



American Journal of Applied Research and AI (AJARAI)

VOLUME 1 ISSUE 2 (2026)



PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

From Supported Transitions to Carried State a Controlled Ledger Evaluation

Ayush Ojha^{*}

Article Information

Received: August 09, 2026**Accepted:** October 05, 2026**Published:** October 24, 2026

Keywords

*Accumulated State, Conservation,
Ledger Evaluation, Representation,
Synthetic Workflow*

ABSTRACT

When does one-step prediction support accurate accumulated state? We compare five small multilayer perceptron representation packages on a synthetic ledger controller with all output statuses represented in training. The primary endpoint is mean bounded six-balance error at horizon eight with equal family weights. The original frozen study fits 120 models over eight corpora and three paired seeds. Canonical sign-and-observed-amount prediction reaches 99.9975% mean in-distribution exactness but only 34.24% complete-state exactness at horizon eight. Its bounded accumulated-error contrast with canonical regression is -0.1450 (97.5% interval [-0.1562, -0.1337]); the conservation-package contrast is -0.0056 (97.5% interval [-0.0163, 0.0051]). The conservation package conserves every horizon-eight state while attaining 0.09% state exactness. An operational heuristic has lower numerical error but produces capacity-invalid states. The original twelve-fit support diagnostic fails its capacity gate in all six augmented fits. A separately frozen reviewer follow-up adds 192 fits addressing parameter budgets, loss weighting, an identical-head objective, a role-equivariant architecture and matched one-step versus four-step carried training. None of the 192 follow-up fits passes the balance-flip capacity-status gate. Mathematically equivalent flattened and teacher-forced controls exhibit different recorded outcomes, limiting isolated temporal interpretation. The new oracle-state and decoder diagnostics remain separate from the original primary evidence. Results are bounded to the specified generator and training packages; they establish neither a general failure of conservation methods nor real financial utility.

INTRODUCTION

An evaluator of a learned simulator must decide whether its predicted states are useful, not merely well formed. A ledger-like transition provides a particularly explicit distinction. A posting can preserve the sum of six balances while crediting the wrong role. A last-step posting can be correct while earlier mistakes remain in the current balances. Conversely, a regression model can make useful numerical progress without reproducing every integer cent exactly. Conservation, transition exactness, and accumulated-state error therefore answer different questions.

The logical distinction alone does not establish a useful empirical result. A comparison is weak if the training task is not learned, if a supposedly compositional test introduces a previously unseen output class, or if a trivial or domain heuristic already performs better. Our question is narrower: after status support and basic fitting competence are established, how do controlled representation packages rank on one-step transitions and on their own carried state under a specified stress distribution? We compare five small neural conditions and three explicit references, retaining an adverse heuristic comparison whenever it occurs.

The experiment separates three measurements. Supported ID examples test learning within the generator's scheduled cells. Four withheld primitive/review-trigger cells test a limited recombination of supported labels and triggers. Accumulated rollouts test repeated use under a different distribution of states and actions. These are

not interchangeable definitions of out-of-distribution generalization. In particular, the rollout challenge changes initial headroom and later state occupancy, so it cannot isolate horizon alone.

Our contribution is a controlled diagnostic study and an auditable evaluation package: input-disjoint splits with full status support, a representation comparison with equal training opportunity, true model-carried rollouts, competent operational references, corpus-level contrasts, and complete raw prediction records. The empirical conclusions are reported with the limitations that make them interpretable. The study neither proposes a general-purpose architecture nor tests a language model or a deployed financial system.

Tables 1–6 contain the main and auxiliary comparisons. Tables 7–15 record controller details, support counts, component exactness, corpus contrasts, pilot history, checkpoint losses and prefix strata. Appendix B reports post-collection reviewer analyses in Tables 16–18; Appendix C reports the additional controlled training study in Tables 19–22. Appendix D adds the secondary correctness intervals in Table 23.

LITERATURE REVIEW

Conservation can be a useful inductive bias. MC-LSTM routes mass through recurrent storage and redistribution using left-stochastic matrices and separate mass and auxiliary inputs (Hoedt *et al.*, 2021, Section 2). ConCerNet learns invariant representations from trajectories, then projects a predicted continuous derivative onto the

¹ Georgia Institute of Technology, United States

^{*} Corresponding author's email: aojha47@gatech.edu

tangent space of those invariants (Zhang *et al.*, 2023, Section 3.2). Its construction differs from our supplied zero-sum invariant, discrete statuses and integer-cent decoder. Neither architecture is reproduced here. Our results therefore neither rank these methods nor contradict their reported physical-system findings; they concern one small fixed-rule ledger and its specified training and decoding packages.

Hard projection is also used in physics-informed surrogate architectures. KKT-hPINN attaches fixed layers that orthogonally project neural outputs onto a supplied system of linear equality constraints (Chen *et al.*, 2024). Our mean centering is a special zero-sum linear projection followed by integer conserving rounding. We do not reproduce KKT-hPINN, its chemical-process tasks or its broader input-output constraints; our fixed-checkpoint ablation tests only the stated ledger postprocessing package.

Arithmetic priors and induced programs offer different hypotheses about representation. Neural arithmetic logic units combine addition/subtraction and multiplicative operations through learned gates (Trask *et al.*, 2018, Section 2). DreamCoder learns a reusable program library together with a neural search policy for inductive synthesis (Ellis *et al.*, 2021, Section 2). These approaches differ from predicting ledger transitions with a fixed MLP. Neither is evaluated here: the exact rule-table controller is supplied, not induced, and the sign-and-observed-amount head uses known magnitude rather than learning general arithmetic. Our results therefore do not rank arithmetic modules or program-induction systems, and the added role-projected MLP is not a substitute for those untested baselines.

Object-oriented world modeling makes compositional generalization an explicit formal problem (Zhao *et al.*, 2022). Our four-cell primitive/trigger holdout is less general and should be read at that scope. Neural algorithmic reasoning studies demonstrate that representation, generator design and the unit of correctness can substantially affect generalization assessments (Mahdavi *et al.*, 2023). Evaluations of LLM-based world models also distinguish model components from decision tasks (Yang *et al.*, 2026). We adopt the measurement discipline without treating our MLP study as evidence about those model classes or decision benchmarks.

Related controlled financial evaluations reinforce the distinction between interface success and the endpoint of interest. A paired simulator study measures post-state consequences of fixed language-model proposals under execution-time mediation (Ojha, 2026b). A separate typed-field study isolates a value-only deterministic executor under a shared support certificate and gate (Ojha, 2026c). A payment-diagnosis study distinguishes response validity from correctness and limits its presentation-order conclusions in the presence of generator cues (Ojha, 2026a). These are different tasks and model classes. They contextualize the present separation of posting invariants, transition correctness

and accumulated state; their results are not pooled with this ledger study, and none establishes competence of the small neural conditions evaluated here.

Long-horizon prediction also depends on the relationship between training and inference states. Scheduled sampling explicitly mixes observed and model-generated previous tokens during sequence training (Bengio *et al.*, 2015, Section 2.4). This motivates separate oracle-state and model-carried diagnostics. Our additional fixed four-step feedback condition uses a declared straight-through numeric surrogate and discrete categorical choices; it does not implement a scheduled-sampling curriculum or reproduce that work's token-generation tasks.

MATERIALS AND METHODS

Controller predictions and correctness

The deterministic controller has six labeled roles: cash, receivable, customer liability, merchant payable, fee revenue, and loss expense. A state contains one slot per role, each with an integer-cent balance, absolute capacity and lock flag; it also contains a workflow stage, risk band, approval limit, sanctions flag and books-open flag. Slot order is randomly permuted when examples are generated. Initial balances are generated by eight equal-and-opposite transfers and thus sum to zero; individual balances need not be zero.

An action exposes its primitive, positive integer amount, authentication strength, and expedited flag. The eight primitives are submit, approve, capture, settle, merchant payout, fee assessment, refund, and chargeback. Submit and approve advance the stage without posting. Each other successful primitive increases one known role and decreases another by the observed amount. For example, capture increases receivable and decreases customer liability. The learned input contains role labels and the primitive, but not the target routing, output status or next stage.

The controller checks sanctions, books availability, stage validity, affected-role locks, review triggers, then capacity in that order. A high-risk expedited action or a low-authentication action exceeding the approval limit triggers review after earlier checks pass. A posting is rejected if any resulting absolute balance is at least its capacity. The four statuses are applied, review, rejected, and deferred. All non-applied transitions preserve the state. The complete primitive table is in Appendix A; an independent rule-table implementation agrees with every generated oracle transition before fitting.

Let $b_t \in \mathbb{Z}^6$ and g_t be oracle balances and stage, with u_t the current transition status. The model starts from the same state and then evolves $\hat{b}_t, \hat{g}_t$ from its own predictions. Context fields remain fixed in both paths. We measure

$$M_t = \frac{1}{6} \sum_{j=1}^6 |\hat{b}_{t,j} - b_{t,j}|, \quad B_t = \min(M_t/100000, 1),$$

$$X_t = \mathbf{1}[\hat{b}_t = b_t, \hat{g}_t = g_t, \hat{u}_t = u_t], Q_t = \prod_{s=1}^t X_s.$$

M_t is accumulated six-balance MAE in cents; B_t is the bounded primary error; X_t is full-state exactness, including

the current status; Q_t requires exactness at every prefix. Conservation is $1[\sum_i \hat{b}_{(t_i)} = 0]$ Current transition-tuple exactness compares the predicted delta, next stage and status with the oracle tuple and is reported separately. The status is included in X_t as an operational output, although it is not a persistent state field.

For example, predicting a capture posting of +a to cash and -a to customer liability conserves the total but misroutes the asset. If a later delta matches the oracle, the earlier cash/receivable discrepancy remains in M_t . This illustrative example motivates direct state measurement; it is not counted as empirical evidence.

Earlier aggregate evaluation

An earlier version of this project disclosed that review was absent from training, validation and ID data, while its conjunction test required that status. That stratum was an unsupported-label stress test, not a clean supported-label composition test. It also disclosed that a zero-delta predictor achieved aggregate OOD MAE 1176.88 cents, lower than every learned condition (best learned value 4117.09). A sign-and-observed-amount diagnostic reached 99.994% ID exactness and 79,999/80,000 exact recursive transition tuples. These historical results are retained as motivation; none is pooled into the present estimates.

Three changes follow. First, training must include every evaluated output status. Second, useful regression competence must be demonstrated against a zero-posting reference before scaling a comparison. Third, a recursive transition tuple must not be renamed accumulated-state

accuracy. The new measurement retains each model's balances and stage and compares them with the oracle's separately evolved state. The prior result and this revision answer different questions; the revision does not retrospectively turn the earlier evaluation into a stronger test.

