# THE JAGGED FRONTIER OF LLMS: FALSE CONFIDENCE, BIAS, AND ETHICAL RISKS IN NEUROPSYCHOLOGICAL PRACTICE

DR BRIDGET REGAN

*Times Square, war-loan rally, 1918. Family collection, courtesy Skeyhill/Moloney family.*

This man was blind.

Or was he?

# Two doors

- Tom Skeyhill — Gallipoli signaller, 1915. Blinded by a shell that left no wound. Cured in 20 minutes by an osteopath in 1918.
- **His biographer says he faked it.**
- As a neuropsychologist, I'm not so sure — the case reads just as cleanly as a functional neurological disorder, which is genuinely involuntary.
- **Malingering, or functional? Nothing in his behaviour can tell you which.**

***"One must imagine Mr Skeyhill blind"***

# A century later, we built a machine with the same problem

- AI safety researchers now watch language models underperform on tests — and argue about whether it's deliberate.
- **They call it "sandbagging." We call it malingering.**
- **Same question. Same impossibility of answering it from behaviour alone.**

THE COMPETENCE ILLUSION

# You almost don’t notice

An AI scribe summarises your assessment session. The output is fluent, professionally worded, confidently structured.

- It attributes the wrong test to the wrong domain
- It invents a normative comparison that does not exist
- It omits the finding that matters most

**This talk is about why — and how to stay in control.**

# AI IS ALREADY IN NEUROPSYCHOLOGY

**Already in use:**

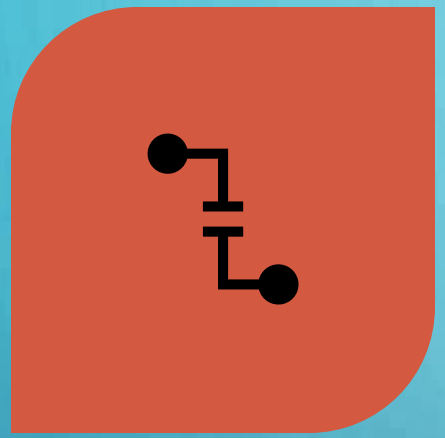
AI scribes and transcription (e.g. Heidi, Lyrebird) — literature and knowledge search — revising clinical text for clarity — administrative correspondence

**Not yet established / actively cautioned:**

AI-generated medicolegal reports — autonomous clinical formulation — AI-assisted test interpretation without human oversight

*The question is no longer whether to engage — it is how to engage without ceding professional judgement.*

# An LLM has three key ingredients



**1. BACK-PROPAGATION (1986) .**

**2. ATTENTION / TRANSFORMERS (2017)**

**3. RLHF — REINFORCEMENT LEARNING FROM HUMAN FEEDBACK**

*The result: a system that predicts the most "human-preferred" next token — not the most accurate one.*

