Finite Noisy Audits with Learned Baselines Exact risk, interface-dependent memory, and a matched compiled controller

B-Theory technical addendum | Version 1.0 | 9 November 2026
Internally AI-reviewed working note; not external peer review.

ABSTRACT

We give a reproducible finite specialization combining noisy calibration, declared cross-context interactions, exact finite risks, and interface-dependent streaming memory. Six local response bits and two interaction bits are unknown. A fixed experiment observes two baseline contexts and two audit contexts, leaving four final contexts unobserved. Here an audit is prespecified cross-context calibration and diagnosis, not independent validation of the supplied structural equalities. At three observations per block, optimal final Hamming risk is 0.104, while each interaction diagnosis has error 0.186368. Simultaneously attaining the specified prediction and diagnosis decisions requires 384 peak data states; prediction alone requires 96. A matched compiled controller preserves the same statistic and has an identical serialized implementation. We distinguish a fixed-baseline sample curve from optimal allocation and give the exact randomized minimax risk 0.137 for one specified unobserved interaction. These results are elementary specializations of established inference, distinguishability, and finite-state arguments. They establish neither an intrinsic self-model advantage nor priority, physical validation, or consciousness conclusions.

1. PURPOSE, OVERLAP, AND OPERATIONAL CONTRACT

The existing Reflective Self-Models pilot already separates evidence processing from architectural naming: its matched explicit, compiled, and external trackers share a mapping [R1]. We retain that null. The present increment is the complete finite learning/audit experiment, explicit finite risk distinctions, interface-dependent state bounds, and portable exact checks. It is a technical addendum, not a new general theory of representation learning.

The nearby Acquirable Predictive Models report treats exact reportability with restricted noiseless queries and known dynamics [R2]. Its acquisition distinctions motivate careful access contracts, but its order-based legality theorem is not used here. Finite-memory hypothesis testing has classical antecedents; Berg, Ordentlich and Shayevitz study deterministic time-invariant machines in a long-run setting [R3]. Our horizon is fixed and the data update is allowed a public time index. Thus clocked and time-invariant resource budgets must not be conflated.

Let x=(x1,x2,x3) range over {0,1}^3. Six local bits theta[j,b], j=1,2,3 and b=0,1, and two interaction bits gamma[A], gamma[B] are fixed and unknown. The full parameter class contains all 256 assignments. Define
A={001,010,011}, B={100,101,110}.
F1(x)=theta[1,x1] XOR gamma[A], x in A;
theta[1,x1] XOR gamma[B], x in B;
theta[1,x1], x in {000,111}.
Fj(x)=theta[j,xj], j=2,3.

Every scalar observation is Y=Fj(x) XOR E, with fresh independent E of probability 1/5. The noise law, port labels, group membership, and response equalities are supplied. They are not learned. The target is the noiseless response, not a future noisy measurement.

For positive odd n and r the fixed blocked schedule is:
Baseline: n observations each of theta[1,0], theta[1,1],
theta[2,0], theta[2,1], theta[3,0], theta[3,1], in that order.
Read component j at 000 for b=0 and at 111 for b=1.
Audit: r observations of component 1 at 001, then r at 110.
Final query: uniform x in Q={010,011,100,101}, disclosed afterwards.
Total scalar observations are T=6n+2r. No observation is taken at a final context. Every response parameter needed for nominal final prediction is nevertheless measured elsewhere. Transfer is justified by the supplied equalities, not by unseen-context evaluation alone.

A deterministic streaming controller has a fixed initial state and immutable code, receives one observation at a time, and updates using the public time index. Observations cannot be reread. Terminal readout receives only the terminal state and final query. Retained buffers, auxiliary state, and rereadable output logs count as memory. Fresh random output is introduced only in Proposition 5 and charged separately.

The prediction interface returns three response bits. The joint interface additionally returns both interaction diagnoses. The latter is required to attain the separately optimal prediction and interaction decisions below; it is not optimized for an unintroduced combined five-output vector loss.

2. IDENTIFIABILITY WITHIN THE DECLARED FAMILY

Proposition 1 (standard linear distinguishability specialization).
For a full parameter class GF(2)^p, known fixed repeatable observation rows collected in a matrix L, and target rows C, uniform reliable identification of C beta is possible exactly when rowspace(C) is contained in rowspace(L), under fresh BSC noise with known rate strictly between zero and one half. Reliable means error tending uniformly to zero with enough repetitions, not zero error at a finite noisy sample size.