Supported label design and prospective protocol Splits and five conditions

Each main corpus contains 30,000 training cases, 5,000 validation cases, 5,000 supported ID cases, 4,000 withheld-cell cases, and 384 length-32 trajectories. Training, validation and ID cycle over 88 eligible primitive/condition cells: the original 76 cells plus 12 supported conjunction cells. The four withheld cells are capture and merchant payout with high risk plus expedited execution, and refund and chargeback with a large amount plus low authentication. Both conjunction rules and all eight primitives occur in training. This is a supported-label primitive/trigger holdout; it does not require inventing a new status.

Canonical input fingerprints include the role-sorted initial state and complete action sequence. Duplicates are rejected within the scheduled cell, continuing its seeded generator. Train, validation, ID and withheld examples are disjoint within each corpus. Main generation excludes all 44,384 distinct development inputs. Independently seeded corpora may contain the same finite-generator inputs; they are repeated generator draws, not distinct real institutions. Slot permutations of the same underlying case cannot cross a within-corpus split.

Table 1: Controlled conditions. All use the same two-layer, width-64 ReLU backbone and observed state/action fields. Sign conditions have a larger head and a different numeric loss, so cross-head contrasts compare packages.

Condition	Slot order	Posting output	Conservation intervention
A	Random	Six continuous values	None
B	Canonical roles	Six continuous values	None
C	Random	Six three-way signs $\times$ amount	None
D	Canonical roles	Six three-way signs $\times$ amount	None
E	Canonical roles	Six continuous values	Centering + conserving decode

A–D cross role ordering with output representation (Table 1). Canonicalization sorts observed labels before encoding and maps predictions back; it uses no target labels. Regression predicts normalized continuous deltas with smooth- L_1 loss, then rounds to integer cents. Sign conditions classify each slot as negative, zero, or positive and multiply by the observed integer amount. The controller always produces deltas in this set, so this supplies a correct, task-specific output restriction. It does not enforce that exactly one sign is positive and one negative. E subtracts the mean delta during training and uses role-tied largest-remainder rounding at decoding, guaranteeing a zero-sum integer posting. B–E is a training-and-decoding conservation package comparison, not an isolated mechanistic effect of a conservation loss. All conditions predict stage and status through categorical heads, with cross-entropy losses weighted

0.5 each. Every predicted non-applied status forces zero delta and unchanged predicted stage. A, B and E have 10191 parameters; C and D have 10971. Backbone width, optimizer, learning-rate schedule, data order and training budget are matched; parameter count and numeric loss scale across output types are not. Arm order is deterministically randomized within each corpus and paired seed.

Normalization and numerical handling

All monetary input fields are divided by the fixed scale of 100,000 cents: slot balances, slot capacities, the approval limit and the action amount. Regression targets use the same scale. Role, stage, risk, primitive and authentication fields are one-hot encoded; boolean flags use zero or one. The scale is fixed rather than estimated from training data. There is no per-feature standardization, learned

normalization layer, input clipping, output clipping or gradient clipping. Training uses the stated AdamW settings without a learning-rate scheduler. The bound $\min(M/100000, 1)$ applies only to the reported bounded-error metric and does not constrain predicted balances. These choices do not equalize the numeric objectives across output types. The default smooth-L1 regression loss and the sign cross-entropy loss have different scales and geometry. C and D also have 780 more parameters than A, B and E and receive exact observed magnitude in decoding. Canonical ordering is compared within each output family, but the cross-head comparison cannot attribute the result to representation, parameter count, magnitude access or loss scaling independently. We report the actual counts and losses rather than treating a matched backbone as a fully matched intervention.

Learning adequacy and the retained pilot failure

The prospective development specification used one separate corpus and two seeds per arm, ten fits in total. D had to reach at least 99% ID tuple exactness and at least 95% recall for each status in each seed. B had to beat the zero-posting reference on applied, nonzero ID delta MAE and meet the same recall threshold. These are study adequacy criteria, not deployment or journal standards. The 20-epoch first pilot failed B's numerical criterion: its two applied-posting MAEs were 10881.49 and 10121.52 cents, versus 9109.16 for zero posting. D passed, and B's numeric validation loss was still declining. The one allowed amendment increased every condition to 100 epochs while preserving all other choices and gate thresholds. Both pilots, their curves, checkpoints and prediction rows are retained. In the amended pilot, B's MAEs were 4856.64 and 4947.74, all B/D status recalls were 1, and D achieved exact ID prediction in both seeds. This passed the fixed gate but did not claim that B was an optimal regressor.

The main configuration fixes eight new corpus seeds and three new paired training seeds, yielding 120 fits. AdamW uses learning rate 10^{-3} , weight decay 10^{-4} , batches of 512 and one CPU thread. Each run selects the lowest validation-loss checkpoint within 100 epochs. There is no rollout loss, rollout-based checkpoint selection, architecture search, extra seed selection, or outcome-dependent extension. Configuration, source, tests, analysis, amendment and development-exclusion data were hashed before the first main fit. This is a local prospective freeze, not public preregistration. The completed main study contains 120 fits and 3,126,912 independently reconstructed prediction rows.

Carried-state stress test and operational references

Each corpus contains 128 trajectories from each of three families. The cycle family repeats submit, approve, capture, settle, merchant payout, starting at draft. The fee-capacity family repeats fee assessment from captured or settled, with one initial headroom value selected from 20,000, 60,000 or 120,000 cents and shared by all slots in that trajectory. Accumulated balances can therefore

change later capacity decisions. The mixed family samples primitives, amounts, authentication and expedited flags, with a randomly selected initial stage and risk. Actions are generated once, independently of model predictions. The oracle and all evaluated models receive identical sequences.

At each step the model receives its own predicted balances and stage. No predicted state is replaced by the oracle. Predicted non-applied statuses retain the model's current state; finite capacity-invalid states continue to be evaluated and carried. A nonfinite prediction becomes an absorbing failure: $B_t=1$, exactness is zero for all remaining steps, and raw MAE is unavailable with its denominator explicitly recorded. Invalid or nonfinite cases are never removed from bounded or exactness endpoints. Evaluation saves all 32 steps; prespecified displays use horizons 1, 2, 8 and 32.

The controller reference executes the independent rule-table oracle. Its transition rules are known and cheap to execute in this study, so the exact controller already solves this specified task. Learned prediction is evaluated as a diagnostic; no practical advantage over that exact executor is established. Persistence predicts zero delta, unchanged stage and applied status; it is a numerical and task-floor reference, not a strong categorical classifier. A workflow heuristic obeys primitive stage preconditions and exact role postings but ignores sanctions, closed books, locks, review and capacity checks. This domain-informed heuristic is an important competitor: it can be poor on the balanced one-step control schedule yet strong on rollout families that rarely exercise its omitted rules. It is not trained or presented as a learned system.

Primary comparison and uncertainty

The primary endpoint is mean B_8 , equally weighting the three trajectory families. The two primary contrasts are D minus B and E minus B; negative values favor D or E. For each corpus, we average three paired-seed differences, then compute a Student-t interval across the eight corpus means (7 degrees of freedom). The same seed trio is reused in every corpus: inference concerns new draws from this generator conditional on that fixed trio, not unseen initialization seeds. Each marginal interval uses 97.5% confidence, applying Bonferroni to the two-contrast family for nominal 95% joint coverage. Unadjusted 95% intervals are retained as sensitivity summaries. With only eight corpora, coverage relies on an approximate normal-mean assumption and cannot establish broad population generality. Neither prediction rows nor nested seeds are counted as independent corpora.

A prespecified interval half-width of 0.02 is a reporting precision reference, not a stopping or success condition. We retain all eight differences, adverse contrasts and failures without expanding the sample. Applied-posting routing, magnitude, status recall, exactness, conservation, capacity validity, other horizons and families are descriptive secondary measurements, without further familywise claims.

Post collection reviewer analyses

Additional analyses were completed on October 11, 2026 after the original outcomes were known. They reuse every original corpus, paired training seed and selected checkpoint; no fitting, new checkpoint selection, data-dependent stopping or replacement of the prespecified primary result occurred. We evaluated all 120 checkpoints on oracle states at horizons 1, 2, 8 and 32 of the same held-out trajectories, retaining 184,320 predictions. The carried-state values remain those of the frozen original evaluation. Teacher-forced tuple exactness tests the current delta, next stage and status on an oracle input; carried-state exactness tests accumulated balances, stage and current status. Their difference cannot identify a causal fraction due to accumulation.

For weighting sensitivity, each original corpus retains its three fixed paired seeds, and we apply seven stated family-weight scenarios. A paired corpus percentile bootstrap uses 20,000 draws with seed 20261011; leave-one-corpus-out means omit each of the eight corpora in turn. These are descriptive post-collection checks, not replacements for the two prespecified 97.5% corpus-t intervals or measured deployment proportions. Appendix B reports the scenarios and complete family-by-horizon diagnostic. A final reporting analysis on October 12, 2026 adds uncertainty for the secondary correctness endpoints in Table 16. For every arm, family and scheduled horizon, we first average the three fixed paired seeds within each corpus, then report an exploratory unadjusted 95% Student-t interval over the eight corpus means (7 degrees of freedom). Teacher tuple exactness is recomputed from the retained 184,320 predictions; carried full-state exactness uses the original 120 saved summaries. Appendix D reports all 60 cells. These approximate, unclipped intervals can extend outside the unit range, and a zero-width interval indicates no observed corpus variation rather than certainty about other populations. They provide neither familywise coverage nor untouched confirmation.

A separate fixed-checkpoint ablation evaluates all 24 B fits with test-time mean centering followed by either independent cent rounding or conserving integer rounding, and all 24 E fits with its original centering but independent rounding. Each intervention carries its own changed predictions through all 32 steps. The original B and E results remain comparison conditions. B versus centered B isolates inference postprocessing on identical weights; E versus centered B with the same conserving decoder compares centering during training while holding that inference package fixed; conserving versus independent rounding within one trained arm isolates the integer decoder. These interventions add 72 checkpoint evaluations and 884,736 retained rollout rows. They do not test multi-step training, matched head/loss training or larger neural baselines.

Additional controlled training study

A reviewer-requested follow-up was locally frozen

before its first fit on October 11, 2026, after the original test distribution and inference diagnostics had been inspected. It therefore supplies additional controlled evidence on an already inspected benchmark, not untouched confirmation. Eight new conditions each use all eight original corpora and the same three paired seeds, producing 192 fits. The one-step data and held-out ID, primitive/trigger and 32-step rollout evaluations remain unchanged. All conditions use 100 epochs, the original AdamW settings and one CPU thread per fit; three isolated spawn workers execute independent fits. All scheduled identities, source and input hashes, histories, selected checkpoints and raw rows are retained.

