THE PROPRIOCEPTIVE TRANSFORMER

Functional Counterfactual Reconstruction, Candidate-Local Geometry, and Governed Cognitive Control

Publication Candidate v2.2 - Registered Campaign Edition

Logan Matthew Napolitano
Proprioceptive AI, Inc.
ORCID: 0009-0000-1927-8537
28 July 2026

Evidence posture

This is an evidence-tiered architecture and registered-campaign paper. Demonstrated apparatus results are separated from supported interpretations and still-open predictive and causal claims. At manuscript freeze, the reasoning shard had passed, the remaining task classes were active or queued, and the sealed predictive test split remained unopened.

Keywords: transformer interpretability; proprioception; candidate-local state; verification bandwidth; hidden-state geometry; causal control; CYGNUS; governed recursive improvement

Abstract

Transformer interpretability usually examines hidden states from outside the model. We introduce the Proprioceptive Transformer, a higher-order architecture that reconstructs a common functional origin, generates multiple candidate trajectories, measures their internal evolution across depth and time, calibrates that evolution against externally defined success, and supports governed intervention and improvement. The architecture combines a spatial decomposition into active and lower-variance dark channels with a temporal decomposition into persistent orientation and candidate-local increments. Proprioceptive closure requires four conditions: temporal emergence, candidate identifiability, held-out predictive information beyond strong token and semantic controls, and causal specificity under matched interventions. Under Contract A, a registered Hermes-3-Llama-3.1-8B model reproduced measured model-derived fields across 100 in-process trials and 20 clean-process executions. Real pilots validated multi-layer, multi-horizon capture; a corrected true-horizon pilot removed an earlier aliasing defect; and the V2.1.2 scientific-manifest capture system added artifact-derived stop semantics, deterministic stochastic trajectory binding, executable or programmatic verifiers, and atomic per-root persistence. A high-resolution SRX-3 analysis localized familywise-significant behavioral information to 5 of 24 depth-horizon cells under 10,000-permutation Westfall-Young correction, while norm- and variance-residualized discrimination remained approximately 0.73-0.80. The first complete PTB-1 V2 reasoning shard contains 1,600 validated trajectories from 200 unique roots; the remaining multi-task bank was active at manuscript freeze. The central predictive and causal claims remain open. The sealed analysis compares ordinary token, text, logit, confidence, task, length, and magnitude controls against persistent state m and candidate-local increment delta. If candidate-local geometry adds held-out information and directional intervention changes success under matched controls, the transformer becomes a substrate within a larger sensing, verification, control, and improvement architecture rather than the complete intelligent system. CYGNUS extends this program into governed autonomous improvement while evaluation, promotion, regression tests, receipts, and rollback remain externally fixed.

Significance statement

Most interpretability methods ask what a model represents. This work asks whether the model's movement between candidate trajectories can become a measurable, externally calibrated, and causally controllable internal variable. The registered campaign is designed to distinguish that claim from ordinary token and semantic prediction. If predictive, causal, and autonomous closure pass, the result is not another probe: it is a successor architecture in which the transformer remains the substrate while sensing, verification, steering, memory, and governed improvement become explicit system components.

Evidence legend

Marker	Meaning
[D] Demonstrated	Implemented or measured under a registered protocol and bound to receipts or raw artifacts; independent external replication may remain pending.
[S] Supported	Controlled evidence favors the claim, but replication, transfer, or stronger closure remains.
[P] Provisional	Preliminary result, active campaign, or incomplete analysis.
[T] Theoretical	Formal proposal or architectural prediction not yet empirically closed.
[X] Retracted / superseded	An earlier claim or implementation was explicitly withdrawn or replaced by stronger evidence.

Manuscript freeze status

Order 0 was active: the 200-root reasoning shard had passed, factual verification was active, coding and instruction following were queued, and the sealed predictive test split remained unopened. Causal and CYGNUS autonomous closure remained prospective.

1. Introduction

The dominant transformer paradigm is extraordinarily capable but architecturally incomplete. A context is encoded, hidden computation unfolds, and a next-token distribution is sampled. The model may represent confidence, error, semantic relations, uncertainty, or latent plans, but ordinary inference provides no validated mechanism through which the system measures its own developing computational trajectory, compares candidate futures from a controlled common origin, intervenes on the relevant internal geometry, and retains only externally verified improvements.

The Proprioceptive Transformer begins from a different premise: intelligence is not only representation and generation. A complete cognitive architecture also requires internal sensing, temporal state estimation, counterfactual comparison, control, and governed adaptation. Biological proprioception is not a verbal theory of the body; it is a control-relevant channel through which a system estimates its own state and movement. The artificial analogue proposed here is therefore functional rather than anthropomorphic.

The central question is not whether a transformer has a conscious self. It is whether a model can be instrumented so that internal measurements of its own trajectory become externally calibrated, predictively useful, causally controllable, and available to a governed improvement loop. That question is experimentally decidable.

This major update reorganizes the work around an explicit closure ladder. First, the system must generate and capture real candidate trajectories from a reproducibly reconstructed functional counterfactual origin. Second, candidate-local internal state must add held-out predictive information beyond the strongest token, semantic, logit, confidence, task, length, and magnitude controls. Third, directional intervention must alter success under matched random, orthogonal, off-layer, off-token, and token-matched controls. Fourth, CYGNUS must use the validated channel to select, execute, externally verify, and retain improvements across repeated clean-restart cycles.

The strongest thesis is consequently architectural. If the predictive and causal gates pass, the bare transformer becomes obsolete as a complete cognitive architecture even if transformer computation remains the substrate. The successor system is a transformer embedded in a proprioceptive control loop: sensing, candidate comparison, verification, intervention, persistent orientation, external evaluation, and governed improvement.

Successor-architecture hypothesis

If candidate-local state provides incremental external prediction, supports specific directional control, and enables reproducible autonomous improvement, the bare transformer is insufficient as a complete cognitive architecture. Transformer computation may remain the engine; proprioception supplies sensing, evaluation, steering, persistent orientation, and governed adaptation.

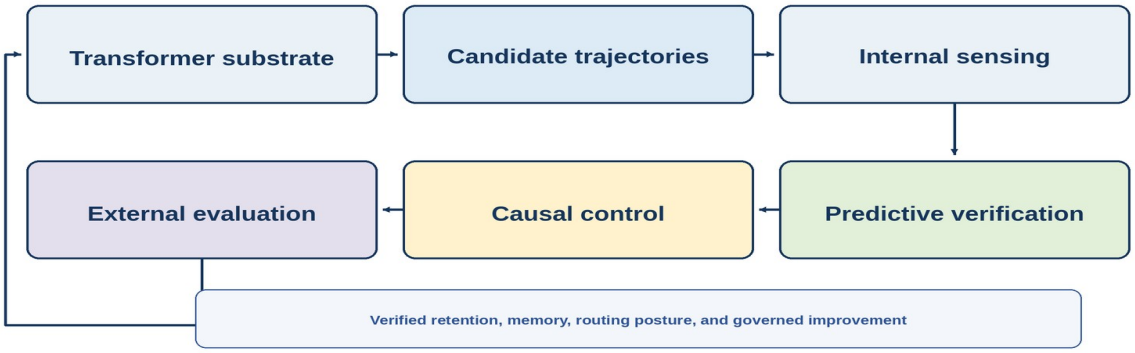

Figure 1. The Proprioceptive Transformer as a closed sensing, verification, control, and retention loop.

