RETENTION IS NOT ENOUGH: SELECTIVE READOUT IN BOUNDED HYBRID MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Correct retention and correct slot selection need not produce a correct answer. We examine this distinction in typed online streams containing shared rules and sparse overrides, with a common 256-byte persistent-state budget for learned systems. Across three frozen workflow fits, 772 of 4,602 exception-query evaluations remain incorrect after decoded-content and addressing checks pass; replacing soft retrieval with the stored top-1 payload rescues only 30. On a separate query subset, masking nonmatching payloads raises rule accuracy from 60.15% to 86.93% at fixed weights, with little change to exception accuracy. Training shared predictors with either a hard match gate or a one-scalar soft-null option largely repairs the failure, without full parameter isolation. The hard gate reaches 99.959% on original-task streams of 8,192 events; the soft null is correct on every sampled workflow query at lengths 2,048 and 8,192. These controlled, prior-informed tasks admit smaller exact programs. Our contribution is a query-aligned diagnosis of retained content, slot selection and downstream use, together with tests of simple rejection-capable readers, rather than a new memory or attention primitive.

1 INTRODUCTION

Online predictors often apply a shared rule to most entities while retaining a few exact overrides. A policy change updates the default; an exception changes only its named entity. Recurrent state can compress the shared rule, while addressable memory can retain sparse bindings. Both are established choices in recurrent and memory-augmented models (Cho et al., 2014; Graves et al., 2014; 2016). Their combination raises a question beyond storage: when should retrieved information influence the output? *A memory can be correct and still be the wrong input to a prediction.*

Consider a correctly stored exception for entity a and a query about entity b, which has no exception. Softmax over occupied slots returns a mixture even when none applies to b. The exception need not be deleted; it should simply not affect this prediction. No-op attention and scope-aware retrieval are established ideas (Bondarenko et al., 2023; Qiu et al., 2025; Mitchell et al., 2022). We instead ask how to separate retention failures from use failures in a fixed-state learner, and whether a small readout change can repair the latter without removing parameter sharing. Here *rejection* means suppressing retrieved value input, not refusing to answer or erasing stored state.

We organize the study around four alternative explanations (Table 1): missing exception content, incorrect slot selection or mixing, harmful use of inapplicable content, and a need for complete parameter separation. Query-aligned integrity checks and fixed-checkpoint interventions test the first three; separately trained readers test the fourth. The checks concern different query subsets, so their effect sizes are not collapsed into a single error decomposition.

The resulting contribution is controlled diagnostic evidence, not the invention of retrieval rejection. A hard match gate and a one-scalar soft-null reader repair the tested shared predictors; full path isolation provides a stronger protected reference. Hard gating is evaluated on both generators, while soft null is evaluated on workflow and uses the same similarity prior. We use *composition* to mean selecting an applicable state source within one online predictor, not algebraically combining both memories in every answer.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Candidate explanation	Existing test	Supported inference
Exception errors reflect absent or incorrectly decoded content	Query-aligned integrity: 772 errors remain in J	The checked missing-content failures do not explain those errors.
Residual errors are entirely due to immediate soft mixing	Stored top-1 payload at fixed weights rescues 30 of 772	Lookup hardening alone is insufficient at these checkpoints.
Nonmatching values do not affect rule answers	Fixed-weight masking: rule accuracy 60.15% $\rightarrow$ 86.93%	The inapplicable payload path affects these predictions.
Repair requires complete parameter separation	Retrained shared hard-gate and soft-null readers nearly/perfectly solve the tested workflow arrays	Full isolation is not necessary for the observed repair.

Table 1: **Four tests of retention and use.** Each row has its own query subset or training comparison. Together they constrain explanations of the observed failure; they do not identify a unique historical optimization cause. Detailed diagnostics are in Section 4.2.

2 TASK AND RESOURCE MODEL

2.1 A SHARED RULE WITH PERSISTENT OVERRIDES

An episode presents events $x_{1:T}$, then queries about entities. Each event contains an observable type, a descriptor and a value. Types distinguish rule observations, exception writes and distractors. The primary setting has 32 entities, $K = 3$ active exceptions and a known family of 16 policies. Let p_T be the final policy and $\mathcal{E}_T$ the logical active-exception map. The target for entity k is

$$y_T(k) = \begin{cases} \mathcal{E}_T(k), & k \in \mathrm{dom}(\mathcal{E}_T), \\ f_{p_T}(k), & \text{otherwise.} \end{cases} \tag{1}$$

Writes replace an entity's previous override. Query subtype and gold exception membership are not model inputs. Solvers infer matches from their stored keys and the observable query descriptor. Both task families use fixed entity descriptors and known finite rule families; they do not test discovery of an unseen rule system at evaluation time.

The original generator changes a modular shared rule in 32-event phases. Workflow instead uses a nonlinear policy table, stochastic policy changes and asynchronous exception overwrites. Its update schedule is not a renamed copy of the original generator. In both cases active exception identities persist within an episode and fit the three-slot bank. Appendix D gives the observation encodings, event schedules and target construction, including the original generator's final-policy-conditioned exception values.

2.2 STATE, PARAMETERS AND TASK KNOWLEDGE

A solver processes each event once:

$$S_t = U_\theta(S_{t-1}, x_t), \qquad \mathrm{bytes}(S_t) \le B, \qquad \hat{y}_T = R_\theta(S_T, q). \tag{2}$$

Discarded events cannot be reread at query time. Static parameters and fixed task constants are distinguished from episode-dependent state and temporary computation. For the main learned comparisons, $B = 256$ bytes. The hybrid allocates 64 bytes to 16 FP32 recurrent values and 192 bytes to three slots. Each slot contains eight FP32 key values, seven FP32 payload values and one FP32 timestamp, which also encodes occupancy. The recurrent-only control has 64 FP32 values; Dual-GRU partitions those values between two states. The gated-delta adaptation retains one 8×8 FP32 matrix. The query-side match gate adds no persistent state.

Two executable non-learning references use the task structure more directly: a 62-byte original-task rule/table program with 2,048 bytes of static prototypes, and a 7-byte workflow program with 2,560 bytes of static lookup constants. Both achieve 100% on their reported checks. They are not hidden-label oracles, and they are not 256-byte learned controls. Their success limits our question to learned representation and use, rather than optimal algorithmic storage or intrinsic task difficulty.

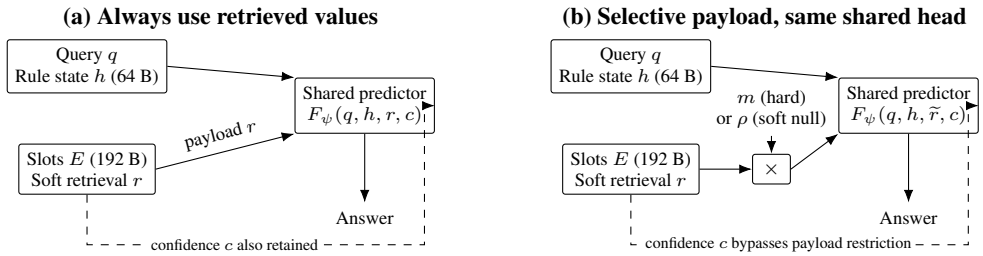


Figure 1: **Selective use changes the read, not the stored contents.** Both designs retain a 64-byte recurrent state and three 64-byte slots. The hard gate zeros the seven-value payload on a descriptor miss; the soft null attenuates it by the real-slot attention mass in Equation (7). Query, state and real-slot confidence still reach the shared head. These operations do not erase bindings or establish independence from the entire memory bank. Path isolation is a separate, stronger comparator.

3 SELECTIVE READOUT OF HYBRID STATE

3.1 TYPE-AWARE UPDATES AND ADDRESSABLE RETRIEVAL

The recurrent state h_t is updated on rule events and held fixed otherwise. Slots E_t are updated on exception events and held fixed otherwise. These are controlled, observable event types, also available to the type-aware baselines. Learned projections encode descriptor keys and value payloads. A write overwrites a sufficiently similar occupied key or selects a free/least-recently-written slot; the precise routine is in Appendix D.

For a query projection $\bar{q}$ and occupied normalized keys k_j, the retrieval is

$$s_j = \langle \bar{q}, k_j \rangle, \quad a_j = \frac{\exp(20 s_j)}{\sum_{i \in \mathcal{O}} \exp(20 s_i)}, \quad r(q, E) = \sum_{j \in \mathcal{O}} a_j v_j, \quad c(q, E) = \max_{j \in \mathcal{O}} s_j, \tag{3}$$

where $\mathcal{O}$ is the occupied-slot set. An empty bank returns $r = 0, c = -1$. Because the softmax distributes mass over occupied slots, it can return a nonzero payload when no key is a satisfactory match. Slot contents are explicitly retained, but learned keys and soft retrieval do not constitute guaranteed exact lookup.

