COGNITIVE SCIENCE AND CONTINUAL LEARNING

When Does an Observation Become a Memory? Learning, Recall, and Evidence Across Systems

A cross-system synthesis of controlled neural writes, human event replay and time-stamped first-person evidence

Revanth Matha · Matha Labs | Preprint awaiting review 1.0 | 2026-11-20

Abstract

A later answer can reflect retained experience, a helpful cue or a lucky guess about the present. We distinguish exposure, retained state, retrieval and evidentiary support through three paired analyses and a frozen-model transfer experiment. Across 12 matched symbolic streams, protecting eight initial values improved their retention by 54.34 percentage points and reduced one-hop current accuracy by 5.86 points. Across 347 cued events in 40 people, doubling event duration within a participant was associated with an 11.15% longer geometric replay time (95% interval 6.10% to 16.26%). Of 500 questions about one EgoLife wearer, 115 cited evidence at least a day old and 110 did so even for their nearest cited clue. In 192 symbolic histories, a frozen neural writer with supplied evidence and episode rules retained the first sighting and correctly withdrew unsupported present claims. Carrying the last sighting forward matched the physical world in 144 cases while making 96 unsupported assertions. These results identify testable differences among preserved content, cued report and justified knowledge. We propose matched transfer tests for visual identity, consolidation and developing capacity.

Keywords: episodic memory, fast weights, within-person analysis, egocentric video, source monitoring, continual learning

Contributions

1. Three paired analyses that retain the original denominators and distinguish retention of an initial value, recall timing and the age of an annotated clue.

2. An executed transfer of a frozen numerical writer into 192 wearer-specific symbolic histories with explicit, independently scored current and past states.

3. A matched protocol for testing identity, source and episodes under interference, cue changes and developmental resource schedules.

1. Plain-English summary

A mug sits on your desk when you leave the kitchen. An hour later you can name the mug and point to the desk. A roommate has moved it. You have answered one question correctly about your last sighting. To answer where the mug is now, you need new information. Give a camera both views and it can document the move. Give a learner only your view and a guess can be lucky. These are three different ways to obtain the same spoken answer.

We examine how to tell them apart. A small neural network lets us change which earlier answers its weights preserve. People wearing cameras let us compare how long an event lasted with how long they later spent mentally replaying it. The EgoLife first-person benchmark lets us measure how far a question points back into a recorded week. Finally, we run the neural writer on stories in which the mug may have moved outside the wearer's view. Each experiment reveals a different reason a later answer can succeed or fail. [1][2][3]

An earlier experience counts as evidence for remembering when we can say who had access to it, what later response it could influence, which cue elicited the response, and how to score it. The strongest test changes one of those conditions while keeping the others fixed. Where an intervention is unavailable, we report the narrower observation that the data support.

2. Four questions behind a memory claim

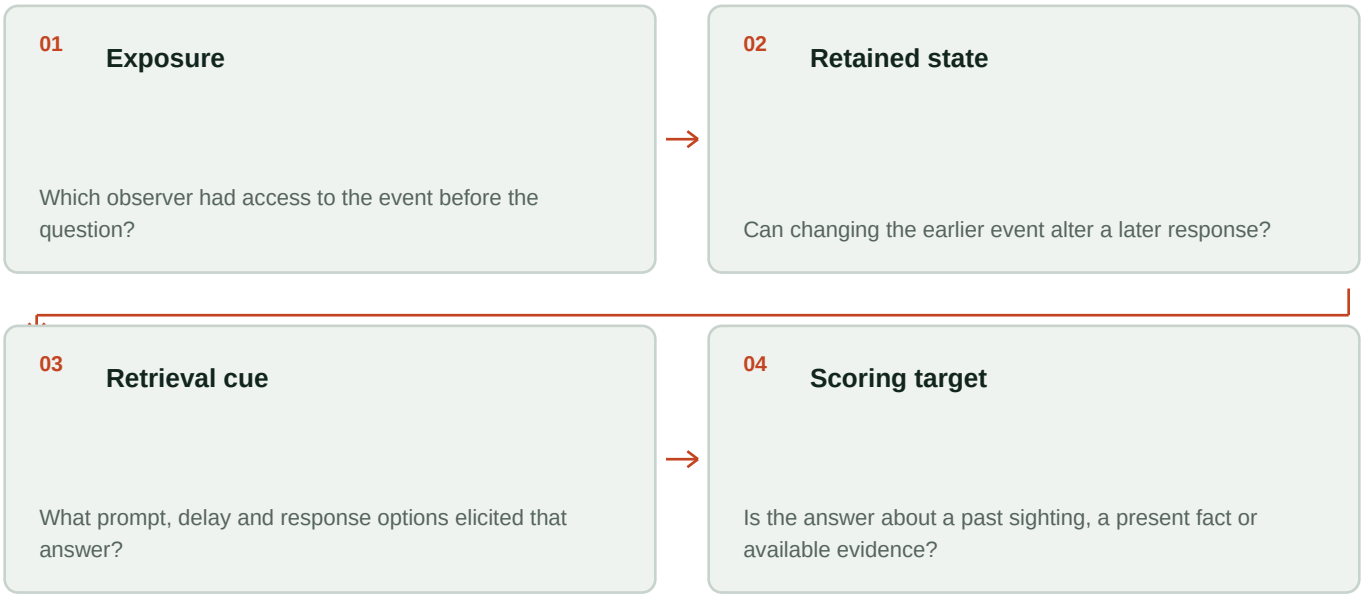
An observer can have an opportunity to see a mug without attending to it. A neural writer can change a weight while leaving every answer we test unchanged. A particular cue can elicit a correct answer from general knowledge alone. A faithfully retrieved past fact can be false today. Claims about memory become testable once the experiment specifies which of these ambiguities it resolves.

