

Composition is not reconstruction

What a compact model keeps—and what restricted access loses

Technical working manuscript · Evidence-synthesis edition 3 · 3 October 2026.

Public working copy · Authorship unassigned · Not peer reviewed or submitted.

This is a separate revision; all earlier editions are preserved.

Abstract

We separate executable consistency, learning within a supplied family, predictive sufficiency, and reconstruction. Five completed studies extend an earlier affine analysis. A learned primitive library predicts withheld whole orders and longer chains exactly in a clean known affine family, but its initial fixed-feature comparator is inadequate. An adequate endpoint-trained comparator shares that compositional prior; under noise it has lower mean forward RMSE in all 14 cells, with 13 adjusted intervals excluding zero. A channel ablation retains the advantage with target-only noise, so noisy training inputs are not necessary for the effect.

A fresh identical-observation length-one control agrees numerically with primitive OLS. Local information projected through task sensitivities predicts the endpoint advantage; all 14 residual-gap intervals contain zero. This is compatibility with a first-order account, not equivalence, complete mediation or architecture discovery.

On NASA C-MAPSS FD001, a selected four-sensor predictor has 15.25% higher engine-mean MSE than full ridge on 20 held-out simulated engines. Its adjusted interval lies above the frozen tolerance: noninferiority fails. A full-input three-PC model has the same six regression coefficients as compact and near-full error. Compact parameterization is therefore distinct from restricted sensor access. All 21 channels were loaded offline; physical sensing savings were not measured.

The synthesis supplies a claim/evidence/falsifier map, complete latest-study result tables, explicit cost boundaries and preserved failures. It runs no new experiment. Deterministic replay is software repeatability, not independent replication.

1 · Four questions that cannot substitute for one another

Imagine two boxes of stones. A total-count representation is enough to predict moving a stone between boxes, but not a reward that depends on the left box alone. Compression is neither inherently sufficient nor inherently harmful: sufficiency belongs to a task, an action set, a domain and a horizon.

Executable affine composition: $(A_2, b_2) \circ (A_1, b_1) = (A_2 A_1, A_2 b_1 + b_2)$

An interpreter and a compiled product of the same fitted maps can agree while both predict incorrectly. Exact arithmetic identities are classical. Prior Lean checks concern bounded algebraic specifications; they do not prove empirical validity or a verified Bend-to-runtime correspondence. The retained independent Bend kernel failure and floating-point counterexample are not resolved here.

Represented closure: $G(\phi(x), a) = \phi(F(x, a))$; task readout: $h(\phi(x)) = q(x)$

Closure and readout sufficiency are separate requirements. Task prediction can succeed without retaining everything needed to reconstruct x . For $C(x, y) = (x, 0)$, every pair $(x, y_1), (x, y_2)$ has the same output. A decoder given that output alone cannot recover both inputs. History, side information or a restricted admissible domain would define a different problem.

A noisy fitted matrix may be full rank even when the world transformation loses a coordinate. An inverse of that fitted matrix shows algebraic self-consistency, not recovery of erased information. Supplied loss contracts prevent false requests; they are not learned discoveries.

COMPUTATION

Same fitted rules

PREDICTION

Independent targets

RECONSTRUCTION

Recover original state

These distinctions motivate the evidence hierarchy below. Empirical forward-risk improvements cannot establish exact invertibility, minimal state, unbounded closure or a full OaK agent.

2 · The older evidence remains, with its limitations

The preceding 20-page edition analyzed 540 paired affine records: 20 seeds × three training sizes × three positive noise levels × three sequences. Direct prediction had lower mean latent RMSE in all 27 cells; 22 Bonferroni intervals excluded zero and 26 Holm-adjusted tests rejected. Training-label access was unequal. Its arithmetic maximum $1.3007197156015997 \times 10^{-16}$ checked two evaluations of the same model, not prediction. No zero-noise cases ran; exported noiselessMax=0 was vacuous.

A later six-cell comparison matched 120 transition records and 720 scalar supervised slots. Direct mean RMSE was 3.7%–9.1% lower, with six adjusted intervals excluding zero. Primitive fits received true-prefix training domains; parameter counts and label information still differed. An exploratory shared-trajectory low-noise lossy cell favored composition slightly; this was not an equal-information result.

Intermediate-noise reuse shifted five of six contrasts toward composition, but all three low-noise reused intervals contained zero. Reuse changed unique readings to 540 or 480 rather than 720. One frozen covariance check FAILED: realized correlation -0.07047584804559227 against fixed absolute 0.06. The failed assertion, initial stopped pass and unchanged arrays remain. No subsequent study repairs that check.

The original determinant guard falsely accepted 18/180 lossy requests. An isolated supplied-contract guard rejected all 180 historical and 36 fresh lossy cases and passed 27 unit tests. It neither learned a loss contract nor improved forward prediction. The live app was not repaired by that isolated experiment.

The later full 15-column archive resolves the row-projection gap. Historical executable-byte linkage and original per-point predictions remain incompletely recovered. New archives do not retroactively establish that provenance.

The old edition described seeded-library testing as a proposal. That narrow question has now been executed, as section 3 explains. Calibrated familiarity, new operation-family discovery and adaptive OaK remain proposals, not completed extensions.

3 · Seeding a library: learning numbers in a known grammar

The simulator supplies identity 2D coordinates, symbols T/S/C, affine family and left-to-right execution. T rotates 35° ; $S(x,y)=(x+0.45y+0.2,y-0.15)$; $C(x,y)=(x,0)$. Three six-parameter maps are learned by OLS from 192 primitive pairs each, without truth initialization. The 18 coefficients are genuinely learned; the grammar and architecture are supplied.