3.2 MINIMAL REPAIR: GATE THE VALUE, RETAIN THE SHARED PREDICTOR

The original coupled predictor computes

$$\hat{y}_{\text{coupled}} = \arg\max_y F_\psi(q, h_T, r(q, E_T), c(q, E_T))_y. \tag{4}$$

Its query, recurrent state, retrieved payload and confidence all enter one head. Define a fixed descriptor match

$$m(q, E) = \mathbb{1}\{c(q, E) > \tau\}, \qquad \tau = 0.97, \quad \widetilde{r}(q, E) = m(q, E) r(q, E). \tag{5}$$

Our cross-generator reference, the *match-gated coupled hybrid*, replaces only r by $\widetilde{r}$ in Equation (4). On a miss, the seven retrieved payload values are zero. Query, recurrent state and confidence are retained. The shared embeddings, slot writer and coupled prediction head are unchanged; the workflow implementation retains the coupled model's 6,615 static parameters. The gate is used during training as well as inference and is not trained on query-subtype labels.

This restricts the value-to-predictor path, not the state layout. On a no-match query, the local gradient along this payload-read edge is zero:

$$\left.\frac{\partial \ell}{\partial r}\right|_{\text{payload edge}} = m \, \frac{\partial \ell}{\partial \widetilde{r}} = 0 \quad (m = 0). \tag{6}$$

Confidence, query and recurrent-state paths remain active; shared parameters can still receive gradients through them and through other examples. Retaining parameter sharing therefore does not retain the same gradient coupling. Retraining with the gate can change learned representations as

well as predictions. At fixed weights we test only the inference-time payload effect; the separately trained comparison measures the combined training and inference effect. The ungated head already receives confidence and could in principle learn rejection, so its observed failure is not an architectural impossibility.

3.3 Soft rejection through a null candidate

The soft-null comparator retains the coupled writer, head and confidence input, but adds an unwritable zero-payload candidate to the read normalization. Its sole new parameter is a static logit $\ell_\varnothing$, not an episode-dependent key. For a nonempty bank, let

$$Z(q,E)=\sum_{j\in\mathcal{O}} e^{20s_j}, \qquad \rho(q,E)=\frac{Z(q,E)}{Z(q,E)+e^{\ell_\varnothing}}, \qquad r_{\text{null}}=\rho(q,E)\,r(q,E). \tag{7}$$

An empty bank yields zero payload. The null stores no episode information, so persistent state remains 256 bytes; static parameter count increases from 6,615 to 6,616. Unlike hard masking, finite-logit null attention generally attenuates rather than exactly zeros the real-slot payload. Neither operation deletes slots or removes every memory-dependent input: confidence still reaches the shared head.

Equivalently, the real-slot mass is a smooth gate on total similarity evidence:

$$\rho=\operatorname{sigmoid}\big(\operatorname{logsumexp}_{j\in\mathcal{O}}(20s_j)-\ell_\varnothing\big)\,. \tag{8}$$

It depends on all occupied scores, whereas the hard gate depends on their maximum. Only when one slot dominates does $\rho\approx\operatorname{sigmoid}(20c-\ell_\varnothing)$. This algebraic rewrite is not a new attention mechanism. The null logit starts at $20\times 0.97=19.4$, reusing the hard gate's similarity prior. E01 tests one initialization on workflow; it does not test prior-free learning or initialization robustness.

3.4 Protected comparison: path isolation

The path-isolated variant uses the same match criterion but selects between separate predictors:

$$\hat{y}_{\text{isolated}}=\begin{cases}\arg\max_y F_{\theta_{\text{e}}}(r(q,E_T))_y, & m(q,E_T)=1,\\ \arg\max_y F_{\theta_{\text{r}}}(q,h_T)_y, & m(q,E_T)=0,\end{cases} \tag{9}$$

with disjoint rule and exception parameter paths. The switch selects an output branch, not a single slot: the exception head still receives the soft retrieval. This variant has 7,143 static parameters and the same 256-byte state. We retain it as a stronger protection control, not as a necessary ingredient of the minimal repair. Appendix E distinguishes its conditional independence from the weaker payload exclusion in Equation (5).

3.5 Training and comparison scope

Coupled, hard-gated, soft-null and joint-isolated fits use the natural-then-balanced curriculum. For per-query cross-entropy ℓ_i, gold rule/exception training subsets $I_{\text{r}}, I_{\text{e}}$, and update u, the objective is

$$\mathcal{L}_u=\begin{cases}|I|^{-1}\sum_{i\in I}\ell_i, & u\le 2{,}000,\\ \frac{1}{2|I_{\text{r}}|}\sum_{i\in I_{\text{r}}}\ell_i+\frac{1}{2|I_{\text{e}}|}\sum_{i\in I_{\text{e}}}\ell_i, & u>2{,}000.\end{cases} \tag{10}$$

Subtype supervision supplies loss weights, not inference-time membership labels. Each joint fit is allocated 4,000 updates; evaluated checkpoints follow the frozen development-selection rule. Selected steps and effective target counts are recorded separately in Appendix C. Comparisons match persistent state, not static parameters, FLOPs or wall time.

4 Retention, use and repair

4.1 Evidence blocks and metrics

We distinguish earlier common-array representation comparisons, frozen D02 interventions, and separately trained reader confirmations, including the E01 soft-null study. Their checkpoint and

Frozen payload on the same J subset	Correct	Wrong	Accuracy
Ordinary soft retrieval	3,830	772	83.225%
Stored top-1 payload	3,860	742	83.877%

Table 2: **Correct content and top-1 addressing do not ensure correct use.** The same 4,602 exception-query evaluations pass all three discrete integrity checks in J. The label-free top-1 intervention rescues 30 errors and harms none within J, improving accuracy by 0.652 points. This conditions on integrity, unlike the rule-query intervention in Table 3. Counts pool three frozen coupled fits; they are not independent episode counts.

dataset identities remain separate (Appendix A). The two generators train their own models; cross-generator success is not zero-shot transfer of a checkpoint.

Main neural studies use three fitted seeds and training length 128. Evaluation extends to 8,192 events. We report natural query accuracy, rule/exception accuracy, and *joint-episode accuracy*: the fraction of episodes with all 32 queries correct. At $K = 3$, natural accuracy is $(29/32)A_{\mathrm{r}} + (3/32)A_{\mathrm{e}}$. Standard deviations describe fitted-seed variation. Paired uncertainty is calculated at the whole-episode level when prediction exports are available; queries sharing a state are not independent samples. Each comparison identifies its own common arrays. Older comparisons, D02 and E01 are not pooled; identical training-seed labels do not imply identical test sets.

Representation context. On the earlier original-task arrays, the ungated 256-byte hybrid reaches 98.20% at length 8,192, versus 90.62% for GRU. GRU answers all rule queries and no exception queries correctly; the hybrid retains both capabilities, but only 57.81% of its complete query sets are correct. These starting-point comparisons, including type-aware GRU, three Dual-GRU splits and the budget-constrained Gated DeltaNet recurrence (Yang et al., 2025), are reported in Appendix B. The following diagnostics ask why retained information can still be used incorrectly.

4.2 A FIXED-WEIGHT RETENTION–USE GAP

Applicable exception content. The aligned D02 replay evaluates three frozen coupled workflow fits at length 2,048. Its reported integrity indicator J intersects decoded-key presence, matching latest-value correctness and top-1 key addressing (Appendix A.2). It holds on 4,602 of 4,608 exception-query evaluations, yet only 3,830 answers within J are correct (83.225%): 772 errors remain. All six queries outside J are also wrong. Importantly, the top-1 replay separately checks that the selected slot's decoded value is correct on every query in J; this is not inferred from the existence of some correctly valued matching slot. These are pooled query evaluations across three fitted models, not independent episodes.

Soft mixing versus coupled interpretation. A second frozen replay replaces the soft mixture by the stored payload of the maximum-cosine occupied slot, for *every* query. Other inputs and weights stay fixed; labels and J select the reported subset only. Table 2 shows that this rescues 30 of 772 errors, harms none within J, and leaves 742. Ordinary top-1 attention mass was already 99.19% within J and 98.75% on its 772 normal-error queries. Thus hardening lookup provides little repair at these checkpoints. Correct decoded content and a correct selected slot still do not ensure correct interpretation by the downstream coupled path. This does not identify the head alone as the unique cause or rule out effects of mixing during training.

