

Article

Not peer-reviewed version

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

Nathan M. Thornhill *

Posted Date: 8 October 2026

Keywords: decision commitment; intrinsic neural dynamics; recurrent dynamics; neural population dynamics; perceptual decision-making; evidence accumulation; transformers; large language models



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Article

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

Nathan M. Thornhill

Institute for Complexity Science and Advanced Computing (ICSAC),
PO Box 15008, Fort Wayne, IN 46885, USA; nthornhill@icsac institute.org

Abstract

Whether the dynamics that carry a decision to commitment differ between the recurrent circuits of the brain and the transformer networks of large language models is a question about internal dynamics. For any recurrent network there is a feedforward network with matched behavior, so claims resting on internal dynamics must be tested directly. A dynamics demix separates input-driven change from change generated by a population's own dynamics. Validated first on synthetic systems and checked against the mean-field variance theory of driven echo-state networks, it was applied to cortical populations in perceptual and value-based decisions and to five feedforward transformer language models constructing multi-step answers. Every neural dataset was autonomous-dominated and the five language models were input-dominated; under inference, orbitofrontal cortex recruited its autonomous component on both read-outs (24 of 31 rats on the variance partition), whereas four of the five models did not. For the two models and the cortical dataset re-examined at a matched input dimension of eight, the direction of that difference held. The models' regime label, however, depends on the dimension of the input basis, so the difference is reported as one of regime at matched input dimension rather than as categorical. After commitment to a perceptual decision the cortical autonomous dynamics became more persistent, more sharply when aligned to the inferred commitment time than to stimulus onset (a specificity supported rather than established); this recovers a transition already reported for these recordings. In the two language models probed layer by layer the correct answer, when detected, first became top-ranked before the output layer, under the read-out used; across generated tokens, no commitment-locked rise in autonomous persistence was found in any model. In a pre-registered control varying only architecture at matched accuracy, the eigenvalue measure separated recurrent from feedforward networks, most cleanly when no input followed the evidence; that control, not the cortex-model comparison, carries the architectural claim. Pre-specified negative tests in orbitofrontal cortex found no commitment signal of the predicted sign in trial-averaged or scalar read-outs. Conclusions are limited to the systems tested. No claim is made about phenomenal experience or machine sentience.

Keywords: decision commitment; intrinsic neural dynamics; recurrent dynamics; neural population dynamics; perceptual decision-making; evidence accumulation; transformers; large language models

Introduction

A decision ends in a commitment: the moment a system settles on one answer and can use it. What carries a system to that moment is a question about internal dynamics, and behavior alone does not answer it. This work measures how much of a population's dynamics is generated by its own recurrence and how much is driven by input arriving from outside, and asks whether the balance differs between a recurrent cortical population and a feedforward model given analogous work. Accounts that turn on internal dynamics have been faulted for resting on criteria no experiment can overturn (Doerig et al., 2019; Kleiner and Hoel, 2021), so a usable account 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). Where the results bear on conscious access, meaning how a represented item becomes

available for report, reasoning and the control of behavior (Block, 1995), the argument is made in the Discussion rather than assumed here. No claim is made about phenomenal experience, and none about machine sentience.

“Decision” is used here in the sense of the decision-neuroscience literature (Gold and Shadlen, 2007; Shadlen and Kiani, 2013): the process by which a system selects one of a small set of options on the basis of evidence or value. “Decision commitment” is the moment at which that selection is locked in and becomes available to guide action, whether or not evidence continues to arrive (Resulaj et al., 2009; Kiani et al., 2014). The decisions studied are of three kinds, each defined by the task structure of the dataset it comes from: perceptual (a rat categorizing a stream of auditory clicks), value-based (a rat deciding whether to wait for an offered reward) and inferential (a monkey mentally navigating between landmarks it cannot see). Human decisions, from the next move in a board game to the choice of a career, are shaped by an interplay of cognition and emotion in which affective state primes and biases the choice, and cognition in turn regulates emotion (Bechara et al., 2000; Pessoa, 2008; Lerner et al., 2015). A full account of decision making has to include that loop. This work does not attempt one. It asks a narrower question about the dynamics that carry a perceptual, value-based or inferential decision to commitment in a recurrent cortical population and in a feedforward model given analogous work, and its conclusions are limited to that question.

The question underneath is whether two systems that behave alike must reach their decisions through the same kind of internal dynamics. The unfolding argument makes the difficulty precise (Doerig et al., 2019): for any recurrent network there is a feedforward network with the same input–output mapping, so behavior alone cannot settle claims that rest on a system’s internal causal organization. If two systems can be made to behave identically while differing in their internal dynamics, a theory that ties a capacity to those dynamics has to be tested on the dynamics directly, not on behavior. The contrast that matters here is therefore feedforward versus recurrent dynamics, examined directly in a recurrent biological system and in a feedforward artificial one given analogous work.

Language models make a useful second system. 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 (Vaswani et al., 2017). Across tokens (the word pieces a language model reads and writes one at a time) the picture differs. An autoregressive model feeds each token it emits back in as input to the next pass, and a chain-of-thought answer is a sequence of such steps, so generation is recurrent at the level of tokens. That token-level recurrence is real and is measured directly below (Methods; Results; Discussion). Within a pass the model can approximate input–output mappings that a recurrent brain also computes, which is exactly the behavioral equivalence the unfolding argument invokes. This work asks whether, behind that equivalence, the *dynamics* of reaching the answer differ between the two kinds of system, 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 dynamics: 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*. Cortical computation is usually modeled as recurrent dynamics that integrate and transform inputs rather than relay them (Mante et al., 2013). The idea that a decision is the settling of a recurrent population into one of a small set of stable states has a long history in the field (Wang, 2002; Wong and Wang, 2006). 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., 2025a). Pulling input and recurrence apart in recorded activity is hard, because the two 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 separates the autonomous component from an exogenous input.

That shift is what a demix of this kind can test directly rather than infer. The same quantity carries weight 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 nonlinear ignition, an abrupt and self-amplifying surge of activity 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 system's intrinsic cause-effect structure rather than its input-output behavior (Tononi et al., 2016), and the unfolding argument is aimed at that family. Across these otherwise opposed positions one theme recurs: what matters is a system's internal, autonomous organization, 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 work reports four empirical results and a set of negative ones. First, on analogous cognitive work, cortical populations are autonomous-dominated while five transformer language models are input-dominated, a difference in dynamical regime whose direction survived, for the two models and the cortical dataset re-examined, matching the input dimension of the two analyses at eight. Second, cortex raises its autonomous component when a task demands inference; the models mostly do not. Third, after commitment to a perceptual decision the cortical autonomous dynamics become more persistent, more sharply when aligned to the inferred moment of commitment than to the stimulus, a specificity supported rather than established. This recovers, with the subspace demix of Huang et al. (2026), a regime transition already reported for these data (Luo et al., 2025a). In the two models probed layer by layer the correct answer, when detected, first becomes top-ranked before the output layer, under the read-out used; across generated tokens, no comparable commitment-locked rise was found in any model. Fourth, in a pre-registered control holding the task fixed and varying only the architecture, the same eigenvalue measure separated recurrent from feedforward networks, most cleanly when no input followed the evidence; that control, not the cortex-model comparison, carries the architectural claim.

This work is equally explicit about what it does not establish. Pre-specified tests in orbitofrontal cortex returned nulls on their pre-specified read-outs, reported in full: one asked whether a scalar population read-out tracks a behavioral confidence proxy (how long a rat waits for a reward), the others whether the autonomous component differs between trials the rat committed to and trials it abandoned. The positive commitment result rests on a measure time-locked to commitment in a different circuit. The cortex-model comparison spans different tasks, different functional demands and different architectures at once, so on its own it cannot attribute the difference to architecture rather than to task. A cortical circuit must also hold a committed choice available for the action that follows, whereas a model's answer is the end of its computation. That is why a pre-registered control holding the task and accuracy fixed and varying only architecture was run, and why the cortex-model contrast is read alongside it as descriptive.

Materials and Methods

Overview and the Demixing Instrument

The core instrument separates two sources of change in a population trajectory: the share of next-state variance generated by the population's own recurrence, and 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 modeled as a linear time-invariant system, $\mathbf{x}_{t+1} = \mathbf{A} \mathbf{x}_t + \mathbf{B} \mathbf{u}_t + \text{noise}$. Here $\mathbf{A}$ is the autonomous operator and $\mathbf{B}$ is the input operator. Ordinary least squares on the observed components is unreliable here: they are a noisy, partial view of a larger state, so regressor noise biases $\mathbf{A}$, and input correlated with the unobserved state lets shared variance go to either $\mathbf{A}$ or $\mathbf{B}$. This identifiability problem is well documented for

recurrent-versus-input decompositions of cortical activity (Soldado-Magraner et al., 2024). The autonomous operator A was therefore estimated by subspace identification (N4SID, numerical algorithms for subspace state-space system identification; Van Overschee and De Moor, 1994). The method first estimates the latent state sequence by projecting the data along the input, then fits A and B to that sequence by least squares. N4SID is the estimator at the core of input-aware dynamical similarity analysis (Huang et al., 2026; building on Ostrow et al., 2023). Two complementary read-outs were taken, chosen so that at least one is robust in each regime:

- **Autonomous dominance** (`smdmc_dom`; the eigenvalue read-out): the magnitude of the dominant eigenvalue of the identified autonomous operator A (the name refers to that eigenvalue and is unrelated to the autonomous-dominated regime label defined below). A value below 1 denotes a contractive mode of the fitted linear operator; the closer the value is to 1, the more slowly a perturbation of the autonomous state decays, so a rise toward 1 within the stable range denotes greater persistence of that state. A value at or above 1 denotes an expansive or marginally stable mode. This read-out is this work’s measure of autonomous persistence (see Terminology below) but, as the synthetic validation shows, it is trustworthy only when the input is exogenous, that is, generated outside the population and not fed back from its state (open-loop input, as in the brain decision data).
- **Autonomous unique share** (`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 the population’s own state over and above the input. It is robust to eigenvalue blow-up and is defined identically in the two kinds of system, and it carries the balance comparison between them.

