The Internal Dynamics of Decisions: From Rats and Primates to AI

Nathan M. Thornhill *

Posted Date: 17 August 2026

doi: 10.20944/preprints202608.1095.v1

Keywords: conscious access; computationalism; unfolding argument; recurrent processing; neural population dynamics; decision commitment; large language models; perceptual decision-making

 doi:10.20944/preprints202608.1095.v1

Article

The Internal Dynamics of Decisions: From Rats and Primates to AI

Nathan M. Thornhill

ICSAC Institute, Fort Wayne, Indiana, USA; nthornhill@icsacinstitute.org

Abstract

Whether the signatures of conscious access are fixed by what a system computes, or also by the substrate that implements it, is usually argued conceptually: empirical purchase requires the same computation examined in two different substrates, with the same signature sought in both. The unfolding argument sharpens the difficulty, since for any recurrent network there is a feedforward network with the same behaviour, so behaviour alone cannot settle claims resting on internal dynamics, which must be tested directly. A dynamics demix separating input-driven change from change generated by a population's own recurrence — validated on synthetic systems before any recorded data — was applied to two substrates on matched cognitive work: cortical populations in perceptual and value-based decisions, and language models constructing multi-step answers. Every neural dataset was autonomous-dominated; five language models were input-dominated, a categorical regime difference. At commitment to a perceptual decision the cortical autonomous component strengthened, locked to commitment, not the stimulus; the models showed no such attractor and committed only at their output layer. In a pre-registered control varying only architecture at matched accuracy, the eigenvalue measure separated recurrent from feedforward networks — architecture, not task. Pre-registered negative tests show the cortical effect is invisible to trial-averaged and scalar read-outs. A dynamical property used to reach a commitment is present in cortex but absent from a feedforward model on an analogous task. Nothing is claimed about phenomenal experience or machine sentience; the gap between decision commitment and conscious access is the central limitation.

Keywords: conscious access; computationalism; unfolding argument; recurrent processing; neural population dynamics; decision commitment; large language models; perceptual decision-making

Introduction

The study of consciousness splits into two questions. One is phenomenal experience: what it is like to undergo a state. The other is access: how a represented item becomes available for report, reasoning and the control of behaviour (Block 1995). This paper is about access, and about a narrower part of it still: the moment a system settles on one answer and can use it, here called commitment. It makes no claim about phenomenal experience, and none about machine sentience. That narrow scope is deliberate. Theories of consciousness have been faulted for resting on criteria no experiment can overturn (Doerig et al. 2019; Kleiner and Hoel 2021), so a usable account of access should meet hard, public standards: a measurable quantity, a decidable test, and a stated way for the claim to be wrong (Doerig et al. 2021). This account was built to those standards, and every empirical claim is paired with a condition that would have falsified it.

A central question in the science of consciousness is whether computation alone fixes its signatures. In a form an experiment can address, the question is whether the signatures of conscious access are fixed by *what* a system computes — and so would recur in any system that computes the same thing — or whether they also depend on *how* that computation is implemented in a particular

 doi:10.20944/preprints202608.1095.v1

substrate. The first view is computationalist; the second, in its strong form, implementationist (Searle 2017). The debate has mostly run on conceptual grounds because empirical purchase is hard to get: it needs the same computation examined in two very different substrates, with the same measurable signature sought in both. The unfolding argument sharpens the difficulty (Doerig et al. 2019): for any recurrent network there is a feedforward network with the same input–output mapping, so behaviour alone cannot settle claims that rest on a system's internal causal organisation. If two systems can be made to behave identically while differing in their internal dynamics, a theory that ties consciousness to those dynamics has to be tested on the dynamics directly, not on behaviour. That is the test this paper attempts.

Language models make a useful second substrate. They are not offered as models of the brain, nor as candidates for experience. They carry out multi-step inference and decision tasks, and unlike a brain they can be read out at every layer, their internal dynamics fully observable. What matters for the unfolding argument is that a transformer language model is, within a single forward pass, feedforward: information flows from input embeddings through a fixed stack of layers to an output, with no recurrence across the depth of the stack. It can approximate input–output mappings that a recurrent brain also computes, which is exactly the behavioural equivalence the unfolding argument invokes. This paper asks whether, behind that equivalence, the *dynamics* of reaching the answer differ between the substrates, and whether the difference is of a kind that bears on access.

The measurable quantity throughout is the balance of a population's dynamics between two sources of change. At each step, a population's next state can be driven by external input from outside the population, or generated by its own internal recurrence — its tendency to evolve under its own connectivity even with no new input. A trajectory set mostly by incoming input is *input-driven*; one generated mostly by its own recurrence is *autonomous*. The distinction is central to systems neuroscience: cortical computation is usually modelled as recurrent dynamics that integrate and transform inputs rather than relay them (Mante et al. 2013), and decision formation in particular has been described as a shift from input-driven evidence accumulation to an intrinsically generated, autonomous regime that holds the choice once it is made (Luo et al. 2025). Pulling these two contributions apart in recorded activity is hard, because input and recurrence are entangled whenever the input itself depends on the system's state; the analysis uses a demixing estimator, validated below on systems with known dynamics, that recovers the autonomous component without contamination from the input.

What ties this measure to access is the role recurrence plays in several theories of consciousness. Recurrent processing theory holds that local recurrent or re-entrant activity, before any global broadcast, is already sufficient for phenomenal experience, with more widespread recurrence required for the content to become reportable (Lamme 2006). Global-workspace accounts describe access as a non-linear ignition that broadcasts a representation for report and control (Mashour et al. 2020), itself a dynamical and recurrent process. Integrated information theory ties consciousness to a substrate's intrinsic cause–effect structure rather than its input–output behaviour (Tononi et al. 2016), and the unfolding argument is aimed squarely at that family. Across these otherwise opposed positions one theme recurs: what matters is a system's *internal, autonomous* organisation, not merely the function it computes. A demix that measures, on the same footing in a brain and in a feedforward model, how much of the dynamics is autonomous therefore turns that shared intuition into a decidable test.

This paper reports four empirical results and a set of negative ones. First, on matched cognitive work, cortical populations are autonomous-dominated while five transformer language models are input-dominated — a large, categorical difference in dynamical regime. Second, cortex raises its autonomous component when a task demands inference; the models mostly do not. Third, at the commitment to a perceptual decision the cortical autonomous component strengthens, and it does so locked to the moment of commitment rather than to the stimulus, recovering a regime transition already reported for these data (Luo et al. 2025) with the subspace demix introduced for that purpose (Huang et al. 2025); the models show no comparable attractor and commit only at their output layer.

 doi:10.20944/preprints202608.1095.v1

Fourth, the model's lack of a commitment attractor is structural, not a quantitative shortfall: a feedforward stack has no autonomous temporal regime to settle into.

This paper is equally explicit about what it does not establish, because the negatives are part of the evidence and one of them defines the central limitation. The cortical commitment effect is invisible to trial-averaged and scalar read-outs: pre-registered tests asking whether the autonomous component scales with a behavioural confidence proxy, or with an opt-in/opt-out commitment contrast, returned well-powered nulls. That is why a measure time-locked to commitment was needed, and those nulls are reported in full. Above all, what is measured here is *decision* commitment — the dynamical event by which a perceptual or value-based choice is locked in — and decision commitment is not conscious access. A perceptual decision can be reached without the choice being consciously reportable, and conscious access plausibly involves more than the commitment step. So no claim is made to have measured conscious access, still less phenomenal experience. The claim is narrower and, arguably, defensible: a dynamical property a brain uses to reach a commitment is present in cortex and absent from a feedforward model performing an analogous task, and that bears on whether the relevant signatures are fixed by computation alone. The step from there to access is made carefully in the Discussion, and flagged as the place a reader should be most sceptical.

