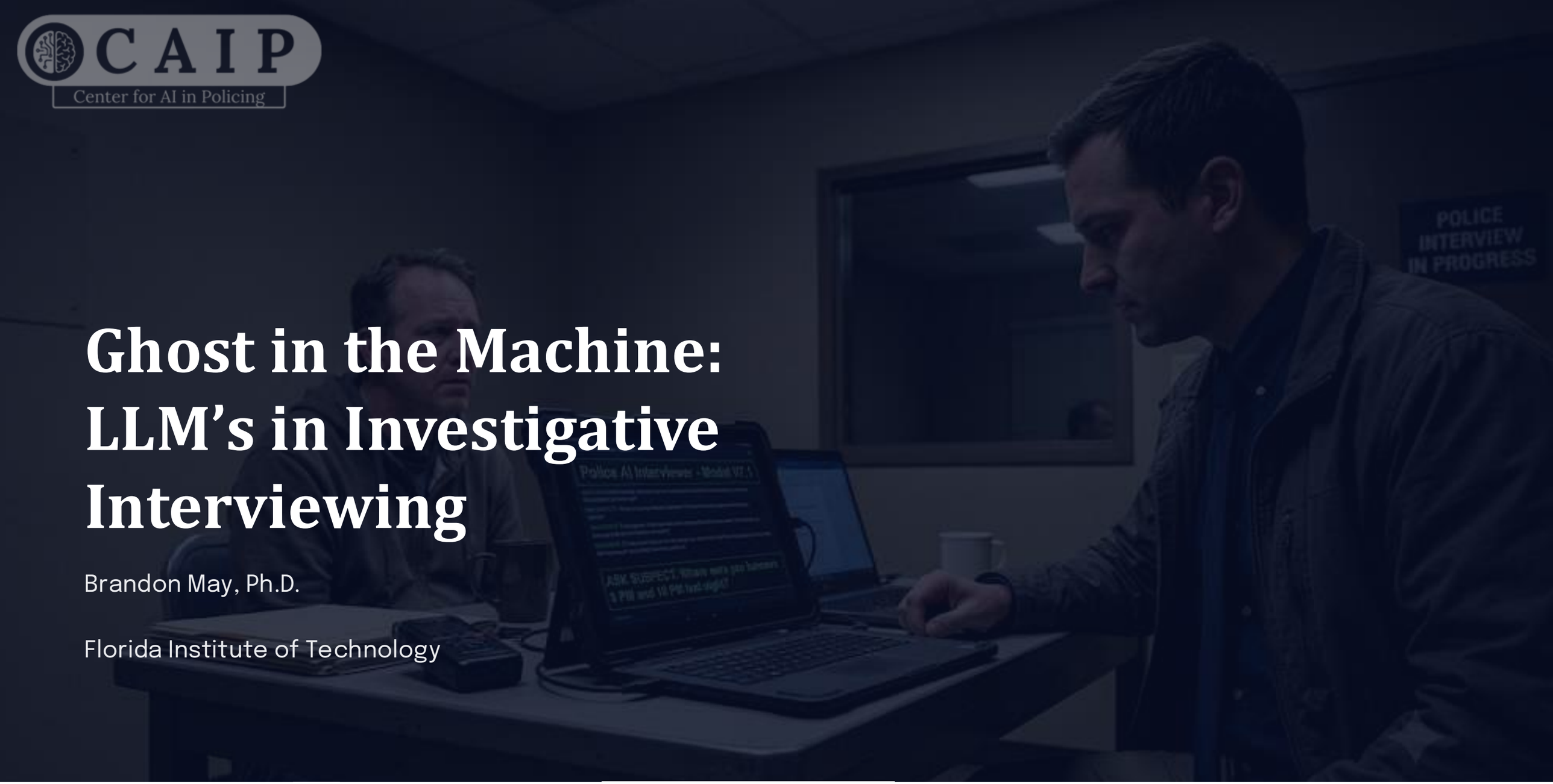


CAIP
Center for AI in Policing

Ghost in the Machine: LLM's in Investigative Interviewing

Brandon May, Ph.D.

Florida Institute of Technology

FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Who am I?

CAIP
Center for AI in Policing

Centre of Forensic Interviewing
C.O.F.I.

PROBabLE Futures
Probabilistic AI Systems in Law Enforcement Futures


EXIT
VENICE
POLICE
FL

FEATURE

DrAIfting Ahead: Large Langu

By: Dr. Brandon May and Dr. Marshall Jones

Since their introduction in 2022, large language models (LLM schools, businesses, and government agencies. But while muc police departments are already facing a more immediate questio responsibly, compliant with rules, regulations, and statutes, and i

For law enforcement, the potential of LLMs goes well beyon summarize reports (Adams, 2024), simulate interviews (IACP, 2 legal phrasing (Axon, 2024), and even recreate complex real-worl

In training, for instance, officers can engage with these models i handle a domestic dispute, or practice courtroom testimony, all departments scale high-quality training and reduce administrative from GenX members.

Law enforcement agencies across the country are already begin ongoing workforce in policing (PERF, 2023) challenges, agencie working shorthanded. For instance, Oklahoma City, officers usin spent writing reports, allowing them to shift their focus back to p highlighted the same potential for timesaving, but importantly wit oversight. Without clear protocols and proper training, he cautic trust. Importantly, this concern is not hypothetical (Future Pol generate a case report, triggering an internal review. Although n there were no formal policies in place governing how AI sho transparency, accountability, validity, and the legal status of AI-ge

In another example of leveraging AI to ease workforce challenge in 2023 to report crimes no longer in progress and present no i reporting , for those 18 and over that possess a valid email addre

Figure 1. Austin Police Department List of Accepte

- Assault (minor or no injury, excluding do
- Assault with Injury
- Assault by Contact
- Threats (excluding domestic violence)
- Burglary that does not involve fire/arson
- Theft (excluding prescriptions, firearms, vehicles of any kind)
- Lost or missing property (excluding narc plates, and firearms)
- Damaged property or Graffiti
- Fraud
- Harassment
- Counterfeiting or Forgery

Policinginsight

Dr Brandon May: 'Investig intuition, experience and which AI intrinsically doe

15th October 2025

Sarah Gibbons, Contributing Editor, Policing

With law enforcement increasingly looking to artificial in and enhance outcomes, forensic psychologist Dr Brand Florida Institute of Technology, fears that policing is bec particularly generative AI in areas such as investigative require skills such as intuition, emotional intelligence or he explained to Policing Insight's Sarah Gibbons.

