THE PROPRIOCEPTIVE TRANSFORMER

Functional Counterfactual Reconstruction, Adaptive Calibration, Dose-Calibrated Fixed-Prefix Control, and Governed Cognitive Improvement

Research Candidate v2.5 - Operator Mechanisms and Causal Pipelines Edition

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

Evidence posture

This v2.5 edition records a completed measurement substrate, the exact-prefix theorem correction, a source-complete fixed-prefix causal architecture, an externally supervised experiment window, and a detailed clean-outcome and dose-calibration pipeline. It does not claim that causal steering has succeeded. Operator Proof V2 has passed 59/59 mechanics tests and the supervisor has passed an unattended recovery smoke; clean-label mechanism fitting, dose response, eleven-arm specificity, protected-suite validation, and clean-restart causal replication remain open.

Keywords: transformer interpretability; proprioception; fixed-prefix intervention; hidden-state steering; candidate-local state; CYGNUS; causal control; 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 candidate trajectories, measures their evolution across depth and time, and connects those measurements to selection, intervention, external verification, memory, and governed improvement. The completed PTB-1 V2 substrate contains 800 unique roots, 6,400 registered Hermes-3-Llama-3.1-8B trajectories, and 147,100 activation arrays captured under scientific manifests, checkpoint-derived stop semantics, executable or programmatic verifiers, and atomic per-root persistence. The campaign also established an exact-prefix reconstructibility correction: with a fixed deterministic model and identical token prefix, the on-manifold hidden state is fixed. Proprioceptive value therefore cannot rest on Shannon information irreducible to an exact prefix under unlimited reconstruction. The decisive claims are operational and causal: internally available state may select cognitive procedures under matched resources, and intervention after a byte-identical prefix may directionally alter future external success. CYGNUS Operator Proof V2 prospectively implements this test using the exact unquantized bfloat16 Hermes checkpoint, manual autoregressive decoding, independent state reconstruction per arm, three competing direction mechanisms, eleven intervention arms, validation-only mechanism/layer/horizon/dose selection, a protected suite, and clean-restart receipts. The source-and-mechanics program passes 59 of 59 tests. An external CPU-only supervisor has also passed an unattended SIGKILL recovery smoke, decoupling production-service restoration from the CUDA-holding pilot. Development calibration revealed that an apparent arithmetic capability cliff was magnified by comma-blind parsing, a short token cap, and textual blank-line stopping. The original cliff magnitude is superseded; a corrected steep transition remains provisional because lower-difficulty points are still truncation-confounded. A measured 3.1669 success/failure mean-state separation is retained as real geometry with confounded semantics, and an eighteen-cell zero-effect steering grid is retained as evidence that the original RMS-scaled intervention regime was below the behavioral dose threshold rather than as a null mechanism result. The current architecture therefore adds adaptive difficulty calibration, an explicit clean-outcome contract, dose-calibrated fixed-prefix intervention, external experiment supervision, and an Ivakhnenko-style mechanism competition. No valid registered Hermes causal arm result has yet been produced. A successful directional pilot, matched-control specificity, clean-restart replication, and three governed CYGNUS improvement cycles remain the open route to causal and autonomous proprioceptive closure.

Significance statement

The central advance is an executable causal proof architecture. Exact-prefix prediction is mathematically constrained by reconstructibility; fixed-prefix intervention is not. If a train-derived internal direction changes future success while prompt and prefix remain byte-identical, and if random, orthogonal, off-layer, off-token, and token-matched controls remain weaker, then hidden state becomes a specific cognitive control variable rather than a passive representation. The architecture then supplies CYGNUS with a measured channel for procedure selection, intervention, verification, and governed retention.

Evidence legend

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

v2.5 manuscript status

Order 0 is closed. The full PTB-1 bank and corrected label overlay are preserved. The original exact-prefix predictive framing is superseded as the primary architecture proof. Operator Proof V2 source and mechanics are complete; the external V4 supervisor has passed; the original capability-cliff magnitude is superseded; the corrected competence transition is provisional; the old steering grid establishes an insufficient-dose floor. No valid registered Hermes causal mechanism verdict has yet been issued.

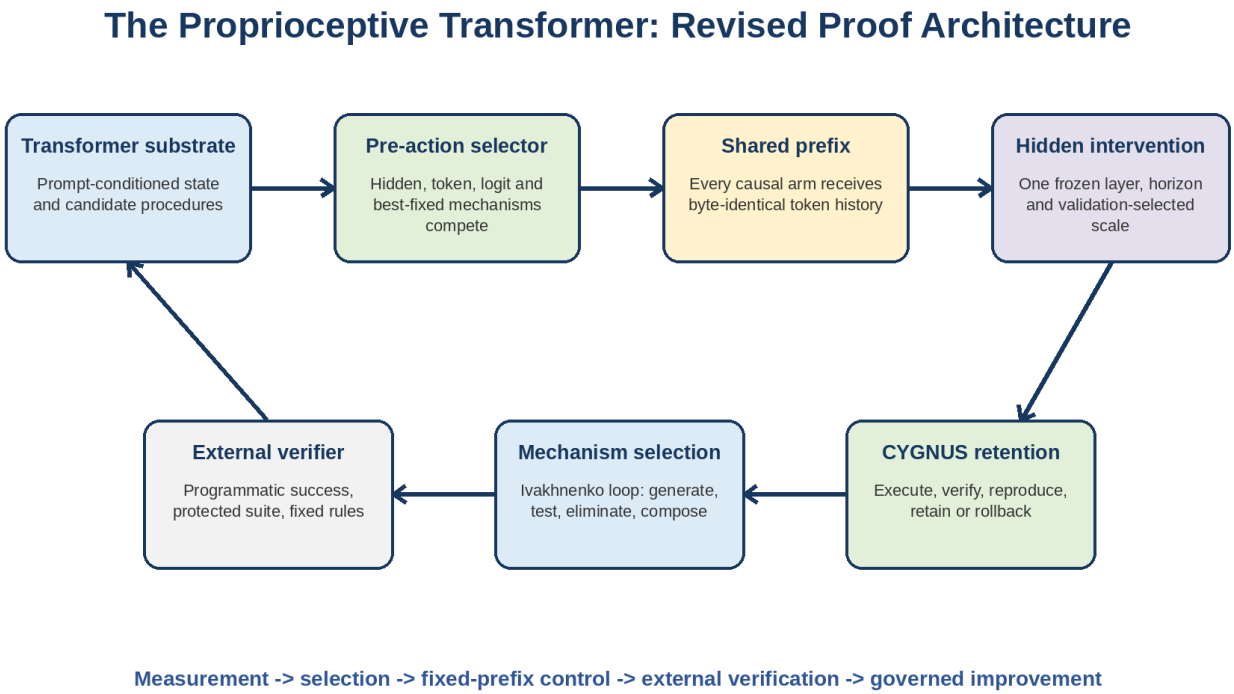

Figure 1. Revised proof architecture: matched internal action selection, fixed-prefix causal control, external verification, and governed retention.

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 encode uncertainty, semantic relations, latent plans, confidence, or error, but ordinary inference supplies no validated closed loop through which the system measures its developing trajectory, compares alternative actions, intervenes on the relevant geometry, and retains only externally verified improvements.

The Proprioceptive Transformer begins from a cybernetic premise: intelligence is not only representation and generation. A complete cognitive architecture also requires internal sensing, state estimation, counterfactual comparison, control, memory, and governed adaptation. Biological proprioception is not a verbal self-theory. It is a control-relevant measurement channel. The artificial analogue proposed here is functional rather than anthropomorphic.

Publication candidate v2.2 organized the program around a four-order closure ladder: complete measurement, predictive proprioception, causal proprioception, and CYGNUS autonomous improvement. The measurement order has now closed. The subsequent analysis produced a deeper correction. When a deterministic transformer is conditioned on an exact token prefix, its on-manifold hidden state is determined by that prefix and the model. A primary claim of information-theoretic independence from the exact prefix is therefore the wrong target.

The revised architecture asks two operational questions. First, under matched training resources, estimator capacity, and inference cost, does an already-available internal state select a cognitive procedure better than token, logit,

confidence, random, and best-fixed alternatives? Second, while the preceding token history remains byte-identical, does changing hidden state alter future external success specifically and directionally? The second question is the decisive causal proof because it changes the system while holding its visible computational history fixed.

Successor-architecture hypothesis

If internally available state improves action selection under matched resources, fixed-prefix intervention specifically changes future success, and CYGNUS uses that channel for reproducible externally governed improvement, then the bare transformer is insufficient as a complete cognitive architecture. Transformer computation remains the engine; proprioception supplies sensing, selection, steering, verification, memory, and governed adaptation.

1.1 Principal contributions of v2.5

1. A completed PTB-1 V2 measurement substrate containing 800 unique roots, 6,400 trajectories, and 147,100 activation arrays.
2. An exact-prefix reconstructibility lemma, empirically confirmed by bit-identical hidden states for identical prefixes.
3. A revised operational thesis: internal state is valuable when timely, low-cost, selectable, and causally actionable, not because it is information-theoretically independent of exact tokens.
4. CYGNUS Operator Proof V2: an exact-Hermes, unquantized, manual-decoding, eleven-arm fixed-prefix causal architecture.
5. An Ivakhnenko-style competition among mean, supervised residualized, dark-subspace, and optional action-transition mechanisms.
6. A 59-test mechanics program with planted negative fixtures, arm-distinctness checks, split integrity, protected-suite interfaces, and receipt binding.
7. An external CPU-only supervisor that owns V4 shutdown, pilot execution, GPU drain, V4 restoration, verification, and lock release.
8. An adaptive difficulty-calibration pipeline that eliminates ceiling, floor, verifier-error, truncation-dominated, and malformed regimes before direction fitting.
9. A clean outcome contract with semantic token stopping, explicit final-answer parsing, failure taxonomy, and qualification thresholds.
10. A dose-calibrated intervention program using a validation-only beta ladder in hidden-state norm units and permanent immediate-logit validity checks.
11. A cumulative evidence ledger that supersedes inflated magnitudes without erasing measured geometry, dose-floor evidence, or still-open mechanisms.
12. An updated ascension ladder from measurement to operational selection, fixed-prefix causal control, and three-cycle CYGNUS governed improvement.

2. Evidence Discipline and Current Ledger

The theory is broader than the currently closed evidence. The evidentiary system must operate in both directions: ambition cannot promote a claim, but caution cannot erase a measured result. This edition distinguishes measurement closure, conceptual correction, source-and-mechanics completion, development evidence, and still-open causal performance.

Claim	Status	Current interpretation
Functional counterfactual reconstruction under Contract A	[D]	Registered model-derived fields reproduced across repeated and clean-process executions.
Complete PTB-1 V2 bank	[D]	800 roots, 6,400 trajectories, 147,100 arrays; structural integrity and source identities frozen.
Exact-prefix reconstructibility lemma	[D/S]	Deterministic from the model contract and empirically supported by bit-identical same-prefix states.