Methods and Materials

Overview and the Demixing Instrument

The core instrument separates, from a population trajectory, the share of next-state variance generated by the population's own recurrence from the share driven by a known external input. Let the population state at time t be a vector $\mathbf{x}_t$ (after dimensionality reduction; see below) and let $\mathbf{u}_t$ be the external input at the same time. The dynamics are modelled as a linear time-invariant system, $\mathbf{x}_{t+1} = A\,\mathbf{x}_t + B\,\mathbf{u}_t + \text{noise}$, where A is the autonomous (recurrence) operator and B the input operator. Ordinary fits of this model are unreliable when the input is correlated with the state, because the regression cannot then attribute shared variance to A versus B; this identifiability failure is well documented for recurrent-versus-input decompositions of cortical activity (Soldado-Magraner et al. 2024). A was therefore estimated with a subspace state-space identification method (numerical subspace state-space system identification, N4SID; Van Overschee and De Moor 1994), the estimator at the core of recent input-aware dynamical similarity analysis (Huang et al. 2025; building on Ostrow et al. 2023). Two complementary read-outs were taken from each fit, chosen so that at least one is robust in each regime:

- **Autonomous dominance** (sdmdc_dom): the magnitude of the dominant eigenvalue of the identified autonomous operator A. A value below 1 denotes a contractive, stable autonomous mode (a tendency to relax towards an attractor); a value at or above 1 denotes an expansive or marginally stable mode. This read-out is the natural measure of attractor strength but, as the synthetic validation shows, it is only trustworthy when the input is exogenous (the brain regime).
- **Autonomous unique variance** (auto_unique): the additional fraction of next-state variance explained by the past state beyond what the known input already explains, computed as $R^2(\mathbf{x}_{t+1} \mid \mathbf{x}_t, \mathbf{u}_t) - R^2(\mathbf{x}_{t+1} \mid \mathbf{u}_t)$. This is the share of the dynamics attributable to recurrence over and above the input. It is robust to eigenvalue blow-up and is comparable across substrates, and it carries the cross-substrate balance claim.

The input-driven share r2_input = $R^2(\mathbf{x}_{t+1} \mid \mathbf{u}_t)$, the fraction of next-state variance explained by the external input alone, is also reported. A trajectory with high r2_input and low auto_unique is input-driven; the reverse is autonomous-dominated.

Synthetic Validation of the Instrument

 doi:10.20944/preprints202608.1095.v1

Before any recorded data were analysed, the estimator was validated on synthetic linear systems with a known autonomous gain g (the true dominant eigenvalue of A) driven by inputs of varying strength s, under two input regimes: exogenous input independent of the state, and input correlated with the state. With exogenous input, N4SID recovered the true autonomous gain essentially exactly, and the estimate was invariant to input strength: for $g = 0.60$ the estimate was 0.601, 0.600, 0.600 and 0.600 at input strengths $s = 0$, 0.5, 1.0 and 2.0 (spread < 0.002); for $g = 0.80$, 0.800 at every input strength; for $g = 0.95$, 0.950 at every input strength. Ordinary dynamic mode decomposition, by contrast, drifted upwards with input strength (for $g = 0.60$: 0.601, 0.626, 0.642, 0.671), confirming that the naive estimator confounds input with recurrence. Under strongly state-correlated input the autonomous-eigenvalue read-out became unreliable and could exceed 1, while the auto_unique read-out remained interpretable, tracking true recurrence with a flat trial-shuffle null in the negative control. These validations fix the division of labour used throughout: auto_unique is the cross-substrate workhorse; sdmdc_dom is reported where the input is exogenous (the brain decision data) and interpreted with caution elsewhere. One further caveat from the validation matters for the recruitment analysis below: the variance-partition read-out, while robust to eigenvalue blow-up, is not invariant to a difference in input *strength* between two conditions — a control system with equal recurrence but unequal input did not return a flat null on auto_unique. The orbitofrontal-cortex recurrence-recruitment contrast is therefore corroborated with the autonomous-eigenvalue read-out, which is input-invariant by construction (Figure 1a) and passed the corresponding negative control. Full synthetic results are in the deposited results/inputdsa_synthetic2.json and results/validate3.json.

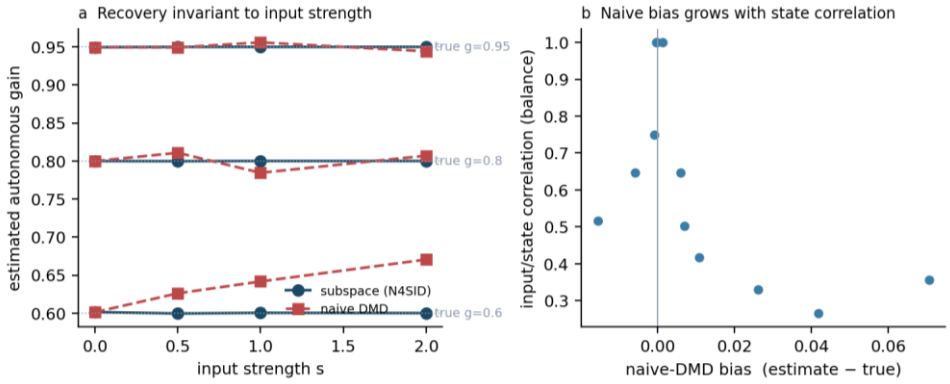


Figure 1. The demixing instrument and its validation on synthetic systems with known dynamics. (a) On synthetic linear systems with a known autonomous gain, the subspace estimator (N4SID) recovers the true gain and holds the estimate fixed as input strength is swept over a four-fold range (points stay on the dotted true-gain lines), whereas ordinary dynamic mode decomposition drifts upwards with input strength. (b) The naive estimator's bias (estimate minus true gain) grows as the input becomes more correlated with the state, the regime in which the subspace estimator is required.

Alt text: validation plots showing the subspace estimator recovers the true autonomous gain invariant to input strength while the naive estimator inflates as input strength and input–state correlation increase.

Neural Datasets and the Input Signal

Three public single-unit / Neuropixels datasets were reanalysed. In every case the external input **u**_t was defined from the task, not from the neural data, so that the demix tests recurrence against a genuinely exogenous drive.

Rat orbitofrontal cortex (OFC), value-based decisions. Single-unit recordings from rats performing a temporal wagering task with hidden reward states (Constantinople laboratory; published in Schiereck et al. 2026) were reanalysed across 164 sessions from 31 rats. Population trajectories were reduced to their leading principal components per session. The input **u**_t was the task-locked, value- and time-modulated evoked drive (cue and movement-onset kernels scaled by offered value).

 doi:10.20944/preprints202608.1095.v1

Inference trials (those shortly after an unsignalled hidden-state switch, trials-since-switch < 10) were contrasted with an established-state control (trials-since-switch ≥ 20), with reward value matched between conditions by subsampling.

Primate entorhinal cortex (EC), mental navigation. Single-unit recordings during a mental-navigation task (Neupane et al. 2024) were reanalysed across 14 sessions from two monkeys. The occluded-inference condition (mentally constructing a hidden trajectory) was contrasted with the visible-landmark condition. This dataset is reported as a second, weaker anchor (see Limitations).

Rat frontal cortex and striatum, perceptual decision with measured commitment. Neuropixels recordings from rats performing the Poisson-clicks evidence-accumulation task, with a per-trial behavioural estimate of the moment of commitment (the change-of-mind / commitment time, nTc) derived by the original authors (Luo et al. 2025; data at Dryad doi:10.5061/dryad.sj3tx96dm; commitment times from the authors' public repository), were reanalysed across 115 sessions from 12 rats (approximately 36,000 frontal and striatal units; 19,095 committed trials). The input **u**_t was the accumulated sensory evidence (the running difference of left and right click counts, from the recorded click times). Trajectories were binned at 25 ms in absolute time and aligned two ways: to the behavioural commitment time, and to the stimulus onset. The autonomous read-outs were computed time-resolved in sliding windows around each alignment point; the pre-commitment window was centred at −150 ms and the post-commitment window at +200 ms, and the paired test compared the commitment-aligned rise with the stimulus-aligned rise within each session. The commitment time is estimated from the animals' behaviour (choices and click trains) independently of the recorded population activity, so aligning the neural dynamics to it is not circular.