The parameter-budget control B67 uses 10,869 parameters versus the sign head's 10,971, a difference below 1%; changing width remains an architectural intervention. B025 and B4 multiply the original regression numeric training loss by 0.25 and 4 while retaining the original unweighted validation criterion. DBrier retains D's exact architecture, 18-output head, observed amount and decoder, replacing numeric sign cross-entropy with 1.5 times the class-summed Brier loss. That factor matches cross-entropy's logit gradient at uniform three-class probabilities only, without asserting full optimization equivalence. RoleProjected uses the inherited role-tied projected network at width 101 with 10,963 parameters. It is an additional structured architecture package, not an isolated equivariance or head intervention.

The matched sequence suite uses canonical B with the original 30,000 one-step cases plus 3,000 new four-step training trajectories per corpus, balanced across the three families. A separate 600-trajectory validation set is shared. Canonical sequence/prefix inputs are globally disjoint from 561,254 original main, pilot and auxiliary inputs and from other added groups. All original structural primitive/trigger HOLDOUT combinations are excluded from added training and validation, with zero retained exposure. The proof retains rejected conjunction counts rather than equating exact-case disjointness with semantic holdout.

BSequenceFlat independently processes 512 oracle-state prefixes in one forward batch. BSequenceTF processes four chronological 128-case batches, resets each input to its oracle prefix, then averages losses before one optimizer update. BSequenceCarry instead carries its own balances and stage through those four prefixes and propagates numeric gradients through earlier balances. All three use the same transition groups, update counts, decoded-next-state loss, categorical supervision and common checkpoint criterion: one half ordinary validation loss and one half model-carried four-step validation loss on the same additional held-out sequences. Oracle targets follow their separately evolved trajectory. Integer-cent forward rounding uses an identity straight-through numeric derivative; categorical argmax and applied-status gates detach while retaining cross-entropy supervision. Finite capacity-invalid states continue unclipped, without oracle reset. Original-B comparisons change

data exposure, update budget and validation and remain package comparisons.

The flattened and teacher-forced objectives are mathematically equivalent for this network, which has no dropout or batch-dependent normalization. Their separate implementations are a batch/reduction control; observed differences cannot be advertised as a temporal-training benefit. The controlled carried comparisons are reported against both. A separate 2,000-case balance-flip probe per corpus holds action and context fixed within 1,000 applied/rejected pairs. A fit passes the declared status gate only if each status recall is at least 95% and both members' statuses are correct in at least 90% of pairs. These thresholds remain unchanged. Follow-up

uncertainty averages the fixed paired seeds inside each corpus and reports exploratory 95% t intervals over eight corpus differences, without a new familywise or untouched-confirmation claim.

RESULTS AND DISCUSSION

Single-step learning and the withheld review ceiling

Table 2 separates learning a supported transition from predicting its numeric components. D has mean ID exactness 99.9975%, with minimum 99.98% across its 24 main fits. B's applied ID posting MAE is 5,746.61 cents; the zero-posting reference is 9,165.69. Exact regression scores must be read alongside these numerical errors: independent cent rounding makes small continuous

Table 2: One-step outcomes. Learned scores average 24 fits per condition; references average eight corpora. Posting metrics condition on true-applied, nonzero ID deltas. Posting MAE averages absolute delta error over six slots, then over those cases; MAEs are cents. Routing requires unique positive and negative extrema at the correct roles; it does not require the other slots to be zero. Magnitude error is half the absolute difference between predicted and target L_1 posting mass.

Condition	ID exact (%)	Posting MAE	Dominant routing (%)	Magnitude MAE	Holdout (%)
A	54.3975	10,194.31	11.62	22,952.13	100.00
B	54.3975	5,746.61	87.26	8,593.33	100.00
C	93.7500	908.43	81.80	2,091.89	99.97
D	99.9975	0.00	100.00	0.00	100.00
E	54.3975	5,691.36	85.05	8,393.74	100.00
Workflow heuristic	54.72	0.00	100.00	0.00	0.00
Persistence	0.00	9,165.69	0.00	27,497.08	0.00
Controller	100.00	0.00	100.00	0.00	100.00

errors fail an exact match. Full per-status recalls appear in the appendix.

The four-cell holdout is deliberately reported in a separate column. Its oracle targets are all review with zero posting and unchanged stage. A correct status activates the shared non-applied interface, which supplies the remaining outputs. A constant-review rule with that interface would therefore score perfectly on this isolated stratum by construction. A learned ceiling here is compatible with narrow primitive/trigger transfer but cannot by itself demonstrate even that rule was used, difficult numerical

composition, or complete controller reasoning.

Primary carried-state comparison

For D-B, the mean difference is -0.14495; the first condition is lower throughout the reported interval. The half-width 0.01129 meets the prespecified 0.02 reporting reference. No sample was added in response.

For E-B, the mean difference is -0.00558; the first condition has an interval spanning both improvement and deterioration. The half-width 0.01072 meets the prespecified 0.02 reporting reference. No sample was

Table 3: Prespecified horizon-8 bounded accumulated error contrasts. Each of eight observations averages three paired training seeds. Negative values favor the first condition. Marginal 97.5% intervals apply Bonferroni to the two-contrast family for nominal 95% joint coverage, subject to the eight-corpus approximation

Contrast	Mean difference	97.5% corpus-t interval	Half-width
D-B	-0.14495	[-0.15625, -0.13366]	0.01129
E-B	-0.00558	[-0.01630, 0.00513]	0.01072

added in response.

E's conservation rate is 100.00%, as intended by its decoder, while its full-state exactness is 0.09%. D's corresponding state exactness is 34.24%, despite its much higher single-step ID score. These are measured

outcomes of the specified packages and distributions; they do not attribute the difference to a single isolated mechanism.

The prespecified descriptive representation diagnostics also retain adverse comparisons. B minus A in horizon-8

Table 4: Carried-state outcomes at horizon 8, on identical exogenous actions. Exactness includes all six balances, workflow stage, and current status; trajectory exactness requires every prefix. Finite capacity-invalid states remain in every score. The error and exactness columns are distinct from current transition-tuple exactness.

Condition	MAE (cents)	B_8	X_8 (%)	Q_8 (%)	Conserved (%)	Capacity invalid (%)
A	16,067.24	0.1607	0.09	0.00	36.11	0.00
B	23,291.74	0.2310	0.11	0.00	35.68	0.77
C	9,120.28	0.0912	19.93	19.22	67.26	0.01
D	8,608.44	0.0861	34.24	33.33	84.58	0.00
E	22,696.22	0.2255	0.09	0.00	100.00	0.71
Controller	0.00	0.0000	100.00	100.00	100.00	0.00
Persistence	10,370.28	0.1037	0.00	0.00	100.00	0.00
Workflow heuristic	4,112.68	0.0411	72.95	72.33	100.00	18.42

bounded error is 0.07037, while B minus A in applied ID posting MAE is -4447.70 cents. Thus the one-step numerical ranking must not be substituted for the carried-state ranking. D minus C is -0.00511 and D minus A is -0.07459 in bounded horizon-8 error. These comparisons are descriptive, without additional familywise inferential claims.

The heuristic comparison changes the practical interpretation

The workflow heuristic has horizon-8 MAE 4,112.68 cents, lower than D's 8,608.44; its full-state exactness is 72.95%, versus D's 34.24%. Its capacity-invalid rate is 18.42%, versus D's 0.00%, so the numerical advantage is not overall dominance or safety. Its one-step ID exactness

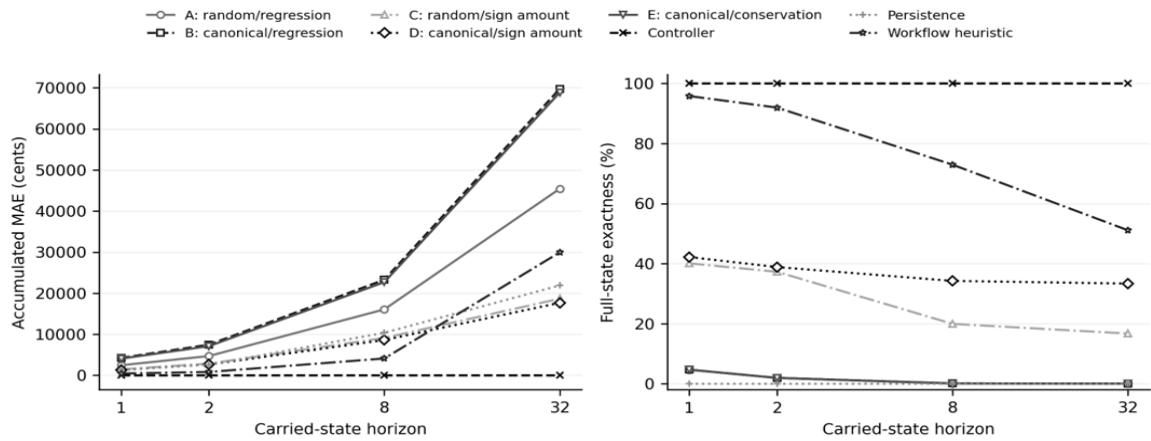


Figure 1: Each model carries its own state. Left: mean six-balance accumulated MAE, including finite invalid states. Right: full-state exactness, including current status. Points average the same fixed eight corpora and three nested seeds for learned conditions; lines connect reported horizons and are not interpolation experiments. The controller is perfect by construction

is only 54.72%. This reversal in what a benchmark rewards is central: the heuristic ignores important controls that appear frequently in the balanced one-step schedule, but obeys exact workflow transitions and posting roles

useful in the rollout families. A better learned-to-learned contrast therefore cannot by itself establish the strongest operational baseline.

Table 5 keeps the three families visible. The fee-capacity

Table 5: Family-specific horizon-8 outcomes; exactness columns are percentages. Each family has 128 trajectories per corpus. These descriptive breakdowns expose where a pooled difference arises rather than treating all rollouts as one homogeneous challenge.

Condition	Cycle B_8	Cycle X_8	Fee B_8	Fee X_8	Mixed B_8	Mixed X_8
A	0.1734	0.00	0.1707	0.00	0.1378	0.26
B	0.1483	0.00	0.1707	0.00	0.3740	0.33
C	0.0099	57.68	0.1707	0.00	0.0929	2.12
D	0.0000	100.00	0.1707	0.00	0.0875	2.73
E	0.1367	0.00	0.1707	0.00	0.3690	0.26
Workflow heuristic	0.0000	100.00	0.0797	44.73	0.0437	74.12
Persistence	0.0794	0.00	0.1707	0.00	0.0610	0.00
Controller	0.0000	100.00	0.0000	100.00	0.0000	100.00

family directly exercises accumulated-balance-dependent rejection under initial headroom unlike the usual training cases. Mixed paths produce many rejected or review transitions, so unchanged-state cases can be easier than repeated applied postings. Figure 1 reports all prespecified horizons without resetting a model to the oracle.