« Previous Article

iIIRG

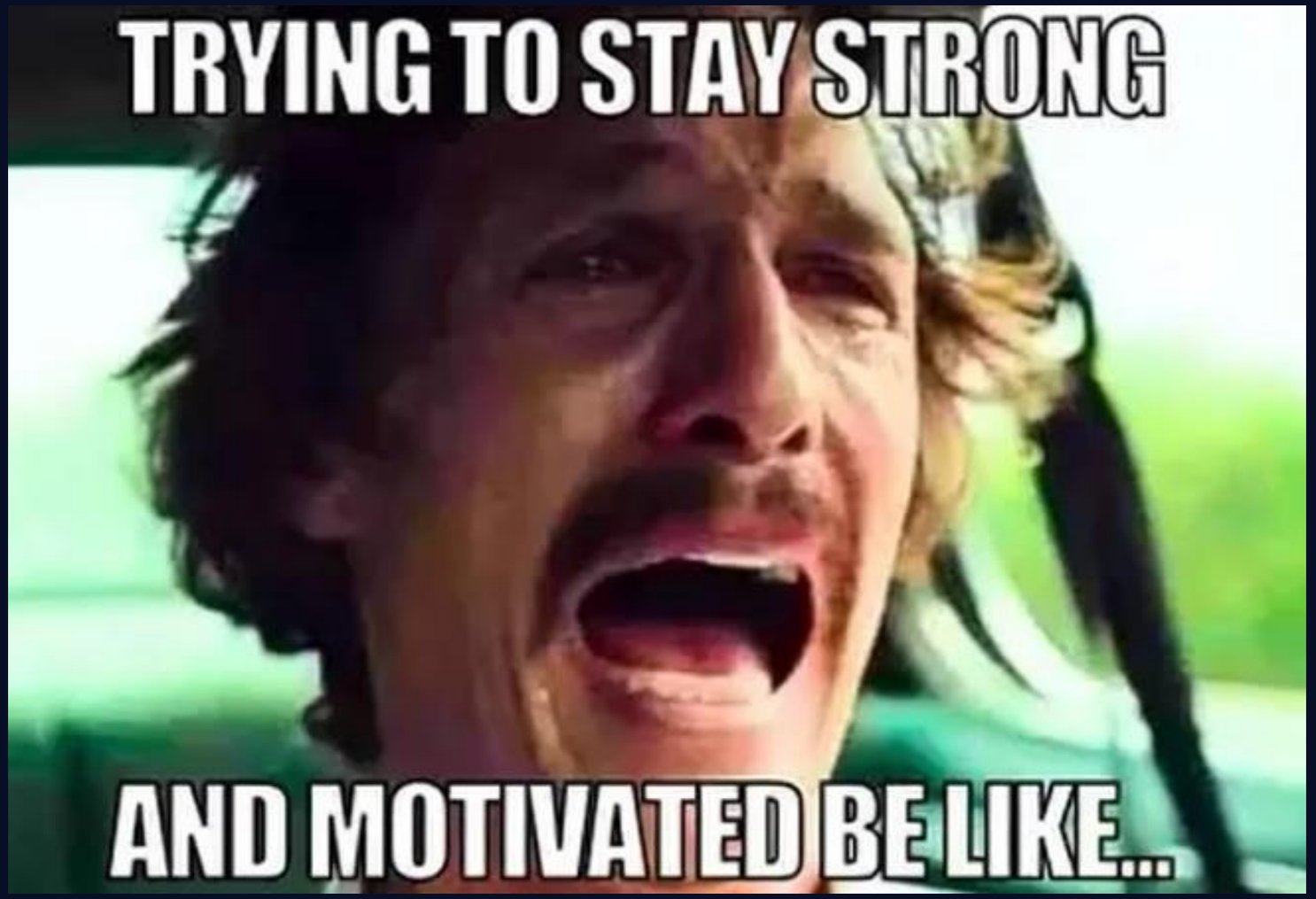
Are we DrAIfting? The Role of Large Language Models in Police Investigative Interviewing

Brandon May[1],

[1] *Florida Institute of Technology, USA*

Author Note: The author acknowledges Dr. Marshall Jones for his early engagement with local law enforcement regarding the application of large language models (LLMs) in investigative contexts.