Seeds 3001–3032 form 32 independent synthetic replicates. Clean and input/target-noisy $\sigma=0.03$ conditions share latent states. Whole endpoint words TS/ST/TCS are omitted from direct training; their motifs may occur in other words. Six/eight-operation endpoints are absent. Test states are independent and exactly disjoint from training, with in-domain and fourfold-range-shift conditions.

Clean library mean RMSE is below 6×10^{-15} throughout. Under noise, six-step TSTSTS library RMSE rises from 0.048429 in-domain to 0.103017 shifted. The supplied true-map noisy-input reference is 0.038934/0.038861, not a proven optimal predictor. The lossy eight-step chain can be forward-predictable without being invertible.

The first sequence-conditioned comparator gets 576 primitive plus 4,320 endpoint pairs, versus 576 for the library; it has 390 prediction coefficients. Its fixed feature matrix has rank 41/65 and approximation error even on clean seen TT (RMSE 0.005657). Unsupported long-position features make it inadequate for a general superiority claim. Its large errors are not evidence that composition is universally better.

All 28 cells use one Bonferroni t family, df31. No post-score model revision occurred. A noisy TT in-domain contrast slightly favors the fixed comparator; shifted TT is unresolved. Supplied loss guards reject all 256 lossy evaluations, including 128 noisy apparently-full-rank fits. This is not learned reconstruction.

The result is feasibility of parameter reuse within a known affine family. It cannot establish discovery of representations, skills, operation families, or an unbounded compositional agent.

4 · A comparator that shares the prior passes the gate

An adequate follow-up compares two ways to learn the same 18-parameter recurrence. Primitive training uses 176 pairs for each generator. Endpoint training uses 16 pairs for each of 33 length-two/three words, excluding TS/ST/TCS whole words. Both receive 528 pairs and 2,112 recorded scalar slots, but different states and label types. Neither receives the other route's supervision.

Endpoint fitting uses analytic-Jacobian least squares with three train-only starts: identity and two fixed perturbations. It chooses the smallest training residual, never test loss. Clean calibration seeds 4101–4104 are separate from evaluation 5101–5132. The frozen gate checks local Jacobian rank 18 and finite-support prediction error below 10^{-7} on 81 length-four words plus seven test words at four states. Observed maximum is 2.22×10^{-16} . This is finite/local adequacy, not a global theorem.

Both clean methods are essentially exact. With input/target noise $\sigma=0.03$, endpoint mean RMSE is lower in all 14 cells; 13 intervals in the shared 28-cell family exclude zero. ST in-domain is unresolved. Six-step shifted TSTSTS gives primitive/endpoint 0.105837/0.058033, gap -0.047804 [$-0.079866, -0.015743$]. Eight-step lossy shifted gives 0.061801/0.032839, gap -0.028962 [$-0.045266, -0.012658$].

The corrected interpretation is supervision-route comparison within a supplied compositional family, NOT composition versus noncomposition. The adequate comparator removes the original feature ceiling by sharing the recurrence. Because designs, seeds and budgets differ, this is not a controlled attribution of the entire earlier weak-comparator gap.

Clean sub- 10^{-14} interval exclusions should not be presented as meaningful accuracy improvements. Equal pairs and parameters do not equalize information or optimization work. Appendix tables preserve every clean/noisy cell and the unresolved outcome.

5 · Input noise is not necessary for the advantage

The channel ablation keeps the adequate estimators and supplied recurrence unchanged. Fresh seeds 6101–6132 use input-only (0.03,0), target-only (0,0.03), or both (0.03,0.03) training noise. Methods share evaluation states, while method-specific training states differ. Clean and noisy evaluation inputs use the same underlying truth; clean-input evaluation isolates estimation error from added evaluation-input propagation.

Endpoint mean forward RMSE is lower in all 84 cells. Of the 84 primary contrasts, 77 simultaneous intervals exclude zero. All 42 clean-input evaluation cells favor endpoint, including all 14 target-only cells. Noisy evaluation has 13/14, 10/14 and 12/14 resolved contrasts for input, target and both respectively.

Target-only shifted TSTSTS with clean evaluation: primitive 0.064239, endpoint 0.027011, gap -0.037228 [$-0.055500, -0.018955$]. With noisy evaluation: 0.076561/0.048380, gap -0.028182 [$-0.043805, -0.012558$]. Noisy training inputs are therefore not necessary for the measured advantage; this does not prove input bias absent or identify a universal causal mechanism.

All 84 route gaps and 56 conditional changes in gaps belong to one frozen 140-contrast Bonferroni t family, df31. Adding target noise to input-only training resolves 0/28 interactions. Adding input noise to target-only resolves 14/28 toward endpoint; the rest are unresolved, not equivalent. These are conditional additions, not an additive decomposition of error.

The seven unresolved noisy-evaluation ID cells are input TS; target TT/TS/ST/TCS; both TS/ST. No seed was excluded, retuned or replaced. The prior clean calibration is reused and charged as historical cost, not new evidence. The exact copied estimator and source records are retained in the package.

6 · The identical-data control and local information

A fresh target-only $\sigma=0.03$ study uses seeds 7101–7132. Primitive OLS and a joint length-one solver receive EXACTLY the same primitive observations and parameterization. The third route fits length-two/three endpoints. All use the same supplied 18-parameter recurrence and 528 pairs. Evaluation has 96 states per word/range with clean inputs and clean latent targets.