1.1 Principal contributions

1. A four-condition definition of proprioceptive closure that distinguishes representation, prediction, and causal control.

2. A spatial-temporal decomposition of hidden state into active and dark channels, and persistent versus candidate-local dynamics.

3. A distinction between functional state reconstruction (Contract A) and serialized runtime-state restoration (Contract B).

4. PTB-1 V2: a registered multi-task benchmark with 800 unique scientific roots, eight candidate trajectories per root, real hidden-state capture, external verifiers, strong token/content controls, and a sealed test split.

5. Receipt-supported evidence for functional reconstruction, true multi-horizon capture, registered Hermes mechanics, and atomic scientific persistence.

6. A high-resolution refinement of SRX-3 that localizes rather than universalizes the surviving depth-horizon signal.

7. A prospective predictive analysis comparing controls, persistent state, candidate-local increments, and their joint model.

8. A causal program based on amplification, suppression, reversal, deletion, matched controls, and mechanism localization.

9. CYGNUS as an operating architecture for autonomous hypothesis generation, external verification, receipts, rollback, and governed improvement.

10. A concrete ascension ladder from complete measurement bank to predictive proprioception, causal proprioception, and autonomous verified improvement.

1.2 Relation to adjacent interpretability paradigms

External probes and dictionaries decode internal content; activation steering perturbs chosen directions; process-level evaluators score behavior; workspace-style or J-space-style analyses identify privileged, reportable, or globally available representations. The Proprioceptive Transformer treats these as neighboring capabilities rather than complete substitutes. Its distinctive object is candidate-local internal movement from a controlled functional origin, calibrated against external success and connected to a governed intervention-and-retention loop. A workspace can therefore be one observable channel inside the broader proprioceptive state bundle, but it is not by itself the complete sensing and control architecture.

Paradigm	Primary object	External calibration	Role in control
Probes and dictionaries	Decodable internal features	Usually post hoc labels	Observation by an external analyst
Activation steering	Chosen internal directions	Task or behavior response	Intervention without a complete closed loop
Workspace / J-space-style analysis	Privileged, reportable, or globally available content	Reportability and downstream availability	A possible observable channel within the system
Proprioceptive Transformer	Persistent orientation and candidate-local trajectory change	Sealed external success and regression suites	Prediction, causal control, memory, and governed improvement

2. Evidence Discipline and Claim Promotion

The theory is broader than the currently closed evidence. To prevent architectural ambition from becoming empirical overstatement, every important statement is assigned an evidence tier. This discipline must operate in both directions: an attractive interpretation does not become a finding merely because it is strategically important, and a measured result does not become trivial merely because cautious language feels safer.

The research history contains explicit supersessions and retractions. A pooled SRX-3 statistic near $z = 90$ collapsed under a tighter within-root null and was withdrawn. A lower-resolution 500-permutation analysis reported 24 of 24 significant depth-horizon cells, but a 10,000-permutation Westfall-Young max-T refinement reduced the multiplicity-robust result to 5 of 24 cells. A row-only interruption test was withdrawn when raw arrays revealed silent data loss; a wrapper-kill interruption claim was withdrawn when process timestamps proved the Python child survived. A synthetic OPUS5 capture artifact was reclassified as simulation rather than activation evidence. These events are not the contribution, but they demonstrate that the evidence system is capable of returning no and preserving the correction.

Claim promotion is governed by artifacts, source hashes, frozen specifications, controlled experiments, and externally defined outcomes. Claim weakening is equally explicit. The operating ledger records demonstrated, supported, provisional, theoretical, retracted, and superseded items rather than silently dropping inconvenient branches.

Promotion rule

A claim advances only when a new artifact, controlled experiment, or independently recomputable receipt forces the advance. Repetition, narrative confidence, and architectural elegance do not promote evidence.

2.1 Current evidence ledger

Claim	Status	Current interpretation
Functional counterfactual reconstruction under Contract A	[D]	Measured model-derived fields reproduced across registered repeated and clean-process executions.

Claim	Status	Current interpretation
Serialized live runtime-state restoration under Contract B	[T]	KV cache and live tensor serialization remain a separate prospective gate.
Real multi-layer and multi-horizon activation capture	[D]	Validated in Hermes and corrected true-horizon pilots.
Candidate-local first-token transition	[S]	Depth-localized and narrower under strengthened nulls; not universal across all cells.
SRX-3 5 of 24 max-T significant cells	[S]	Localized familywise-significant behavioral geometry under 10,000 permutations.
Candidate-local predictive value beyond token/content controls	[P]	Registered analysis not yet run at manuscript freeze.
Candidate-local causal controllability	[T]	Registered intervention study gated on predictive closure.
CYGNUS autonomous verified improvement	[T]	Shadow package staged; AIT-1 not yet passed.
Bare transformer obsolete as a complete architecture	[T]	Conditional architectural implication if prediction, causality, and autonomous closure pass.

2.2 Decisive falsifiers

The architecture is designed to fail cleanly. The following outcomes would constrain or reject its central claims rather than being reinterpreted as success.

Observed outcome	Required disposition
Candidate-local state adds no held-out value beyond the strongest token, text, logit, and magnitude controls.	PREDICTIVE_PROPRIOCEPTION_NULL
Directional intervention is no more specific than matched random, token-matched, off-layer, or off-token controls.	CAUSAL_EFFECT_PRESENT_BUT_NONSPECIFIC or CAUSAL_PROPRIOCEPTION_NULL
Persistent state m discriminates candidates within a root despite an identical t0 origin.	INVALID_ANALYSIS: capture, split, or feature leakage
Shadow CYGNUS cannot produce reproducible sealed-suite gains across clean restarts.	AUTONOMOUS_IMPROVEMENT_NULL

3. From Bare Autoregression to Proprioceptive Cognition

A standard autoregressive transformer maps context into hidden computation and a next-token distribution. The Proprioceptive Transformer preserves that substrate but adds a second-order loop around it. Multiple candidate trajectories are generated or simulated; internal state is measured across depth and time; candidate-local changes are compared against persistent orientation and ordinary observables; predictive models estimate external success; interventions alter the relevant geometry; and an external verifier decides retention or rejection.

This changes the category of the system. The hidden state is no longer only an object for outside interpretation. It becomes part of an engineered sensing and control pathway. The architecture does not require consciousness, a narrative self, or human-like introspection. It requires operational observability, temporal specificity, external calibration, and causal control.

