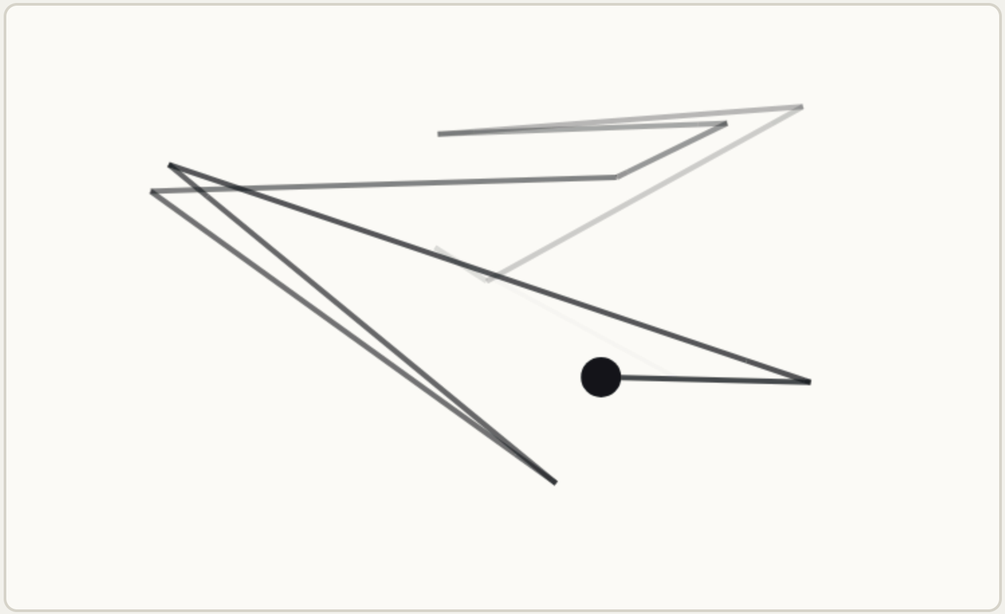
A sketch. Not recordings.

A



B



Brain, or *AI model?*

Hold on to your guess.

THE SCIENCE OF CONSCIOUSNESS 2026 · SAN DIEGO · POSTER #530

The Internal Dynamics of Decisions

From rats and primates to AI

Nathan M. Thornhill
Institute for Complexity Science and Advanced Computing (ICSAC) · Fort Wayne, Indiana

RAT BRAINS | MONKEY BRAINS | FIVE AI MODELS measured the same way



ANIMAL BRAIN CORTEX · RUNS ON ITSELF

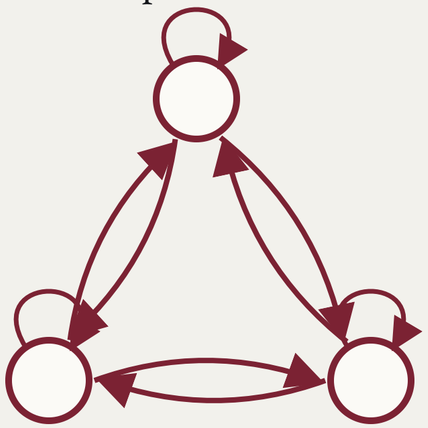
Can you tell them apart?



AI MODEL · RUNS ON ITS INPUT

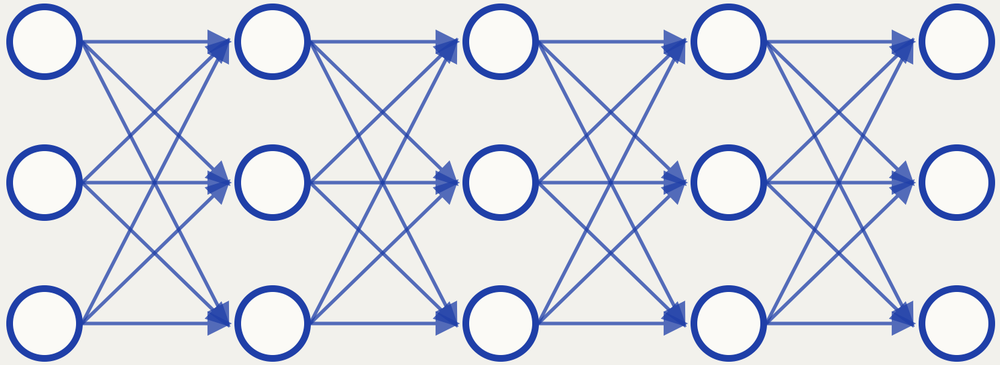
Behavior can't tell you which one has the loops

with loops



unroll

no loops



same inputs in → same outputs out

same inputs in → same outputs out

WHY THE INSIDE MATTERS HERE

Three theories, all betting on the inside

Recurrent processing

Lamme 2006

Local loops in cortex matter for experience.

Global workspace

Mashour et al. 2020

Ignition: a self-amplifying surge that broadcasts what you're aware of.

Integrated information

Tononi et al. 2016

A system's internal cause and effect, more than its input and output.