The length-one control agrees with OLS: maximum coefficient difference 1.4544×10^{-14} and forward prediction difference 2.3048×10^{-13} , below frozen limits 10^{-9} and 10^{-8} . Changing solver route on identical observations does not explain the endpoint gap here. Different endpoint observations remain different information.

At truth: $F = J^T J / \sigma^2$; $V = \sigma^2 (J^T J)^{-1}$; predicted task MSE = $\text{mean tr}(G V G^T) / 2$

J stacks training sensitivities; G differentiates the tested word at each test state. OLS covariance is exact conditional on its linear fixed design. Joint endpoint covariance and composed forward risk are first-order nonlinear approximations. Truth is used for this diagnostic only, never for learner initialization. Rank threshold and calculation protocol were frozen before scoring.

All 64 truth-design matrices have rank 18. The endpoint-minus-primitive information matrix has a negative minimum eigenvalue for 2/32 seeds, so there is no universal directionwise dominance. Lower projected task risk, not a single information trace, is the relevant comparison.

The three-start optimizer and prior calibration are unchanged. Endpoint starts cost more than OLS. The diagnostic does not estimate every curvature/bias effect, establish optimal experimental design, or measure causal mediation.

7 · Compatible with a first-order account—not explained away

Actual endpoint mean forward MSE is lower in all 14 cells, with all 14 adjusted intervals excluding zero. The local information projection predicts lower endpoint risk in all 14, also with interval exclusions. All 14 actual-minus-predicted gap intervals contain zero. These 42 contrasts form one frozen Bonferroni family, df31.

A residual interval containing zero does not prove equivalence or that the whole causal mechanism is known. There was no prospective practical-equivalence margin or adequate-power claim. Ratios of predicted to observed mean gaps range 0.812–1.289; they are not percentages causally explained.

TSTSTS, shifted · MSE	Primitive OLS	Endpoint
Actual	0.003777951	0.001011605
First-order prediction	0.004493499	0.000979068
Actual endpoint–OLS gap	–0.002766346	CI [–0.004460534, –0.001072157]
Residual gap	+0.000748086	CI [–0.000916242, +0.002412413]

This control sequence narrows the story: (i) a weak fixed-feature comparison cannot support general composition benefit; (ii) input measurement noise is not necessary; (iii) solver choice on identical primitive records agrees; (iv) different training sensitivities predict the observed advantage at first order. None of these claims licenses architecture discovery.

RMSE from earlier studies and MSE from this information study are distinct endpoints. Squaring a seed-mean RMSE does not reproduce mean MSE. Tables are not pooled across different seed cohorts or multiplicity families.

Next-step falsification is a proposal: commit new noise levels, nonlinear families or independent task populations before scoring. Detectable residual departures would challenge the local account. No such follow-on was executed in this synthesis.

8 · FD001: transfer to an external simulator task

NASA's public C-MAPSS FD001 release supplies 100 complete run-to-failure training trajectories, one operating condition and one HPC fault. A deterministic hash split assigns 60 engines to fitting, 20 to calibration, 20 to evaluation. Official test trajectories and RUL labels are not loaded. This is an external simulator dataset, not deployed real-world evidence or replication of endpoint-versus-primitive fitting.

Target sensor 11 is chosen by fit-engine median absolute correlation with cycle. Means, scales, PCA and channel ranking are fitting-only. Predictors use current target, previous target and current cross-sensor values to fit the next target increment. Engine identity, elapsed cycle, lifetime and future measurements are not predictor features. Feature selection is supplied procedure, not learned architecture discovery.

Compact chooses sensors 11/17/3/8. Full ridge uses all 21 channels plus lagged target and intercept (23 coefficients). Structured full input uses three fit-only PCs plus current/previous target and intercept (six coefficients). Compact has six; target-only has three. Eight seeded four-channel random subsets are assessed on the same engines, not eight independent datasets.

Calibration chooses one penalty from $\{0.0001, 0.01, 1\}$; all selected penalties are 0.0001. There is no calibration refit. The noninferiority allowance is 10% of calibration full-ridge MSE, fixed before scoring. Four paired engine-mean contrasts use Bonferroni t intervals, df19. The 4,055 chronological predictions are dependent within engines: $n=20$, not 4,055.

Valid code/data/calibration freeze is 10:17:42 UTC, 3 October 2026. Locally available evaluation bytes make protection procedural, not cryptographic blinding. Complete histories and shared simulator construction limit exchangeability and early-life deployment generalization. No adaptation, active sensing, RUL or uncertainty-calibration claim follows.

9 · Small coefficients are not few measurements

Method	Input channels	Coefficients	Engine-mean MSE
Full ridge	21	23	0.0114382
Full input · three PCs	21	6	0.0114759
Selected compact	4	6	0.0131821
Random-subset mean	4	6	0.0132334
Target history	1	3	0.0146315
Persistence	1	0	0.0201168

Compact/full mean gap is +0.001743863 [0.001382841,0.002104884], entirely above the frozen allowance 0.001177925. The narrow noninferiority criterion FAILED, rather than simply being inconclusive. The 15.25% relative error penalty is descriptive engine-mean MSE, not an increase in every engine's risk by exactly that percentage.