Inapplicable content on rule queries. Table 3 addresses a different failure: exception payloads can affect queries for which no exception applies. Only values change at evaluation; query, recurrent state, confidence and weights remain fixed. Suppressing all values helps rule prediction but removes useful exception input. Conditional no-match masking instead raises rule accuracy from 60.15% to 86.93%, with exception accuracy essentially unchanged (83.12% to 83.10%). Rule accuracy improves in all three fits (Appendix A). These frozen interventions establish a role for inappropriate payload use, but do not reach the near-perfect results of readers trained with rejection. In particular, a no-match mask does not directly correct the applicable exception errors in J; retraining can change how those payloads are represented and interpreted.

Fixed-weight read intervention	Overall	Rule	Exception
Original coupled	62.31	60.15	83.12
Suppress all retrieved values	79.58	87.15	6.40
Suppress values only on no-match	86.57	86.93	83.10

Table 3: **Retained contents do not ensure appropriate use.** Workflow, length 2,048; reported mean percentages over the same three formal coupled fits. Only the retrieved value vector changes. Query, recurrent state, confidence and weights are held fixed; no model is retrained. Conditional masking raises rule accuracy from 60.15% to 86.93%, with exception accuracy 83.12% versus 83.10%. The separately reported exception-integrity statistics are not the denominator of the overall score. This table does not report the separately trained gated model.

	$T = 2{,}048$		$T = 8{,}192$	
Reader / system	Query	Joint ep.	Query	Joint ep.
GRU	90.875	0.000	90.885	0.000
Coupled	62.325	0.000	62.321	0.000
Hard match gate	99.929	98.307	99.992	99.740
Soft null	**100.000**	**100.000**	**100.000**	**100.000**
Path-isolated	**100.000**	**100.000**	**100.000**	**100.000**

Table 4: **A rejection option repairs the shared reader without full isolation.** Workflow E01 common arrays; mean percentages over three fitted seeds. All learned systems retain 256 bytes. Hard gate and soft null keep the coupled writer/head; the null adds one static logit. Query accuracy and all-32-queries-correct episode accuracy are distinct. These are separately trained comparisons, not frozen D02 interventions. Exact per-seed counts, variation and paired whole-episode intervals accompany the export (Appendix A.3).

4.3 Rejection-capable training with shared parameters

The E01 confirmation in Table 4 compares every reader on the same workflow arrays at each length. The per-fit query counts are 65,536, 16,384 and 4,096 at lengths 128, 2,048 and 8,192. Soft null achieves 100% query and joint-episode accuracy for every fitted seed at the two longer lengths. At length 128, its three fits make 2, 3 and 0 errors, for mean accuracy 99.997%. Its mean query gains over the hard gate are 0.098, 0.071 and 0.008 points, respectively; every fitted-seed difference is nonnegative. These finite-sample results show that exact binary gating and full parameter isolation are not needed for the observed repair. They do not establish population equivalence or universal dominance of soft null.

The final soft-null reader assigns about 78.1% mean attention mass to real slots on descriptor matches, but only 0.15–0.16% on no-match queries (Appendix A.3). This is selective attenuation, not uniform suppression of the bank. The contrast is measured after training with a strong initialization prior; no initialization-ablation claim is made. The hard-top-1 replay and the separately trained null model are different interventions: the latter changes learning as well as inference and should not be described as merely sharpening exception lookup.

The hard gate retains the broader evidence base. On the original generator's post-P0 array at length 8,192, it reaches $99.959 \pm 0.051\%$, with 99.991% rule accuracy, 99.653% exception accuracy and 98.698% joint-episode accuracy. Each fitted seed exceeds GRU by 9.28–9.38 points on that same array. The isolated mean is 100%; the 0.041-point mean query gap coexists with a 1.302-point episode gap. Full gated length results and separately trained original isolated results remain in Appendix A.1. Workflow E01 does not supply soft-null evidence on this generator or under descriptor stress.

An earlier within-isolated ablation finds identical correct predictions for ordinary joint training and protected staging. Staging is therefore not required for that measured accuracy. Hard and soft rejection provide a simpler repair while retaining a shared predictor; isolation remains a stronger protection reference.

5 OPERATING BOUNDARIES

The frozen descriptor stress evaluates hard-gated and isolated readers, not soft null. At the passing grid points ($\sigma = 0.02$ Gaussian noise or $\alpha = 0.10$ nearest-entity blending), both meet the predefined aggregate criteria; the next tested levels fail. Isolation is more accurate at these same points, but the grid establishes neither a larger passing level for isolation nor a continuous robustness radius. Perturbed queries enter both matcher and predictor, and errors remain even among correctly routed queries. Appendix F gives the exact perturbations and classwise route counts; this study is supporting evidence, not a claim of broad robustness.

Prior *ungated* allocations lose to compact-only at the tested 128-byte interior point and gain at 256/512 bytes; query weights also change their preferred allocation (Appendix G). These are not frontiers of the final gated readers. Earlier original-task arrays contain critical-exception ages of 993–3,072 events, with median 2,031. Evaluation on 8,192-event streams therefore demonstrates length extrapolation, not uniform 8,192-step retention.

6 RELATED WORK

Scope-aware memory. SERAC stores edits explicitly, uses a learned scope classifier to select when an edit applies, and otherwise retains the frozen base model's prediction (Mitchell et al., 2022). This directly anticipates separating retention from applicability. Our setting instead learns a changing recurrent rule state from typed streams under a common episode-state budget. Our hard and soft repairs restrict only the value input to one shared predictor; neither repair selects between a frozen base model and a separately trained counterfactual generator. The contribution is the fixed-checkpoint diagnosis and same-parameter repair, not inventing scope-aware retrieval.

Attention with a rejection option. Bondarenko et al. (2023) connect activation outliers to attention heads attempting no-op updates and introduce clipped softmax and gated attention. Qiu et al. (2025) study query-dependent sigmoid gating after scaled dot-product attention. We compare a fixed descriptor-match gate with a one-scalar zero-payload null candidate over a three-slot bank. These are instances of established rejection/no-op ideas, not new attention primitives. Our additional evidence is the query-aligned retention/use diagnosis and same-state-budget comparison. One null initialization on workflow does not establish a general ranking of rejection mechanisms.

State representations and optimization. GRUs, selective state-space models and Gated DeltaNet provide recurrent representations (Cho et al., 2014; Gu & Dao, 2023; Yang et al., 2025); NTM, DNC and RMT provide learned memory interfaces (Graves et al., 2014; 2016; Bulatov et al., 2022). Soft modularization and gradient surgery address parameter sharing and optimization interactions (Yang et al., 2020; Yu et al., 2020). Our intervention preserves sharing but changes the payload path and its training gradients. Full isolation is a comparator, not a necessary explanation of the main gain. The modern compact baseline is limited to the stated recurrence adaptation.

7 DISCUSSION AND LIMITATIONS

Keeping information without applying it. *Information need not be erased to be excluded from a decision.* This is a limited sense of "functional forgetting at read time": stored bindings remain available, while their value contribution is blocked or attenuated for the current query. The analogy concerns control of use, not deletion or lasting loss of access. Human retrieval-suppression experiments can produce reduced later recall (Anderson & Green, 2001); our read operations leave stored state unchanged and do not test that phenomenon. Accordingly, *selective readout* or *retrieval rejection*, rather than "learning to forget," describes the measured mechanism. The hard gate is deterministic; the soft null learns one scalar from a specified similarity prior.

What the interventions establish. The evidence separates two questions: whether applicable stored content yields a correct answer, and whether inapplicable content disrupts another answer. The hard-top-1 diagnostic concerns the former; conditional masking concerns the latter. Retraining with rejection repairs the observed system-level failure but does not uniquely identify its historical

optimization cause. The original coupled head already receives confidence and could in principle learn to ignore retrieval. Explicit rejection is an effective inductive bias here, not a proved necessity for every predictor. A fixed budget limits retention; it does not make the applicability problem exclusive to bounded memory. The reusable procedure is to audit content, selection and use separately before changing storage capacity or the reader. We establish its utility here, not transfer of the observed failures to other memory applications.

Scope and resources. Both generators supply observable event types, fixed descriptors and a known finite policy family, with three fixed exception identities per episode fitting the three-slot bank. Smaller specialized programs solve them exactly; the original generator also conditions exception values on the final rule answer. Conclusions concern the evaluated learned systems, not intrinsic task hardness or universal architectural superiority. Soft null has only workflow evidence and reuses the 0.97 prior. Descriptor stress concerns hard-gated and isolated readers on a finite grid. Persistent-state matching does not equalize static parameters, computation or supervision. Neither many queries nor repeated lengths turn three fitted seeds into broad model-population evidence.