D's horizon-8 full-state exactness is 100.00% on cycle, 0.00% on fee-capacity, and 2.73% on mixed trajectories. All five learned conditions have the same reported fee-capacity bounded error, 0.1707, as persistence. The pooled D-B advantage is therefore not evidence of learned capacity competence.

No nonfinite predictions occurred in the retained main evaluation; all raw-error denominators are therefore complete. The independent artifact audit verified 3,126,912 rows across 144 model/reference streams and reconstructed every saved summary. This checks recorded-state consistency and accounting, not independent retraining or real-world validity.

Exploratory capacity state support diagnostic

Inspection during main collection identified a specific support restriction: all 2,046 original capacity-block training cases in the first main corpus had zero balances. The remaining 27,954 cases had nonzero balances, so the problem was not absence of balance input. Capacity decisions had been trained on a narrow boundary fixture. We therefore specified a separate, exploratory support diagnostic before fitting any auxiliary model, preserving every main result and its original interpretation.

The diagnostic has one fresh corpus, two new paired training seeds, B/D/E, and original versus augmented training: twelve additional fits. Both training sets contain 30,000 examples and receive the same 100-epoch opportunity. Augmented training retains a shared 20,000 original cases and replaces the other 10,000 with capacity-state pairs. Both conditions select checkpoints on the identical validation mixture of 2,500 original cases and 2,500 capacity-pair examples. All statuses remain

supported. Canonical pilot and planned-main inputs are excluded, and planned main fingerprint roots are checked against the completed main corpora. This changes training support and its weighting; it is not a controlled estimate of one feature's isolated mechanism.

Each pair holds the action, stage, capacities and other context fixed. An outward-posted role's balance moves across the capacity decision boundary, with an inactive role compensating to preserve zero sum. Both initial states are legal; all earlier checks and the other posting role remain safe. One member must apply and the other reject. Both posting directions and all six posting primitives are included. Learning examples have margins between one-quarter and three-quarters of the amount; one-cent boundaries are separate oracle correctness fixtures. A predictor insensitive to balances sees identical remaining inputs, so its status accuracy is at most 50% and its both-members-correct accuracy is zero on these balanced pairs.

Shared evaluation contains 5,000 supported ID cases, 4,000 review holdouts, 2,000 held-out capacity pairs and 384 new length-32 trajectories. On the same held-out trajectory IDs, we additionally evaluate one-step predictions from the oracle state at every prefix. These teacher-forced rows are separate from the self-carried path and do not reset it. Their original prefix is retained in the case ID; the one-step evaluator's local step index is always one. We report prefix-specific results, not an all-prefix average mislabeled as horizon eight.

For every augmented B/D/E seed, capacity applied and rejected status recall must each reach 95% and both members must be correct for at least 90% of pairs. D must also retain 99% original ID tuple exactness. These are decision-competence gates; exact continuous regression cents are reported separately. Original-data models may fail: their support gap is the question. Every outcome is retained, with no further tuning or sample expansion. One corpus and two seeds support descriptive comparisons only.

Table 6: Every auxiliary fit, without selection or exclusions. ID, capacity applied/rejected recall (A/R), both-pair-status correctness, teacher-forced (TF) prefix-8 exactness and carried-state exactness are percentages. MAE is carried six-balance error at horizon 8 in cents. TF uses oracle states on the same trajectories and is a separate diagnostic.

Training	Arm	Seed	ID	Cap. A	Cap. R	Both	TF tuple@8	Carried X_8	MAE
augmented	B	811	54.40	59.50	74.40	34.10	14.32	0.00	20,392.79
augmented	B	821	54.40	64.45	64.10	28.70	15.62	0.52	17,925.30
augmented	D	811	100.00	69.00	63.70	32.70	59.64	42.19	4,712.46
augmented	D	821	100.00	84.55	54.85	39.40	63.02	44.01	4,530.16
augmented	E	811	54.40	65.15	67.40	32.60	10.94	0.26	18,824.66
augmented	E	821	54.40	73.40	54.20	27.75	16.41	0.26	17,934.44
original	B	811	38.12	20.15	79.85	0.00	22.14	0.00	13,664.27
original	B	821	49.84	19.80	80.15	0.00	21.61	0.00	12,846.13
original	D	811	49.14	19.95	79.95	0.00	22.40	1.04	9,898.00
original	D	821	51.16	19.85	65.60	0.00	4.69	1.04	9,891.49
original	E	811	38.52	20.25	79.70	0.00	22.14	0.00	12,606.70
original	E	821	49.80	19.80	80.00	0.00	21.61	0.26	12,470.18

The auxiliary original-data conditions are not unchanged copies of the competent main baselines. Their shared validation set contains 50% new capacity-state cases, changing checkpoint selection: original B selects epochs 2/3, D 3/4, and E 2/3. Augmented B selects 99/99, D 99/99, and E 99/98. Original D has only 49.14/51.16% ID exactness at those selected checkpoints. Its mixed validation losses are 1.1935/1.1938 at selection and 9.4470/9.6472 at epoch 100. These observed losses support a selection-distribution explanation, not an assertion about unmeasured later-checkpoint ID accuracy. Both auxiliary training conditions use the identical selection rule, but changes relative to the main study also involve new corpus/seed and validation distributions; they cannot be attributed to capacity training alone.

The augmented competence gate fails for B/811, B/821, D/811, D/821, E/811, E/821. This outcome is retained without further fitting. The auxiliary therefore does not establish that all compared conditions are competent on the state-sensitive capacity task; any numerical improvements below remain descriptive.

For B, original-to-augmented mean both-pair correctness changes from 0.00% to 31.40%, while carried horizon-8 MAE changes from 13,255.20 to 19,159.04 cents. Under augmented training, teacher-forced prefix-8 transition exactness is 14.97% and self-carried state exactness is 0.26%. These average two fixed seeds on one corpus and carry no cross-corpus interval.

For D, original-to-augmented mean both-pair correctness changes from 0.00% to 36.05%, while carried horizon-8 MAE changes from 9,894.75 to 4,621.31 cents. Under augmented training, teacher-forced prefix-8 transition exactness is 61.33% and self-carried state exactness is 43.10%. These average two fixed seeds on one corpus and carry no cross-corpus interval.

For E, original-to-augmented mean both-pair correctness changes from 0.00% to 30.18%, while carried horizon-8 MAE changes from 12,538.44 to 18,379.55 cents. Under augmented training, teacher-forced prefix-8 transition exactness is 13.67% and self-carried state exactness is 0.26%. These average two fixed seeds on one corpus and carry no cross-corpus interval.

The auxiliary workflow heuristic has carried horizon-8 MAE 4,531.25 cents and state exactness 69.27%. Its omitted capacity checks make both-members-correct capacity accuracy zero. The artifact audit reconstructed all 563,640 model/reference rows. The main study is unchanged, and this post-design support diagnostic cannot retroactively remove the main distribution shift.

Reviewer analyses of oracle state and carried state

The new main checkpoint diagnostic makes the family boundary explicit. At horizon 8, D has 100.00% teacher-forced tuple exactness and 100.00% carried-state exactness on cycle paths. On fee-capacity paths the corresponding values are 43.95% and 0.00%; on mixed paths they are 17.19% and 2.73%. B and E also have 43.95% teacher-forced tuple exactness on the fee-capacity prefix. Errors therefore occur even when the oracle supplies the current

state. This reinforces the missing capacity-competence finding; it does not isolate error accumulation after correct state-sensitive one-step decisions. Table 16 includes every condition, family and scheduled horizon rather than only the favorable cycle family.

The equal-weight D minus B bootstrap interval is $[-0.15412, -0.13778]$, and its leave-one-corpus-out means range from -0.14677 to -0.14154. E minus B has a bootstrap interval $[-0.01381, 0.00188]$, again spanning zero, and leave-one-corpus-out means from -0.00776 to -0.00295. These small-sample sensitivity summaries do not improve or establish frequentist coverage. Family-only means for D minus B are -0.14834 on cycle, 0 on fee-capacity and -0.28652 on mixed paths. Thus increasing fee-capacity weight attenuates the observed pooled advantage and a fee-only evaluation supplies none. No deployment weights are inferred from these constructed families.

The additional secondary-endpoint intervals preserve every Table 16 mean. They distinguish uncertainty in teacher tuple correctness from uncertainty in carried full-state correctness rather than treating either endpoint as independent prediction-level replication. The small corpus sample and fixed seed trio remain limitations; zero-width intervals for uniformly successful or failed cells are not evidence of a universal guarantee. Table 23 reports all arm, family and horizon cells, including unfavorable ones.

The projection ablation separates numerical accuracy from the imposed invariant. Centering and conserving decoding on unchanged B checkpoints reduces mean bounded horizon-8 error by 0.000709, while full-state exactness remains 0.1085%. Conserving rounding raises conservation from 51.71% for independently rounded centered B to 100.00%, with a mean bounded-error difference of approximately 1.0×10^{-8} . For E, independent rounding gives 52.21% conservation and the same 0.0868% full-state exactness as its original conserving decoder; the bounded-error difference is approximately 7.2×10^{-10} . Table 18 reports the full numerical and validity metrics. These small rounding differences are properties of the recorded toy task, not a universal irrelevance of projection. Centering during training, under the common conserving decoder, has a mean E minus centered-B difference of -0.004874 with mixed signs across corpora. No broadly beneficial effect or capacity repair follows from that descriptive contrast.

Results of the additional training controls

All 192 scheduled fits completed; recorded training failures were 0. The additional raw-row reconstruction verified 4,766,208 new-fit rows and 144,000 original-checkpoint capacity-probe rows. Reloading the selected follow-up checkpoints reproduced 98,304 retained ID predictions, and chosen epochs match the declared minimum validation criterion. These checks establish recorded arithmetic and inference consistency, not an independently retrained replication. Tables 19–22 retain every condition and adverse outcome. None of the 192

follow-up fits passes the balance-flip capacity-status gate. For bounded horizon-8 error, the exploratory means and 95% corpus intervals are B67 minus B: -0.00897 [-0.02080, 0.00286]; B025 minus B: -0.01288 [-0.01934, -0.00642]; B4 minus B: -0.01617 [-0.02423, -0.00811]; DBrier minus D: 0.00191 [-0.00141, 0.00523]; RoleProjected minus E: -0.08826 [-0.10005, -0.07648]. Negative differences favor the first condition. These alternatives test the stated budgets or objectives, not all possible losses, parameterizations or conservation architectures. The original primary D minus B and E minus B estimates are unchanged and are not replaced by a retrospectively selected favorable comparison.