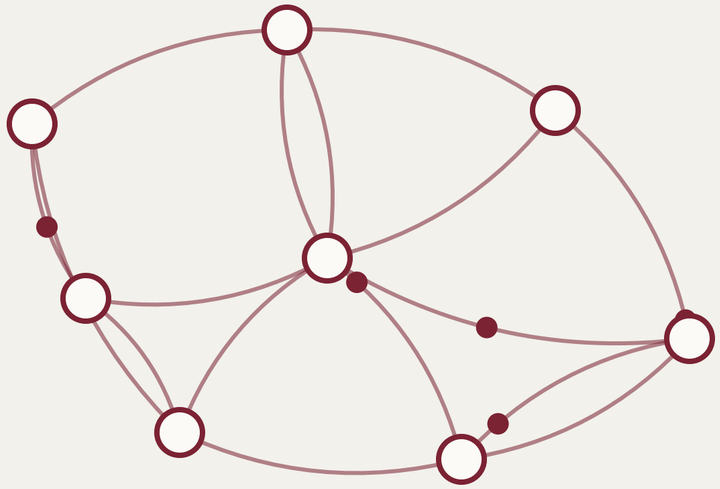
> If matching behavior can't decide it, measure the inside.

TWO KINDS OF SYSTEM

When each one works toward a decision, how much of what it does next comes from *itself*?

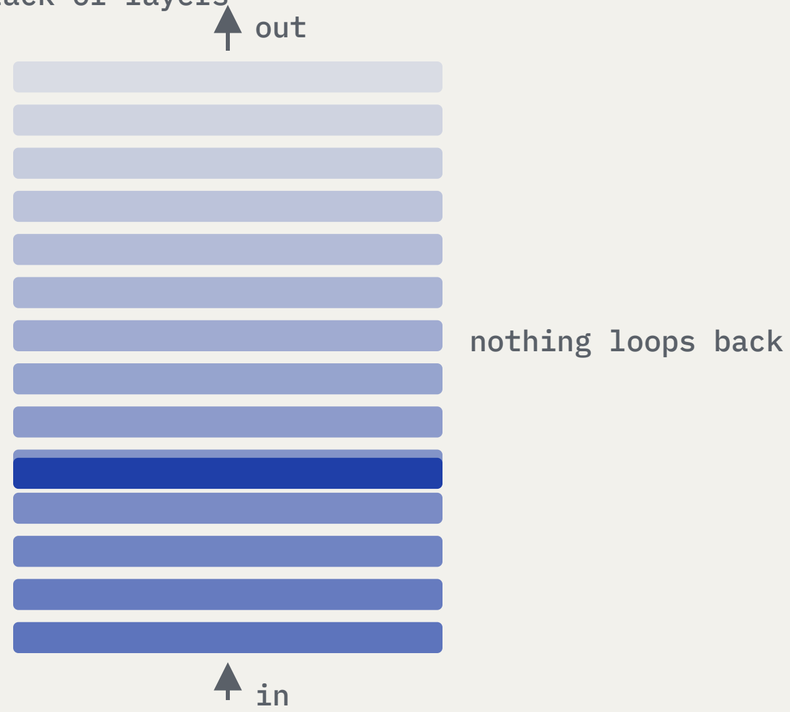
Brain

loops on loops

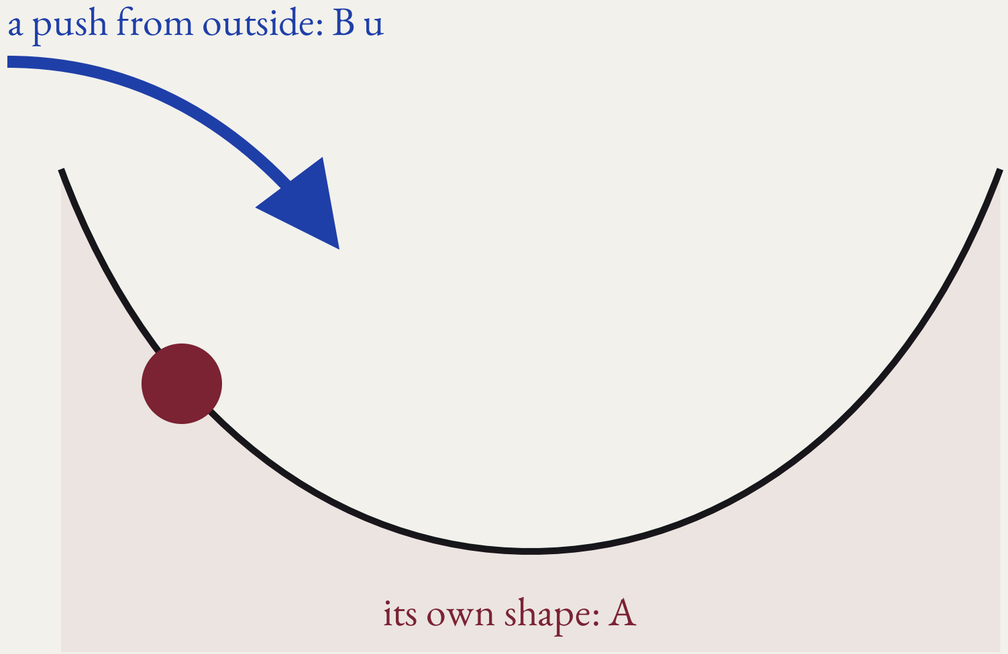


AI model, one pass

a stack of layers



Itself, or its input?