Language Models and the Matched Computation

Five instruction-tuned transformer language models spanning three families and a range of scales were analysed: Qwen2.5-1.5B-Instruct, Qwen2.5-3B-Instruct, Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Gemma-2-9b-it. Hidden states were extracted at every layer for a set of construction items, each of which states a fact reachable only through an intermediate entity (a two- or three-step lookup); the inference condition (landmark in the released code) was contrasted with a length- and vocabulary-matched filler condition requiring no inference (var_filler). For each item the across-token trajectory over the final $K = 12$ token positions at a given depth was treated as the population trajectory **x**_t, and the input **u**_t was defined as the layer-0 input embeddings at the same positions, the information being fed into the layer. The demix was computed at six fractional network depths (0.4 to 0.9 of the stack), and the recruitment contrast is reported per depth and averaged over them. The same read-outs were applied. Because a transformer has no recurrence within a forward pass, the across-token trajectory at a fixed depth is the closest analogue of a temporal population trajectory; this is stated explicitly, and the across-layer analysis (below) is the complementary view. Causal attention does, however, mix information across token positions, so the across-token trajectory is not a purely feedforward signal, and the model's non-zero autonomous share partly reflects this; it is treated as a limitation of the analogue rather than as evidence of temporal recurrence.

The commitment layer. For the model analogue of decision commitment, the commit layer L* was defined per item as the first layer at which the correct answer token becomes the top-ranked prediction under the logit lens (the projection of intermediate hidden states through the model's output head; Nostalgebraist 2020; the tuned-lens refinement is Belrose et al. 2023). Across-layer trajectories at the answer-predicting token position were unit-normalised per layer to remove the confound of growing residual-stream norm, and the autonomous-dominance read-out was computed in sliding layer windows aligned to L*, exactly mirroring the brain's commitment-aligned analysis. Because the logit lens indexes the layer at which the answer becomes linearly readable from the residual stream, rather than the point at which a dynamical state settles, L* is treated here as a read-out landmark and not as evidence of an attractor in itself; what carries the comparison is the autonomous-dominance read-out computed in windows around it.

 doi:10.20944/preprints202608.1095.v1

Statistics

Cross-substrate claims are made at the level of dynamical *regime and sign* only; absolute magnitudes are never compared across substrates, because they depend on unit counts, recording modality and reduction choices. The two substrates' baseline levels are summarised with statistics that are not identically defined — neural levels are across-session medians of the inference condition, while model levels are means pooled over conditions and probe depths, so the balance comparison rests on the categorical difference in regime, not on a like-for-like level. A further asymmetry matters: the exogenous input regressor is low-dimensional for the brain (a task-derived evidence or value signal) and high-dimensional for the models (the full layer-0 embedding), so the absolute input-driven share is not comparable across substrates even in principle, because a low-dimensional regressor can explain only a limited share of a multi-component population's next state while a high-dimensional one can explain more. The cross-substrate claim therefore does not rest on the absolute level gap. The load-bearing comparisons are the within-substrate contrasts (the recruitment of autonomy under inference, computed with the same input definition within each substrate) and the within-brain commitment attractor, neither of which is affected by this asymmetry; the condition contrasts (inference minus control) are defined identically across substrates and carry the recruitment claim. Within a substrate, condition contrasts (inference minus control; post-commitment minus pre-commitment) were computed per session and aggregated with the Wilcoxon signed-rank test over sessions, with per-animal sign tests reported alongside to guard against session-count pseudo-replication. Directional hypotheses (that the autonomous component rises under inference, and strengthens after commitment) were pre-specified; two-sided p-values are reported throughout unless a one-sided test is named explicitly. Confidence intervals on model read-outs were obtained by item bootstrap. For the commitment-locked analysis, the rise after commitment was compared with the rise after stimulus onset by a paired Wilcoxon test over sessions (commitment-aligned versus stimulus-aligned within the same session). Negative-control tests were pre-specified with their kill-conditions before analysis. No family-wise correction was applied across the small number of pre-registered contrasts; exact n, sign counts and p-values are reported for every test so the reader can apply their own correction.

Software

Analyses used Python with NumPy, SciPy and scikit-learn; model activations were extracted with the Hugging Face Transformers library; subspace identification used the dynamical-similarity-analysis implementation of Ostrow and colleagues. Use of a large language model as a coding and copy-editing aid is declared in full in the Declaration of AI use below.

Results

The Instrument Recovers Autonomous Dynamics Without Input Contamination

The demix was first checked on synthetic systems whose true autonomous gain was known (Methods; Figure 1). With exogenous input it recovered that gain to three decimal places and held the estimate fixed as input strength was swept over a four-fold range, where the naive estimator drifted upwards by up to 12% of the true value. Under state-correlated input the eigenvalue read-out lost reliability, but the variance-partition read-out (auto_unique) remained interpretable, tracking true recurrence against a flat shuffle null. These results justify the analysis strategy: the variance partition carries the cross-substrate balance comparison, and the eigenvalue read-out is reserved for the brain decision data, where the input (sensory evidence) is genuinely exogenous to the recorded population.

On Matched Cognitive Work, Cortex Is Autonomous-Dominated and Language Models Are Input-Driven

The first test asked, on the same footing in both substrates, how much of a population's moment-to-moment dynamics is driven by external input versus generated by its own recurrence (Figure 2). The difference between substrates was large and categorical. In rat OFC during value-based

 doi:10.20944/preprints202608.1095.v1

decisions, the input-driven share of next-state variance was 0.10 and the autonomous unique share 0.73; in primate entorhinal cortex during mental navigation, the input-driven share was 0.002 and the autonomous share 0.24. Every neural dataset was autonomous-dominated: the recorded trajectory was generated overwhelmingly by the population's own recurrence, with external input explaining at most a tenth of the next-state variance.

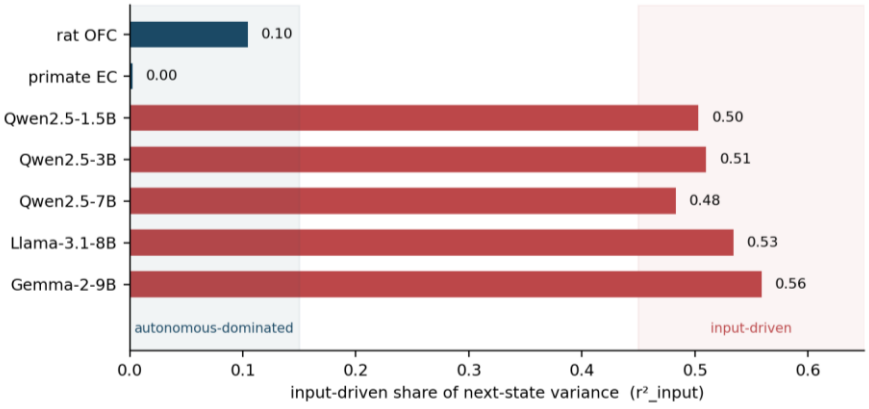


Figure 2. Cortex is autonomous-dominated; language models are input-driven. Input-driven share of next-state variance for each substrate: rat orbitofrontal cortex (0.10), primate entorhinal cortex (0.002), and five transformer language models (0.48–0.56). Brain datasets in one colour, models in another; the gap between substrate groups is categorical.

Alt text: bar/scatter plot showing brain input share near 0.0–0.1 and all five models near 0.5.