The matched sequence-suite contrasts are BSequenceCarry minus BSequenceFlat: 0.00741 [-0.00404, 0.01885]; BSequenceCarry minus BSequenceTF: 0.00602 [-0.00613, 0.01818]; BSequenceTF minus BSequenceFlat: 0.00138 [-0.00232, 0.00509]. Flat versus TF is the implementation control, despite equivalent mathematical objectives. The retained first corpus/seed selects epochs 97 and 96 and has respective bounded horizon-8 errors 0.101824 and 0.092243. The initial validation values differ by approximately 1.3×10^{-8} , but no causal explanation of the later divergence has been established. Any carried-versus-control effect must be read alongside this observed control drift and the two separate contrasts. The suite does not demonstrate that temporal feedback alone repairs capacity decisions, and a comparison with original B cannot attribute changes solely to four-step training.

Reading ordinary competence and the capacity failure together

The ordinary ID gate and the auxiliary capacity gate refer to different predicates over different inputs. In the amended development pilot, the canonical regressor had to improve on zero posting for applied examples and recognize the supported statuses. The sign condition had to reproduce the complete transition tuple on the scheduled ID cases. Passing those requirements supports collection of the specified main comparison; it does not assert competence on every legal balance configuration of the controller. In particular, the capacity-block examples in the original schedule use zero balances. They provide the output label and a boundary fixture without requiring the model to distinguish the legal balance-flip pairs introduced later. Output-label support therefore removes an unsupported-class problem while leaving a state-sensitive support problem possible.

This distinction explains why the near-perfect ID result must be read beside the family results and auxiliary gates. The sign condition reproduces the ordinary ID tuple with the observed amount supplied to its output representation. The fee-capacity trajectories instead change how remaining headroom interacts with repeated postings. Those trajectories ask a model to respond to a balance-dependent control boundary that the ID score does not establish it has learned. The auxiliary holds the action and other relevant context fixed while varying

balances across that boundary. Its both-members-correct endpoint requires the model to distinguish the two outcomes within the same pair, rather than succeed on an aggregate that can reward a constant status. Failure of that endpoint is a substantive diagnostic result, but it cannot be converted into evidence of accumulated error after the missing decision competence has already been established.

Teacher-forced prefix evaluation provides another useful separation. It presents the oracle state from the held-out trajectory to the one-step predictor, whereas the carried path presents that model's own state. These rows can locate difficulties on the shifted state distribution without overwriting the carried trajectory. Their exactness scores nevertheless refer to a current transition tuple, while carried exactness refers to accumulated balances, stage and current status. The two percentages therefore measure different objects. A numerical difference between them cannot, by subtraction alone, assign a causal fraction of failure to error accumulation. Relevant capacity competence, state occupancy, categorical errors and the stricter accumulated endpoint remain distinct considerations. The present diagnostic preserves these distinctions instead of treating a single one-step percentage as a certificate for the full rollout.

What the operational references do and do not establish

The exact controller is particularly important in interpreting the applied setting. Its rules and observed inputs are available, and direct execution already returns the correct transition at each step. Nothing in this experiment makes state observation costly or the transition function unknown. There is consequently no demonstrated reason to substitute a learned predictor for that controller in this specified task. The empirical contribution is the comparison and diagnosis of learned prediction under explicit evaluation conditions, not an efficiency, autonomy or deployment claim. A future use case with genuinely unavailable or expensive transitions would need its own evidence and cost model; the present study does not create such a cost to make learned prediction appear useful.

The other references serve different purposes. Persistence reveals how much numerical error can be avoided by making no posting, without supplying a strong classifier or a faithful workflow. The workflow heuristic uses the known primitive stages and postings but omits several controls. Its favorable carried numerical and exactness scores show that a learned-to-learned improvement does not by itself establish the best operational performance within the suite. Its capacity-invalid states simultaneously expose an adverse outcome omitted by numerical error alone. These observations justify reporting multiple endpoints together. They do not establish a universal ranking of the heuristic and learned conditions, and they do not turn a zero observed invalid-state rate into a guarantee for untested actions or distributions.

The fee-capacity breakdown also limits the pooled contrast. The main D-minus-B result averages the three scheduled families with equal weight. This is the specified estimand, not the prevalence of those families in a financial institution. D's success on cycle paths and failure of complete exactness on fee-capacity paths can coexist with a favorable pooled numerical contrast. Reweighting those families could change the operational interpretation; no empirical deployment weights are available here. Reporting the complete family table makes that dependence visible without retrospectively changing the primary comparison, selecting only favorable paths, or claiming that the main contrast measures a general financial capability.

Limits on the conservation comparison and uncertainty

The conservation condition is a specific training-and-decoding package, rather than a test of all possible invariant-aware learning methods. It changes the canonical regressor through centering and conserving integer decoding. The main comparison measures that package under the fixed data, optimization and state distributions. The interval for E minus B contains both improvement and deterioration in bounded accumulated error. This is neither proof of equal performance nor proof that conservation is useless. Likewise, perfect conservation in decoded outputs verifies the imposed zero-sum property, while the accompanying state and numerical metrics evaluate other aspects of task correctness. A conserved posting can still have the wrong routing, stage or amount, and a preserved total cannot reconstruct a state that earlier decisions have made incorrect.

The uncertainty calculation also has a deliberately limited unit of replication. It uses eight generated corpora and averages the same three paired initialization seeds within each corpus. The millions of retained rows support faithful reconstruction of those recorded paths; they do not create millions of independent controllers, deployments or training environments. The reported corpus intervals concern draws from the declared generator conditional on the fixed seed trio and use an approximate small-sample mean calculation. Keeping the raw rows, corpus contrasts and failed outcomes makes this boundary inspectable. It does not supply independent retraining or a broader sample of domain semantics. The useful reading of the result is therefore a bounded diagnostic of this evaluation design, with the failure of capacity competence remaining central to its interpretation.

Interpretation and limitations

The experiment asks whether a representation that fits the one-step task also predicts its own evolving state. The supported ID results establish the extent of ordinary task learning. The review-only holdout establishes a much narrower observed behavior: producing a supported review output on four recombinations, where the interface supplies zero posting and unchanged stage after that status is chosen. A constant-review predictor can

achieve the same isolated ceiling, so this stratum alone does not establish compositional rule use. The rollout comparison then tests a shifted use regime, not a stronger version of the same one-step score. Its differences can reflect capacity/headroom shift, accumulated errors, state-distribution shift and categorical mistakes together. The sign-and-amount representation incorporates exact magnitude from the input and changes the loss, head size and representable outputs. It is deliberately task specific. Its original comparison with regression cannot identify the independent contribution of those changes. The reviewer follow-up adds an approximately matched parameter budget, fixed loss-weight sensitivities, an identical-sign-head alternative objective and an existing role-equivariant architecture. These controls test specific alternatives without establishing a complete mechanistic decomposition or optimal training. Within each output family, canonical ordering is more controlled, but the finite architecture and fixed optimization budget still limit claims. Passing an ordinary numerical/status gate does not certify convergence, cent-exact prediction or state-sensitive capacity competence. Arithmetic modules, learned action policies, longer trained horizons and systematic architecture or loss searches remain untested. Conservation is mechanically guaranteed for E's decoded finite predictions. High conservation therefore verifies its interface invariant, not successful accounting decisions. We retain the package's empirical numerical effects, including a null or adverse result, without claiming the invariant caused every difference. Exact cent matching also imposes a much stricter criterion than low continuous loss. The report keeps raw MAE, bounded error and exactness together so that either metric cannot replace the others.

The domain is a finite, synthetic workflow using simplified role names and rules. It is not an institution's ledger, legal accounting model, market simulator, financial safety test, or evidence about language-model reasoning. Review, sanctions, lock and capacity rules are chosen toy semantics. No private records, monetary actions, external APIs or model downloads are involved. Long-horizon trajectories were generated exogenously and do not evaluate a learned policy, planning quality or counterfactual action selection. Mixed rollouts contain no deferred outcomes under the chosen static context, despite deferred being represented and tested in one-step splits.

Finally, the development corpus influenced the sole training-budget amendment, and historical results motivated the design. Only the fixed main evidence is used for corpus-level estimates. The random corpora sample the same generator and fixed cell schedule; they do not measure variation across tasks, controllers or real deployments. Numerical replay validates internal evidence consistency, while oracle agreement checks the specified controller. Neither establishes the external correctness of the domain, an independent training reproduction, human review, or publication readiness.

CONCLUSION

The resulting evidence is a diagnostic comparison with clear boundaries. Status support removes one earlier confound; a retained failed pilot establishes why learning adequacy must precede interpretation; separate one-step and carried-state endpoints test different capabilities; and an operational heuristic prevents learned-model comparisons from standing alone. In this run, D's mean ID exactness is 99.9975%, its horizon-8 state exactness is 34.24%, and the exact-conservation condition E has horizon-8 state exactness 0.09%. The auxiliary capacity-support repair fails its stronger competence gate. Consequently, the evidence does not establish a learned baseline competent on the relevant capacity decisions, nor isolate long-horizon accumulation after that competence has been achieved. The practical recommendation is to audit these quantities together before using invariant satisfaction or transition scores as evidence of a useful evolving simulator. The follow-up now tests specific parameter, loss, architecture and training alternatives with the original evidence preserved. None of the 192 follow-up fits passes the balance-flip capacity-status gate. Observed differences between equivalent Flat/TF controls further limit isolated temporal interpretation. The recommendation remains to audit numerical error, exact state, validity and relevant decision competence together; neither a favorable package contrast nor invariant satisfaction certifies the full evolving controller.

Data and code availability

The study uses generated synthetic records and a deterministic controller. The original reproduction archive retains source, configurations, protocols, hashes, histories, checkpoints, input corpora and prediction rows. The revision supplements add the October 11 teacher-forced, weighting, uncertainty and projection analyses, and the separately frozen 192-fit training follow-up with all added inputs, histories, checkpoints, outputs, verification and reporting scripts. The materials distinguish original prospective evidence from a follow-up amendment on an already inspected test set. Public repository deposition is not claimed. Source replay and numerical reconstruction do not constitute independently retrained replication or validation of financial-domain realism. A separately identified October 12 reporting supplement supplies the secondary correctness intervals, all 480 corpus-cell means and the analysis script.