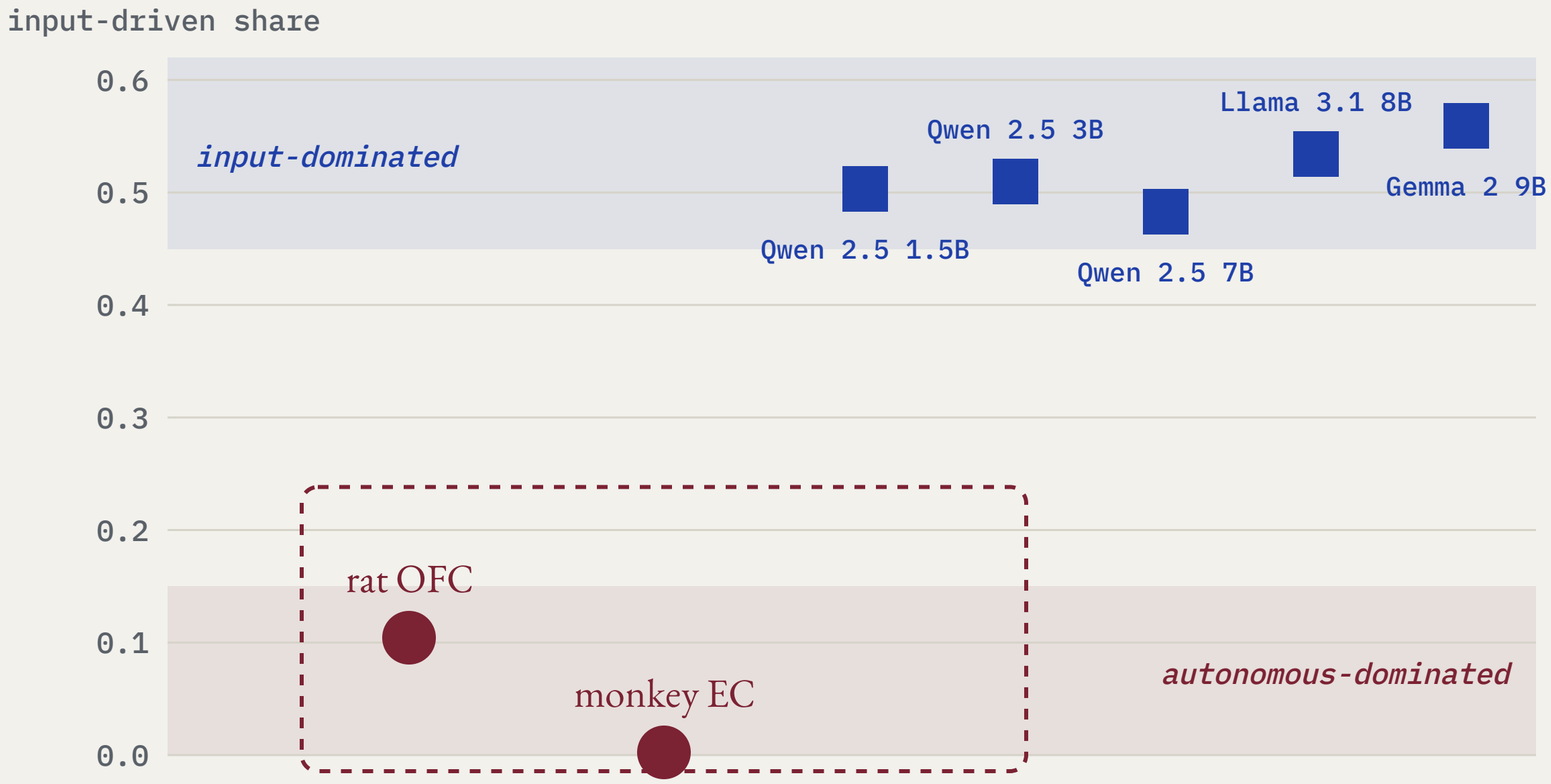
$$x_{t+1} = A\,x_t + B\,u_t + \text{noise}$$

A: what the system does to itself
B u: what the input does to it

In a recording the two come mixed together, and you only see part of the system.
Ordinary methods hand back a blend.

FINDING ONE

Every brain dataset leans on itself. Every AI model leans on its input.



Side = this share against the system's own share. Bands are a guide.
Read the sides, not the sizes: the two sides' input is counted differently.

A · AI models



B · brains



The opening sketch, now colored.