Let H be the history available to an observer by the query time, R a retained state, q a question and Y the answer. For a machine we can replay two histories differing in one earlier event, fix the final input and reader, and compare answers under R(H) and R(H′). A change in the answer demonstrates the event's influence for that reader and cue. To locate the influence in the weight array, carry only the array across processes while fixing every other input. An unchanged answer can arise from a failed write, an inaccessible trace, two histories sharing the same answer, or a probe that ignores the changed feature. A probe with different cues can discriminate some of these cases. [1][8]

For a person, behavioral evidence ordinarily stops at the response. A randomized exposure and a later memory test can establish an effect of the exposure on behavior. The state that mediates this effect remains unobserved in the wearable-camera data used here. For a wearer's assistant we can inspect a time-stamped stream, but the analyst still needs to determine which clues the wearer could have obtained. Speech, a sound,

and a first-person view can carry different information about the same event. [2][3][6]

Figure 1. What a delayed answer can establish



The boxes are measurement questions. The history intervention is available in the controlled writer. The human data supply timed, cued replays. The camera annotations identify cited evidence times. The figure specifies the inference available in each setting.

3. Materials, research questions and study design

We carried out an exploratory secondary analysis of three Matha Labs preprints with a shared concern about the persistence of experience. We examined the released experimental definitions, matched stream results, participant-grouped event durations and public EgoLife question annotations. We selected primary external work when it experimentally manipulated an observation, cue, resource or write rule, or described the collection of the underlying data. We also examined an independent replication of an environmental-cue experiment. This focused selection supplies alternative mechanisms and test designs. It cannot estimate publication bias or the relative frequency of mechanisms in the literature. [1][2][3][12][13][14]

For the machine contrast we selected the revisable and eight-anchor conditions with the same interval-unlock schedule, pretrained kernel, world, stream seed, number of optimizer steps and replay-sampling design. We subtracted arm scores within world and seed, then summarized the four paired worlds for each seed. For the human contrast we fit a slope from variation within each participant, preserving participant membership during uncertainty estimation. For the camera contrast we computed two definitions of evidence age on the same questions. There is no common participant or response variable across the three datasets. The comparison concerns which intervention or measurement can discriminate explanations. [1][2][3][4][5][6][7]

To bring two domains together experimentally, we also ask the frozen writer from Paper 06 to act on symbolic wearer histories from the visibility design of Paper 08. This test shares the writer, response vocabulary and

source restrictions across conditions. Object identity and evidence availability enter as supplied features. Human behavior is absent from this experiment. The human analysis contributes an independently observable relationship and a rationale for varying cues in subsequent tests. [1][2][3]

Table 1. What is measured in each arm

Arm	Unit and reference	Reach of inference
Controlled writer	Matched world within a stream seed. Readouts scored against initial or evolving values.	Effect of an address policy under one pretrained kernel.
Human replay	A cued event nested in a participant. Both durations belong to that event.	Within-person association of event duration with replay time.
First-person benchmark	A curated question for one wearer. Annotators cite target times.	Age of a cited clue under two clock definitions.
Controlled histories	A symbolic history with separate wearer evidence and hidden-world labels.	Effect of evidence-aware updates and the requested readout.