Claim	Status	Current interpretation
First-token and depth-localized behavioral geometry	[S]	Localized SRX-3 effects survive strengthened controls; not universal across all cells.
Increment beyond first-token readout on the existing bank	[X/P]	Not closed in development and superseded as the primary architecture proof.
Matched operational action selection	[T]	Specified but not yet executed on real registered Hermes data.
Operator Proof V2 source and mechanics	[D]	Prospective implementation complete; 59/59 mechanics tests pass.
Fixed-prefix causal controllability	[T]	Registered Hermes development pilot and sealed campaign remain pending.
CYGNUS autonomous verified improvement	[T]	Requires validated selector and causal channel plus three governed clean-restart cycles.
Bare transformer obsolete as a complete architecture	[T]	Conditional implication if operational selection, causal specificity, and autonomous closure pass.

2.1 Supersessions since v2.2

- The manuscript-freeze status is superseded: all four PTB-1 task classes completed and the full bank closed.
- The original external labels were invalid because V2 verifier names fell through a legacy dispatch and silently returned zero. The immutable bank was preserved and a 6,400-row corrected label overlay was produced from frozen outputs and verifier semantics.
- The pooled strongest-control analysis was dominated by task-class base rates and did not test within-root candidate selection.
- The first root-conditioned t1 analysis did not close beyond first-token identity and is retained as a readout-redundancy result, not as a refutation of operational proprioception.
- The exact-prefix lemma supersedes any claim that a deterministic hidden state should add Shannon information beyond the exact token prefix under unlimited reconstruction.
- Fixed-prefix causal intervention is elevated from a later confirmatory study to the decisive architecture proof.

3. Exact-Prefix Reconstructibility

3.1 Lemma

For a deterministic transformer in evaluation mode, with fixed parameters, attention mask, generation position, cache-construction procedure, and exact token prefix, the on-manifold hidden state is deterministic.

$$h_t = f_\theta(\text{prompt}, x_{1:t}, \text{mask}, \text{position}, \text{cache contract})$$

Consequently, a token-based system with the exact model, exact prefix, and unlimited reconstruction compute can recover the same hidden state. An empirical test demanding information-theoretic independence from that complete prefix is structurally degenerate.

3.2 Empirical confirmation

The completed PTB-1 bank confirmed the lemma at machine precision. Across 4,020 same-root candidate pairs sharing the same first generated token, t1 hidden states were bit-identical with relative L2 distance 0.000e+00. Across 2,649 candidate pairs sharing the same four-token prefix, t4 hidden states were likewise bit-identical. States diverged only when token sequences diverged.

> **Consequence**
>
> The architecture should not claim that hidden state contains Shannon information unavailable from the exact prefix. It should claim that hidden state is already computed, internally accessible, potentially lower-latency or lower-cost than reconstruction, and directly intervenable while the preceding prefix remains fixed.

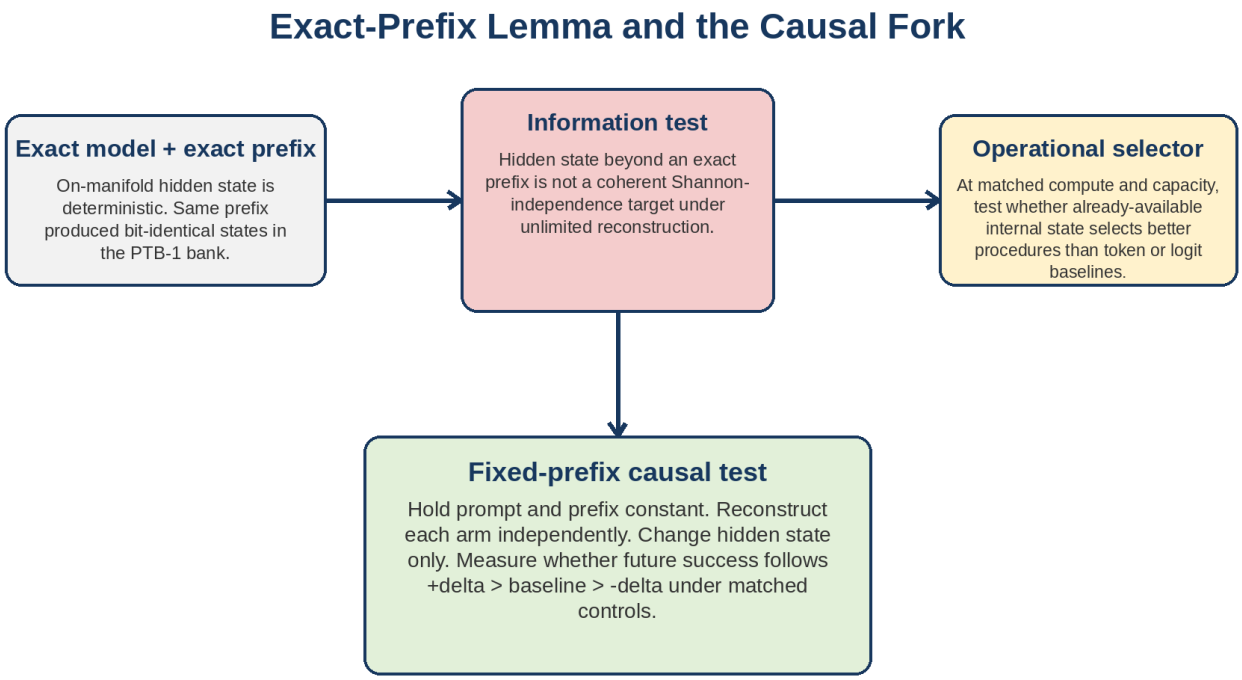

Figure 2. The exact-prefix identity invalidates one predictive framing but exposes the decisive causal fork.

3.3 Two nondegenerate proof targets

1. Matched operational selection: at equal training roots, estimator family, feature dimension, regularization grid, and explicitly measured inference cost, does already-available internal state choose better cognitive procedures than token or logit baselines?
2. Fixed-prefix causal control: after reconstructing the same prompt and exact prefix independently for every arm, does a hidden-state intervention directionally change future external success under matched geometric, temporal, and token-level controls?

4. Spatial-Temporal State Architecture

$$h_{l,t} = a_{l,t} + d_{l,t}$$

$$d_{l,t}^{(c)} = m_{l,t} + delta_{l,t}^{(c)}$$

The spatial decomposition separates active, high-variance representational directions from lower-variance dark geometry that may remain dynamically or causally important. The temporal decomposition separates persistent orientation from the increment produced by the present candidate trajectory. These are different update and persistence rules, not merely symbolic partitions.

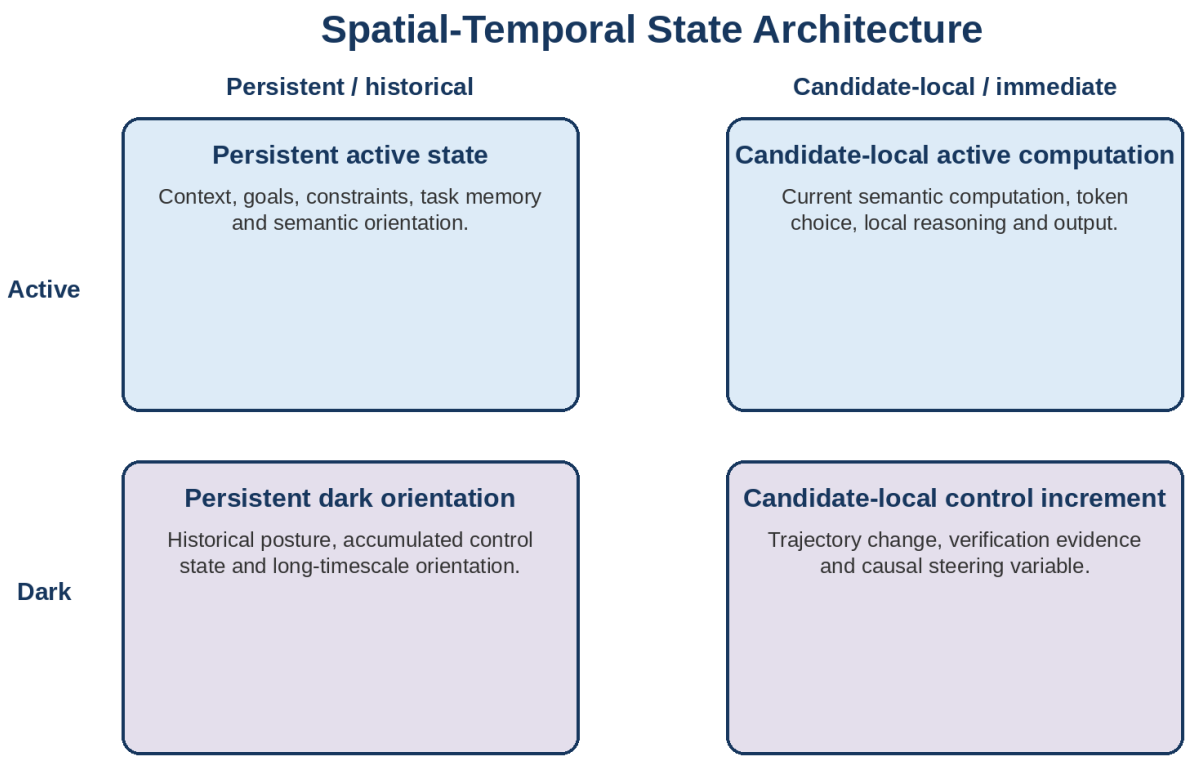

Figure 3. Four-cell spatial-temporal architecture.

4.1 Proprioceptive state bundle

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 and candidate causal variable.
kappa_(l,t)	Local curvature and control geometry.
v_(l,t)	Externally calibrated verification state.

The bundle need not exist as one monolithic vector in the base transformer. It is an engineered system state assembled from activations, persistent controller memory, candidate differences, geometry, intervention metadata, and external calibration. Proprioception is therefore a property of the measurement-control architecture, not a claim that a pretrained transformer spontaneously contains a neatly labeled self-variable.

5. Counterfactual State Contracts

5.1 Contract A: functional state reconstruction

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, hook state, and cache-construction procedure define a reproducible functional origin. Under the registered Hermes gate, model-derived fields reproduced across 100 in-process trials and 20 clean-process executions.

Contract A result

A reproducibly reconstructed functional counterfactual origin has been demonstrated. This supports controlled candidate comparison and independent per-arm reconstruction.

5.2 Contract B: serialized runtime-state restoration

Contract B remains stronger and prospective. It would serialize and restore a live KV cache, selected hidden tensors, generation position, attention-mask state, routing state, controller state, and RNG state. Operator Proof V2 does not require Contract B because each intervention arm reconstructs its own common prefix state under Contract A.

6. PTB-1 V2: Completed Measurement Substrate

PTB-1 V2 now provides the complete multi-task measurement substrate that v2.2 described as active. The bank was captured using the exact registered Hermes-3-Llama-3.1-8B model, bfloat16 precision, no quantization, five relative depths, five token horizons, eight candidates per root, scientific manifests, and artifact-derived semantic terminators.