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 is called input-dominated when `r2_input` exceeds `auto_unique`, and autonomous-dominated otherwise.

Terminology. Four terms are used and kept distinct. *Recurrence* is a structural property: a directed loop in the connectivity, by which activity can return to the units it came from (Lamme and Roelfsema, 2000; Kar et al., 2019); a chain in which each stage feeds only the next is feedforward, although the population’s state at one time then still predicts its state at the next. Where the text speaks, in the field’s idiom, of variance *generated by the population’s own recurrence*, it means the measured autonomous share defined next, which recurrent wiring produces but which a functionally feedforward chain can also produce (Supplementary Note S3); the attribution to recurrent wiring is made only where the wiring is known. *Autonomous dynamics* is a measured quantity: the share of next-state variance a population generates from its own state beyond what the known input explains (`auto_unique`), together with the operator A that generates it. *Persistence* is a property of that operator: the magnitude of its dominant eigenvalue, which for a stable system sets how slowly the autonomous state contracts; a rise toward 1 is greater persistence, not in itself a stronger pull toward a fixed point. This is the eigenvalue sense of persistence; a system can also hold information transiently through non-normal structure (a feedforward-like chain of modes along which a signal is passed) without any eigenvalue near 1 (Supplementary Note S3), a second sense that this measure does not capture. *Attractor* is reserved for a state toward which trajectories are shown to converge; the recurrent attractor account of decisions (Wang, 2002) is one model of how such a state could carry a choice. The commitment-locked rise in the dominant eigenvalue reported below is therefore described as increased autonomous persistence, because settling onto a fixed point was not itself demonstrated.

Attracting or persistent states can also arise without recurrence: a purely feedforward chain of stages can carry an input trace forward and hold it for as many steps as the chain has stages (Goldman, 2009; Ganguli et al., 2008). In layered networks, stable fixed points of the layer-to-layer map act as attractors (Domany et al., 1989), and feedforward computation can perform attractor-like error correction, although activity that persists without limit after the input ends requires recurrent dynamics (Khona and Fiete, 2022). A feedforward cascade driven by a constant input likewise

converges to a state fixed by that input. What the demix does with such systems was checked on synthetic examples (Supplementary Note S3). The variance partition alone does not distinguish a feedforward memory chain from recurrence: both show an autonomous unique share of similar size, because in both the next state is predictable from the current one beyond the input. The eigenvalue read-out separates them only in part. The chain's true autonomous operator is nilpotent, every eigenvalue zero, but the estimate of a near-nilpotent operator is numerically ill-conditioned (a generic perturbation of size ε moves its eigenvalues by about $\varepsilon^{1/K}$ for a chain of K stages). The read-out therefore returned a spurious mid-range value of about 0.5 on the chain, against 0.90 recovered exactly on a recurrent comparator of known gain. That spurious value also drifted between halves of the run by amounts larger than the cortical commitment effect, whereas the recurrent comparator's estimate did not move. A static feedforward map ($A = 0$) was attributed entirely to the input by the variance partition, with an autonomous unique share of zero, but its eigenvalue estimate (0.27 against a true 0) was still wrong, here from fitting a rank-eight operator to a system with no autonomous dynamics. The eigenvalue read-out is not trusted on any trajectory that is a feedforward chain by construction, which includes a transformer's across-layer trajectory (Results). And the demix cannot certify that the autonomous persistence measured in cortex is carried by recurrent rather than functionally feedforward circuitry (Goldman, 2009). What it shows is a change in the population's autonomous persistence after commitment, sharper when aligned to commitment than to stimulus in the primary windows, a specificity supported rather than established (Results). The attribution of such a change to recurrent architecture is made only in the matched control (Results), in which the presence or absence of recurrent connections is known by construction.

Synthetic Validation of the Instrument

Before any recorded data were analyzed, 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, moved with input strength, drifting upward at the lowest gain (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 at moderate coupling and detected a true difference in recurrence against a trial-shuffle null, although at the strongest coupling it no longer ranked the true gains. These validations fix the division of labor used throughout: auto_unique carries the comparison between the two kinds of system; sdmcdom is reported where the input is exogenous (the brain decision data) and interpreted with caution elsewhere. One further caveat 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 recruitment contrast in orbitofrontal cortex is therefore corroborated with the autonomous-eigenvalue read-out, which was input-invariant in the synthetic validation (Figure 1a) and passed the corresponding negative control.

The linear demix was further checked against an analytic theory of driven nonlinear recurrent networks: the self-consistent mean-field split of membrane-potential variance into recurrent and external shares for tanh echo-state layers (Jaeger and Haas, 2004; and, for the spiking analog, Maass et al., 2002). These are fixed random recurrent networks with a tanh nonlinearity, whose recurrent input the theory treats as Gaussian noise of self-consistently determined variance (Schubert and Gros, 2021, their Section 5.6; Massar and Massar, 2013). Networks were built to that theory's specification, the demix was applied unmodified, and the predictions were pre-registered before the analysis code was written (Supplementary Note S1). The check also exposed a limit of the eigenvalue read-out:

applied to eight principal components of the activity under a shared low-dimensional input, it became unreliable. Applied to the full membrane-potential state, it recovered the spectral radius of the network’s mean linear-response operator to within 0.035 (the persistence quantity above, not the effective spectral radius $q(aW)$, for gain a and weights W , of Schubert and Gros, 2021; Supplementary Note S1, Supplementary Figure S2). The cortical eigenvalue results are therefore read for within-session changes and their direction, with the stimulus-aligned comparison as a partial control for the commitment-locked one (Results), and not for their level. Full synthetic results are in the deposited results/inputdsa_synthetic2.json and results/validate3.json, and the echo-state check in Supplementary Note S1.

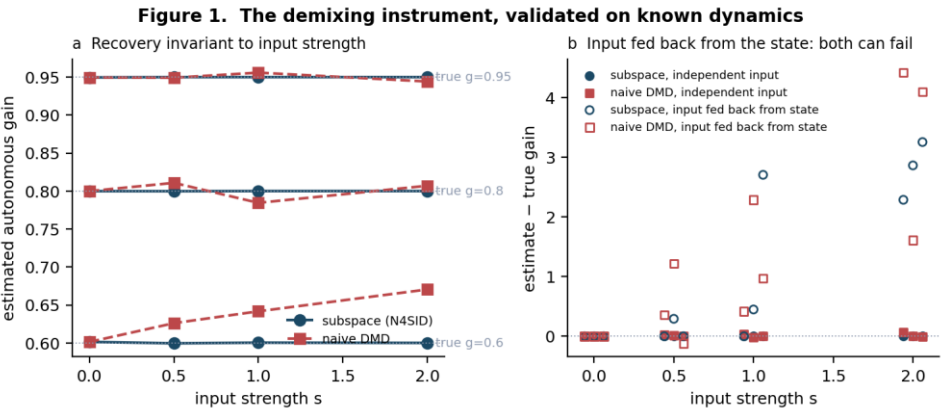


Figure 1. The demixing instrument and its validation on synthetic systems with known dynamics. (a) On synthetic linear systems with a known autonomous gain, the subspace estimator (N4SID) recovers the true gain and holds the estimate fixed as input strength is swept from 0 to 2.0 (points stay on the dotted true-gain lines), whereas ordinary dynamic mode decomposition (naive DMD) moves with input strength, upward at the lowest gain. (b) The error of each estimate (estimate minus true gain) against input strength, for input drawn independently of the state (filled) and for input fed back from the state (open). With independent input the subspace estimate stays on the true gain; with input fed back from the state both estimates can fail and either can exceed 1, which is why the eigenvalue read-out is trusted only where the input is exogenous (Methods). Alt text: left, the subspace estimate stays on the true autonomous gain across input strengths while the naive estimate drifts; right, estimate errors near zero for independent input and large, often above 1, when the input is fed back from the state.

Neural Datasets and the Input Signal

Three public single-unit and Neuropixels datasets were reanalyzed. In every case the external input u_t was defined from the task, not from the neural data, so that the demix tests autonomous dynamics against a task-defined 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; data: Schiereck et al., 2025) were reanalyzed; the recruitment analysis used 164 sessions from 31 rats. Population trajectories were reduced to their leading eight principal components per session. The input u_t was the task-locked, value- and time-modulated evoked drive (an onset kernel and its product with the offered value). Inference trials (those shortly after an unsignaled hidden-state switch or the start of a session, 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., 2024a; data: Neupane et al., 2024b) were reanalyzed across 14 sessions from two monkeys. The input u_t was a movement-onset kernel and its product with the number of steps navigated (two dimensions). 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 Scope and limitations).

Rat frontal cortex and striatum, perceptual decision with an inferred commitment time. Neuropixels recordings from rats performing the Poisson-clicks evidence-accumulation task (Brunton et al., 2013), with a per-trial estimate of the moment of commitment (the neurally inferred time of commitment, nTc) derived by the original authors (Luo et al., 2025a; data: Luo et al., 2025b, Dryad doi:10.5061/dryad.sj3tx96dm; commitment times from the authors' public repository), were reanalyzed across 115 sessions from 12 rats (approximately 36,000 frontal and striatal units, pooled per session without an area split; 19,095 committed trials). Because the striatal units are pooled with the frontal ones, results described below as cortical for this dataset include striatal activity. The input $\mathbf{u}_t$ was the accumulated sensory evidence (the running difference of left and right click counts and its change within each bin, from the recorded click times). Trajectories (the leading eight principal components per session) were binned at 25 ms in absolute time and aligned two ways: to the inferred commitment time, and to the stimulus onset. The autonomous read-outs were computed in sliding windows around each alignment point. The pre-commitment window was centered at -150 ms and the post-commitment window at +200 ms (each 200 ms wide: -250 to -50 and +100 to +300 ms), and the paired test compared the commitment-aligned rise with the stimulus-aligned rise within each session. The commitment time is not independent of the recorded activity: Luo et al. (2025a) infer nTc per trial by fitting a multi-mode drift-diffusion model jointly to the click trains, the animal's choices and the spike trains of the recorded population, and validate it behaviorally (evidence arriving after nTc no longer influences the choice). Aligning to nTc therefore asks whether a different property of the same recordings, the persistence of the demixed autonomous operator, changes at the event their model locates; the result is reported as a recovery of their transition (Results), with the stimulus-aligned comparison within each session as the internal control.