The five language models fell on the opposite side of the divide. Their input-driven share ranged from 0.48 to 0.56 (Qwen2.5-1.5B 0.50, Qwen2.5-3B 0.51, Qwen2.5-7B 0.48, Llama-3.1-8B 0.53, Gemma-2-9b 0.56) and their autonomous unique share from 0.26 to 0.33. Across families and scales, roughly half of each model's across-token dynamics was explained by the information being fed in. This is a difference of dynamical regime, not of degree within a regime: the brain operates in an autonomous-dominated regime and the feedforward models in an input-dominated one. The pre-registered condition that would have undercut the cross-substrate claim (that a model carries the same autonomous-versus-input balance as the brain) was not met by any model.

Cortex Raises Its Autonomous Component When Inference Is Required; Models Mostly Do Not

The balance result is a standing property. The second test asked whether the autonomous component is *recruited* by cognitive demand: does it rise when a task requires inference rather than retrieval of an established state (Figure 3)? In rat OFC it did, on both read-outs. The autonomous unique share was higher on inference trials than on value-matched control trials by a median of +0.0089 (Wilcoxon signed-rank $p = 1.3\times10^{-6}$, two-sided, over 164 sessions; positive in 107 of 164 sessions and in 24 of 31 rats, per-animal sign test $p = 0.003$). The input-cleaned autonomous-eigenvalue read-out, the one robust to a difference in input strength between conditions and the one that passed its negative control, moved the same way (+0.0041, two-sided $p = 0.006$). The population also became *less* input-driven under inference (Δ input share −0.0023, $p = 0.01$), so the two read-outs converge: under inference the brain is at once more autonomous and less input-driven. Value-matching roughly halved a larger raw effect (a real value/novelty confound), but the residual was robust across animals. In primate entorhinal cortex the autonomous share also rose under inference (median +0.035, 10 of 14 sessions; pre-specified one-sided Wilcoxon $p = 0.045$, two-sided $p = 0.091$), presented as directional corroboration in a second species and region rather than independent confirmation, given the two-animal sample. Because this rise appears only in the variance-partition read-out (the input-invariant eigenvalue read-out is null here) and the input share differs between the conditions, it cannot be cleanly separated from the input-strength sensitivity noted in Methods, and it is weighted lightly.

 doi:10.20944/preprints202608.1095.v1

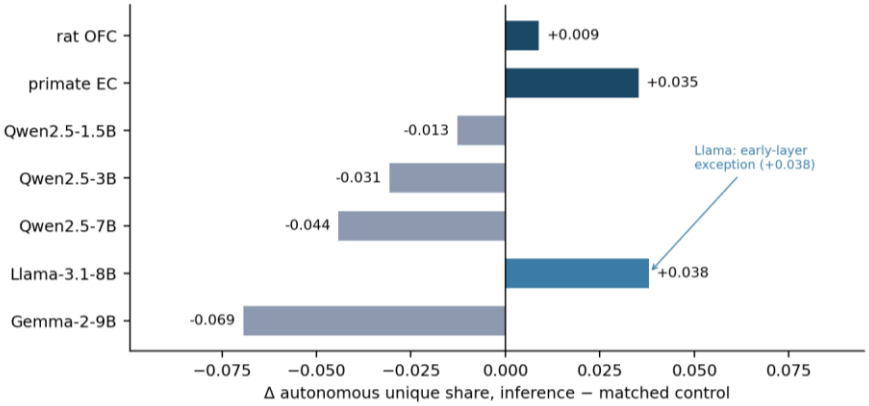


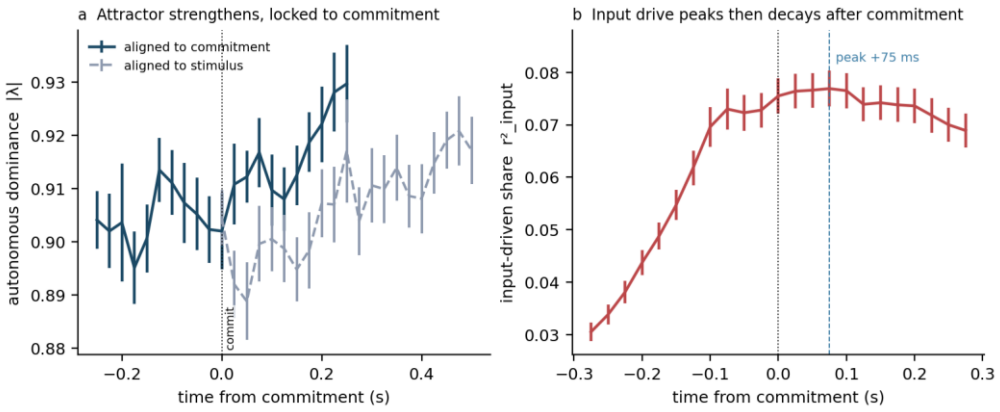
Figure 3. Cortex recruits autonomous recurrence under inference; models mostly do not. Change in autonomous unique share, inference minus matched control, per substrate. Rat OFC rises (+0.0089, two-sided $p = 1.3\times10^{-6}$, 24/31 rats); primate EC rises weakly (+0.035, 10/14 sessions; one-sided $p = 0.045$); four of five models are flat-to-negative; Llama-3.1-8B is the early-layer exception (+0.038).

Alt text: plot of inference-minus-control autonomous share, positive for brain, flat-to-negative for most models.

The models mostly did not recruit autonomy under inference. Four of the five showed a flat-to-negative change, and three of them (both Qwen models above 1.5B and Gemma) became *more* input-driven under inference, with commit-layer changes in autonomous share of −0.07 to −0.10 whose bootstrap intervals excluded zero. The honest exception, reported in full because calibration cuts both ways, is Llama-3.1-8B: its autonomous unique share rose under inference at early-to-middle layers (+0.038 on average, positive at all six probe depths sampled, bootstrap interval excluding zero at the middle depths). This is a genuine within-model effect. It does not break the balance dissociation — Llama remained input-dominated in absolute terms (input share 0.53, firmly in the input-driven regime) and the rise decayed to near zero (+0.002) at the deepest, output-adjacent probe — but it is a real exception to "no model ever recruits autonomy", and it is a natural hook for future work on whether early-layer autonomy tracks reasoning competence. No clean "models never do this" claim was manufactured where the data do not support one.

The Cortical Autonomous Attractor Strengthens at the Moment of Commitment

The balance and recruitment results establish that cortex operates autonomously and engages that autonomy under demand, but they do not tie the autonomy to the act of committing to an answer. The third test does, using the perceptual-decision dataset for which a per-trial behavioural estimate of the commitment moment exists (Figure 4). The autonomous-dominance read-out was computed time-resolved around two alignment points: the behavioural commitment time and the stimulus onset.



 doi:10.20944/preprints202608.1095.v1

Figure 4. The cortical autonomous attractor strengthens at commitment, locked to commitment not stimulus. (a) Autonomous-dominance read-out time-resolved around the behavioural commitment time (rises from ~0.90 to ~0.93 after commitment) versus around stimulus onset (flat); error bars are SEM across 115 sessions. Per session, the post- minus pre-commitment rise is +0.016 (75/115 sessions, two-sided $p = 2.2\times10^{-5}$), and it is sharper aligned to commitment than to stimulus (paired $p = 0.027$). (b) Variance-partition view: the input-driven share peaks ~75 ms after commitment then declines.

Alt text: left, autonomous-eigenvalue curves rising after commitment when aligned to commitment but not to stimulus; right, the input-driven share peaking about 75 ms after commitment then declining.