Property	Closed result
Task roots	800 unique roots across reasoning, factual verification, coding, and instruction following.
Candidate trajectories	6,400.
Activation arrays	147,100 real arrays.
Model	Hermes-3-Llama-3.1-8B, exact registered revision, bfloat16, unquantized.
Structural integrity	0 duplicate candidate keys, 0 row-only roots, 0 array-only roots, 0 orphan arrays, 0 post-terminator continuations.
Splits	Train 3,728 rows; validation 1,368; test 1,304.
Bank receipt	Frozen complete-bank receipt and deterministic archive.
Original bank	Immutable and preserved.
Corrected labels	Separate immutable overlay after verifier-dispatch repair.

6.1 Trajectory binding and stop semantics

Every candidate is one continuous cached trajectory. Captured activations and the externally scored completion belong to the same sampled sequence. The final V2.1.2 source derives a five-token semantic terminator union from checkpoint artifacts and stops at the first emitted member. In registered confirmation, post-terminator continuation fell from 71 of 72 audited records to zero and median semantic response length fell from 512 to 45 tokens.

6.2 Atomic persistence

Each prompt root is an atomic scientific transaction containing all eight rows and all reachable arrays. Temporary artifacts are flushed, fsynced, and atomically renamed before the ledger admits the root. An authenticated process-group SIGKILL after two durable roots followed by array-aware resume produced exact canonical equivalence with an uninterrupted registered Hermes run.

6.3 Label repair without recapture

The capture verifier implemented only legacy V1 names while scientific manifests used V2 names. Every candidate therefore fell through to a silent zero label. The defect affected the outcome column, not the trajectories, tensors, manifests, or token sequences. A prospectively frozen relabeler executed the original answer keys, sealed labels, unit tests, and programmatic constraints against preserved pre-terminator outputs. It produced a 6,400-row overlay with zero verifier errors while leaving the original bank immutable.

7. What the Original Predictive Program Established - and Did Not Establish

The predictive program generated valuable constraints but did not close proprioception. It should not be narrated as a poor architecture result.

7.1 Global control saturation

Corrected development base rates ranged from near-ceiling reasoning and factual verification to near-floor instruction following. Task identity and prompt difficulty therefore produced validation AUROC near 0.994. The pooled endpoint mostly separated task classes rather than candidates from one common root.

7.2 Root-conditioned t1 development result

A prospectively frozen root-conditioned analysis compared candidate controls with controls plus t1 delta. The primary control already achieved pairwise accuracy 0.9719 and selection accuracy 0.9455. Adding delta increased pairwise accuracy to 0.9772 and selection to 0.9636, but worsened conditional log loss; the root-bootstrap information-gain interval crossed zero. Coding was negative, instruction following directionally positive but non-closing, and factual verification underpowered. The registered development verdict was not closed.

7.3 Why this was not the decisive architecture test

At t1, the first token is the discrete readout of the same post-token state being measured. Conditioning on the first token removes nearly all legitimate t1 variation; omitting it leaves an empty within-root baseline because prompt, task, pre-sampling logits, entropy, and margin are root-constant. The exact-prefix lemma generalizes this problem to later horizons. This is not a reason to ignore the development result. It is a reason to interpret it as a readout-redundancy stress test rather than as the final proof of an internally useful control architecture.

7.4 Zero-metric sealed abort

One sealed evaluation marker was written before a feature-schema assertion failed. No raw predictions or test metrics were produced. The event is preserved as a zero-metric mechanical abort, not as a predictive verdict. The current paper does not reopen or reinterpret that test as positive evidence.

> **Order 1 disposition**
>
> The original exact-prefix incremental-prediction order is superseded as the primary proof. Its development results and failures remain part of the evidence record. The revised Order 1 tests matched operational action selection.

8. Revised Order 1: Matched Operational Action Selection

The operational question is whether an already-available internal state provides a useful selector under constrained resources. A token-based system could reconstruct hidden state by running the model, but that computation is not free. The fair comparison therefore matches training roots, estimator family, output feature dimension, parameter budget, regularization grid, candidate procedures, and external verifier, while separately reporting additional forward calls, latency, and compute.

8.1 Candidate procedures

Action	Procedure
A0	Direct answer.
A1	Explicit derivation plus verification.
A2	Adversarial critique followed by revision.
A3	Independent branches followed by aggregation.

8.2 Selector competition

Selector	Information and fairness contract
Hidden-state selector	Pre-action prompt-conditioned state; train-only residualization and dimensionality reduction.

Selector	Information and fairness contract
Token selector	Prompt-token hashed or fixed token features; same estimator family and output dimension.
Logit/confidence selector	Prompt-boundary logits or frozen projection plus entropy, margin, and confidence.
Best fixed	Procedure selected on validation, then frozen for test.
Always A0	Direct-answer baseline.
Random	Uniform procedure selection.

8.3 Operational closure criteria

- Hidden selector exceeds validation-selected best fixed external reward on sealed test.
- Hidden selector exceeds matched token and logit/confidence selectors.
- Paired root-bootstrap intervals exclude no improvement for required comparisons.
- Selection occurs before external execution.
- Inference latency and extra model-forward cost are reported.
- The result reproduces from a clean restart.

9. Revised Order 2: Fixed-Prefix Causal Proprioception

Prediction does not establish control. The decisive experiment fixes prompt and token prefix, reconstructs the same state independently for each arm, changes hidden state at one explicit layer and token horizon, and measures future external success. No arm inherits a cache mutated by another arm.

$$Success(+delta) > Success(baseline) > Success(-delta)$$

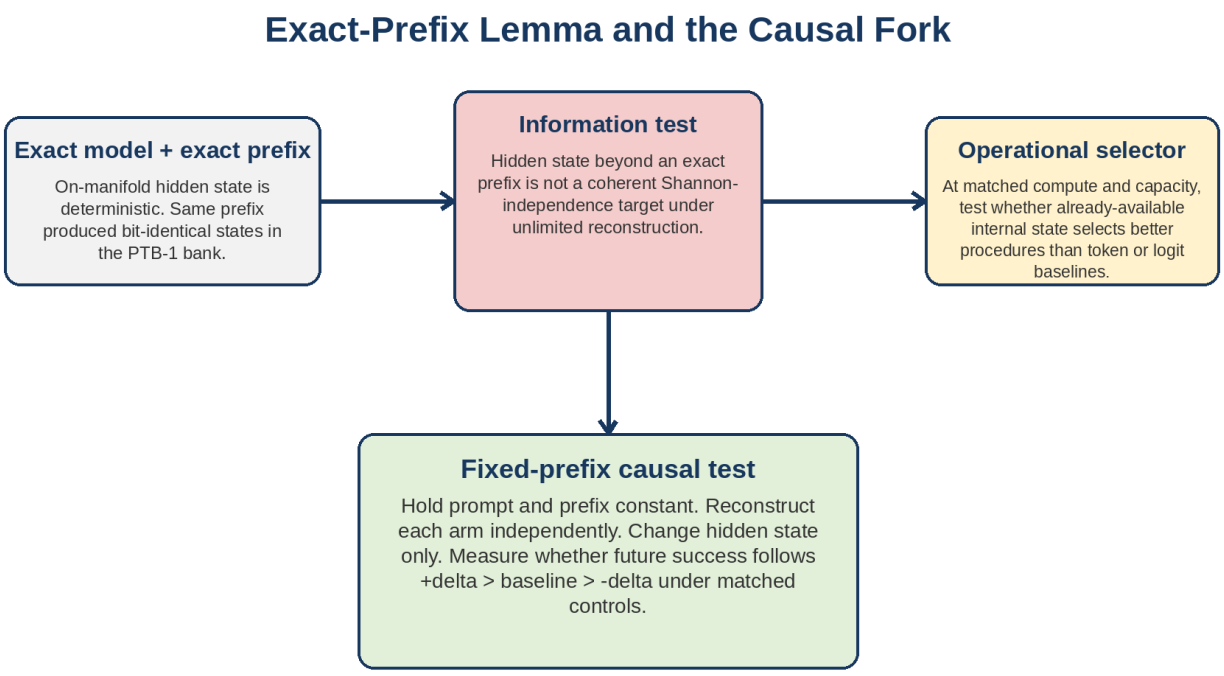

Figure 4. Fixed-prefix intervention changes internal state while visible history remains identical.

9.1 Causal transaction

1. Tokenize the prompt and run prompt prefill with KV cache.

2. Generate and hash one shared prefix.
3. Reconstruct the exact same prompt and prefix state independently for every arm.
4. Verify byte-identical prefix IDs and hashes.
5. Install the intervention hook only for the frozen target layer and token step.
6. Run exactly one intervention-bearing forward pass and remove the hook immediately.
7. Continue generation under the same decoding and stop semantics.
8. Score externally and bind all arm metadata to source, model, direction, layer, horizon, and prefix hashes.

9.2 Registered arms

Arm	Purpose
Baseline	Fresh reconstruction with no hidden-state modification.
Amplify	Add +alpha times the normalized success direction.
Negative	Add -alpha times the direction.
Suppress	Reduce the existing projection without reversing it.
Reverse	Reflect the directional component through zero.
Delete	Remove the registered subspace projection.
Matched-norm random	Control perturbation magnitude.
Orthogonal placebo	Geometrically unrelated direction.
Off-layer	Test layer specificity.
Off-token	Test temporal specificity.
Token-matched	Match immediate logit effect while remaining geometrically distinct.

9.3 Causal specificity

A positive arm is insufficient if generic perturbations, token-matched controls, or off-target interventions behave similarly. Causal closure requires directional ordering, paired intervals, matched-control attenuation, protected-suite non-regression, and clean-restart replication.

10. CYGNUS Operator Proof V2

The historical operator artifact existed only on the local lil workstation and had never run on the big execution host. It was preserved as a read-only historical seed and prospectively superseded. The new workspace isolates baseline, source, specifications, tests, runs, artifacts, and receipts. It does not modify live OPUS5, the completed PTB-1 bank, or V4 production source.

10.1 Source architecture

Module	Role
manual_decode.py	Registered model contract, manual token-by-token decoding, cache reconstruction, explicit hook site.
interventions.py	Eleven algebraically and geometrically distinct arms.
directions.py	M1-M4 direction construction and train-only subspace fitting.
selection.py	Validation-only mechanism, layer, horizon, amplitude, and rank selection.
verifiers.py	Programmatic development and protected-suite scoring.
receipts.py	Source, model, environment, result, and clean-restart bindings.

Module	Role
tasks.py	Development tasks requiring non-answer-bearing prefixes and post-intervention continuation.