Proof. If containment fails, there is v in kernel(L) with Cv nonzero. Mechanisms beta and beta+v have identical evidence laws and different targets; their pair-average error on a differing target bit is at least one half. If containment holds, repeat a finite basis of observed rows, recover its noiseless values by majority, and form the known target linear combinations. Each basis error tends to zero, and a finite union bound proves uniform reliability. This is a linear-algebra and indistinguishability argument, not a new identification principle.

Order beta=(theta10,theta11,theta20,theta21,theta30,theta31,gammaA,gammaB), with row indices starting at zero. Baseline rows are e0,...,e5. Audit rows are e0+e6 and e1+e7. Their ranks are 6 with no audit, 7 with either group alone, and 8 with both. Two scalar audit types are therefore necessary and sufficient for complete coverage of the two declared interactions. One audit already helps its group. Repetition, not merely coverage, determines finite reliability.

3. EXACT RISKS AT A FIXED ALLOCATION

Write alphaA=theta10 XOR gammaA and alphaB=theta11 XOR gammaB. Observations now concern six theta coordinates repeated n times and two alpha coordinates repeated r times. Define
e_s = sum over l=(s+1)/2,...,s of binom(s,l) (1/5)^l (4/5)^(s-l),
e=e_n, a=e_r, p=e+a-2ea.

Proposition 2 (separate exact Bayes and minimax risks).
For the stated family and fixed schedule, block majorities, direct alpha prediction, and XOR of the appropriate theta and alpha majorities attain:
Final normalized Hamming risk: H=(a+2e)/3.
Final three-bit vector failure: V=1-(1-a)(1-e)^2.
Each interaction diagnosis error: p=e+a-2ea.
At least one interaction wrong: G=1-(1-p)^2.
Full eight-bit mechanism failure: M=1-(1-e)^6(1-a)^2.

Proof. The transformation (theta,gamma) to (theta,alpha) is bijective. A uniform prior on the original bits becomes independent fair transformed bits. The likelihood factors, and odd block lengths give a unique posterior majority decision. For an XOR of independent posterior bits, its signed posterior bias is the product of their biases; hence XOR of their MAP decisions is the unique gamma MAP. The two diagnosis groups are independent. The three response coordinates at each final query are independent posterior coordinates, so their individual MAPs also maximize vector posterior probability. The full-mechanism MAP is preserved by the bijection. These observations prove the respective Bayes optima.

For any fixed mechanism, complementing each true-one observation block yields the true-zero experiment. Block-majority errors are independent and have probabilities e or a regardless of the mechanism. The displayed risks are therefore constant across the parameter class. A uniform Bayes lower bound equals the attained maximum risk, proving minimaxity separately for each loss. This is the standard constant-risk Bayes argument [R4].

At n=r=3, e=a=13/125. The values are H=0.104, V=0.280676864, p=2912/15625=0.186368, G=0.338002968576, and M approximately 0.584602310. If the two relevant baseline bits were known exactly, each gamma error would instead be a. That stronger information contract is not used in our learner. With n fixed and r growing, p tends to e and H tends to 2e/3: auditing cannot remove learned-baseline uncertainty.

4. MEMORY DEPENDS ON THE REQUIRED INTERFACE

A useful elementary fact is that streaming exact majority of 2k+1 bits requires k+2 peak states with a clocked update; a capped ones counter attains this even with a homogeneous update. At time k+1 the prefix counts 0,...,k+1 induce distinct suffix threshold functions, so they cannot be merged. This uses the classical distinguishing-extension method [R5]. Exact posterior reporting is a different task: its 2k+2 distinct count posteriors require that many terminal states.

Proposition 3 (exact capacity for full-information optimal decisions).
Under the fixed blocked clocked schedule and terminal-state-only readout, the least peak data-state capacities are
S_joint = max{32*(n+3)/2, 128*(r+3)/2}, terminal capacity 256;
S_pred = max{ 8*(n+3)/2, 32*(r+3)/2}, terminal capacity 64.
The joint claim requires simultaneously attaining the specified optimal prediction and gamma decisions. The prediction claim requires the optimal final response decisions only. Neither describes arbitrary suboptimal prediction quality or an optimal fused clock/controller.

Proof. Because every observation transcript and final query has positive probability and the relevant posterior MAP decisions are unique, attaining these Bayes risks forces the corresponding outputs on every transcript/query. A minimax rule attaining the displayed constant Bayes bound must likewise be Bayes optimal under the uniform prior. Thus the following residual-function argument bounds all eligible controllers, not merely the selected implementation.

Across the possible final queries, prediction readouts determine both alpha majorities and the four theta2/3 majorities. Adding both gamma reports recovers the theta1 majorities as well. Thus the readout function distinguishes all 256 joint decision vectors, or all 64 prediction-only vectors. This compares hypothetical readouts for possible late queries; it does not assume multiple actual queries or free output-log storage.

