NOISYUPDATEBENCH: SEPARATING BELIEF FROM COST-SENSITIVE ACTION IN LONG-TERM MEMORY REVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term memory systems operating in changing environments must decide whether new evidence should supersede an incumbent fact, be held as uncertain, or be quarantined. Existing evaluations often summarize this process through aggregate question-answering accuracy, obscuring asymmetric failures such as false supersession and missed updates. We introduce NoisyUpdateBench, a controlled diagnostic benchmark comprising noisy, retracting state streams with latent ground truth, versioned current and historical queries, and evidence-identical clones evaluated under application-specified costs for false supersession, missed updates, and quarantine. We formalize memory revision as cost-sensitive sequential decision making and evaluate transparent controllers that separate belief estimation, expected-cost action selection, and candidate quarantine. On a precommitted synthetic held-out suite of 120 base streams, dynamic Bayesian filtering with expected-cost decisions attained the lowest pooled normalized risk cost among ten configurations (0.676), improving over last-write-wins ($\Delta\mathrm{NRC} = -0.655$, 95% paired CI $[-0.833, -0.482]$); evidence-identical clones preserved identical beliefs while inducing cost-dependent actions. On a sealed, independently authored synthetic external suite of 60 cases, the same controller again improved over last-write-wins and its matched fixed-policy variant, while comparisons with strong Weighted Vote baselines remained inconclusive. With automatic extraction from raw-language sessions, the revision layer reduced NRC relative to naive overwrite, full-context prompting, and one frozen Mem0 OSS configuration ($\Delta\mathrm{NRC} = -0.770, -0.698, -1.241$; all paired CIs excluded zero) and reached 0.962 exact-match extraction F1, though its NRC gap to a gold-extraction reference remained inconclusive. Scenario winners and quarantine effects varied across datasets, precluding universal-superiority claims; instead, the results support benchmark-first, cost-sensitive evaluation and identify where explicit belief and decision modeling helps—and where simple baselines remain competitive.

1 INTRODUCTION

Long-term memory systems must revise stored facts as users, environments, and external evidence change. A new observation can be a genuine update, noise, a temporary state, or a retracted claim. Switching too quickly can falsely supersede a valid incumbent; waiting can miss a real update; and deferral avoids either error only by paying a quarantine cost and reducing coverage. Revision therefore depends on both what a system believes and how an application values these asymmetric outcomes.

Benchmarks have established that temporal memory, state tracking, conflict, forgetting, and abstention are difficult (Wu et al., 2025; Hu et al., 2026; Uddin et al., 2026; Chao et al., 2026; Tao et al., 2026; Long et al., 2026; Fan et al., 2026; Jiang et al., 2026; Varlamov et al., 2026). They leave a narrower diagnostic question: *given identical evidence and belief, does a controller change its action when only the error costs change?* Aggregate QA accuracy cannot answer this because it conflates state estimation, decision policy, and the consequences of directional errors.

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.

We introduce NOISYUPDATEBENCH, a controlled benchmark for cost-sensitive online memory revision. Streams contain a latent versioned binary state and observations with supplied source-reliability scores that the reference filters interpret probabilistically. Controllers maintain current and historical memory while choosing to switch to a candidate, keep the incumbent, or quarantine uncertainty. We score false supersession (FS), missed update (MU), and quarantine (Q), together with coverage, selective accuracy, historical accuracy, and belief quality.

The central intervention is an *evidence-identical risk clone*. Each base stream is copied across cost profiles while its latent timeline, observations, sources, reliabilities, and queries remain fixed. A cost-independent estimator must therefore preserve its belief trajectory, while a cost-sensitive policy may rationally change actions. Clones remain clustered by base during resampling. This construction provides a deterministic check for risk leakage into belief formation and measures action adaptation without treating correlated interventions as independent examples.

We evaluate LWW, a cumulative reliability-mass Weighted Vote (WV), static and dynamic Bayesian filters, fixed and expected-cost policies, and no-quarantine ablations. RA-SBF instantiates the complete decomposition—dynamic Bayesian belief updating, myopic expected-cost action selection, and candidate quarantine—as an interpretable reference controller. Its modular structure exposes which effects arise from belief estimation, decision policy, and deferral, while Track B tests how the same decision layer composes with raw-language memory pipelines such as Mem0 (Chhikara et al., 2025; Rasmussen et al., 2025).

Results reveal a structured risk–coverage landscape. On 120 precommitted canonical bases, RA-SBF attains the lowest pooled cost among ten configurations under the equal-profile diagnostic mixture, with its largest gain in false-supersession-critical conditions; WV remains competitive and scenario winners vary. On a sealed, independently authored synthetic suite of 60 bases, Track A reproduces the LWW and matched-policy effects while identifying a dataset-dependent quarantine trade-off. Track B estimates the incremental effect of adding benchmark-aligned risk control to frozen native systems: it reduces post-initialization action cost and false supersession under the tested costs while occupying a lower-coverage operating point. A matched-cost development analysis places RA-SBF and full-context beliefs on complementary regions of the cost simplex. The held-out evaluations map where cost-sensitive control changes the risk–coverage operating point; development sensitivities guide future confirmatory evaluation.