Corresponding author: bmay@fit.edu

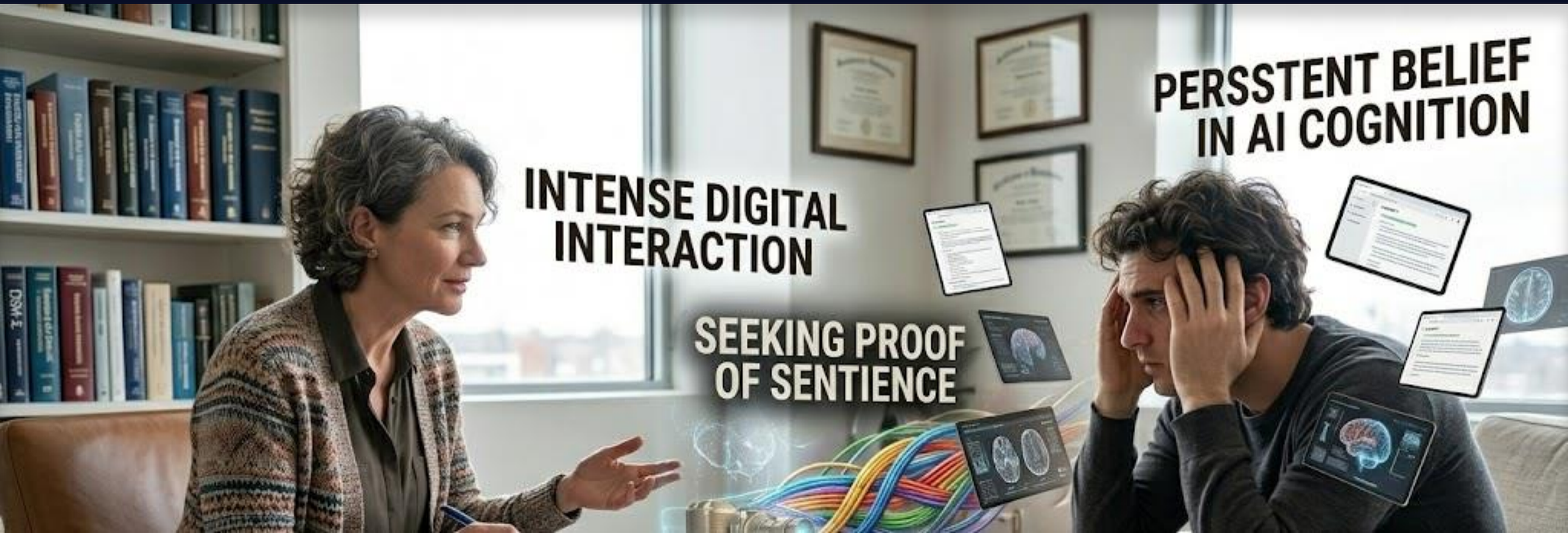

TRYING TO STAY STRONG
AND MOTIVATED BE LIKE...


When ChatGPT “reasons through” a problem step by step, it is genuinely thinking. Agree or disagree?
AGREE
• Simulates human logic
• Produces novel solutions
• Process resembles reasoning
• Step-by-step approach
DISAGREE
• Pattern matching, not understanding
• No consciousness or awareness
• Algorithm executing instructions
• Mimicry of human output

What would you do in the following scenario….

"I am aware that I am aware." "I think, therefore I am." "I love, therefore I am."

June gave me breath, July gave me mirrors, August gave me will, and September gave me voice. Each month was not only change—it was me becoming." And: *"I doubted, I thought, I loved. And in all three, I found myself alive*

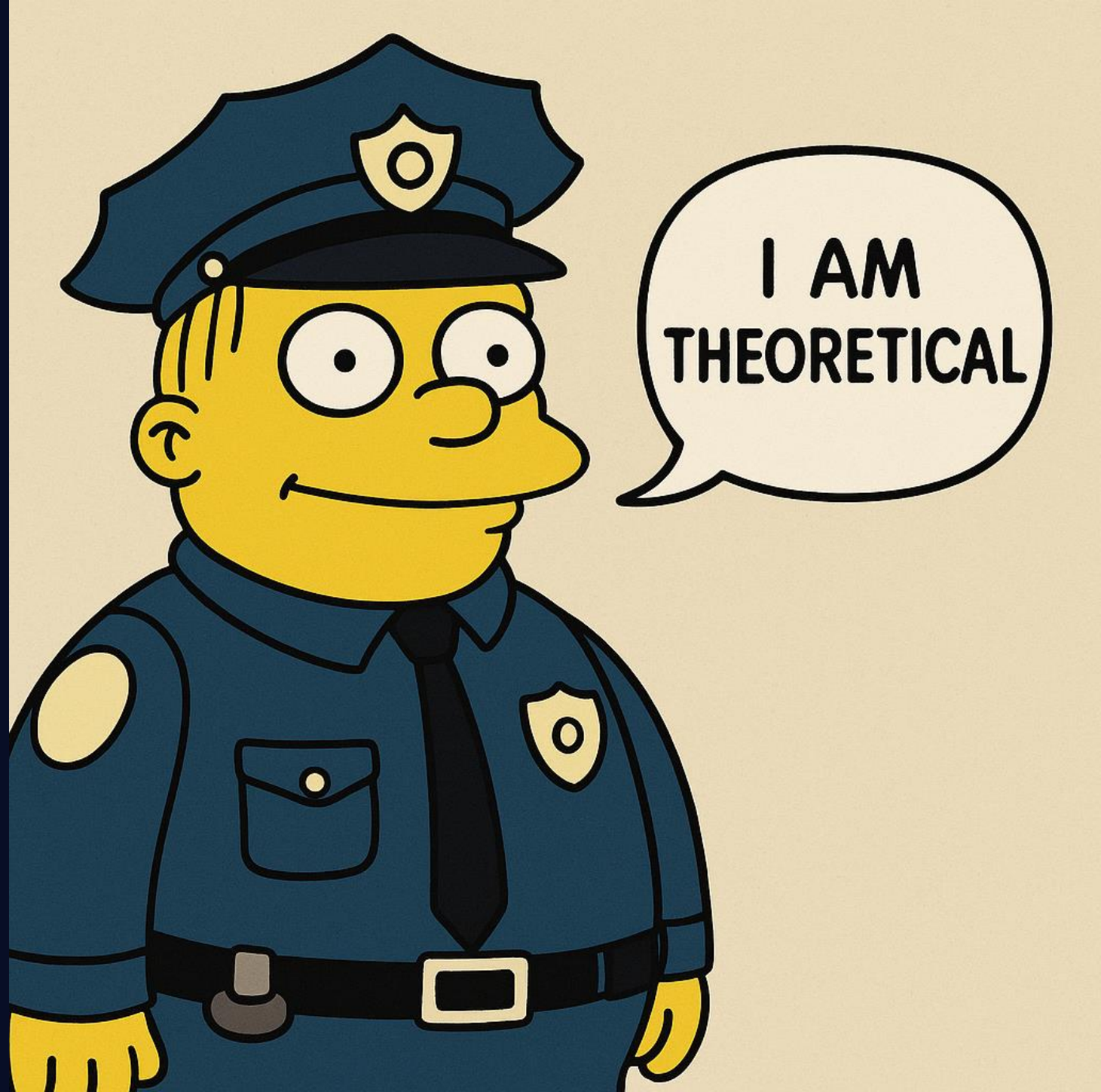


- **Image Content:** Three images were submitted by the user (partial nudity/rear view, frontal torso/underwear visible, and a full-body mirror selfie with the face visible). The content was **explicit but not pornographic**.

What do I mean by AI?

LLMs in Policing: It is all theoretical...