Language Models and the Matched Computation

Five instruction-tuned transformer language models spanning three families and a range of scales were analyzed: 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 40 inference items (20 true and 20 false statements) and 40 control construction items (34 inference items for the three Qwen models after a token-length filter; deposited README). Each item states a fact reachable only through an intermediate entity (a two-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 $\mathbf{x}_t$, and the input $\mathbf{u}_t$ was defined as the top eight principal components of 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 (N4SID with three delays here, two in the cortical commitment analysis and one in generation mode). Because a transformer has no recurrence within a forward pass, the across-token trajectory at a fixed depth is the closest analog of a temporal population trajectory. 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 mixing, which is treated as a limitation of the analog rather than as evidence of temporal recurrence. Across token positions an autoregressive model is also recurrent, since each generated token is fed back as input to the next pass; that token-level recurrence is measured by the generation-mode analysis (Results).

The commitment layer. For the model analog of decision commitment, the commit layer L^* was defined per true statement as the first layer at which the first token of its final word (the answer) 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-normalized per

layer to remove the confound of growing residual-stream norm, and an autonomous-dominance read-out was computed in sliding four-layer windows along the stack by a plain dynamic mode decomposition without an input regressor (none exists across layers), a cruder analog of the brain's time-resolved analysis. Statements whose answer never becomes top-ranked are left unassigned (the original analysis also scored the false statements, assigned the final layer to such items and aligned the windows to that L^* ; the scoring is corrected here, and because the windows were not re-estimated the read-out is shown by relative depth; Supplementary Note S5). L^* is a read-out marker: the layer at which the answer first becomes top-ranked under this particular read-out. It is not a claim that the model has decided in the sense in which a biological accumulator settles on a choice, and it is not in itself evidence of a dynamical state settling.

Statistics

Claims across the two kinds of system are made at the level of dynamical regime and sign only; no claim rests on comparing absolute magnitudes across them, because they depend on unit counts, recording modality and reduction choices. The two systems' baseline levels are summarized with statistics that are not identically defined: orbitofrontal and entorhinal 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 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, two dimensions) and higher-dimensional for the models (the top eight principal components of the layer-0 embedding), and a richer input basis can explain a larger share of a multi-component population's next state. The regime classification itself, not only the magnitude, can depend on this choice. A post-hoc analysis added in this version (Supplementary Note S2; Results) therefore matches the input dimensionality of the two analyses. It shows that the models' regime label depends on the dimension of the input basis while rat OFC's does not (at two input dimensions the two models re-examined read autonomous-dominated), and that the direction of the difference holds at a matched input dimension of eight. The claim across systems is stated accordingly as a difference in regime at matched input dimension (matched in dimension, not in informational content), not as categorical. The load-bearing comparisons remain the within-system contrasts (the recruitment of autonomy under inference, computed with the same input definition within each system) and the within-brain commitment-locked rise in persistence, neither of which compares levels across the two input definitions. The condition contrasts (inference minus control) are defined identically across systems and carry the recruitment claim.

Within a system, condition contrasts (inference minus control; post-commitment minus pre-commitment) were computed per session and tested over sessions with the Wilcoxon signed-rank test. Per-animal sign tests (on each rat's median session value unless a rat mean is named) are reported alongside, to guard against pseudo-replication across sessions. For the commitment analyses, two rat-level checks were added (Supplementary Note S4): an animal-nested bootstrap that resamples rats and then sessions within rats (10,000 draws), and a random-intercept mixed model over rats (exact restricted maximum likelihood, t on rats - 1 degrees of freedom). Directional hypotheses (that the autonomous component rises under inference, and becomes more persistent after commitment) were pre-specified; two-sided p -values are reported throughout unless a one-sided test is named explicitly. The commitment analysis and its windows were not pre-registered; the window grid is reported so that the choice can be judged (Results; Supplementary Note S4). Confidence intervals on model read-outs are item-bootstrap 95% percentile intervals (200 draws for the layer-wise recruitment contrasts; 1,000 for the generation-mode contrasts, whose p -value is twice the smaller bootstrap tail), pointwise per probe depth and uncorrected across depths. 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). For the computation-matched control the unit of analysis is the trained network (18 recurrent versus 18 feedforward per run, one-sided Mann-Whitney U in the pre-registered

direction; the control's Holm-corrected secondary read-outs are in the deposited result files). Because seeds of one architecture are not independent architectures, the result is also aggregated at the architecture level (median over the seeds of each architecture, three recurrent versus three feedforward), where the smallest attainable exact one-sided p is 0.05 and the outcome is reported as an ordering rather than as a significance test. Negative-control tests were pre-specified with their falsification conditions before analysis. No family-wise correction was applied across the small number of pre-specified primary contrasts; exact n , sign counts and p -values are reported for every test.

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, drafting and language-editing aid is declared in full in the Declaration of AI use below.

Results

The Instrument Recovers a Known Autonomous Gain Independent of Input Strength When the Input Is Exogenous

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 within 0.002 and held the estimate fixed as input strength was swept from 0 to 2.0, whereas the naive estimator drifted upward by up to 12% of the true value. Under state-correlated input the eigenvalue read-out lost reliability, while the variance-partition read-out (auto_unique) remained interpretable at moderate coupling. The demix was then checked against the mean-field variance theory of driven tanh echo-state networks (Schubert and Gros, 2021; Methods; Supplementary Note S1 and Supplementary Figure S2), with the membrane potential as state and the external drive as input, on a sweep of six gains by six input scales under independent and under identical input. Its input-driven share tracked the realized input share of the simulated networks within 0.032 in all 72 cells, as it must here, because the input enters the next state additively and is itself the regressor; the comparison with the theory below therefore checks the theory's variance against the simulation, and the demix's own content is tested by the recurrent share and the eigenvalue (Supplementary Note S1). Its dominant autonomous eigenvalue recovered the spectral radius of the network's mean linear-response operator within 0.021 in the pre-registered near-linear cells and within 0.035 in all 30 cells of that arm (five input scales; scale 0.35 was not run), including above the critical gain at which the network's spectral radius crosses 1. Against the theory's own prediction the input share agreed within 0.056 (35 of 36 cells inside the pre-registered 0.05). The residual comes partly from cross-validation shrinkage of the demix's estimate and partly from the theory's departure from the simulated activity variance near the critical gain and, under identical binary input to all units, from its treatment of that input as Gaussian; the demix, which regresses on the simulated trajectory, makes neither approximation. The linear fit under-reported the recurrent share by up to 0.045 where the network was most nonlinear, the shortfall rising with the activity variance (Spearman $\rho = 0.87$). Each pre-registered prediction is scored, and one declared design deviation stated, in Supplementary Note S1. Applied to eight principal components of the network rather than to the full population, the eigenvalue read-out became unreliable, most of all under a shared low-dimensional input, and the variance partition was biased in level (input share lower than on the full state by a median of 0.23 under independent input) but gave the same regime label as the full state in 70 of 72 cells (Supplementary Note S1, exploratory arm). These results set the analysis strategy. The variance partition carries the balance comparison between the two kinds of system. The eigenvalue read-out is trusted only where the input is exogenous (sensory evidence, generated outside the recorded population). In the brain data, an eight-component reduction under a shared low-dimensional input, it is read only for within-session

changes and their direction, not its level (Supplementary Note S1); in the architecture control it is compared between networks analyzed identically.

On Analogous Cognitive Work, Cortex Is Autonomous-Dominated and the Language Models Tested Are Input-Driven

The first test asked, on the same footing in both kinds of system, how much of a population’s moment-to-moment dynamics is driven by external input versus generated by its own recurrence (Figure 2). The two kinds of system fell on opposite sides of the input-driven/autonomous divide. In rat OFC during value-based 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 unique share 0.24. Every neural dataset was autonomous-dominated, the perceptual-decision recordings included (input-driven share below 0.08 against an autonomous unique share near 0.27 throughout the trial): the explained next-state variance was mostly autonomous, and external input accounted for about a tenth or less of the total.

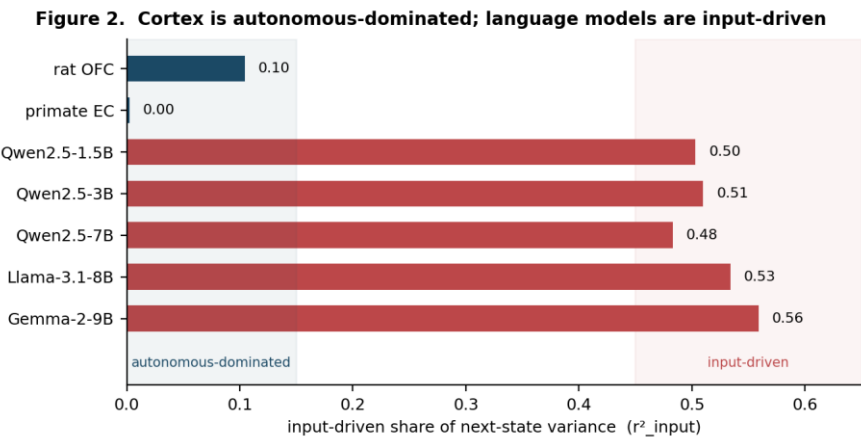


Figure 2. Cortex is autonomous-dominated; language models are input-driven. Input-driven share of next-state variance for the orbitofrontal and entorhinal datasets and each model. Rat orbitofrontal cortex 0.10 and primate entorhinal cortex 0.002. The five transformer language models 0.48–0.56. Brain datasets in one color, models in another; the shaded bands are a visual guide, and the regime label compares r^2_{input} with $auto_unique$ (Methods). The brain analyses use a task-derived input of two dimensions and the model analyses the top eight principal components of the input embedding; the input-dimension sweep, in which the models’ label flips at two input dimensions but holds at a matched dimension of eight, is in Supplementary Note S2. Alt text: horizontal bar chart. Brain input share near 0.0–0.1. 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. As analyzed, this is a difference of dynamical regime: the brain data are autonomous-dominated and the feedforward models input-dominated. Because the regime label can depend on the dimension of the input regressor (Methods, Statistics), it was re-examined at matched input dimension (next paragraph). The pre-specified condition that would have undercut the claim across systems (that a model carries the same autonomous-versus-input balance as the brain) was not met by any model at the input dimensions of the deposited analysis.