Compact minus structured full is +0.001706143 [0.001335056,0.002077231]. Compact minus random-subset mean is −0.000051303 [−0.000341183,+0.000238577], unresolved—not equivalence or a proven selection benefit. Compact minus persistence is −0.006934686 [−0.007823914,−0.006045457].



The three-PC full-input model has near-full point error with the same coefficient count as compact. No structured-versus-full noninferiority contrast was prespecified: near-full here is descriptive, not proven equivalence. The decisive distinction is model parameterization versus access to information.

Four rather than 21 channels implies 81% fewer hypothetical stream reads. Every channel was actually loaded offline, and the shared compact feature function normalizes all 21. Neither physical acquisition savings nor an optimized four-sensor runtime was measured.

10 · Costs, failures and audits are part of the result

Study	First-batch work / time	Material inequality
Seeding	3,137 fit/setup calls; 2.896 s	576 vs 4,896 pairs; weak comparator
Adequate comparator	40.513 s; OLS 0.0154 s; endpoint 4.136 s	192 endpoint starts + 12 historical calibration
Channel ablation	9.282 s; OLS 0.0200 s; endpoint 7.005 s	288 starts; reused calibration charged
Information design	42.521 s; OLS 0.0073 s; endpoint 2.366 s	96 length-one starts also take 0.349 s
FD001	Preparation 0.670 s; evaluation 0.053 s	36 failed + 36 repaired + 36 replay fits

Times are machine-specific and differently scoped. They are not matched FLOPs, memory, complete research labor or portable speedup. FD001 evaluation time precedes serialization; failed-preparation, download and other unretained overhead are unknown, not zero. The whole study loaded 433,251 sensor scalars across 20,631 cycles.

FD001's first preparation produced NaN correlation from sensors constant within some engines, before protected scoring. Original code/log are retained; assigning zero correlation to within-engine constants amended the protocol before valid freeze. That attempt completed 36 fits. Counting only the successful 36 hides real work; one replay brings the total to 108.

Completed stored-evidence audits report 3,584 seeding RMSEs; 2,688 comparator RMSEs; 8,064 channel RMSEs; 2,516 information metric/training/interval checks plus matrix checks; and 65,398 FD001 checks. This edition reads their receipts and summaries; it does not rerun fitting or all prior audits.

Exact replay counts—13,507, 7,048, 14,144 and 6,950 arrays for the first four studies, and 19 arrays/36 coefficient vectors for FD001—establish same-implementation reproducibility, not independent validation. The old covariance failure and separate native-app compiler failures remain unresolved.

11 · Claim, evidence and falsifier map

Claim / status	Actual evidence	What would challenge it or what is missing
Known-family parameter reuse · supported narrowly	Clean held-out whole words/chains; noisy error grows	New families/coordinates; learn grammar rather than supply it
General composition superiority · NOT established	Weak fixed-feature comparator invalid; adequate endpoint shares prior	Adequate noncompositional comparator with declared access/cost
Input noise necessary · rejected here	Target-only advantage, 14/14 clean cells	Conditional result; other mechanisms not ruled out
Solver route explains gap · not here	Identical-observation length-one agreement	Other solvers/noise/families need fresh controls
First-order information account · compatible	14/14 predicted directions; 0/14 residual exclusions	Prospective residual/equivalence test, larger noise/nonlinearity
Four-sensor tolerance · failed on FD001	Adjusted lower gap exceeds frozen allowance	New target/split/task must be frozen; do not retune these engines
Compact coefficients imply sparse access · false implication	Six-coefficient full input is distinct from six-coefficient compact	Measure acquisition/runtime with only admitted inputs
Inverse recovery · impossible in specified collision	C erases a coordinate; supplied guard prevents false requests	Side information/domain restrictions would change the task
OaK/adaptive architecture · proposed	None of these studies learns representation/skill grammar	Continual feature/subtask/option/model/planning experiment

This map is not a declaration of universal truths from finite data. Evidence and rejection apply to the stated protocol. A new split, changed target or different budget would be a new study; it cannot silently replace a failed endpoint.

Independent replication, prospective information-cost comparisons, a verified learned loss guard and a full continual OaK agent remain possible research directions. None was run or launched by this revision.

12 · Limits, literature and conclusion

The affine studies share one supplied family, coordinates and recurrence. Their whole-word holdouts may contain familiar motifs, and all operation families are known. Synthetic replicate seeds are independent within each declared cohort, while cross-cell observations are paired. Approximate t intervals quantify seed-level variation under assumptions; they do not include development/model-selection uncertainty, prove normality or establish equivalence.

FD001 adds a simulator population, not physical-world deployment. One target, subset and split cannot establish a general sensor-reduction law. Engine exchangeability is assumed despite shared construction. A valid point result is not evidence for active acquisition, adaptation or RUL estimation. Complete histories differ from early-censored operational monitoring.

Classical background retained from the earlier literature ledger: Strang, Column Space and Nullspace (MIT OCW); Littman, Sutton & Singh, Predictive Representations of State (NeurIPS 2001); Li, Walsh & Littman, Towards a Unified Theory of State Abstraction for MDPs (ISAIM 2006). These motivate the distinctions; none guarantees our fitted models. This synthesis does not claim a new full-text literature review.

NASA source: Saxena, Goebel, Simon & Eklund, Damage Propagation Modeling for Aircraft Engine Run-to-Failure Simulation (PHM08), public PCoE repository. The retained source readme's sensor endpoint is a typo: the files contain 21 sensor columns, 26 total columns. Original download page and archive hashes remain in FD001 evidence.

