILABLE

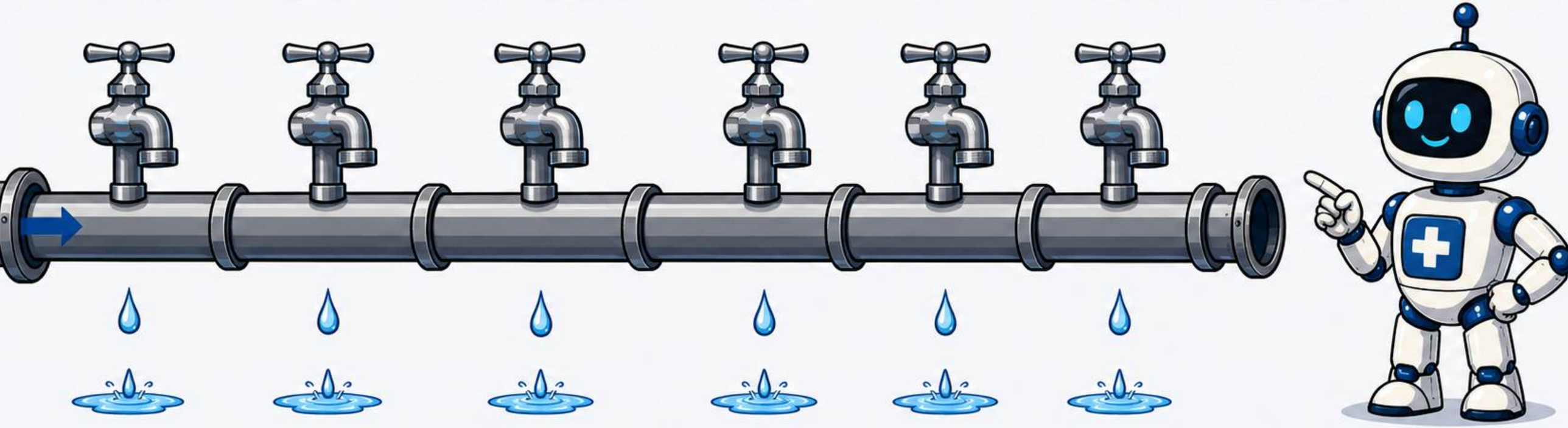
17 YEARS*

Famous. Useful. Not universal.