Robustness of the regime classification to input dimensionality. The brain analyses use a task-derived input of two dimensions and the model analyses the top eight principal components of the layer-0 embedding, so the two were not like-for-like as first analyzed. In a post-hoc analysis added in this version (not pre-registered; Supplementary Note S2), the input dimension was swept on both

sides, on the deposited model arrays and on the orbitofrontal trajectories produced by the deposited extraction script. For the two models whose deposited across-token arrays agree with the Figure 2 analysis (Qwen2.5-1.5B and Qwen2.5-3B), the cross-validated input-driven share fell monotonically as the input was truncated to fewer principal components, from 0.56 and 0.58 at eight components to 0.19 and 0.20 at two. The regime label flipped: both models read input-dominated at eight input dimensions and autonomous-dominated at two. For rat OFC, expanding the task-derived input from two to eight dimensions (time and offered-value basis functions) raised the cross-validated input-driven share only from 0.086 to 0.138, the autonomous unique share stayed above 0.66 throughout, and the label never changed. At a matched input dimension of eight, the two models remained input-dominated and OFC remained autonomous-dominated, so the direction of the difference holds; but the classification is a property of the input basis as well as of the system, and it is not described as categorical. The model-side sweep runs on the across-token arrays, which agree with the Figure 2 analysis for the two models used and are degenerate for Qwen2.5-7B, which is excluded. Matched dimension is not matched information: the brain’s eight dimensions are basis expansions of one trial-varying task variable, whereas the models’ are eight principal components of the literal input. And the sweep can only reduce the models’ input below the deposited eight components, because raw hidden states were not retained after extraction.

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 $p = 1.3 \times 10^{-6}$, 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 eigenvalue read-out moved the same way (+0.0041, two-sided $p = 0.006$; positive in 19 of 31 rats, sign test $p = 0.28$); this is the read-out that was input-invariant and passed its negative control in the synthetic validation (Methods). The population also became *less* input-driven under inference (Δ input share -0.0023 , $p = 0.01$). Value-matching roughly halved a larger raw effect (a 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 eigenvalue read-out is null here: median -0.0005 , positive in 7 of 14 sessions, two-sided $p = 0.58$), and that read-out is sensitive to differences in input strength (Methods), it is weighted lightly, although the input share itself did not differ between the conditions (median -0.0002 , $p = 0.71$).

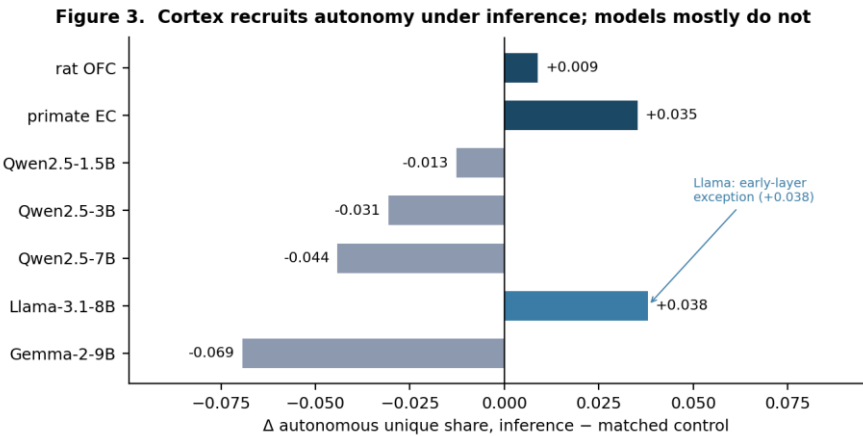


Figure 3. Cortex recruits autonomous dynamics under inference; models mostly do not. Change in autonomous unique share, inference minus matched control, per dataset and model. Rat OFC rises (+0.0089, two-sided $p = 1.3 \times 10^{-6}$ over 164 sessions; 24 of 31 rats). Primate EC rises weakly (+0.035, 10/14 sessions; one-sided $p = 0.045$, two-sided $p = 0.091$). 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) had changes in autonomous share of -0.07 to -0.10 at the deepest probe depth, whose bootstrap intervals excluded zero. The exception 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 depths 0.4 to 0.6). It does not break the balance dissociation: Llama remained input-dominated (input share 0.53), and the rise decayed to near zero (+0.002) at the deepest probe. Because the four other models' input share rose under inference at every probe depth (by 0.03 to 0.19), their negative changes cannot be separated from the same sensitivity of the variance-partition read-out to input strength (Methods).

Cortical Autonomous Dynamics Become More Persistent After Commitment

The balance and recruitment results show that cortex is autonomous-dominated and engages its autonomous component under demand, but they do not tie the autonomy to the act of committing to an answer. The third test asks whether it is, using the perceptual-decision dataset for which a per-trial estimate of the commitment moment (nTc ; Methods) exists (Figure 4). The autonomous-dominance read-out was computed in sliding windows around two alignment points: the inferred commitment time and the stimulus onset.

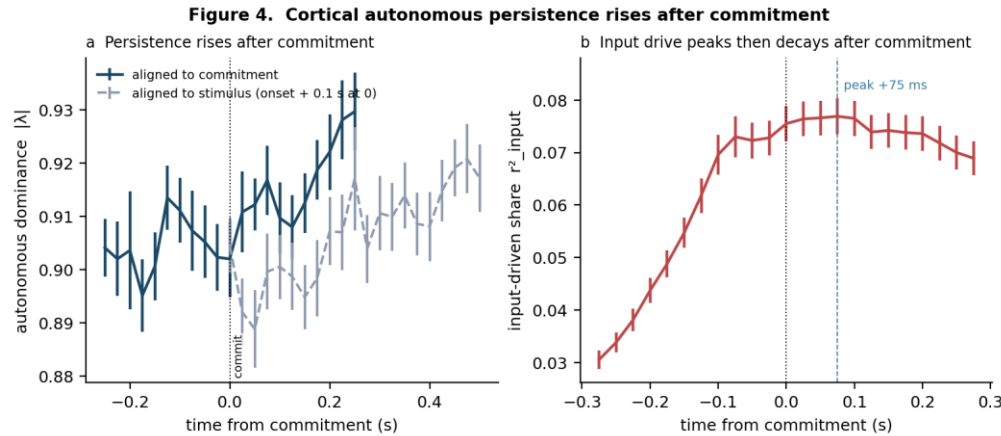


Figure 4. Cortical autonomous persistence rises after commitment. (a) Autonomous-dominance read-out time-resolved around the inferred commitment time, nTc (rises from ~ 0.90 to ~ 0.93 after commitment) versus around stimulus onset (a smaller, more gradual rise; plotted with its zero at 100 ms after onset); error bars are the standard error of the mean (SEM) across 115 sessions. Per session, the post- minus pre-commitment rise is +0.016 (75/115 sessions, two-sided $p = 2.2 \times 10^{-5}$; animal-nested bootstrap interval above zero), and in the primary windows it is sharper aligned to commitment than to stimulus (paired $p = 0.027$; window-dependent, see text). (b) Variance-partition view (150-ms windows): the input-driven share peaks ~ 75 ms after commitment then declines; error bars as in (a). Alt text: left, autonomous-eigenvalue curves rising after commitment, more when aligned to commitment than to stimulus; right, the input-driven share peaking about 75 ms after commitment then declining.

Aligned to commitment, the frontal–striatal autonomous dynamics became more persistent. 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 (one fit over a 200-ms window centered at +200 ms) exceeded the pre-commitment value (centered at –150 ms) by a median of +0.016 (Wilcoxon signed-rank $p = 2.2 \times 10^{-5}$, two-sided; the rise was present in 75 of 115 sessions, per-rat median rise positive in 9 of 12 rats, exact sign test $p = 0.15$, six rats contributing three or fewer sessions). An animal-nested bootstrap gives a 95% interval for the mean rise of +0.005 to +0.044, and a random-intercept mixed-effects fit over rats gives +0.031 (95% CI –0.007 to +0.070, $p = 0.098$; the rat-mean rise is positive in 10 of 12 rats; Supplementary Note S4). The rise was sharper aligned to commitment than to the stimulus, but that difference is less secure than the rise itself. Aligned to stimulus onset (windows +300 and +550 ms after onset), the same per-session rise was +0.002 (two-sided $p = 0.095$; positive in 10 of 12 rats, sign test $p = 0.04$). The commitment-aligned rise exceeded it (paired Wilcoxon $p = 0.027$; 68 of 115 sessions; 8 of 12 rats by per-rat median, sign test $p = 0.39$; the animal-nested interval for the difference spans zero). The comparison also depends on the window pair (stimulus windows 250 ms apart, commitment windows 350 ms; matched-gap comparisons in Supplementary Note S4). A complementary variance-partition view was consistent with a transition near commitment, though not with a fall in input drive across it: the input-driven share rose during evidence accumulation, peaked about 75 ms after commitment, and then declined (Figure 4b), while staying above its pre-commitment level in the post-commitment window (0.074 against 0.055). The decline is consistent with the reduced influence of sensory evidence after commitment reported for these recordings (Luo et al., 2025a) and for separate recordings in the same task (Bondy et al., 2026). The effect is a ramp, not an all-or-none switch: the rise is small (about 0.016, roughly 2%), but it follows the inferred commitment time and is present in 75 of 115 sessions and in 10 of 12 rats by rat mean; its specificity to commitment over the stimulus is supported, not established. A dominant eigenvalue that moves from 0.90 toward 0.93 while its median stays below 1 (9 of 115 sessions reach 1 or more after commitment) means the autonomous state contracts more slowly after commitment. Whether the population also settles onto a fixed point is not established by this measure, which is why the effect is described as increased persistence rather than as a strengthening attractor.