REFERENCES

- Bengio, S., Vinyals, O., Jaitly, N., & Shazeer, N. (2015). Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in neural information processing systems* (Vol. 28). https://proceedings.neurips.cc/paper_files/paper/2015/hash/e995f98d56967d946471af29d7bf99f1-Abstract.html
- Chen, H., Constante Flores, G. E., & Li, C. (2024). Physics-informed neural networks with hard linear equality constraints. *Computers & Chemical Engineering*, 189, Article 108764. <https://doi.org/10.1016/j.compchemeng.2024.108764>
- Ellis, K., Wong, C., Nye, M., Sablé-Meyer, M., Morales, L., Hewitt, L., Cary, L., Solar-Lezama, A., & Tenenbaum, J. B. (2021). DreamCoder: Bootstrapping inductive program synthesis with wake-sleep library learning. In *Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation* (pp. 835–850). Association for Computing Machinery. <https://doi.org/10.1145/3453483.3454080>
- Hoedt, P.-J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G. S., Hochreiter, S., & Klambauer, G. (2021). MC-LSTM: Mass-conserving LSTM. In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 4275–4286). Proceedings of Machine Learning Research. <https://proceedings.mlr.press/v139/hoedt21a.html>
- Mahdavi, S., Swersky, K., Kipf, T., Hashemi, M., Thrampoulidis, C., & Liao, R. (2023). Towards better out-of-distribution generalization of neural algorithmic reasoning tasks. *Transactions on Machine Learning Research*. <https://openreview.net/forum?id=xkrtvHlp3P>
- Ojha, A. (2026a). Presentation order and structured diagnosis in a synthetic payment workflow benchmark. *American Journal of Innovation in Science and Engineering*, 5(3), 19–30. <https://doi.org/10.54536/ajise.v5i3.8542>
- Ojha, A. (2026b). Stateful mediation and selective auditing for financial language-model actions: A paired controlled simulator study. *American Journal of Financial Technology and Innovation*, 4(1), 198–210. <https://doi.org/10.54536/ajfti.v4i1.8543>
- Ojha, A. (2026c). Typed field support with a value-only deterministic executor for synthetic financial records: A multi-seed controlled evaluation. *American Journal of Financial Technology and Innovation*, 4(1), 185–197. <https://doi.org/10.54536/ajfti.v4i1.8551>
- Trask, A., Hill, F., Reed, S. E., Rae, J., Dyer, C., & Blunsom, P. (2018). Neural arithmetic logic units. In *Advances in neural information processing systems* (Vol. 31). <https://proceedings.neurips.cc/paper/2018/hash/0e64a7b00c83e3d22ce6b3acf2c582b6-Abstract.html>
- Yang, C., Wang, X., Jiang, J., Zhang, Q., & Huang, X. (2026). LLM-based world models can make decisions solely, but rigorous evaluations are needed. *Transactions on Machine Learning Research*. <https://openreview.net/forum?id=XmYCERERcD>
- Zhang, W., Weng, T.-W., Das, S., Megretski, A., Daniel, L., & Nguyen, L. M. (2023). ConCerNet: A contrastive learning based framework for automated conservation law discovery and trustworthy dynamical system prediction. In *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 41694–41714). Proceedings of Machine Learning Research. <https://proceedings.mlr.press/v202/zhang23ao.html>
- Zhao, L., Kong, L., Walters, R., & Wong, L. L. S. (2022). Toward compositional generalization in object-

oriented world modeling. In *Proceedings of the 39th International Conference on Machine Learning* (Vol. 162, pp. 26841–26864). Proceedings of Machine Learning Research. <https://proceedings.mlr.press/v162/zhao22b.html>.

Appendix A Controller details and split support

The original ten one-step condition families are regular, single high risk, single expedited, single large amount, single low authentication, sanctions block, closed books, locked affected role, stage mismatch, and capacity block.

Locked-role and capacity-block conditions apply only to the six posting primitives; thus they form 76 cells rather than 80. Twelve of sixteen conjunction cells are added to training. Non-conjunction actions use amounts 2500, 5000, 10000 or 20000 cents; large conditions use 100000 with approval limit 50000. Initial ordinary capacity is the absolute initial balance plus 250000. The capacity-block family sets balances to zero and one posting capacity exactly to the action amount. This restricted training coverage helps explain why the later headroom distribution is a genuine additional shift.

Table 7: Applied primitive semantics, after earlier control checks pass. All amounts are observed positive integers. Roles named in the posting column receive +a and -a, respectively.

Primitive	Allowed stage(s)	Posting (+, -)	Next stage
submit	draft	none	pending
approve	pending	none	authorized
capture	authorized	receivable, customer liability	captured
settle	captured	cash, receivable	settled
merchant payout	settled	merchant payable, cash	draft
fee assessment	captured, settled	receivable, fee revenue	unchanged
refund	captured, settled	customer liability, cash	draft
chargeback	captured, settled	loss expense, cash	draft

Every primary corpus difference and reporting precision

The unadjusted 95% sensitivity interval for D-B is [-0.154351, -0.135557]. It is not the multiplicity-adjusted primary interval.

The unadjusted 95% sensitivity interval for E-B is

[-0.014502, 0.003337]. It is not the multiplicity-adjusted primary interval.

Compute and retained development history

The main collection took 38.62 elapsed minutes in the recorded local environment, including generation, fitting, evaluation and serialization. It ran with one PyTorch CPU

Table 8: Status counts per corpus (ranges when they vary); rollout counts sum all 32 oracle steps. Counts are targets, not independent statistical observations.

Split	Applied	Review	Rejected	Deferred
train	13640	4084	7502	4774
validation	2280	672	1250	798
test_id	2280	672	1250	798
test_ood	0	4000	0	0
rollouts	5689–5914	115–200	6192–6454	0

Table 9: Mean supported ID recall for every status, in percent.

Condition	Applied	Review	Rejected	Deferred
A	100.00	100.00	99.99	100.00
B	100.00	100.00	99.99	100.00
C	99.98	99.99	99.94	99.99
D	100.00	100.00	99.99	100.00
E	100.00	100.00	99.99	100.00

Table 10: Horizon-8 component versus complete-object exactness, in percent. A correct last tuple can coexist with incorrect accumulated balances.

Condition	Status	Stage	Last tuple	Full state	Full trajectory
A	53.06	83.00	16.23	0.09	0.00
B	53.12	82.52	16.13	0.11	0.00
C	52.90	81.63	48.39	19.93	19.22

Condition	Status	Stage	Last tuple	Full state	Full trajectory
D	54.01	83.20	51.43	34.24	33.33
E	53.15	82.29	16.17	0.09	0.00

Table 11: All eight primary corpus differences in mean bounded horizon-8 error. Each averages three paired training seeds; no corpus is excluded.

Corpus seed	D-B	E-B
920301	-0.146174	0.000045
920302	-0.149035	-0.024007
920303	-0.136704	-0.015911
920304	-0.137721	0.002573
920305	-0.168844	-0.007524
920306	-0.147048	-0.001955
920307	-0.132260	0.009641
920308	-0.141847	-0.007525

thread, Python 3.12.10, PyTorch 2.10.0+cpu and NumPy 2.4.2. Other local work could run concurrently, so elapsed duration is not a controlled runtime comparison.

Auxiliary selected checkpoints and validation losses Auxiliary applied and unchanged prefix strata Reproduction and evidence boundaries

The accompanying package retains study.py for

development and study_main.py for the frozen main run; vendor files preserve the original controller, encoders, MLP and cent decoders. Prospective protocols, the single pilot amendment, exact JSON configurations, SHA-256 freezes and runtime metadata precede their corresponding executions. Raw corpus files include initial states, actions and controller targets. Each fit retains its epoch history, chosen checkpoint, checkpoint hash, all prediction rows

Table 12: Retained pilot-v1 development outcomes. The corpus is reused across pilots for the sole budget amendment and excluded from main generation. These rows are not pooled into main inference.

Condition	Seed	ID exact (%)	Applied posting MAE	Minimum status recall (%)
A	601	54.40	10,147.33	100.00
A	607	54.38	10,386.64	99.92
B	601	54.38	10,881.49	99.92
B	607	54.32	10,121.52	99.68
C	601	65.72	8,831.87	99.68
C	607	65.74	8,180.56	99.76
D	601	99.98	0.24	100.00
D	607	99.60	0.00	99.92
E	601	54.40	10,854.60	100.00
E	607	54.30	10,197.89	99.60

Table 13: Retained pilot-v2 development outcomes. The corpus is reused across pilots for the sole budget amendment and excluded from main generation. These rows are not pooled into main inference.

Condition	Seed	ID exact (%)	Applied posting MAE	Minimum status recall (%)
A	601	54.40	10,452.71	100.00
A	607	54.40	10,175.62	100.00
B	601	54.40	4,856.64	100.00
B	607	54.40	4,947.74	100.00
C	601	97.06	427.39	99.68
C	607	93.28	970.27	99.76
D	601	100.00	0.00	100.00
D	607	100.00	0.00	100.00
E	601	54.40	4,527.93	100.00
E	607	54.40	5,213.80	100.00

Table 14: All twelve selected checkpoints. Component losses are unweighted; total loss is numeric +0.5 stage +0.5 status. Numeric loss differs between regression and sign heads, so its cross-head magnitude is not directly comparable. Every model trained for 100 epochs; the selected epoch can be early under the common mixed validation distribution.

Training	Arm	Seed	Epoch	Val. total	Numeric	Stage	Status	Epoch100 val.
augmented	B	811	99	0.2619	0.0086	0.2269	0.2797	0.2637
augmented	B	821	99	0.2731	0.0083	0.2373	0.2923	0.2787
augmented	D	811	99	0.3462	0.0914	0.2261	0.2836	0.3509
augmented	D	821	99	0.3325	0.0893	0.2204	0.2660	0.3350
augmented	E	811	99	0.2652	0.0068	0.2268	0.2900	0.2679
augmented	E	821	98	0.2759	0.0079	0.2324	0.3036	0.2803
original	B	811	2	0.9259	0.0151	0.8551	0.9664	7.9278
original	B	821	3	0.9181	0.0163	0.6718	1.1317	8.1309
original	D	811	3	1.1935	0.3428	0.7304	0.9711	9.4470
original	D	821	4	1.1938	0.3083	0.6547	1.1163	9.6472
original	E	811	2	0.9258	0.0147	0.8560	0.9661	6.8573
original	E	821	3	0.9271	0.0152	0.6737	1.1501	8.0110

Table 15: Descriptive auxiliary prefix strata, averaging the two fixed seeds. Applied n is the number of true-applied nonzero postings among 384 trajectories at that prefix. TF uses oracle-state one-step inputs; its numerical error is cents per balance. Carried columns measure complete accumulated state within each true-outcome stratum. No-post includes zero-posting applied primitives and non-applied outcomes. Empty strata are marked – and never treated as zero error

Training	Arm	h	Applied n	TF tuple (%)	TF delta MAE	Carried (%)	applied	Carried no-post (%)
original	B	1	140	0.00	3,277.06	0.00		7.17
original	B	2	127	0.00	3,063.07	0.00		3.31
original	B	8	190	0.00	2,896.41	0.00		0.00
original	B	32	10	0.00	10,306.09	0.00		0.00
original	D	1	140	0.36	3,080.36	0.36		6.35
original	D	2	127	0.79	2,892.06	0.39		3.50
original	D	8	190	0.53	2,922.15	0.26		1.80
original	D	32	10	0.00	6,625.00	0.00		0.00
original	E	1	140	0.00	3,230.50	0.00		6.97
original	E	2	127	0.00	3,018.40	0.00		3.31
original	E	8	190	0.00	2,911.41	0.00		0.26
original	E	32	10	0.00	9,837.95	0.00		0.00
augmented	B	1	140	0.00	4,657.15	0.00		5.33
augmented	B	2	127	0.00	4,404.03	0.00		3.50
augmented	B	8	190	0.00	6,228.52	0.00		0.52
augmented	B	32	10	0.00	11,726.11	0.00		0.00
augmented	D	1	140	92.14	238.10	92.14		59.22
augmented	D	2	127	90.16	173.88	85.83		56.03
augmented	D	8	190	87.63	207.24	83.68		3.35
augmented	D	32	10	80.00	4,187.50	0.00		32.35
augmented	E	1	140	0.00	3,989.85	0.00		5.53
augmented	E	2	127	0.00	3,857.10	0.00		3.50
augmented	E	8	190	0.00	6,182.32	0.00		0.52
augmented	E	32	10	0.00	12,029.15	0.00		0.00

and summaries. Reference rows are saved under the same conventions. A completion record lists every scheduled fit; no successful-only subset is selected.