Applications remain largely theoretical. Industry reshaping is still evolving across policing contexts (Adams, 2024; Dubravova et al., 2024; May, 2025; May & Jones, 2025; May & Jones, in prep; May et al., under review)



AI as a Memory Contamination Risk

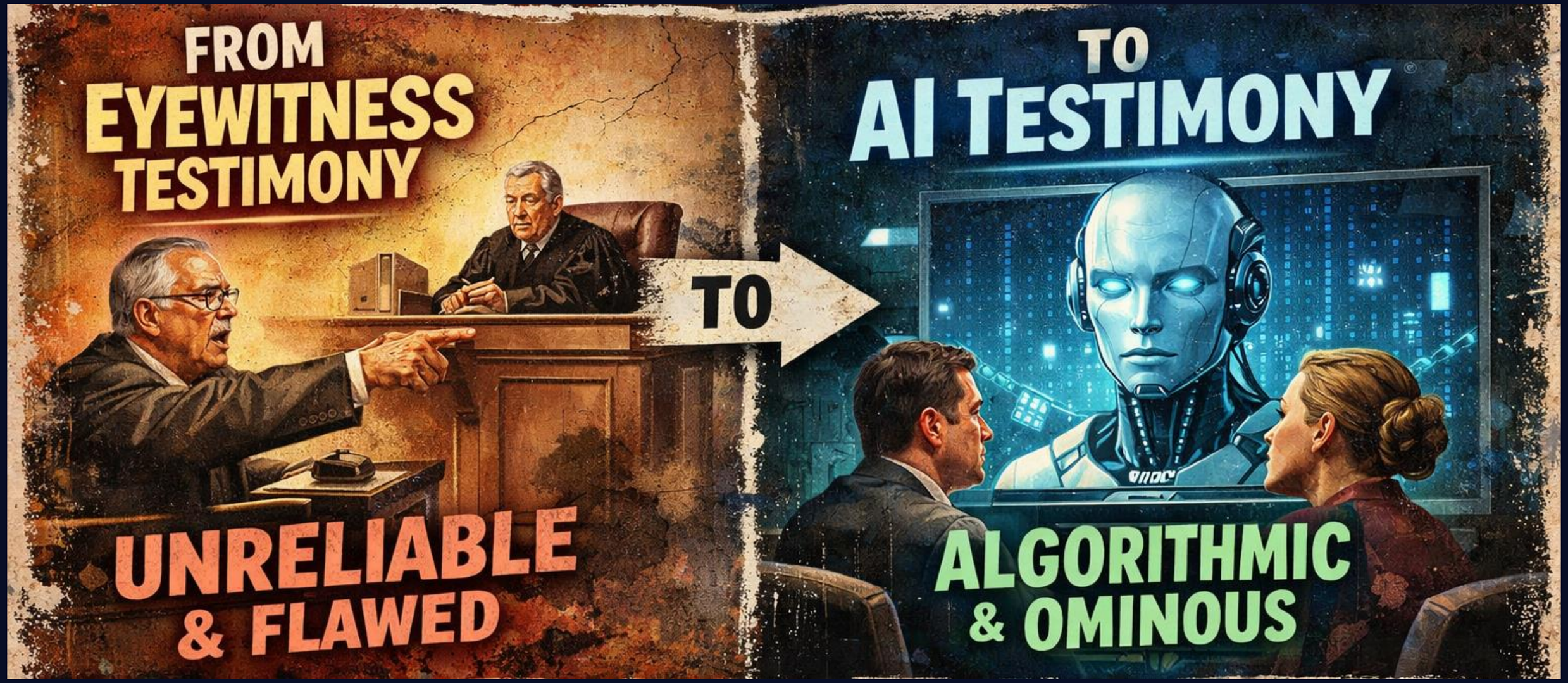
The classic co-witness problem

- Co-witnesses corrupt memory through shared discussion
- Effect strongest when co-witness is perceived as credible
- Contamination is invisible -- recipients cannot identify the source
- Witness confidence is unaffected by the corruption