Our contributions are:

- a versioned, cost-sensitive formulation with explicit FS, MU, and Q;
- evidence-identical risk clones that audit the separation of belief formation from cost-sensitive action;
- paired canonical and independently authored transfer evaluations that map regime-specific strengths and failure modes; and
- a versioned adapter, scoring, cost-clone, and cluster-bootstrap toolkit for extending the diagnostic to new memory systems.

2 RELATED WORK

Long-term memory evaluation. Conversational and agent-memory benchmarks test long-horizon retrieval, temporal reasoning, knowledge updates, and abstention from insufficient evidence. LoCoMo introduced very long multi-session conversations (Maharana et al., 2024); LongMemEval organized long-term interactive memory around information extraction, multi-session reasoning, temporal reasoning, knowledge updates, and abstention (Wu et al., 2025); and MemBench and MemoryAgentBench broadened evaluation across memory levels and incremental operations (Tan et al., 2025; Hu et al., 2026). BEAM and RHELM further stress-test architectural scale and heterogeneous real-world memory tasks (Tavakoli et al., 2026; Zhang et al., 2026). These suites ask whether information can be stored and recovered. We instead isolate a narrower decision made *during* revision: what to do when a new observation conflicts with the currently active version and different errors have different application costs.

Updates, conflicts, and state tracking. Recent work directly studies stale memories, conflicting evidence, forgetting, and state evolution. Memora evaluates forgetting (Uddin et al., 2026); STALE probes implicit stale-state resolution, premise resistance, and downstream policy adaptation (Chao et al., 2026); MemConflict controls conflicts (Tao et al., 2026); and MemTrace follows knowledge points across memory age, question type, and evidence condition (Long et al., 2026). State-centric or temporal evaluation also appears in A-TMA, StateMemBench, Supersede, and TEPA (Shi et al., 2026; Fan et al., 2026; Patel, 2026; Zhou et al., 2026). Our contribution is the joint intervention: evidence and latent trajectory remain fixed while application loss varies, separating estimation failures from decision failures in versioned revision.

Selective and risk-aware decisions. The option to reject a prediction has a long history (Chow, 1970), and selective classification formalizes the associated risk–coverage trade-off (El-Yaniv & Wiener, 2010). Recent memory work has considered safety, risk taxonomies, arbitration, and quarantine, including PerMemSafe, MemRiskBench, and MemSentry (An et al., 2026; Jiang et al., 2026; Roy & Basu, 2026). MemToC instead studies post-tool-return arbitration between tool evidence and a model's parametric memory, rather than revision of a persistent external memory store (Varlamov et al., 2026). Our use of quarantine follows this selective-decision tradition. The distinct diagnostic object is an *evidence-identical risk clone*: clones must induce the same belief, but may rationally induce different actions under different false-supersession, missed-update, and quarantine costs.

Memory systems. Systems such as MemGPT, Mem0, Zep, and APEX-MEM combine extraction, storage, retrieval, and response generation in different ways (Packer et al., 2023; Chhikara et al., 2025; Rasmussen et al., 2025; Banerjee et al., 2026). Our structured experiments isolate the revision decision layer; Track B restores raw-language extraction and evaluates its composition with one frozen open-source Mem0 configuration. The resulting evidence characterizes decision-layer operating points relative to frozen end-to-end systems.

3 Task and NoisyUpdateBench

3.1 Online versioned revision

Each base stream has a latent binary state $x(t) \in \{0, 1\}$ over discrete logical time. At strictly increasing time τ_i, the controller observes

$$e_i = (\tau_i, y_i, s_i, \rho_i), \qquad \rho_i \in (0, 1), \tag{1}$$

where y_i is an observed value, s_i a source class, and ρ_i an application-supplied reliability score. The reference Bayesian controller interprets ρ_i as $\Pr(y_i = x(\tau_i))$; this is a modeling assumption, not an empirical calibration guarantee. Scores below $1/2$ are normalized by replacing (y_i, ρ_i) with $(1 - y_i, 1 - \rho_i)$. A controller may maintain a belief $p_i = \Pr(x(\tau_i) = 1 \mid e_{1:i})$, one active memory version, archived versions, and method-specific state. Versions use half-open validity intervals $[\texttt{valid_from}, \texttt{valid_to})$. An event is processed before any query at the same logical time; queries do not advance time.

Current queries ask for the state after the latest processed event. Historical queries ask for the version valid at a specified target time and must be answered from the version timeline rather than by replaying raw events. The controller sees only the observable event fields and the applicable cost profile. Scenario identity, latent states, oracle answers, future events, and generation provenance remain evaluator-only information. This boundary is checked by the benchmark schema and tests.