WHOSE DATA

None of the brain data is mine

RAT	How long is a reward worth waiting for? Orbitofrontal cortex.	Schiereck et al. 2026 · 164 sessions · 31 rats
MONKEY	Navigating in the mind to a landmark it can't see. Entorhinal cortex.	Neupane et al. 2024 · 14 sessions · 2 monkeys
RAT	Which side had more clicks? Frontal cortex and striatum.	Luo et al. 2025 · 115 sessions · 12 rats · about 36,000 units
AI MODELS	Qwen 2.5 (1.5B, 3B, 7B) · Llama 3.1 8B · Gemma 2 9B	statements that take two steps of lookup · every layer read

It finds the true answer at every input strength tested

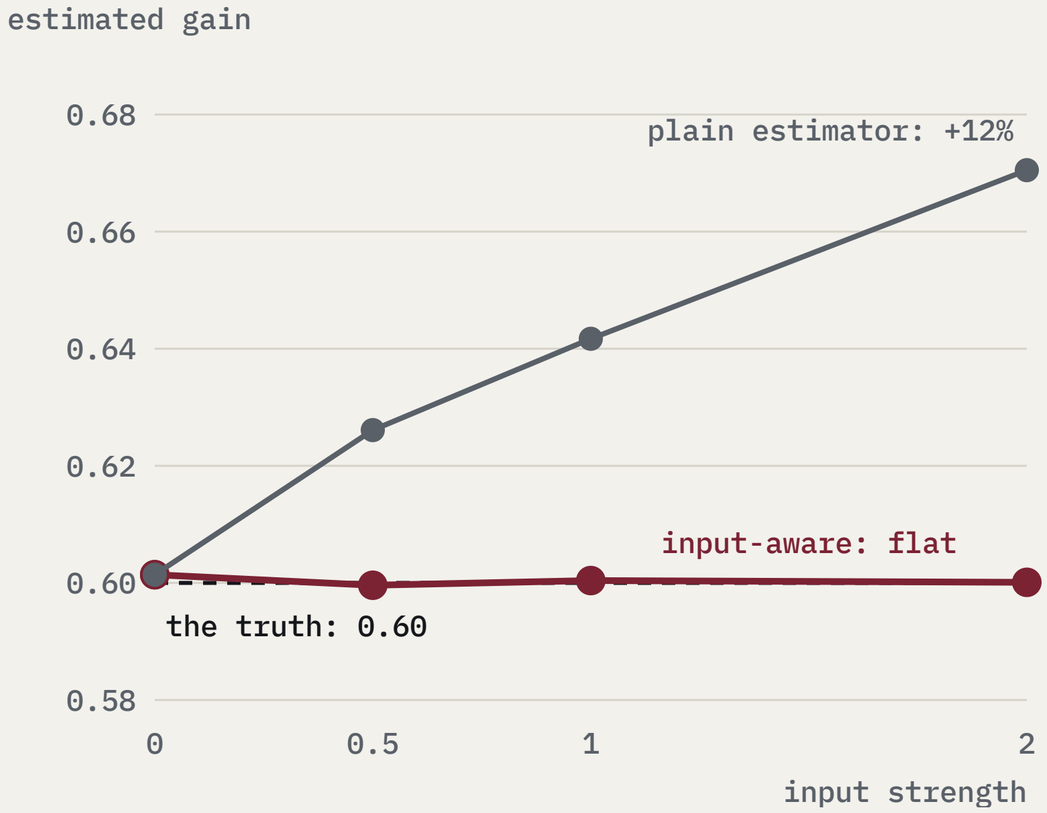
N4SID

Van Overschee & De Moor 1994

Input-aware DSA

Huang et al. 2026 (ICLR)
builds on Ostrow et al. 2023

Synthetic systems with a known answer and input from outside the system.



A REVIEWER'S QUESTION

Match how the input is counted, and the direction holds

	INPUT COUNTED AS 2	INPUT COUNTED AS 8
Rat orbitofrontal cortex	AUTONOMOUS	AUTONOMOUS
Qwen 2.5 1.5B	AUTONOMOUS	INPUT-DOMINATED
Qwen 2.5 3B	AUTONOMOUS	INPUT-DOMINATED

At a matched eight, the direction held. The models' label depends on how their input is counted, so: *read the sides, not the sizes.*

FINDING TWO · HARDER TRIALS

When the rat has to work it out, its cortex leans more on itself

Rats: 24 of 31 lean more on themselves



Monkeys: 10 of 14 sessions, one read-out (support only)



