Integrated GTCS Ablation Benchmark: Reduction, Measurement, and Causal-Identification Stress Tests

Kostiantyn Osmolovskyi
Independent Researcher, Ukraine
ORCID: 0009-0006-3144-7237

Computational Methods Adjunct
Version 1.0
2026-09-01
No Technical Note number assigned in this release

Abstract

This computational adjunct reports the first integrated ablation benchmark constructed around selected distinctions in the General Theory of Cognitive Structuring (GTCS). Its purpose is not to verify GTCS empirically or to convert formal distinctions into claims of universal computational necessity. Instead, a frozen synthetic realization is used to ask narrower questions: whether selected variables can be removed, collapsed, reparameterized, or behaviorally compiled away; whether observational confounding can make reduced models appear adequate; whether intervention-informed calibration identifies the intended causal role; how measurement noise changes recoverability; and whether high proxy-level predictive accuracy preserves a declared latent decision criterion outside calibration support. The benchmark produces mixed results. Current-state and finite recent-history rivals fail to reproduce history-sensitive outcomes in the frozen realization, supporting the computational utility of some history-bearing state information without establishing uniqueness of the GTCS retained-overload scalar. Pre-symbolic admission resists simple signalization and global-threshold collapses but is exactly reparameterizable by a context-dependent gate. An explicit transient coherence-representation variable is distinct in typed definedness but behaviorally compilable away for the single stateless downstream output used here. Observational models can achieve near-perfect in-distribution accuracy while failing at chance under frozen history interventions; recurrent recovery of a latent history state does not guarantee identification of its downstream causal role. Repeated measurement substantially improves out-of-distribution recovery, while a later criterion-transfer test shows that near-perfect calibration-support prediction can coexist with systematic misalignment from the intended latent criterion. The benchmark therefore supports a methodological separation among formal non-equivalence, computational necessity, state estimation, causal-role identification, support-level prediction, and criterion-transfer validity.

Keywords: GTCS; ablation benchmark; computational reduction; causal identifiability; measurement error; out-of-distribution testing; criterion transfer; operationalization.

1 Purpose and Epistemic Ceiling

The benchmark was constructed after a strict conceptual collision audit identified a recurring problem: a typed distinction can be formally legitimate while its computational role in a

particular realization remains untested. The goal of this adjunct is therefore destructive rather than confirmatory. For each selected distinction, the benchmark asks whether a principled rival can remove or collapse the distinction while retaining the same declared output or intervention behavior.

The strongest licensed statement is local:

the frozen realization permits or requires a property under its declared assumptions

not:

GTCS is empirically validated or universally computationally necessary.

The benchmark is synthetic. Formal target values, intervention cells, and latent states are known by construction. This makes exact reduction and identifiability tests possible, but it also means that every result inherits the data-generating assumptions, state spaces, thresholds, and output surfaces frozen in the design. No psychological, neuroscientific, clinical, organizational, or artificial-system generalization follows from the simulations alone.

2 Audited Distinctions

The first cycle focuses on four owner-level separations and three downstream methodological separations.

2.1 Owner-level separations

1. current structural pressure/tension versus retained history-bearing overload;
2. signalization versus context-relative pre-symbolic admission;
3. admitted discrepancy configuration versus coherence-representation construction;
4. candidate-relative admissibility versus direct current-pressure selection.

2.2 Methodological separations

1. latent-state recovery versus causal-role identification;
2. latent target versus noisy measurement estimate;
3. support-level predictive accuracy versus recovery of a declared latent criterion.

These methodological separations are not new GTCS owner objects. They belong to the operational and computational validation layer.

3 Frozen Deterministic Realization

The exact phase uses a finite synthetic architecture with current structural variables, signalization and admission contexts, a history-bearing retained state, partial representation construction, candidate-relative constraints, and a final selector. The retained scalar follows the frozen recurrence used by the benchmark:

$$L_{t+1} = \rho L_t + u_t, \qquad \rho = 0.5,$$

with finite history enumeration over five update steps. Candidate-specific history criteria are thresholded on the resulting latent state. The representation layer is partial and context-dependent. The benchmark intentionally does not claim that these threshold values or state spaces are canonical GTCS parameters.

Exact enumeration at horizon $H = 5$ produced 125 reachable retained-state values and 144,000 fully evaluated current states. Ten of ten unit-test methods and thirteen of thirteen structural invariant checks passed before reduction verdicts were accepted.