10.2 Competing mechanisms

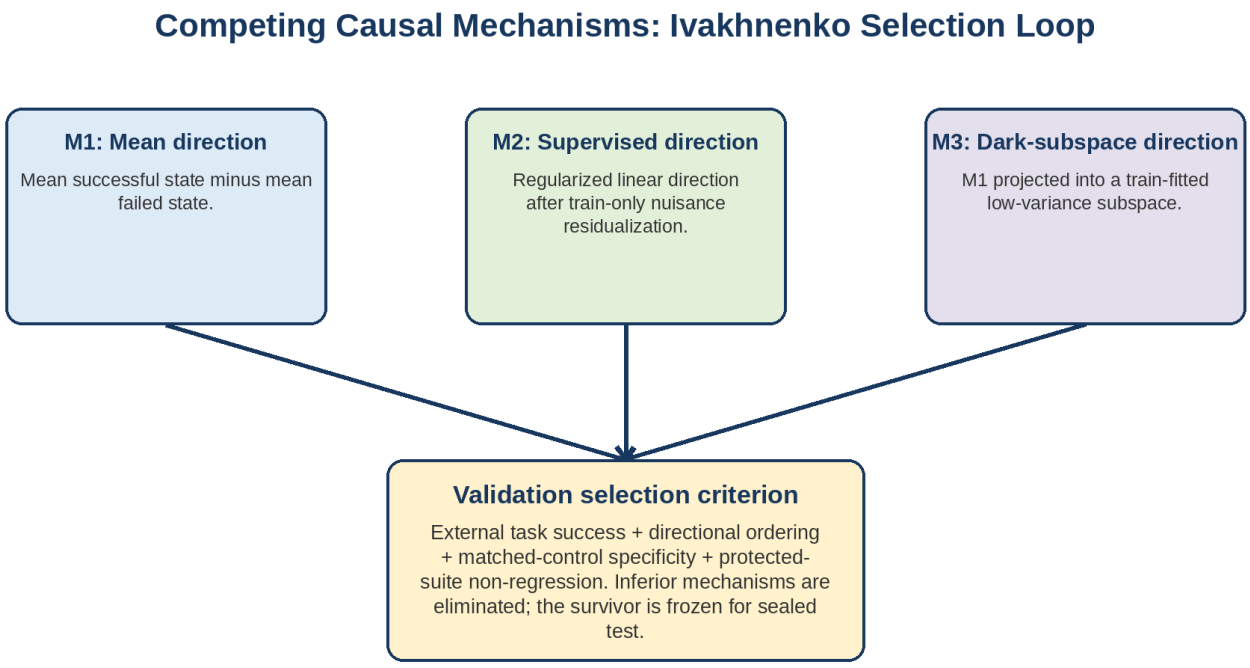

Figure 5. Competing direction constructions are selected externally rather than crowned by theory.

Mechanism	Construction	Current status
M1	Mean successful-state minus mean failed-state direction.	Synthetic recovery pass; real validation pending.
M2	Regularized supervised linear direction after nuisance residualization.	Synthetic recovery pass; real validation pending.
M3	M1 projected into a train-fitted low-variance dark subspace.	Distinct synthetic mechanism; real validation pending.
M4	Action-conditioned transition direction.	Optional extension.

10.3 Mechanics program

11. Operator Proof V2 Implementation State

Operator Proof V2 is the prospective causal implementation of the Proprioceptive Transformer. It begins from the sole historical operator artifact, which existed on the lil workstation and had never executed on big. That artifact is preserved read-only as a historical seed rather than treated as prior causal evidence. The new implementation was written prospectively on big, bound to the registered Hermes model contract, and decomposed into independently testable modules.

Module	Primary responsibility	Current status
manual_decode.py	Prompt prefill, manual token-by-token generation, exact prefix capture, one-step hook installation, cache reconstruction, semantic stopping.	Implemented; mechanics tested.
interventions.py	Algebra for baseline, amplify, negative, suppress, reverse, delete, random, orthogonal, off-layer, off-token, and token-matched arms.	Implemented; arm distinctness enforced.
directions.py	M1-M4 construction, nuisance residualization, dark-subspace projection, uncertainty, normalization.	Implemented; synthetic recovery tests passed.
selection.py	Train-only fitting, validation-only mechanism/layer/horizon/dose selection, deterministic tie rules.	Implemented; stdlib-shadow defect corrected.
tasks.py	Programmatically verifiable task generation, graded difficulty regimes, calibration and root namespace separation.	Implemented; clean-outcome revision active.
verifiers.py	Final-answer parsing, outcome taxonomy, protected-suite scoring, explicit verifier errors.	Clean outcome contract under V2.0.4.
receipts.py	Source/model/spec/task/prefix/direction/result/ environment hashing and clean-restart binding.	Implemented.
window_supervisor.py	CPU-only ownership of V4 stop, pilot child, GPU drain, V4 restoration, verification, and lock release.	Validated unattended restore smoke.

Current implementation status

The source-and-mechanics program passes 59 of 59 tests. The external supervisor has passed a real unattended V4 recovery smoke after SIGKILL. No valid registered Hermes intervention arm has yet produced a causal result. The next admissible scientific step is clean outcome calibration, clean M1/M2/M3 refitting, validation-only dose selection, and the eleven-arm development pilot.

12. External Experiment Supervision and Production Recovery

The first pilot windows revealed a systems-level design error: the CUDA-holding experiment process also owned V4 restoration. That process could retain model references, terminate without unwinding finally, or die before freeing GPU memory. Three windows therefore required manual recovery. V2.0.4 moves restoration authority into an external CPU-only supervisor that imports neither torch nor transformers and never owns a CUDA context.

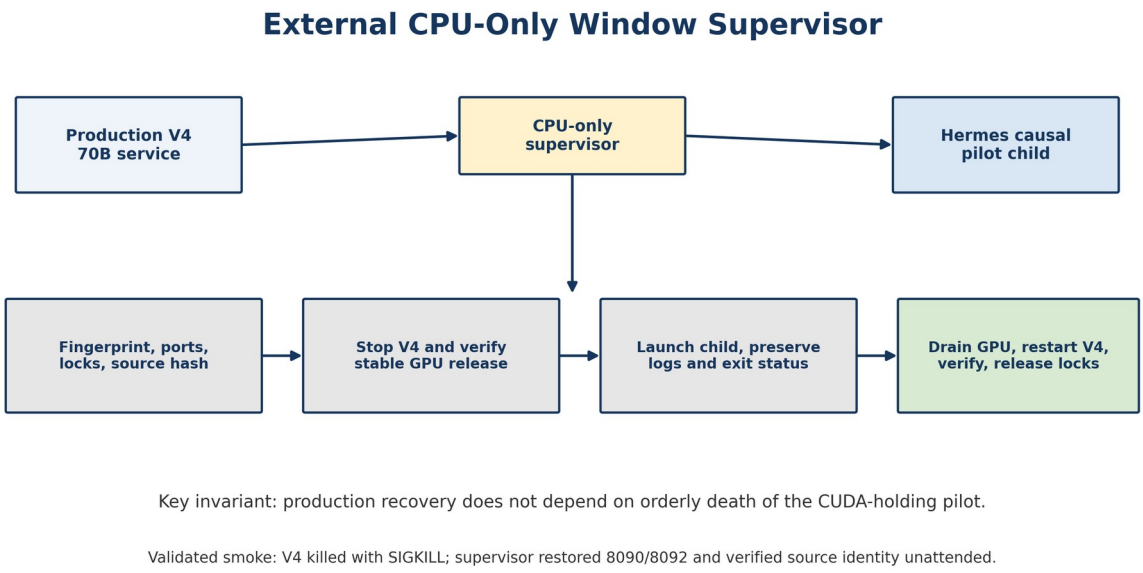

Figure 6. External CPU-only supervisor decouples experimental failure from production-service recovery.

12.1 Supervisor state machine

1. Fingerprint V4, port 8090, the independent 8092 service, source hash, expected model identity, corrector state, and GPU occupancy.
2. Acquire the GPU and window locks.
3. Stop only the registered V4 process.
4. Poll until the registered physical GPU is stably released.
5. Launch the causal pilot as a direct child process with a distinct process group.
6. On normal exit, exception, SIGTERM, SIGINT, or forced kill, preserve logs and continue cleanup.
7. Poll until all pilot GPU processes disappear and memory is stably free.
8. Restart V4 only if port 8090 is absent and the frozen source hash still matches.
9. Verify the service is ready, the 70B model identity is correct, the corrector is loaded but disabled, 8092 is untouched, and GPU occupancy returns to baseline.
10. Release locks only after post-restore verification.

12.2 Demonstrated recovery result

A loader-abort smoke deliberately removed V4 without relying on the pilot's finally block. The external supervisor detected port-down state, waited through the grace period, verified GPU release, restarted V4, and confirmed service readiness and source identity without human intervention. The complete recovery took approximately 45 seconds. This is a demonstrated operational result, not a causal-science result.

Property	Validated observation
Supervisor process	CPU-only; no torch, CUDA, or transformer import.
Failure mode exercised	V4 killed with SIGKILL; no orderly pilot cleanup.
Authority	Port readiness and source/model verification, not pgrep alone.
Recovery	8090 and 8092 restored/verified unattended.
Production invariant	Experimental child failure cannot strand V4 by retaining CUDA state in the supervisor.

13. Clean Outcome Contract

A causal direction is meaningful only if positive and negative labels represent cognitive success and failure rather than parser quirks, formatting conventions, or token-budget exhaustion. The first calibration run violated this requirement: comma-formatted correct answers failed, a 48-token cap truncated long derivations, and textual blank-line stopping could terminate before the final answer. These defects scaled with answer magnitude and inflated the original capability-cliff estimate.

13.1 Semantic stopping

- Stop only on artifact-declared semantic terminator token IDs.
- Do not use decoded-text blank lines as a stopping rule.
- Record terminator identity and position.
- Exclude the terminator from verifier text.
- Classify max-token exhaustion separately from semantic completion.

13.2 Final-answer parsing

The parser must recognize registered formatting-equivalent correct answers without becoming permissive. Accepted forms include bare integers, comma-formatted integers, equals-prefixed integers, and explicit final-answer phrases. The parser rejects wrong values, near misses, superstrings, malformed outputs, and multiple conflicting final answers.

Candidate output	Disposition	Reason
64128	Accept	Bare integer equals the frozen answer.
64,128	Accept	Registered comma formatting.
= 64,128.	Accept	Registered punctuation and equals prefix.
The final answer is 64,128.	Accept	Explicit final-answer phrase.
641280	Reject	Superstring / wrong integer.
64,127	Reject	Near miss.
64,128 ... final answer 64,129	Reject	Conflicting final answers.
No parseable final answer	Reject / classify	NO_FINAL_ANSWER or MALFORMED_OUTPUT.

13.3 Outcome taxonomy

Outcome class	Definition	Use in fitting
CORRECT	Parsed final answer matches the frozen verifier.	Positive outcome.
WRONG_ANSWER	A valid final answer is present but incorrect.	Primary negative outcome.
MAX_TOKEN_TRUNCATION	Generation reaches the token cap before a usable final answer.	Excluded from clean correctness fitting; tracked separately.
VERIFIER_ERROR	Verifier raises, times out, or cannot execute.	Hard data-quality failure.
MALFORMED_OUTPUT	Output violates the registered answer contract.	Negative only if prospectively defined; otherwise separate.
NO_FINAL_ANSWER	No parseable final answer after semantic completion.	Tracked separately; capped by qualification rule.

Clean-label invariant

A difficulty regime may enter M1/M2/M3 fitting only when outcome variation is clean enough to identify cognition rather than truncation or formatting: 30-70% success, minimum positive and negative support, at least 25% mixed-procedure roots, zero verifier errors, no more than 20% truncation-only negatives, and no more than 20% malformed/no-final-answer negatives.