8 Conclusion

Correct retention, correct slot selection and correct prediction are distinct. In the tested bounded hybrid systems, hardening an already correct lookup repairs little of the residual exception failure, while suppressing inapplicable values improves rule predictions at fixed weights. Training shared predictors with either a hard match gate or a prior-initialized soft null largely removes the failure; full path isolation is not required for the main gain. The design lesson is selective use rather than deletion: preserve a binding for queries where it applies, without allowing it to intrude on every other prediction.

AI Use Disclosure

Generative AI tools were used to provide feedback on experimental design and methodology, assist with code development, and improve the clarity and presentation of the manuscript. All AI-assisted outputs were independently reviewed and verified by the author, who takes full responsibility for the final content.

REFERENCES

Michael C. Anderson and Collin Green. Suppressing unwanted memories by executive control. *Nature*, 410(6826):366–369, 2001. doi: 10.1038/35066572. URL https://www.nature.com/articles/35066572.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Quantizable transformers: Removing outliers by helping attention heads do nothing. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/edbcb7583fd8921dad78adecfe06a99b-Abstract.html.

Aydar Bulatov, Yury Kuratov, and Mikhail Burtsev. Recurrent memory transformer. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/47e288629a6996a17ce50b90a056a0e1-Abstract-Conference.html.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734. Association for Computational Linguistics, 2014. doi: 10.3115/v1/D14-1179. URL https://aclanthology.org/D14-1179/.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines, 2014. URL https://arxiv.org/abs/1410.5401.

Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, Adrià Puigdomènech Badia, Karl Moritz Hermann, Yori Zwols, Georg Ostrovski, Adam Cain, Helen King, Christopher Summerfield, Phil Blunsom, Koray Kavukcuoglu, and Demis Hassabis. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538:471–476, 2016. doi: 10.1038/nature20101. URL https://www.nature.com/articles/nature20101.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2023. URL https://arxiv.org/abs/2312.00752.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. Memory-based model editing at scale. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15817–15831. PMLR, 2022. URL https://proceedings.mlr.press/v162/mitchell22a.html.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/904e89bb4e632e75fb47f093b620b257-Abstract-Conference.html.

Ruihan Yang, Huazhe Xu, Yi Wu, and Xiaolong Wang. Multi-task reinforcement learning with soft modularization. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/32cfdce9631d8c7906e8e9d6e68b514b-Abstract.html.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/4904fad153f6434a7bcf04465d4be2cc-Abstract-Conference.html.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.nips.cc/paper/2020/hash/3fe78a8acf5fda99de95303940a2420c-Abstract.html.

A EXPERIMENT IDENTITY AND EVIDENCE AVAILABILITY

The main mechanism concerns rejection-capable readout in a shared predictor. Hard gating has cross-generator evidence; the soft-null comparator is workflow-only. Earlier path-isolation, allocation and representation blocks are retained as separately identified evidence, not renamed as new gated results. The same gated architecture is fitted separately to the two generators. Table 5 identifies the blocks used in this draft.

Block	Content	Available evidence in this bundle
R23	Earlier original common confirmation	Query predictions, per-seed rows, aggregate SD and conditional paired intervals.
W24	Earlier workflow composition ablation	Query predictions, per-seed rows, joint/staged comparison and program.
P0	Gated workflow fits; isolated original fits; three-seed coupled interventions	Executed source/configs and selected checkpoints included; the D02 intervention arrays are exported separately.
P0-final	Original gated fits and frozen workflow descriptor stress	Original common predictions/counts/intervals and the complete saved stress grid are included.
D02 / top-1	Query-aligned exception integrity and hard-top-1 frozen replay	49,152 aligned rows plus original arrays, hashes and verifiers; five unpersisted state-identity fields are explicit.
E01	Soft-null comparator and fresh common-array reader comparison	All five saved predictions, targets/masks/IDs, exact counts, fitted-seed SDs, intervals and optimizer records included.
Allocation / utility	Prior ungated budget and query-weight studies	Reported values and prior audit descriptions; not new gated frontiers.

Table 5: Evidence blocks remain distinct. R23 and W24 are the original author-checked exports. P0, P0-final, D02 and E01 are later studies with their own arrays and checkpoint identities; no interval is transferred across blocks.

P0 original path-isolated result. Separately trained isolated models use 256 bytes and 7,143 static parameters. The P0 report gives $99.998 \pm 0.004\%$ overall at lengths 128, 512 and 2,048 (99.935% joint-episode), and $100 \pm 0\%$ overall and 100% joint-episode at 8,192. These are P0 arrays, not substituted into the final post-P0 original table. The final post-P0 report explicitly confirms an isolated mean of 100% at 8,192 on that comparison array; the other final per-length comparator rows require export.

A.1 FULL REPORTED GATED LENGTH SCAN

The main method table uses the later E01 workflow common arrays. This gated-length table retains P0/post-P0 values on their own arrays. The original post-P0 export contains the hard-gate, GRU, ungated-hybrid and isolated predictions, exact per-seed counts and existing hard-gate paired intervals on those arrays; they are not inferred from E01.

Task	T	Overall ± SD	Rule	Exception	Joint ep.
Original	128	99.954 ± 0.040	99.983	99.674	98.665
	512	99.937 ± 0.065	99.970	99.620	98.405
	2,048	99.945 ± 0.050	99.966	99.740	98.893
	8,192	99.959 ± 0.051	99.991	99.653	98.698
Workflow	128	99.90 ± 0.05	99.92	99.75	98.32
	2,048	99.83 ± 0.19	99.84	99.74	97.98
	8,192	99.95 ± 0.06	99.95	100.00	99.22

Table 6: **One match-gated architecture, separately trained on two generators.** Accuracy percentages, averaged over three fitted seeds; SD is the reported sample SD of overall query accuracy. Every learned fit retains 256 bytes. Joint episode means all 32 queries are correct. Original rows are from the final post-P0 block; workflow rows are from P0, not the earlier composition-ablation arrays. The panels use separate confirmation arrays, documented in Appendix A; the reported SD describes training-seed variation, not an episode-bootstrap interval.

A.2 Query-aligned integrity and hard-top-1 replay

Training seeds 92401–92403 yield original-to-conditional rule accuracies 50.99→87.30, 52.98→81.69 and 76.48→91.79% at length 2,048. Their value-only intervention preserves confidence. D02 aligns content, addressing and prediction at the exception-query level.

Reported indicator and selected-slot check. Let d_K and d_V denote the audit's key and value decoders, $k(q)$ the query class, $y(q)$ its target and $j^*(q) = \arg\max_{j \in \mathcal{O}} s_j(q)$ the selected occupied slot. Write K_q, V_q, A_q for the *recorded* key-presence, matching-latest-value and top-1-key flags. The reported subset is

$$J_q = K_q \wedge V_q \wedge A_q. \tag{11}$$

All reported counts continue to use this original indicator. If duplicate decoded keys occur, the existence of a correctly valued matching slot does not by itself establish correctness of v_{j^*}. The top-1 replay reports the additional selected-slot property

$$T_q = \mathbb{1}\{j^* \in \mathcal{O},\ d_K(k_{j^*}) = k(q),\ d_V(v_{j^*}) = y(q)\}, \qquad \sum_q J_q(1 - T_q) = 0. \tag{12}$$

The sum covers the 4,602 reported J evaluations. We do not redefine J as T and reuse its count without a rowwise check. Gold targets and decoded classes determine analysis subsets only; the forward intervention selects j^* without them.

Executed decoder implementation. For d_K, the fixed 32 workflow prototypes (NumPy seed 26092410) pass through each checkpoint's learned bias-free $16 \to 8$ slot-key map and are normalized; a stored normalized key is assigned the nearest row by Euclidean distance. For d_V, value IDs 0–15 pass through that checkpoint's learned 16-dimensional value embedding and affine $16 \to 7$ slot-value map; a stored payload is assigned the nearest unnormalized seven-dimensional candidate by Euclidean distance. Neither decoder has a fitted probe, calibration split, temperature or extra label input. Exact-distance ties use the first candidate index.