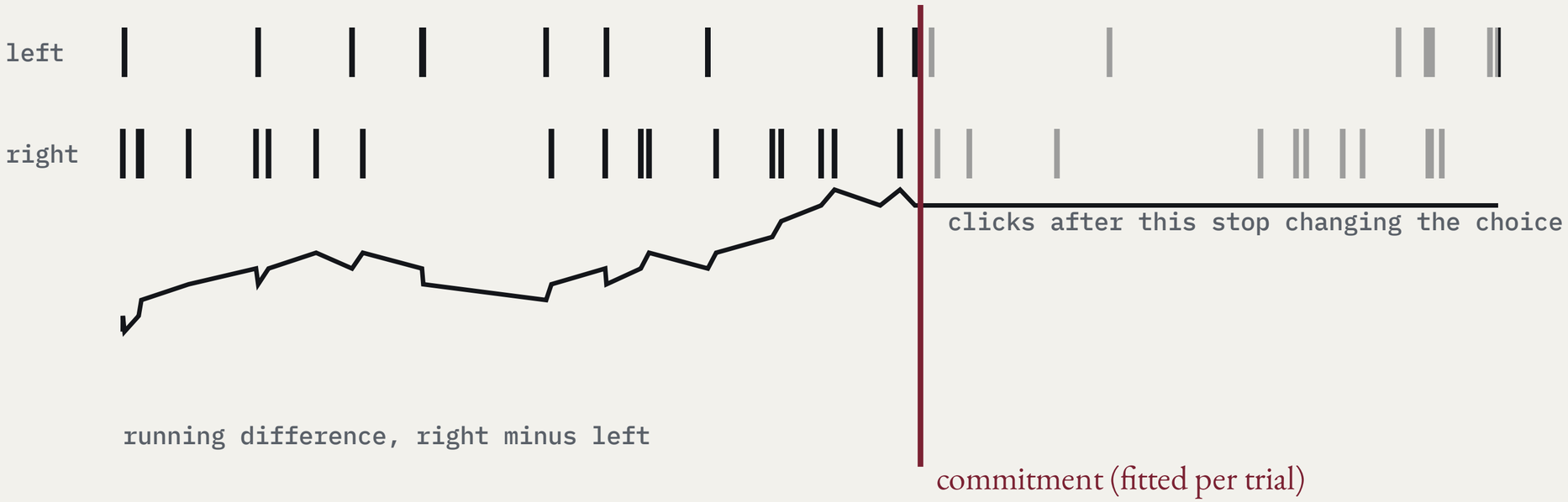
AI models: 4 of 5 did not

Qwen 2.5 1.5B	Qwen 2.5 3B	Qwen 2.5 7B	Llama 3.1	Gemma 2

Llama 3.1: early and middle layers, nearly gone by the deepest probe, still input-dominated