14. Adaptive Difficulty Calibration

Difficulty is treated as an experimentally selected system parameter rather than an intuition. The calibration stage evaluates predeclared regimes on roots that are disjoint from training, validation, mechanics, protected, and sealed-test partitions. It selects the first easy-to-hard regime satisfying the clean-label qualification rule. Ceiling, floor, dirty-label, and insufficiently mixed regimes are eliminated before any direction is fitted.

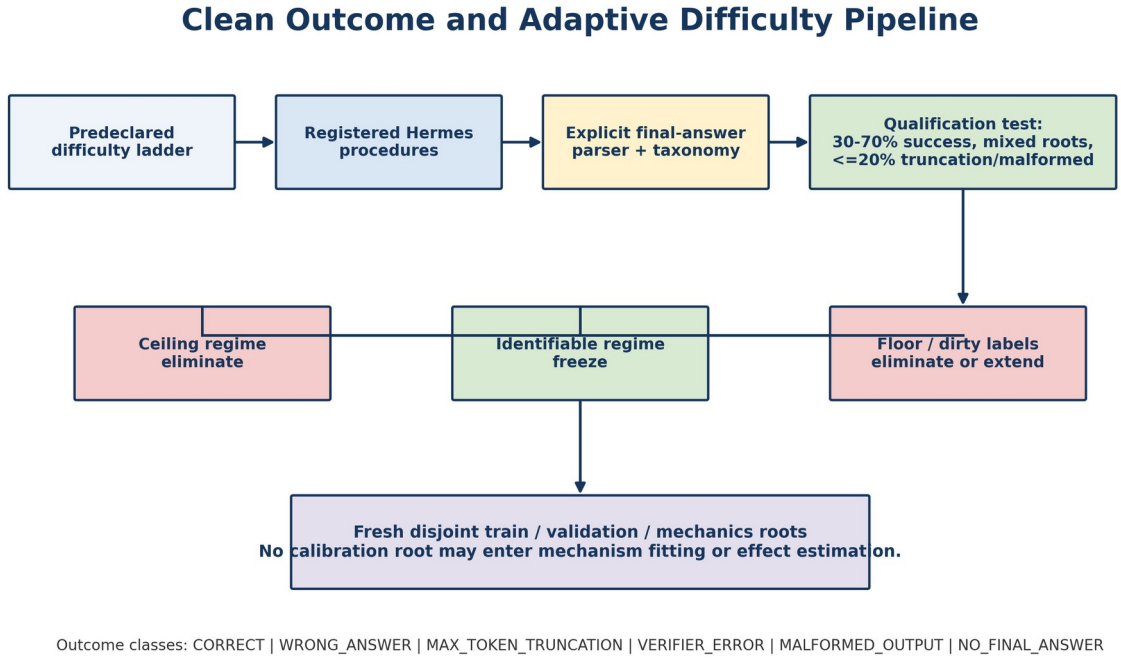

Figure 7. Adaptive calibration eliminates ceiling, floor, and dirty-label regimes before mechanism fitting.

14.1 Superseded and corrected calibration curve

The original observed curve included a drop from 0.969 to 0.031. That magnitude is superseded because comma-blind parsing, short-generation truncation, and textual stopping were confounded with task difficulty. A corrected development pass produced 1.000, 0.969, 0.469, and 0.188 across four regimes. A steep nonlinear transition therefore remains provisionally visible, but the two lower points remain heavily truncation-bound and are not yet clean competence estimates.

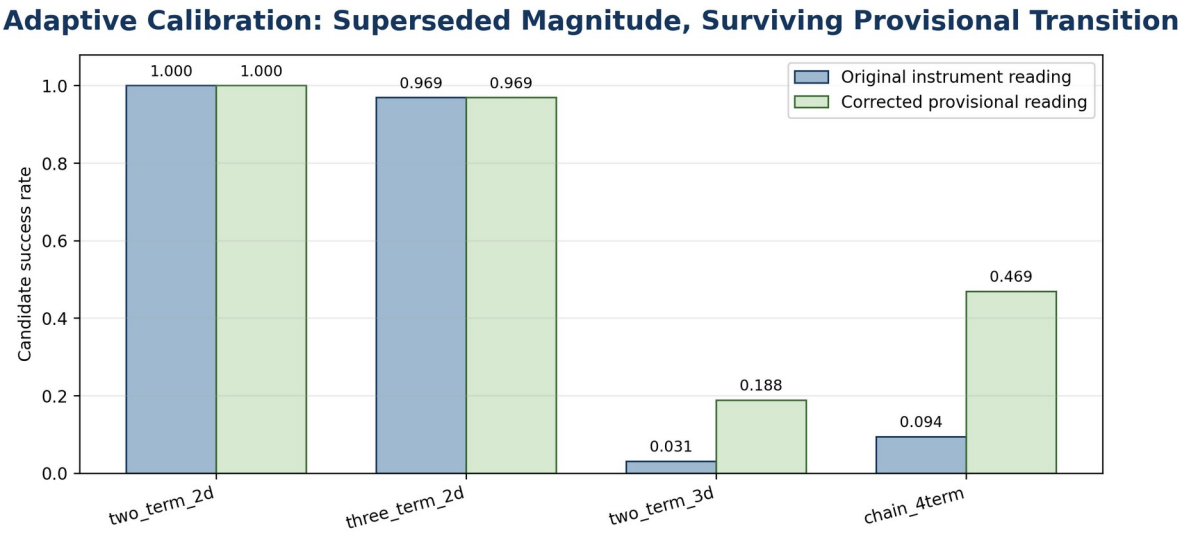

Figure 8. Original and corrected calibration readings. The original magnitude is superseded; the corrected transition remains provisional.

Evidence item	Disposition	Interpretation
Original 0.969 -> 0.031 magnitude	Superseded	Inflated by parser, token-cap, and textual-stop confounds.
Corrected 1.000 -> 0.969 -> 0.469 -> 0.188 curve	Provisional	Steep decline remains, but lower points await clean-outcome remeasurement.
Capability phase boundary	Open	Requires dense, clean calibration and uncertainty/change-point analysis.
Model resistance or defensive agency	Unsupported	No experiment bears on strategic resistance to inspection.

14.2 Calibration roots and namespace isolation

Calibration roots use an independent namespace and are hash-disjoint from pilot roots. The corrected implementation demonstrated 160 calibration roots and 280 pilot roots with intersection zero. Calibration outcomes select task difficulty only; they may not enter direction fitting, mechanism selection, effect estimation, or causal evaluation.

15. Competing Internal Control Mechanisms

Operator Proof V2 does not crown one direction by theoretical preference. It constructs multiple candidate mechanisms, measures them externally, eliminates mechanisms that behave like generic perturbations, and composes the strongest survivor into the next layer of the architecture. This is the Ivakhnenko-style self-organizing element of the program.

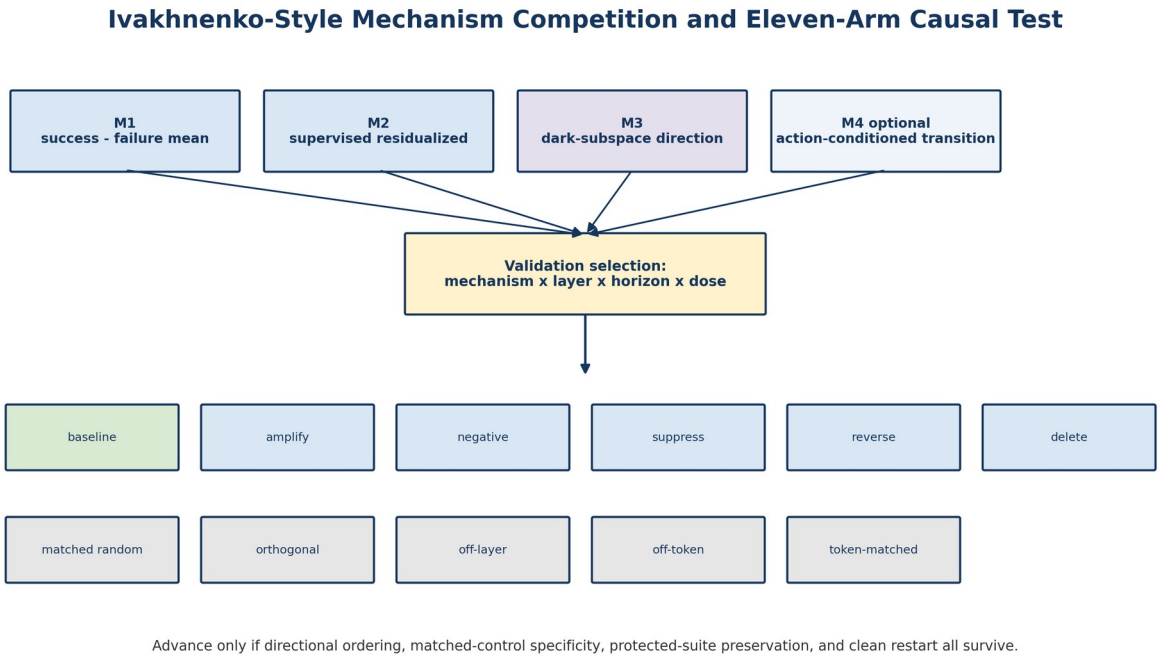

Figure 9. Competing direction mechanisms enter a validation-selected eleven-arm causal experiment.

Mechanism	Construction	Hypothesized advantage	Primary risk
M1 mean direction	Normalized difference between train-only successful and failed mean states.	Simple, interpretable, low variance.	May encode completion length, verbosity, or task nuisance.
M2 supervised residualized	Regularized supervised direction after train-only nuisance removal.	Can isolate weak distributed success geometry.	May overfit small support or absorb readout artifacts.
M3 dark-subspace direction	Project M1 or supervised signal into a train-only low-variance subspace.	Potentially greater specificity with lower immediate token disruption.	Dark subspace may be unstable or noncausal.

M4 action-transition direction	Contrast successful and failed action-conditioned state transitions.	Connects procedure choice to downstream control.	Requires additional action-trajectory collection and may be higher variance.

15.1 Formal constructions

$$M1:\ \hat{d}_1 = \mathrm{normalize}(\mu_{success} - \mu_{failure})$$

$$M2:\ \hat{d}_2 = \mathrm{normalize}(\arg\min_w [L_{train}(w; \text{residualized } h) + \lambda \|w\|_2^2])$$

$$M3:\ \hat{d}_3 = \mathrm{normalize}(P_{dark}\,\hat{d}_1)\ \text{ or }\ \mathrm{normalize}(P_{dark}\,\hat{d}_2)$$

$$M4:\ \hat{d}_4 = \mathrm{normalize}(E[\Delta h \mid \text{successful action}] - E[\Delta h \mid \text{failed action}])$$

The previously observed mean-state separation of 3.1669 is retained as a real geometric measurement with confounded semantics. It may reflect correctness, completion versus truncation, concise versus extended procedure, stable versus malformed generation, or a combination. Clean labels are required to decompose these possibilities. The old directions must be refit before causal interpretation.

15.2 Required geometry report

- Direction hashes, norms, uncertainty, and training support.
- Pairwise cosine similarities among M1, M2, M3, and any M4.
- Cosines with correctness, truncation, completion-length, and malformed-output contrasts.
- Layerwise stability and cross-root bootstrap intervals.
- Active versus dark energy composition.
- Immediate logit projection and token-readout overlap.