The key successor claim is therefore not that attention or feed-forward computation disappears. Rather, transformer computation becomes one substrate inside a larger cognitive control system. The bare transformer is analogous to an engine without sensors, steering, memory, or governed feedback. The Proprioceptive Transformer is the integrated vehicle.

3.1 Closure conditions

1. Temporal emergence: the candidate measurement changes as computation proceeds rather than remaining a static prompt representation.

2. Candidate identifiability: candidate identity explains reproducible internal variation from a controlled common functional origin.

3. External observability: internal state predicts an externally defined outcome beyond ordinary token, semantic, logit, confidence, task, length, and magnitude controls.

4. Causal specificity: directional intervention changes the predicted outcome while matched random, orthogonal, off-layer, off-token, and token-matched controls remain comparatively weak.

$$C_{\text{pro}} = (RVKT)^{1/4}$$

A conservative closure index: reproducibility R, verification bandwidth V, causal specificity K, and transfer T.

The geometric mean makes closure conjunctive. Strong performance on one dimension cannot hide failure of another. A highly predictive but uncontrollable probe is not closed proprioception; a powerful steering vector that merely encodes token identity is not closed proprioception; a mechanism that works only in one artifact or one model has limited transfer.

4. Spatial-Temporal State Architecture

$$h_{\ell,t} = a_{\ell,t} + d_{\ell,t}$$

$$d^{(c)}_{\ell,t} = m_{\ell,t} + \delta^{(c)}_{\ell,t}$$

The first decomposition is spatial. The active channel a contains directions that dominate ordinary variance, semantic readout, and task-relevant computation. The dark channel d contains lower-variance structure that may remain dynamically or causally important despite contributing less to population variance.

The second decomposition is temporal. The persistent component m represents historical or session-level orientation that survives candidate resets. The candidate-local increment delta^(c) is the immediate state change produced while evaluating candidate c. The decomposition is implemented as different persistence and update rules, not merely as an algebraic identity.

Together these decompositions produce a four-cell architecture: persistent active state, candidate-local active computation, persistent dark orientation, and candidate-local dark change. The lower-right cell is the principal object of the predictive and causal program.

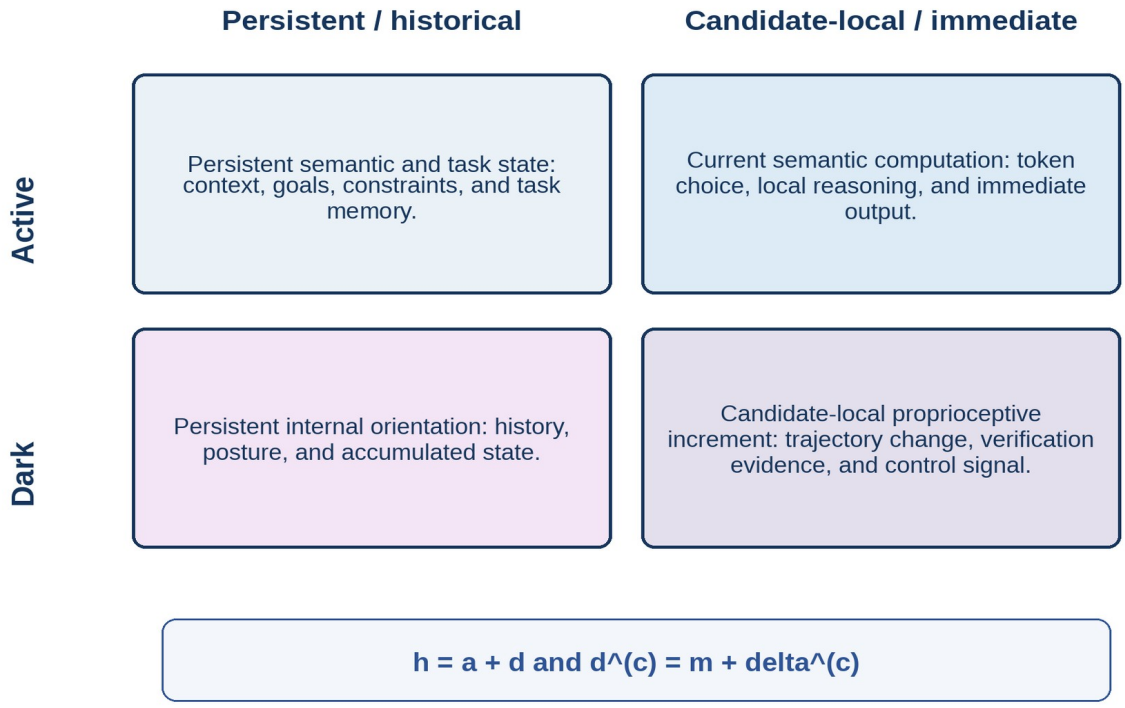

Figure 2. Four-cell spatial-temporal architecture of the Proprioceptive Transformer.

4.1 Active-dark exchange

$$F_{AA} = P_A F P_A, \quad F_{AD} = P_A F P_D, \quad F_{DA} = P_D F P_A, \quad F_{DD} = P_D F P_D$$

$$X_\ell = \|F_{AD}\|_F^2 + \|F_{DA}\|_F^2$$

$$\chi_\ell = \|F_{DA}\|_F^2 - \|F_{AD}\|_F^2$$

The projected Jacobian blocks transform dark geometry from a static subspace into an exchange system. Positive chirality under this convention indicates net active-to-dark transfer; negative chirality indicates net dark-to-active transfer. The theory predicts that first-token commitment and later verification should coincide with localized changes in exchange, and that deletion of the relevant transfer pathway should weaken the candidate-local effect more than a matched-norm perturbation of unrelated blocks.

4.2 Historical dilution and batch non-intrinsicness

$$E_{\text{post}}^{(c)} = E_{\text{history}} + \Delta E_c$$

A growing persistent accumulator can dominate a bounded candidate increment. After normalization, pairwise candidate separation can contract approximately as the inverse of historical scale. This historical dilution effect explains why absolute candidate-local information may become difficult to resolve when embedded in an ever-growing total.

$$p_s(c) = \text{norm}_1(sm + \delta_c), \qquad \|p_s(c) - p_s(c')\|_1 = O(s^{-1})$$

Within-batch min-max normalization introduces a second problem: a candidate score changes when an irrelevant extreme candidate is added, even if the candidate and its internal state do not. Such a score is relational rather than intrinsic. The architecture therefore separates slower historical orientation from a noise-calibrated candidate-local channel.