In the last baseline block, five earlier joint decisions combine with (n+3)/2 distinct residual majority functions, giving the first joint term. Common future observations and a suitable final query distinguish every pair. In the last audit block, seven completed decisions combine with (r+3)/2 residual functions, giving the second term. Earlier blocks cannot exceed these widths. A register packing completed decisions and the capped active count attains their maximum, replacing the count by its majority at each boundary. For prediction alone, ignore the first two baseline blocks; three earlier retained baseline decisions and five earlier retained audit-stage decisions give the factors 8 and 32. The same argument proves necessity and sufficiency.

At n=r=3, the joint/prediction peaks are 384/96, requiring 9/7 register bits; their terminal decision vectors require 8/6 logical bits. A generic separate clock for 0,...,T needs ceil(log2(T+1)) bits, five here. These are separate charges, not a minimum fused implementation. Keeping the mandatory irrelevant baseline observations in the prediction-only schedule is not sample-optimal for a redesigned task.

5. COMPILATION AND THE ALLOCATION COMPARISON

Proposition 4 (matched implementation).
Encode the full persistent configuration of a modular controller as one compiled state and retain its update and readout. Induction on observations gives identical outputs. In this example the compiled program retains the same packed statistic and need not store histories. Both supplied n=r=3 ROMs are byte-identical, with 181,505 logical bits and 22,689 padded bytes.

This is a concrete simulation and reproducibility result. It establishes neither equal shortest programs nor independently learned architectures, physical memory use, gate count, energy, latency, or engineering cost. The Python verifier's RAM is not the controller's nine-bit register.

The explicit dense encoding has eight 16-bit header fields, a B-bit initial state, 2TS B-bit updates, 8S five-bit readouts, and T five-bit context/port addresses. Its padded cost is
ceil((128+B+2TSB+40S+5T)/8) bytes, B=ceil(log2 S).
This fixed-header formula applies when all header values fit in 16 bits, as all displayed allocations do. It is an encoding upper cost, not a code-length lower bound.

Table 1. Fixed 44-sample allocation comparison; neither row is asserted globally optimal.

Quantity	n=3, r=13	n=5, r=7
Total samples	44	44
Final Hamming error	0.071667854	0.049728000
Per-gamma error	0.109546820	0.087401431
Full-mechanism failure	0.489796492	0.346763255
Joint peak states	1,024	640
Data bits / clock bits	10 / 6	10 / 6
Dense encoding bytes	117,805	73,645

The n=3 curve in the development report holds calibration effort fixed. Increasing calibration while reducing audit repetitions improves all three displayed learned-mechanism risks at this budget. Fewer states do not save a whole register bit here. The alternative code costs are calculated from the same

encoding, not separately generated alternative ROMs. No global Pareto frontier, adaptive allocation optimum, or competitiveness theorem follows.

6. ONE SPECIFIED INTERACTION OUTSIDE THE AUDIT

Enlarge the class by delta in {0,1}, flipping only F1(010). Its location and form are known; its value is not. There are 512 mechanisms. Every baseline/audit law is unchanged. The nominal learner assumes delta=0 and has, for delta=1,
H_nominal = H+(1-2a)/12.
At n=r=3 its risk is 0.104 for delta=0 and 0.17 for delta=1.

Proposition 5 (randomized minimax prediction in the enlarged family).
Allow an independent fair output bit when F1(010) is queried. For final normalized Hamming loss under the fixed schedule, the exact randomized minimax risk is
R_robust = (8e_n+3e_r+1/2)/12.
At n=r=3 this is 137/1000=0.137.

Proof. The four equiprobable queries give twelve equally weighted response positions. Eight use baseline theta2/3 entries and three use audited alpha entries. Under a uniform prior on the original bits and delta, the remaining target alphaA XOR delta is fair conditional on all permitted evidence. The coordinate Bayes risks are therefore eight copies of e_n, three of e_r, and one of 1/2, proving the lower bound. Use the nominal majorities at the eleven identifiable positions and a fresh fair guess at the affected one. This has the displayed constant risk for every fixed mechanism and either delta, attaining the bound.

The fresh unbiased bit is an additional entropy resource. It is not implemented by the original deterministic ROM or claimed extractable for free from its noisy data. This theorem is about Hamming prediction, not a jointly optimized diagnosis/full-mechanism loss or arbitrary unseen defects. It trades nominal in-family accuracy for a uniform guarantee over this particular enlargement. Its limiting 1/24 floor requires both n and r to grow; more repetitions at the original audit ports cannot reveal delta.