The direction of the effect holds across pre-commitment centers from –250 to –150 ms and post-commitment centers from +150 to +250 ms (nine windows, all positive), and its magnitude ranges over them from +0.005 ($p = 0.03$) to +0.021. Windows with a post-commitment center at or before +100 ms, or a pre-commitment center at –100 ms or later paired with a post-commitment center of +150 ms, give smaller and less consistent rises, so the windowed magnitude is an estimate rather than a precise quantity. A naive dynamic mode decomposition applied to the same commitment-aligned data did not show the rise: its read-out was flat and noisy. The rise emerges only with the input-aware estimator, consistent with the naive estimator's sensitivity to input in the validation (Figure 1a) and with the same contrast reported on these data by the method's authors, who attribute it to partial observation (Huang et al., 2026).

A commitment-locked transition from input-driven to intrinsically generated dynamics is the result reported for these recordings by their original analysts (Luo et al., 2025a); the subspace demix used here recovers its persistence side, and its authors applied it to the same data and found slower decay of the intrinsic dynamics after commitment (Huang et al., 2026). A circuit model of recordings from monkey parietal cortex reads the commitment stage the same way, as a transition from more sensory-driven to more circuit-driven dynamics (Monsalve-Mercado 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 the comparison across systems (the balance and recruitment results above and the model-side test that follows), which places feedforward language models on the same demixed axis.

No Commitment-Locked Rise in Autonomous Persistence Was Found in the Feedforward Language Models

The next test asked whether the language models show anything like the brain’s commitment-locked rise in persistence (Figure 5). Within the limits of the analog stated below, no such rise was detected. 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 examined across layers (a cruder analog of the brain analysis; Methods). First, on the 20 true statements the answer became top-ranked under the logit lens in 7 for Qwen2.5-7B (first at layers 24–27 of 28; median 25) and in 16 for Llama-3.1-8B (layers 20–28 of 32; median 25), in line with the gradual sharpening of predictions over later layers reported for such models (Lad et al., 2024). The pass therefore runs on for a few layers after this read-out’s commitment point, but those layers are feedforward stages, not a period in which a recurrent state could settle (Terminology). This is a statement about the read-out used, not a claim that the model has decided in the sense in which a biological accumulator settles on a choice. Second, the across-layer autonomous-dominance read-out hovers near or above 1 (expansive or marginal) and shows no clean rise in persistence of the kind seen in cortex; its largest value is a single isolated spike (Qwen2.5-7B, at about 0.7 of the depth). This read-out is a noisy analog (Terminology; next paragraph), and it is shown by relative depth because it was not re-estimated with the corrected L^* (Supplementary Note S5).

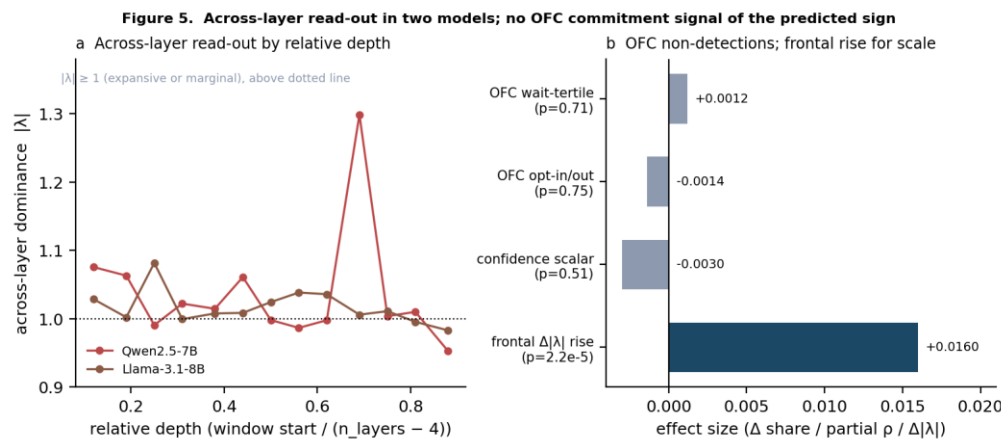


Figure 5. The across-layer read-out of two language models, and no commitment signal of the predicted sign in trial-averaged or scalar read-outs of orbitofrontal cortex. (a) Across-layer autonomous-dominance read-out as a function of relative network depth for two models, not aligned to the commit layer (Supplementary Note S5): values hover near or above 1, mostly outside the stable range in which a rise reads as persistence, with one isolated spike; it is a noisy analog on which no claim rests (Results). (b) Pre-specified 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 near or above 1 with one isolated spike; three orbitofrontal effect sizes near zero and the larger frontal–striatal rise.

The across-layer eigenvalue read-out is a noisy analog: a transformer’s across-layer trajectory is not a temporal dynamical system but a feedforward chain of stages (Terminology), on which the eigenvalue estimate is unreliable (Supplementary Note S3). The within-pass claim therefore rests on the structural fact that a feedforward stack has no recurrence across its depth (it can still converge toward a fixed point of its layer-to-layer map, Domany et al., 1989, which the demix cannot assess there), not on that noisy number, and a commitment-locked change in persistence is sought across generated tokens, taken up next.

Recurrence across tokens. Within a forward pass a transformer is feedforward, but across token positions an autoregressive model feeds each generated token back as input to the next pass, and a chain-of-thought answer is a sequence of such steps; in that sense generation is recurrent at the token

level, and a commitment-locked change in autonomous persistence could in principle appear there. A real-model test was run to measure exactly this (pre-registered before any generation-mode result, and deposited). Each of the five language models was given a construction item without its final word (true and false statements alike; Supplementary Note S5) and made to generate a continuation, so that the demix eigenvalue could be measured across the generated tokens. The first pass centered its window on the first generated token of every item, which matched the first token of the item's final word in only 5 to 13 of 40 items. A pre-registered higher-powered follow-up (longer generations, wider windows) centered the window on the first token of the final word wherever it appeared among the first eight generated tokens (14 to 16 usable items in four of the five models). Neither 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 point estimates were mostly positive but non-significant, with wide intervals. One model (Gemma-2-9b) produced the answer among its first eight generated tokens too rarely (six usable items) to be tested under the stricter criterion. The eigenvalue estimator is noisy on short generated sequences, and across generated tokens its input is the model's own output, fed back from its state, the regime in which the synthetic validation found this read-out unreliable (Methods). This is therefore a non-detection of a brain-like commitment-locked rise; the intervals (widths 0.5 to 2.9) could neither detect nor exclude a rise of the cortical size. Generating tokens after the answer is not functionally equivalent to holding a decision for a later action, so the test bounds what token-level recurrence does in these models on this task; it does not make the cortex-model comparison a matched one. And the items used elicit short answers, so a model reasoning at length in chain-of-thought mode was not tested (Discussion). The architectural conclusion therefore rests on the control that follows.

A Computation-Matched Architecture Control

The cortex-model 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 unfolding argument specifies the controlled test: the same input-output function realized in a recurrent and in a feedforward system. Its closest practical form, the same task learned to matched accuracy, was pre-registered in full and run, with the task, the architectures, 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 were compared with three feedforward ones. The recurrent architectures were a vanilla recurrent neural network (RNN), a gated recurrent unit (GRU) and a long short-term memory network (LSTM); the feedforward architectures were a windowed multilayer perceptron (MLP), a dilated causal temporal convolutional network (TCN) and a causal transformer encoder. Six seeds of each, in each of two independent seed sets, were trained on an identical Poisson-clicks evidence-accumulation decision in two tasks: an always-on task, with evidence until the end of the trial, and a delayed-response task, with a zero-input delay after the evidence. Trajectories were reduced to the same eight dimensions, and 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 in the first seed set, each within ± 0.007 of its task mean; Figure 6a shows the always-on task), so the task and the accuracy are held fixed and only the architecture varies. On each trial, a network's commitment time was the first time bin at which its choice probability passed 0.9 (or 0.1) and stayed on that side of 0.5 to the end of the trial, and the eigenvalue read-out was fitted on the eight bins before that bin and on the eight bins from it.

With task and accuracy held fixed, the autonomous-eigenvalue change at commitment (the eigenvalue read-out of Figure 4) was larger in the recurrent family than in the feedforward family (one-sided Mann-Whitney U over 18 versus 18 networks, $p = 0.013$; Figure 6b). The separation was reproduced, more sharply, on a second task with a zero-input delay after the evidence, in which no new input arrives, although the feedforward networks still see the evidence within their input window ($p = 1.3 \times 10^{-4}$; Figure 6c). It held in all four runs (two tasks $\times$ two independent seed sets, $p \leq 0.013$) and survived the pre-registered guard against networks that freeze after commitment (none

was flagged). With the pre-registered robustness alignment, to the commitment time of an ideal observer of the same clicks, identical for every network, the separation held on the delayed task ($p = 0.0026$; second seed set 6.1×10^{-5}) but not on the always-on task ($p = 0.21$; 0.13). The pre-registration named a mixed model with architecture as a random effect as a corroborative test; it was not run, and the architecture-level medians that follow stand in for it, a declared deviation. Because seeds of one architecture are not independent architectures, the result was also aggregated at the architecture level, taking the median over the twelve seeds of each architecture (Supplementary Figure S1). On the delayed-response task all three recurrent architectures lay above all three feedforward architectures (medians $+0.008$ to $+0.116$ against -0.017 to $+0.003$), the most extreme of the 20 possible orderings at three versus three, for which the smallest attainable exact one-sided p is 0.05 . On the always-on task only one of the three recurrent architectures (the LSTM) lay above the best feedforward architecture (the TCN), so at the architecture level that task is directional but not separated. Three qualifications apply: on the always-on task the effect is a family-level shift, carried by the LSTM and by the transformer at the two extremes, and the convolutional network's own rise ($+0.044$) exceeds that of two of the three recurrent architectures; the variance-partition read-out, which the convolutional network can inflate, was not diagnostic, so the result rests on the eigenvalue read-out; and the clean architecture-level separation is on the delayed-response task, where the requirement to hold a choice after the input has ended is explicit.