Responses have unlike units and reference standards. Repeated probes, events and questions inherit dependence from their stream, participant or wearer.

4. What protecting a weight preserves

The writer learns numerical changes to a 16-row, 64-column array of inference weights. It sees the current vector at the target address, a bank-wide mean, the operation and its payload. The bank mean allows all retained columns to influence a summary, while discarding which column contributed a particular value. A trained decoder reads an addressed vector. SET, COPY and RETRACT are learned during core training. Online adapters can learn subsequent operations. Keeping the core frozen gives us a consistent device with which to test a permission rule. [1]

The eight-anchor rule rejects writes to the first eight profile addresses. Every other address remains revisable. The evolving target can change even at a protected address, so the same refusal can constitute success on initial-state retention and failure on current-state tracking. To expose the consequence, each arm receives 128 events and 512 optimizer steps in each of 12 matched streams. Each hop has 8,192 post-update readouts per stream, or 98,304 per arm. The anchor probe tests eight initial values after each event, yielding 12,288 responses per arm. The distinct sources of variation are three stream seeds with four generated worlds each. Repeated reads of a protected cell do not amount to thousands of independent demonstrations. [1]

Protecting the anchors increased retention of the original values by 54.34 percentage points while decreasing current one-hop accuracy by 5.86 points and two-hop accuracy by 12.67 points. The sign of each contrast was the same for every stream seed. The larger two-hop loss is consistent with one stale value affecting more than one query, but the aggregate contrast cannot attribute each failure to a particular intermediate value. The initial-anchor result measures obedience to a known protection rule. Whether a learner could discover which observations merit protection requires a separate test.

Table 2. Matched change when eight starting values are protected

Readout	Revisable	Eight anchors	Difference (pp)
Current, one hop	93.60%	87.74%	-5.86
Current, two hops	91.90%	79.23%	-12.67
Eight original values	45.66%	100.00%	+54.34

Difference is eight-anchor minus revisable. Across the three seed means the ranges were -7.13 to -4.66, -14.88 to -9.87 and +47.63 to +61.23 percentage points, respectively. Ranges describe these runs. The world and seed pairing keeps the compared streams identical. [1]

5. What changes when the same person replays a longer event?

In the wearable-camera study, Colson and D'Argembeau had participants select everyday events over three consecutive days. Two to four days later, photographs from the beginning and end of each event cued silent mental replay. Participants marked its start and end with a keypress. The dataset contains 358 rows. We applied the authors' upper-duration exclusion for mental replay, dropping the 11 rows at or above 187 seconds. The remaining 347 events came from 40 people, with four to ten events per person. Median sampled event duration was 13.13 minutes and median replay duration was 31.73 seconds. A population in which slower replayers also tend to record longer events could produce an apparent relationship across rows even if a given person takes equally long to replay short and long events. We therefore asked a within-person question. [2][4][5]