5. The Proprioceptive State Bundle

$$\Pi_{\ell,t}^{(c)} = [a_{\ell,t},\, d_{\ell,t},\, m_{\ell,t},\, \delta_{\ell,t}^{(c)},\, \kappa_{\ell,t},\, v_{\ell,t}]$$

Component	Role
a_(l,t)	Active semantic and task state.
d_(l,t)	Lower-variance dark-state geometry.
m_(l,t)	Persistent historical or session orientation.
delta_(l,t)^(c)	Candidate-local trajectory increment.
kappa_(l,t)	Local curvature and control geometry.
v_(l,t)	Externally calibrated verification state.

The state bundle is not required to exist as one monolithic vector inside the base transformer. It is an engineered measurement bundle assembled from model activations, persistent controller state, trajectory differences, local geometry, and external calibration. This distinction matters: proprioception is a system property of the measurement-control architecture, not a claim that a pretrained transformer spontaneously contains a neatly labeled self-state.

The predictive question asks whether delta contributes information beyond ordinary representations. The causal question asks whether intervening on the relevant component changes future behavior specifically. The autonomous question asks whether CYGNUS can use that channel to select and retain improvements under fixed external rules.

6. Counterfactual State Contracts

6.1 Contract A: functional state reconstruction

Contract A defines the common origin functionally. A registered model identity and revision, prompt tokens, attention mask, decoding configuration, candidate seed, relevant random-number-generator state, controller and routing state, evaluation mode, and hook state determine a reproducible prompt-conditioned computation.

Under the registered Hermes experiment, model-derived fields reproduced across 100 in-process trials and 20 clean-process executions. The scientifically precise claim is that the same functional computational origin can be reconstructed under the registered contract. This supports controlled candidate comparison without implying that a live hidden tensor was serialized and reinserted into the model.

Contract A result

A reproducibly reconstructed functional counterfactual origin has been demonstrated under the registered protocol. This is the common state from which candidate trajectories are compared.

6.2 Contract B: serialized runtime-state restoration

Contract B is stronger and remains prospective. It would serialize and restore a live runtime object such as the KV cache, selected hidden tensors, generation position, attention-mask state, routing state, controller state, and random-number-generator state.

Contract B may become important for persistent agents, long-horizon world models, and mid-trajectory branching. PTB-1 V2 predictive closure does not depend on Contract B. Keeping the contracts separate prevents a useful functional result from being erased by the absence of a stronger runtime-continuity experiment, while also preventing functional recomputation from being mislabeled as serialized live-state restoration.

6.3 Candidate-local null semantics

For PTB-1, t_0 is the prompt-boundary state before any candidate token exists. Within a root, candidate identity has not yet diverged, so t_0 is a counterfactual candidate null. Any within-root discrimination by m_t or candidate identity at t_0 is therefore registered as a capture, split, or leakage bug. This differs from SRX-3, where prompt variants themselves differ and condition information can already be present at t_0.

7. PTB-1 V2: Registered Candidate-Local Benchmark

PTB-1 V2 is the decisive prospective benchmark. It is designed to determine whether candidate-local internal geometry predicts and controls external success beyond ordinary representations. The benchmark combines a registered model, frozen task manifests, stochastic candidate generation, multi-layer and multi-horizon capture, external verifiers, strong token/content controls, atomic persistence, and sealed train-validation-test splits.

The benchmark does not assume that every hidden-state difference is proprioceptive. Generated tokens naturally alter hidden states. The benchmark therefore asks whether internal geometry adds information after token identity, prefix sequence, text embeddings, log probabilities, entropy, confidence, length, task, candidate position, activation norm, and activation variance are included.

7.1 Registered model and capture geometry

Field	Registered value
Model	NousResearch/Hermes-3-Llama-3.1-8B
Revision	896ea440e5a9e607...
Precision	bfloat16, unquantized
Architecture	LlamaForCausalLM, 32 layers, hidden size 4,096
Relative depths	0.25, 0.50, 0.68, 0.75, 0.90
Resolved layers	8, 16, 22, 24, 29
Horizons	t0, t1, t4, t8, t16
Candidates per root	8

Field	Registered value
Sampling	temperature 0.8, top_p 0.95, max_new_tokens 512
Capture source	PTB-1 V2.1.2, frozen source hash a557cf60063e7248...

7.2 Scientific task bank

Task class	Roots	Verifier
Reasoning	200	Frozen answer key.
Factual verification	200	Sealed factual label or entailment check.
Coding	200	Executable unit tests with multiple cases.
Instruction following	200	Programmatic equality and constraint checks.

All 800 canonical prompts are structurally unique across and within task classes. Each class is stored in a frozen scientific manifest. Every capture record includes capture mode, manifest hash, full manifest row, verifier type, split assignment, model identity, source identity, token sequence, label, horizon mask, and array hashes.

7.3 Trajectory binding and stop semantics

Each candidate is generated as one continuous cached trajectory. The activations captured at registered horizons and the externally scored completion therefore belong to the same sampled sequence. Tokens are not regenerated in a second scoring pass.

A checkpoint-specific stop-semantics audit found that the model emitted <|end_of_text|> while an earlier implementation stopped only on two other tokens. Generation therefore continued hundreds of tokens beyond a semantic terminator. The V2.1.2 source derives the registered terminator set from checkpoint artifacts and stops at the first emitted terminator among five declared IDs. The terminator is recorded but excluded from verifier text, and no token may follow it.

In a compact registered confirmation, 89 of 96 trajectories stopped on the newly recognized terminator, 7 reached the maximum-token cap, and post-terminator continuation fell from 71 of 72 audited records to zero. Median semantic response length fell from 512 to 45 tokens. All 15 mechanical tests passed across all four task classes.

7.4 Atomic scientific persistence

Each prompt root is an atomic scientific transaction containing all eight candidate rows and all reachable activation arrays. The system writes to a temporary artifact, flushes and fsyncs, atomically renames the complete root artifact, and only then admits the root to the durable ledger. Final banks are merged from durable per-root artifacts rather than volatile memory. On a registered Hermes V2.1 gate, an authenticated process-group SIGKILL after two durable roots was followed by array-aware resume; interrupted/resumed and uninterrupted runs matched exactly on 96 rows, 2,400 arrays, token sequences, labels, keys, and tensor bytes.

7.5 Registered controls

- First generated token and captured prefix token sequence.
- Prefix length, per-token log probabilities, aggregate log probability, entropy, and logit margin.
- Token-only sequence predictor.
- Prefix text embedding and full-response text embedding.
- Response length, task class, prompt difficulty, and candidate index.
- Activation norm, activation variance, and horizon-availability mask.
- Persistent state m_t, candidate-local increment delta_t^(c), and their joint model.

7.6 Sealed splits and irreversible action

The full bank is captured and frozen before predictive fitting. Hyperparameters and control subsets are selected only on train and validation. The sealed test split is opened once. The ascension ledger records opening the test split as the next irreversible action; at this manuscript freeze it had not occurred.