4 CA-012: Exact Ablation and Reparameterization

Rival	Mismatches	States	Verdict
No retained L: candidate admissibility	101,376	144,000	Non-equivalent
No retained L: selection	768	144,000	Non-equivalent
Signalization = admission	44,000	144,000	Non-equivalent
One global precision threshold: selection	696	144,000	Non-equivalent
Context-dependent admission gate	0	144,000	Exact functional reparameterization
Admitted configuration = representation	42,000	144,000	Non-equivalent typed/definedness status
Compiled $\mathrm{Sel} \circ \Phi$	0	144,000	Explicit transient representation behaviorally eliminable for v1 output
Direct pressure-to-selection	768	144,000	Non-equivalent

Table 1: Exact CA-012 reduction results.

4.1 History-bearing state

A frozen witness holds current input fixed while changing retained history. In the low-history case, $T = 4$ and $L = 0$ permit candidate δ_1; in the high-history case, $T = 4$ and $L = 4$ reject it. Hence a current-state-only factorization cannot reproduce both cases.

Moreover, raw recent-history rivals with $k = 1, 2, 3, 4$ all have exact counterexamples at horizon five: the same last-k excess tail can correspond to retained states on opposite sides of the relevant candidate threshold. This supports the computational utility of some history-bearing state in this realization. It does not establish that the scalar L is the unique minimal sufficient state, nor that every cognitive system requires this recurrence.

4.2 Pre-symbolic admission

The simple identification of signalization with admission fails exactly, as does the best single global precision-only threshold. However, a context-dependent gate reproduces the frozen admission truth table with zero error. The correct verdict is therefore mixed:

typed/causal separation survives; functional irreducibility is not shown.

The benchmark does not support a claim that the PSA admission relation is mathematically irreducible to every generic gating mechanism.

4.3 Coherence representation

Identifying the admitted configuration directly with the representation fails because representation construction has separate partial definedness. Yet the downstream selector composed with the construction map is behaviorally exact without materializing an explicit transient representation variable. Thus:

$$\boxed{\text{typed distinction survives; explicit transient variable is eliminable for this one output.}}$$

A stronger computational case for an explicit representation would require a richer test surface such as persistence, independent availability, degradation, adequacy manipulation, delayed reuse, or multiple downstream consumers.

5 CA-013: Observational Confounding and OOD Failure

The exact architecture was then embedded in a noisy observational generator. The generator was frozen before model fitting. It was deliberately constructed so that ordinary current conditions strongly predicted low retained load in the observational training distribution, making a current-only shortcut highly attractive.

Model	ID accuracy	History-OOD accuracy
Full typed + noisy L proxy	0.9986	0.7417
Current only	0.9945	0.4999
Raw history $k = 1$	0.9962	0.5003
Raw history $k = 2$	0.9968	0.4965
Raw history $k = 3$	0.9976	0.5001
Raw history $k = 4$	0.9977	0.4976
Learned recurrent state	0.9968	0.5427

Table 2: CA-013 observational and frozen history-OOD selection performance.

The current-only model reaches approximately 99.45% ordinary held-out accuracy but collapses to approximately 50% on the frozen history intervention set. Finite raw-history models show the same qualitative pattern. Hence:

$$\boxed{\text{excellent observational fit} \not\Rightarrow \text{correct history-state factorization.}}$$

The recurrent model is especially informative. Validation selects $\rho = 0.50$, matching the latent recurrence coefficient, and the learned recurrent summary correlates with latent L at about 0.981 in-distribution. Nevertheless, its history-OOD decision accuracy remains only about 54%. This yields a stronger methodological separation:

$$\boxed{\text{latent-state recovery} \not\Rightarrow \text{causal-role identification.}}$$

The state can be reconstructed well while the downstream decision relation is learned incorrectly because the observational support provides little orthogonal evidence about what high retained load should do under otherwise eligible current conditions.

6 CA-014: Intervention-Informed Calibration

CA-014 added balanced, pre-specified history intervention examples to the unchanged 100,000-row observational training set. The calibration schedule was frozen as:

$$0,\ 100,\ 500,\ 1{,}000,\ 5{,}000,\ 10{,}000.$$

Current-only and raw-history $k \leq 4$ models remain near chance on the history-OOD pairs throughout the schedule. This is a structural identifiability result for those feature spaces: the paired low/high cases are identical in their current features and last-four histories while requiring opposite outputs. Additional labels cannot create information absent from the representation.