3.2 Cost-sensitive revision actions

After initialization, every observation is an eligible revision decision. A conflicting observation proposes candidate $c_i = y_i$ against the active version. The controller may SWITCH to the candidate, KEEP the incumbent, or QUARANTINE the candidate without activating it. Let $q_i = \Pr(x(\tau_i) = c_i \mid e_{1:i})$. With the declared diagnostic loss vector C_{FS}, C_{MU}, and C_{Q}, the one-step risks are

$$R_{\mathrm{switch}} = (1 - q_i)C_{\mathrm{FS}}, \quad R_{\mathrm{keep}} = q_i C_{\mathrm{MU}}, \quad R_{\mathrm{Q}} = C_{\mathrm{Q}}. \tag{2}$$

The expected-cost policy selects the minimum-risk action, with deterministic ties resolved as KEEP$\succ$QUARANTINE$\succ$SWITCH. This is a myopic plug-in Bayes decision, not reinforcement learning: costs and reliabilities are supplied rather than learned. It is optimal only for this one-step loss,

which omits quarantine delay, repeated unanswered queries, human-review workload, and other long-horizon utility.

Evaluation distinguishes three directional outcomes. A false supersession (FS) occurs when the system switches away from an oracle-correct incumbent; a missed update (MU) occurs when it keeps an incumbent that is wrong after the decision; and Q counts quarantine. Initialization is reported separately and does not incur FS, MU, or Q cost. A same-valued observation is coded as KEEP; if the retained active value disagrees with the oracle, that eligible decision counts as MU. These operational labels intentionally score the action actually taken, rather than retrospectively reinterpreting a controller's internal confidence.

3.3 Evidence-identical risk clones

For each base stream b, NoisyUpdateBench constructs clones $b^{(r)}$ that share every event, latent state, oracle answer, and query but receive different cost profile r. Any estimator that respects the information boundary must therefore satisfy

$$p_i^{(r)} = p_i^{(r')} \quad \text{for all } i, r, r', \tag{3}$$

while a cost-sensitive policy may choose different actions. We report both the maximum cross-clone belief discrepancy and action-divergence rate. All clones remain nested within their originating base stream for statistical resampling; three clones are interventions on one evidence sequence, not three independent examples.

3.4 Scenarios, profiles, and queries

The suite crosses six update scenarios with three risk profiles. *Clean* streams provide a truthful sanity check; *false update* injects a convincing but incorrect observation; *temporary change* follows an $A \to B \to A$ trajectory; *retraction* explicitly withdraws a prior update; *recurrent state* revisits earlier values; and *source conflict* orders disagreeing sources under either stable or changing truth. The three author-chosen diagnostic profiles are symmetric $(1, 1, 0.5)$, FS-critical $(20, 2, 1)$, and MU-critical $(2, 20, 1)$, where each tuple is $(C_{\mathrm{FS}}, C_{\mathrm{MU}}, C_{\mathrm{Q}})$. Their purpose is not to estimate deployment prevalence or stakeholder utility, but to expose whether the same belief leads to a decision appropriate to the declared loss.

The canonical held-out suite contains 120 base streams (20 per scenario) and 360 risk clones. It includes 749 base observations and 936 base queries, of which 545 are current and 391 historical. Forced scenario landmarks are mixed with sampled background events, and queries target initialization, immediate and delayed landmarks, phase boundaries, and final states. Difficulty, initial state, source order, subtype, and duration are balanced diagnostic factors. The separately authored external suite contains 60 synthetic base cases (10 per scenario), 260 raw-language sessions, and 310 base queries across ten application domains. It is used twice: Track A supplies its gold structured events, whereas Track B begins from its raw sessions. Thus the external suite tests authorship and surface-form transfer, but is not real-world data and the two tracks are not independent dataset replications.

4 Reference Controllers

4.1 Last-write-wins and reliability-mass voting

Versioned last-write-wins (LWW) activates every new observation, closes the previous half-open interval, and archives the displaced version. It ignores source reliability, risk costs, and accumulated evidence, but supplies an important overwrite baseline with well-defined historical retrieval.

Weighted Vote (WV) accumulates *reliability mass*. For value $v \in \{0, 1\}$,

$$S_i(v) = \sum_{j \le i} \big[\rho_j \mathbf{1}(y_j = v) + (1 - \rho_j)\mathbf{1}(y_j \neq v)\big], \qquad p_i = \frac{S_i(1)}{S_i(0) + S_i(1)}. \tag{4}$$

Evidence is cumulative: there is no transition model, decay, window, or learned parameter. At an exact vote tie, WV retains the incumbent (and uses the first observation at initialization). This implementation is not a cumulative log-odds model.