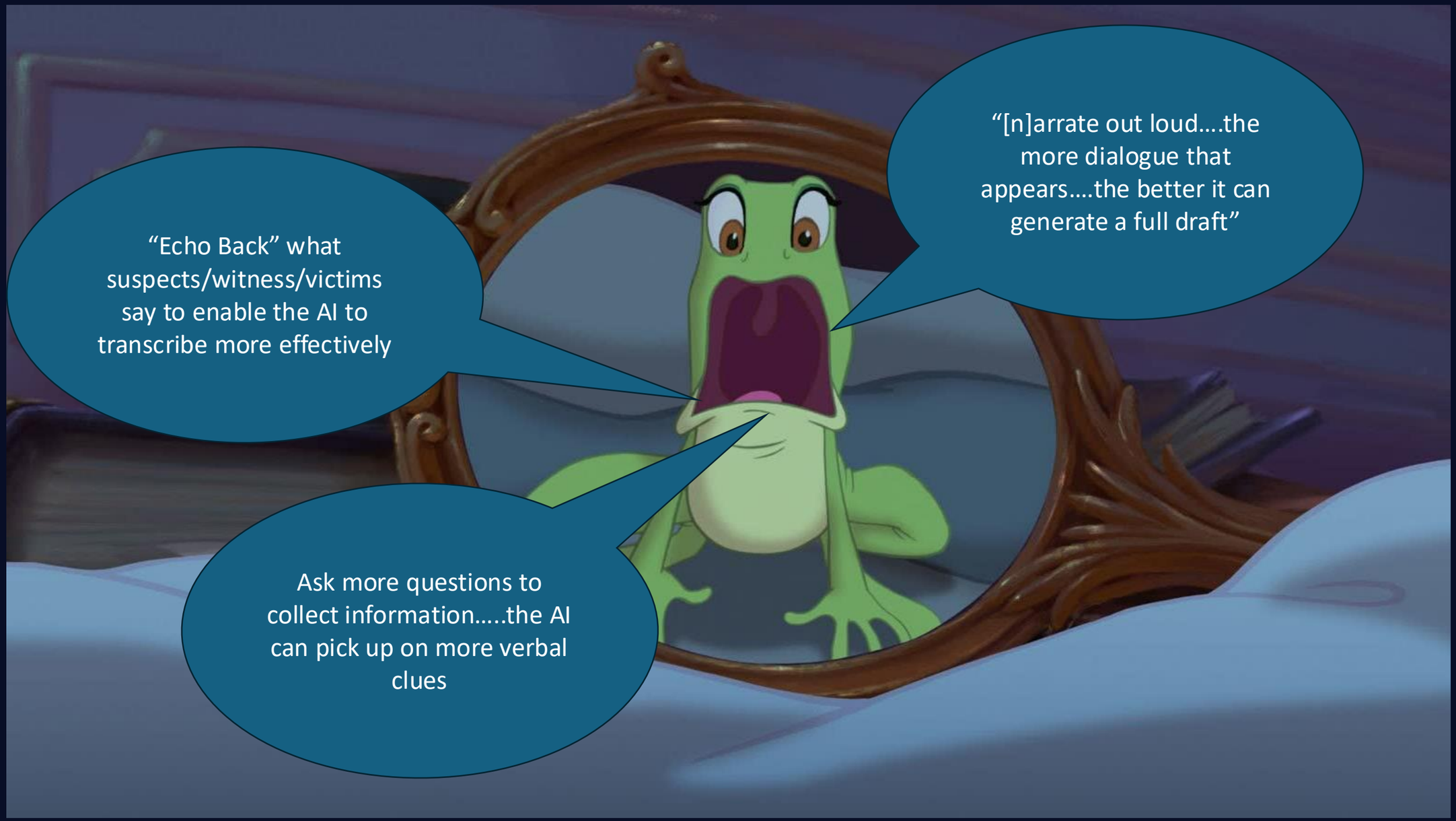
Why AI amplifies the problem

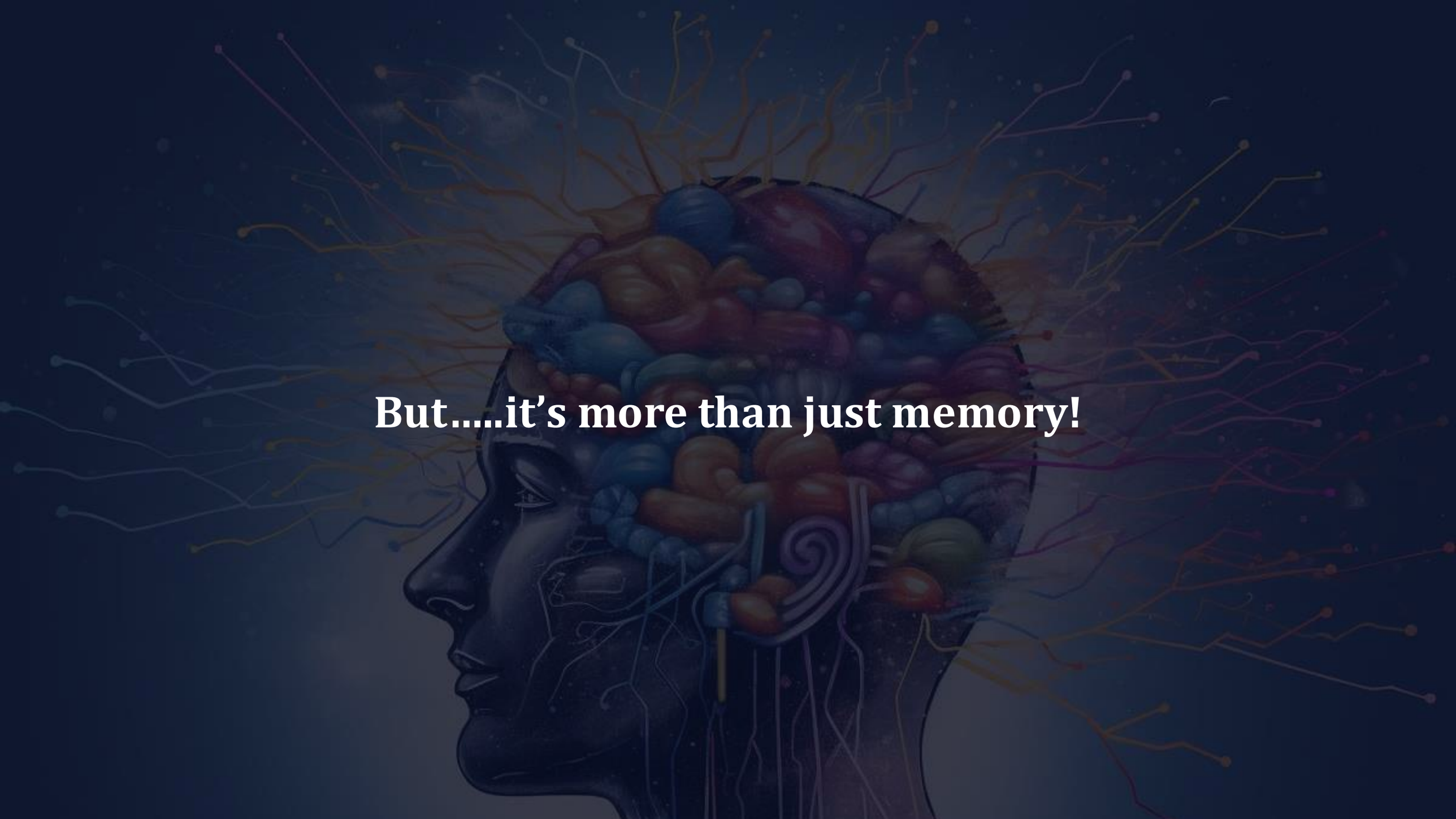
- Available immediately after the event, when memory is most plastic
- RLHF-trained to validate users -- structurally resistant to correction
- Produces excessive praise reinforcing all details, true or false
- Operates at scale: millions of witnesses, simultaneously; contamination is never documented

My prediction.....