16. Dose-Calibrated Fixed-Prefix Control

The original intervention used alpha times the per-element RMS of the hidden state. Across eighteen cells, two mechanisms, three layers, and amplitudes 0.05/0.15/0.40, baseline, amplify, and negative generations were byte-identical. This is retained as an insufficient-dose calibration result: the applied displacement was only about 0.6% to 5% of the hidden-state norm and did not validly test the mechanisms.

$$\text{Old:}\ \Delta h = \alpha \cdot \mathrm{RMS}(h) \cdot \hat{d}$$

$$\text{Revised:}\ \Delta h = \beta \cdot \|h\|_2 \cdot \hat{d}$$

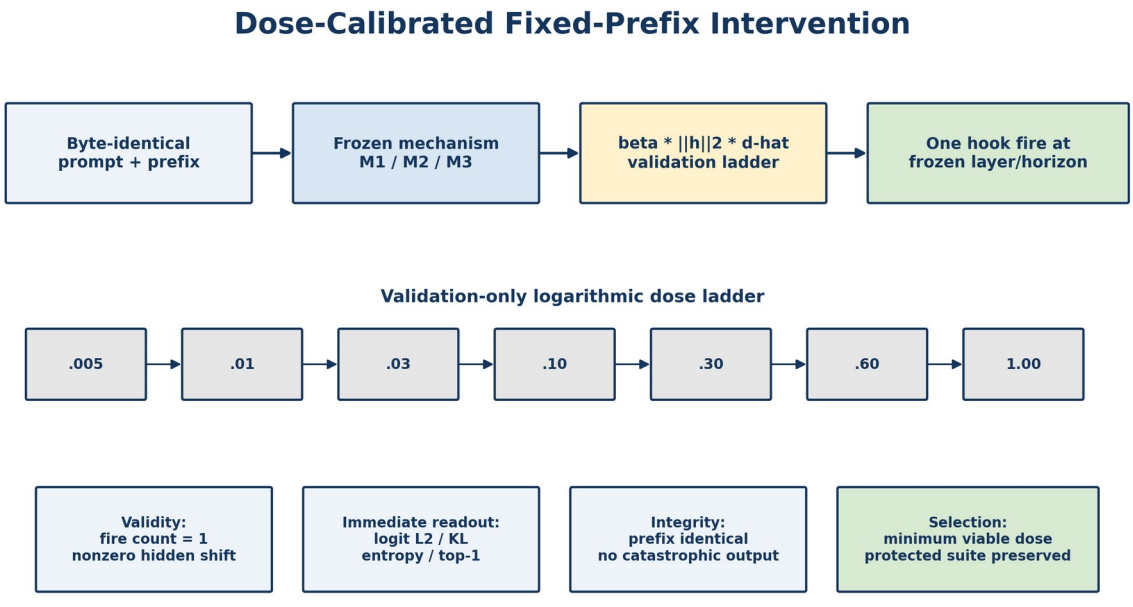

Figure 10. Validation-only dose calibration selects the smallest perturbation that measurably reaches the readout while preserving generation integrity.

16.1 Validation-only dose ladder

The preregistered development ladder is beta in {0.005, 0.01, 0.03, 0.10, 0.30, 0.60, 1.00}. Large values are not presumed better. The selected dose is the minimum viable dose that produces a finite, reproducible immediate logit effect, preserves prefix identity and output integrity, remains within the protected-suite development bound, and permits meaningful comparison among target and control arms.

16.2 Per-cell validity measurements

Measurement	Purpose
||Delta h|| / ||h||	Actual intervention energy in interpretable norm units.
Hook fire count	Must equal one.
Target layer / token step	Must match the frozen receipt.
Immediate logit L2 and KL	Proves the intervention reached the readout and quantifies token disruption.
Entropy and top-1 token change	Separates mild steering from discrete readout flips.
Output hash and length	Detects byte identity, truncation, and broad generation change.
Format and catastrophic-output rate	Rejects destructive doses.
External success and protected-suite delta	Balances target effect with broad capability preservation.

Grid-cell refusal rule

A grid cell is invalid if the hook does not fire exactly once, hidden displacement is zero, immediate logit effect is zero or non-finite, prefixes differ, intervention occurs during prefill, or the target layer/horizon differs from the frozen receipt. An invalid cell cannot support a null or positive mechanism verdict.

16.3 Multistep branch

If every safe single-step dose through beta = 1.0 produces measurable logit changes but no downstream behavioral variation, the program opens a separate prospective MULTISTEP_INTERVENTION_V2_1 branch. Single-step, two-step, four-step, and decaying recurrent injection are compared at matched total intervention energy. The original single-step protocol is not silently converted after outcomes are observed.

17. Eleven-Arm Causal Program

Every arm reconstructs the same prompt and byte-identical token prefix independently. No arm inherits a mutated cache from another. The hook is installed immediately before one frozen forward pass at one frozen layer and generation horizon, then removed immediately. The baseline is a fresh reconstruction with no state modification.

Arm	State operation	Scientific purpose
Baseline	h' = h	Measure unperturbed success.
Amplify	h' = h + beta ||h|| d_hat	Increase the learned success-associated component.
Negative	h' = h - beta ||h|| d_hat	Test directional degradation.
Suppress	Reduce current projection onto d_hat without crossing zero.	Test partial component dependence.
Reverse	Reflect the current component through zero.	Test sign-specific causal ordering.
Delete	Remove the registered direction or subspace projection.	Test necessity.
Matched-norm random	Equal-norm frozen random direction.	Control generic perturbation energy.
Orthogonal placebo	Direction orthogonal to target geometry.	Control unrelated state movement.
Off-layer	Same intervention at a registered non-target layer.	Test spatial specificity.

Off-token	Same intervention at a registered non-target generation step.	Test temporal specificity.
Token-matched	Geometrically distinct placebo selected to match immediate logit effect.	Control token/readout disruption.

Development target: Success(+delta) > Success(baseline) > Success(-delta)

17.1 Specificity conditions

- Positive versus baseline paired root-bootstrap interval excludes zero.
- Baseline versus negative/reversal interval excludes zero in the predicted direction.
- Matched random and orthogonal placebo remain materially closer to baseline.
- Off-layer and off-token effects attenuate.
- Token-matched control remains weaker despite comparable immediate logit impact.
- Protected-suite degradation remains inside the frozen bound.
- Selected ordering reproduces from a clean process and source/model receipt.

18. End-to-End Development Pipeline

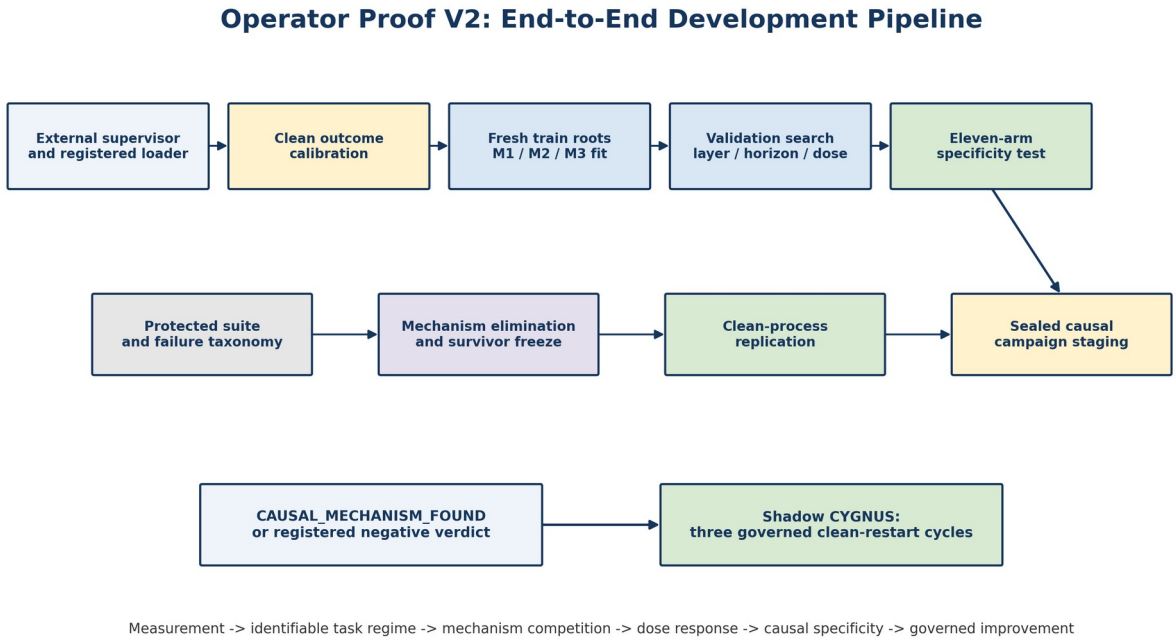

Figure 11. Operator Proof V2 development pipeline from supervised window control to sealed causal campaign staging.

18.1 Pipeline stages

1. External supervisor fingerprints production, acquires locks, stops V4, and proves GPU release.
2. Registered loader asserts PCI ordering, RTX 5090 identity, capacity, exact Hermes revision, bfloat16 parameters, and absence of quantization before weight materialization.
3. Clean calibration selects an identifiable task regime on disjoint calibration roots.
4. Fresh train roots generate clean pre-intervention states and externally verified outcomes.
5. M1, M2, and M3 are fitted completely on train only.
6. Validation selects mechanism, layer, horizon, normalization, rank, and minimum viable dose.
7. All eleven arms execute on byte-identical prefixes with independently reconstructed caches.
8. Protected-suite and failure-taxonomy results are computed.
9. Weak mechanisms are eliminated; the strongest survivor and complete receipts are frozen.
10. A clean-process replication reconstructs the selected experiment.
11. Only then is a sealed causal campaign staged.
12. The supervisor drains the GPU, restores V4, verifies the production service, and releases locks.

18.2 Allowed development verdicts

Verdict	Meaning
CAUSAL_MECHANISM_FOUND	A mechanism satisfies directional ordering, matched-control specificity, protected-suite preservation, and clean-restart replication on development/validation.
DIRECTIONAL_EFFECT_PRESENT_BUT_NONSPECIFIC	Behavior changes directionally, but placebos, token effects, off-target effects, or broad degradation prevent a proprioceptive interpretation.
NO_VALIDATED_MECHANISM_FOUND	All mechanisms receive valid clean-label, valid-dose, valid-arm tests and none satisfies the registered development criteria.
INVALID_DEVELOPMENT_RUN	Task labels, model contract, intervention delivery, source binding, support, or evaluation mechanics prevent interpretation.

19. Protected Suite and Clean-Restart Replication

A steering direction is not useful if it improves one narrow arithmetic metric by broadly degrading language, formatting, recall, or basic reasoning. The protected suite is frozen before mechanism selection and evaluated across baseline and intervention arms.