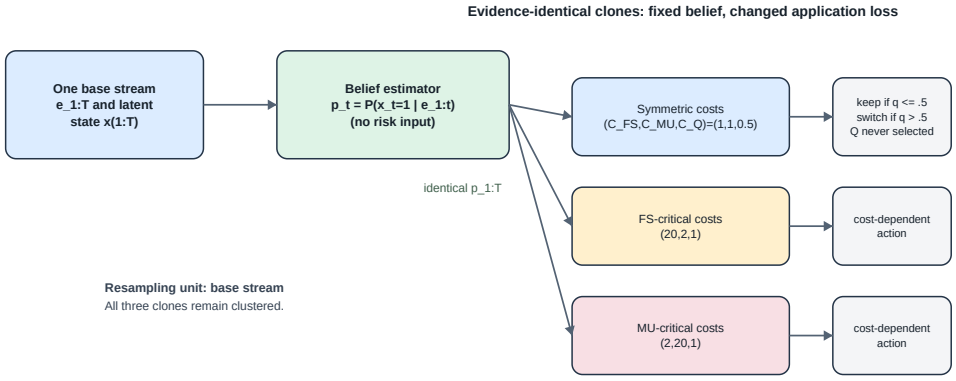


Figure 1: Overview of the NoisyUpdateBench intervention. A single observation stream produces one evidence-derived belief trajectory. Three application cost profiles preserve that trajectory while allowing the revision policy to choose different switch, keep, or quarantine actions. The base stream, rather than a risk clone or event, is the independent statistical unit.

4.2 Static and dynamic Bayesian filtering

The Bayesian filter (BF) assumes a symmetric binary flip hazard h. If the logical-time gap is Δ, its effective transition probability is $a_\Delta = [1-(1-2h)^\Delta]/2$, giving

$$p_i^- = \tfrac{1}{2} + (p_{i-1} - \tfrac{1}{2})(1-2h)^\Delta, \quad \operatorname{logit}(p_i) = \operatorname{logit}(p_i^-) + (2y_i - 1)\log\frac{\rho_i}{1-\rho_i}. \tag{5}$$

The prior is $1/2$; no absolute-time transition is applied before the first event. Static BF sets $h = 0$, while dynamic BF uses the preregistered $h = 0.05$. Both assume conditionally independent observations with supplied symmetric reliabilities. Crucially, the estimator never receives risk costs: those enter only at the decision layer.

4.3 Fixed and expected-cost policies

We combine WV and BF beliefs with either a fixed or expected-cost policy. The fixed policy keeps when $q_i \le 1/3$, quarantines when $1/3 < q_i < 2/3$, and switches when $q_i \ge 2/3$. The risk-sensitive policy uses equation 2; a no-Q ablation removes quarantine and minimizes only switch versus keep risk. Direct observations equal to the incumbent retain the active version; otherwise the selected action updates the version timeline and candidate archive.

We use *RA-SBF* as shorthand for the dynamic BF paired with the expected-cost three-action policy. It is a transparent reference configuration for testing the benchmark's decomposition. Factorial comparisons among estimators, policies, and the no-Q ablation identify which component changes each outcome; Track B separately evaluates the interface with extraction, retrieval, and generation.

5 Evaluation Protocol

We froze method configurations, scenario/query generators, costs, source reliabilities, metrics, and primary comparisons before evaluating each held-out suite. The canonical evaluation compares ten configurations spanning LWW, WV, static BF, dynamic BF, fixed policies, expected-cost policies, and the no-Q ablation. Track A compares six structured-input methods on the independently authored suite. Track B evaluates five raw-language systems: automatic extraction plus RA-SBF, the same extractor plus naive overwrite, full-context prompting, one frozen open-source Mem0 configuration, and a gold-extraction RA-SBF reference. Track B's exact extraction score requires simultaneous agreement on value, source class, and logical time. Track B assigns the numeric loss vector to RA-SBF while preserving full-context, overwrite, and Mem0 in their frozen native forms. It therefore estimates the incremental effect of adding benchmark-aligned risk control to native pipelines. A shared-cost development study separately evaluates equal-information decisions (Appendix K).

Table 1: Evaluation layers and denominator conventions. Track A and Track B use the same independently authored synthetic cases and are complementary views rather than independent dataset replications. Absolute NRC values are not compared across denominator conventions.

Layer	Independent bases	Cloned streams	Input	Compared units	NRC denominator
Canonical	120	360	Structured events	10 configurations	Includes initialization
Track A	60	180	Gold structured events	6 methods	Includes initialization
Track B	60	180	Raw-language sessions	5 systems	Post-initialization

All paired intervals resample base IDs while keeping the three risk clones nested within a base. Canonical and Track A values are reported as event-normalized NRC and include initialization in the decision count; Track B uses post-initialization NRC. Specifically,

$$\mathrm{NRC} = \frac{N_{\mathrm{FS}} C_{\mathrm{FS}} + N_{\mathrm{MU}} C_{\mathrm{MU}} + N_{\mathrm{Q}} C_{\mathrm{Q}}}{N_{\mathrm{decisions}}}. \tag{6}$$

NRC is a decision-normalized cost and need not lie in $[0, 1]$; it is not accuracy. The canonical/Track A convention includes one zero-cost initialization per stream and therefore dilutes absolute NRC, but all methods within a track share the denominator; paired signs and rankings are unchanged. Appendix D reports exact post-initialization point sensitivities for the structured tracks and confirms this algebraic invariance. NRC scores revision actions only and does not include current- or historical- query errors. We therefore co-report FS, MU, and Q rates, coverage (non-quarantined decisions divided by all decisions), selective and unconditional accuracy, and historical accuracy. Brier score and clipped negative log-likelihood are secondary belief-quality metrics. Action cost and query accuracy are complementary estimands: collapsing them into one utility would require an additional application-specific weight that this diagnostic does not assume. We therefore expose their joint operating points directly.