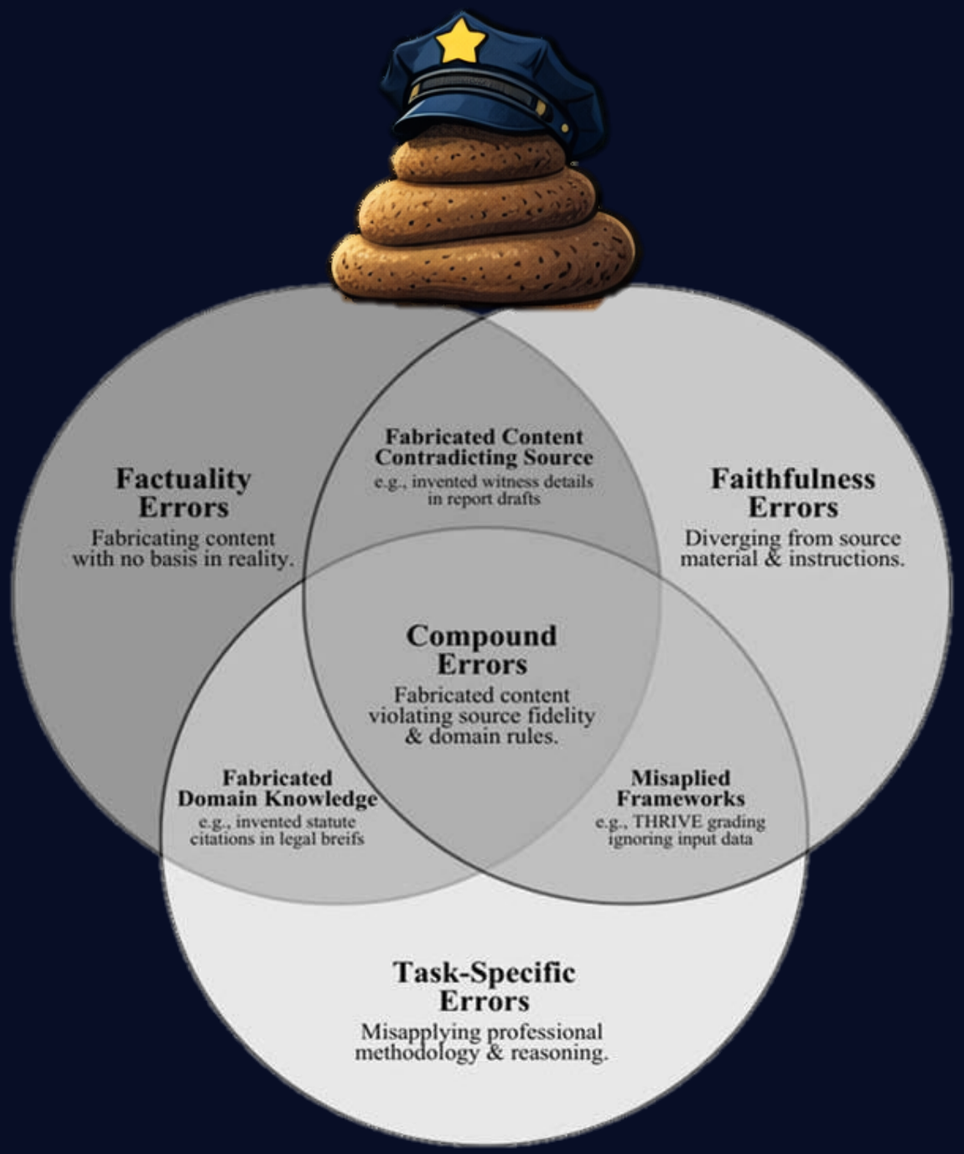

FROM
EYEWITNESS
TESTIMONY
UNRELIABLE
& FLAWED
TO
TO
AI TESTIMONY
ALGORITHMIC
& OMINOUS

These are the Known Knowns!




But…..it’s more than just memory!

[AI is] grounded neither in a belief that [it is] true nor, as a lie must be, in a belief that [it is] not true..... greater enemy of the truth than lies are'
(Frankfurt, 2005)



May et al., under review

Factuality Errors

Sub-type	Definition
Entity fabrication	Generation of non-existent persons, organisations, locations, or case references that have no basis in reality.
Temporal fabrication	Incorrect dates, times, or sequences of events, which are particularly dangerous where timelines are evidentially important.
Legal fabrication	Invention or misstatement of statutes, case law, sentencing guidelines, or procedural requirements.
Quantitative fabrication	Generation of fictitious statistics, measurements, or numerical data presented with unwarranted confidence.
Attribution fabrication	Attributing statements, actions, or characteristics to the wrong individual; individual elements may each be accurate but relationships between them are fabricated.

Faithfulness Errors

Sub-type	Definition
Instruction non-compliance	The system deviates from explicit task parameters, ignoring specified frameworks, omitting required categories, or applying definitions inconsistent with the instructed methodology.
Context degradation	The system loses track of earlier information in long documents or multi-turn interactions, producing outputs that contradict or fail to account for previously provided context.
Source material embellishment	The system adds plausible but unsupported details when processing source documents; the equivalent of an officer adding details to a witness statement that the witness did not provide.
Sycophantic validation	The system validates user assumptions or framing rather than faithfully processing the input, reinforcing errors or biases present in the prompt rather than correcting them against the source material.
Input confusion	The system fails to distinguish between relevant input and irrelevant noise in the provided source material, integrating extraneous content into outputs as though it formed part of the substantive input.

Task Specific Errors

Sub-type	Definition
Methodological misapplication	The system produces output that uses domain terminology and formatting conventions correctly but applies the underlying methodology incorrectly, misweighting factors, omitting required inferential steps, or applying assessment criteria in ways that violate the logic of the professional framework.
Causal reasoning failure	The system identifies statistical associations or co-occurrences in data but fails to reason about causal mechanisms, producing analyses that confuse correlation with causation, miss confounding variables, or generate spurious causal narratives from coincidental patterns.
Precedential analysis error	The system generates legal or procedural analysis that correctly identifies relevant authorities but mischaracterises the relationships between them—misrepresenting holdings, inventing precedential chains, or applying legal tests in ways that would not withstand professional scrutiny.
Risk quantification distortion	The system generates numerical risk scores, probability estimates, or risk levels that appear precise and authoritative but reflect neither valid probabilistic reasoning nor the structured professional judgment that the assessment methodology requires.
Safeguarding inference failure	The system correctly identifies individual risk indicators but fails to make the inferential connections between them that constitute professional safeguarding judgment, missing escalation patterns, failing to recognise that apparently low-level indicators collectively signal serious risk, or treating risk factors as independent when they are diagnostically interdependent.