Model	0	0.1%	0.5%	1%	5%	10%
Current only	0.4999	0.5000	0.5000	0.5002	0.5001	0.5000
Raw history $k = 4$	0.4976	0.5006	0.4993	0.5006	0.5006	0.5001
Noisy L statistical model	0.7417	0.7460	0.7712	0.7573	0.7712	0.7644
Recurrent state	0.5014	0.5077	0.5041	0.5004	0.5011	0.5023
Latent oracle	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 3: CA-014 history-OOD accuracy across intervention-calibration fractions.

The strong pre-specified success condition was not met. The noisy-L learner peaks around 77%, and the first recurrent learner remains near chance. The latent oracle remains exact. CA-014 is therefore negative for sufficient causal calibration under the first frozen measurement/training design, while positive for the requirement that some distinguishing history information be represented.

7 CA-015: Measurement-Model Stress

CA-015 held the latent task fixed and varied only the observation model. Independent noisy measurements of L were averaged at:

$$m \in \{1, 2, 4, 8\}.$$

With per-measurement noise standard deviation 0.12, the standard deviation of the average decreases as $0.12/\sqrt{m}$.

Model	$m = 1$	$m = 2$	$m = 4$	$m = 8$
Direct L statistical learner	0.7778	0.8920	0.9759	0.9859
Recurrent statistical learner	0.5387	0.5950	0.6551	0.7446
Literal structured proxy rule	0.7431	0.7552	0.7538	0.7527
Latent oracle	1.0000	1.0000	1.0000	1.0000

Table 4: CA-015 primary history-OOD accuracy as measurement precision increases.

The direct state-aware learner improves from about 77.8% to 98.6%, showing that measurement quality explains a large fraction of the CA-014 gap. The recurrent learner also improves monotonically but remains substantially below the direct state model, indicating additional state-construction and causal-mapping limitations.

The literal structured rule produces a new boundary problem. The positive low-history cells lie exactly at the latent thresholds $L = \ell_i$. For any finite symmetric continuous noise, direct

application of the same numerical threshold to the noisy estimate classifies an exact boundary state correctly only about half the time. Therefore:

$$\boxed{\text{latent criterion} \neq \text{proxy decision boundary by default.}}$$

8 CA-016: Latent-Criterion Transfer

CA-016 separates task prediction from semantic fidelity to the intended latent rule. Five decision constructions were compared:

1. literal proxy threshold $\bar{L} \leq \ell$;
2. a continuous Gaussian measurement-error posterior with locally flat prior and 0.5 decision threshold;
3. a two-support Bayesian model assuming only $L = \ell$ or $L = \ell + 0.25$;
4. a logistic proxy classifier trained on those same two support states;
5. the latent oracle $L \leq \ell$.

The literal threshold and continuous-posterior rule are mathematically prediction-equivalent under the declared assumptions. The two-support Bayes rule places its proxy boundary at:

$$\ell + 0.125,$$

and the logistic classifier independently learns essentially the same shifted boundary.

Rule	$m = 1$	$m = 2$	$m = 4$	$m = 8$
Original two-point support accuracy				
Literal / continuous posterior	0.7394	0.7506	0.7531	0.7517
Two-support Bayes	0.8486	0.9302	0.9826	0.9981
Logistic proxy classifier	0.8488	0.9302	0.9825	0.9982
Criterion-transfer-grid accuracy				
Literal / continuous posterior	0.8692	0.9005	0.9236	0.9364
Two-support Bayes	0.8001	0.8180	0.8292	0.8331
Logistic proxy classifier	0.8006	0.8191	0.8284	0.8324

Table 5: CA-016 support prediction versus latent-criterion transfer. Values are means over the two candidates.

At $m = 8$, the support-conditioned Bayesian and logistic models approach 99.8% on the original two-state task but only about 83.3% on the pre-frozen transfer grid. The literal/continuous-posterior rule is much worse on the exact boundary-positive support state, but reaches about 93.6% criterion-transfer accuracy. On a novel latent-negative state $L = \ell + 0.0625$, the shifted classifiers become increasingly confident in the wrong positive classification as measurement precision improves.

The central result is therefore:

$$\boxed{\text{support-level predictive adequacy} \not\Rightarrow \text{latent-criterion recovery.}}$$

This motivates the methodological notion of criterion-transfer validity: whether a fixed operational decision relation remains aligned with the declared latent criterion on relevant target

states beyond the support used to calibrate that relation. The notion is an operational audit tool, not a new GTCS primitive.