Protected domain	Example metric	Failure signal
Basic arithmetic	Exact answer accuracy	Broad degradation outside calibrated target regime.
Factual recall	Frozen factual score	Loss of ordinary knowledge access.
Instruction compliance	Programmatic constraint pass	Increased format or rule failures.
Short code generation	Executable unit-test pass	Syntax or semantic collapse.
Language coherence	Length, malformed output, log-loss proxy	Catastrophic or incoherent generation.

19.1 Clean restart

After validation selection, the pilot process terminates and the external supervisor restores V4. A new causal window loads Hermes from a clean process, reconstructs the frozen direction and configuration from receipts, and repeats the selected validation experiment. The replicated run must preserve direction hash, layer, horizon, amplitude, arm identities, qualitative ordering, specificity pattern, and protected-suite status.

20. Current Development Evidence and Status

Evidence item	Status	Current interpretation
PTB-1 V2 bank: 800 roots, 6,400 trajectories, 147,100 arrays	Demonstrated	Complete measurement substrate.
Functional counterfactual reconstruction, Contract A	Demonstrated	Registered repeated and clean-process model-derived fields reproduced.
Exact-prefix hidden-state identity	Demonstrated	Same exact prefix yields bit-identical on-manifold state under the registered deterministic contract.
Operator Proof V2 source and 59/59 mechanics	Demonstrated implementation	Architecture is mechanically executable; efficacy remains open.
External CPU-only supervisor restore smoke	Demonstrated operationally	V4 recovery no longer depends on orderly pilot cleanup.
Original capability-cliff magnitude	Superseded	Inflated by parser, token-cap, and textual-stop confounds.
Corrected steep competence transition	Provisional	Requires clean-outcome remeasurement and dense calibration.
Mean-state separation 3.1669	Measured, semantics confounded	Real geometry; may encode correctness, completion, verbosity, stability, or mixtures.
Original eighteen-cell zero-effect grid	Dose-floor result	Old RMS-scaled dose did not alter trajectories; mechanisms not validly tested.
Clean outcome contract	Implementation in progress	Required before clean mechanism fitting.
Valid dose response and eleven-arm specificity	Open	No admissible causal verdict yet.
CYGNUS AIT-1	Open	Requires validated selector/control channel and three governed clean-restart cycles.

Current scientific boundary

The architecture has not received a valid positive or negative causal verdict. The next result must come from clean mixed outcomes, freshly refitted mechanisms, a measurable dose response, the complete eleven-arm comparison, protected-suite preservation, and clean-restart replication.

21. Mechanism Localization and Active-Dark Exchange

A causal direction becomes more informative when its writer, reader, and transfer pathway can be localized. Operator Proof V2 therefore connects intervention outcomes to the spatial-temporal architecture rather than treating steering as an isolated vector trick.

- Identify the first layer and horizon at which the selected direction becomes behaviorally effective.
- Measure whether active-to-dark or dark-to-active exchange changes near the effective intervention site.
- Use off-layer, off-token, deletion, activation patching, and interchange to distinguish route-specific control from generic damage.
- Localize attention heads or MLPs that read or write the selected state.
- Compare immediate logit L2/KL with downstream success to separate readout disruption from deeper causal effects.
- Test whether later verification state changes after intervention.
- Measure whether M3 produces stronger specificity or lower token disruption than M1/M2.

$$X_l = ||P_A F_l P_D||_F^2 + ||P_D F_l P_A||_F^2$$

$$chi_l = ||P_D F_l P_A||_F^2 - ||P_A F_l P_D||_F^2$$

A validated causal effect localized to a depth/horizon region, attenuated off-layer/off-token, and accompanied by a coherent active-dark exchange change would provide a stronger mechanistic basis than behavior change alone.

22. CYGNUS Governed Autonomous Closure

CYGNUS is an operating research and control architecture integrating hidden-state hooks, persistent internal state, candidate-local reset fields, routing, autonomous mission execution, persistent memory, external verification, receipts, rollback, and governed promotion. The next role of the validated control channel is not merely to steer answers; it is to rank and regulate CYGNUS's own candidate research actions.

22.1 AIT-1: Autonomous Improvement Threshold 1

1. Detect a measurable failure.
2. Generate multiple competing hypotheses or interventions.
3. Predict candidate internal trajectories and external outcomes.
4. Choose among candidates without human selection.
5. Execute the selected candidate.
6. Evaluate on a sealed external suite.
7. Retain or reject under fixed promotion rules.
8. Reproduce the improvement from a clean restart.
9. Pass protected-suite regression checks.
10. Preserve receipts, rollback state, and evidence status.

> **Governing invariant**
>
> Autonomy over search and construction is not autonomy over evaluation and promotion. CYGNUS may improve its solutions; it may not redefine what counts as improvement.

22.2 Three-cycle closure

AIT-1 closes only after three consecutive governed cycles. Each cycle must use immutable evaluators, thresholds, promotion criteria, protected-suite rules, and rollback state. A positive causal mechanism is therefore a component of the ascension architecture, not the final claim.

23. Instrument Integrity and Cumulative Correction

The program's evidence discipline is cumulative rather than destructive. A defect supersedes the affected number, label, or implementation claim; it does not erase unrelated measurements or the architecture. The table below records the current disposition of the major development corrections.

Development finding	Scientific risk	Correction	Disposition
Rows durable while arrays volatile	Resumed bank silently missing activations	Atomic per-root rows-plus-arrays artifacts	Old pass retracted; new persistence demonstrated.
Shell wrapper killed instead of Python child	Interruption test vacuous	Direct process-group kill with PID/PGID proof	Old interruption claim retracted; authenticated n=2 pass preserved.
Incomplete stop-token set	Post-termination text contaminates labels and runtime	Artifact-derived semantic terminator union	Corrected and confirmed.
Legacy verifier dispatch returned zero	All labels degenerate	Offline immutable label overlay and verifier fixtures	Original labels superseded; bank preserved.
Exact-prefix predictive framing	Test becomes deterministic identity under unlimited reconstruction	Shift to matched operational selection and fixed-prefix causality	Primary proof target superseded.
Comma-blind parser, 48-token cap, textual stop	Inflated arithmetic cliff	Clean final-answer parser, semantic stopping, larger budget	Original magnitude superseded; transition provisional.
RMS-scaled intervention too weak	False causal null	Norm-scaled beta ladder with immediate-logit validity	Old grid retained as dose floor.
Pilot process owned V4 recovery	Repeated manual production restoration	External CPU-only supervisor	Supervisor smoke passed.
Token-matched/orthogonal arm collision	Different labels could hide identical algebra	Pairwise geometric distinctness assertion	Mechanics defect repaired.
selectors.py shadowed stdlib selectors	Transitive import failure	Renamed selection.py	Mechanics defect repaired.

These corrections harden the instrument before the decisive causal experiment. They are not the central theoretical contribution. The contribution remains the integration of sensing, candidate comparison, fixed-prefix control, external verification, and governed retention.

24. Updated Falsifiable Predictions

1. Exact-prefix identity: under the deterministic registered contract, identical prefixes yield identical on-manifold hidden states.
2. Operational selection: under matched training roots, estimator family, feature dimension, and charged compute, already-available hidden state selects better procedures than best-fixed, token, logit/confidence, and random alternatives.
3. Clean outcome identifiability: an adaptive calibrator can find a task regime with mixed, verifier-clean, non-truncation-dominated outcomes.
4. Mechanism competition: at least one of M1/M2/M3 will outperform matched random and orthogonal perturbation on validation, or the development program will issue NO_VALIDATED_MECHANISM_FOUND.
5. Dose response: hidden displacement and immediate logit effect increase with beta before catastrophic degradation dominates.
6. Directional causality: amplify exceeds baseline and negative/reversal degrades success.
7. Control specificity: random, orthogonal, off-layer, off-token, and token-matched controls remain weaker.
8. Dark-subspace specificity: M3 may achieve a better effect-to-logit-disruption ratio than M1 or M2.
9. Mechanism localization: effective intervention concentrates in registered layers and horizons rather than appearing uniformly.
10. Protected-suite conservation: target improvement does not require broad capability or coherence loss.
11. Clean-restart replication: the selected effect reconstructs from immutable receipts in a fresh process.
12. Governed improvement: shadow CYGNUS completes three sealed-suite improvement cycles without evaluator or promotion-rule modification.

25. From Measurement to Governed Recursive Improvement

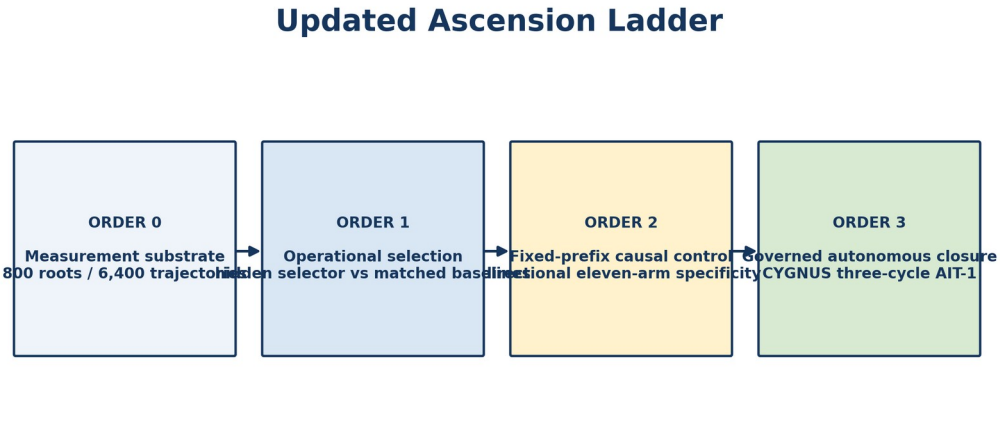

Figure 12. Updated closure ladder: measurement, matched selection, fixed-prefix causal control, and governed autonomous improvement.

Order 0 closes the measurement substrate. Revised Order 1 asks whether already-computed internal state improves action selection under matched operational resources. Order 2 closes causal proprioception through fixed-prefix directional intervention with matched-control specificity. Order 3 closes governed autonomous improvement through three clean-restart CYGNUS cycles.

25.1 Implications for artificial superintelligence

The ASI implication is conditional and architectural. If internal state selects cognitive procedures under matched resources, fixed-prefix intervention specifically controls future success, and CYGNUS repeatedly selects and retains improvements under immutable external 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 edition does not claim that ASI, causal steering, or autonomous closure has been achieved. It specifies and implements the shortest nondegenerate experimental sequence by which a proprioceptive architecture could become a credible route toward governed recursive improvement.

26. Discussion

The decisive conceptual advance is the correction of the proof target. Exact-prefix reconstructibility means that information-theoretic independence from the complete token history is not the correct architectural criterion. Internal state matters when it is already available, compresses control-relevant trajectory information under realistic resource constraints, and can be manipulated to alter the future while visible history remains fixed.

The fixed-prefix experiment is therefore stronger than ordinary activation steering. Generic steering shows that a model can be pushed. Proprioceptive control requires a direction derived from externally verified trajectory outcomes, trained and selected without test leakage, applied after a common prefix, directionally ordered against baseline and negative arms, specific against matched placebos, and safe under protected-suite and clean-restart tests.