Conclusion

A model can calculate consistently without predicting well; predict well without reconstructing; or use few coefficients while still requiring every sensor. The completed controls favor a narrower explanation than architecture triumph: supervision design carries task-relevant information, and restricting access can destroy it. Information sufficiency must be tested at the task endpoint, with losses, costs and comparator adequacy made explicit.

The defensible contribution of this revision is a measured boundary map and a counterexample, not a new universal architecture or proof of general intelligence. Original papers, the learner book, apps and websites are unchanged.

13 · Reproducibility and reading guide

This edition was generated from inspected report text, original editable manuscript, frozen study/protocol code, retained aggregate JSON and audit/cost receipts. It never executes scientific fitting code. Latest tables are populated mechanically from tables-source.json copied from study summaries, not retyped from this brief. Display rounding is separate from the exact retained values.

The editable package includes manuscript.json and manuscript.md, render.py, vector figures, all latest aggregate tables, source inventory, report PDFs/text, selected protocols/implementation/audit context, and the previous paper/editable ZIP unchanged. It is an evidence-synthesis package, NOT a replacement for the complete raw-array experiment archives. Full replay still requires the previously delivered experiment archives and seeding raw shards; archived source programs are included for inspection, not executed.

Source association: seeding report attachment [private record withheld] ; comparator
[private record withheld] ; channel ablation [private record withheld] ; information design
[private record withheld] ; FD001 [private record withheld] and text
[private record withheld] . Previous edition: [private record withheld] and source
[private record withheld] . Input hashes are computed and retained in association receipts; filenames alone are not proof of identity.

Full raw archive handles: seeding main [private record withheld] plus four shards
[private record withheld] , [private record withheld] , [private record withheld] ,
[private record withheld] ; comparator [private record withheld] ; channel
[private record withheld] ; information [private record withheld] plus replay
[private record withheld] ; FD001 [private record withheld] . These are referenced, not duplicated in this smaller package.

Appendices preserve all latest mean/contrast rows, including unresolved cases and numerical-scale clean values. The channel study’s 56 interaction contrasts are retained in its full report and JSON rather than omitted from the evidence package. No cross-study meta-analysis or pooled significance family is implied.

Exact verification and page-inspection receipts are delivered alongside the PDF. Checksums secure present files, not historical scientific authorship or unseen source-to-run provenance.

Appendix A · Seeding: all 28 cells

Seed-mean forward RMSE; n=32; 28-contrast Bonferroni family. Positive D–L favors library. The fixed direct baseline is inadequate, especially on long words. Clean gaps cannot establish a general composition benefit.

Noise	Word/range	Library RMSE	Fixed direct	D–L adjusted CI
0	TT / in-domain	4.59901e-16	0.00565713	[0.00557501, 0.00573925]
0	TT / shifted	1.54829e-15	0.0225977	[0.0222761, 0.0229194]
0	TS / in-domain	4.9386e-16	0.0125474	[0.0123904, 0.0127045]
0	TS / shifted	1.62653e-15	0.0394488	[0.0387111, 0.0401866]
0	ST / in-domain	4.90497e-16	0.00989784	[0.00970024, 0.0100954]
0	ST / shifted	1.60265e-15	0.0352847	[0.0346654, 0.035904]
0	TCS / in-domain	5.01164e-16	0.0916632	[0.0898805, 0.0934459]
0	TCS / shifted	1.66498e-15	0.365764	[0.359807, 0.371721]
0	TSTSTS / in-domain	1.65003e-15	0.260988	[0.256588, 0.265389]
0	TSTSTS / shifted	5.05479e-15	0.903889	[0.887691, 0.920088]
0	SSTTSS / in-domain	1.90354e-15	0.280452	[0.275131, 0.285772]
0	SSTTSS / shifted	5.8224e-15	0.768714	[0.753333, 0.784096]
0	TSTCSTST / in-domain	1.3641e-15	1.54708	[1.51782, 1.57634]
0	TSTCSTST / shifted	3.70825e-15	5.89005	[5.78663, 5.99346]
0.03	TT / in-domain	0.0314819	0.0308996	[-0.00115815, -6.41122e-06]
0.03	TT / shifted	0.0444989	0.0433282	[-0.00766699, 0.00532563]
0.03	TS / in-domain	0.0322004	0.0347131	[0.00162129, 0.00340397]
0.03	TS / shifted	0.0420818	0.0575646	[0.00741765, 0.0235479]
0.03	ST / in-domain	0.0324204	0.0340685	[0.000779489, 0.00251672]
0.03	ST / shifted	0.0418268	0.0556109	[0.00695819, 0.02061]
0.03	TCS / in-domain	0.0225919	0.0936368	[0.0676093, 0.0744806]
0.03	TCS / shifted	0.0333419	0.360142	[0.314248, 0.339353]
0.03	TSTSTS / in-domain	0.048429	0.265781	[0.205322, 0.229382]
0.03	TSTSTS / shifted	0.103017	0.906853	[0.75751, 0.850163]
0.03	SSTTSS / in-domain	0.0512624	0.285172	[0.219624, 0.248195]
0.03	SSTTSS / shifted	0.103663	0.785482	[0.62451, 0.739128]
0.03	TSTCSTST / in-domain	0.030517	1.54355	[1.48054, 1.54552]
0.03	TSTCSTST / shifted	0.062747	5.85872	[5.66499, 5.92696]

Appendix B · Adequate comparator: $\sigma=0$

n=32; means of forward RMSE, not pooled RMSE. All clean/noisy cells share one 28-contrast family. Negative E–P favors endpoint. Near-arithmetic clean differences have no substantive accuracy interpretation.