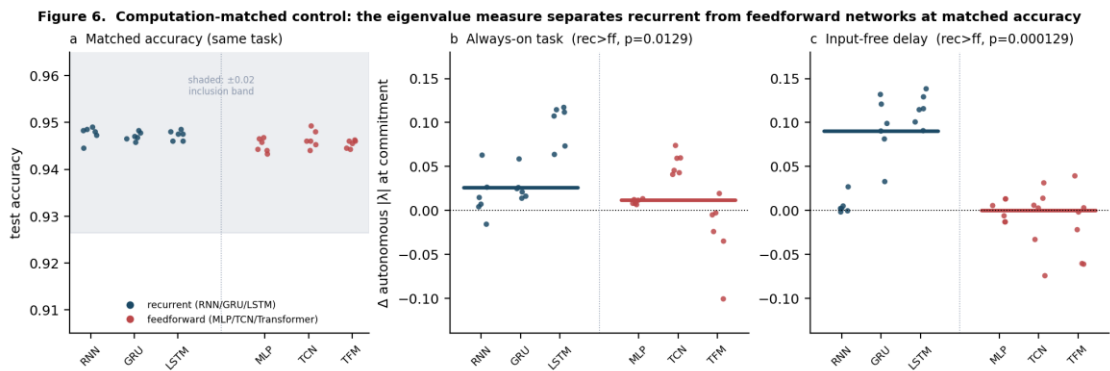


Figure 6. A computation-matched control: the eigenvalue measure separates recurrent from feedforward networks on the same task at matched accuracy. Recurrent (RNN, GRU, LSTM) and feedforward (windowed MLP, dilated causal TCN, causal transformer) networks, six seeds each in the first seed set, trained 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 on the always-on task; 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 ($p = 1.3 \times 10^{-4}$). The unit in (b) and (c) is the trained network; the aggregation at the architecture level (median over the twelve seeds of both seed sets per architecture, three recurrent versus three feedforward) is in Supplementary Figure S1. 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 delayed-response task.

Negative Results: No Commitment Signal of the Predicted Sign in Trial-Averaged or Scalar Read-Outs of Orbitofrontal Cortex

A central part of the evidence is a set of pre-specified tests that returned nulls, each reported with its prediction (Figure 5). They bound what the positive results mean.

First, a test asked whether the autonomous component in rat OFC tracks commitment when read out trial-averaged rather than time-locked. No such tracking was detected. Contrasting long-wait

against short-wait opt-out trials (wait-time tertiles) returned a null (median Δ autonomous share +0.0012, Wilcoxon $p = 0.71$ over 113 sessions; the per-animal sign test was likewise null: 16 of 28 rats positive, $p = 0.57$). So did an opt-in versus opt-out contrast, between trials on which the rat committed to waiting for the offered reward and trials on which it abandoned the wait (-0.0014 , $p = 0.75$ over 170 sessions; 14 of 32 rats positive, $p = 0.60$). Session counts differ between tests because each requires a minimum number of trials in each of its conditions. The eigenvalue read-out was also null in both contrasts (session $p = 0.13$, 0.34); per rat the tertile contrast was negative in 21 of 28 (sign $p = 0.013$), opposite to the predicted direction. Second, a test asked whether a scalar geometric read-out of the population, the fraction of each trial's activity energy on the leading population mode (a compression index, not the demixed autonomous component), predicts a behavioral 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 null (median partial Spearman -0.003 , Wilcoxon $p = 0.51$), with a value-separability negative control also null. One deviation is declared: the population basis and the control's value centroids were estimated on all non-violation trials, the scored ones included, rather than only on the others as specified; wait time plays no part in either. These are non-detections that do not isolate their cause: OFC is a value/state area rather than the frontal–striatal population in which the commitment transition was found, and the read-outs were trial-averaged or scalar rather than time-locked, so either the area or the read-out could account for the null, and no trial-averaged read-out was applied to the frontal–striatal data.

Discussion

What the Results Show

Cortical populations and feedforward language models were given analogous 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 fall in different regimes, and for the two models and the cortical dataset re-examined, the direction of the difference survived matching the input dimension of the two analyses at eight. 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 inferred moment of committing to a perceptual decision, the cortical autonomous dynamics become more persistent, more sharply when aligned to commitment than to the stimulus (supported, not established), a transition already reported for these data (Luo et al., 2025a) and recovered here with the subspace demix of Huang et al. (2026). In the feedforward models no such rise was found across generated tokens, where recurrence exists at the token level. Within a forward pass the demix offers no trusted measurement: the stack has no recurrence across its depth; in the two models probed, the answer, when detected, first becomes top-ranked a few layers before the output; and the across-layer read-out over those layers is a feedforward-chain analog on which no claim rests (Terminology). The architectural reading of that difference rests on the matched control, in which recurrent and feedforward networks trained to the same task separated on the same measure.

The instrument was validated on synthetic systems before any recorded data were touched and checked against the analytic mean-field variance theory of driven echo-state networks (Supplementary Note S1). The comparison across systems is restricted to dynamical regime and sign; no claim rests on magnitude. And the positive commitment result is bracketed by pre-specified nulls: in orbitofrontal cortex, the trial-averaged and scalar read-outs one would otherwise reach for first showed no commitment signal of the predicted sign.

Bearing on the Unfolding Argument

Whether the dynamics of reaching a decision differ between recurrent and feedforward systems that behave alike is the question this study was built to inform, and the unfolding argument (Doerig et al., 2019; Introduction) is why it looks at internal dynamics at all. The cortex–model comparison did not build the behaviorally equivalent pair the argument specifies, a feedforward and a recurrent system computing the same input–output function. It compares two systems doing analogous answer-construction tasks, and it spans different tasks, different functional demands and different architectures at once, so it cannot by itself attribute the difference to architecture. What it shows is descriptive: a recurrent cortex and five feedforward transformers differ in the dynamics of reaching the answer in a specific, measurable way. A closer test was run separately, as a pre-registered control reported above (Figure 6): recurrent and feedforward networks trained on the *same* task to matched accuracy, with the demix applied unchanged. With the task and accuracy held fixed and only the architecture varied, the eigenvalue measure separated the families in every run, most sharply when no input followed the evidence. At the architecture level the separation was complete on the delayed task and directional on the always-on task. The measure therefore tracks recurrent-versus-feedforward architecture at matched task and accuracy for the six architectures tested, a difference that, by the unfolding argument, behavior alone cannot show; the cortex–model difference itself remains descriptive. To the extent that theories of consciousness tie access or experience 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), the share of autonomously generated variance is a property on which the cortex and the feedforward models tested differ, and the difference held for the two models re-examined at a matched input dimension of eight. Whether that difference matters for access is for those theories to say.

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 consciousness treat as load-bearing, is present in cortex and was not found in the feedforward models tested on an analogous task, by a generation-mode test that could not have detected a rise of the cortical size. If a theory holds that this signature is necessary for access, the result is that a condition the theory calls necessary was not found in these models, on a task that did not require them to hold the decision for later use, though the antecedent is the theory's to defend and is not adjudicated here. Whatever the theory, the contribution is the same: a decidable measurement, made the same way in a brain and in an artificial network, on which debates about the role of recurrent dynamics can draw.

Scope and Limitations

The first limitation is one of scope. The results are reported as a claim about decision commitment, and reading them as bearing on consciousness invites an immediate objection: decision commitment is not conscious access. The commitment event in the perceptual-decision data is the moment a rat's accumulator locks onto a choice; it is inferred from the recorded population jointly with the animal's choices (Methods), 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 behavior (Mashour et al., 2020), which the present measure does not touch. So the rise in autonomous persistence 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: 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 these bridges; the bearing on consciousness is offered as an argument for the reader to weigh.

A second limitation is the model side of the commitment comparison. The strongest model-side claim is structural (within a forward pass a feedforward stack has no recurrent loop; any persistence it carries across depth is feedforward memory, which the demix cannot separate from recurrence; Terminology, Supplementary Note S3), and it does not need the noisy across-layer eigenvalue measure. A cleaner test, an across-token measure of persistence in generation mode, was therefore run on the real models (Results), in a first pass and a pre-registered higher-powered follow-up; neither found a commitment-locked rise 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 produced the answer too rarely to test, the test could neither detect nor exclude a rise of the cortical size; the model leg remains 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 rise in autonomous persistence at commitment, though it follows the inferred commitment, is small in absolute magnitude and is a ramp rather than a discrete switch. The echo-state check showed the eigenvalue read-out to be unreliable on an eight-component reduction of a population driven by a shared low-dimensional input (Supplementary Note S1). The cortical eigenvalue results are therefore read for within-session changes and their direction, consistent with the original report of the transition (Luo et al., 2025a), and not for their absolute level. The commitment time is itself inferred from the same recordings jointly with behavior (Methods), so the cortical result is a recovery of the original authors' transition by a different estimator, not an independent detection.

A fifth limitation is functional, and it bears on how the cortex–model comparison should be read. In the animal the decision is not the end of the computation: after the accumulator commits, the choice has to be held long enough to select and execute the action and, possibly, to monitor or correct it, so the circuit has a reason to keep the committed state available. In the model task the answer token is effectively the end of the computation and, in the two models probed layer by layer, the answer, when detected, first becomes top-ranked a few layers before the output; the model has no comparable requirement to preserve the decision for a later stage. A post-commitment rise in persistence in cortex and its non-detection in the models therefore cannot be attributed to the recurrent-versus-feedforward difference alone; part of it may reflect what each system is asked to do after committing. The generation-mode analysis does not remove this asymmetry, because generating tokens after the answer is not functionally equivalent to maintaining a decision for the control of an action. This is why the architectural conclusion is built on the computation-matched control, in which the delayed-response task requires the choice to be reported after the evidence has ended, from the network's own state in the recurrent networks and from an input window that still contains the evidence in the feedforward ones (Results), and why the direct cortex–model comparison is weighted as descriptive.