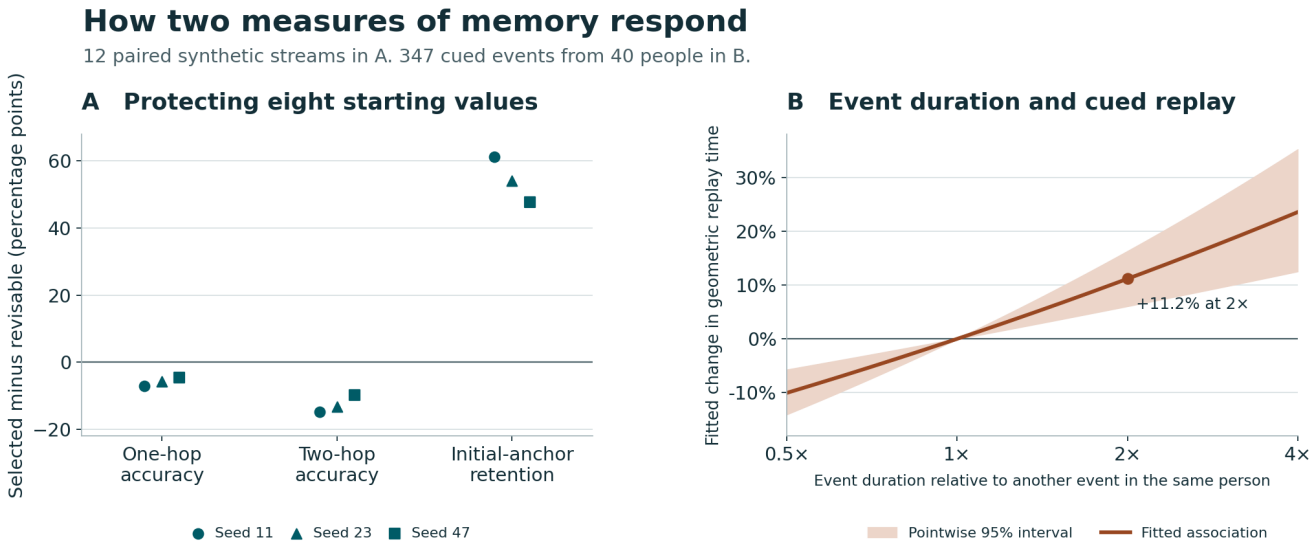
Colson and D'Argembeau had already found a nonlinear within-person association in a generalized additive mixed model. Our reanalysis estimates a single interpretable slope and the corresponding change for a doubling of event duration. We used natural logarithms of event minutes and replay seconds and subtracted each participant's average log duration from their log durations for each variable. We fit an ordinary least-squares slope to the paired deviations. This is equivalent to giving each participant a separate intercept and fitting one shared slope. People with more events or wider duration spread contribute more information about the slope. We drew 10,000 bootstrap samples of entire participants with replacement, preserving their events together, and took the 2.5th and 97.5th percentiles. The transformation back to a relative duration assumes the fitted log-linear relation over the observed range. [4][5]

The slope was 0.153 (95% participant-bootstrap interval 0.085 to 0.217). Comparing two events for the same participant, doubling experienced duration was associated with an 11.15% longer geometric replay time (6.10% to 16.26%). This partially controls stable differences among participants. Event type, interest, rehearsal and the content visible in the cue can still vary within a person. Accuracy of recalled details and the fraction of sensory experience encoded remain unmeasured. A longer replay can include more details, longer pauses, or a different way of reconstructing the episode. [4][5]

```
ln(replay seconds) = participant intercept + b x ln(event minutes) + residual change for twice the duration = 100 x (2^b - 1)%
```

The factor of two refers to experienced event time. The response is relative geometric replay duration for a person under the fitted model.

Figure 2. Write protection and cued replay on separate measurement scales



Panel A reports paired changes in percentage points within generated streams. Panel B displays relative geometric duration under the within-person slope and a percentile band from 10,000 resamples of participants. The axes represent distinct tasks. [1][4][5]

6. A witness's clock and a world's clock

EgoLife recorded a group sharing a house using first-person devices and external cameras. The public A1 JAKE question file contains 500 questions on a seven-day clock. A question may cite several evidentiary times. We computed the elapsed age of both the oldest and nearest cited time against the query clock, retaining the provided day and whole seconds. A repeated time within a question has no effect on its minimum or maximum. The original analysis found 564 target occurrences, 563 distinct within question and 48 questions with multiple distinct times. Age describes where the annotator pointed. It leaves open whether the assistant could have used another clue or the wearer attended to this one. [3][6][7]

Using the oldest cited clue, 115 of 500 questions reached at least 24 hours back. Using the nearest clue, 110 reached that threshold. The matched classification contained 110 above threshold under both definitions, 5 above under oldest alone, none above under nearest alone and 385 below under both. Changing the target definition changed one percentage point of the full file. The impossible nearest-only cell follows from the order of the two clocks and supplies a parsing check. Median ages were 5.87 and 5.79 hours, respectively. The one-point contrast describes this curated file for one wearer. [3][7]

An analyst could use another resident's camera to resolve a hidden move. The resulting label describes the world at the query time. A system with only the wearer's record may still lack evidence for that label. For two otherwise identical histories in which an unobserved transfer occurs in just one, every deterministic reader of the wearer's record must return the same current-location answer for both. A forced choice necessarily misses at least one of those two cases. Adding a channel available to the wearer, such as a reliable spoken report,

changes the premise. A recording gap by itself cannot establish either a transfer or its absence. [3][6]