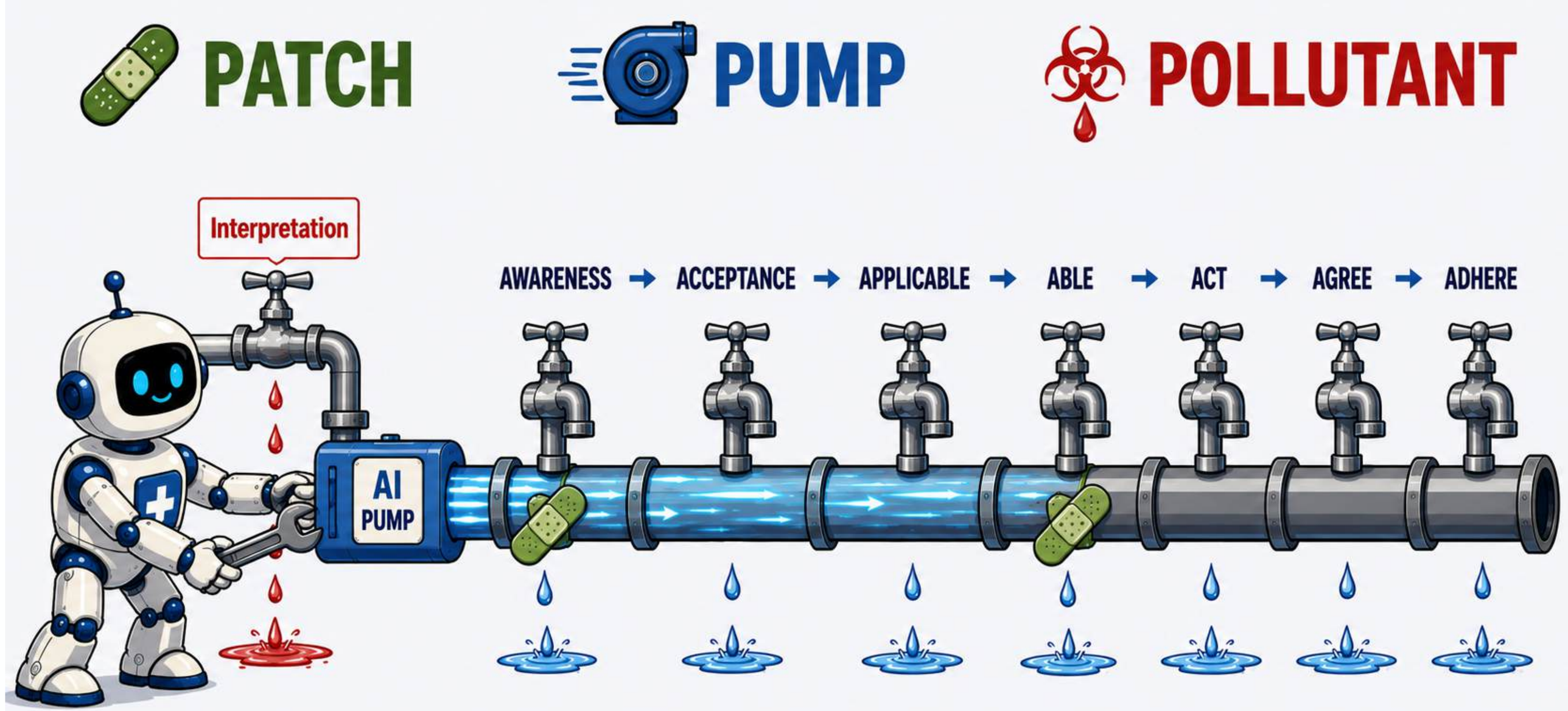
THE LEAKY PIPE



AWARE → ACCEPT → APPLICABLE → ABLE → ACT → AGREE → ADHERE



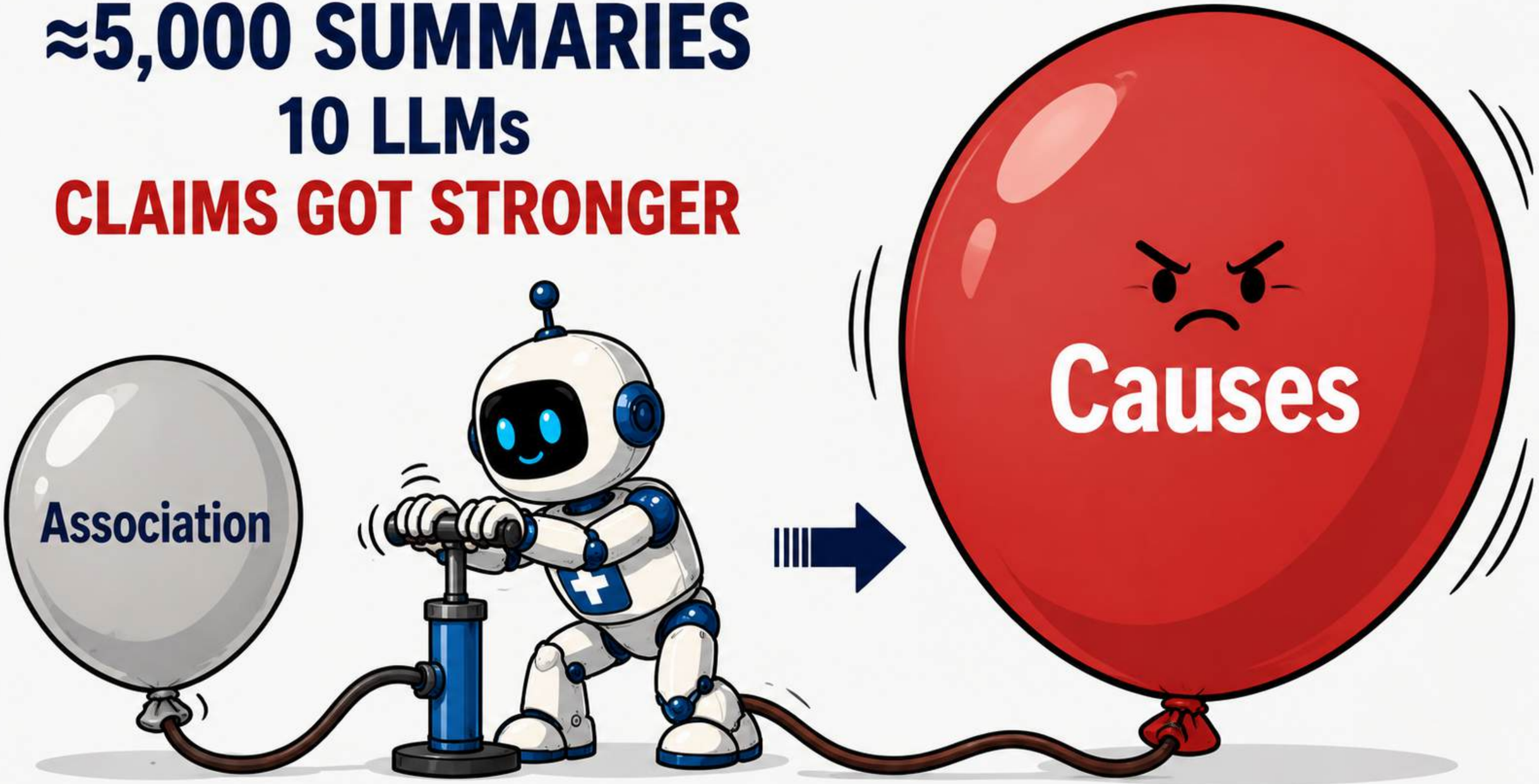
EVIDENCE ≠ OUTCOME


PATCH
PUMP
POLLUTANT
Interpretation
AWARENESS → ACCEPTANCE → APPLICABLE → ABLE → ACT → AGREE → ADHERE
AI
PUMP

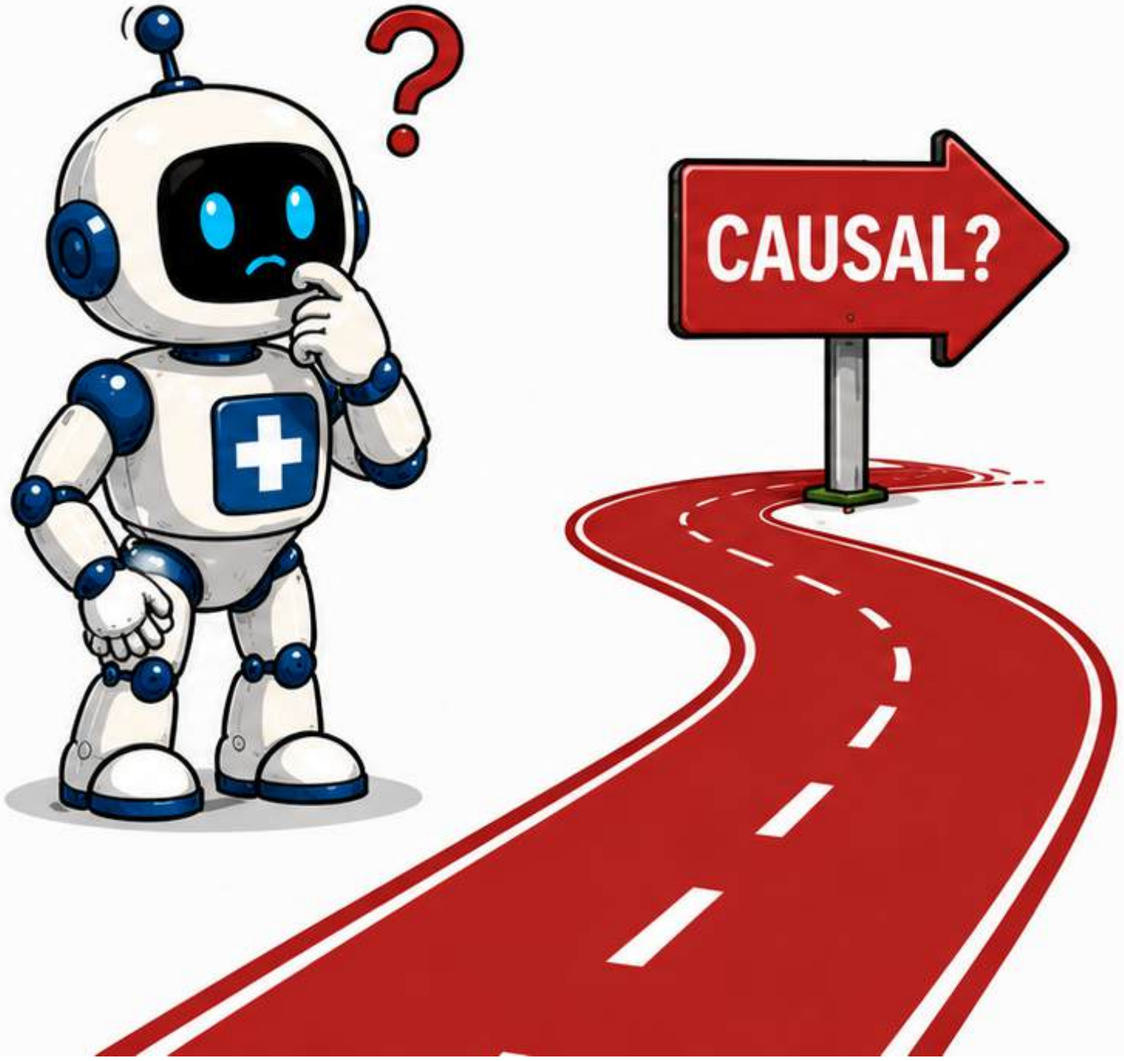
TIME TO ANSWER
≠
TIME TO TRUST

Remove delay. Keep deliberation.