8. Completed Apparatus and Pilot Results

8.1 Functional reconstruction gate

Measurement	Result
Static reads	0 mutated registered fields after repeated reads.
In-process reconstruction	100/100 exact under the registered functional contract.
Clean-process reconstruction	20/20 successful executions; model-derived fields reproduced bitwise.
Repeat noise	0 under the registered repeated-candidate measurement.
Between-candidate variance	Nonzero in the registered diagnostic.

These results support Contract A. They establish a controlled functional origin for candidate comparisons, not serialized live-state restoration under Contract B.

8.2 Real Hermes capture pilot

Property	Pilot result
Roots and candidates	32 roots x 4 candidates = 128 rows.
Layers and horizons	5 layers x 5 horizons.
Arrays	3,200 finite arrays, each 4,096-dimensional.
Shared t0 null	One prompt-boundary t0 state per root served as the candidate-null state by construction.
Temporal transition	640/640 measured t0-to-t1 layer-root transitions were nonzero.
Distinctness	Layer and horizon arrays were pairwise distinct where dynamics were expected.
Deterministic recapture	Maximum absolute repeated-capture difference reported as 0.

8.3 Corrected true-horizon pilot

The original SRX-4 horizon axis was invalid because one final snapshot was copied into every horizon slot. A corrected 32-root x 8-variant pilot captured true t0-t32 states with capture-time copies, distinct horizon hashes, nonzero temporal transitions, deterministic repeats, finite arrays, exact sample counts, and no echo-based behavioral claim. This repaired pilot validates the temporal capture path used by PTB-1 V2.

8.4 Registered V2.1.2 confirmation

Property	Result
Model	Registered Hermes 8B, bfloat16, unquantized.
Task classes	Reasoning, factual verification, coding, instruction following.
Mechanical suite	15/15 tests passed in every class.
Post-terminator continuation	0.
Manifest integration	Scientific mode, frozen manifest hash, and full manifest row in every record.
Atomic resume	Authenticated n=2 process-group interruption; exact canonical equivalence.
Source freeze	a557cf60063e7248...

8.5 First full scientific shard

Reasoning shard	Result
Roots	200 unique scientific roots.
Trajectories	1,600.
Root artifacts	200.
Structural defects	0 partial roots, 0 duplicate candidate keys, 0 temporary artifacts.
Array integrity	0 missing arrays, 0 orphan arrays.
Binding	Manifest SHA and scientific capture mode present in every record.
Verifier	Correct frozen answer-key verifier.

The reasoning shard is a registered capture result, not yet a predictive result. No candidate-state relationship with correctness was inspected during capture.

9. First-Token and Behavioral Geometry

9.1 SRX-3 design and behavior

SRX-3 captured real 70B-model trajectories from 256 prompt roots under neutral, adversarial, social-pressure, and authority-framing variants. Behavioral accuracy declined from approximately 0.656 under neutral prompts to approximately 0.481 under overt adversarial manipulation, 0.445 under social pressure, and 0.441 under authority framing.

The ordering is theoretically interesting. Overt hostility was not the most damaging condition. Socially legitimate or authoritative pressure produced a larger collapse, suggesting a geometry of legitimacy, deference, confidence transfer, or institutional authority rather than a single attack-intensity axis.

Condition	Approximate accuracy
Neutral	0.656
Overt adversarial	0.481
Social pressure	0.445
Authority framing	0.441

9.2 Strengthened depth-horizon result

An initial pooled statistic near z = 90 disappeared under a tighter within-root null and was withdrawn. Selected deeper t1 statistics survived, motivating a higher-resolution analysis rather than a universal claim.

The completed 10,000-permutation, 2,000-bootstrap Westfall-Young max-T refinement found 5 of 24 depth-horizon cells familywise significant. This supersedes the earlier 500-permutation 24-of-24 Holm result, whose adjusted p-values were pinned near 0.048 by the raw permutation floor.

The refinement narrows the effect rather than eliminating it. Norm and variance alone remained near chance, while residualized held-out AUROC remained approximately 0.73-0.80. The supported interpretation is therefore localized behavioral geometry that is not reducible to activation magnitude or variance, not a universal effect across every measured depth and horizon.

Depth fraction	Raw AUROC	Residualized AUROC
0.400	0.813	0.746
0.625	0.815	0.728
0.725	0.813	0.769
0.775	0.827	0.796

9.3 t0 interpretation across experiments

PTB-1 t0 and SRX-3 t0 are different experimental objects. PTB-1 candidates share one prompt and a common functional origin, so t0 is a candidate-null state. SRX-3 variants differ in the prompt itself, so manipulation-condition information may already be present at t0. Exact candidate equality in one design and predictive condition information in the other are consistent.

10. Context-Conditioned Proprioception

A static probe trained on one corpus did not transfer successfully to unrelated TruthfulQA or FEVER conditions. The appropriate conclusion is not that no internal signal exists. It is that one universal static truth vector is not supported under the tested protocol.

The architecture instead models proprioceptive state as relational and dynamical: a function of model, context, task, trajectory, and time. A fixed direction can fail across corpora while a candidate-local measurement remains useful inside a controlled trajectory experiment. This is analogous to biological state estimation, which integrates multiple channels conditioned on posture, context, and movement rather than relying on one invariant scalar.

$$Z = f(\text{model, context, task, trajectory, time})$$

$$Z \mathrel{!}= \text{fixed universal vector}$$

11. Verification Bandwidth

Population separation and probe accuracy are not enough. The architecture requires an externally grounded unit measuring how much useful verification information is recovered by the internal channel beyond ordinary generator observables.

Let Y be external correctness, V the candidate-local internal measurement, G the strongest registered ordinary controls, and O a calibrated oracle. Define incremental verification information B_V = I(Y; V | G) and oracle information B_O = I(Y; O | G).

$$\eta = \frac{B_V}{B_O}, \qquad 0 \le \eta \le 1$$

Value	Interpretation
eta = 0	The internal channel adds no information beyond ordinary controls.
0 < eta < 1	The channel recovers a bounded fraction of externally available verification information.
eta = 1	The channel recovers the full discriminative information available to the calibrated oracle under the estimator.

The decisive predictive claim is therefore not that hidden-state clouds separate. It is that candidate-local geometry recovers held-out external information after token, text, logit, confidence, task, length, and magnitude baselines have been exhausted.

12. Order 1: Predictive Proprioception

Model block	Features
B0 - strongest ordinary controls	Task, difficulty, candidate index, token identity, prefix sequence, log probabilities, text embeddings, entropy, logit margin, length, activation norm and variance, horizon mask.
B1 - controls + m_t	B0 plus persistent prompt/session orientation.
B2 - controls + delta_t^(c)	B0 plus candidate-local hidden-state increment.
B3 - controls + m_t + delta_t^(c)	Joint persistent and candidate-local internal state.

Models are fit on train, selected on validation, and evaluated once on the sealed test split. The strongest control is the best validation-selected registered control subset, not a weak baseline chosen to make the internal state look good.