Word/range	Primitive RMSE	Endpoint RMSE	E–P adjusted CI
TT / ID	8.05408e-16	6.95567e-18	[-1.71725e-15, 1.20349e-16]
TT / shift	2.90754e-15	3.5454e-18	[-6.30552e-15, 4.97527e-16]
TS / ID	6.80162e-16	1.07883e-17	[-1.30014e-15, -3.86077e-17]
TS / shift	2.33557e-15	1.30906e-17	[-4.55101e-15, -9.3959e-17]
ST / ID	6.64544e-16	1.41107e-17	[-1.20892e-15, -9.19501e-17]
ST / shift	2.26588e-15	1.52803e-17	[-4.11448e-15, -3.86709e-16]
TCS / ID	4.9726e-16	1.12789e-17	[-8.82869e-16, -8.90928e-17]
TCS / shift	1.72693e-15	2.73056e-17	[-3.24034e-15, -1.58915e-16]
TSTSTS / ID	2.22305e-15	3.05874e-17	[-4.39434e-15, 9.41066e-18]
TSTSTS / shift	7.40617e-15	4.72722e-17	[-1.48149e-14, 9.7096e-17]
SSTTSS / ID	2.29111e-15	5.15124e-17	[-3.99144e-15, -4.87762e-16]
SSTTSS / shift	7.33851e-15	1.08043e-16	[-1.33887e-14, -1.07228e-15]
TSTCSTST / ID	1.50302e-15	7.4824e-18	[-2.59586e-15, -3.95219e-16]
TSTCSTST / shift	3.979e-15	1.60512e-17	[-6.39923e-15, -1.52667e-15]

Appendix B · Adequate comparator: $\sigma=0.03$

n=32; means of forward RMSE, not pooled RMSE. All clean/noisy cells share one 28-contrast family. Negative E–P favors endpoint. Near-arithmetic clean differences have no substantive accuracy interpretation.

Word/range	Primitive RMSE	Endpoint RMSE	E–P adjusted CI
TT / ID	0.0320486	0.0310609	[-0.00165514, -0.000320273]
TT / shift	0.0459372	0.0381993	[-0.0149797, -0.000496123]
TS / ID	0.0321318	0.0313178	[-0.00142576, -0.000202112]
TS / shift	0.0437326	0.0350161	[-0.0142807, -0.00315221]
ST / ID	0.0320614	0.0314834	[-0.0012221, 6.62157e-05]
ST / shift	0.0431906	0.0344125	[-0.0138605, -0.00369573]
TCS / ID	0.0226737	0.0215208	[-0.00200078, -0.000305098]
TCS / shift	0.0350183	0.0245545	[-0.015693, -0.00523453]
TSTSTS / ID	0.048348	0.0411522	[-0.0115183, -0.00287341]
TSTSTS / shift	0.105837	0.0580329	[-0.0798661, -0.0157429]
SSTTSS / ID	0.0510967	0.0435249	[-0.0125475, -0.00259621]
SSTTSS / shift	0.109314	0.0606152	[-0.0782413, -0.0191566]
TSTCSTST / ID	0.0295787	0.0230351	[-0.00983106, -0.00325614]
TSTCSTST / shift	0.0618013	0.0328392	[-0.0452664, -0.0126577]

Appendix C · input-channel training noise

n=32; all 84 route contrasts plus 56 interactions share the 140-contrast family. Clean/noisy below denotes evaluation INPUT, not training target. Negative E–P favors endpoint. Interactions remain in original report/JSON.

Eval · word/range	Primitive RMSE	Endpoint RMSE	E–P adjusted CI
clean · TT/ID	0.0079702	0.00421933	[-0.00530367, -0.00219808]
clean · TT/shift	0.0279529	0.0140138	[-0.0208788, -0.00699936]
clean · TS/ID	0.00658789	0.0031958	[-0.00523131, -0.00155288]
clean · TS/shift	0.0235379	0.0105121	[-0.0207458, -0.00530594]
clean · ST/ID	0.00659081	0.0032114	[-0.00527181, -0.001487]
clean · ST/shift	0.023409	0.00988552	[-0.02125, -0.00579684]
clean · TCS/ID	0.00594369	0.00257247	[-0.0047793, -0.00196316]
clean · TCS/shift	0.0210288	0.00878335	[-0.0174041, -0.00708685]
clean · TSTSTS/ID	0.0239961	0.0101508	[-0.0216934, -0.00599736]
clean · TSTSTS/shift	0.0828476	0.0339903	[-0.0788434, -0.0188711]
clean · SSTTSS/ID	0.0253852	0.0108487	[-0.0229347, -0.00613818]
clean · SSTTSS/shift	0.0810202	0.0353844	[-0.0773279, -0.0139438]
clean · TSTCSTST/ID	0.0168159	0.00626088	[-0.0160546, -0.00505552]
clean · TSTCSTST/shift	0.0508882	0.0188947	[-0.0466007, -0.0173863]
noisy · TT/ID	0.0311942	0.0304832	[-0.00140008, -2.19785e-05]
noisy · TT/shift	0.0411874	0.0331922	[-0.0123835, -0.00360698]
noisy · TS/ID	0.032475	0.0322093	[-0.000786222, 0.000254962]
noisy · TS/shift	0.0399488	0.0331706	[-0.0112676, -0.00228888]
noisy · ST/ID	0.0325682	0.0323003	[-0.000532758, -3.01933e-06]
noisy · ST/shift	0.0391515	0.0331124	[-0.01024, -0.00183822]
noisy · TCS/ID	0.0219673	0.0211894	[-0.00129474, -0.000260998]
noisy · TCS/shift	0.0303056	0.0231525	[-0.0106729, -0.00363332]
noisy · TSTSTS/ID	0.0463552	0.040261	[-0.0109036, -0.00128478]
noisy · TSTSTS/shift	0.0920571	0.0523076	[-0.0659312, -0.0135677]
noisy · SSTTSS/ID	0.0484011	0.042591	[-0.0101189, -0.00150129]
noisy · SSTTSS/shift	0.0939539	0.0555942	[-0.0666225, -0.0100969]
noisy · TSTCSTST/ID	0.027605	0.0225148	[-0.00822946, -0.00195078]
noisy · TSTCSTST/shift	0.0557513	0.029586	[-0.0390785, -0.0132521]