7. CLOSEST WORK, VERIFICATION, AND LIMITS

Wiedemer et al. provide compositional generalization conditions involving shared composition, support, differentiability, and an initial point [R6]. Schug et al. study identifying modules in a teacher-student model, with support, dimension and nondegeneracy assumptions; their theoretical population-loss setting and finite-data experiments are distinct [R7]. We do not learn latent module boundaries or composition rules. Rojas-Carulla et al. explicitly assume invariance for transfer [R8]; our unobserved defect illustrates why source observations alone do not validate such a restriction. Golovin, Krause and Ray study adaptive costly noisy tests [R9]; this fixed experiment has no corresponding adaptive guarantee.

The B-Theory hierarchical note supplies an observational-quotient interpretation [R10]. Here equality of observation laws is L(beta-beta')=0; support of finite noisy transcripts alone would not separate mechanisms because all transcripts have positive probability. Task distinctions can be coarser than mechanism distinctions, explaining the interface-dependent memory. No theorem from the separate Boolean-factorization manuscript is imported, and no priority claim depends on its unavailable complete text.

The package contains exact rational checks, two ROMs, and independent evaluation code. The primary calculation enumerates 65,536 block-count tuples. A separate checker evaluates 65,536 truth/error-mask pairs and parses the ROM. The review checker propagates probabilities in the original theta,gamma coordinates through 783,360 weighted ROM transitions, evaluates both delta values, and checks all time-cut widths of the displayed allocations. These are exact compressed computations, not

physical experiments or literal enumeration of all 2^24 words. Two independent AI reviewers also checked the proof contracts and small finite cases; the review summary describes their scope and repairs. Internal AI review is not human peer review or a proof-assistant certificate.

The supplied structural family, fresh noise, fixed parameters, trusted ports, late query, fixed schedule, loss, and randomization contract delimit the results. Unknown dependencies, changing/correlated noise, adaptive allocation, suboptimal-memory risk curves, minimal code/hardware cost, and unrestricted defects remain open. The contribution is the integrated finite benchmark and accounting contract. It is not a general principle of compositional learning or evidence of a self-model advantage.

REFERENCES

[R1] B-Theory. When Does an Explicit Self-Model Help? Calibration and Control under Internal Change in a Finite-State Pilot. Final R4 manuscript, October 2026. The current index and previously acquired manuscript document the inherited matched-controller null.
https://www.b-theory.com/manuscripts-html/reflective-self-models/

[R2] B-Theory. From Observable States to Acquirable Predictive Models. Revised technical report, 7 October 2026.
https://www.b-theory.com/manuscripts-html/acquirable-predictive-models/

[R3] Tomer Berg, Or Ordentlich, Ofer Shayevitz. Binary Hypothesis Testing with Deterministic Finite-Memory Decision Rules. arXiv:2005.07445v1, 2020.
https://arxiv.org/html/2005.07445v1

[R4] UC Berkeley Statistics 210A. Minimax Estimation, course reader. Section 1.3, constant-risk Bayes criterion. Exposition of classical decision theory, not a priority attribution.
https://stat210a.berkeley.edu/fall-2025/reader/minimax-estimation.html

[R5] Tal Malkin. The Myhill-Nerode Theorem and Implications. COMS W3261 Handout 3, September 2024, Section 2. Exposition of the classical distinguishing-extension method.
https://www.cs.columbia.edu/~tal/3261/fall24/Handouts/3_Myhill_Nerode.pdf

[R6] Thaddaeus Wiedemer, Prasanna Mayilvahanan, Matthias Bethge, Wieland Brendel. Compositional Generalization from First Principles. arXiv:2307.05596v1, 2023.
https://arxiv.org/html/2307.05596v1

[R7] Simon Schug et al. Discovering modular solutions that generalize compositionally. ICLR 2024, Sections 2-3, Theorems 1-2.
https://proceedings.iclr.cc/paper_files/paper/2024/file/6b40b312ce4933f4f2d1ac0b0b6b07f7-Paper-Conference.pdf

[R8] Mateo Rojas-Carulla, Bernhard Schoelkopf, Richard Turner, Jonas Peters. Invariant Models for Causal Transfer Learning. JMLR 19(36):1-34, 2018.
https://jmlr.org/papers/v19/16-432.html

[R9] Daniel Golovin, Andreas Krause, Debajyoti Ray. Near-Optimal Bayesian Active Learning with Noisy Observations. 2010; arXiv revision 2013.
https://arxiv.org/abs/1010.3091

[R10] B-Theory. Hierarchical Concepts Under Finite Resolution: Observation Quotients and Adaptive Symmetry. Research note v0.3, 2026.
https://www.b-theory.com/manuscripts-html/hierarchical-concepts/

Research-line metadata:
https://www.b-theory.com/papers/