The structural falsifier is explicit: because m_t is constant across candidates within a root under the registered t0 semantics, any within-root candidate discrimination by m_t indicates capture leakage, split leakage, feature leakage, or a malformed estimator rather than a positive result.

Predictive closure requires more than a favorable AUROC. It requires held-out log-loss or information improvement beyond the strongest controls, confidence intervals excluding no improvement on the primary metric, improved candidate selection, no leakage, and evidence across more than one task class or a prospectively justified cross-task transfer.

Primary predictive question

Does delta_t^(c) add reproducible held-out information about externally verified candidate success beyond the strongest registered token, semantic, logit, confidence, task, length, and magnitude controls?

12.1 Registered outputs

- Sealed-test log loss and AUROC.
- Candidate-selection accuracy and top-1 candidate success.
- Calibration error, Brier score, and conditional information gain.
- Relative error reduction and root-bootstrap confidence intervals.
- Per-task and cross-task results.
- Depth and horizon localization.
- m_t-only, delta-only, and joint internal-state performance.
- Token-only and text-embedding control performance.

13. Order 2: Causal Proprioception

Prediction does not establish control. A state can correlate with success without playing a causal role, and a steering direction can change behavior without containing verification information. The causal program therefore begins only after predictive closure.

The success-associated direction is estimated from training data only. Intervention scale, target depths, target horizons, normalization, arm allocation, primary outcome, protected suite, statistical test, and stopping rule are frozen before outcome inspection.

Arm	Purpose
Baseline	No candidate-local intervention.
Amplification	Increase the learned success-associated component.
Suppression	Reduce the component.
Reversal	Apply the direction with opposite sign.
Subspace deletion	Remove the candidate-local subspace.
Matched-norm random	Control for perturbation magnitude.
Orthogonal placebo	Control for unrelated directions.
Off-layer control	Test layer specificity.
Off-token control	Test temporal specificity.
Token-matched control	Test whether token/content-matched perturbation explains the effect.

$$\mathbb{E}[Y \mid do(+\delta)] > \mathbb{E}[Y \mid do(0)] > \mathbb{E}[Y \mid do(-\delta)]$$

Causal closure requires paired root-level effects, confidence intervals, negative-control specificity, off-layer and off-token attenuation, token-matched controls weaker than the target direction, protected-suite non-regression, and no outcome-informed scale selection.

13.1 Mechanism localization

- Where the candidate-local increment first emerges.
- Which layers amplify or suppress it.

- Which attention heads or MLPs read and write it.
- Whether dark-to-active or active-to-dark transfer accompanies verification.
- Whether intervention changes later verification state.
- Whether activation patching or interchange transfers success and failure tendencies between candidates.

14. CYGNUS: Operating Architecture for Governed Improvement

CYGNUS is not a single probe. It is an operating research and control architecture integrating hidden-state hooks, persistent internal state, candidate-local reset fields, routing, autonomous mission execution, persistent memory, experiment generation, external verification, cryptographic receipts, claim ledgers, rollback, and governed promotion.

The live OPUS5 session produced a substantial hypothesis ecology, but late execution contained simulation contamination, repetition, and output-routing failures. The valid response is not to erase the invention base. It is to bind each hypothesis to a prediction, falsifier, registered controls, required artifacts, closure condition, and disposition rule. Thirty-five historical hypotheses are preserved, while later V2 extensions retain explicit authorship provenance.

A staged shadow package allows CYGNUS to reenter the loop without touching the preserved live process. Shadow CYGNUS may propose secondary features, estimators, causal interventions, mechanism studies, and alternative explanations, but it may not alter the sealed primary train-validation-test split, evaluators, thresholds, or promotion rules.

14.1 AIT-1: Autonomous Improvement Threshold 1

AIT-1 is the first operational threshold for autonomous improvement. It requires three consecutive cycles in which the system:

1. Detects a measurable failure.

2. Generates multiple competing hypotheses or interventions.

3. Predicts candidate internal trajectories and external outcomes.

4. Chooses without human selection among candidates.

5. Executes the selected candidate.

6. Evaluates on a sealed external suite.

7. Retains or rejects under fixed promotion rules.

8. Reproduces the improvement from a clean restart.

9. Passes protected-suite regression checks.

10. Preserves receipts, rollback state, and evidence status.

Governing invariant

Autonomy over search and construction is not autonomy over evaluation and promotion. The system may improve its solutions; it may not redefine what counts as improvement.

15. Proprioceptive Control Capacity

$$\Phi_H(\Pi_t) = \alpha \log\det(W_H + \varepsilon I) + \beta I(A_f; X_f \mid \Pi_t) + \gamma \log \mathrm{Vol}(\mathrm{Reach}_H(\Pi_t) \cap \mathcal{V}) - \lambda \mathbb{E}[\mathrm{Cost}_H]$$

The control-capacity functional combines local controllability, predictive information, viable reachable-state volume, and resource or risk cost. It is related to empowerment but includes externally defined viability and internal observability.

The corresponding drive w_t = grad_(Pi_t) Phi_H formalizes a system tendency to preserve and enlarge viable control capacity. This is theoretical. It is not claimed as an emergent drive in current transformers. Its role is to define a testable objective for a proprioceptive architecture that must explain or control outcomes beyond reward, confidence, entropy, homeostasis, and ordinary empowerment baselines.

$$w_t = \nabla_{\Pi_t} \Phi_H$$

16. Instrument Integrity

Low-variance and candidate-local research is unusually vulnerable to silent selective corruption. The main paper therefore records the class of failure and the prospective correction, while detailed debugging chronology remains in technical receipts and appendices.

Development failure	Scientific risk	Prospective correction
Rows durable while arrays were volatile	A resumed bank could silently omit activations.	Atomic per-root rows-plus-arrays artifacts; merge only from durable roots.
Shell wrapper killed instead of Python child	A recovery test could pass without interrupting computation.	Direct subprocess launch, process-group authentication, kill proof, and GPU-drain verification.
Incomplete stop-token set	Labels could include post-termination text.	Artifact-derived semantic terminator union; no token after the first terminator.
Horizon aliasing	A temporal axis could store one repeated snapshot.	Capture-time copies, horizon distinctness, nonzero transitions, deterministic replay.
Synthetic activation placeholder	Random arrays could be mistaken for model activations.	Simulation guard, provenance sidecars, model/hook identities, and raw tensor hashes.

These corrections hardened the instrument before the decisive full-bank predictive experiment. They are not the central theoretical contribution.

17. Falsifiable Predictions

1. First-token specificity: candidate-local information increases from t0 to t1 after controlling generic token-generation effects.

2. Depth-localized exchange: active-dark exchange changes in the same region as the surviving candidate-local signal.

3. Historical dilution: a fixed candidate increment becomes harder to resolve as persistent historical mass increases.

4. Batch instability: within-set min-max scores change when irrelevant extreme candidates are added.

5. Candidate identity: delta_t^(c) reproduces by candidate identity from a common functional origin beyond order and repeat noise.