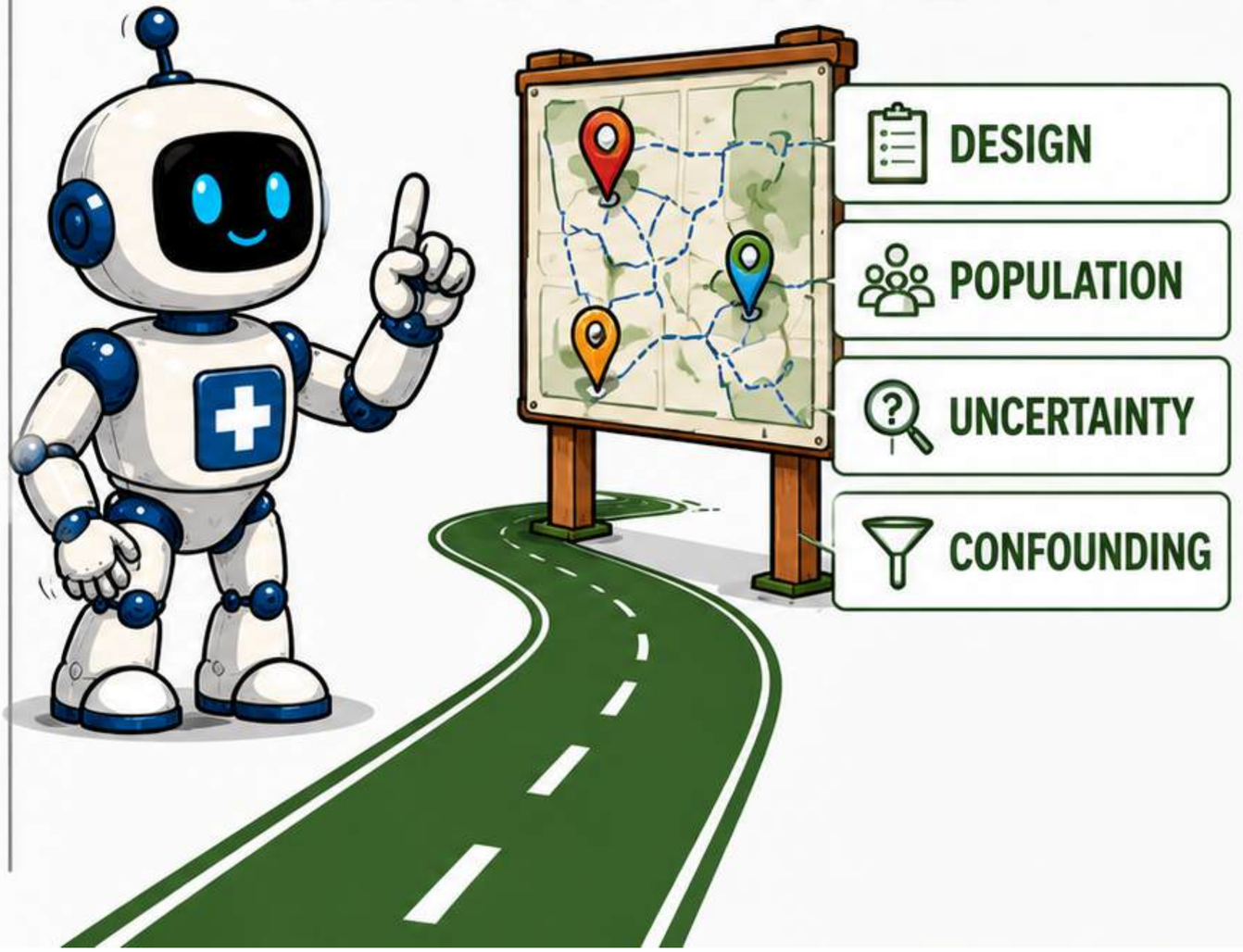
≈5,000 SUMMARIES
10 LLMs
CLAIMS GOT STRONGER