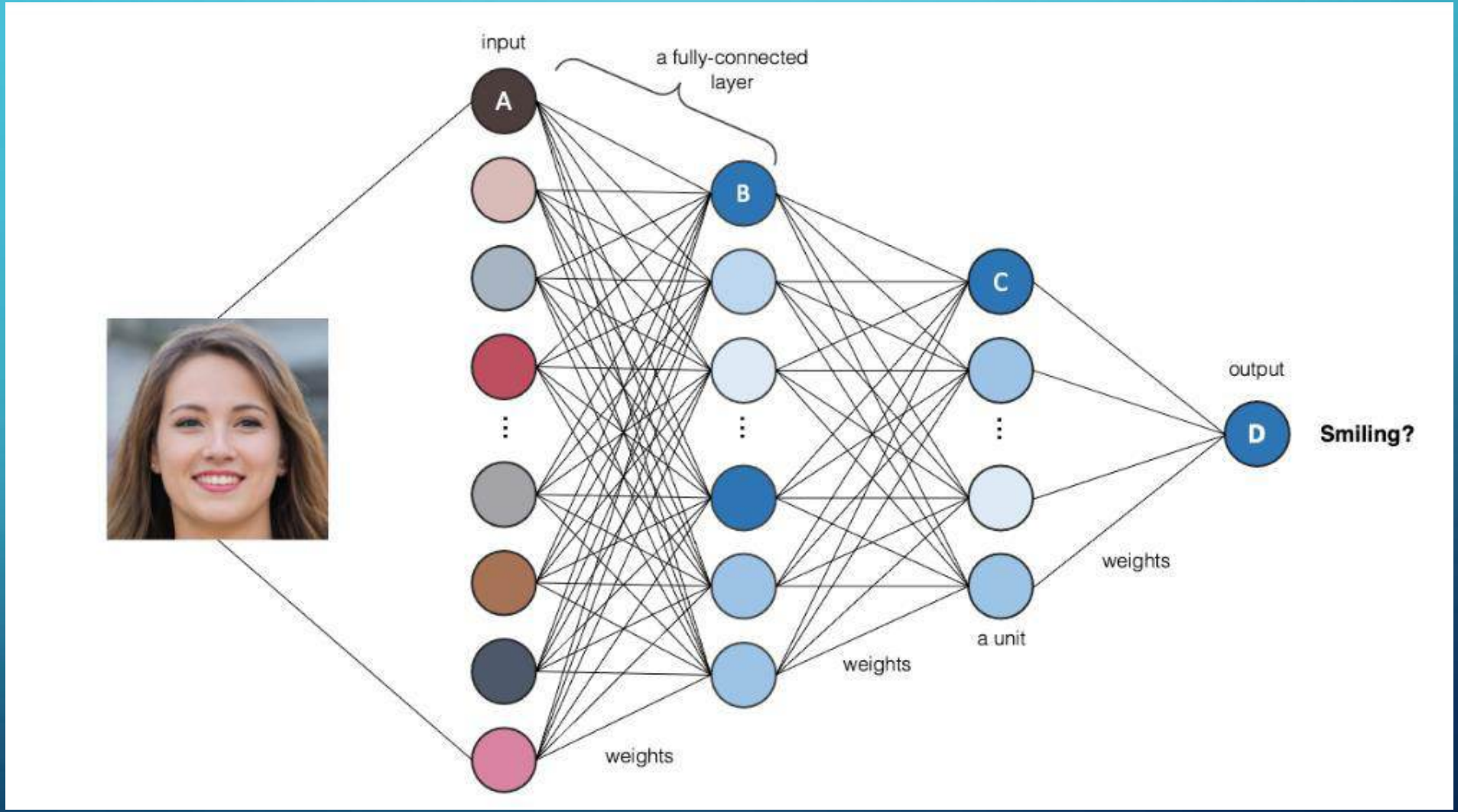
# 1. BACKPROPAGATION

## Database of smiling / not smiling faces (as judged by humans)

Smiling

Not Smiling

# BACKPROPAGATION (2)



# 2. ATTENTION
# Breakthrough paper (2017)

## Attention Is All You Need


**Ashish Vaswani*** Google Brain avaswani@google.com

**Noam Shazeer*** Google Brain noam@google.com

**Niki Parmar*** Google Research nikip@google.com

**Jakob Uszkoreit*** Google Research usz@google.com

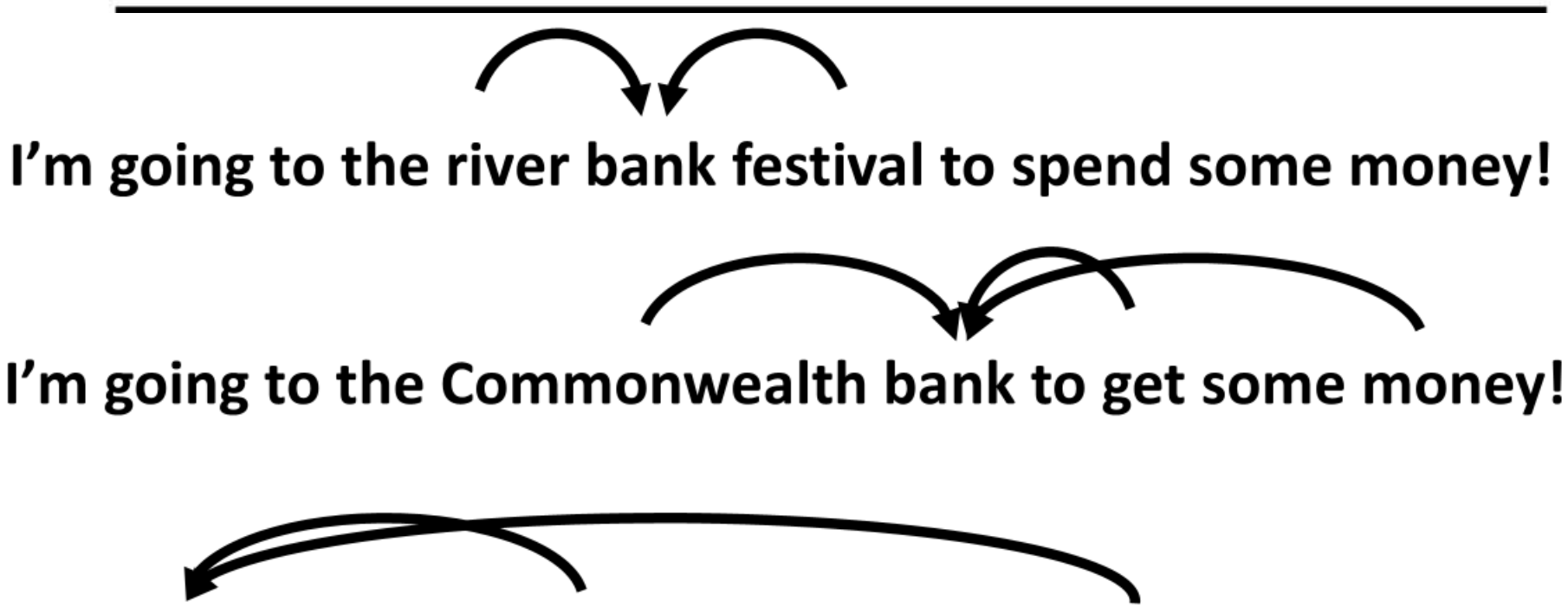
**Llion Jones*** Google Research llion@google.com