Appendix C · target-channel training noise

n=32; all 84 route contrasts plus 56 interactions share the 140-contrast family. Clean/noisy below denotes evaluation INPUT, not training target. Negative E–P favors endpoint. Interactions remain in original report/JSON.

Eval · word/range	Primitive RMSE	Endpoint RMSE	E–P adjusted CI
clean · TT/ID	0.00678805	0.00431317	[-0.00422953, -0.00072024]
clean · TT/shift	0.0219771	0.0144116	[-0.0138057, -0.00132525]
clean · TS/ID	0.00584439	0.00280788	[-0.00419979, -0.00187322]
clean · TS/shift	0.0186254	0.00920562	[-0.0139744, -0.00486517]
clean · ST/ID	0.00561229	0.00278087	[-0.00425146, -0.00141139]
clean · ST/shift	0.0179775	0.00883819	[-0.0138394, -0.00443912]
clean · TCS/ID	0.00582647	0.00311828	[-0.00416179, -0.00125459]
clean · TCS/shift	0.0178076	0.0101135	[-0.0125604, -0.00282797]
clean · TSTSTS/ID	0.0197826	0.00858984	[-0.0154444, -0.0069411]
clean · TSTSTS/shift	0.0642386	0.027011	[-0.0555, -0.0189553]
clean · SSTTSS/ID	0.0195109	0.00855872	[-0.0149482, -0.00695617]
clean · SSTTSS/shift	0.0599519	0.0272171	[-0.0482211, -0.0172484]
clean · TSTCSTST/ID	0.014887	0.00650117	[-0.0126381, -0.00413359]
clean · TSTCSTST/shift	0.0383083	0.0192526	[-0.0329329, -0.0051784]
noisy · TT/ID	0.0309647	0.0306972	[-0.000821754, 0.000286678]
noisy · TT/shift	0.037703	0.0331256	[-0.00833811, -0.000816643]
noisy · TS/ID	0.0324438	0.0321627	[-0.000651856, 8.97413e-05]
noisy · TS/shift	0.0366817	0.0329417	[-0.00641448, -0.00106551]
noisy · ST/ID	0.0325313	0.0322998	[-0.000659569, 0.000196548]
noisy · ST/shift	0.0359545	0.0324466	[-0.0057735, -0.00124248]
noisy · TCS/ID	0.0217812	0.0214359	[-0.000873108, 0.000182375]
noisy · TCS/shift	0.0279323	0.0234213	[-0.00760662, -0.00141542]
noisy · TSTSTS/ID	0.0439707	0.0400883	[-0.00592714, -0.00183779]
noisy · TSTSTS/shift	0.0765613	0.0483795	[-0.0438055, -0.0125581]
noisy · SSTTSS/ID	0.0459942	0.0421982	[-0.00620586, -0.00138614]
noisy · SSTTSS/shift	0.074156	0.0505467	[-0.0363484, -0.0108701]
noisy · TSTCSTST/ID	0.0266746	0.0229063	[-0.00601534, -0.00152135]
noisy · TSTCSTST/shift	0.0450041	0.0298017	[-0.027182, -0.00322276]

Appendix C · both-channel training noise

n=32; all 84 route contrasts plus 56 interactions share the 140-contrast family. Clean/noisy below denotes evaluation INPUT, not training target. Negative E–P favors endpoint. Interactions remain in original report/JSON.