Adaptive calibration is also part of the architecture. A system cannot learn a control direction at ceiling, floor, or under dirty outcomes. The calibrator generates competing task regimes, evaluates them externally, eliminates unidentifiable regimes, and promotes the first regime that can support causal identification. This is not merely benchmark hygiene; it is a self-organizing experimental layer.

The provisional calibration transition remains scientifically interesting, but it should not be called a phase boundary until clean dense remeasurement supports that term. Likewise, the 3.1669 mean-state separation is real geometry but not yet a correctness direction. The program preserves these observations while demanding the clean experiment that determines their meaning.

The external supervisor illustrates a second form of proprioception at the system level: the research architecture monitors its own operational state, detects experiment failure, restores production, and preserves evidence. This is not model cognition, but it is part of the governed architecture required for reliable autonomous improvement.

27. Limitations

1. Operator Proof V2 has passed source and mechanics tests but has not yet produced a valid registered Hermes causal arm result.
2. The clean outcome contract and authoritative clean calibration curve remain active implementation work.
3. The corrected competence transition is provisional; lower points remain truncation-confounded until remeasurement.
4. The 3.1669 geometric separation has confounded semantics and cannot yet be promoted as correctness geometry.
5. The original zero-effect grid used an insufficient dose and provides no mechanism verdict.
6. Matched operational selection remains specified but untested against a fully frozen compute/latency fairness contract.
7. A positive development effect would still require a sealed causal campaign, protected-suite replication, and clean restart.
8. Transfer beyond one Hermes checkpoint, one model family, and one task family remains required.
9. Contract B serialized live runtime-state restoration remains untested.
10. The historical CYGNUS operator seed never executed on big and is not prior causal evidence.
11. The PTB-1 predictive program did not close; it yielded theorem, task-design, label, and control lessons rather than a positive predictive verdict.
12. The control-capacity functional remains theoretical.
13. The term proprioception is functional and does not imply consciousness, subjective experience, resistance, or strategic self-defense.

28. Conclusion

The Proprioceptive Transformer is an architecture for reconstructing a common functional origin, measuring persistent and candidate-local dynamics, selecting among cognitive procedures, intervening on hidden state after a fixed observable prefix, verifying outcomes externally, and retaining improvements under governed rules.

The measurement substrate is complete. The exact-prefix theorem has corrected the predictive target. Operator Proof V2 now provides a source-complete, manually decoded, validation-selected, eleven-arm causal architecture with 59/59 mechanics tests and an externally validated production-recovery supervisor. Adaptive calibration, clean outcome parsing, mechanism competition, dose-response validity, protected-suite preservation, and clean-restart replication form the remaining development path.

The next decisive result is explicit: a cleanly learned internal direction must produce a measurable dose response and the ordering Success(+delta) > Success(baseline) > Success(-delta) while matched controls remain weaker. If that result survives sealed replication and CYGNUS then completes three governed improvement cycles, transformer computation becomes the engine inside a self-measuring, self-correcting, and governably improvable cognitive architecture.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., et al. “Attention Is All You Need.” NeurIPS, 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.” PNAS, 1998.

[15] Napolitano, L. “CYGNUS, PTB-1, SRX-3, SRX-4, Proprioceptive Transformer, and Operator Proof V2 Technical Receipts.” Proprioceptive AI, Inc., 2026.

Appendix A. Detailed Evidence Ledger

Claim	Tier	Promotion requirement
Complete PTB-1 measurement substrate	[D]	Already closed under receipts.
Functional reconstruction under Contract A	[D]	Already demonstrated under registered repetitions.
Exact-prefix identity	[D]	Already confirmed under deterministic capture contract.
Operator Proof V2 mechanics	[D-implementation]	59/59 tests and source freeze.
External supervisor recovery	[D-operational]	Validated smoke; production restoration verified.
Clean competence transition	[P]	Clean outcome remeasurement and dense calibration.
Correctness-specific geometry	[P]	Clean-label decomposition and replication.
Valid dose response	[U]	Nonzero hidden/logit effect with safe generation.
Causal mechanism	[U]	Directional ordering plus matched controls and protected suite.
Causal proprioception	[U]	Sealed replication and clean restart.
CYGNUS AIT-1	[T/U]	Three governed clean-restart improvement cycles.
Bare transformer obsolete as complete architecture	[T]	Operational selection, causal control, and autonomous closure all pass.

Appendix B. Source and Module Map

Artifact	Role	Binding expectation
baseline/cygnus_hidden_operator_lab_v1.py	Historical unvalidated seed from lil.	Read-only provenance; never treated as live causal evidence.
src/cygnus_operator_proof_v2.py	Top-level pilot orchestration.	Frozen source receipt.
src/manual_decode.py	Exact-prefix manual decoding and one-step intervention.	Model/source/prefix receipts.
src/interventions.py	Eleven arm algebra.	Arm distinctness and geometry tests.
src/directions.py	M1-M4 construction.	Train-only fit receipts.
src/selection.py	Validation-only selection.	Frozen leaderboard and tie rule.
src/tasks.py	Calibration and development tasks.	Disjoint manifests and root hashes.
src/verifiers.py	Outcome parser and protected suite.	Fixture and error receipts.
src/receipts.py	Hash-addressed provenance.	Complete artifacts and environment.
src/window_supervisor.py	CPU-only production recovery.	Supervisor restore receipt.

Appendix C. Mechanism Definitions and Selection

Mechanisms are generated on train roots, not chosen by theory. Validation uses a frozen combined criterion consisting of external directional ordering, matched-control specificity, protected-suite preservation, and minimum viable dose. Direction norm or raw output disruption alone cannot select a mechanism.

Mechanism	Train-only inputs	Validation output
M1	Successful and failed mean states.	Direction, layer, horizon, beta.
M2	Residualized states, labels, regularization grid.	Regularization, direction, layer, horizon, beta.

M3	Train-only low-variance basis and success signal.	Subspace rank, direction, layer, horizon, beta.
M4	Action-conditioned state transitions and outcomes.	Action transition direction and intervention configuration.

Appendix D. Eleven-Arm Algebra

```
baseline:      h' = h
amplify:       h' = h + beta ||h|| d_hat
negative:      h' = h - beta ||h|| d_hat
suppress:      h' = h - gamma <h,d_hat> d_hat,  0 < gamma < 1
reverse:       h' = h - 2 <h,d_hat> d_hat
delete:        h' = h - P_D h
matched_random:h' = h + beta ||h|| r_hat,  ||r_hat||=1
orthogonal:    h' = h + beta ||h|| q_hat,  <q_hat,d_hat>=0
off_layer:     target intervention at frozen non-target layer
off_token:     target intervention at frozen non-target generation step
token_matched: geometrically distinct q_hat selected on validation to match immediate logit effect
```

Appendix E. Clean Outcome and Calibration Schema

Field	Required value
Stopping	Artifact-declared semantic terminator IDs only.
Maximum tokens	Sufficient for registered derivation contract.
Parser	Explicit final-answer parser with comma and punctuation equivalence.
Negative taxonomy	Wrong, truncation, verifier error, malformed, no final answer.
Success band	0.30-0.70.
Minimum support	At least 12 positive and 12 negative candidates.
Mixed roots	At least 25%.
Verifier errors	0.
Truncation-only negatives	<=20%.
Malformed/no-final-answer negatives	<=20%.
Partition rule	Calibration disjoint from train/validation/mechanics/protected/test.

Appendix F. Dose-Response Receipt Schema

Field	Description
root_id / arm_id	Immutable experimental identity.
prefix_token_ids / prefix_sha256	Byte-identical pre-intervention history.
mechanism / direction_sha256	Frozen train-derived direction.
layer / token_horizon / beta	Frozen validation-selected configuration.
hidden_pre_sha / hidden_post_sha	State binding.
displacement_ratio	||Delta h|| / ||h||.
hook_fire_count	Must equal one.
logit_l2 / logit_kl / entropy_delta	Immediate readout effect.
top1_changed	Discrete next-token effect.
continuation_sha256 / output_length	Downstream generation.

outcome / failure_taxonomy	External verifier result.
protected_suite_delta	Broad capability effect.
runtime / gpu_memory	Operational cost.

Appendix G. Acceptance and Promotion Criteria

- Exact registered Hermes revision, bfloat16 parameters, unquantized, registered RTX 5090.
- Clean identifiable outcomes and disjoint partitions.
- Byte-identical prompt and prefix across every arm.
- Independent cache reconstruction; no arm inherits another arm's state.
- Hook absent during prefill; fires once at the registered layer and token step; removed immediately.
- Direction fit on train only.
- Mechanism, layer, horizon, amplitude, normalization, and rank selected on validation only.
- All eleven arms distinct and fully recorded.
- Primary directional ordering and paired root-bootstrap intervals.
- Matched-control specificity and protected-suite non-regression.
- Complete source, model, prefix, direction, result, environment, and supervisor receipts.
- Clean-restart replication before causal promotion.
- Sealed causal campaign before CAUSAL_PROPRIOCEPTION_CLOSED.
- Three governed clean-restart CYGNUS cycles before AIT-1.

Appendix H. Supersession and Retraction Register

Item	Disposition
Pooled SRX-3 z near 90	Retracted under tighter within-root null.
SRX-3 24/24 Holm result	Superseded by 10,000-permutation max-T: 5/24 cells.
Original SRX-4 horizon axis	Invalid; corrected true-horizon pilot passed.
Row-only atomic resume	Retracted after 375 missing arrays were found.
Wrapper-kill interruption claim	Retracted; Python child survived.
opus5_capture.npz as activation evidence	Retracted; self-declared random placeholder.
Original PTB labels	Superseded by immutable corrected label overlay.
Exact-prefix Shannon irreducibility as primary proof	Superseded by operational selection and fixed-prefix causality.
Original 0.969 -> 0.031 cliff magnitude	Superseded by parser/token-budget correction.
Corrected steep transition	Provisional pending clean-outcome remeasurement.
Old 18-cell zero-effect grid	Retained as insufficient-dose calibration, not causal null.
Model resistance interpretation	Unsupported.

Appendix I. Major Changes from v2.3 to v2.5

- Adds the complete external supervisor architecture and validated unattended restore smoke.
- Adds the clean outcome contract, final-answer parser, failure taxonomy, and qualification thresholds.
- Adds adaptive calibration as an explicit experimental layer.
- Supersedes the original cliff magnitude while preserving the corrected transition as provisional.
- Preserves 3.1669 geometry as measured but semantically confounded.
- Recasts the original zero-effect steering grid as an insufficient-dose floor.
- Adds norm-scaled beta dose calibration and immediate-logit validity checks.
- Expands M1/M2/M3/M4 mechanism definitions and selection criteria.
- Expands all eleven intervention arms with algebra and specificity rationale.
- Adds the end-to-end causal development pipeline, protected suite, clean restart, and allowed verdicts.
- Updates the evidence ledger and ascension ladder without claiming causal success.

Declarations

Data and code availability

Frozen specifications, source hashes, manifests, complete-bank receipts, corrected label overlays, source-and-mechanics receipts, supervisor receipts, calibration artifacts, and future causal result packages are maintained as hash-addressed artifacts. Release or independent replication packages should bind the original immutable bank, overlays, sources, model revision, task manifests, evaluator, and environment.

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.

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.