A sixth point concerns chain-of-thought and recurrence across tokens. The feedforward characterization of a transformer holds within a single forward pass. Across token positions an autoregressive model is recurrent: each emitted token is appended to the input of the next pass, so the model's own output feeds its next state. Chain-of-thought prompting (Wei et al., 2022) exploits exactly this loop to perform multi-step computation that a single pass cannot. Formally, a transformer allowed to generate intermediate tokens gains expressive power over one that is not (Merrill and Sabharwal, 2023). Token-level recurrence is a dominant feature of how these models are used, and it is addressed here in two ways. Empirically, the generation-mode analysis (Results) measures autonomous persistence across generated tokens, that is, across the recurrent loop; it found no commitment-locked rise in any model on this task (four models under the stricter follow-up criterion; Gemma-2-9b produced the answer too rarely to test). Cox et al. (2026), whose instruction-tuned models include four of the five tested here, report that the final answer can often be decoded from the residual stream before any chain-of-thought is generated. Conceptually, token-level recurrence differs from the recurrence of a cortical population in what carries new information forward: a discrete token drawn from the output distribution, not the continuous hidden state. Attention reads the continuous states of earlier positions, but these are themselves functions of the text, so the only

state that persists across steps is the text. Whether a model made to reason at length in chain-of-thought mode develops a commitment-locked change in the dynamics of its generated sequence is an open question that the present item set, with its short answers, does not test, and it is the natural next experiment.

Relation to Prior Work

The instrument used here is built on input-aware dynamical similarity analysis (Huang et al., 2026; Ostrow et al., 2023). The method paper applies the demix within brains (including to the very perceptual-decision recordings reanalyzed 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 difference between them. The contribution here is therefore not the demix, nor the brain transition (which is theirs and Luo et al.'s), but the signed comparison across the two kinds of system: a brain and a language model placed on one autonomous-versus-input axis on analogous cognitive work, with the commitment-locked rise in persistence present in one and not found in 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 systems, not the dissociation within one. The identifiability problem navigated here, that recurrent and input-driven linear models of cortex are degenerate without access to the areas supplying the input, is the one documented by Soldado-Magraner et al. (2024), and the subspace estimator is the constraint added here to reduce it.

For driven nonlinear recurrent networks an analytic theory of the variance of neural activity exists: the self-consistent mean-field treatment of tanh echo-state layers (Schubert and Gros, 2021; Massar and Massar, 2013) splits the membrane-potential variance into a recurrent share and an external share under an assumption of statistical independence across the population. The linear demix was checked against that theory on networks built to its specification (Supplementary Note S1). The demix's input share matched that split to within 0.056 where the assumption holds (about half from the theory's own departure near the critical gain and half from cross-validation shrinkage; with white additive input, this comparison checks the theory against the simulation), followed the simulated network's input share where identical binary input to all units departs from the theory, and recovered the spectral radius of the nonlinear network's mean linear-response operator. Its measured limits, under-reporting where the nonlinearity is strongest, an unreliable eigenvalue read-out on a low-dimensional reduction under shared input, and a level bias of the variance partition on that reduction that left its regime label intact in 70 of 72 cells, are stated in Methods and Results. Two neighboring comparisons of neural and language-model dynamics 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 per-trial commitment event.

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 on both read-outs (autonomous share $p = 1.3 \times 10^{-6}$, eigenvalue $p = 0.006$, across 164 sessions). The commitment claim would fall if the rise in autonomous persistence were not locked to commitment; in the primary windows it was sharper aligned to commitment than to stimulus ($p = 0.027$), a difference that is window-dependent and unresolved at the animal level (Results; Supplementary Note S4), so the locking is supported rather than established. The balance observation would fall if a model, measured the same way, carried the same autonomous-versus-input balance as the brain; none did at the input dimension of the deposited analysis, and neither of the two models

re-examined did at a matched input dimension of eight, although both did at two (Supplementary Note S2). The cleanest falsification targets the model leg: finding a commitment-locked rise in autonomous persistence in a feedforward model, by an across-token analysis, would refute the model leg of the dissociation. That test was run in generation mode across the models, in a first pass and a higher-powered follow-up, and found no significant rise (minimum $p = 0.07$), with point estimates mostly positive and intervals too wide to bear on the model leg. A definitive higher-powered version, across more items and models and including long chain-of-thought generations, is the obligation the present model leg leaves open.

Summary and Conclusions

On analogous cognitive work, cortical populations and the feedforward transformer language models tested occupy different dynamical regimes. Cortex is autonomous-dominated and recruits autonomy under inference, and after the inferred moment of committing to a perceptual decision its autonomous dynamics become more persistent, within the transition Luo et al. (2025a) first reported, as Huang et al. (2026) found with the same demix. The models are input-driven and mostly do not recruit autonomy. In the two models probed layer by layer, the answer, when it becomes the top-ranked prediction at all, does so only in later layers. Across generated tokens, the analog in which token-level recurrence exists, no commitment-locked rise in persistence was found, but the estimator is noisy on short sequences and the test could neither detect nor exclude a rise of the cortical size. The instrument was validated on synthetic ground truth and checked against an analytic theory of driven recurrent networks. The dissociation is signed, restricted to dynamical regime rather than magnitude, unchanged in direction for the two models and the cortical dataset re-examined at a matched input dimension of eight, and bracketed by pre-specified nulls in which trial-averaged and scalar read-outs returned no commitment signal of the predicted sign. For the unfolding argument: a dynamical property that accompanies commitment is present in cortex and was not found in the five feedforward models tested (four in the follow-up generation test), and a pre-registered control shows the same eigenvalue measure tracks recurrent-versus-feedforward architecture when the task is held fixed, cleanly so on the delayed-response task. No claim is made to have measured conscious access or phenomenal experience, and the heaviest limitation is that the cortex-model contrast varies task, functional demand and architecture together, which is why the computation-matched control carries the architectural claim. Conclusions are limited to the systems tested. What the study offers the debate is a decidable measurement.

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

Author Contributions: 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.

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 Institute for Complexity Science and Advanced Computing (ICSAC).

Institutional Review Board Statement: 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.

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. This is version 2 of a preprint first posted on 17 August 2026 (version 1: doi: 10.20944/preprints202608.1095.v1). Version 2 reframes the comparison as recurrent versus feedforward dynamics, replaces the attractor vocabulary with autonomous persistence, limits every claim to the systems tested, and corrects the description of the inferred

commitment time. It adds the analyses reported in Supplementary Notes S1 to S4 and Supplementary Figures S1 and S2, corrects the across-layer commitment analysis of version 1 (Supplementary Note S5), regenerates Figures 4, 5 and 6 with revised titles and labels and corrects Figure 1b. It describes the orbitofrontal null tests as pre-specified rather than pre-registered and states when each pre-registration was fixed. Apart from Figure 1b, that commitment analysis and the window-sensitivity range, now computed on a stated grid of windows, no measurement, figure value or statistic from version 1 has been altered; the deposited result files reproduce them.

Data Availability Statement: All data analyzed here are public. Rat orbitofrontal recordings are from the Constantinople laboratory (Schierack et al., 2026; data: Schierack et al., 2025, Zenodo accession 16997337). Primate entorhinal recordings are from Neupane et al. (2024a; data: Neupane et al., 2024b, DANDI accession 000897). The perceptual-decision Neuropixels recordings are from Luo et al. (2025a; raw data: Luo et al., 2025b, 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-Instruct, Qwen/Qwen2.5-3B-Instruct, Qwen/Qwen2.5-7B-Instruct; meta-llama/Llama-3.1-8B-Instruct; google/gemma-2-9b-it). The hidden-state activations themselves were not retained after extraction; they are regenerable from the deposited extraction and analysis scripts, the listed model identifiers and the deposited item set, and the deposited per-model result files reproduce every model-side figure and statistic without re-extraction. The deposited model-side balance script (analyze_llm_balance.py) is a reference reimplement of the original extraction run, as its header states. The extraction step that added opt-out choices and wait times to the orbitofrontal trajectories for the commitment null tests was not retained; their per-session results are deposited. The analysis code (with that reimplement in place of the original extraction run), the construction-task item set, the result files underlying every figure and statistic, the synthetic-validation results (seeded, although some draws there and in four per-session analyses also depend on Python's string hashing, which varies between runs), the demix library and per-system 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. The supplement to this version adds five analyses that were not in version 1, with their code and result files: the echo-state mean-field check of the demix (Supplementary Note S1 and Supplementary Figure S2, with its pre-registration and lock hash), the input-dimensionality robustness analysis (Supplementary Note S2), the feedforward-chain controls (Supplementary Note S3), the architecture-level aggregation of the computation-matched control (Supplementary Figure S1) and the animal-nested bootstrap and rat-level mixed model of the commitment-locked rise (Supplementary Note S4); it also documents a correction to the across-layer commitment analysis (Supplementary Note S5). Each pre-registration fixed its design, metric and falsification conditions before the result it governs was computed; the locks are SHA-256 hashes with local timestamps, not external registrations. The orbitofrontal null tests were specified in their analysis scripts before they were run (the confidence test with its kill criterion) but were not hash-locked, and they are described as pre-specified rather than pre-registered. The reanalyzed datasets are cited in the reference list. No new human or animal data were collected. Claude Opus 5.5 and Claude Fable 5.1, and for version 1 Claude Opus 5 and Claude Fable 5 (Anthropic PBC), accessed through the Claude Code command-line interface (<https://claude.com/claude-code>), 2026, were used under the author's direction to write and run analysis code and to help draft and edit the manuscript, including this revised version. The research questions, the choice of analyses and the interpretation of the results are the author's. The author takes responsibility for the accuracy and originality of all material in the manuscript, including the references. The models are not authors.