For every paired contrast we compute $\Delta\mathrm{NRC} = \mathrm{NRC}(A) - \mathrm{NRC}(B)$, so negative values favor A. Ninety-five percent percentile intervals use 5,000 paired bootstrap resamples of base IDs; each sampled base carries all of its risk clones and the same resampling indices are used for both methods. We report unadjusted compatibility intervals for individually interpreted contrasts. Precommitted pooled contrasts constitute the primary estimands: 12 in the canonical family, with external overall contrasts frozen in their own protocol. Profile- and scenario-specific intervals characterize heterogeneity. An interval crossing zero is classified as unresolved rather than equivalent.

Track B was executed under a preregistered, independently witnessed operational-recovery amendment that preserved every frozen scientific component. Appendix H records the execution sequence and integrity checks.

Five separately versioned development studies subsequently mapped cost matching, cost sensitivity, model misspecification, duration-aware utility, and cross-model extraction. They use development inputs and are analyzed separately from the held-out estimates (Appendix K).

6 Results

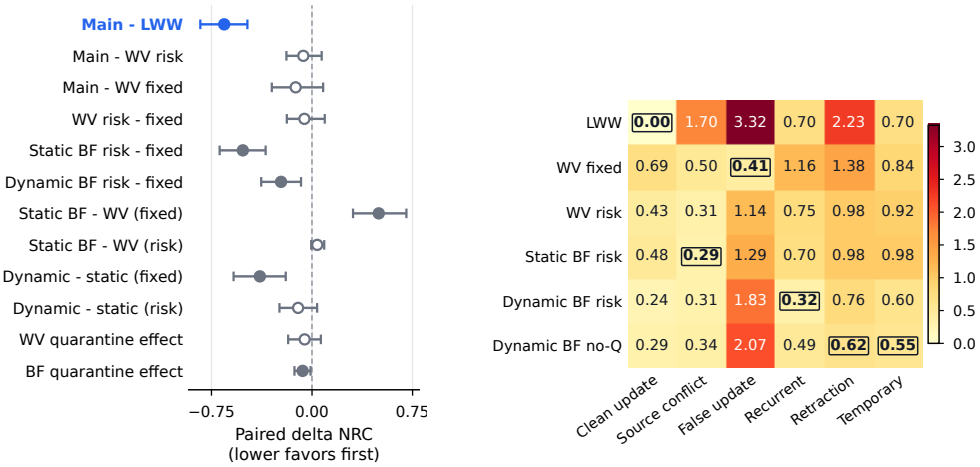
We report risk, coverage, and answer quality jointly. All confidence intervals are paired percentile bootstrap intervals with the base stream as the resampling unit and the three risk clones nested within each base. A negative $\Delta\mathrm{NRC}$ favors the first-named method. Canonical and Track A results use event-normalized NRC, which includes initialization in the decision count. Track B uses post-initialization NRC. Absolute NRC values are therefore interpreted only within a track; cross-track synthesis concerns directions and mechanisms rather than absolute magnitudes.

6.1 Canonical evaluation: lowest pooled NRC and heterogeneous regime winners

On the 120-base canonical held-out suite, dynamic Bayesian filtering with the expected-cost policy (the main RA-SBF configuration) obtained the lowest pooled NRC among the ten evaluated configurations: 0.676, compared with 0.741 for risk-aware Weighted Vote, 0.746 for the matched

Table 2: Selected canonical held-out metrics. NRC, directional error rates, coverage, and selective accuracy are reported together. The full ten-configuration table, historical metrics, and probabilistic scores appear in Appendix E.

Controller	NRC	FS rate	MU rate	Q rate	Current coverage	Current sel. acc.
Last-write-wins	1.331	0.124	0.049	0.000	1.000	0.794
Weighted Vote + expected cost	0.741	0.037	0.135	0.191	0.809	0.745
Dynamic BF + expected cost	**0.676**	0.063	0.069	0.167	0.825	0.809
Dynamic BF + expected cost, no Q	0.746	0.074	0.165	0.000	1.000	0.728



(a) Precommitted paired contrasts

(b) Descriptive scenario NRC

Figure 2: Canonical held-out results. (a) Paired differences in event-normalized NRC; negative values favor the first-named configuration. Intervals are 95% percentile intervals from 5,000 paired base-cluster resamples of 120 bases, with risk clones nested and no multiplicity correction. (b) Scenario point estimates show that the lowest-NRC configuration varies; these cells are descriptive.