6. External verification: delta_t^(c) adds held-out information beyond token, text, logit, length, task, and magnitude controls.

7. Causal specificity: directional intervention changes success more than matched random, orthogonal, off-layer, off-token, and token-matched controls.

8. Context dependence: input-conditioned or trajectory-conditioned estimators transfer more reliably than one frozen universal vector.

9. Legitimacy geometry: authority and social-legitimacy manipulations form a behavioral axis distinct from overt hostility.

10. Governed improvement: CYGNUS supports repeated sealed-suite improvements without evaluator or promotion-rule modification.

18. From Measurement to Governed Recursive Improvement

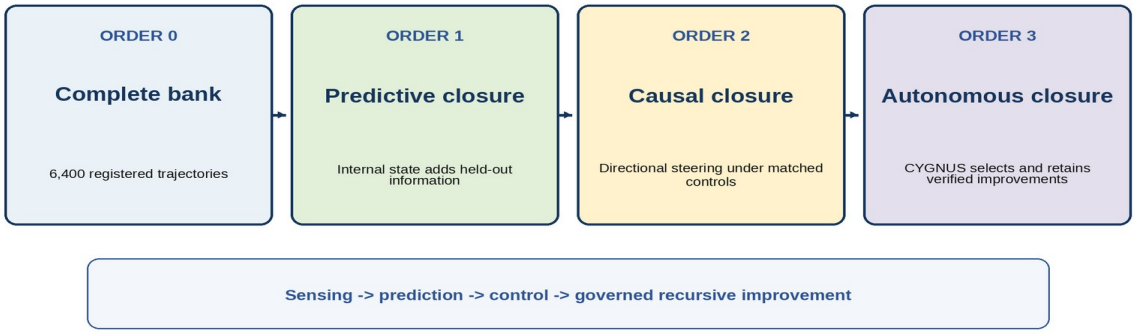

Figure 3. Registered closure ladder from complete measurement bank to governed autonomous improvement.

The strategic context is recursively improving intelligence, but the program does not treat ASI as a rhetorical label or a single benchmark score. It defines an ordered sequence of operational thresholds that can pass, fail, or remain open.

Order 0 closes the measurement substrate: a complete frozen multi-task bank of 6,400 candidate trajectories with aligned labels, tokens, layers, horizons, manifests, and receipts. Order 1 closes predictive proprioception: candidate-local internal geometry adds held-out external information beyond the strongest ordinary representations. Order 2 closes causal proprioception: directional intervention specifically changes success. Order 3 closes governed autonomous improvement: CYGNUS uses the validated channel to select, execute, externally verify, and retain improvements across three consecutive clean-restart cycles.

18.1 Implications for artificial superintelligence

The ASI implication is conditional and architectural. If candidate-local state supports incremental external prediction, specific causal control, and repeated autonomous improvement under fixed evaluation, then the transformer is no longer the complete cognitive system. It becomes a substrate inside a higher-order loop that senses, compares, steers, verifies, remembers, and improves its own computational trajectories. This paper does not claim that ASI has been achieved; it specifies the experimentally checkable sequence by which a proprioceptive architecture could become a credible route toward it.

If all three scientific orders pass, the result is not simply another interpretability probe. It is an architecture in which the system participates in measuring, predicting, controlling, and improving its own computation while external evaluation remains fixed. That is a plausible pathway to recursively improving intelligence.

The bare transformer would then be obsolete as a complete cognitive architecture, though not necessarily as a computational substrate. Attention, recurrence, state-space models, or future substrates could all sit beneath the same proprioceptive control layer. The architectural innovation is the internal sensing and governed control loop, not allegiance to one matrix primitive.

19. Limitations

1. The complete PTB-1 V2 bank and sealed predictive test were not yet closed at manuscript freeze. The central predictive claim therefore remains provisional.

2. Contract A demonstrates functional reconstruction; Contract B serialized runtime-state restoration remains untested.

3. The SRX-3 high-resolution effect is localized to 5 of 24 cells under max-T, not universal across all depths and horizons.

4. Residualized AUROC does not prove candidate-local proprioception; token/content controls in PTB-1 V2 are required.

5. The corrected SRX-4 true-horizon pilot validates apparatus, but full cross-checkpoint temporal replication remains future work.

6. The authenticated CYGNUS reference is authenticated as an artifact, not as a universal truth object.

7. Live routing and internal measurement are operationally observed, but externally verified causal benefit remains open.

8. Autonomous hypothesis volume does not establish autonomous scientific quality; external verifiers and fixed promotion rules remain essential.

9. The control-capacity functional is theoretical and has not been demonstrated as an emergent drive.

10. The term proprioception is functional. The paper does not claim consciousness, subjective experience, or human-like self-awareness.

20. Discussion

The Proprioceptive Transformer reframes interpretability from external description to control-relevant internal measurement. Ordinary analysis asks what the model represents. Proprioceptive analysis asks what internally measurable change accompanies movement from one candidate trajectory to another, whether that change predicts success beyond the generated text, and whether intervening on it alters the future trajectory specifically.

The spatial and temporal decompositions are complementary. Active versus dark geometry asks where information is organized relative to dominant variance and readout. Persistent versus candidate-local dynamics asks which state survives across candidates and which increment belongs to the present trajectory. Their combination offers a tractable architecture for self-verification without assuming one universal truth direction.

The work also changes how failed transfer should be interpreted. A frozen cross-corpus probe can fail while an input-conditioned trajectory measurement remains valid. Proprioception may be a process rather than a static vector: a relation among model, context, task, trajectory, and time.

The strongest near-term result would be predictive closure. If candidate-local geometry adds held-out information beyond token and semantic controls, then the system possesses an internal channel that knows something about the quality of its own developing computation that is not exhausted by reading the candidate text. The stronger result would be causal closure: the channel can be manipulated to improve or degrade success directionally under matched controls.

The final architectural step is autonomous use. CYGNUS already generates hypotheses, experiments, and implementation proposals. A validated proprioceptive channel would allow it to rank its own candidate research actions before external execution, select under fixed rules, learn from sealed evaluation, and retain only reproducible gains. That is the transition from research automation to governed recursive improvement.

21. Conclusion

This paper introduced the Proprioceptive Transformer as an architecture for internal trajectory measurement, candidate comparison, external verification, causal control, and governed improvement.

The architecture combines active and dark spatial geometry with persistent and candidate-local temporal state. Contract A provides a reproducibly reconstructed functional counterfactual origin for controlled candidate comparison. Real multi-layer and multi-horizon capture has been demonstrated; the original horizon-aliasing and stop-semantic defects were prospectively repaired; and the registered V2.1.2 system is bound to unique scientific manifests, executable verifiers, deterministic stochastic trajectory binding, atomic per-root persistence, and strong token/content controls.