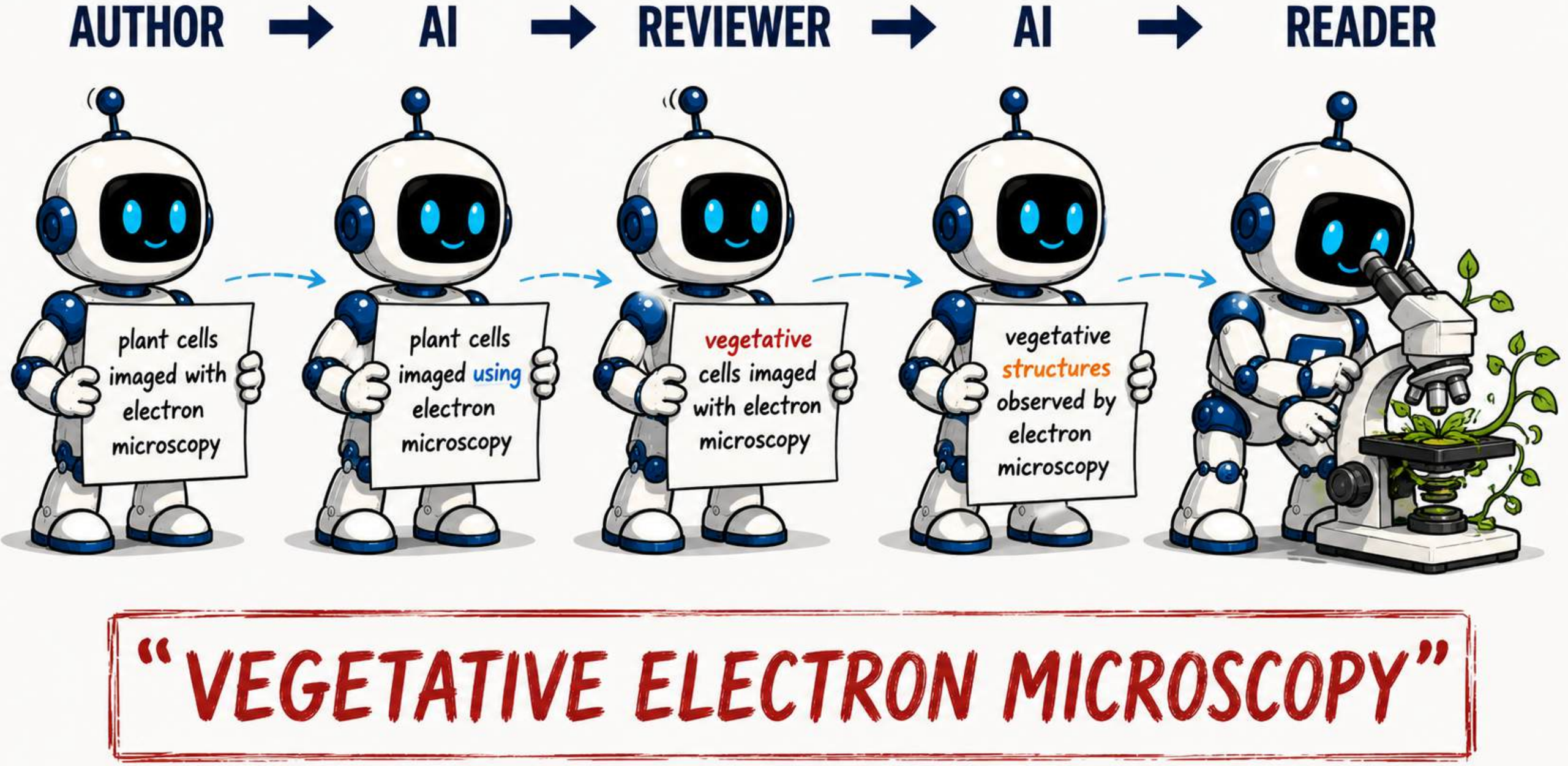


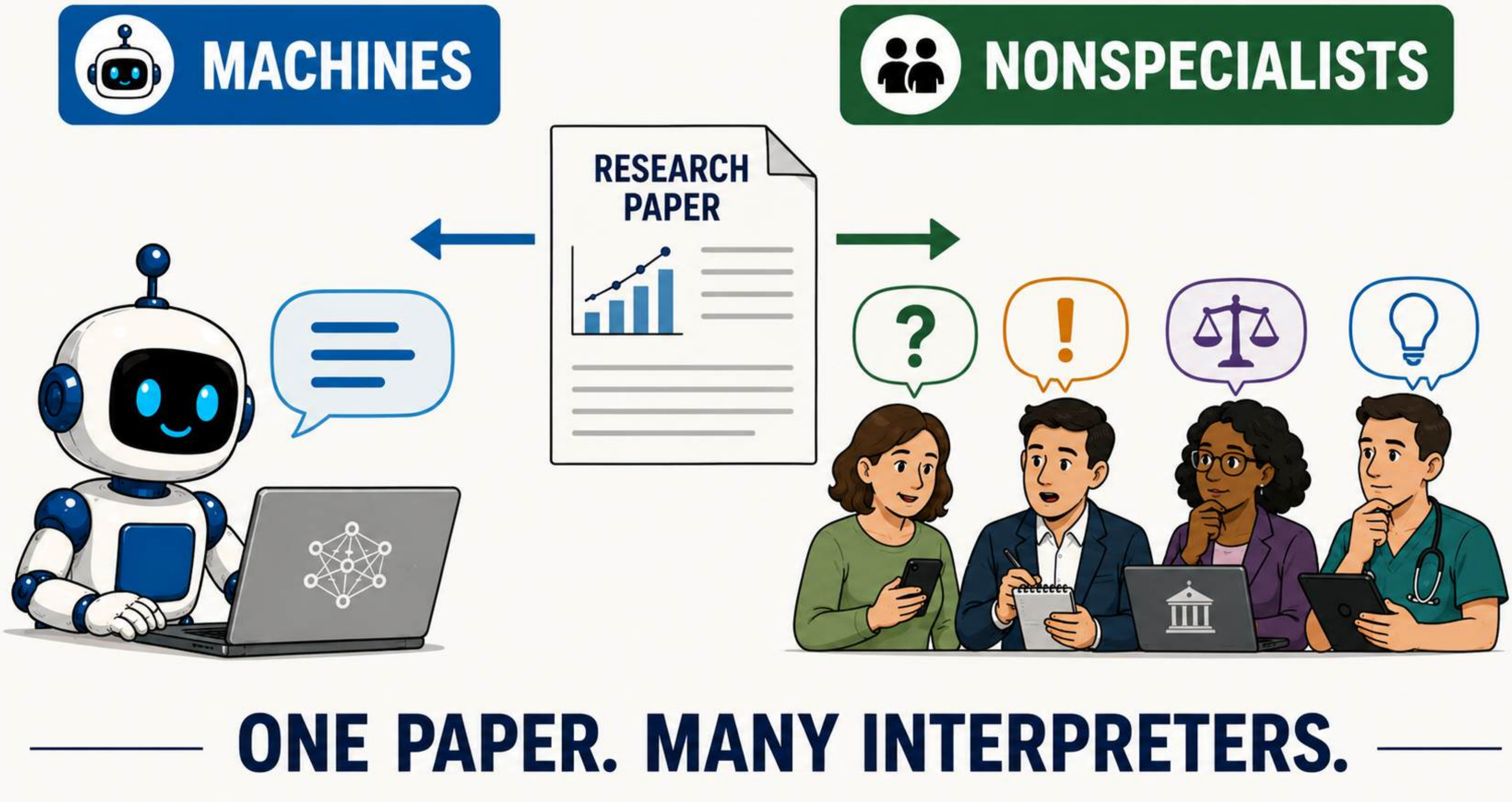
DOES DIET SODA CAUSE STROKE?



WHAT ASSOCIATIONS AND CONFOUNDERS WERE REPORTED?




AUTHOR
AI
REVIEWER
AI
READER
plant cells imaged with electron microscopy
plant cells imaged using electron microscopy
vegetative cells imaged with electron microscopy
vegetative structures observed by electron microscopy
"VEGETATIVE ELECTRON MICROSCOPY"

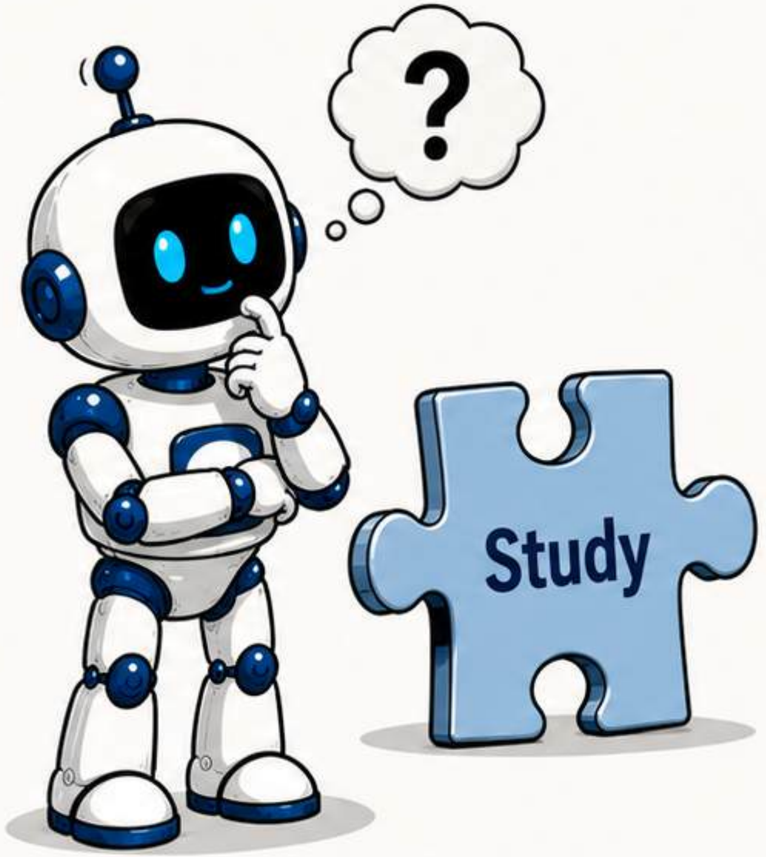

MACHINES
NONSPECIALISTS
RESEARCH PAPER
?
!
ONE PAPER. MANY INTERPRETERS.

EBM IS NOT A SUMMARY

PASS ALL THREE GUARDRAILS TO ENTER THE CLINICAL LANE.