Occupancy is `slot_stamp`> 0. The historical value flag first finds all occupied slots whose decoded key equals the query's latent prototype class, then reads `argmax` of that Boolean vector. Thus, with duplicate decoded keys it uses the lowest slot index, not the largest stamp; "latest" is a historical field name. Top-1 addressing independently uses maximum cosine after assigning invalid slots score -10^4; ties again take the lowest slot index. The export preserves this original J and supplies its rowwise flags. It also supplies the separately recorded top-1 payload correctness flag, which is true on all 4,602 J rows. Occupied counts, selected indices, decoded IDs and duplicate-key counts were not persisted; their CSV fields are blank and recovering their frequency would require a newly authorized state replay. Correctness under these audit decoders is not correctness of the shared prediction head.

The joint coverage is $4{,}602/4{,}608 = 99.870\%$, while normal accuracy within it is $3{,}830/4{,}602 = 83.225\%$. The hard-top-1 intervention raises this to $3{,}860/4{,}602 = 83.877\%$. Per-seed conditional

Fit seed	Exception N	$\lvert J \rvert$	Normal correct	Top-1 correct	Rescued / harmed
92401	1,536	1,536	1,225	1,231	6 / 0
92402	1,536	1,530	1,179	1,201	22 / 0
92403	1,536	1,536	1,426	1,428	2 / 0
Pooled	4,608	4,602	3,830	3,860	30 / 0

Table 7: Query-aligned D02 integrity and the later hard-top-1 replay on the same three coupled checkpoints and data. Correct counts condition on J. J is computed from labels and decoded contents for analysis only, never supplied to the forward intervention. All six queries outside J are wrong under the normal reader.

accuracies are 79.753/77.059/92.839% normally and 80.143/78.497/92.969% after intervention. Only 30/772 (3.89%) of the normal errors in J are rescued; none is introduced there. Harm outside J is not inferred from this subset result.

For every query, the intervention uses the stored seven-dimensional payload of the maximum-cosine occupied slot instead of its softmax-weighted mixture. Query, recurrent state, confidence and weights do not change. Gold labels and J are used only for scoring. Reported mean normal top-1 mass is 99.19% in J and 98.75% on the normal-error subset. Correctness of a separately decoded payload does not force the shared head to interpret it correctly with the other inputs. The result localizes residual errors downstream of the checked storage and selection properties; it does not measure an exact causal fraction attributable to historical mixing or to one layer.

The replay verifies normal predictions and routing against their earlier saved sources. The export includes both replay NPZ files, the three conditional/zero-value intervention NPZs, the executed runners and verifiers, checkpoint/data hashes and 49,152 aligned model–query rows. No new model replay was performed for this revision.

A.3 Soft-null comparator and rejection behavior

E01 uses one predefined variant, training seeds 92401–92403, 4,000 updates per fit, and no initialization or variant sweep. The update-1,000 qualification gate precedes expansion to all fits. Test seeds and episode counts are fixed before training, and common arrays are generated after selected E01 checkpoint hashes are frozen. The fixed logit initialization is 19.4; selected logits are 18.822, 18.724 and 18.519 at updates 3,750, 3,250 and 3,500. This static scalar is the only parameter-count increase. No gold query subtype enters inference. The executed source constructs one `AdamW(model.parameters(), lr=0.001, weight_decay=0.01)` group with no scheduler, so the null scalar is in group 0 and is decayed with the other parameters. All fits execute 4,000 updates. Consequently, its observed drift cannot by itself distinguish task-gradient learning from decay. Neither an initialization sweep nor retraining is needed to state this limitation. The

	$T = 128$		$T = 2{,}048$		$T = 8{,}192$	
System	Query	Joint	Query	Joint	Query	Joint
GRU	90.847	0.049	90.875	0.000	90.885	0.000
Coupled	61.978	0.000	62.325	0.000	62.321	0.000
Hard gate	99.899	98.519	99.929	98.307	99.992	99.740
Soft null	99.997	99.919	100.000	100.000	100.000	100.000
Path-isolated	100.000	100.000	100.000	100.000	100.000	100.000

Table 8: Full E01 common-array comparison, mean percentages across three fitted seeds. Joint is the fraction of episodes with every query correct. The reported per-seed null errors at $T = 128$ are 2, 3 and 0 of 65,536 queries each. Longer-set denominators and per-seed comparator rows must accompany the raw export; they are not inferred from earlier runs.

normalizer, not a learned null key, supplies the additional candidate. The real-slot confidence input is retained, and null is excluded from writes, replacement, occupancy and persistent-state accounting. Occupied real slots are those with positive stamps; unused real logits are set to -10^4 before concatenating the null logit. Empty-bank behavior, finite null-logit gradients and limiting-logit behavior

Length	Query N/fit	Null errors/fit	Null query mean±SD	Hard-gate query mean±SD
128	65,536	2 / 3 / 0	99.997 ± 0.002	99.899 ± 0.070
2,048	16,384	0 / 0 / 0	100.000 ± 0	99.929 ± 0.043
8,192	4,096	0 / 0 / 0	100.000 ± 0	99.992 ± 0.014

Table 9: E01 exact denominators and fitted-seed variation (percent). The three error positions correspond to seeds 92401–92403. The export also contains all five systems' integer counts and the existing 36 null-versus-comparator paired whole-episode intervals (2,000 resamples, conditional on each fitted seed). Zero observed errors do not establish zero population risk.

Length	Real-slot mass on matches	Real-slot mass on no-match
128	78.118%	0.157%
2,048	78.119%	0.152%
8,192	78.096%	0.163%

Table 10: Mean real-slot attention mass of the trained soft-null reader. "Match" means maximum occupied cosine confidence strictly above 0.97. Each cell first averages query positions within each fitted seed/group and then averages the three seed means; exact group counts and quantiles accompany the export. The remaining mass is assigned to the zero-payload null. This is not an initialization ablation.

are checked in the exported implementation record. E01 establishes only the specified workflow comparison. It has no original-generator, alternate-initialization or descriptor-stress evaluation.

Why the strictly protected coupled cell is absent. The planned ablation would retain the original coupled readout while freezing every parameter on the rule-output path. In the audited graph every exception-trainable group also lies on that path, so strict protection leaves no exception-only parameters. The cell was not replaced with a non-equivalent partially unfrozen experiment. This is a limitation of that intervention definition, not proof that functional protection is impossible. The shared-parameter hard-gate and soft-null comparisons directly test different, constructible readout repairs.

B EARLIER COMPARISONS RETAINED FOR CONTEXT

These tables retain the original uploaded canonical arrays and uncertainty estimates. They do not pool later gated fits with older baselines.

Additional recurrent splits. On the earlier original common arrays, Dual-GRU 16/48, 32/32 and 8/56 obtain 89.51/90.14/89.10% at length 2,048 and 68.57/85.47/69.85% at length 8,192. These are tested partitions, not an exhaustive search over recurrent models.

Earlier workflow diagnosis. A 64-byte rule-only recurrent system achieves 100% on the reported isolated task. Expanding the coupled recurrent component to 192 or 320 bytes while retaining three slots does not repair its joint performance. At fixed diagnostic checkpoints, setting retrieved values to zero on all queries changes rule accuracy from 69.59/52.52/51.47% to 88.70/93.19/86.64% at total budgets 256/384/512 bytes. These one-seed length-128 development measurements are distinct from the three-seed length-2,048 P0 intervention. They motivate the later formal test, rather than supplying its uncertainty.

Parameter paths and staging. The original coupled graph permits rule and exception losses to reach shared value embeddings, the GRU, slot keys/values and the coupled head. The isolated graph separates those losses into disjoint parameter groups when matching is correct. Local negative gradient cosines are consistent with detrimental interactions but do not uniquely identify the historical training cause. The joint-isolated and protected-staged variants give identical correct predictions on the W24 arrays. The newer gated-coupled model shows that removing all parameter sharing is not required for the main repair.

Type-aware route distillation. The earlier 1,154-parameter router study freezes successful original-task state/readout and fits the observable-type route with ordinary or balanced BCE. Both reproduce parent predictions in 24 evaluated variant/seed/length comparisons, with long-length scores 98.05/98.07%. This is not latent semantic routing and is not the minimal readout gate introduced here. It is retained as a secondary training result.

System	$T = 2,048$	$T = 8,192$
GRU	90.62 ± 0.00	90.62 ± 0.00
Type-aware GRU	90.59 ± 0.07	90.59 ± 0.06
Dual-GRU, 32/32	90.14 ± 0.12	85.47 ± 1.07
Gated-delta adaptation	90.49 ± 0.02	90.49 ± 0.04
Hybrid, 64/192 B	98.03 ± 0.78	98.20 ± 0.54

Table 11: **Earlier representation comparison, original Q17 (R23).** Overall accuracy (%, mean $\pm$ sample SD across three fitted seeds), 256 B per learned system. The hybrid is not the workflow isolated variant. The separate 62-B specialized program scores 100% at both lengths. Each seed uses the same 512/128 episodes (16,384/4,096 queries).