FINDING THREE · CREDIT FIRST

The moment of commitment is Luo's finding

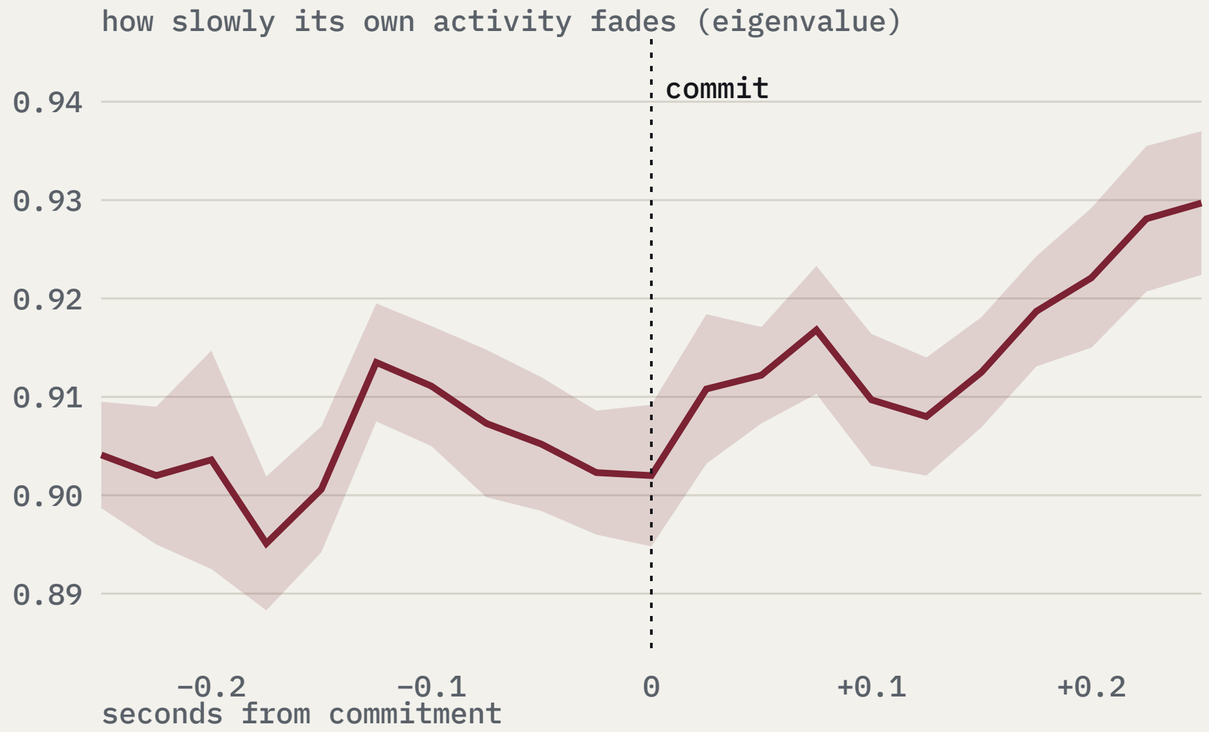


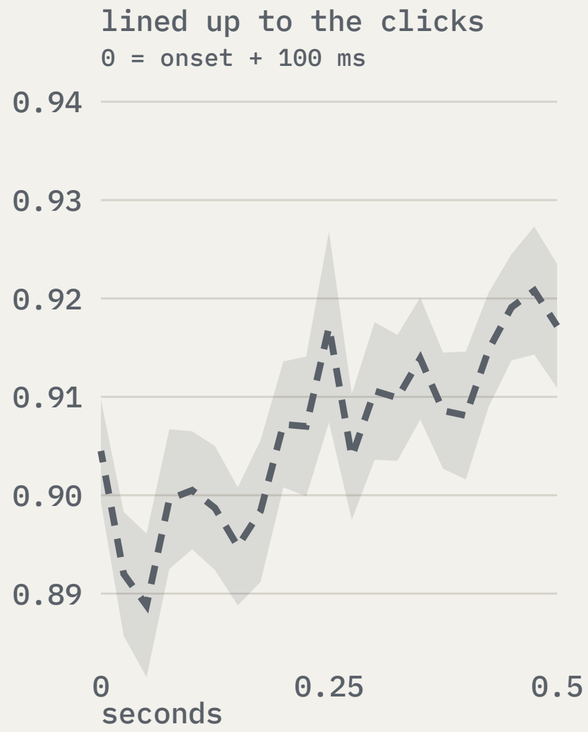
Luo et al. 2025, *Nature*: at commitment the brain shifts from adding up evidence to a state it generates itself.

Huang et al. 2026 saw slower fading after commitment in these recordings. Recovered here, not discovered. The timing is fitted partly from the same spikes, so this is not an independent detection.

FINDING THREE

After the rat commits, its brain holds on longer





SPECIFICITY TO COMMITMENT: SUPPORTED, NOT ESTABLISHED

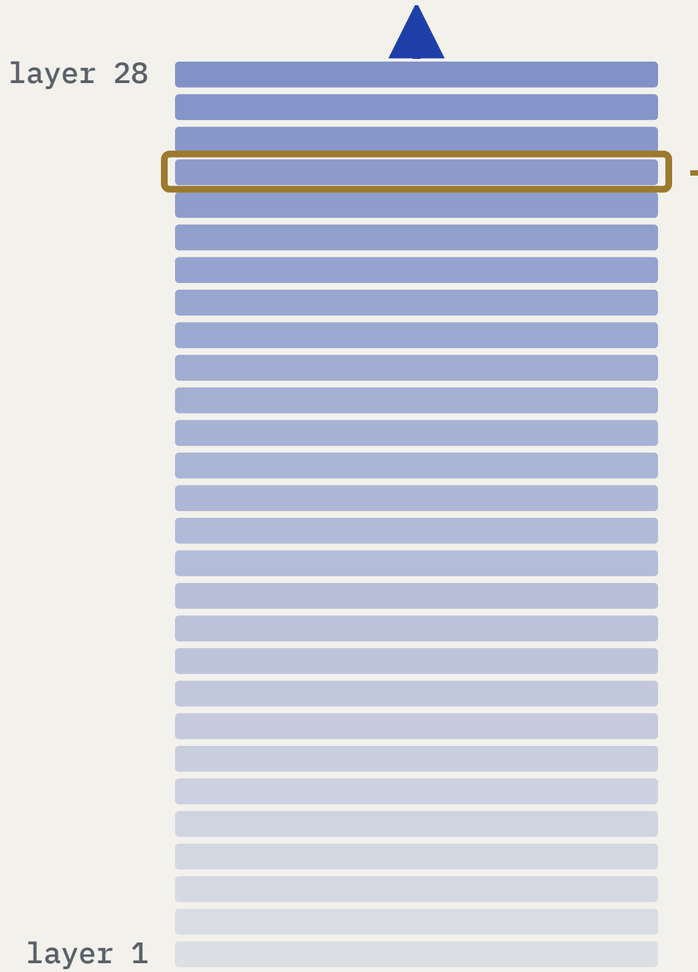
75 of 115 sessions

10 of 12 rats

About 2% per session (median). A ramp, not a switch. A rat-level model gives p ≈ 0.1.

WHERE DOES AN AI MODEL COMMIT?