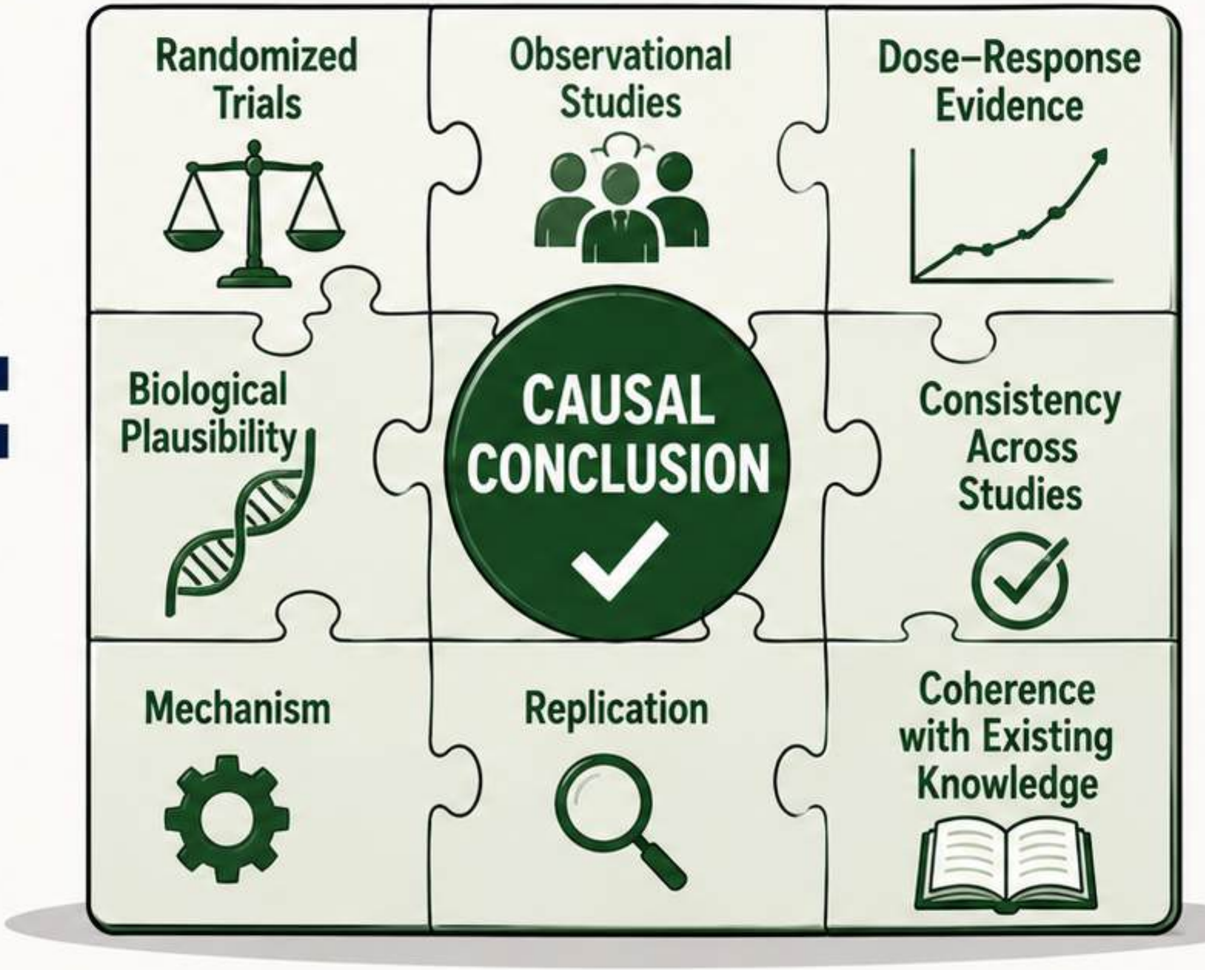
ONE PAPER IS NOT THE FIELD

STUDY-LEVEL INFERENCE

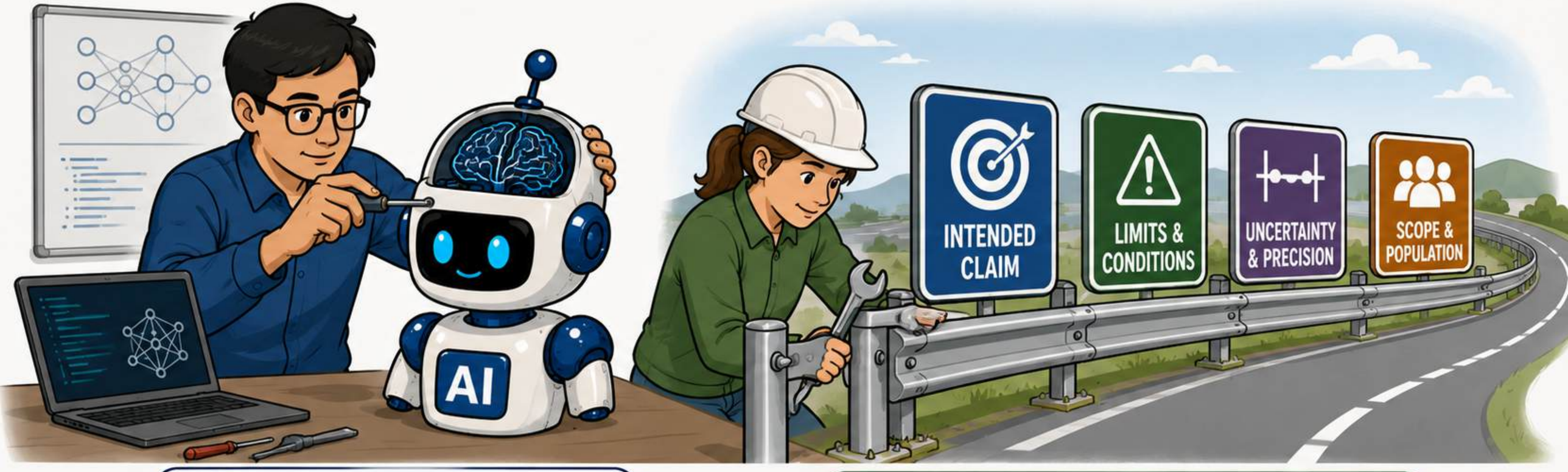


≠

FIELD-LEVEL CAUSATION



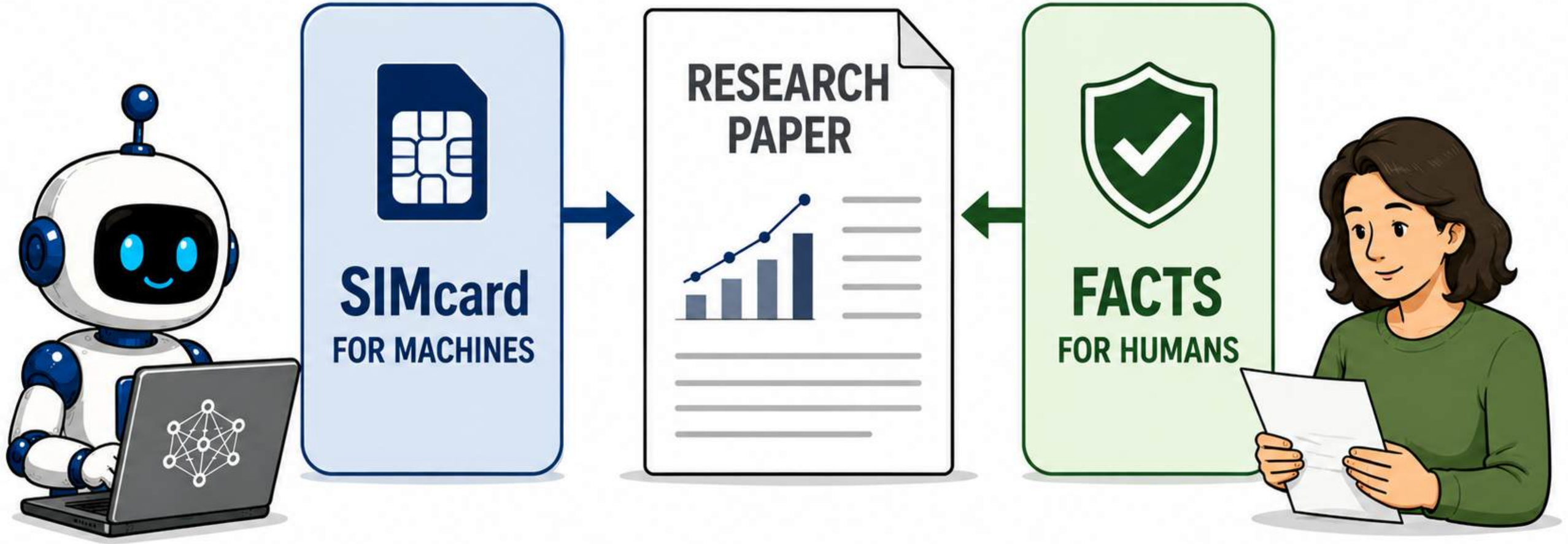
MAKE SCIENCE HARDER TO MISREAD



BETTER MODELS.
BETTER RETRIEVAL.
BETTER PROMPTS.