Nine main tests and four auxiliary tests cover controller boundaries, support/disjointness, canonical equivariance and exact rounding, model-carried accumulation, failure handling, pilot exclusion, randomized training order, routing ties, matched capacity pairs and teacher-prefix exclusion. Generation checks every target against the independently written rule table. The prespecified analyzer checks complete corpus/seed/arm identities and reconstructs primary endpoints from raw balances. A separate artifact audit reconstructs both accumulated paths from initial states and recorded deltas, checks every saved summary, and validates checkpoint hashes. It does not rerun optimization. The figure and table builder reads those verified summaries; no result numbers are manually substituted.

The package documentation provides PowerShell-safe commands to inspect hashes, run tests, reproduce numerical analysis, regenerate manuscript tables/figures and train into a new output directory. Re-executed training can depend on library, hardware and floating-

point environment; the original runtime is retained in the freeze. Frozen source and numerical artifacts are distinct from prose, visualization and audit scripts written after outcomes. The latter are labeled post-collection reporting.

Appendix B Post collection reviewer analyses

All tables in this appendix are October 11, 2026 post-collection analyses of the original eight corpora and fixed seed trio. Percentages refer to different stated correctness objects. Family rows each aggregate 3,072 predictions per horizon across 24 fits, with 128 trajectories per family per fit. Teacher inputs are oracle states; carried values come from each model's own evolving state.

Appendix C Additional controlled training study

Each condition averages 24 fits across eight original corpora and the three fixed paired seeds. Exactness columns are percentages; MAE is cents per balance. The original fixed primary study remains separate. New capacity probes have 2,000 cases per fit and 1,000 balance-flip pairs. All intervals below are exploratory 95% corpus-t intervals without a new familywise guarantee.

Appendix D Secondary correctness uncertainty

Table 23 is an October 12, 2026 post-collection reporting

Table 16: Main oracle-state teacher-forced transitions and model-carried states for every family and scheduled horizon. TF is current transition tuple exactness and local posting MAE in cents; carried X is accumulated full-state exactness and carried B is bounded accumulated six-balance error. There were no nonfinite teacher predictions.

Arm	Family	h	TF tuple %	TF MAE cents	Carried X %	Carried B
A	Cycle	1	0.00	2,143.57	0.00	0.0214
A	Cycle	2	0.00	2,324.72	0.00	0.0413
A	Cycle	8	0.00	4,515.08	0.00	0.1734
A	Cycle	32	0.00	2,348.99	0.00	0.6788
A	Fee capacity	1	5.37	2,669.27	5.37	0.0267
A	Fee capacity	2	9.08	2,599.28	1.46	0.0527
A	Fee capacity	8	43.95	1,489.26	0.00	0.1707
A	Fee capacity	32	99.80	1.63	0.00	0.2315
A	Mixed	1	8.53	2,573.87	8.53	0.0257
A	Mixed	2	8.11	2,444.10	4.39	0.0470
A	Mixed	8	8.14	1,753.63	0.26	0.1378
A	Mixed	32	5.47	1,987.85	0.00	0.4448
B	Cycle	1	0.00	3,035.39	0.00	0.0304
B	Cycle	2	0.00	2,790.11	0.00	0.0436
B	Cycle	8	0.00	4,135.19	0.00	0.1483
B	Cycle	32	0.00	2,844.58	0.00	0.5877
B	Fee capacity	1	5.37	2,669.27	5.37	0.0267
B	Fee capacity	2	9.08	2,599.28	1.46	0.0527
B	Fee capacity	8	43.95	1,489.26	0.00	0.1707
B	Fee capacity	32	99.80	1.63	0.00	0.2315
B	Mixed	1	8.63	7,068.36	8.63	0.0707
B	Mixed	2	8.04	7,458.48	4.33	0.1281
B	Mixed	8	7.91	7,465.75	0.33	0.3740
B	Mixed	32	5.44	7,776.13	0.00	0.8288
C	Cycle	1	99.71	0.00	99.71	0.0000
C	Cycle	2	99.74	0.00	99.45	0.0000

Arm	Family	h	TF tuple %	TF MAE cents	Carried X %	Carried B
C	Cycle	8	94.04	102.54	57.68	0.0099
C	Cycle	32	99.51	0.00	50.03	0.0538
C	Fee capacity	1	5.37	2,669.27	5.37	0.0267
C	Fee capacity	2	9.05	2,599.28	1.46	0.0527
C	Fee capacity	8	43.91	1,489.26	0.00	0.1707
C	Fee capacity	32	99.74	1.63	0.00	0.2315
C	Mixed	1	15.20	1,718.75	15.20	0.0172
C	Mixed	2	15.53	1,559.11	10.90	0.0326
C	Mixed	8	13.70	866.83	2.12	0.0929
C	Mixed	32	10.22	1,141.76	0.07	0.2727
D	Cycle	1	100.00	0.00	100.00	0.0000
D	Cycle	2	100.00	0.00	100.00	0.0000
D	Cycle	8	100.00	0.00	100.00	0.0000
D	Cycle	32	100.00	0.00	100.00	0.0000
D	Fee capacity	1	5.37	2,669.27	5.37	0.0267
D	Fee capacity	2	9.08	2,599.28	1.46	0.0527
D	Fee capacity	8	43.95	1,489.26	0.00	0.1707
D	Fee capacity	32	99.80	1.63	0.00	0.2315
D	Mixed	1	21.19	1,317.68	21.19	0.0132
D	Mixed	2	21.26	1,191.00	15.07	0.0265
D	Mixed	8	17.19	725.64	2.73	0.0875
D	Mixed	32	12.92	1,005.18	0.03	0.2942
E	Cycle	1	0.00	2,739.99	0.00	0.0274
E	Cycle	2	0.00	2,409.71	0.00	0.0361
E	Cycle	8	0.00	4,151.68	0.00	0.1367
E	Cycle	32	0.00	2,448.08	0.00	0.5398
E	Fee capacity	1	5.37	2,669.27	5.37	0.0267
E	Fee capacity	2	9.08	2,599.28	1.46	0.0527
E	Fee capacity	8	43.95	1,489.26	0.00	0.1707
E	Fee capacity	32	99.80	1.63	0.00	0.2315
E	Mixed	1	8.66	6,818.95	8.66	0.0682
E	Mixed	2	8.11	7,131.64	4.23	0.1246
E	Mixed	8	7.88	7,051.41	0.26	0.3690
E	Mixed	32	5.57	7,118.70	0.00	0.8424

Table 17: Descriptive horizon-8 family-weighting sensitivity. Weight order is cycle, fee-capacity, mixed. Half scenarios assign one half to the named family and one quarter to each other family. Negative contrasts favor the first condition. Weights are constructed sensitivity scenarios, not deployment prevalence.

Scenario	Family weights	D minus B	E minus B
Equal thirds	0.333, 0.333, 0.333	-0.14495	-0.00558
Cycle half	0.500, 0.250, 0.250	-0.14580	-0.00711
Fee half	0.250, 0.500, 0.250	-0.10872	-0.00419
Mixed half	0.250, 0.250, 0.500	-0.18034	-0.00546
Cycle only	1.000, 0.000, 0.000	-0.14834	-0.01168
Fee only	0.000, 1.000, 0.000	0.00000	0.00000
Mixed only	0.000, 0.000, 1.000	-0.28652	-0.00507

Table 18: Fixed-checkpoint centering and integer-decoder ablation at horizon 8. Every condition averages 24 fits and carries its own intervened state on identical exogenous actions. Original B and E are unchanged original results. Nonfinite predictions were zero in all interventions.

Inference condition	State MAE cents	Bounded B	State X %	Conserved %	Capacity invalid %
B original	23,291.74	0.231038	0.1085	35.68	0.77
B centered independent	23,218.74	0.230329	0.1085	51.71	0.77
B centered conserving	23,218.74	0.230329	0.1085	100.00	0.77
E centered independent	22,696.22	0.225455	0.0868	52.21	0.71
E original conserving	22,696.22	0.225455	0.0868	100.00	0.71

analysis of the original eight corpora and fixed three-seed trio. Each percentage mean and interval uses eight corpus observations after averaging paired seeds within corpus. Intervals are exploratory unadjusted 95% corpus-t intervals, not binomial intervals over prediction rows.

They are approximate and unclipped; endpoints outside 0–100% and degenerate intervals expose the normal-mean and finite-sample limitations. No familywise coverage or population certainty is asserted.

Table 19: All follow-up conditions. ID numerical error conditions on true-applied, nonzero postings. Ordinary gates are the unchanged numerical/status gate for regressors and ID exact/status gate for the sign head; passing does not certify capacity competence.

Condition	Params	ID exact %	ID posting MAE	H8 B	H8 X %	H32 X %	Ordinary gates
B67	10869	54.3992	5374.64	0.22207	0.09	0.00	24/24
B025	10191	54.3967	8066.03	0.21816	0.12	0.00	20/24
B4	10191	54.3992	3344.03	0.21486	0.09	0.00	24/24
DBrier	10971	100.0000	0.00	0.08799	34.22	33.34	24/24
RoleProjected	10963	54.4000	934.42	0.13719	0.14	0.00	24/24
BSequenceFlat	10191	54.3992	5622.93	0.10283	8.79	1.18	24/24
BSequenceTF	10191	54.3992	5470.58	0.10422	8.83	1.19	24/24
BSequenceCarry	10191	54.3992	6487.23	0.11024	8.75	1.22	24/24

Table 20: Follow-up bounded horizon-8 contrasts. Negative differences favor the first condition. Flat versus TF is an implementation/reduction control; its observed difference limits isolated temporal interpretation. Original-condition comparisons are the stated architecture/objective or broader training packages.