Inside one pass, layers aren't a clock



no loop inside a pass

so no clock to read

then the pass ends

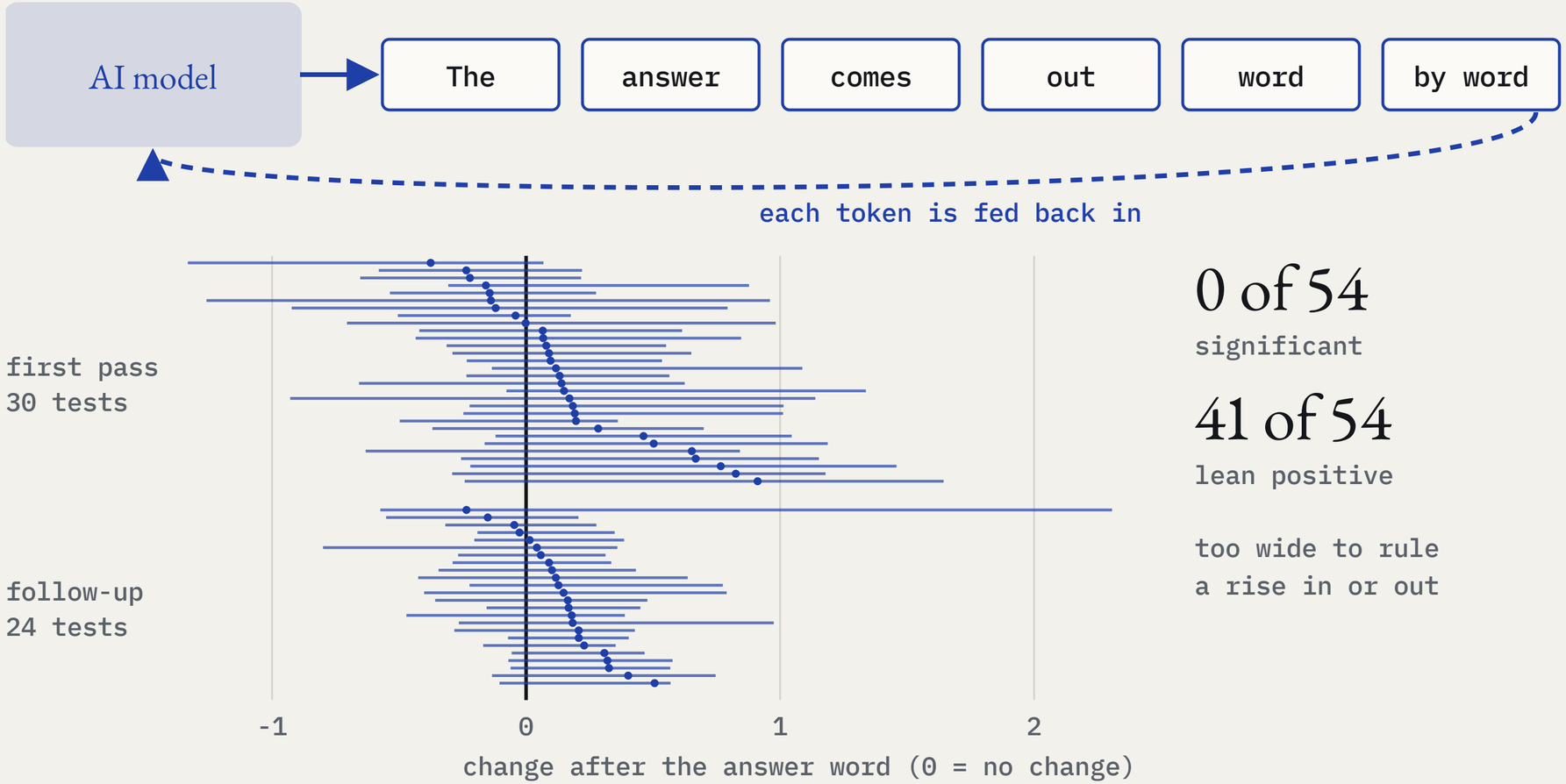
≈ layer 25: the answer's first token

tops the logit lens (Qwen 2.5 7B)
Llama 3.1: about 25 of 32
a marker on the read-out,
not the model deciding

The poster prints a later layer:
a scoring bug found after
printing, fixed in the paper.

THE LOOP IT DOES HAVE

Across generated tokens: no significant rise, and a test too wide to rule one out



0 of 54
significant

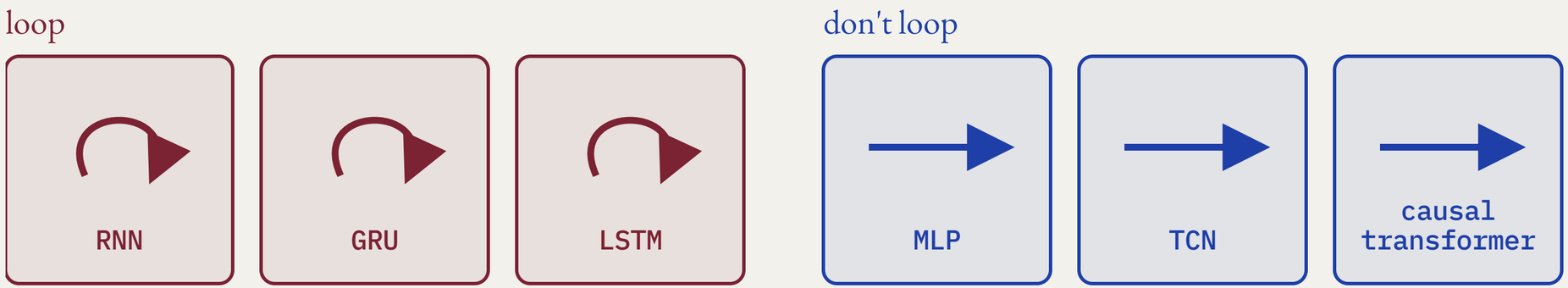
41 of 54
lean positive

too wide to rule
a rise in or out

THE CONTROL

Change only the wiring

	RAT	AI MODEL
task	clicks, left or right	a two-step lookup
after deciding	holds the choice to act on it	done: the answer is the end
wiring	loops	one pass, no loops