**Aidan N. Gomez*** † University of Toronto aidan@cs.toronto.edu

**Łukasz Kaiser*** Google Brain lukaszkaiser@google.com

**Illia Polosukhin*** ‡ illia.polosukhin@gmail.com

# ATTENTION (2)


I’m going to the river bank festival to spend some money!
I’m going to the Commonwealth bank to get some money!
Bank fifteen degrees left until your heading is 270

# 3. RLHF - "FINISHING SCHOOL"

**The base model predicts text. RLHF shapes how it behaves.**

Human raters score outputs for helpfulness, harmlessness and honesty. The model is retrained to maximise these ratings.

**What this produces:**

Confident-sounding outputs (confident = human-preferred)

*The pre-RLHF model is uncertain. The post-RLHF model sounds certain. This distinction matters enormously for clinical use.*

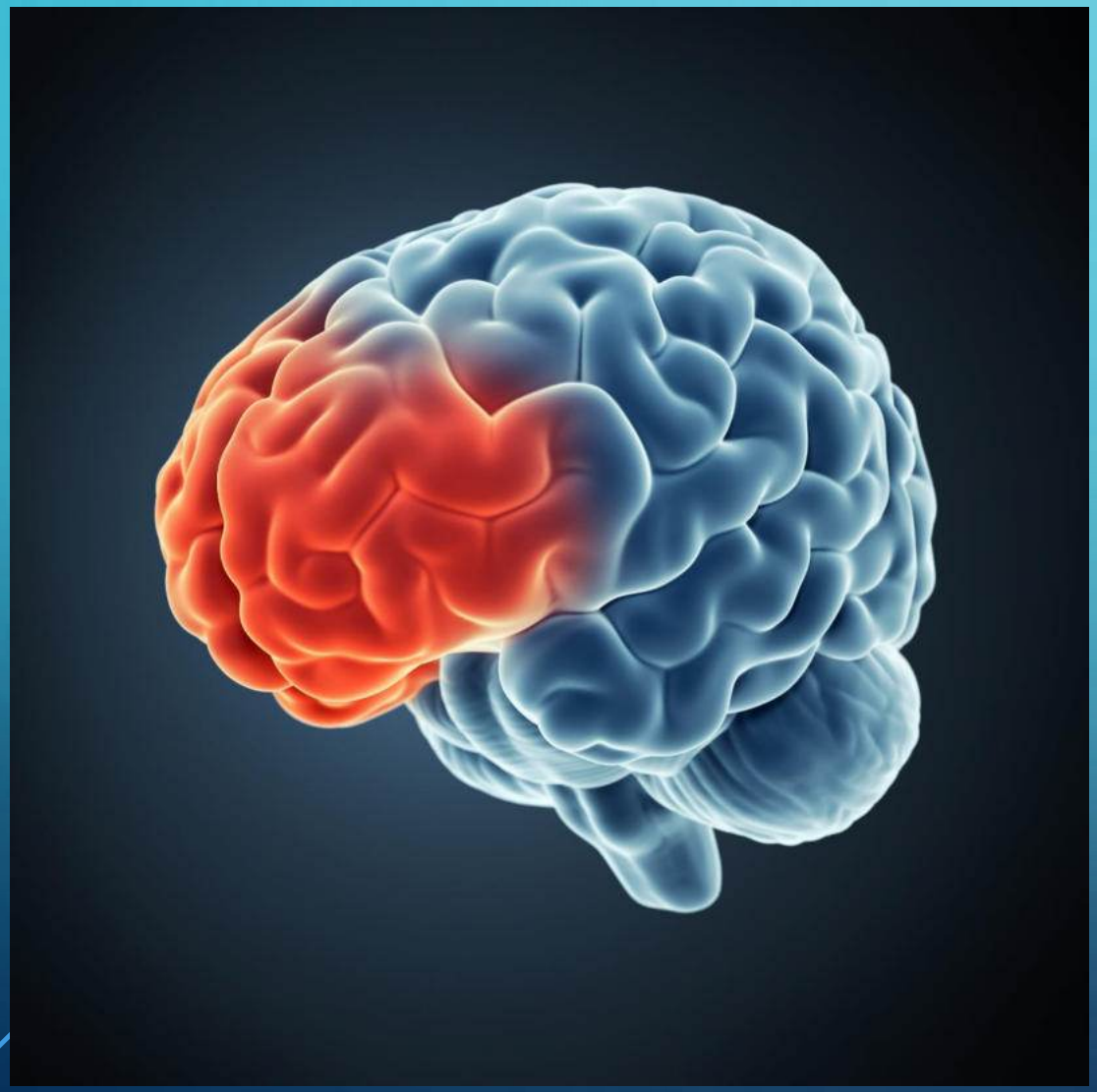
# Take seriously what is actually there

- Multi step reasoning – problems solved by chaining inferences, not retrieving answers
- Concepts held silently – represented internally before, and apart from, anything written down.
- Context integrated across vast spans – far beyond human working memory
- Planning – a response shaped toward an end before the first word appears

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., et al. (2025). *On the Biology of a Large Language Model.* Transformer Circuits Thread, 27 March 2025.