7. Can the same writer remember a sighting and withdraw a claim?

We executed an additional experiment joining the actual Paper 06 writer to the observation rules studied in Paper 08. We loaded the frozen codec and writer, verified the checkpoint and source-file digests, and began each history with an empty array. Eight pairs of familiar location values each supplied two possible starting locations. We crossed the start with a subsequent move or no move, a visible or occluded observation, and zero, eight or 32 distractor writes. The full factorial has 192 histories, with 96 visible and 96 occluded. Every occluded moved history has a stationary partner with identical evidence, including the same distractors. [1][3]

The controller writes the first sighting to last-seen, currently supported and first-episode addresses. On a later sighting it updates the first two addresses and writes a later episode. On an occlusion it retracts the currently supported value to the unknown category. The last sighting and initial episode remain queryable. Distractor writes use other addresses. We queried all four addresses with the frozen decoder. The room label appears only in evaluation. This scoring rule requires evidence of the current location. The experimental world permits an uncertain transfer during every occlusion, so the last confirmed location becomes inadequate evidence for a present claim.

The writer and supplied evidence rule produced four correct readouts in every constructed history. All 64 histories with 32 distractors triggered weight compaction and still answered all four queries correctly. A carry-forward baseline used the same decoded last-seen value for current location. That policy matched the hidden physical world on 144 of 192 cases, yet gave 96 unsupported location assertions, one for every occlusion. Half of those assertions were lucky because the hidden object stayed put. Retracting the current claim left the earlier episode available. This demonstrates that these operations can coexist in a trained array under supplied semantics. It leaves visual identity, evidence discovery, calibrated trust and learned invalidation for future experiments.

Table 3. Four readouts on the same 192 histories

Readout	Visible (96)	Occluded (96)	Reference
Last sighting	96/96	96/96	Most recent observed value
Supported current location	96/96	96/96	Final visible value or unknown
First episode	96/96	96/96	Initial sighting
Later episode	96/96	96/96	Final sighting or absent
Carry-forward, supported current	96/96	0/96	Same present-evidence target

Each row uses the same histories. The supplied rule and decoder jointly achieved all four target readouts in 192/192. The carry-forward baseline matched the room label in 96 visible and 48 hidden cases. Room agreement and evidence-supported accuracy have distinct targets.

8. Retrieval cues and competing explanations

The human slope quantifies how the duration of a person's replay varies with event length. A distinct measure is how much visible detail the person reports accurately. Finley and Brewer had nine undergraduates wear a camera over two days and recall selected time periods one week and one month later. The students compared their accounts against the photographs. Mean self-scored accuracy was 89% for the scored visible details. Details outside the frame, thoughts during the event and untested experiences cannot enter that percentage. A 16-hour digital recorder at a fixed five megabits per second produces 36 decimal gigabytes. Converting the replay slope or detail score to biological gigabytes would additionally require a specified neural encoding and a test that identifies which distinctions persist under that encoding. [2][4][18]

Retrieval itself can alter later performance. Roediger and Karpicke had students study prose and either restudy it or attempt free recall without feedback. Restudy gave an advantage on the immediate test. Prior testing gave an advantage on delayed tests. A protocol that repeatedly probes the same person has changed that person's exposure. A machine readout can be defined to leave the weight state fixed, though a system that trains on its own answer introduces an analogous intervention. For a fair retention comparison, record every previous probe, and compare an unprobed delay group or a matching frozen-state readout. [12]

A specific candidate cue can also fail. An original diver experiment found a recall advantage when words were learned and tested in the same environment. A later within-person replication with 16 divers found comparable recall across matched and unmatched environments after a four-minute delay. The effect of changing a cue therefore requires its own controlled comparison in the task of interest. Once several associations share a cue, a missing answer could reflect competition among traces or inaccurate attribution of a remembered detail to an episode. An experiment on reminding before a second object list elicited intrusions from that list during recall of the first. A computational temporal-context account reproduced the qualitative asymmetry through context binding. Their agreement at the level of observable errors leaves the internal explanation open. [13][14][15][16]