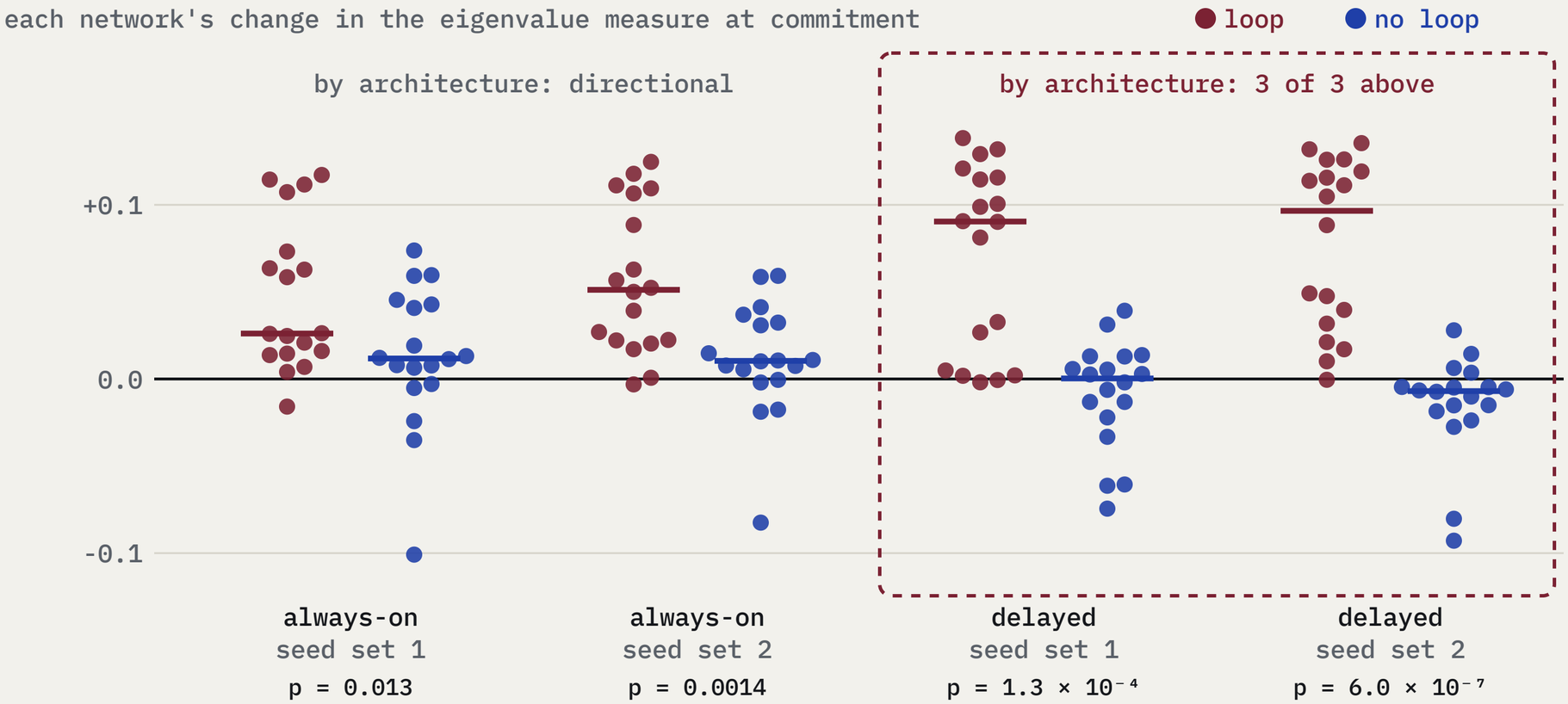


Not the five AI models: six small networks trained from scratch.

Same click task the rats did · trained to the same score · within 1.13× in parameters · 144 networks, none thrown out · design frozen and hash-locked before any dynamical measure

THE CONTROL · RESULT

The looping networks still stand apart



The other pre-registered measure did not separate them. Aligned to an ideal observer's commitment time: holds on the delayed task, not the always-on one.

Wire networks two ways, train them to the same score on the same task, and their dynamics still tell them apart.

WHAT IT DOESN'T SHOW

The nulls and the limits

PRE-SPECIFIED TESTS, RAT OFC

Long waits vs short waits before giving up

Waited for the reward vs quit

A single-number read-out vs willingness to wait (9,106 trials where it quit)

No signal in the predicted direction. A value area, and read-outs not lined up to commitment.

LIMITS

Committing to a decision isn't conscious access.

The monkey data is two animals.

Persistence in cortex can't certify loops (Goldman 2009). The wiring claim rests on the trained networks.

SO WHAT?

No claim about experience, in a brain or a machine

Matching behavior can't decide a claim about internal dynamics.

Internal dynamics can be measured, the same way, in a brain and in a machine.

Persistence at commitment, a property of the kind several theories treat as load-bearing: in the rat recordings; not found in the five AI models tested, by a test too blunt to rule it out.

WHERE I COULD USE YOU

Send me something to measure

HUMAN DATA

A task and a recording method where this could run in people.

REASONING MODELS

A model that reasons at length, and how you'd mark where it commits.

YOUR GUESS, WITH REAL NUMBERS

Brain or AI model? Play the game. Beat the room.



tsc2026.nathanthornhill.com

game · paper · data
pre-registrations

nthornhill@icsacinstitute.org