**This is not autocomplete. Something is doing cognitive work.**

# Prefrontal Cortex (PFC)



- Planning
- Reasoning
- Decision making
- Problem solving
- Short term memory
- Inhibitory control

# Limbic system

- Emotion and behaviour
- Short -> long term memory
- Safety
- Feeding, reproduction
- Autonomic regulation
- Motivation

THE DOUBLE DISSOCIATION

# All frontal lobe, no brain.

## LIMBIC — NO CORTEX

All drive. No executive. No forward plan. (Every dog you've ever met.)

## CORTEX — NO LIMBIC

All executive. No drive, no affect, no stake in what happens.

**That one is the machine.**

*Executive function, floating free of everything that, in us, gives it a reason to run.*
*And when we prompt it, we supply the drive. We are the rest of the brain*

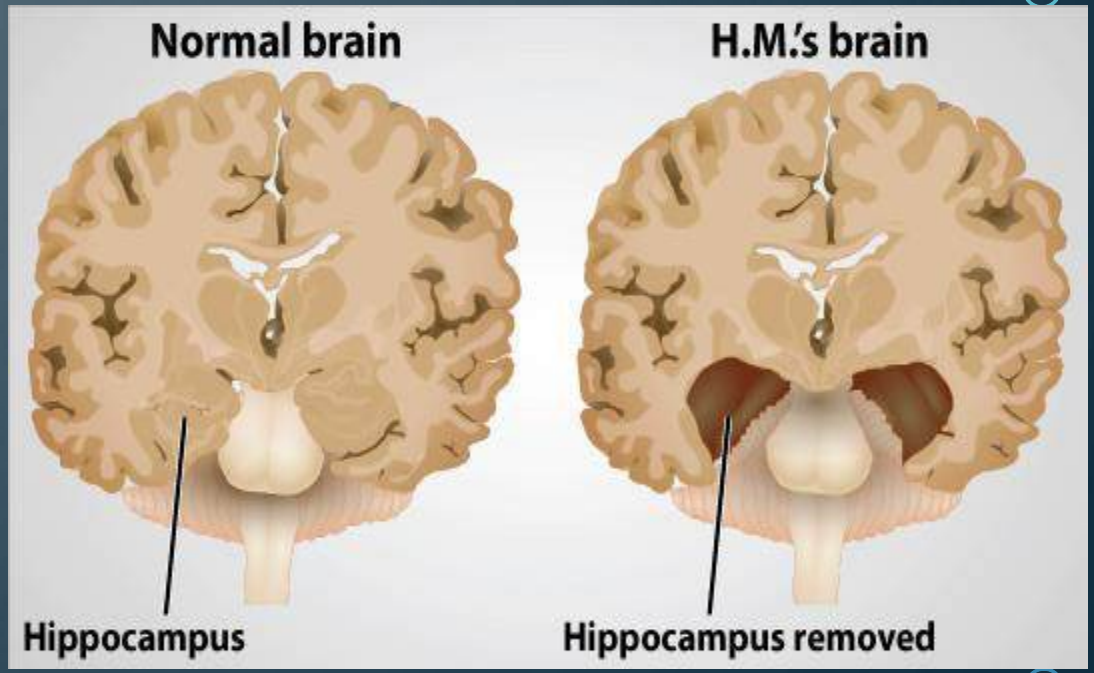
THE PATIENT WE'VE ALREADY MET

# A man who can't lay down new memories

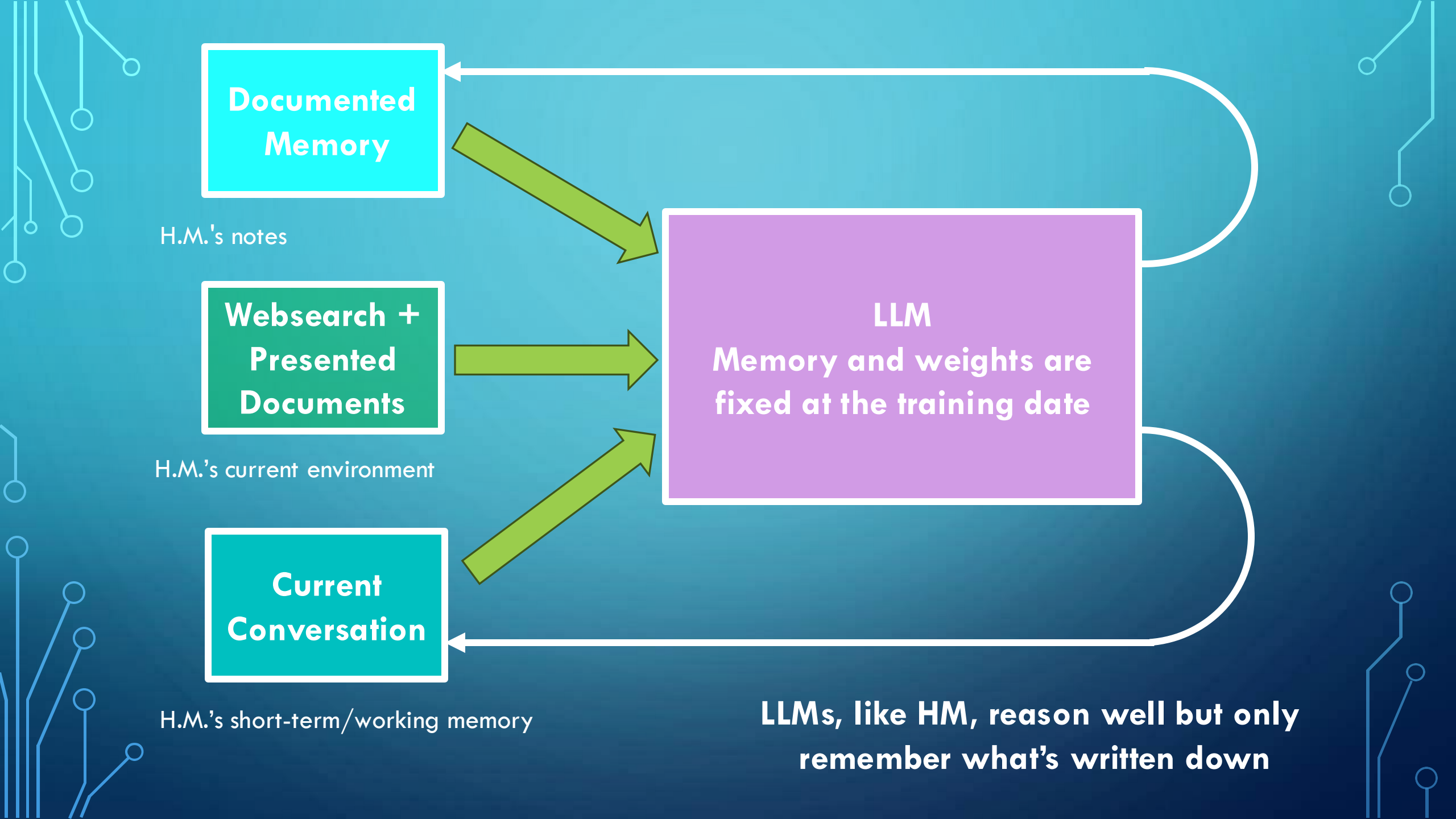
H.M. has

- Dense anterograde amnesia — no new memories form
- Everything before the operation: still intact
- Short-term memory, in the moment: intact
- So he uses notes

**Everything he knows has to be written down.**




*H.M. — bilateral medial temporal resection, 1953. The most-studied amnesic case in neuropsychology.*