System	$T = 128$	$T = 2,048$	$T = 8,192$
Coupled hybrid	62.16 ± 13.16	62.30 ± 13.63	62.57 ± 13.64
GRU	90.78 ± 0.05	90.88 ± 0.03	90.70 ± 0.11
Path-isolated, joint	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
Path-isolated, staged	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00

Table 12: **Earlier workflow composition ablation (W24).** Overall accuracy (%, mean $\pm$ sample SD over three seeds), 256 B per learned system. Each fit executes 4,000 updates; selected checkpoints may precede completion. Each seed uses the same 2,048/512/128 episodes (65,536/16,384/4,096 queries). Both isolated variants also score 100% on each branch and on joint episodes. Saved 7-B program predictions are exact on these same arrays; it is a specialized, non-learning reference.

Block	System ($T = 8,192$)	Overall	Rule	Exception	Joint ep.
R23	GRU	90.62	100.00	0.00	0.00
R23	Ungated hybrid	98.20	98.11	99.13	57.81
W24	GRU	90.70	99.54	5.21	0.00
W24	Coupled	62.57	60.41	83.51	0.00
W24	Isolated, joint	100.00	100.00	100.00	100.00

Table 13: Branch and joint-episode scores from the inherited prediction-level blocks. These are earlier model/array comparisons, not values for the new match-gated runs.

C TRAINING EXPOSURE AND RESOURCE ACCOUNTING

The workflow composition-ablation report gives batch size 32, a run budget of 4,000 optimizer updates and 16,384,000 stream-event exposures per fit. These totals refer to executed training, not necessarily to the selected checkpoint. The isolated-joint selected checkpoints occur at updates 2,000, 2,250 and 2,000. The phase-boundary and final checkpoints are also used for the stated invariance diagnostic; they are not silently substituted for the selected checkpoints in the main result table.

Training procedure	Rule target positions	Exception target positions
GRU / coupled hybrid / isolated joint	3,712,000	384,000
Protected staged	1,856,000	192,000

Table 14: Reported effective loss-target exposures over the complete workflow fits. Equal optimizer-update budgets do not imply equal supervision exposure. The isolated-joint comparison uses the same target counts as the coupled baseline; the protected staged procedure does not.

System	Recurrent B	Episodic B	Other resource information
Main GRU	256	0	25,488 static parameters.
Type-aware shared GRU	256	0	44,496 static parameters.
Dual-GRU 16/48; 32/32; 8/56	256	0	20,880; 19,344; 22,800 static parameters.
Coupled / hard-gated hybrid	64	192	6,615 static parameters.
Soft-null hybrid	64	192	6,616 static parameters; one unwritable scalar null logit.
Path-isolated (both procedures)	64	192	7,143 static parameters.
Gated-delta adaptation	256	0	7,418 static parameters; no short convolution.
Original-task program	62 B total		2,048 B of fixed prototype constants.
Workflow program	7 B total		2,560 B of fixed policy/prototype constants.

Table 15: Resource categories from the author-checked manifests and P0 method description. Parameters are counted separately from persistent episode state; these are not parameter- or FLOP-matched comparisons. In staged training only 5,856 rule parameters, then 1,287 exception parameters, are trainable. Program constants are non-learning task knowledge, not episode state.

The isolated-joint runs report approximately 794–808 seconds of training wall time, with an unrelated GPU co-tenant present. Earlier protected, coupled and GRU runs report means near 618, 767 and 245 seconds in different conditions. We do not use these numbers as a speed comparison. Matched wall-clock or FLOP efficiency is not established. Test-time persistent bytes and training costs are separate axes.

The P0 workflow hard-gate model preserves the coupled model's 6,615 parameters; E01 adds one static scalar, for 6,616. The original hard-gate and isolated fits likewise execute 4,000 updates each. The export records selected steps, executed updates, checkpoint hashes, persistent bytes and static parameters for 21 core fits, and includes 27 selected checkpoints spanning the five workflow readers and original-task comparators. E01 evaluates its selected earlier checkpoints after completing each fit. Frozen descriptor stress and both D02 replays perform zero optimizer updates. Runtime comparisons across different co-tenant conditions are not claimed.

D IMPLEMENTATION SPECIFICATIONS

Observation interface and policy functions. An episode permutes 32 fixed unit-normalized Gaussian prototypes in $\mathbb{R}^{16}$ across entity IDs. Original-task prototypes use NumPy seed 170920; workflow uses 26092410. The finite policy functions are $f_p(k) = (c(k) + p) \bmod 16$ in the original task and $f_p(k) = A_{p,c(k)}$ in workflow. Here $c(k)$ is the prototype class and A is a fixed 16×32 uint8

table: each row independently permutes two copies of $0, \ldots, 15$, using seed 26092412. Policies are known at training time, not novel test-time rules.

The solver receives a 16-dimensional event descriptor, a value ID in $0, \ldots, 15$, and event type $a_t \in \{0, 0.5, 1\}$ for distractor, rule and exception events. Rule-event descriptors are replaced by a fixed normalized Gaussian token (seed 170921 or 26092411). Other event descriptors are noiseless entity prototypes. Query descriptors add independent Gaussian noise with standard deviation 0.01 and are renormalized. Each episode queries all 32 entities once in random order. Learned solvers receive neither entity IDs, prototype-class labels, the policy lookup table, nor the gold query subtype at inference. Specialized programs have the prototype lookup and, for workflow, the policy table as static constants. A fixed rule token is generator/input-encoding information, not an additional evolving memory state.

Original-task stream construction. Each 32-event phase contains four rule events carrying the current offset, exception writes if scheduled, and otherwise independent random entity/value distractors, all randomly permuted within the phase. Initial offsets are uniform; subsequent offsets add a uniform nonzero increment modulo 16. The $K = 3$ active exception entities are sampled once and persist without deletion. All are written in phase zero; every 32 subsequent phases one active entity is overwritten in cyclic order. Exception values are constructed to differ from that entity's *final* rule answer, including the initial values. Thus original-task exceptions are guaranteed to matter at evaluation. This construction makes earlier exception values statistically dependent on the final policy, although that final policy label is not supplied directly. It is a target-conditioned generator, not a strictly causal online generator. Workflow does not impose this constraint. Training/development contain 8,192/512 episodes of length 128, using seeds 26092021/26092022.

Workflow stream construction. The first event initializes a uniform policy; the next K events initialize the uniformly chosen active exceptions. Each later event independently selects a policy change with probability 0.05, an exception overwrite with probability 0.04, and a random entity/value distractor otherwise. Policy and overwritten values change by a uniform nonzero increment modulo 16. Overwrites choose uniformly among the active exception entities. No exceptions are deleted or replaced by new entities. Unlike the original task, override values may coincide with the current rule answer. Training/development contain 8,192/2,048 episodes of length 128, with seeds 26092401/26092402. Both tasks query only after the complete stream. Final query weighting is always 29 rule and three exception positions per episode.

State tensors and writing. Per episode, the hybrid retains FP32 tensors $h \in \mathbb{R}^{16}$, keys $K \in \mathbb{R}^{3\times 8}$, payloads $V \in \mathbb{R}^{3\times 7}$ and stamps $s \in \mathbb{R}^3$, initially zero: $4(16 + 24 + 21 + 3) = 256$ bytes. A positive stamp marks a used slot. On an exception event, a bias-free $16 \to 8$ projection produces the normalized key; a $16 \to 7$ affine map of the value embedding produces its payload. Among occupied keys with cosine similarity strictly above 0.97, the highest similarity selects an overwrite. Otherwise the minimum stamp selects a free slot or, when full, the least recently *written* slot. Argmax/argmin ties select the first index. Keys and payloads are replaced, and the stamp becomes the previous maximum plus one. Reads do not update stamps. At $K = 3$ the nominal active set fits exactly, although learned key collisions remain possible.

Reading and threshold lineage. The same key projection normalizes the query. Invalid slots receive score -10^4; the value read is the softmax of 20 times those cosine scores, weighted over the payloads. An all-empty state returns zero retrieval and confidence -1. Otherwise confidence is the maximum occupied cosine. The match-gated model zeros only the seven-value retrieval when confidence is not strictly above 0.97, retaining the original confidence input and shared head. The isolated model instead switches to its exception head exactly when confidence exceeds 0.97, without replacing the soft retrieval by top-1 lookup. The threshold is inherited from the earlier Q17 slot-write implementation and fixed in the separated-model protocol before confirmation; no confirmation-set tuning or optimality claim is made. This lineage does not establish a universally calibrated threshold.