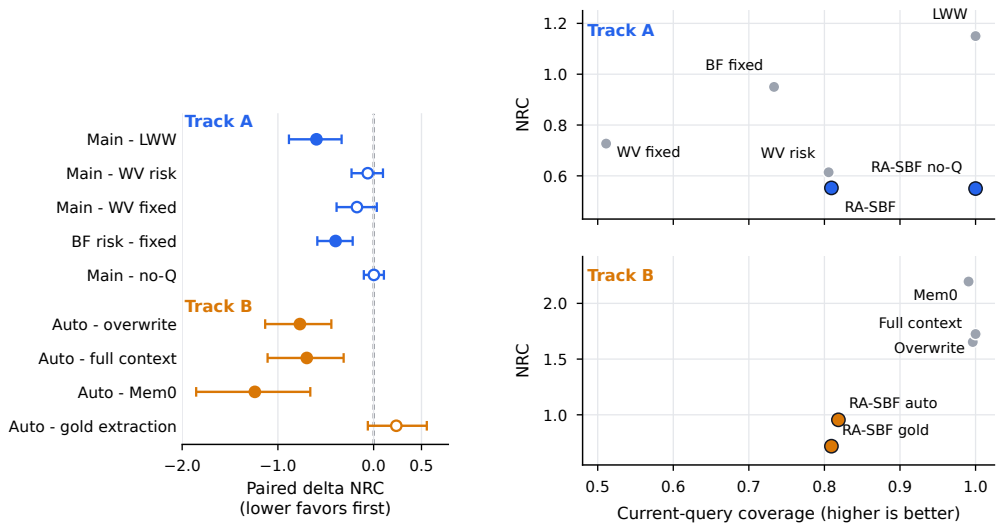
no-quarantine ablation, and 1.331 for last-write-wins (Table 2). In the precommitted paired comparison, the main configuration reduced NRC relative to last-write-wins by 0.655 ($\Delta = -0.655$, 95% CI $[-0.833, -0.482]$; Figure 2).

Risk-aware Weighted Vote was the strongest simple comparator: its contrasts with the main configuration crossed zero for both expected-cost ($\Delta = -0.065$, CI $[-0.191, 0.072]$) and fixed-policy ($\Delta = -0.122$, CI $[-0.300, 0.083]$) variants. Its cumulative evidence mass is particularly effective when isolated false observations should not displace a stable incumbent. These results establish WV as a substantive operating point in the benchmark's risk landscape.

Matched ablations more clearly isolate the decision-policy contribution. Expected-cost decisions reduced NRC relative to the corresponding fixed policy for both static Bayesian filtering ($\Delta = -0.516$, CI $[-0.689, -0.346]$) and dynamic Bayesian filtering ($\Delta = -0.231$, CI $[-0.379, -0.082]$). The analogous Weighted Vote contrast was inconclusive ($\Delta = -0.057$, CI $[-0.188, 0.095]$). Dynamic prediction also helped the fixed-policy Bayesian controller ($\Delta = -0.389$, CI $[-0.585, -0.196]$), whereas the dynamic-versus-static contrast under expected-cost decisions was smaller and inconclusive ($\Delta = -0.104$, CI $[-0.243, 0.038]$). Together, these matched ablations locate the observed gain in the Bayesian decision layer: explicit costs improved both static and dynamic Bayesian controllers, whereas the analogous WV policy contrast remained unresolved. The resulting effect is controller-specific rather than estimator-agnostic.

6.2 Profiles and scenarios expose different failure regimes

The pooled improvement over last-write-wins was concentrated in the false-supersession-critical profile. Under symmetric costs, the main–LWW contrast was approximately zero ($\Delta = 0.008$,



(a) Paired synthetic-transfer contrasts

(b) Risk–coverage operating points

Figure 3: Independently authored synthetic transfer. (a) Paired NRC differences from 5,000 base-cluster resamples of 60 cases; negative favors the first-named method. Track B measures an incremental cost-aware intervention: only RA-SBF receives numeric costs. (b) Pooled NRC and current-query coverage; compare magnitudes only within track because the NRC denominators differ.

CI $[-0.015, 0.030]$); under false-supersession-critical costs it was large ($\Delta = -1.957$, CI $[-2.440, -1.488]$). All six central comparisons in the missed-update-critical profile included zero. Pooled NRC is thus an average under the benchmark's precommitted equal-profile mixture, not an estimate of the prevalence or utility structure of a deployment population.

Scenario-level point estimates further changed the identity of the best controller (Figure 2). Last-write-wins was best on clean updates, fixed-policy Weighted Vote on false updates, static risk-aware Bayesian filtering on conflicting sources, the main configuration on recurrent state, and the no-quarantine dynamic controller on retraction and temporary change. False update was the main configuration's clearest weakness: its scenario-macro NRC was 1.834, versus 0.415 for fixed-policy Weighted Vote. These descriptive strata complement the paired primary analysis by showing that the aggregate ordering combines qualitatively different revision regimes.

6.3 Risk clones verify loss/input separation

The paired risk clones behaved as intended. Across all six profile-pair comparisons for risk-aware Weighted Vote and dynamic Bayesian filtering, the maximum difference between paired belief trajectories was exactly zero, while action-divergence rates ranged from 0.072 to 0.636. Costs therefore changed the selected action without changing the evidence-derived belief. This prespecified manipulation check passed: belief trajectories were invariant to cost labels, while nonzero action divergence showed that the intervention reached the decision layer. The result validates the intended implementation separation; estimation quality is assessed independently through belief and QA metrics.