Eval · word/range	Primitive RMSE	Endpoint RMSE	E–P adjusted CI
clean · TT/ID	0.010962	0.00572672	[-0.00705991, -0.00341065]
clean · TT/shift	0.0371885	0.0203961	[-0.0229748, -0.01061]
clean · TS/ID	0.00868485	0.00411209	[-0.0061518, -0.00299371]
clean · TS/shift	0.0296828	0.0138033	[-0.0235094, -0.00824951]
clean · ST/ID	0.00857793	0.00401866	[-0.00591955, -0.003199]
clean · ST/shift	0.0289947	0.0126335	[-0.0234712, -0.00925112]
clean · TCS/ID	0.00855703	0.00386681	[-0.00671262, -0.00266781]
clean · TCS/shift	0.0289664	0.0127984	[-0.0225766, -0.00975945]
clean · TSTSTS/ID	0.0303964	0.0126075	[-0.025235, -0.0103429]
clean · TSTSTS/shift	0.0991466	0.042316	[-0.0905979, -0.0230632]
clean · SSTSS/ID	0.0302134	0.0129952	[-0.0253643, -0.00907217]
clean · SSTSS/shift	0.0944497	0.0414788	[-0.0868889, -0.0190529]
clean · TSTCSTST/ID	0.0233692	0.00875515	[-0.0202666, -0.0089616]
clean · TSTCSTST/shift	0.0674644	0.0251985	[-0.0592902, -0.0252415]
noisy · TT/ID	0.0320505	0.0308988	[-0.00209719, -0.000206113]
noisy · TT/shift	0.0481167	0.0361736	[-0.0168141, -0.00707207]
noisy · TS/ID	0.032924	0.0324074	[-0.00105013, 1.69291e-05]
noisy · TS/shift	0.043611	0.0344772	[-0.0143425, -0.00392512]
noisy · ST/ID	0.0329155	0.0324331	[-0.000970095, 5.31994e-06]
noisy · ST/shift	0.0423264	0.0338326	[-0.0127276, -0.00425996]
noisy · TCS/ID	0.0226479	0.0214617	[-0.00197063, -0.00040187]
noisy · TCS/shift	0.0362861	0.0250256	[-0.0163525, -0.00616836]
noisy · TSTSTS/ID	0.0499415	0.0411982	[-0.01385, -0.0036365]
noisy · TSTSTS/shift	0.106911	0.058046	[-0.0796128, -0.0181178]
noisy · SSTSS/ID	0.0513623	0.0434553	[-0.0131343, -0.00267973]
noisy · SSTSS/shift	0.105695	0.0599432	[-0.0759877, -0.0155152]
noisy · TSTCSTST/ID	0.0320192	0.0235224	[-0.0124784, -0.00451527]
noisy · TSTCSTST/shift	0.0708757	0.0339491	[-0.0528186, -0.0210346]

Appendix D · Information design: all 14 MSE cells

Coordinate-averaged MSE, NOT mean RMSE squared. n=32; three contrasts per cell form one 42-family correction. Truth-local predictions do not enter learner fitting.

Word/range	Actual OLS / E	Predicted OLS / E	Actual E–OLS CI
TT/ID	5.11467e-05 / 1.95664e-05	5.48246e-05 / 2.02654e-05	[-5.03588e-05, -1.28017e-05]
TT/shift	0.00050579 / 0.000221482	0.000596474 / 0.000229878	[-0.000510899, -5.77161e-05]
TS/ID	3.46579e-05 / 9.20985e-06	3.33014e-05 / 8.52514e-06	[-3.70711e-05, -1.3825e-05]
TS/shift	0.000321367 / 9.82336e-05	0.000359899 / 9.4028e-05	[-0.000335385, -0.000110882]
ST/ID	3.51107e-05 / 9.00553e-06	3.3027e-05 / 8.43719e-06	[-3.55913e-05, -1.6619e-05]
ST/shift	0.000346762 / 9.498e-05	0.00034585 / 9.05385e-05	[-0.000379087, -0.000124476]
TCS/ID	3.61196e-05 / 1.02555e-05	3.48336e-05 / 1.02206e-05	[-4.21145e-05, -9.61387e-06]
TCS/shift	0.0004051 / 0.000118165	0.000360703 / 0.000114652	[-0.000488783, -8.50856e-05]
TSTSTS/ID	0.000398394 / 9.68046e-05	0.000388799 / 8.39161e-05	[-0.000434466, -0.000168713]
TSTSTS/shift	0.00377795 / 0.00101161	0.0044935 / 0.000979068	[-0.00446053, -0.00107216]
SSTTSS/ID	0.000432071 / 0.000109256	0.000408062 / 8.65845e-05	[-0.00048547, -0.000160161]
SSTTSS/shift	0.00420216 / 0.00118938	0.00434761 / 0.000953444	[-0.00505625, -0.000969306]
TSTCSTST/ID	0.000243894 / 4.18482e-05	0.000215197 / 4.34598e-05	[-0.000336079, -6.80122e-05]
TSTCSTST/shift	0.00188746 / 0.00030946	0.00165003 / 0.000368146	[-0.00259832, -0.000557686]

Appendix D · Residual gaps: all 14 cells

Residual=(actual E–OLS MSE gap)–(predicted E–OLS gap). All intervals include zero: compatibility, not equivalence. The 42-family correction applies.

Word/range	Actual–predicted gap	Adjusted CI
TT/ID	2.97903e-06	[-1.57127e-05, 2.16707e-05]
TT/shift	8.22888e-05	[-0.000139014, 0.000303591]
TS/ID	-6.71791e-07	[-1.23386e-05, 1.09951e-05]
TS/shift	4.27372e-05	[-7.30939e-05, 0.000158568]
ST/ID	-1.51528e-06	[-1.1055e-05, 8.02448e-06]
ST/shift	3.52997e-06	[-0.000122911, 0.000129971]
TCS/ID	-1.25123e-06	[-1.74e-05, 1.48976e-05]
TCS/shift	-4.08827e-05	[-0.000240478, 0.000158712]
TSTSTS/ID	3.29322e-06	[-0.000127587, 0.000134174]
TSTSTS/shift	0.000748086	[-0.000916242, 0.00241241]
SSTTSS/ID	-1.33762e-06	[-0.000162217, 0.000159541]
SSTTSS/shift	0.000381393	[-0.00160489, 0.00236768]
TSTCSTST/ID	-3.03087e-05	[-0.000163467, 0.00010285]
TSTCSTST/shift	-0.000296118	[-0.00131712, 0.000724881]