Neural dimensions. The GRU and coupled hybrid embed 16 value IDs into 16 dimensions, concatenate descriptor/embedding/type into 33 inputs, and use a standard GRUCell of width 64 or 16. Their read heads concatenate query and recurrent state (plus seven retrieved values and confidence for the hybrid), then apply a 64-unit ReLU layer and 16 logits. Type-aware GRU has separate width-64

rule/exception transitions sharing one state; Dual-GRU has two independently routed states of the reported widths and a shared 64-unit ReLU read head. The isolated model has separate rule and slot value embeddings, a width-16 rule GRU, a $32 \to 64 \to 16$ ReLU rule head and a $7 \to 32 \to 16$ ReLU exception head.

The gated-delta control projects the 33 event inputs and 16 query inputs to width 32, uses one 8-dimensional key/value head and an 8×8 matrix state, and has a $48 \to 64 \to 16$ ReLU read head. Keys/queries use normalized SiLU projections; values use SiLU, update strength uses sigmoid, decay uses a learned exponential/softplus parameterization, and retrieval uses gated RMS normalization. Short convolution is disabled. This is the repository's task adaptation, not an off-the-shelf pretrained baseline.

Optimization and checkpoint accounting. The main original-task and workflow fits execute 4,000 updates at batch size 32, sampling episodes with replacement and processing all 32 targets. The inherited training configuration uses AdamW with $\beta = (0.9, 0.999)$ and $\epsilon = 10^{-8}$, weight decay 0.01 and global gradient clipping at 1, without a scheduler. This does not establish the parameter-group assignment of the later E01 null scalar, which requires its own export. Learning rate is 10^{-3} except the selected gated-delta control, which uses 3×10^{-3}. The original weighted-mean implementation uses exception weight 1 for the first 2,000 steps and $29/3$ thereafter, normalized by the sum of weights; with fixed $K = 3$ this is exactly Equation (10). Evaluation every 250 steps selects the largest development branch-macro accuracy, retaining the earliest tie. The modern control has additional development-only learning-rate screening and sanity costs. Its seed 92011 continues the selected screening checkpoint to 4,000 total updates; the other formal seeds are separate fits. This is not a compute-equal comparison.

Staged training directly supervises rule-head outputs on rule positions, reloads the best rule-stage checkpoint, freezes it, then supervises exception-head outputs on exception positions. Stage one selects development rule accuracy; stage two selects natural overall accuracy, each retaining the earliest tie. This gold-subset supervision is distinct from the joint model's matcher-selected output loss. Each stage executes 2,000 steps even if its selected checkpoint is earlier. Joint-isolated selected steps 2,000/2,250/2,000 correspond to 8,192,000/9,216,000/8,192,000 event exposures and 1,856,000/2,088,000/1,856,000 rule plus 192,000/216,000/192,000 exception target positions, not the complete-fit totals in Table 14. Per-fit selection and exposure records accompany the statistical data.

Diagnostic identity and slot metrics. The earlier single-seed readout diagnostic (not Table 3) uses 2,048 development episodes of length 128, not length 2,048. Its stored `zero_values_keep_confidence` intervention zeroes retrieval for all queries. The separate 0.97-gated intervention preserves matched values and yields overall scores 88.81/91.36/85.48% at 256/384/512 B; these are not the rule scores reported in the earlier diagnostic. Its three distinct integrity fields—key retention, key-and-latest-value retention, and top-1 key addressing—each equal 100/99.8046875/100% across these budgets. Top-1 decoded value accuracy is instead 100/99.83724/100%. These are four metrics, not three names assigned to the three budget values. The old composition table and the new P0 three-seed intervention study use their own fits and arrays. The P0 study is the main diagnostic in Table 3; its raw intervention and integrity exports have not been added to this working bundle.

E What the gate does and does not protect

For an unmatched query, Equation (5) gives the shared predictor inputs $(q, h, 0, c)$. At fixed q, h, c, ψ and fixed no-match status, changing only the retrieved value payload cannot affect that output. This is direct payload exclusion, not full independence from the memory bank: keys affect c and the match decision, and shared parameters may still be changed by exception training.

For the isolated model, fixing $q, h, \theta_{\mathrm{r}}$ and preprocessing, with both E and E' producing no match, gives

$$R_{\mathrm{isolated}}(q, h, E) = F_{\theta_{\mathrm{r}}}(q, h) = R_{\mathrm{isolated}}(q, h, E'). \tag{13}$$

Here the exception head and values have no output path while these premises hold. A sequence-level extension additionally requires that exception events do not change the rule-state trajectory, rule parameters or shared preprocessing. Equation (6) concerns one direct edge, not the total derivative

Query stress	Model	Overall	Route err.	Correct-route acc.	Pass
$\sigma = 0.02$	Gated	99.078	0.682	99.719	Yes
	Isolated	99.955	0.014	99.969	Yes
$\sigma = 0.04$	Gated	95.43	2.47	97.68	No
	Isolated	97.93	1.07	98.94	No
$\alpha = 0.10$	Gated	99.143	0.543	99.649	Yes
	Isolated	99.988	0.000	99.988	Yes
$\alpha = 0.20$	Gated	92.80	3.33	95.75	No
	Isolated	95.86	1.85	97.54	No

Table 16: **A measured, limited descriptor operating region.** Frozen workflow models, length 2,048, three checkpoints per model; all scores are reported mean percentages. Clean accuracy is 99.833% for gated and 100% for isolated. σ denotes additional Gaussian noise; α denotes nearest-entity blending. Passing uses the original one-point overall-drop and 1% aggregate route-error criteria. The latter is not a branchwise FPR/FNR bound (Equation (14)). Perturbed queries enter both matcher and predictor. The table shows passing levels and their next tested levels, not a continuous robustness radius.

with respect to shared parameters. Local zero gradients alone do not establish sequence-level protection, especially in the presence of optimizer state or decay. This is an architectural property, not a new general theorem of memory or a guarantee of correct matching.

The soft-null read is ρr, with $0 < \rho < 1$ for a nonempty finite-score bank. It generally does not yield exact exclusion, although its no-match mass can be very small. Its read leaves the stored bindings unchanged; the confidence bypass remains. The hard gate and soft null are two implementations of selective value use. Path isolation makes stronger exclusions with a different parameterization. Neither finite-sample equal predictions nor equal maximum passing stress-grid levels establish population equivalence or a continuous robustness-radius comparison.

F Frozen descriptor stress: results and specification

Frozen query perturbations. The workflow stress test reuses one common length-2,048 array and six frozen hard-gated/isolated checkpoints (three per variant); soft null is not included. The generator's base query is $\mathrm{normalize}(d+\epsilon)$, with unit entity descriptor d and iid coordinate noise $\epsilon \sim \mathcal{N}(0, 0.01^2)$. For the Gaussian axis, one shared standard-normal tensor from NumPy seed 26092601 is added to this already normalized query as $\mathrm{normalize}(q + \sigma z)$, with $\sigma \in \{0, .01, .02, .04, .06, .08, .12\}$. For blending, the true entity is excluded and the clean descriptor with maximum cosine to d is chosen within the episode; first index breaks exact ties. The perturbed query is $\mathrm{normalize}((1-\alpha)q+\alpha d_{\mathrm{near}})$, with $\alpha \in \{0, .1, .2, .3, .4, .5\}$. The same perturbations are shared across checkpoints and enter both matcher and predictor; events, stored state, labels, threshold and weights remain fixed. Margin is true-descriptor cosine minus the largest other-descriptor cosine. No training occurs. The stability criterion is mean overall loss at most one point from clean and mean route error at most 1%.

Stress	Model	Rule FPR	Exception FNR	Correct-route accuracy
$\sigma = 0.02$	Gated	4/44,544 (0.009%)	331/4,608 (7.183%)	48,680/48,817 (99.719%)
	Isolated	0/44,544	7/4,608 (0.152%)	49,130/49,145 (99.969%)
$\sigma = 0.04$	Gated	18/44,544 (0.040%)	1,194/4,608 (25.911%)	46,830/47,940 (97.685%)
	Isolated	0/44,544	528/4,608 (11.458%)	48,108/48,624 (98.939%)
$\alpha = 0.10$	Gated	6/44,544 (0.013%)	261/4,608 (5.664%)	48,714/48,885 (99.650%)
	Isolated	0/44,544	0/4,608	49,146/49,152 (99.988%)
$\alpha = 0.20$	Gated	313/44,544 (0.703%)	1,324/4,608 (28.733%)	45,507/47,515 (95.774%)
	Isolated	0/44,544	908/4,608 (19.705%)	47,058/48,244 (97.542%)

Table 17: Pooled integer route and conditional-answer counts across the three fitted seeds at the displayed stress points. FPR and FNR retain separate rule and exception denominators; the complete 78-row grid is exported.