Contrast	Mean difference	Descriptive 95% interval
B67 minus B	-0.008972	[-0.020805, 0.002861]
B025 minus B	-0.012879	[-0.019340, -0.006418]
B4 minus B	-0.016174	[-0.024233, -0.008115]
DBrier minus D	0.001910	[-0.001410, 0.005230]
RoleProjected minus E	-0.088263	[-0.100050, -0.076476]
BSequenceCarry minus BSequenceFlat	0.007407	[-0.004035, 0.018849]
BSequenceCarry minus BSequenceTF	0.006022	[-0.006135, 0.018179]
BSequenceTF minus BSequenceFlat	0.001385	[-0.002324, 0.005093]

Table 21: Additional held-out balance-flip capacity-status probe, with original B/D/E checkpoints evaluated on the identical probe. The gate requires at least 95% recall for each status and 90% both-members-correct pairs in each fit. Aggregate means cannot waive a failed per-fit gate.

Condition	Applied recall %	Rejected recall %	Both pair %	Minimum fit pair %	Gate passes
Original E	19.18	80.44	0.00	0.00	0/24
Original D	19.13	80.44	0.00	0.00	0/24
Original B	18.81	80.44	0.00	0.00	0/24
B67	19.32	80.44	0.00	0.00	0/24
B025	17.57	80.44	0.00	0.00	0/24
B4	19.08	80.44	0.00	0.00	0/24
DBrier	18.99	80.44	0.00	0.00	0/24

Condition	Applied recall %	Rejected recall %	Both pair %	Minimum fit pair %	Gate passes
RoleProjected	19.56	80.44	0.00	0.00	0/24
BSequenceFlat	42.04	58.62	1.22	0.10	0/24
BSequenceTF	40.98	59.67	1.23	0.10	0/24
BSequenceCarry	47.93	52.40	0.75	0.00	0/24

Table 22: Matched sequence-suite oracle-state transition exactness versus model-carried full-state exactness and bounded accumulated error, every family and scheduled horizon. Each row aggregates 3,072 teacher predictions. These percentages measure different objects and are not a causal decomposition of accumulation.

Condition	Family	h	Teacher tuple %	Carried state %	Carried B
BSequenceFlat	cycle	1	0.00	0.00	0.01848
BSequenceFlat	cycle	2	0.00	0.00	0.02941
BSequenceFlat	cycle	8	0.00	0.00	0.10074
BSequenceFlat	cycle	32	0.00	0.00	0.42136
BSequenceFlat	fee capacity	1	5.05	5.05	0.02304
BSequenceFlat	fee capacity	2	6.74	1.40	0.04062
BSequenceFlat	fee capacity	8	17.15	0.00	0.13608
BSequenceFlat	fee capacity	32	22.43	0.00	0.38861
BSequenceFlat	mixed	1	78.19	78.19	0.01579
BSequenceFlat	mixed	2	80.31	62.99	0.02814
BSequenceFlat	mixed	8	86.56	26.37	0.07168
BSequenceFlat	mixed	32	87.89	3.55	0.16966
BSequenceTF	cycle	1	0.00	0.00	0.01865
BSequenceTF	cycle	2	0.00	0.00	0.02997
BSequenceTF	cycle	8	0.00	0.00	0.10355
BSequenceTF	cycle	32	0.00	0.00	0.41873
BSequenceTF	fee capacity	1	4.98	4.98	0.02300
BSequenceTF	fee capacity	2	6.77	1.37	0.04074
BSequenceTF	fee capacity	8	17.55	0.00	0.13864
BSequenceTF	fee capacity	32	23.08	0.00	0.38238
BSequenceTF	mixed	1	78.35	78.35	0.01554
BSequenceTF	mixed	2	80.40	63.18	0.02767
BSequenceTF	mixed	8	86.52	26.50	0.07046
BSequenceTF	mixed	32	87.92	3.58	0.16586
BSequenceCarry	cycle	1	0.00	0.00	0.02523
BSequenceCarry	cycle	2	0.00	0.00	0.04012
BSequenceCarry	cycle	8	0.00	0.00	0.12416
BSequenceCarry	cycle	32	0.00	0.00	0.41651
BSequenceCarry	fee capacity	1	4.95	4.95	0.02865
BSequenceCarry	fee capacity	2	6.32	1.37	0.04563
BSequenceCarry	fee capacity	8	13.15	0.00	0.13426
BSequenceCarry	fee capacity	32	17.09	0.00	0.35786
BSequenceCarry	mixed	1	78.19	78.19	0.01613
BSequenceCarry	mixed	2	80.37	63.05	0.02829
BSequenceCarry	mixed	8	86.52	26.24	0.07229
BSequenceCarry	mixed	32	87.86	3.65	0.17368

Table 23: Secondary teacher-forced transition-tuple and carried full-state exactness, percentage means and exploratory 95% corpus-t intervals for every original arm, family and scheduled horizon. The two correctness objects remain distinct.

Arm	Family	h	TF mean %	TF interval %	Carried mean %	Carried interval %
A	Cycle	1	0.00	[0.00, 0.00]	0.00	[0.00, 0.00]
A	Cycle	2	0.00	[0.00, 0.00]	0.00	[0.00, 0.00]
A	Cycle	8	0.00	[0.00, 0.00]	0.00	[0.00, 0.00]
A	Cycle	32	0.00	[0.00, 0.00]	0.00	[0.00, 0.00]
A	Fee capacity	1	5.37	[3.79, 6.95]	5.37	[3.79, 6.95]
A	Fee capacity	2	9.08	[6.93, 11.23]	1.46	[0.51, 2.42]
A	Fee capacity	8	43.95	[39.19, 48.70]	0.00	[0.00, 0.00]
A	Fee capacity	32	99.80	[99.34, 100.27]	0.00	[0.00, 0.00]
A	Mixed	1	8.53	[6.69, 10.37]	8.53	[6.69, 10.37]
A	Mixed	2	8.11	[6.15, 10.06]	4.39	[3.71, 5.08]
A	Mixed	8	8.14	[6.83, 9.45]	0.26	[-0.02, 0.55]
A	Mixed	32	5.47	[4.57, 6.36]	0.00	[0.00, 0.00]
B	Cycle	1	0.00	[0.00, 0.00]	0.00	[0.00, 0.00]
B	Cycle	2	0.00	[0.00, 0.00]	0.00	[0.00, 0.00]
B	Cycle	8	0.00	[0.00, 0.00]	0.00	[0.00, 0.00]
B	Cycle	32	0.00	[0.00, 0.00]	0.00	[0.00, 0.00]
B	Fee capacity	1	5.37	[3.79, 6.95]	5.37	[3.79, 6.95]
B	Fee capacity	2	9.08	[6.93, 11.23]	1.46	[0.51, 2.42]
B	Fee capacity	8	43.95	[39.19, 48.70]	0.00	[0.00, 0.00]
B	Fee capacity	32	99.80	[99.34, 100.27]	0.00	[0.00, 0.00]
B	Mixed	1	8.63	[6.80, 10.45]	8.63	[6.80, 10.45]
B	Mixed	2	8.04	[6.14, 9.94]	4.33	[3.76, 4.90]
B	Mixed	8	7.91	[6.77, 9.05]	0.33	[0.00, 0.65]
B	Mixed	32	5.44	[4.41, 6.47]	0.00	[0.00, 0.00]
C	Cycle	1	99.71	[99.57, 99.85]	99.71	[99.57, 99.85]
C	Cycle	2	99.74	[99.39, 100.09]	99.45	[99.09, 99.80]
C	Cycle	8	94.04	[92.12, 95.96]	57.68	[50.50, 64.86]
C	Cycle	32	99.51	[99.00, 100.02]	50.03	[43.06, 57.00]
C	Fee capacity	1	5.37	[3.79, 6.95]	5.37	[3.79, 6.95]
C	Fee capacity	2	9.05	[6.89, 11.21]	1.46	[0.51, 2.42]
C	Fee capacity	8	43.91	[39.17, 48.66]	0.00	[0.00, 0.00]
C	Fee capacity	32	99.74	[99.27, 100.21]	0.00	[0.00, 0.00]
C	Mixed	1	15.20	[13.13, 17.27]	15.20	[13.13, 17.27]
C	Mixed	2	15.53	[12.96, 18.09]	10.90	[9.20, 12.61]
C	Mixed	8	13.70	[11.74, 15.67]	2.12	[1.39, 2.84]
C	Mixed	32	10.22	[8.42, 12.02]	0.07	[-0.04, 0.17]
D	Cycle	1	100.00	[100.00, 100.00]	100.00	[100.00, 100.00]
D	Cycle	2	100.00	[100.00, 100.00]	100.00	[100.00, 100.00]
D	Cycle	8	100.00	[100.00, 100.00]	100.00	[100.00, 100.00]
D	Cycle	32	100.00	[100.00, 100.00]	100.00	[100.00, 100.00]
D	Fee capacity	1	5.37	[3.79, 6.95]	5.37	[3.79, 6.95]
D	Fee capacity	2	9.08	[6.93, 11.23]	1.46	[0.51, 2.42]
D	Fee capacity	8	43.95	[39.19, 48.70]	0.00	[0.00, 0.00]
D	Fee capacity	32	99.80	[99.34, 100.27]	0.00	[0.00, 0.00]

Arm	Family	h	TF mean %	TF interval %	Carried mean %	Carried interval %
D	Mixed	1	21.19	[18.48, 23.90]	21.19	[18.48, 23.90]
D	Mixed	2	21.26	[18.26, 24.25]	15.07	[12.37, 17.77]
D	Mixed	8	17.19	[14.29, 20.09]	2.73	[1.67, 3.79]
D	Mixed	32	12.92	[11.28, 14.56]	0.03	[-0.04, 0.11]
E	Cycle	1	0.00	[0.00, 0.00]	0.00	[0.00, 0.00]
E	Cycle	2	0.00	[0.00, 0.00]	0.00	[0.00, 0.00]
E	Cycle	8	0.00	[0.00, 0.00]	0.00	[0.00, 0.00]
E	Cycle	32	0.00	[0.00, 0.00]	0.00	[0.00, 0.00]
E	Fee capacity	1	5.37	[3.79, 6.95]	5.37	[3.79, 6.95]
E	Fee capacity	2	9.08	[6.93, 11.23]	1.46	[0.51, 2.42]
E	Fee capacity	8	43.95	[39.19, 48.70]	0.00	[0.00, 0.00]
E	Fee capacity	32	99.80	[99.34, 100.27]	0.00	[0.00, 0.00]
E	Mixed	1	8.66	[6.87, 10.45]	8.66	[6.87, 10.45]
E	Mixed	2	8.11	[6.11, 10.10]	4.23	[3.65, 4.81]
E	Mixed	8	7.88	[6.75, 9.00]	0.26	[-0.09, 0.61]
E	Mixed	32	5.57	[4.55, 6.58]	0.00	[0.00, 0.00]