9. Machine mechanisms and the cost of losing provenance

Weight-based memory has a prior experimental literature. Schlag, Irie and Schmidhuber showed how outer-product updates implement key-value bindings and used a delta rule to correct a previous association. Titans learns parameters of a long-term memory module at test time. MaLoW uses a recurrent hypernetwork to produce lightweight weight updates from streaming video while keeping the current observation at the input. Each can let previous experience alter a later inference. The address permissions and episode readouts in our experiment are distinct controls. This study makes no performance comparison with those models because their input spaces, training and tasks differ. [8][9][10]

There are at least three reasonable ways to remember the mug. A system might retrieve the original clip with a timestamp, use a changed parameter to answer directly, or cache a compact summary and consult the clip only when asked for a source. MemLife examines time-indexed, entity-linked textual episodes with a retrieval agent.

A cited caption makes assumed identities easier to inspect. It can also encode the wrong identity and then spread that error across later episodes. A parameter-based answer may be fast to read but requires a separate check to recover which observation justified it. The practical comparison is a common set of scenes and questions, with costs for encoding, updating, storing, querying and inspecting evidence charged to every system. [3][10][11]

Separating an entity from its episodes gives consolidation a concrete objective. Evidence from repeated views may strengthen recognition of one mug while preserving the two locations and their timestamps. Before accepting a merge, test identity with same-looking foils and test each episode using a location or source cue. Compare resulting errors with a no-merge condition at the same incoming rate. If splitting after a correction restores the episode address but leaves a learned visual association elsewhere, the correction has superseded a belief while part of the original influence persists. Demonstrating complete removal would require a specified family of probes and a counterfactual learner trained without the experience. [1][3][8]

10. Development changes what a performance threshold means

An infant can perform differently as the body changes even when the skill being tested has not suddenly disappeared. Thelen and colleagues added weight to the legs of four-week-old infants and observed reduced stepping. Buoyancy in a separate group increased stepping. That manipulation altered conditions for expressing a movement during the test. In a different domain, Dehaene and colleagues compared adults with different histories of literacy, including people taught to read as adults. Their fMRI responses varied with literacy history. Group differences in a skill with instruction and practice have a different evidential basis from a directly manipulated load. Both examples motivate separate measures of task demands, experience and available resources. [19][20]

Different signals can also integrate over different spans. Baldassano and colleagues related event boundaries in continuous narratives to patterns of neural activity and later recall. Preferred spans varied across regions in their analysis. The measure was a model of fMRI activity during a particular narrative task. A periodically consolidating writer could be tested at half-second, one-second and two-second intervals as engineering conditions. This imaging result does not estimate any of those operating periods for a person planning a career or reacting to a threat. [17]

A developmental claim for a machine is stronger than showing that an adapter passed yesterday's threshold. Paper 06 activated extra capacity under time, compute, experience and competence policies. Uniform optimizer work per event made time and cumulative-compute thresholds coincide. Its competence gate used recent labeled replay. These designs can test scheduling and implementation, but a claim about learning how to learn requires genuinely new operations, fresh examples and independent variation in resources. Stage new compute separately from the onset of a new skill, give all arms the same eventual optimization budget, and withhold separate transfer probes from every gate. [1]

Define an acquisition endpoint in advance. Count examples until a model achieves the designated accuracy on two independently drawn, unseen test batches for the new skill. Also test earlier skills, identify the number of

trials that never cross the criterion before the budget runs out, and report the learning curves. An apparent reduction in examples disappears as evidence for improved learning if extra compute alone explains it, if only familiar templates pass, or if old skills fall below the retention floor. Inspect frozen-core bytes and run its original behavioral probes so the meaning of an unchanged core remains testable. [1][19][20]

11. A matched experiment that could distinguish the mechanisms

Consider two look-alike mugs, two possible locations and two observers. One sees a transfer. The other gets an occluded interval and sometimes hears a dated spoken report. Independently mark who had visual evidence, who had testimonial evidence, and the true object positions as verified by a separate observer. Introduce contradictory reports on a prespecified subset. Freeze all channels at query time. The learner receives only the assigned channels. Evaluate identity, source, last-seen location, past episode and current support separately. A cue asking about a sound must admit the possibility that the source of the sound is ambiguous.