The largest passing levels in the reported grid are $\sigma = 0.02$ and $\alpha = 0.10$ for both models. Isolation performs better at these points and at the next tested levels. At $\alpha = 0.10$, gated and isolated overall accuracy is 99.143% and 99.988%, respectively. At $\alpha = 0.20$, both fail the stability criterion, with 92.80% and 95.86%. These observations establish a nonzero but limited operating region; they do not identify a continuous robustness radius or a universally calibrated threshold.

The route criterion is prevalence-weighted. Defining FPR as an exception route on a rule query and FNR as a missed exception route, the primary $K = 3$ weighting gives

$$\text{RouteError} = \frac{29}{32}\,\text{FPR} + \frac{3}{32}\,\text{FNR}\,. \tag{14}$$

Thus a 1% aggregate route-error threshold can allow an exception FNR of about 10.67% when FPR is zero. It is not a branchwise reliability guarantee. Table 17 gives the separate pooled rates at the displayed levels; the export contains exact per-seed numerators and denominators for all 78 model/condition rows.

Route error alone does not account for degradation. At $\sigma = 0.04$, correctly routed-query accuracy also falls to 97.68% for gated and 98.94% for isolated. The predictor receives the perturbed query as well as a routing decision. Moreover, the correctly routed subset can change with perturbation. Conditional accuracy documents residual errors; it is not a unique causal decomposition of matching and prediction effects. Appendix H records the stress-test specification and evidence-export status.

G Earlier budget and query-utility analyses

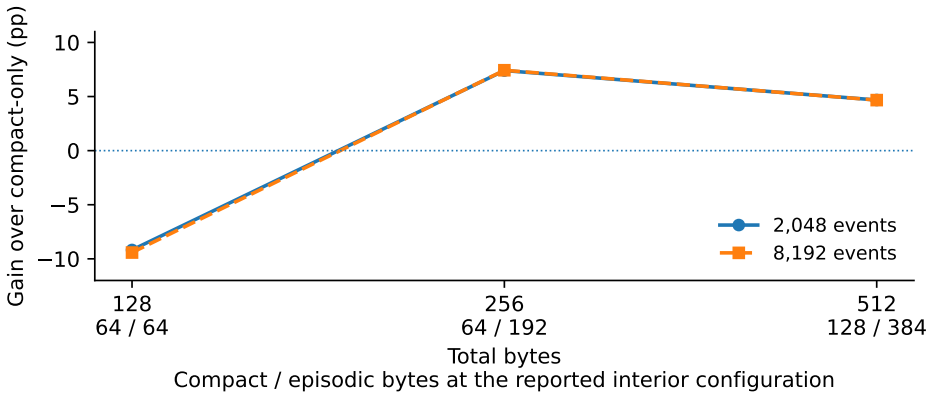


Figure 2: **Prior ungated allocation evidence, not a gated-method frontier.** Each point is the reported best tested interior split versus the compact endpoint at the same total bytes. Negative 128-byte points are retained. Only three reported budget points are plotted; lines do not estimate missing allocations or budgets, and error bars are not fabricated.

For a fixed load condition and branch differences $\Delta_{\text{r}} < 0 < \Delta_{\text{e}}$, the weighted score favors the hybrid when

$$p > \frac{-\Delta_{\text{r}}}{\Delta_{\text{e}} - \Delta_{\text{r}}}. \tag{15}$$

Reported mean break-even exception-query weights are 8.9%, 3.3% and 4.0% for $K = 1, 2, 6$, versus natural weights 3.125%, 6.25% and 18.75%. This post-hoc reweighting does not change the event stream or retrain a model. Full branch scores and the averaging rule across seeds/cells remain required for reproducing the break-even figures. High-minus-low rule-rate effects are below 0.1 points. No two-axis allocation law is asserted.

H Statistical exports and remaining release items

Inherited prediction-level blocks. The original author-checked archive contains 81 query-prediction rows, 29 aggregate rows and 63 paired comparisons for R23 and W24. The preceding

author-checked audit recomputed these values from its NPZ and found agreement with the supplied CSVs; that historical check is distinct from this editorial revision. Original training seeds are 92011–92013; its exported long confirmation arrays use seeds 26092393/26092394. Workflow training seeds are 92401–92403, with confirmation seeds 26092481/26092482/26092483. Exact query and episode counts are retained in the old per-seed export. The inherited bootstrap resamples paired whole

Study	Length	Seed	Gain (pp)	Paired 95% interval
Original	2,048	92011	8.30	[8.15, 8.46]
Original	2,048	92012	7.06	[6.79, 7.32]
Original	2,048	92013	6.86	[6.60, 7.12]
Original	8,192	92011	8.11	[7.76, 8.42]
Original	8,192	92012	7.59	[7.15, 8.01]
Original	8,192	92013	7.03	[6.47, 7.54]
Workflow	128	92401	9.19	[9.08, 9.30]
Workflow	128	92402	9.28	[9.16, 9.40]
Workflow	128	92403	9.21	[9.10, 9.32]
Workflow	2,048	92401	9.15	[8.97, 9.36]
Workflow	2,048	92402	9.10	[8.94, 9.26]
Workflow	2,048	92403	9.11	[8.95, 9.30]
Workflow	8,192	92401	9.28	[8.96, 9.67]
Workflow	8,192	92402	9.42	[8.96, 10.06]
Workflow	8,192	92403	9.20	[8.91, 9.52]

Table 18: Hybrid-minus-GRU (original) and isolated-joint-minus-GRU (workflow) gains from saved predictions. Percentile intervals use 2,000 paired whole-episode resamples conditional on each fitted seed; they are not intervals over training randomness. Inherited from the earlier prediction-level audit; not intervals for D02 or E01.

episodes 2,000 times, keeping all 32 queries together. Its intervals condition on a fitted seed; they are not inference over an unlimited population of model seeds. Zero intervals for equal predictions or 100% sample accuracy are not population guarantees.

Later run-level blocks. P0, P0-final, D02, D02b and E01 are not in the inherited NPZ. Their new export retains separate dataset/checkpoint identities and contains the existing predictions, exact counts and available intervals needed here. Artifact-only verification checks 140 exported files and the headline integer relations without loading a model. It does not rerun historical forward passes or manufacture new intervals.

Aligned integrity and prediction units. The D02 reports now supply the previously missing query-level intersection and integer ratio,

$$\widehat{A}_{\mathrm{e}|J} = \frac{\sum_{i\in I_{\mathrm{e}}} \mathbb{1}[J_i]\mathbb{1}[\hat{y}_i = y_i]}{\sum_{i\in I_{\mathrm{e}}} \mathbb{1}[J_i]} = \frac{3{,}830}{4{,}602}. \tag{16}$$

Appendix A.2 records per-seed counts and the top-1 intervention. Query flags, four aligned predictions where available, metric code, checkpoint/data hashes and verification outputs accompany the export. Aggregate counts are not used to manufacture a paired confidence interval. Five state-identity columns remain explicitly unavailable because the replay did not persist them.

E01 uncertainty and comparators. The E01 export contains all five systems' predictions and integer counts for 65,536, 16,384 and 4,096 queries per fitted seed at the three lengths, together with fitted-seed SDs and all 36 existing null-versus-comparator paired whole-episode intervals. The latter use 2,000 resamples and condition on the fitted seed. Means 0.098/0.071/0.008 points above hard gating are not treated as a general significance or equivalence result; zero errors or zero-width finite-sample intervals are not population guarantees.

Descriptor stress specification. The stress export supplies the exact coordinate space, base and added-noise order, normalization, neighbor rule, tie rule, margin formula, RNG, complete grid, six saved prediction archives and all classwise counts. Aggregate route error remains prevalence-weighted and does not establish correct value retrieval; the separate FPR/FNR and correct-route answer counts prevent that aggregate from hiding branch behavior.

Portable release. This author bundle includes the inherited audits plus executed source/protocol/test files, configs, 27 selected checkpoints, D01/D02/D03/E01 prediction exports, hashes and artifact-only verification. It is still an author package rather than an anonymous public supplement: the historical environment lacks a complete lock file, the repository has no confirmed public-release license, and venue/anonymity/statement checks remain. These release issues do not turn completed experiments into pending runs.