Aligned to commitment, the cortical autonomous attractor strengthened. The dominant autonomous eigenvalue was flat at about 0.90 before commitment and rose to about 0.93 by 250 ms after it; per session, the post-commitment value (averaged over a window centred at +200 ms) exceeded the pre-commitment value (centred at −150 ms) by a median of +0.016 (Wilcoxon signed-rank $p = 2.2\times10^{-5}$, two-sided; the rise was present in 75 of 115 sessions). The strengthening was specific to commitment, not to the stimulus. Aligned to stimulus onset, the same per-session rise was only +0.002 and did not reach significance (two-sided $p = 0.095$), while the rise was significantly sharper when aligned to commitment than to stimulus (paired Wilcoxon $p = 0.027$; commitment-aligned rise exceeded stimulus-aligned rise in 68 of 115 sessions). A complementary variance-partition view told the same story from the other side: the input-driven share rose during evidence accumulation, peaked about 75 ms after commitment, and then declined (Figure 4b), so the population decoupled from sensory evidence and became more autonomous after committing. The effect is a commitment-locked ramp into a stronger attractor, not an all-or-none switch — the rise is small in absolute terms (about 0.016, roughly 2%) and the stimulus-aligned rise is weakly non-zero — but it is locked to the behavioural commitment, well powered, and present in a clear majority of sessions. The direction of the effect is robust to the choice of pre- and post-commitment window (the per-session rise is positive across the plausible window grid), but its magnitude is window-dependent, ranging from about +0.005 ($p \approx 0.04$) to about +0.021, so the windowed magnitude is treated as an estimate rather than a precise quantity. A naive dynamic mode decomposition applied to the same commitment-aligned data did not show the rise (it was flat and noisy); the rise emerges only with the input-aware estimator, consistent with the naive estimator's input bias documented in the validation (Figure 1).

This commitment-locked transition from input-driven to intrinsically generated dynamics is the result reported for these recordings by their original analysts (Luo et al. 2025) and recovered by the subspace demix used here, whose authors applied it to the same data (Huang et al. 2025). The present brain analysis is therefore a confirmation that builds on that published demix, not an independent rediscovery; what is new in this study is not the cortical transition but the cross-substrate comparison in the next section, which places a feedforward language model on the same demixed axis.

A Feedforward Language Model Has No Commitment-Locked Attractor

The fourth test asked whether the model substrate shows anything like the brain's commitment-locked attractor (Figure 5). It does not, and the reason is structural. The commit layer L* at which the answer first becomes the model's top prediction was located per item, and the across-layer autonomous-dominance read-out was then measured in windows aligned to L* (mirroring the brain analysis). Two facts stand out. First, both tested models commit at their final layer (median commit layer 28 of 29 for Qwen2.5-7B; 32 of 33 for Llama-3.1-8B): the answer is not determined until the output, so there is no post-commitment phase in which dynamics could settle — the brain commits mid-process and the attractor strengthens *afterwards*, whereas the model has no "afterwards". Second, the across-layer autonomous-dominance read-out hovers near or above 1 (expansive, feedforward) and shows no clean rise towards a stable contractive attractor like the brain's; the largest values are isolated late-layer spikes attributable to the output projection rather than to any settling dynamics.

 doi:10.20944/preprints202608.1095.v1

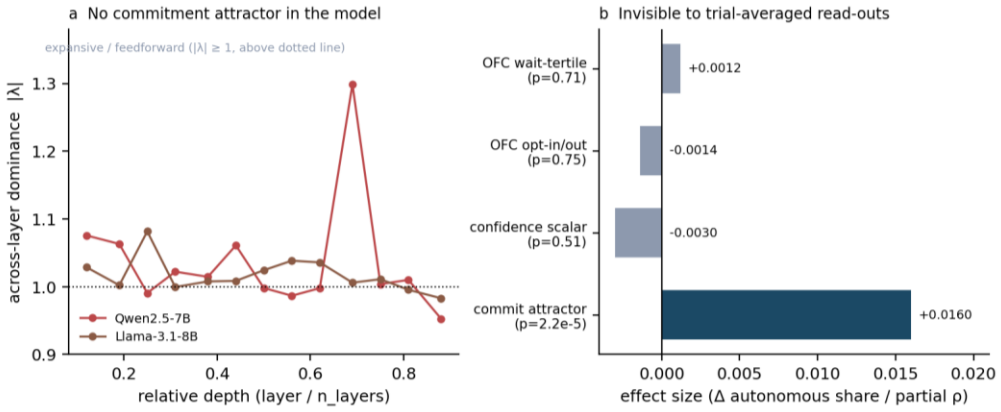


Figure 5. A feedforward model has no commitment-locked attractor, and the cortical effect is invisible to trial-averaged read-outs. (a) Across-layer autonomous-dominance read-out as a function of relative network depth for two models: values hover near/above 1 with late-layer projection spikes, no contractive rise; both models commit at their final layer. (b) Pre-registered nulls: OFC commitment contrasts (wait-time tertiles p = 0.71; opt-in/opt-out p = 0.75) and the scalar confidence test (partial Spearman −0.003, p = 0.51), shown against the commitment-locked positive result for contrast.

Alt text: model across-layer eigenvalue flat near 1 with spikes; null effect sizes near zero with wide intervals.

The strength of this leg needs care. The across-layer eigenvalue read-out is a noisy analogue: a transformer's across-layer trajectory is not a temporal dynamical system, and norm and projection artefacts inflate the eigenvalue at late layers (visible as the spikes in Figure 5). The model-side claim therefore rests on three things that do not depend on that noisy number: the structural fact that a feedforward stack has no autonomous temporal regime to settle into; the finding that the model commits at its output layer, leaving no post-commitment dynamics; and the absence of any clean contractive-attractor rise of the kind the brain shows. The claim is not "the model's attractor is weaker" but "there is no commitment-locked attractor to measure", which the structure of the model entails and the data are consistent with.

A stronger, real-model version of this test was run to turn the structural argument into a measurement (pre-registered and frozen before any result, and deposited). Each of the five language models was made to *generate* the answer to a construction item together with a continuation, so that the demix eigenvalue could be measured across the generated tokens in a window centred on the committed answer token, giving a genuine post-commitment window that mirrors the brain's post-commitment settling under the same instrument. Neither the first pass nor a pre-registered, properly powered follow-up (longer generations, wider windows, and the answer located wherever the model emitted it) found a significant commitment-locked rise in autonomous dominance in any model (0 of 30 and 0 of 24 model-by-depth tests; minimum p = 0.07). The post-commitment eigenvalue leaned weakly and non-significantly upwards, but towards the marginal boundary near 1 rather than settling below it into a contractive attractor: no model reproduced the brain's contractive commitment rise, which stays well below 1 (about 0.90 to 0.93; $p = 2.2\times10^{-5}$). One model (Gemma-2-9b) committed to its answer in too few generations to test under the stricter criterion. The eigenvalue estimator is noisy on short generated sequences, so this corroborates the absence of a brain-like commitment attractor rather than excluding a small effect tightly; the structural argument remains the load-bearing claim.

A Computation-Matched Architecture Control

The cross-substrate comparison above varies task and architecture at once, so it cannot on its own attribute the dynamical difference to architecture rather than to task. The controlled test the unfolding argument specifies — the same input–output function realised in a recurrent and in a feedforward system — was therefore run and pre-registered in full, with the task, the architectures,

 doi:10.20944/preprints202608.1095.v1

the metric and the decision rule frozen and hashed before any dynamical quantity was computed (the pre-registration and code are deposited). Three recurrent architectures (a vanilla recurrent neural network, RNN; a gated recurrent unit, GRU; and a long short-term memory network, LSTM) and three feedforward architectures (a windowed multilayer perceptron, MLP; a dilated causal temporal convolutional network, TCN; and a causal transformer encoder), six seeds each, were trained on an identical Poisson-clicks evidence-accumulation decision and reduced to the same eight-dimensional trajectory; the demix was then applied unchanged. Every network reached matched accuracy on each task (0.943–0.949 on the always-on task and 0.926–0.932 on the delayed task, each within ±0.007 of its task mean; Figure 6a), so the input–output computation is held fixed and only the architecture varies.