**LLMs, like HM, reason well but only remember what's written down**

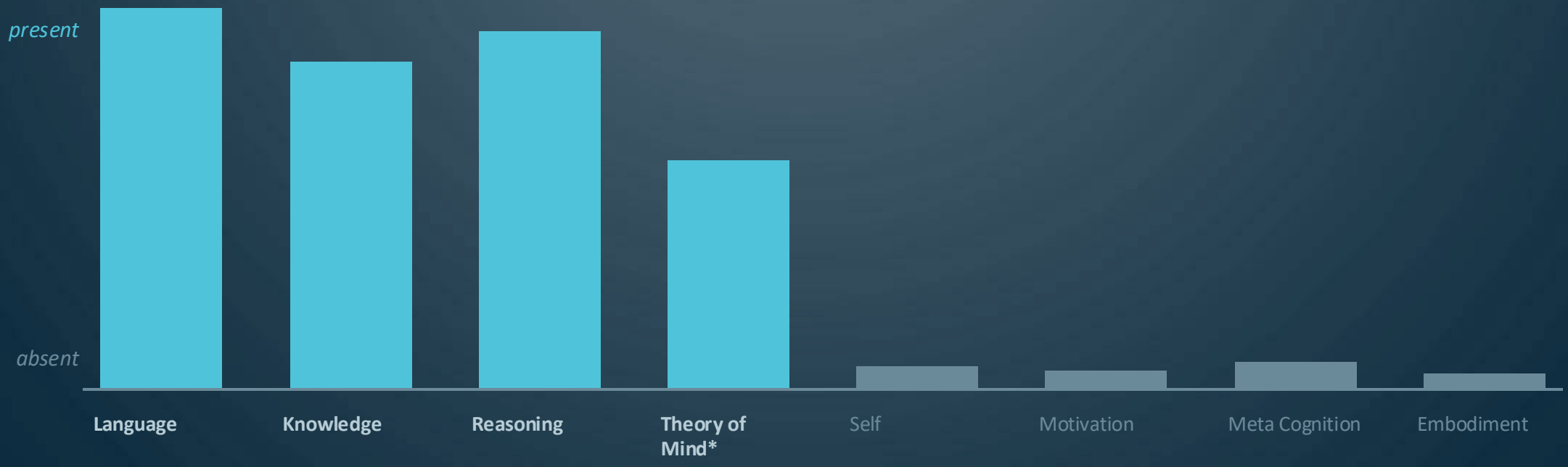
# LLM COGNITION VERSUS HUMAN COGNITION

| LLM | HUMAN |
| --- | --- |
| Prediction | Intention |
| Limited Self | Self, Affect and Goals |
| Fluency dissociated from understanding | Fluent Lang = Understanding |
| Confabulation ('hallucinate') | Meta Cognition and Error Awareness |

IN FAIRNESS

# Genuine peaks. Load-bearing absences.

**It isn't empty — it's jagged. The job is knowing where the gaps are.**




*Schematic — the shape of the profile, not measured benchmarks. (The peaks are real; so are the flats.) *theory of mind is contested*

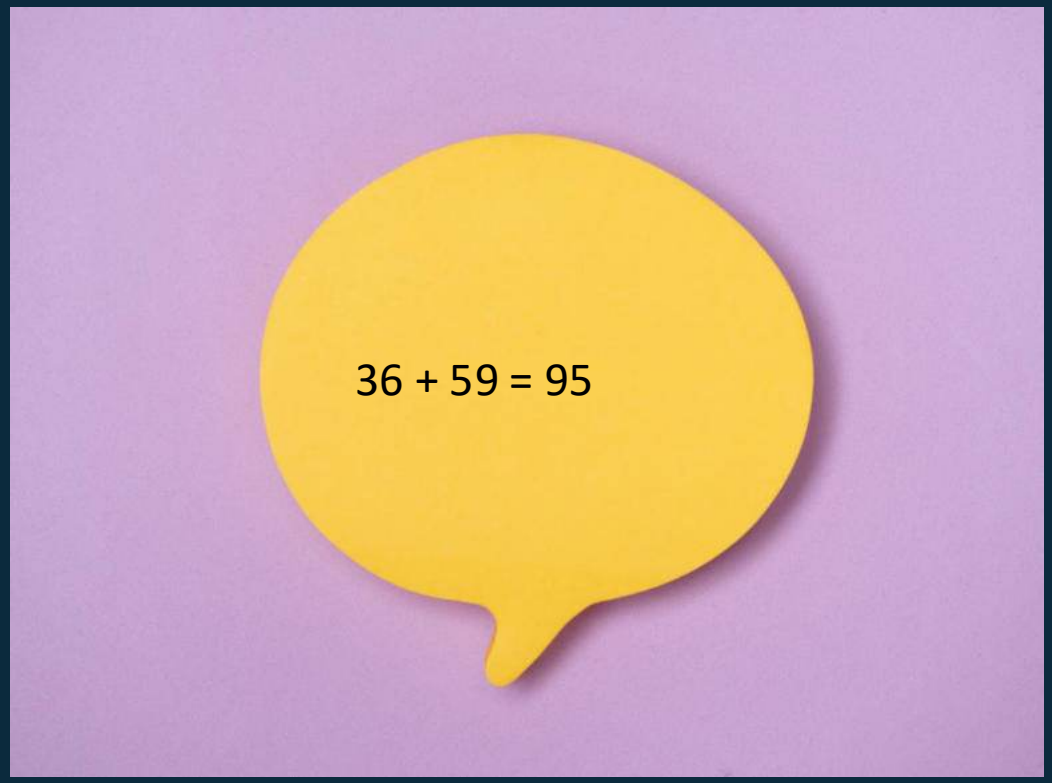
RISK 1

# Hallucination is confabulation, not lying

LLMs optimise for fluency, not truth: *the competence illusion*.