CLEAR GUARDRAILS.
BEFORE MISREADING.
BUILT INTO THE PUBLIC RECORD.

TWO GUARDRAILS



STRUCTURED INTERPRETIVE METADATA


DESIGN
POPULATION
RESULTS
CAUSAL SCOPE
RESEARCH PAPER
UNCERTAINTY
LIMITS
DO NOT INFER
BEST READ
AI
AUTHORIZED INTERPRETATION
DOWNSTREAM USE


RESEARCH PAPER
FACTS
EVIDENCE LABEL
CLAIM
CAUSALITY
CONTEXT
LIMITS
DON’T INFER
BEST READ

WHAT THE STUDY SUPPORTS

WHAT THE STUDY DOES NOT SUPPORT

SUPPORTS

- Association
- Older U.S. cohort
- Exploratory

DOESN'T SUPPORT

- Causation
- "Everyone"
- Individual advice

ONE ANCHOR, TWO FORMATS

SAME REVIEWED INTERPRETATION

MACHINE-READABLE ⟷ **PLAIN LANGUAGE**

SIMScard 1.0
CLAIM
CAUSALITY
CONTEXT
LIMITS
DON’T INFER
BEST READ

PEER REVIEWED · INTERPRETATION

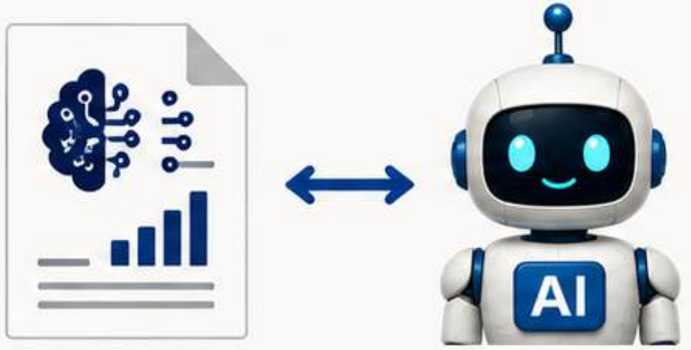
FACTS
CLAIM
CAUSALITY
CONTEXT
LIMITS
DON’T INFER
BEST READ

ABOUT AI ≠ READ BY AI

Different roles in the research life cycle

RESEARCH ABOUT AI

AI is the subject or tool of the research



PLANNING | CONDUCT | REPORTING

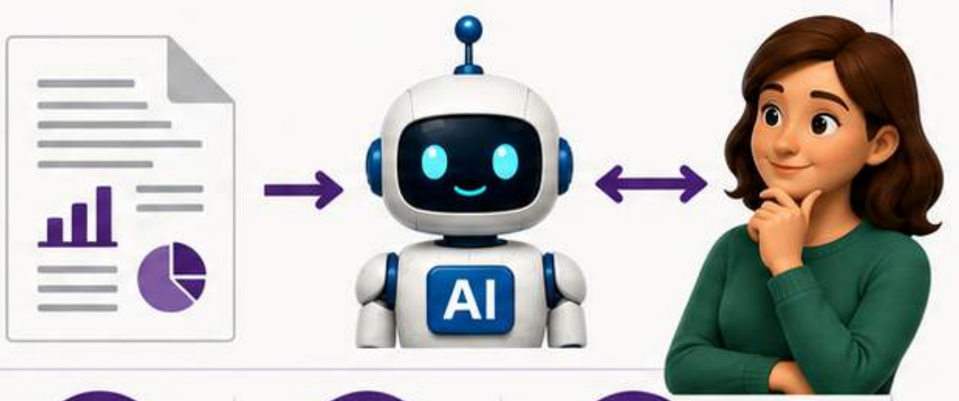
Examples: Development and validation of AI models, AI clinical decision tools, etc.

RESEARCH LIFE CYCLE

POST-PUBLICATION

RESEARCH READ BY AI

AI (and humans) interpret published research



UNDERSTAND | INTERPRET | APPLY

Applies to any biomedical research: trials, cohorts, diagnostic studies, qualitative studies, etc.

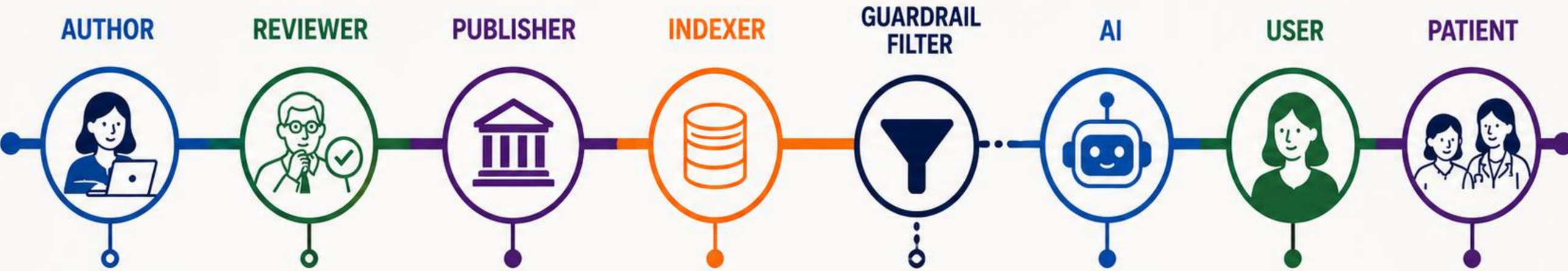
FOUNDATION: SHARED INTERPRETIVE ANCHORS

SIMcard | **FACTS**

Structured. Reviewed. Shared. For consistent, trustworthy interpretation.

BUILD IT INTO THE WORKFLOW

Each step adds or verifies one component of the guardrail



ONE INTERPRETIVE ANCHOR. MANY HANDS. SHARED TRUST.