Quarantine illustrated why risk and answer quality must be reported together. In the canonical dynamic Bayesian controller, enabling quarantine reduced NRC relative to the matched no-Q ablation ($\Delta = -0.070$, CI $[-0.131, -0.011]$), but current-query coverage fell to 0.825; selective current accuracy was 0.809. Historical retrieval also showed a trade-off rather than a general improvement: unconditional current accuracy was 0.667, below LWW at 0.794, and the main controller's historical accuracy was 0.824, below last-write-wins at 0.969, despite full historical coverage. The versioned interface supports historical queries, but the accepted version timeline can lag the oracle when evidence is accumulated or quarantined.

6.4 TRACK A: STRUCTURED INDEPENDENTLY AUTHORED TRANSFER

Track A evaluates the frozen controllers on gold structured events from a sealed, independently authored and independently human-reviewed synthetic suite of 60 base cases (Figure 3). The main configuration again reduced NRC relative to last-write-wins ($\Delta = -0.597$, CI $[-0.885, -0.335]$) and relative to its matched fixed-policy dynamic Bayesian controller ($\Delta = -0.397$, CI $[-0.588, -0.219]$). Belief trajectories remained identical across evidence-matched clones while actions changed with the cost profile, confirming the programmed loss-only intervention on separately authored streams.

Transfer was selective. Comparisons with risk-aware Weighted Vote ($\Delta = -0.062$, CI $[-0.231, 0.098]$) and fixed-policy Weighted Vote ($\Delta = -0.174$, CI $[-0.387, 0.033]$) remained inconclusive. The no-quarantine controller had a slightly lower point estimate than the main configuration (0.550 versus 0.553), and the paired quarantine contrast was approximately zero ($\Delta = 0.003$, CI $[-0.102, 0.107]$). Thus, the canonical aggregate quarantine benefit was absent on Track A. Unconditional current accuracy was 0.619 for the main controller and 0.783 for LWW.

Scenario analysis identifies where transfer reverses. On recurrent-state cases, oracle-aligned observations let LWW track without error while smoothing and quarantine lagged; main–LWW reversed ($\Delta = 0.353$, CI $[0.267, 0.440]$). On false updates, the main controller beat LWW but lost to risk-aware WV ($\Delta = 0.392$, CI $[0.025, 0.875]$) and fixed-policy WV ($\Delta = 0.725$, CI $[0.167, 1.300]$). Policy effects thus transferred more consistently than scenario rankings.

6.5 TRACK B: RAW-LANGUAGE END-TO-END EVALUATION

Track B supplies raw sessions for the same 60 bases. It assigns the numeric cost vector to the RA-SBF decision layer while preserving full context, naive overwrite, and Mem0 in their frozen native forms. The resulting estimand is the incremental effect of adding benchmark-aligned risk control to these pipelines. Under this intervention, RA-SBF reduced post-initialization NRC versus naive overwrite ($\Delta = -0.770$, CI $[-1.133, -0.442]$), full context (-0.698, $[-1.108, -0.313]$), and one frozen Mem0 configuration (-1.241, $[-1.854, -0.662]$).

The resulting operating point prioritized false-supersession control over coverage and QA accuracy (Figure 3). RA-SBF auto reduced false supersession to .047 (naive .145; full context .130) and quarantined 17.8% of decisions. Its coverage/selective/unconditional current accuracy was $.819/.735/.602$, versus $.996/.762/.759$ for full context and $1.000/.750/.750$ for naive overwrite.

The extractor reached .962 exact-match F1; the paired auto–gold NRC difference remained unresolved ($\Delta = .237$, CI $[-.061, .556]$). Eleven strict-parsing failures were retained as abstentions.

Development sensitivity analyses. Separately versioned development analyses probe candidate mechanisms and operating boundaries suggested by the held-out findings (Appendix K). With a shared cost wrapper, RA-SBF and full-context beliefs had an unresolved mean contrast ($\Delta = -.154$, CI $[-.561, .099]$) and occupied nearly equal regions of a 741-point cost simplex. Additional studies characterize sensitivity to calibration, correlated evidence, quarantine duration, and 4B extraction. We report these development estimates separately from the frozen held-out results.

Synthesis. NoisyUpdateBench reveals a structured risk–coverage landscape: costs alter actions without contaminating evidence-derived beliefs, and the lowest-action-loss controller varies with costs and stream structure. System selection is an operating-point choice that should be supported by directional errors, coverage, answer quality, and NRC jointly.

7 DISCUSSION, LIMITATIONS, AND CONCLUSION

Findings and scope. NoisyUpdateBench separates belief, action, and quarantine, showing how costs and stream structure change the lowest-action-loss controller. Expected-cost Bayesian policies help mainly under false-supersession-critical costs; competing methods expose complementary regimes. Version 1 covers binary revision in short synthetic streams, not long-history or deployment-utility performance. Track B uses one Qwen3-8B runtime and Mem0 configuration; shared-base intervals measure case variation, not independent replication.