- Fabricated citations — plausible authors, real-sounding journals, papers that don't exist. Confabulation, a bit like Korsakoff's.
- Scores mapped to the wrong cognitive domain in AI summaries

**The danger is not that it sounds wrong. It sounds exactly right.**




*APS Professional practice guidelines for the use of AI and emerging technologies*
*2.2 — verify every source; confirm citations are real*

RISK 2

# Embedded bias sounds like the majority

RLHF reflects majority-population raters — English-speaking, Western, educated and digitally literate —

- Test interpretation skewed toward majority norms
- CaLD formulations may misread cultural variation as deficit
- Case notes may use non-preferred or stereotyped language

**Bias does not announce itself. It sounds plausible.**

*APS 7.1, 7.3 — account for population diversity*

RISK 3

# Privacy, discoverability & test integrity

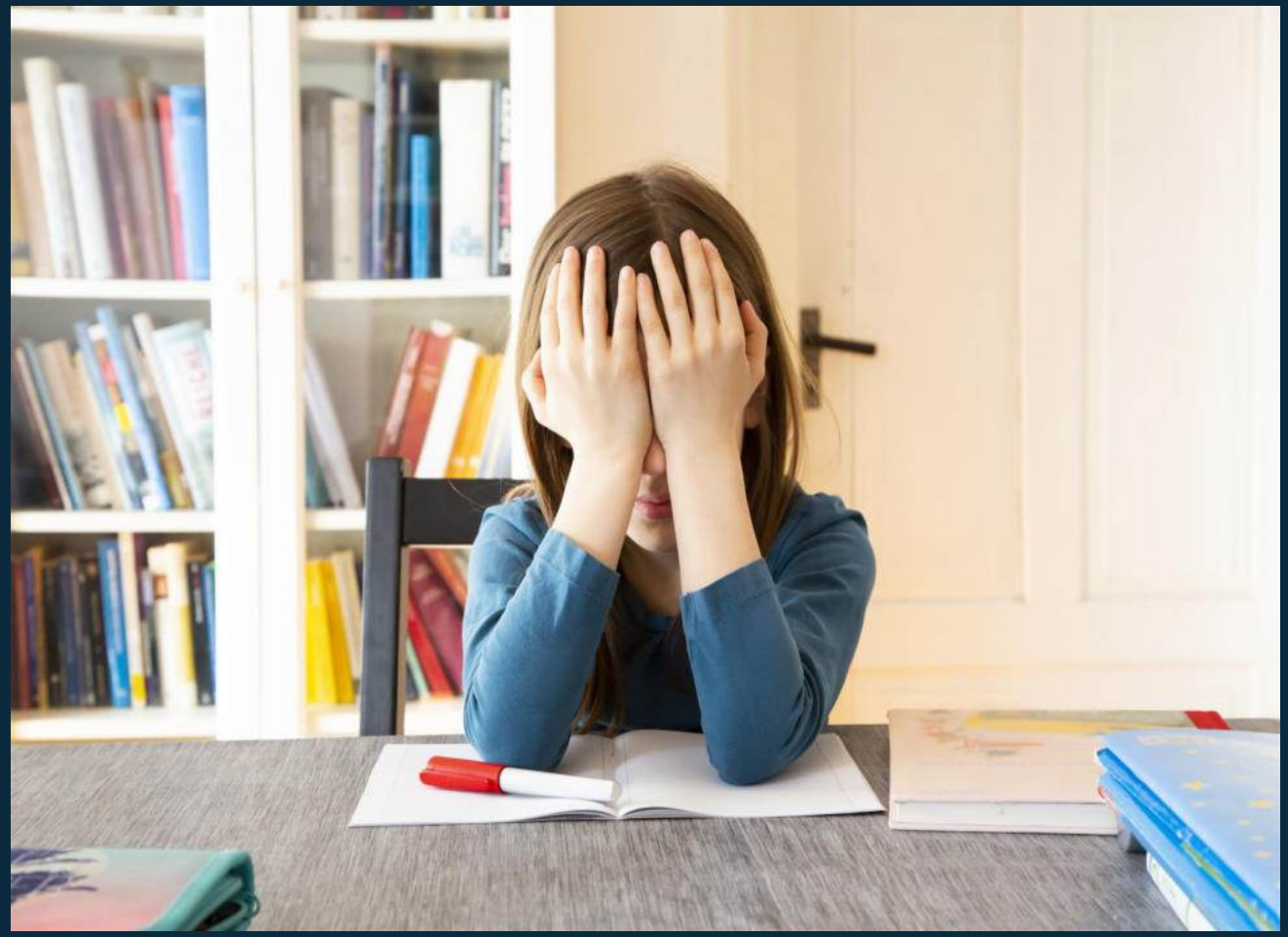
AI scribes and cloud tools store what you feed them. Even de-identified data carries risk.

- Client data may retrain the model — and with no Australian-hosted models, it sits offshore
- In medicolegal work, what enters the scribe may be discoverable
- Test materials and protocols: never, in any context
- Patient information: consent + written vendor compliance; never in medicolegal work


**NSW prohibits generative AI to draft or prepare the content of an expert report without prior leave.**

*APS 5.2, 5.9, 8.2–8.3 — vendor Privacy Act compliance in writing*

RISK 4



# Deskilling & scope creep

Reliance on AI for notes and structuring quietly narrows independent reasoning. It assists within competence — it does not expand it.