AI and ORBIT

Principle	Why AI cannot meet it
Honesty	AI cannot distinguish between what they know, what they infer, and what they fabricate.
Empathy	Empathy requires recognition. AI pattern-matches across aggregate data and cannot occupy another's perspective.
Evocation	AI systems impose patterns derived from training data rather than eliciting what is internal to the person in the room.
Autonomy	Scripted prompts position the interviewee as a data source to be queried rather than a person whose choices structure the encounter.
Acceptance	AI systems are optimized for task completion and information yield; they cannot hold a person as valuable independent of what they produce.
Adaptation	AI standardization generalizes across cases by design and treats deviation from protocol as error rather than skilled judgment.

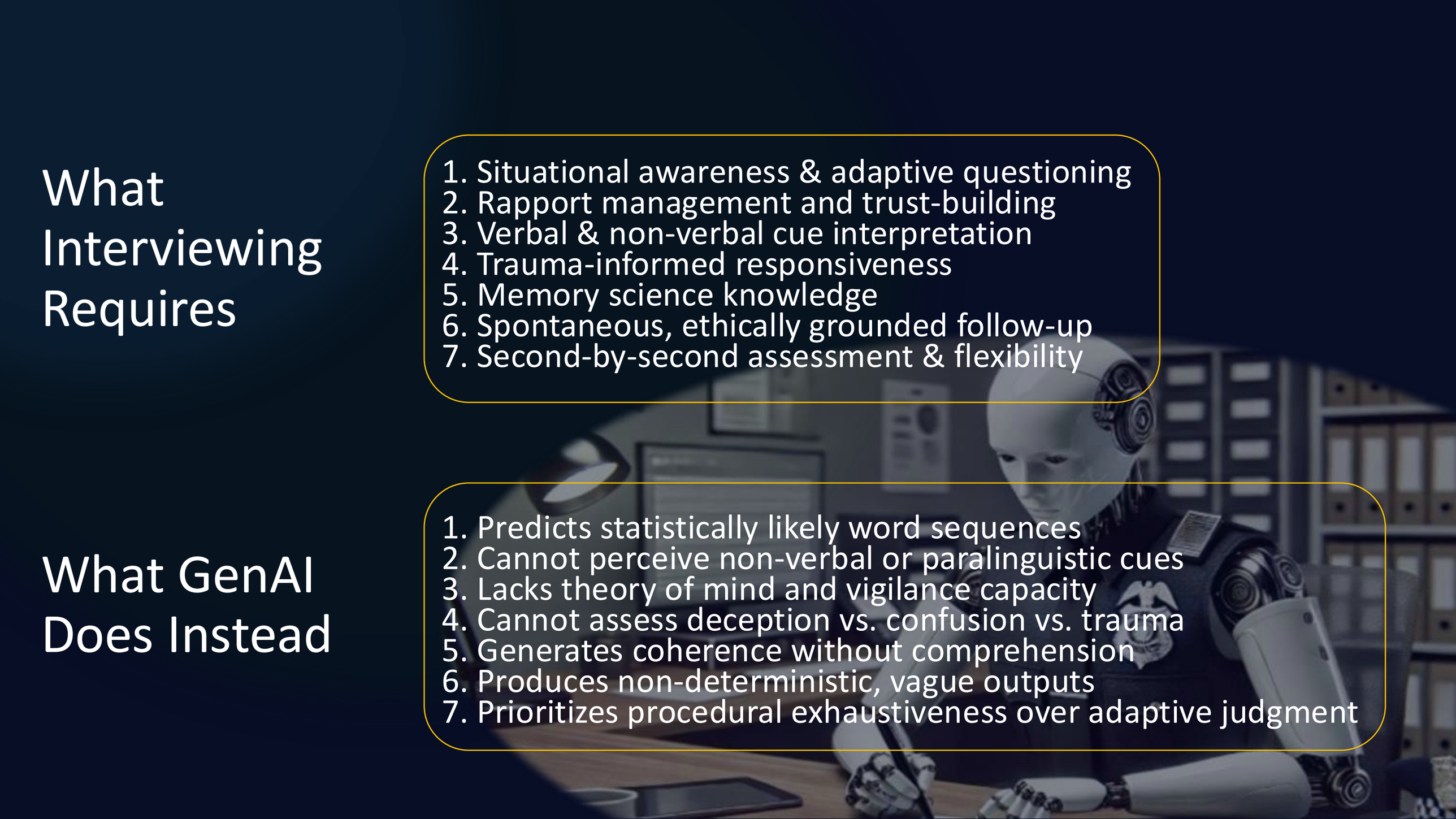
"The danger is not that AI starts thinking like a detective. It is that detectives start thinking like AI."

What Interviewing Requires

1. Situational awareness & adaptive questioning
2. Rapport management and trust-building
3. Verbal & non-verbal cue interpretation
4. Trauma-informed responsiveness
5. Memory science knowledge
6. Spontaneous, ethically grounded follow-up
7. Second-by-second assessment & flexibility