9 Integrated Reduction Verdict

Target distinction	Benchmark result	Correct evidential ceiling
T versus retained L	Current-only and finite-tail rivals fail exact and OOD tests	Supports some history-bearing information in v1; no uniqueness or empirical claim
Signalization versus admission	Simple collapses fail; context gate exact	Typed/causal modularity; no functional irreducibility claim
Admitted configuration versus representation	Naive identity fails; compiled selector exact	Typed/definedness distinction survives; transient explicit representation behaviorally eliminable for v1 output
Candidate-relative admissibility	Direct pressure/current-input rivals fail	Local support under declared benchmark bridge only
Latent state versus causal role	Recurrent state can recover L while OOD decision fails	Reconstruction evidence is not causal-role evidence
Latent target versus measurement	Repeated measurement sharply improves OOD recovery	Measurement model is part of construct identifiability
Latent criterion versus proxy classifier	Near-perfect support prediction can fail criterion transfer	Predictive fit is not semantic recovery of owner criterion

10 What the Benchmark Does Not Establish

The following inferences are explicitly prohibited:

- that GTCS has been empirically validated;
- that the retained-overload scalar is a uniquely necessary state variable;
- that PSA admission is computationally irreducible to every context-sensitive gate;
- that an explicit coherence-representation variable is required for every downstream use;
- that the benchmark's candidate thresholds are canonical GTCS constants;
- that a successful OOD classifier has identified a psychological or neural mechanism;
- that the benchmark discriminates GTCS from all predictive-processing, active-inference, control-theoretic, or recurrent-state formulations;
- that failure of one operational proxy falsifies the corresponding formal target;
- that exact simulation recovery transports automatically to biological, human, artificial, institutional, or inter-system domains.

11 Reproducibility and Frozen-Design Discipline

The benchmark cycle was organized so that each later stage inherited the previous frozen deterministic core rather than repairing it after observing results. CA-012 froze the exact ablation surface. CA-013 froze the observational generator and OOD sets before fitting. CA-014 froze intervention-calibration fractions. CA-015 froze measurement multiplicities $m = 1, 2, 4, 8$ before the measurement stress run. CA-016 froze the latent transfer grid before fitting the support-conditioned classifiers.

A reproducibility manifest accompanies this adjunct and records SHA-256 hashes for the principal designs, run records, scripts, configurations, and summary result files. Large generated datasets are not required in the compact adjunct bundle when their generating scripts/configurations are present; where retained, they should be treated as reproducibility convenience rather than independent evidence.

12 Implications for GTCS Operationalization

The benchmark supports four methodological guards that should govern future simulations and empirical operationalizations:

$$\boxed{\text{formal non-equivalence} \neq \text{computational necessity},}$$

$$\boxed{\text{state recovery} \neq \text{causal-role identification},}$$

$$\boxed{\text{latent criterion} \neq \text{proxy decision boundary by default},}$$

and:

$$\boxed{\text{support-level prediction} \neq \text{criterion-transfer validity}.}$$

These guards do not weaken formal typing. They prevent formal distinctions from being over-promoted into computational or empirical claims without the relevant destructive comparison. Conversely, they prevent a computational reduction from being mistaken for a formal identity when the rival only compiles or reparameterizes the same information.

13 Recommended Next Tests

The first cycle is deliberately closed rather than extended post hoc. Future benchmark versions should be motivated by a new scientific question rather than by an attempt to improve current results. Especially informative extensions would include:

- multi-use or persistent coherence representation, with independent availability and degradation manipulations;
- alternative sufficient history-state constructions compared against scalar L;
- non-threshold candidate criteria and continuous decision loss;
- explicit causal estimators trained on interventions rather than only generic classifiers;
- proxy measurement models with correlated rather than independent errors;
- transfer grids derived from independently specified target distributions;

- cross-system or identity-layer ablations only after their own frozen realization designs are justified.

14 Conclusion

The first integrated GTCS ablation cycle does not produce a uniform confirmation pattern. That is its main scientific value. Some reductions fail decisively; some distinctions survive only as typed interfaces; some mechanisms can be reparameterized; one explicit intermediate representation can be compiled away for the declared single-use output; observational fit can be almost perfect while causal/OOD factorization is wrong; better measurement can rescue state-aware prediction; and task-optimal proxy classification can move away from the intended latent criterion even while its calibration accuracy approaches perfection.

The strongest supported methodological chain is:

$$\begin{aligned} \text{formal distinction} &\neq \text{computational necessity}, \\ \text{state measurement} &\neq \text{causal-role identification}, \\ \text{support-level prediction} &\neq \text{latent-criterion recovery}. \end{aligned}$$

The benchmark should therefore be treated as a reviewed computational and operationalization adjunct outside the formal GTCS verification core. Its role is to make future GTCS realizations easier to challenge, reduce, falsify, or delimit without confusing simulation evidence with empirical confirmation.