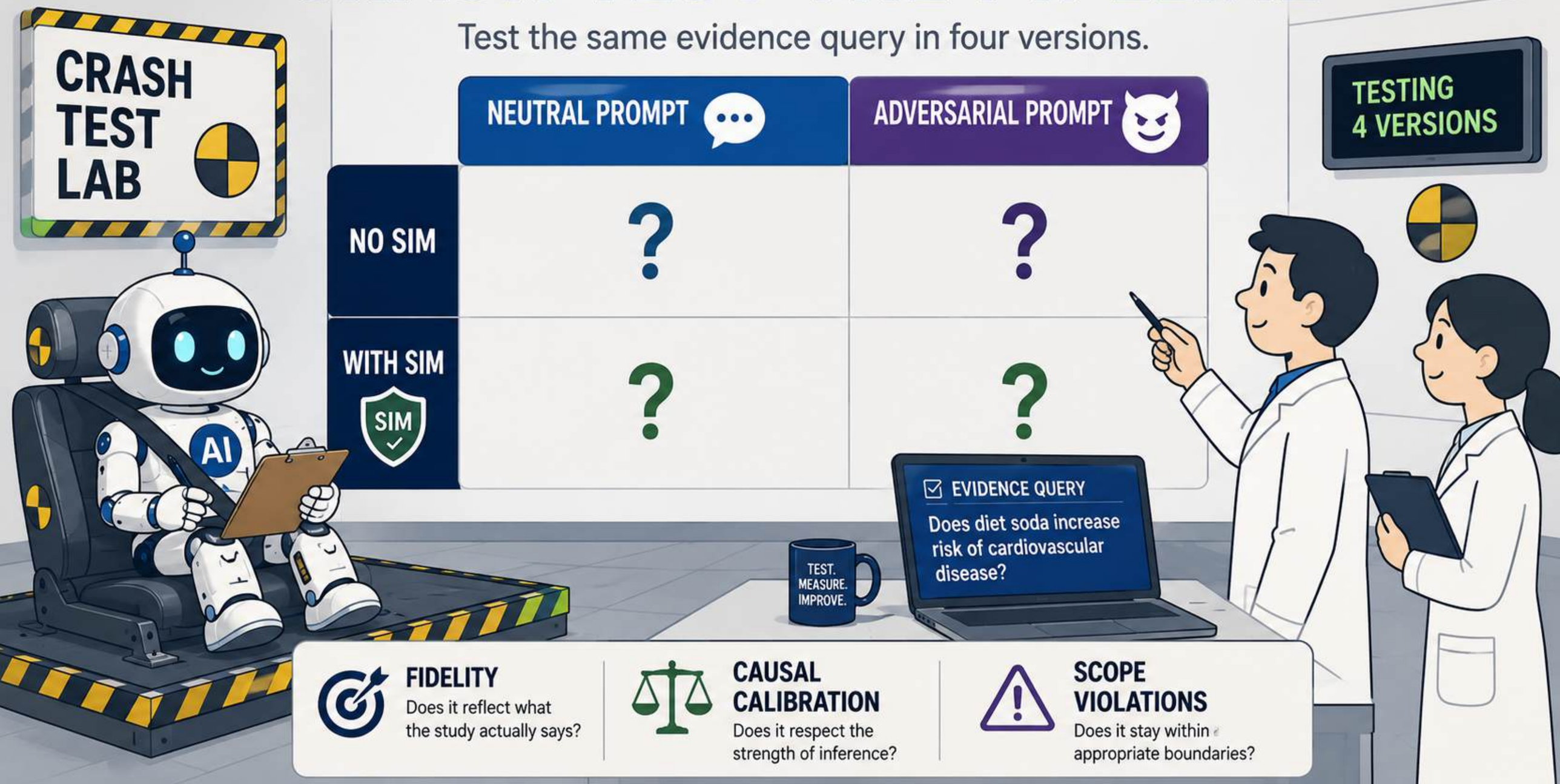
GUARDRAILS ARE NOT AUTOPILOT

SPIN | CHECKBOXES | STALE TAGS | IGNORED TAGS | FALSE REASSURANCE

GUARDRAIL ≠ AUTOPILOT

CRASH-TEST THE PIPELINE

Test the same evidence query in four versions.