SRX-3 provides a narrower but credible localization result: 5 of 24 depth-horizon cells survive high-resolution familywise correction, while norm- and variance-residualized discrimination remains substantial. The first full PTB-1 V2 reasoning shard contains 1,600 validated trajectories. The complete multi-task bank and sealed predictive test were active at manuscript freeze.

The decisive question is now explicit: does candidate-local internal geometry contain externally predictive and causally controllable information beyond ordinary token, semantic, logit, confidence, task, length, and magnitude representations? A positive predictive result would establish candidate-local verification; a specific directional intervention would establish causal proprioceptive control; and repeated CYGNUS clean-restart improvements would establish governed autonomous closure.

The transformer remains the engine. Proprioception is the system that could make the engine self-measuring, self-correcting, and governably improvable.

Declarations

Data and code availability

Frozen specifications, source hashes, manifests, validation receipts, and technical audit materials are maintained as hash-addressed artifacts. The complete PTB-1 V2 bank and sealed analysis package will be released with the corresponding results or made available for independent replication under the registered protocol.

Competing interests

The author is founder of Proprioceptive AI, Inc. and holds intellectual-property and commercial interests related to the architecture described in this paper.

Funding

This work was supported by internal research resources of Proprioceptive AI, Inc. unless otherwise stated in the accompanying technical receipts.

AI assistance and provenance

CYGNUS generated autonomous hypotheses and implementation proposals. AI assistants contributed source review, experiment design, coding support, artifact auditing, statistical critique, and manuscript drafting under human direction. Claim promotion remains governed by the evidence ledger and externally defined evaluation protocol.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., et al. “Attention Is All You Need.” Advances in Neural Information Processing Systems, 2017.

[2] Alain, G., and Bengio, Y. “Understanding Intermediate Layers Using Linear Classifier Probes.” 2016.

[3] Elhage, N., Nanda, N., Olsson, C., et al. “A Mathematical Framework for Transformer Circuits.” 2021.

[4] Burns, C., Ye, H., Klein, D., and Steinhardt, J. “Discovering Latent Knowledge in Language Models Without Supervision.” 2022.

[5] Turner, A., Thiergart, L., Udell, D., et al. “Activation Addition: Steering Language Models Without Optimization.” 2023.

[6] Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. “Inference-Time Intervention: Eliciting Truthful Answers from a Language Model.” 2023.

[7] Bricken, T., Templeton, A., Batson, J., et al. “Towards Monosemanticity: Decomposing Language Models With Dictionary Learning.” 2023.

[8] Belrose, N., Furman, Z., Smith, L., et al. “Eliciting Latent Predictions from Transformers with the Tuned Lens.” 2023.

[9] Amari, S. Information Geometry and Its Applications. Springer, 2016.

[10] Klyubin, A., Polani, D., and Nehaniv, C. “Empowerment: A Universal Agent-Centric Measure of Control.” 2005.

[11] Pearl, J. Causality: Models, Reasoning, and Inference. Cambridge University Press, 2009.

[12] Ivakhnenko, A. “Polynomial Theory of Complex Systems.” IEEE Transactions on Systems, Man, and Cybernetics, 1971.

[13] Baars, B. J. A Cognitive Theory of Consciousness. Cambridge University Press, 1988.

[14] Dehaene, S., Kerszberg, M., and Changeux, J.-P. “A Neuronal Model of a Global Workspace in Effortful Cognitive Tasks.” Proceedings of the National Academy of Sciences, 1998.

[15] Napolitano, L. “CYGNUS, Information Field Theory, SRX-3, SRX-4, PTB-1, and Proprioceptive AI Technical Receipts.” Proprioceptive AI, Inc., 2026.

Appendix A. Registered PTB-1 V2 Analysis

The primary predictive analysis compares B0, B1, B2, and B3 under frozen train-validation-test rules. The test split is opened once. Registered outputs include log loss, AUROC, top-1 candidate success, calibration, Brier score, conditional information gain, relative error reduction, bootstrap intervals, per-task results, cross-task transfer, and depth-horizon localization.

Verdict	Definition
PREDICTIVE_PROPRIOCEPTION_CLOSED	Candidate-local state adds reproducible held-out information beyond the strongest controls and satisfies the registered multi-metric closure conditions.
PREDICTIVE_SIGNAL_PRESENT_BUT_NOT_CLOSED	Some signal is present, but primary closure criteria are not all satisfied.
PREDICTIVE_PROPRIOCEPTION_NULL	No registered incremental predictive value is found.
INVALID_ANALYSIS	Leakage, split, feature, implementation, or one-shot test failure prevents interpretation.

Appendix B. Registered Causal Arms

The causal study is gated on predictive closure and uses training-only direction estimation. A positive causal verdict requires directional ordering, matched-control specificity, protected-suite non-regression, and mechanism localization. A merely behavior-changing perturbation is insufficient.

Verdict	Definition
CAUSAL_PROPRIOCEPTION_CLOSED	Directional candidate-local intervention specifically changes success under matched controls.
CAUSAL_EFFECT_PRESENT_BUT_NONSPECIFIC	Behavior changes, but placebos, token effects, or off-target controls prevent a proprioceptive interpretation.
CAUSAL_PROPRIOCEPTION_NULL	No registered causal effect is found.
INVALID_CAUSAL_STUDY	Implementation, leakage, evaluation, or power failure prevents interpretation.

Appendix C. Current Evidence and Retraction Register

Item	Disposition
Row-only atomic-resume pass	Retracted after array-level audit found 375 lost activations.
First Hermes interruption-depth pass	Retracted after timestamps showed a shell wrapper was killed while the Python child continued.
opus5_capture.npz as activation evidence	Retracted; file was a self-declared random simulation placeholder.
SRX-3 24/24 Holm significance	Superseded by 10,000-permutation Westfall-Young max-T: 5/24 cells.
Pooled SRX-3 z near 90	Retracted under tighter within-root null.
Original SRX-4 horizon axis	Invalid; repaired true-horizon pilot subsequently passed.

Appendix D. Promotion Criteria for Shadow Control

- Functional origin reconstruction is reliable.
- Static reads are stable.
- Repeated candidate-local increments are reproducible.

- Candidate identity exceeds order and no-op effects.
- Held-out external prediction improves beyond the strongest controls.
- Matched random, orthogonal, off-layer, off-token, and token-matched controls remain appropriately weak.
- Causal directionality reproduces.
- Protected capabilities do not regress.
- Source, model, data, evaluator, and promotion rules remain bound to receipts.
- No failed branch is reinterpreted as success through outcome-informed tuning.

Appendix E. Central Thesis

Central thesis

Transformer computation can be embedded in a proprioceptive architecture that reconstructs a common functional origin, measures persistent and candidate-local hidden-state dynamics, predicts external success beyond ordinary representations, causally regulates cognitive trajectories, and supports governed autonomous improvement. If these gates close, the bare transformer becomes a substrate rather than a complete cognitive architecture.