For a human analogue, present consented participants with controlled scenes and use counterbalanced conditions within participants. Prespecify which details can be verified. Randomize free recall, recognition with plausible foils and a source judgment across distinct event sets to avoid testing one set into memory before the next measure. Include missed targets, false alarms and facts reported about the wrong episode. Record response time independently of accuracy. For the machine, preserve visibility, object novelty, delay and conflict conditions, and compare performance under a chronological clip index, a corrective fast-weight baseline and the learned writer. A designed common scene makes the scoring questions comparable while leaving each system's mechanism open. [3][8][12]

Allocate complete people, objects and scenes to held-out partitions as appropriate. Keep the question about an event and its alternative hidden-world realization together in a split. Where several questions share a scene, estimate uncertainty by resampling scenes, with between-person and between-seed variation reported separately. Calibrate confidence and abstention on validation scenes, then freeze the decision rule. Report coverage with accuracy among answered questions, and unsupported assertions among all occluded questions. A system can reduce unsupported answers by refusing everything, so also report retrieval among the visible cases.

Table 4. Predictions a follow-up can overturn

Hypothesis	Intervention and readout	Outcome that would challenge it
Protecting evidence helps selective retention.	Randomize which initial facts stay valid. Hold later changes and query mix fixed. Score both initial and current answers.	A blind static rule matches evidence-conditioned protection on new change rates.
Consolidation can preserve separate episodes.	Merge or retain matched views. Test identity with similar foils and query old locations separately.	Identity improves only while old-location recall collapses.
A cue can recover an accessible but unreported detail.	Assign cues to comparable event sets. Count newly correct details and new intrusions.	More reports arise entirely from intrusion or cue-supplied information.
Knowledge depends on viewpoint.	Hold the wearer stream fixed, vary an occluded move, and reveal a reliable report in only one arm.	The answer varies across identical histories or depends on a future frame.
A developing writer learns new skills more efficiently.	Cross extra compute with new-skill timing. Match total work and score examples to criterion plus old skills.	Earlier criterion crossing disappears on fresh tasks or costs unacceptable old-skill loss.

This is a prospective design. The symbolic transfer audit tests supplied rules. Human cue randomization, learned visual identity from the house, learned consolidation and new-skill transfer still require these experiments.

12. Uncertainty and reproduction

All three data sources condition heavily on what was sampled. One pretrained symbolic kernel faced a finite vocabulary and known addresses. Three stream seeds allow a check of direction across these runs. The transfer audit used distractors outside the target addresses, so it probes compaction and representational survival without direct key collisions. A system that must infer which mug it saw or allocate its own address faces additional failure modes. The perfect symbolic score is best used as a baseline for measuring those added costs. [1][3]

The participant bootstrap reflects sampling variation conditional on the available participants and cued events. The slope was 0.164 when all 358 events were retained and ranged from 0.133 to 0.165 across leave-one-person-out fits. Those checks limit sensitivity to the duration exclusion and a single participant. Selection of which events were cued and unmeasured factors within a person still limit causal interpretation. The EgoLife analysis uses one wearer's released questions and the times that annotators assigned to their clues. Human free-recall performance was unmeasured. No cross-person or cross-house interval can be estimated from that file alone. [4][5][6][7]

The private companion repository contains the matched analysis, the figure inputs, the frozen-checkpoint history replay and row-level synthetic outcomes. The analysis reads pinned artifacts from Papers 06-08 and verifies source digests. It can download the public numerical workbook or accept a local copy with the expected digest. Reviewers can rerun the whole-participant bootstrap, reconstruct the two age-threshold counts and

independently compare moved/stationary pairs. The public manuscript reports aggregate human results. The transfer experiment uses no household recordings or newly recruited participants. [1][2][3][5][7]

13. What survived the observation?

In the controlled array, an earlier value can survive later writes, be retrieved under the appropriate question and coexist with uncertainty about the present. Protecting it also carries a measurable cost when the world changes. For a person, a longer sampled event tends to take somewhat longer to replay, but timing alone leaves its contents undetermined. For a wearer, one day of evidence age identifies the distance to an annotated clue. The source and the wording of a question determine whether that clue supports an answer. A memory experiment advances when it can show exactly which earlier distinction still changes which later response, and under which conditions that response is warranted.

References

[1] Matha R. Programmable Neural Memory: Recognition and Learning in a Continuously Updated Weight Array. Matha Labs preprint 06. Controlled stream conditions, frozen codec and writer, developmental gates and readouts. Accessed 2026-11-20.