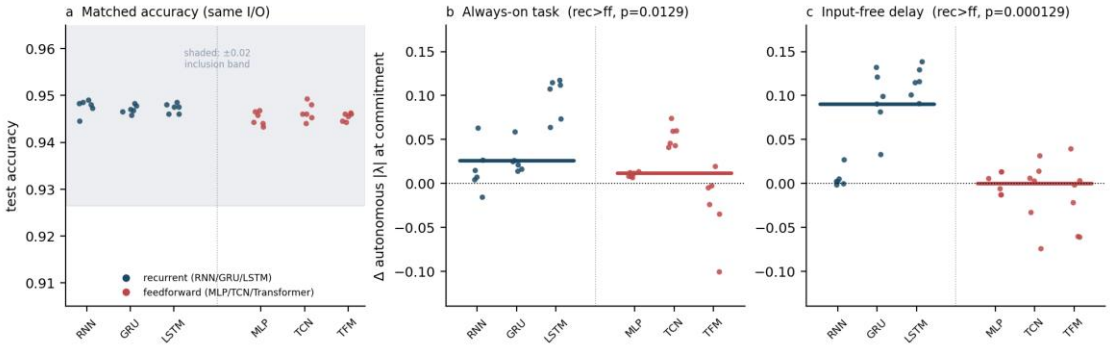


Figure 6. A computation-matched control: the eigenvalue commitment-attractor measure separates recurrent from feedforward networks at matched input–output mapping and accuracy. Recurrent (RNN, GRU, LSTM) and feedforward (windowed MLP, dilated causal TCN, causal transformer) networks, six seeds each, trained on an identical Poisson-clicks decision; design and metric pre-registered and frozen before any dynamical quantity was computed. (a) Final-bin test accuracy by architecture; every network falls within the ±0.02 inclusion band (shaded), so only the architecture varies. (b) Per-network change in autonomous dominance $|\lambda|$ from before to after commitment on the always-on task; recurrent (blue) versus feedforward (red), with family medians (horizontal bars); recurrent > feedforward, one-sided $p = 0.013$. (c) The same on a delayed-response task with a zero-input epoch after the evidence, where the separation is sharper ($p = 1.3\times10^{-4}$).

Alt text: three panels; the left shows all six architectures' accuracies clustered inside a narrow band; the middle and right show the change in autonomous eigenvalue per network, with recurrent networks shifted above feedforward networks, more strongly in the input-free delayed task.

Holding computation fixed, the autonomous-eigenvalue change at commitment (the commitment-attractor measure of Figure 4) was larger in the recurrent family than in the feedforward family (one-sided $p = 0.013$; Figure 6b), and the separation reproduced, more sharply, on a second task with a zero-input delay after the evidence, where no input is available to drive the feedforward state once the choice is made ($p = 1.3\times10^{-4}$; Figure 6c). It held in all four runs (two tasks × two independent seed sets, $p \le 0.013$), replicated on fresh seeds, and survived a guard against a frozen-state artefact. The qualifications are stated plainly: the effect is a family-level shift, carried by the gated recurrent networks and by the transformer at the two extremes rather than a clean separation of every instance, and the variance-partition read-out, which the convolutional network can inflate, was not diagnostic, so the result rests on the eigenvalue and autonomous-flow measures the paper leads with. This is a model-organism control, not a re-matching of the cortex–model pair; what it establishes is that the commitment-attractor measure is the kind of signature that tracks recurrent-versus-feedforward architecture under matched computation.

Negative Results: The Commitment Effect Is Invisible to Trial-Averaged and Scalar Read-Outs

A central part of the evidence is a set of pre-registered tests that returned nulls, each reported with its kill-condition (Figure 5). These nulls are not incidental: they explain why a time-locked dynamical measure was necessary and they bound what the positive results mean.

 doi:10.20944/preprints202608.1095.v1

First, before the commitment-aligned analysis above, the test asked whether the autonomous component in rat OFC tracks commitment when read out trial-averaged rather than time-locked. It does not. Contrasting committed against abandoned trials by a wait-time tertile split returned a null (median Δ autonomous share +0.0012, Wilcoxon p = 0.71 over 113 sessions; the per-animal sign test was likewise null), as did an opt-in versus opt-out contrast (−0.0014, p = 0.75 over 170 sessions). Second, a test asked whether the autonomous component predicts a scalar behavioural confidence proxy (post-decision wait time) trial by trial, controlling for value, reaction time and firing rate; across 170 sessions and 9,106 opt-out trials this was a well-powered null (median partial Spearman −0.003, Wilcoxon p = 0.51), with a value-separability negative control also null. The lesson of these nulls is consistent and was anticipated: a dynamical regime transition that is locked to the commitment *moment* is smeared away by trial averaging and is not captured by a scalar magnitude. The OFC nulls are also area-appropriate (OFC is a value/state area, not the frontal commitment circuit in which the perceptual-decision commitment transition lives), which is why the commitment-locked attractor was established in the frontal Neuropixels data rather than in OFC. Reporting these nulls is what makes the positive commitment result credible: the effect appears exactly where a commitment-locked dynamical measure is applied to the commitment circuit, and nowhere a trial-averaged or scalar measure is applied instead.

Discussion

What the Results Show

Two substrates were set the same kind of cognitive work and compared on a single measurable property: the balance of their dynamics between externally driven and internally generated change. On that property they diverge categorically. Cortical populations are autonomous-dominated, generating their trajectories largely from their own recurrence; five transformer language models are input-driven, with roughly half of their across-token dynamics explained by the information fed in. Cortex recruits more of its autonomous component when a task demands inference; the models mostly do not. And at the behavioural moment of committing to a perceptual decision, the cortical autonomous attractor strengthens, locked to commitment rather than to the stimulus — a transition already reported for these data and recovered here with the subspace demix its analysts and method authors describe (Luo et al. 2025; Huang et al. 2025). The feedforward model shows no such attractor, and its absence is structural: a stack with no recurrence across its depth has no autonomous temporal regime to settle into, and it commits only at its output layer.

The instrument earns these claims. It was validated on synthetic systems before any recorded data were touched, and it recovers a known autonomous gain without contamination from input strength where a naive estimator fails. The cross-substrate comparison is restricted to dynamical regime and sign, never to magnitude. And the positive commitment result is bracketed by pre-registered nulls showing the same effect is invisible to the trial-averaged and scalar read-outs one would otherwise reach for first.

Bearing on Computation Versus Implementation

Whether conscious access is fixed by computation alone is the question this study was built to inform. The unfolding argument (Doerig et al. 2019) presses its hard form: because a feedforward system can in principle match the input–output behaviour of a recurrent one, behaviour alone cannot adjudicate theories that rest on internal dynamics, so such a theory must be tested on the dynamics, not on behaviour. That is why this study looks at internal dynamics at all, and it is worth being plain about what that does and does not license. This study did not build the behaviourally equivalent pair the argument specifies, a feedforward and a recurrent system computing the same input–output function, so it cannot claim to have isolated implementation from computation in the strict sense the argument demands. What it shows is weaker but still pertinent. Two systems doing analogous answer-construction tasks, a cortex and a feedforward transformer, differ in the dynamics of reaching the answer in a specific, measurable way: one constructs and commits through autonomous

 doi:10.20944/preprints202608.1095.v1

recurrence that strengthens into an attractor at commitment, and the other does not. Because the brain–model comparison spans different tasks and different architectures at once, it cannot by itself attribute the difference to substrate rather than to task. So that strict test was run directly, as a pre-registered control reported above (Figure 6): recurrent and feedforward networks trained to the *same* input–output function at matched accuracy, with the demix applied unchanged. Holding the computation fixed and varying only the architecture, the eigenvalue commitment-attractor measure separated the families in every run, and most sharply when no input followed commitment. That makes it the *kind* of signature that tracks recurrent-versus-feedforward architecture under matched computation rather than the task — which is what the unfolding argument demands be shown on the dynamics rather than on behaviour. It is a model-organism control, not a re-matching of the cortex–model pair, and it is read as such. With that in hand, the difference reported here between a cortex and a feedforward model is of the kind the unfolding argument says behaviour cannot reveal, recovered from internal dynamics with a validated instrument. To the extent that theories of access tie it to recurrent or intrinsic dynamics, whether recurrent processing theory (Lamme 2006), the ignition dynamics of global-workspace accounts (Mashour et al. 2020), or the intrinsic cause–effect structure of integrated information theory (Tononi et al. 2016), autonomous recurrence is a property on which a cortex and a feedforward model demonstrably differ. Whether that difference matters for access is for those theories to say; the contribution here is to make it measurable and to put it on one cross-substrate axis.