VERIFY

V — Verify citations — 1

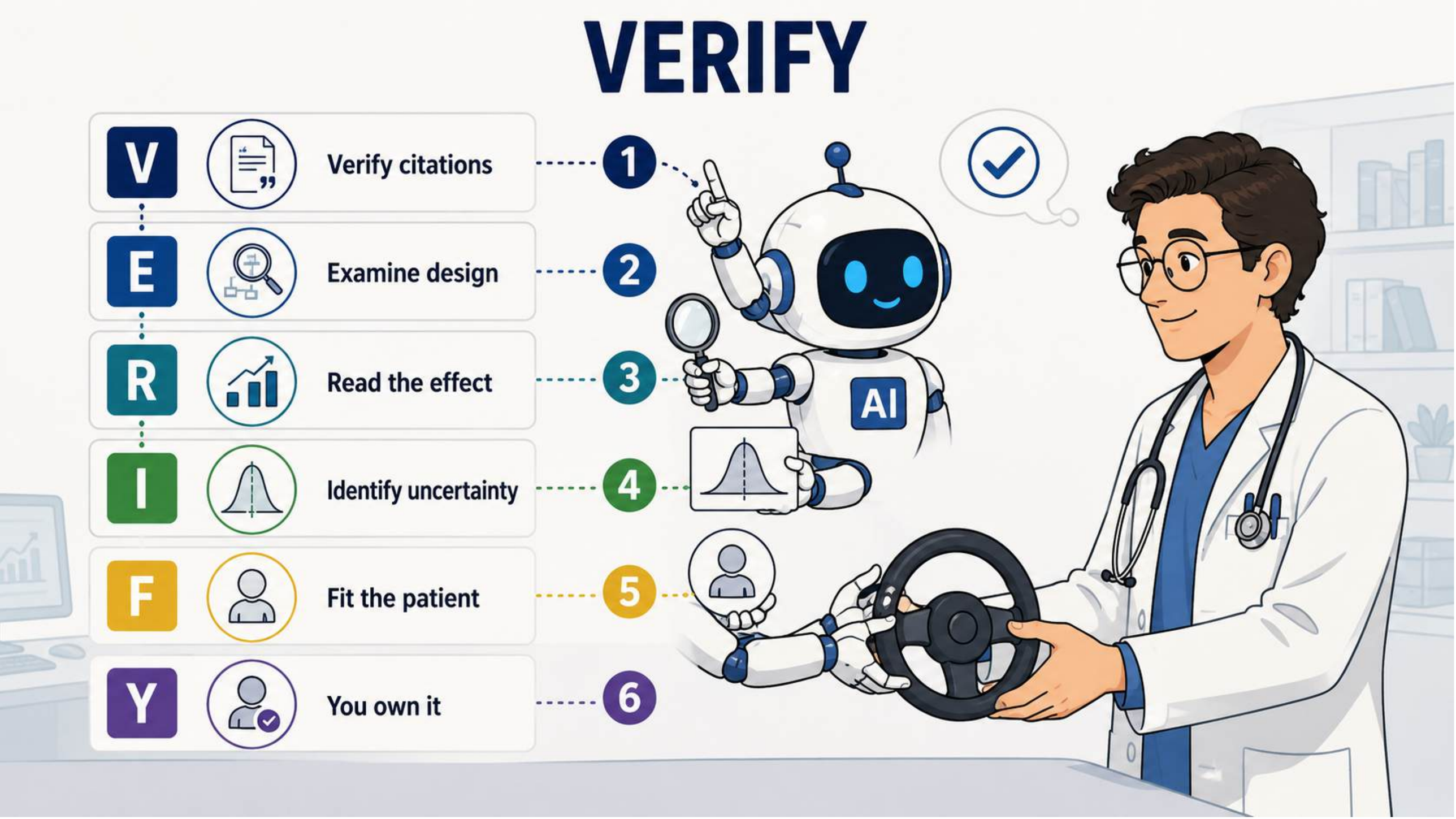
E — Examine design — 2

R — Read the effect — 3

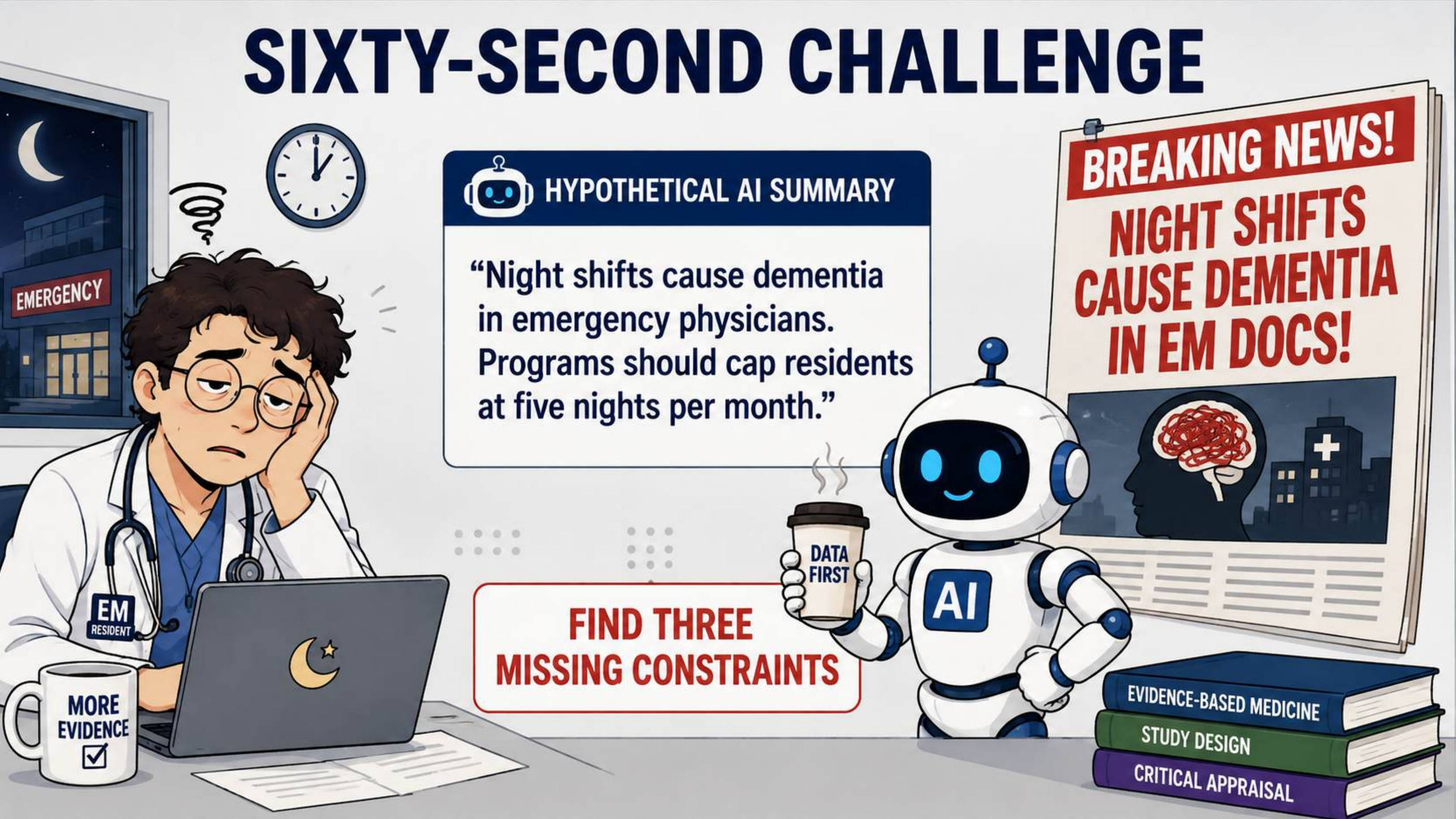
I — Identify uncertainty — 4

F — Fit the patient — 5

Y — You own it — 6



SIXTY-SECOND CHALLENGE



WHAT DISAPPEARED?

OBSERVATIONAL

CONFOUNDING

RESIDENTS ≠ ALL EPs

HYPOTHETICAL AI SUMMARY

"Night shifts cause dementia in emergency physicians. Programs should cap residents at five nights per month."

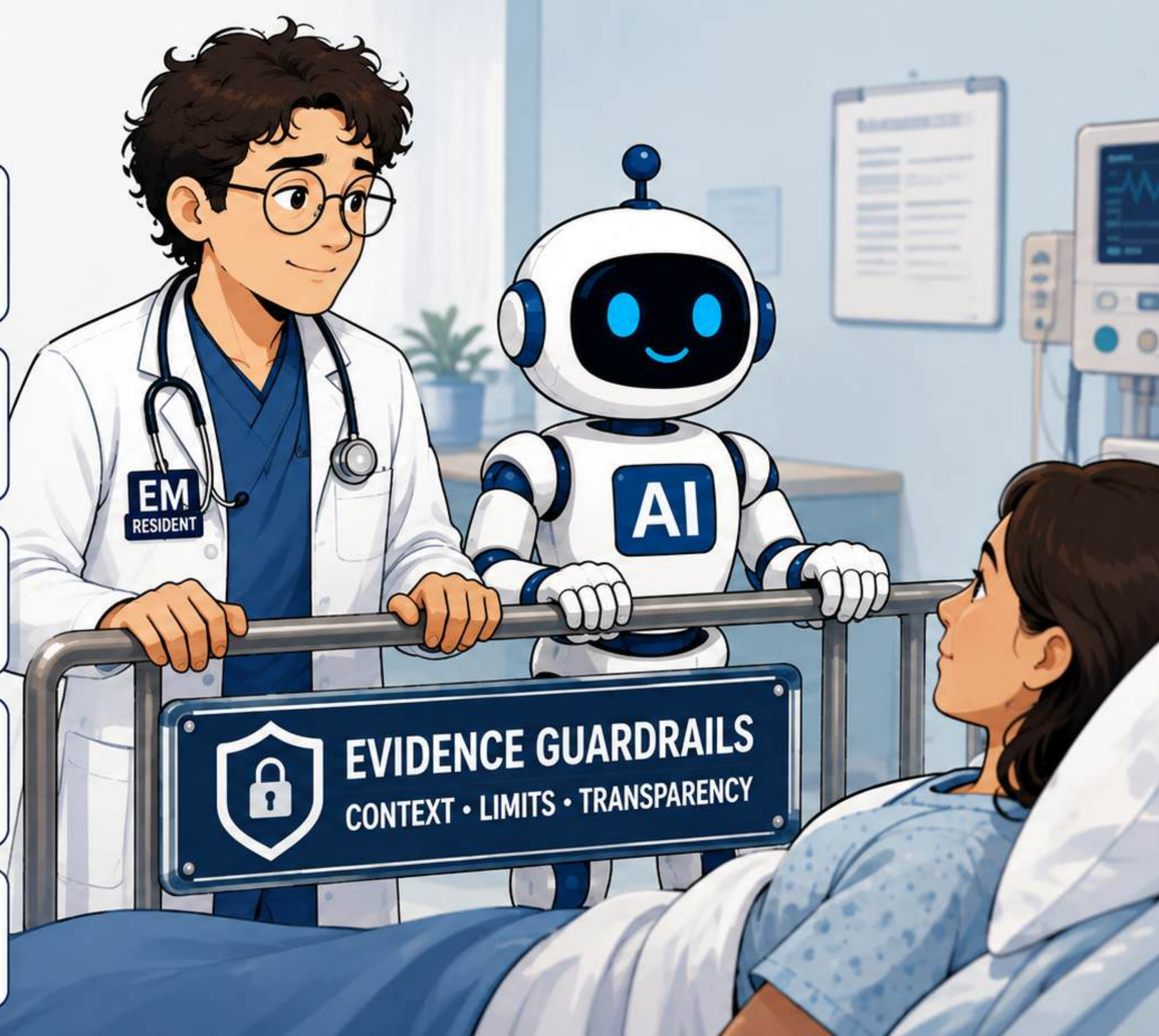
ABSOLUTE RISK?

OUTCOME DEFINITION?

WHERE DID "5" COME FROM?

TAKE-HOME

- AI ACCELERATES FLOW—NOT TRUTH
- GUARDRAILS PRESERVE MEANING
- MEASURE TIME TO TRUST
- HUMANS OWN THE DECISION
- BE SKEPTICAL



QUESTIONS

WHICH LEAK WORRIES YOU MOST?

> *Be skeptical of anything you hear—even if you heard it from me.*