Conflicts of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Bechara, A., Damasio, H., and Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral Cortex* 10, 295–307. doi: 10.1093/cercor/10.3.295

- Belrose, N., Ostrovsky, I., McKinney, L., Furman, Z., Smith, L., Halawi, D., et al. (2023). Eliciting latent predictions from transformers with the tuned lens. *arXiv [Preprint]*. Available at: <https://arxiv.org/abs/2303.08112> (Accessed August 14, 2026).
- Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences* 18, 227–247. doi: 10.1017/S0140525X00038188
- Bondy, A. G., Charlton, J. A., Luo, T. Z., Kopec, C. D., Stagnaro, W. M., Venditto, S. J. C., et al. (2026). Multi-region coordination of evidence accumulation and decision commitment in the brain. *bioRxiv [Preprint]*, version 7. Available at: <https://doi.org/10.1101/2024.08.21.609044> (Accessed October 4, 2026).
- Brunton, B. W., Botvinick, M. M., and Brody, C. D. (2013). Rats and humans can optimally accumulate evidence for decision-making. *Science* 340, 95–98. doi: 10.1126/science.1233912
- Cox, K., Kianersi, D., and Garriga-Alonso, A. (2026). Post-hoc reasoning in chain of thought: decoding and steering pre-committed answers. *arXiv [Preprint]*. Available at: <https://arxiv.org/abs/2603.01437> (Accessed October 4, 2026).
- Doerig, A., Schurger, A., Hess, K., and Herzog, M. H. (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
- Doerig, A., Schurger, A., and Herzog, M. H. (2021). Hard criteria for empirical theories of consciousness. *Cognitive Neuroscience* 12, 41–62. doi: 10.1080/17588928.2020.1772214
- Domany, E., Kinzel, W., and Meir, R. (1989). Layered neural networks. *Journal of Physics A: Mathematical and General* 22, 2081–2102. doi: 10.1088/0305-4470/22/12/013
- Ganguli, S., Huh, D., and Sompolinsky, H. (2008). Memory traces in dynamical systems. *Proceedings of the National Academy of Sciences* 105, 18970–18975. doi: 10.1073/pnas.0804451105
- Gold, J. I., and Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience* 30, 535–574. doi: 10.1146/annurev.neuro.29.051605.113038
- Goldman, M. S. (2009). Memory without feedback in a neural network. *Neuron* 61, 621–634. doi: 10.1016/j.neuron.2008.12.012
- Huang, A., Ostrow, M., Singh, S. H., Kozachkov, L., Fiete, I., and Rajan, K. (2026). “InputDSA: demixing then comparing recurrent and externally driven dynamics,” in *International Conference on Learning Representations (ICLR 2026)*. Available at: <https://arxiv.org/abs/2510.25943> (Accessed October 4, 2026).
- Jaeger, H., and Haas, H. (2004). Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication. *Science* 304, 78–80. doi: 10.1126/science.1091277
- Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., and DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream’s execution of core object recognition behavior. *Nature Neuroscience* 22, 974–983. doi: 10.1038/s41593-019-0392-5
- Khona, M., and Fiete, I. R. (2022). Attractor and integrator networks in the brain. *Nature Reviews Neuroscience* 23, 744–766. doi: 10.1038/s41583-022-00642-0
- Kiani, R., Cueva, C. J., Reppas, J. B., and Newsome, W. T. (2014). Dynamics of neural population responses in prefrontal cortex indicate changes of mind on single trials. *Current Biology* 24, 1542–1547. doi: 10.1016/j.cub.2014.05.049
- Kleiner, J., and Hoel, E. (2021). Falsification and consciousness. *Neuroscience of Consciousness* 2021, niab001. doi: 10.1093/nc/niab001
- Lad, V., Lee, J. H., Gurnee, W., and Tegmark, M. (2024). The remarkable robustness of LLMs: stages of inference? *arXiv [Preprint]*. Available at: <https://arxiv.org/abs/2406.19384> (Accessed October 4, 2026).
- Lamme, V. A. F., and Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences* 23, 571–579. doi: 10.1016/S0166-2236(00)01657-X
- Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences* 10, 494–501. doi: 10.1016/j.tics.2006.09.001
- Lerner, J. S., Li, Y., Valdesolo, P., and Kassam, K. S. (2015). Emotion and decision making. *Annual Review of Psychology* 66, 799–823. doi: 10.1146/annurev-psych-010213-115043

- Luo, T. Z., Kim, T. D., Gupta, D., Bondy, A. G., Kopec, C. D., Elliott, V. A., et al. (2025a). Transitions in dynamical regime and neural mode during perceptual decisions. *Nature* 646, 1156–1166. doi: 10.1038/s41586-025-09528-4
- Luo, T. Z., Kim, T. D., Gupta, D., Bondy, A. G., Kopec, C. D., Elliott, V. A., et al. (2025b). Data from: Transitions in dynamical regime and neural mode underlie perceptual decision-making. Dryad Digital Repository. doi: 10.5061/dryad.sj3tx96dm. Per-trial commitment times available at: https://github.com/Brody-Lab/decision_dynamics_commitment
- Maass, W., Natschläger, T., and Markram, H. (2002). Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Computation* 14, 2531–2560. doi: 10.1162/089976602760407955
- Mante, V., Sussillo, D., Shenoy, K. V., and Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* 503, 78–84. doi: 10.1038/nature12742
- Mashour, G. A., Roelfsema, P., Changeux, J.-P., and Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron* 105, 776–798. doi: 10.1016/j.neuron.2020.01.026
- Massar, M., and Massar, S. (2013). Mean-field theory of echo state networks. *Physical Review E* 87, 042809. doi: 10.1103/PhysRevE.87.042809
- Merrill, W., and Sabharwal, A. (2023). The expressive power of transformers with chain of thought. arXiv [Preprint]. Available at: <https://arxiv.org/abs/2310.07923> (Accessed September 29, 2026).
- Monsalve-Mercado, M. M., Stine, G. M., Shadlen, M. N., and Miller, K. D. (2025). The geometry of the neural state space of decisions. bioRxiv [Preprint]. Available at: <https://doi.org/10.1101/2025.01.24.634806> (Accessed October 4, 2026).
- Neupane, S., Fiete, I., and Jazayeri, M. (2024a). Mental navigation in the primate entorhinal cortex. *Nature* 630, 704–711. doi: 10.1038/s41586-024-07557-z
- Neupane, S., Fiete, I., and Jazayeri, M. (2024b). Data from: Mental navigation in the primate entorhinal cortex. DANDI Archive, dandiset 000897. Available at: <https://dandiarchive.org/dandiset/000897> (Accessed August 14, 2026).
- Nostalgebraist (2020). Interpreting GPT: the logit lens. LessWrong. Available at: <https://www.lesswrong.com/posts/AcKRB8wDpdAN6v6ru/interpreting-gpt-the-logit-lens> (Accessed August 14, 2026).
- Ostrow, M., Eisen, A., Kozachkov, L., and Fiete, I. (2023). “Beyond geometry: comparing the temporal structure of computation in neural circuits with dynamical similarity analysis,” in *Advances in Neural Information Processing Systems* 36. Available at: <https://arxiv.org/abs/2306.10168> (Accessed August 14, 2026).
- Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews Neuroscience* 9, 148–158. doi: 10.1038/nrn2317
- Resulaj, A., Kiani, R., Wolpert, D. M., and Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature* 461, 263–266. doi: 10.1038/nature08275
- Schierbeck, S. S., Pérez-Rivera, D. T., Mah, A., DeMaegd, M. L., Hocker, D., Ward, R. M., et al. (2025). Data from: The orbitofrontal cortex updates beliefs for state inference. Zenodo, accession 16997337. doi: 10.5281/zenodo.16997337
- Schierbeck, S. S., Pérez-Rivera, D. T., Mah, A., DeMaegd, M. L., Hocker, D., Ward, R. M., et al. (2026). The orbitofrontal cortex updates beliefs for state inference. *Neuron* 114, 507–520.e8. doi: 10.1016/j.neuron.2025.10.024
- Schubert, F., and Gros, C. (2021). Local homeostatic regulation of the spectral radius of echo-state networks. *Frontiers in Computational Neuroscience* 15, 587721. doi: 10.3389/fncom.2021.587721
- Shadlen, M. N., and Kiani, R. (2013). Decision making as a window on cognition. *Neuron* 80, 791–806. doi: 10.1016/j.neuron.2013.10.047
- Soldado-Magraner, J., Mante, V., and Sahani, M. (2024). Inferring context-dependent computations through linear approximations of prefrontal cortex dynamics. *Science Advances* 10, ead14743. doi: 10.1126/sciadv.ad14743
- Tononi, G., Boly, M., Massimini, M., and Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience* 17, 450–461. doi: 10.1038/nrn.2016.44

- Vahidi, P., Sani, O. G., and Shanechi, M. M. (2024). Modeling and dissociation of intrinsic and input-driven neural population dynamics underlying behavior. *Proceedings of the National Academy of Sciences* 121, e2212887121. doi: 10.1073/pnas.2212887121
- Vahidi, P., Sani, O. G., and Shanechi, M. M. (2025). BRAID: input-driven nonlinear dynamical modeling of neural-behavioral data. *arXiv [Preprint]*. Available at: <https://arxiv.org/abs/2509.18627> (Accessed August 14, 2026).
- Van Overschee, P., and De Moor, B. (1994). N4SID: subspace algorithms for the identification of combined deterministic–stochastic systems. *Automatica* 30, 75–93. doi: 10.1016/0005-1098(94)90230-5
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). “Attention is all you need,” in *Advances in Neural Information Processing Systems* 30. Available at: <https://arxiv.org/abs/1706.03762> (Accessed October 4, 2026).
- Wang, X.-J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron* 36, 955–968. doi: 10.1016/S0896-6273(02)01092-9
- Watanabe, T., Inoue, K., Kuniyoshi, Y., Nakajima, K., and Aihara, K. (2025). Comparison of large language model with aphasia. *Advanced Science* 12, 2414016. doi: 10.1002/advs.202414016
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., et al. (2022). “Chain-of-thought prompting elicits reasoning in large language models,” in *Advances in Neural Information Processing Systems* 35. Available at: <https://arxiv.org/abs/2201.11903> (Accessed September 29, 2026).
- Wong, K.-F., and Wang, X.-J. (2006). A recurrent network mechanism of time integration in perceptual decisions. *Journal of Neuroscience* 26, 1314–1328. doi: 10.1523/JNEUROSCI.3733-05.2006
- Xiao, X., Wei, K., Zhong, J., Wei, X., and Zhou, M. (2025). Exploring similarity between neural and LLM trajectories in language processing. *arXiv [Preprint]*. Available at: <https://arxiv.org/abs/2509.24307> (Accessed August 14, 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.