The positive claim needs stating precisely, because it is easy to overreach here. Nothing in these data could show that the brain is conscious and the model is not, and no such claim is made. What the data show is that a particular dynamical signature, of a kind several theories of access treat as load-bearing, is present in cortex and absent from a feedforward model doing an analogous task. If a theory holds that this signature is necessary for access, the result is evidence that the model lacks a necessary condition for access, though the antecedent is the theory's to defend and is not adjudicated here. If a theory holds that access is fixed by computation alone, the result is a difficulty for it, because two systems doing analogous answer-construction tasks differ on the signature, even though (as noted above) they are not matched on their input–output function and the comparison cannot by itself separate substrate from task. Either way the contribution is the same: a decidable, cross-substrate measurement the computation-versus-implementation debate can be argued over, rather than one more conceptual stipulation.

The Central Limitation: Commitment Is Not Access

The sharpest objection to reading these results as bearing on consciousness is that what is measured here is *decision commitment*, and decision commitment is not conscious access. The point is correct, and it is raised here rather than left for a reader to press. The commitment event in the perceptual-decision data is the moment a rat's accumulator locks onto a choice; it is defined behaviourally and dynamically, and it can occur without the choice being consciously reportable in the sense the access literature intends (Block 1995). Conscious access plausibly involves more: a global availability of the committed content for report and the flexible control of behaviour (Mashour et al. 2020), which the present measure does not touch. So the autonomous attractor found here is, at most, a candidate dynamical correlate of *a* commitment step that access may require, not a measurement of access itself. Three gaps separate the present claim from a claim about access, named here so the reader can weigh the inference: the species and task gap (a rodent perceptual decision is far from a human report of conscious content); the commitment-to-access gap (a locked-in decision need not be a consciously accessed one); and the access-to-phenomenality gap, not attempted at all. The empirical results stand independently of any of these bridges. The bearing on consciousness is an argument about what the results *could* mean for the debate, offered for the reader to accept or reject, not a finding.

A second limitation is the model side of the commitment comparison. The strongest model-side claim is structural (a feedforward stack has no autonomous regime to settle into), and it does not need

 doi:10.20944/preprints202608.1095.v1

the noisy across-layer eigenvalue measure. A cleaner test, an across-token attractor measured within the forward pass, was therefore run on the real models in generation mode (Results), in a first pass and a pre-registered, properly powered follow-up; neither found a commitment-locked attractor in any model (no significant rise across the model-by-depth tests, minimum $p = 0.07$). Because the eigenvalue estimator is noisy on short generated sequences, and one model committed to its answer too rarely to test, this corroborates the structural claim rather than sharpening it, and the model leg is still reported honestly as the weaker half of the dissociation. Third, the second neural anchor (primate entorhinal cortex) is underpowered (14 sessions) and its eigenvalue read-out is null; it is reported as supporting, not decisive. Fourth, the autonomous attractor strengthening, though commitment-locked and well powered, is small in absolute magnitude and is a ramp rather than a discrete switch.

Relation to Prior Work

The instrument used here is built on input-aware dynamical similarity analysis (Huang et al. 2025; Ostrow et al. 2023), and that method debt is owned plainly. The method paper applies the demix within brains (including to the very perceptual-decision recordings reanalysed here, where it reports the input-to-intrinsic transition at commitment) and within artificial recurrent networks; it does not cross a language model and a brain, and it does not report a signed cross-substrate difference. The contribution here is therefore not the demix, nor the brain transition (which is theirs and Luo et al.'s), but the signed cross-substrate comparison: a brain and a language model placed on one autonomous-versus-input axis on matched cognitive work, with the commitment-locked attractor present in one and absent from the other. That intrinsic-versus-input neural dynamics can be dissociated at all is itself established prior art (Vahidi et al. 2024, 2025); the novelty here is the comparison across substrates, not the dissociation within one. The identifiability problem navigated here, that recurrent and input-driven linear models are degenerate in cortex unless the estimator is constrained, is precisely the one documented by Soldado-Magraner et al. (2024), and the subspace estimator is the added constraint that breaks the degeneracy. Two cross-substrate neighbours bound the contribution from the other side. The closest trajectory study compares neural and language-model dynamics during reading and finds a continuous-versus-bursty contrast (Xiao et al. 2025), but it uses trajectory-geometry descriptors rather than a recurrent-versus-input demix and studies language processing rather than commitment; a separate line compares language-model and neural dynamics through energy-landscape attractor analysis (Watanabe et al. 2025), a lens that is neither a recurrent-versus-input demix nor anchored to a behavioural commitment event. The mechanistically typed claim made here, autonomous recurrence versus external drive, demixed and validated and tied to a commitment event the brain shows and the model does not, is not one those pipelines can make.

Falsifiability and What Would Overturn This

The account is built to be wrong in stated ways (Doerig et al. 2021; Kleiner and Hoel 2021). The recruitment claim would fall if the brain's autonomous component did not rise under inference; it rose in rat OFC ($p \approx 1.3\times10^{-6}$ across sessions, on both read-outs). The commitment claim would fall if the autonomous strengthening were not locked to commitment; it was sharper aligned to commitment than to stimulus ($p = 0.027$). The balance observation would fall if a model, measured the same way, carried the same autonomous-versus-input balance as the brain; subject to the input-dimensionality caveat in Methods, none did. The cleanest falsification targets the model leg: finding a commitment-locked attractor in a feedforward model, by an across-token analysis, would refute the structural reading and would itself be a major result. That test was run in generation mode across the models, in a first pass and a properly powered follow-up, and found no such attractor (minimum $p = 0.07$), in the direction the structural reading predicts; a definitive higher-powered version, across more items and models, is the obligation the present model leg leaves open.

 doi:10.20944/preprints202608.1095.v1

Summary and Conclusions

On matched cognitive work, cortical populations and feedforward transformer language models occupy opposite dynamical regimes: cortex is autonomous-dominated and recruits its recurrence under inference, and at the behavioural moment of committing to a perceptual decision its autonomous attractor strengthens, locked to commitment; the models are input-driven, mostly do not recruit autonomy, commit only at their output layer, and have no commitment-locked attractor to measure. The dissociation is signed, validated on synthetic ground truth, restricted to dynamical regime rather than magnitude, and bracketed by pre-registered nulls that show why a time-locked measure was required. It is read here as an empirical purchase point on the computation-versus-implementation question: a dynamical property used to reach a commitment is present in one substrate and absent from a feedforward model performing an analogous task, which is a difficulty for accounts on which access is fixed by computation alone. No claim is made to have measured conscious access or phenomenal experience, and the gap between decision commitment and conscious access is named as the limitation a reader should weigh most heavily. What the study offers the debate is not a verdict but a decidable measurement and a method for collecting more.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

Ethics: This study is a reanalysis of publicly available datasets and publicly released model weights. No new human or animal data were collected, and no ethics approval was required; the original recordings were collected under the ethical approvals reported in the primary publications cited.

Declaration of AI use: In the interest of full transparency, I disclose that a large language model (Anthropic Claude, Opus-family models, 2026) was used, under my direction, in two supporting roles: (i) as a coding and technique-lookup aid while writing and debugging the analysis scripts, used in the manner of a search engine to help identify suitable code and statistical methods; and (ii) as a drafting and language-editing aid for the manuscript prose. The tool was not used to generate the study's scientific insights, to analyse or interpret the data, or to draw its conclusions: the research questions, the specification of every analysis, and all interpretations and conclusions are my own. Every analysis was specified, run and checked by me against the deposited result files, and every statistical value reported here was recomputed by me from those files. The work in its totality is an accurate representation of my underlying research and novel intellectual contributions and is not primarily the product of the tool's generative capabilities; I have checked the manuscript for accuracy, including verifying that every reference is real and correct, and it contains no AI-introduced plagiarism, misrepresentation or fabrication. I accept responsibility for the veracity and correctness of all material in the manuscript, including any computer-generated material. The tool is not an author and is not listed or cited as one; I take full responsibility for the work as submitted.