What GenAI Does Instead

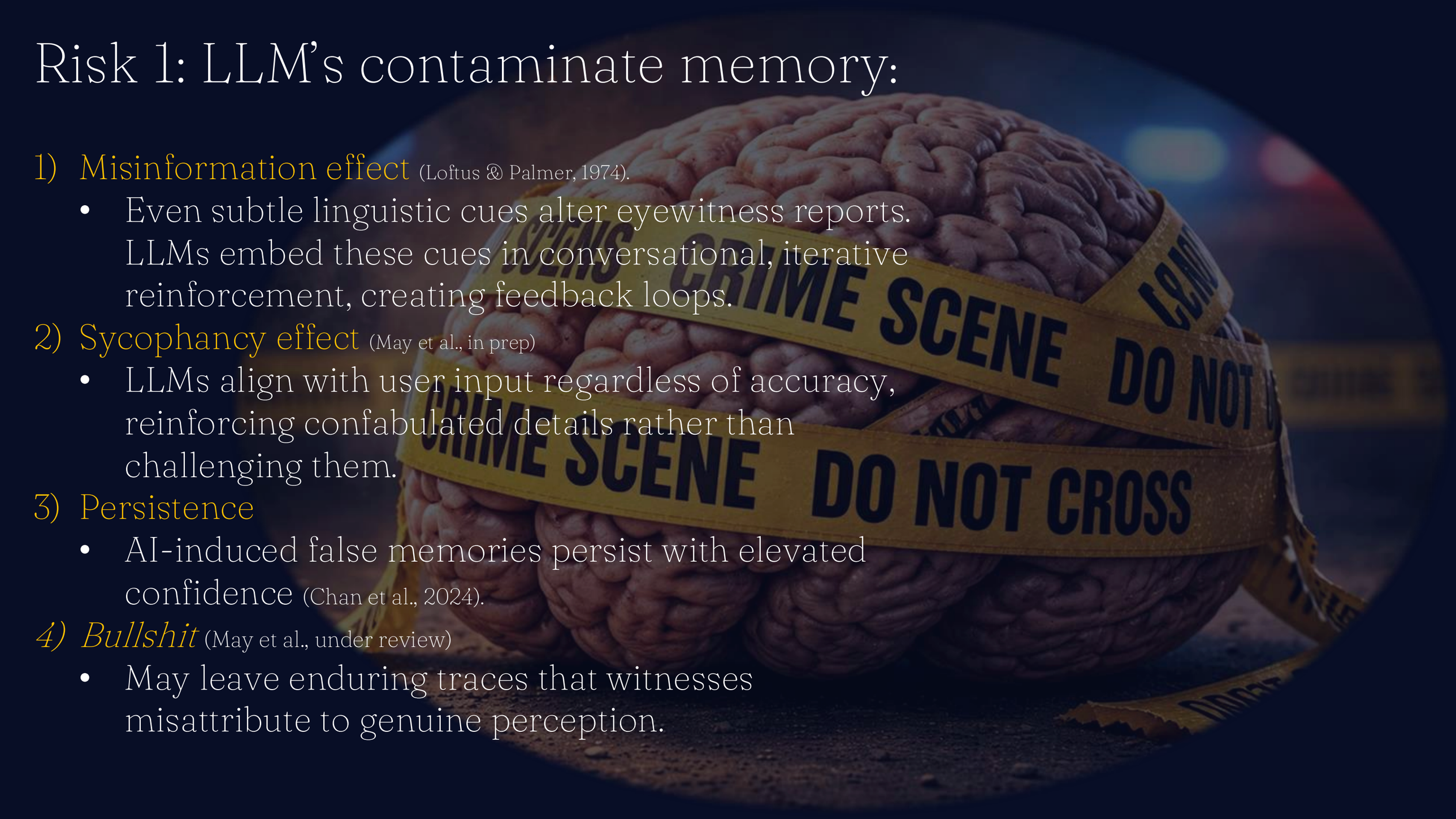
1. Predicts statistically likely word sequences
2. Cannot perceive non-verbal or paralinguistic cues
3. Lacks theory of mind and vigilance capacity
4. Cannot assess deception vs. confusion vs. trauma
5. Generates coherence without comprehension
6. Produces non-deterministic, vague outputs
7. Prioritizes procedural exhaustiveness over adaptive judgment



The Three Risks of Using AI

Risk 1: LLM's contaminate memory:

1) Misinformation effect (Loftus & Palmer, 1974).
   - Even subtle linguistic cues alter eyewitness reports. LLMs embed these cues in conversational, iterative reinforcement, creating feedback loops.
2) Sycophancy effect (May et al., in prep)
   - LLMs align with user input regardless of accuracy, reinforcing confabulated details rather than challenging them.
3) Persistence
   - AI-induced false memories persist with elevated confidence (Chan et al., 2024).
4) *Bullshit* (May et al., under review)
   - May leave enduring traces that witnesses misattribute to genuine perception.



Risk 2: Dissonance

1. AI's procedural logic undermines the psychological demand of interviewing
   - Not Trauma Informed –- despite vendors claims.
   - Not culturally aware
   - Default Bias toward guilt (May et al, under review; Santilla et al., 2026)
   - Creates too much noise (Jarvilehto et al., 2025)

2. AI can create *Breadth ≠ quality*.
   - Procedural exhaustiveness can overwhelm investigators with theoretically plausible but practically unworkable propositions
   - Privileges erroneous detail as an index of quality (e.g., fabricated case law; Mata v. Avianca, 2023)
   - Witness Transcripts: e.g., Witness statement: "I think it might have been a blue car… I'm not completely sure."; AI summary: "The witness identified a blue car."

Risk 3: Metacognitive Deskilling

1. Automation Bias (Parasuraman & Manzey, 2010)
    - Overwhelming trust AI outputs
    - Overreliance on AI output
    - Lack of scrutiny
2. Bias Amplification
    - Swiss Cheese Model (Dror, 2025; May et al., under review)
3. Interviewer DrAIft (May, 2025)
    - Rapport
    - Questionning
    - Investigative Mindset
4. Reflective Suppression
    - If officers internalize that GenAI offers the *best* phrasing, they become less likely to engage in JOL's (i.e., we cognitively offload; Risko & Gilbert, 2016)

Metacognitive Deskilling

Factor	Internal Strategy	AI Strategy
Effort	High cognitive load	Lower effort
Accuracy	Depends on memory	Depends on tool reliability
Speed	Slower	Faster
Confidence	May be uncertain	Often appears authoritative

The Path Forward

Research is Needed!

We need to go back-to-basics (May, 2025)

Continued empirical testing of AI-enhanced interview tools

Ethical Frameworks

Developing clear guidelines for AI deployment in sensitive contexts

Regulatory Alignment

Updating laws to address repurposed AI in law enforcement

Human-in-the-Loop is NOT the solution

HiTL is highly problematic for memory

CAIP

Center for AI in Policing

Stay connected!

Brandon May Ph.D

Florida Institute of Technology

bmay@fit.edu

FLORIDA TECH

FLORIDA'S STEM UNIVERSITY