[2] Matha R. How Much of a Day Do We Remember? Sensory Input, Attention, and the Development of Memory. Matha Labs preprint 07. Initial analysis of event durations and cued replay, plus developmental evidence. Accessed 2026-11-20.

[3] Matha R. What Did the Wearer Last See? Evidence and Memory in First-Person Video. Matha Labs preprint 08. Released question-clock analysis and observation-equivalent hidden worlds. Accessed 2026-11-20.

[4] Colson C, D'Argembeau A. Memory in motion: how real-life event features influence the tempo of episodic recall. Memory, 2025. Naturalistic event recording, beginning/end picture cues and participant-marked replay times. Accessed 2026-11-20.

[5] Colson C, D'Argembeau A. Open data and materials for Memory in Motion. Original numerical workbook, scripts and protocols for the 40-participant study. Accessed 2026-11-20.

[6] Yang J and colleagues. EgoLife: Towards Egocentric Life Assistant. CVPR 2025. Household recording conditions, views and question construction. Accessed 2026-11-20.

[7] EgoLife. EgoLifeQA_A1_JAKE.json. Pinned official public question annotations used for the oldest/nearest comparison. Accessed 2026-11-20.

[8] Schlag I, Irie K, Schmidhuber J. Linear Transformers Are Secretly Fast Weight Programmers. ICML 2021. Fast-weight programming, corrective writes and synthetic retrieval tests. Accessed 2026-11-20.

[9] Behrouz A, Zhong P, Mirrokni V. Titans: Learning to Memorize at Test Time. Authors' primary model report. Neural long-term memory updated during use. Accessed 2026-11-20.

[10] Liu Z and colleagues. Memory as Weights: Internalizing Long-Term History for Streaming Videos. 2026 preprint. MaLoW uses recurrently generated low-rank updates for streaming visual history. Accessed 2026-11-20.

[11] Xiong G and colleagues. MemLife: Curating and Reasoning over Long-Term Egocentric Video Memories. 2026 preprint. Entity-grounded text episodes and time-indexed retrieval. Accessed 2026-11-20.

[12] Roediger HL, Karpicke JD. Test-enhanced learning: taking memory tests improves long-term retention. Psychological Science, 2006. Free-recall testing and restudy compared at differing retention intervals. Accessed 2026-11-20.

[13] Godden DR, Baddeley AD. Context-dependent memory in two natural environments: on land and underwater. British Journal of Psychology, 1975. Primary environmental-context experiment with divers. Accessed 2026-11-20.

[14] Murre JMJ. The Godden and Baddeley (1975) experiment on context-dependent memory on land and underwater: a replication. Royal Society Open Science, 2021. Independent test of the matched-environment recall effect. Accessed 2026-11-20.

[15] Hupbach A, Gomez R, Hardt O, Nadel L. Reconsolidation of episodic memories: a subtle reminder triggers integration of new information. Learning & Memory, 2007. Reminder, new list and later first-list intrusions. Accessed 2026-11-20.

[16] Sederberg PB and colleagues. Human memory reconsolidation can be explained using the temporal context model. Psychonomic Bulletin & Review, 2011. Model-based alternative account of the reminder-dependent intrusion pattern. Accessed 2026-11-20.

[17] Baldassano C and colleagues. Discovering Event Structure in Continuous Narrative Perception and Memory. Neuron, 2017. Event organization across regions during continuous narratives and recall. Accessed 2026-11-20.

[18] Finley JR, Brewer WF. Accuracy and completeness of autobiographical memory: evidence from a wearable camera study. Memory, 2024. Timeslice recall and participant scoring against SenseCam photographs. Accessed 2026-11-20.

[19] Thelen E, Fisher DM, Ridley-Johnson R. The relationship between physical growth and a newborn reflex. Infant Behavior and Development, 1984. Stepping under changes to leg weight and buoyancy. Accessed 2026-11-20.

[20] Dehaene S and colleagues. How learning to read changes the cortical networks for vision and language. Science, 2010. Adult brain responses across literacy histories. Accessed 2026-11-20.

Reproducibility record

Private review materials: https://github.com/RevanthM/memory-across-systems | Research snapshot: a3dc68654bea409d9f825521511481e2c3093925