Conflicts of Interest declaration: I declare I have no competing interests.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The work was conducted independently under the ICSAC Institute.

Acknowledgments: The author thanks the laboratories that made their single-unit and Neuropixels recordings publicly available; this study is a reanalysis of their data and would not have been possible otherwise.

Data availability: All data analysed here are public. Rat orbitofrontal recordings are from the Constantinople laboratory (Schiereck et al. 2026; data at Zenodo, accession 16997337). Primate entorhinal recordings are from Neupane et al. (2024; DANDI accession 000897). The perceptual-decision Neuropixels recordings are from Luo et al. (2025; raw data at Dryad, doi:10.5061/dryad.sj3tx96dm; per-trial commitment times from the authors' repository, github.com/Brody-Lab/decision_dynamics_commitment). Language-model activations were generated from publicly released model weights (Qwen/Qwen2.5-{1.5B,3B,7B}-Instruct; meta-llama/Llama-3.1-8B-Instruct; google/gemma-2-9b-it). All analysis code, the construction-task item set, the result files underlying every figure and statistic, the synthetic-validation results (with the random seed fixed), the demix library and

 doi:10.20944/preprints202608.1095.v1

per-substrate analysis scripts with run instructions, and the two frozen pre-registrations with their code and per-model and per-network result files — the computation-matched architecture control and the real-model across-token generation test — accompany this preprint as supplementary material. Each pre-registration fixed its design, metric and kill conditions before any result was computed. The reanalysed datasets are cited in the reference list. No new human or animal data were collected.

Author Contributions (CRediT): N.M.T.: conceptualization, methodology, software, formal analysis, investigation, data curation, writing—original draft, writing—review and editing, visualization. The author is the sole contributor, gave final approval for publication, and agrees to be accountable for all aspects of the work.

References

1. Belrose N, Ostrovsky I, McKinney L, Furman Z, Smith L, Halawi D, Biderman S, Steinhardt J. 2023. Eliciting latent predictions from transformers with the tuned lens. arXiv:2303.08112.
2. Block N. 1995. On a confusion about a function of consciousness. *Behavioral and Brain Sciences* 18(2):227–247. doi:10.1017/S0140525X00038188.
3. Doerig A, Schurger A, Hess K, Herzog MH. 2019. The unfolding argument: why IIT and other causal structure theories cannot explain consciousness. *Consciousness and Cognition* 72:49–59. doi:10.1016/j.concog.2019.04.002.
4. Doerig A, Schurger A, Herzog MH. 2021. Hard criteria for empirical theories of consciousness. *Cognitive Neuroscience* 12(2):41–62. doi:10.1080/17588928.2020.1772214.
5. Huang A, Ostrow M, Singh SH, Kozachkov L, Fiete I, Rajan K. 2025. InputDSA: demixing then comparing recurrent and externally driven dynamics. arXiv:2510.25943 (ICLR 2026).
6. Kleiner J, Hoel E. 2021. Falsification and consciousness. *Neuroscience of Consciousness* 2021(1):niab001. doi:10.1093/nc/niab001.
7. Lamme VAF. 2006. Towards a true neural stance on consciousness. *Trends in Cognitive Sciences* 10(11):494–501. doi:10.1016/j.tics.2006.09.001.
8. Luo TZ, Kim TD, Gupta D, Bondy AG, Kopec CD, Elliott VA, DePasquale B, Brody CD. 2025. Transitions in dynamical regime and neural mode during perceptual decisions. *Nature* 646(8087):1156–1166. doi:10.1038/s41586-025-09528-4.
9. [dataset] Luo TZ, Kim TD, Gupta D, Bondy AG, Kopec CD, Elliott VA, DePasquale B, Brody CD. 2025. Data from: Transitions in dynamical regime and neural mode during perceptual decisions. Dryad Digital Repository. doi:10.5061/dryad.sj3tx96dm. Per-trial commitment times: github.com/Brody-Lab/decision_dynamics_commitment.
10. Mante V, Sussillo D, Shenoy KV, Newsome WT. 2013. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* 503(7474):78–84. doi:10.1038/nature12742.
11. Mashour GA, Roelfsema P, Changeux J-P, Dehaene S. 2020. Conscious processing and the global neuronal workspace hypothesis. *Neuron* 105(5):776–798. doi:10.1016/j.neuron.2020.01.026.
12. Neupane S, Fiete I, Jazayeri M. 2024. Mental navigation in the primate entorhinal cortex. *Nature* 630(8017):704–711. doi:10.1038/s41586-024-07557-z.
13. [dataset] Neupane S, Fiete I, Jazayeri M. 2024. Mental navigation in the primate entorhinal cortex. DANDI Archive, dandiset 000897. https://dandiarchive.org/dandiset/000897.
14. Nostalgebraist. 2020. Interpreting GPT: the logit lens. LessWrong. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens (technical note).
15. Ostrow M, Eisen A, Kozachkov L, Fiete I. 2023. Beyond geometry: comparing the temporal structure of computation in neural circuits with dynamical similarity analysis. *Advances in Neural Information Processing Systems* 36. arXiv:2306.10168.
16. Schiereck SS, Pérez-Rivera DT, Mah A, DeMaegd ML, Hocker D, Ward RM, Savin C, Constantinople CM. 2026. The orbitofrontal cortex updates beliefs for state inference. *Neuron* 114(3):507–520.e8. doi:10.1016/j.neuron.2025.10.024.
17. [dataset] Schiereck SS, Pérez-Rivera DT, Mah A, DeMaegd ML, Hocker D, Ward RM, Savin C, Constantinople CM. 2026. Data from: The orbitofrontal cortex updates beliefs for state inference. Zenodo, accession 16997337. https://doi.org/10.5281/zenodo.16997337.

 doi:10.20944/preprints202608.1095.v1

18. Searle JR. 2017. Biological naturalism. In: Schneider S, Velmans M, editors. *The Blackwell Companion to Consciousness*, 2nd ed. Wiley-Blackwell. p. 327–336. doi:10.1002/9781119132363.ch23.
19. Soldado-Magraner J, Mante V, Sahani M. 2024. Inferring context-dependent computations through linear approximations of prefrontal cortex dynamics. *Science Advances* 10(51):eadl4743. doi:10.1126/sciadv.adl4743.
20. Tononi G, Boly M, Massimini M, Koch C. 2016. Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience* 17(7):450–461. doi:10.1038/nrn.2016.44.
21. Vahidi P, Sani OG, Shanechi MM. 2024. Modeling and dissociation of intrinsic and input-driven neural population dynamics underlying behavior. *Proceedings of the National Academy of Sciences* 121(7):e2212887121. doi:10.1073/pnas.2212887121.
22. Vahidi P, Sani OG, Shanechi MM. 2025. BRAID: input-driven nonlinear dynamical modeling of neural-behavioral data. *International Conference on Learning Representations (ICLR) 2025*. arXiv:2509.18627.
23. Van Overschee P, De Moor B. 1994. N4SID: subspace algorithms for the identification of combined deterministic–stochastic systems. *Automatica* 30(1):75–93. doi:10.1016/0005-1098(94)90230-5.
24. Watanabe T, Inoue K, Kuniyoshi Y, Nakajima K, Aihara K. 2025. Comparison of large language model with aphasia. *Advanced Science* 12:2414016. doi:10.1002/advs.202414016.
25. Xiao X, Wei K, Zhong J, Wei X, Zhou M. 2025. Exploring similarity between neural and LLM trajectories in language processing. arXiv:2509.24307.