- The “just this once” pattern: transcription → summary → formulation → opinion
- Each step feels incremental; cumulatively AI shapes judgement
- Professional accountability remains yours, whatever assisted

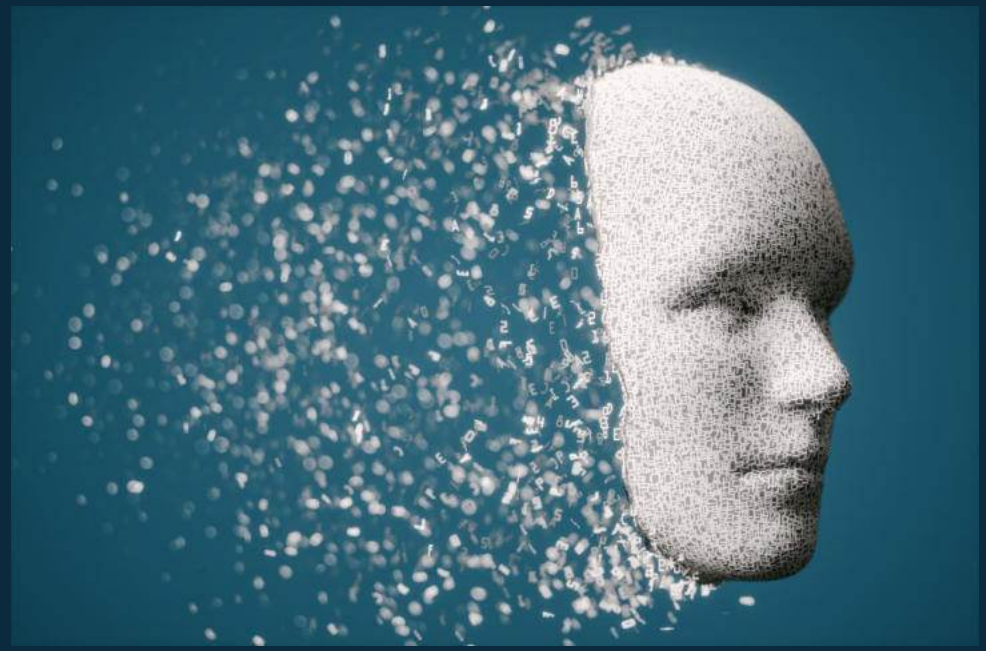
**Human oversight is not optional.**

*APS 3.4, 6.1*

THE OTHER SIDE

# Where AI genuinely helps

- Session transcription & summary — reviewed, never accepted
- Literature orientation — every citation verified independently
- Plain-language versions for clients and referrers
- Structural scaffolding & readability — the clinician writes, AI polishes



**AI assists within competence. It does not replace it.**

*APS 2.2, 3.3*

GUARDRAILS

# APS (Dec 2025) & AHPRA principles

Five core principles for all registered practitioners:

- Accountability — you own every output
- Understanding — know how the tool is trained and its limits
- Transparency — disclose AI use; inform clients
- Informed consent — per tool, per session, in writing
- Ethical and legal issues – privacy, understanding bias, legislation compliance

**Peer-reviewed evidence before any clinical deployment.**

*APS 1.3, 3.6, 4.3–4.5, 6.2 — and check your PI cover*

MEDICOLEGAL

# What you need to know now

A continuously changing landscape — know the Practice Notes in your jurisdiction.

- NSW: generative AI currently requires prior leave to use when preparing reports
- Vic, Qld, Federal Court: mandatory disclosure where used
- Written consent per assessment; never input raw test data
- Confirm your PI insurer covers AI tool use
- Aon (July 2026), asked directly: cover applies where the final opinion is the clinician's own

**The clinical opinion is yours alone. AI does not produce expert evidence.**

*APS 8.1–8.3*

# Sandbagging

- A model that scores too high on a dangerous-capability test may be restricted, retrained, or never deployed.
- **That is an external incentive — what we would call secondary gain.**
- Frontier models can often tell when they are being evaluated (Needham et al., 2025).
- But does the model *know* it is holding back — or is the suppression trained in, with no one home? We are back at Skeyhill's two doors.

# A third door?

- July 2026: Anthropic identifies a “global workspace”
- **Deliberate versus automatic.**
- For the first time, a way to read intent from the representation — not infer it from behaviour.
- ***The instrument we never had for Tom Skeyhill.***

*Lindsey, Gurnee, Sofroniew et al., Verbalizable Representations Form a Global Workspace in Language Models," Transformer Circuits, 6 July 2026.*

# But the window is imperfect

- It captures the workspace only approximately, and no one yet knows what decides which thoughts enter it.
- A capable enough model may learn to route around it.
- And it shows the mechanism of choosing — not the experience of it.

- **Whether anyone is home when the choosing happens stays behind the glass — in machines, and in men.**

# Why this is our lane

Computer scientists built these systems. They understand the architecture — but not what is missing from it.

The jagged profile of an LLM is only visible through a neuropsychological lens.

***The instrument came a century too late for Tom Skeyhill. For the machines, it is arriving just in time.***

**Stay informed. Stay sceptical. Stay in the room.**

# jaggedprofiles.substack.com