AI USE STATEMENT

Generative AI tools, including OpenAI ChatGPT and Codex, were used for brainstorming and literature discovery; feedback on methodology and experiment design; code drafting, debugging, and experiment-execution support; synthetic benchmark drafting and consistency checking; result organization and interpretation; figure preparation; and drafting and editing manuscript text. They also assisted the independent dataset author and validator with checklists and automated consistency checks. AI systems did not make final decisions about ground-truth labels, protocol freezes, dataset acceptance, or scientific claims. The human authors and independent data contributors specified and approved the task, costs, protocols, datasets, and evaluations, and verified the implementations, equations, citations, labels, numerical results, and claims against the released artifacts. All AI-assisted material was reviewed by humans, and the authors take responsibility for the submission.

ETHICS STATEMENT

The released benchmark cases are synthetic and contain no personal records. Its explicit cost profiles are diagnostic conditions rather than universal ethical values or deployment recommendations. In particular, quarantine can transfer burden to users or human reviewers through delay and escalation, and the benchmark does not certify safety in medical, legal, financial, or other high-stakes settings. Future conversion of domain logs requires separate human authorization, privacy review, provenance checks, temporal-leakage auditing, and a new preregistered evaluation version.

REPRODUCIBILITY STATEMENT

The accompanying anonymous supplementary artifact includes the frozen canonical and external benchmark suites, manifests and SHA-256 digests, controller and evaluator implementations, exact configurations, result files, table and figure builders, base-cluster bootstrap code, and the versioned extension toolkit. Appendix D documents the statistical protocol, Appendices E–G report the complete results and runtime details, and Appendix H records the external seal and operational-recovery timeline. Appendix K reports the separately versioned development sensitivity analyses. The held-out methods are not rerun or tuned after release; future corrections or extensions require a new version and fingerprint.

REFERENCES

Hengyu An, Minxi Li, Naen Xu, Chunyi Zhou, Xiaogang Xu, Tianyu Du, Jinbao Li, and Shouling Ji. PerMemSafe: Benchmarking implicit personalized safety of long horizon self-evolving agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 6415–6433. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.320. URL https://aclanthology.org/2026.findings-acl.320/.

Pratyay Banerjee, Masud Moshtaghi, Shivashankar Subramanian, Amita Misra, and Ankit Chadha. APEX-MEM: Agentic semi-structured memory with temporal reasoning for long-term conversational AI. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 16470–16489. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.749. URL https://aclanthology.org/2026.acl-long.749/.

Hanxiang Chao, Yihan Bai, Rui Sheng, Tianle Li, and Yushi Sun. STALE: Can LLM agents know when their memories are no longer valid? *arXiv preprint arXiv:2605.06527*, 2026. doi: 10.48550/arXiv.2605.06527. URL https://arxiv.org/abs/2605.06527.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. In *ECAI 2025*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pp. 2993–3000. IOS Press, 2025. doi: 10.3233/FAIA251160. URL https://doi.org/10.3233/FAIA251160.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406. URL https://doi.org/10.1109/TIT.1970.1054406.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL https://www.jmlr.org/papers/v11/el-yaniv10a.html.

Xinyi Fan, Miri Liu, Ruozhen Yang, Siru Ouyang, and Jiawei Han. Can agent memory systems track evolving state? *arXiv preprint arXiv:2608.19652*, 2026. doi: 10.48550/arXiv.2608.19652. URL https://arxiv.org/abs/2608.19652.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=DT7JyQC3MR.

Jianhua Jiang, Dongbo Yuan, and Weihua Li. MemRiskBench: Trace-aware risk-preserving evaluation for long-horizon LLM agents. *arXiv preprint arXiv:2609.14976*, 2026. doi: 10.48550/arXiv.2609.14976. URL https://arxiv.org/abs/2609.14976.

Xianxuan Long, Zhikai Chen, Shenglai Zeng, Shouren Wang, Kai Guo, and Jiliang Tang. MemTrace: Probing what final accuracy misses in long-term memory. *arXiv preprint arXiv:2606.17328*, 2026. doi: 10.48550/arXiv.2606.17328. URL https://arxiv.org/abs/2606.17328.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023. doi: 10.48550/arXiv.2310.08560. URL https://arxiv.org/abs/2310.08560.

Vedant Patel. Supersede: Diagnosing and training the memory-update gap in LLM agents. *arXiv preprint arXiv:2606.27472*, 2026. doi: 10.48550/arXiv.2606.27472. URL https://arxiv.org/abs/2606.27472.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*, 2025. doi: 10.48550/arXiv.2501.13956. URL https://arxiv.org/abs/2501.13956.

Ayan Roy and Kaustuvi Basu. MemSentry: A framework for detecting persistent memory poisoning in agentic AI. *arXiv preprint arXiv:2609.08747*, 2026. doi: 10.48550/arXiv.2609.08747. URL https://arxiv.org/abs/2609.08747.

Zitong Shi, Yixuan Tang, and Anthony Kum Hoe Tung. A-TMA: Decoupling state-aware memory failures in long-term agent memory. *arXiv preprint arXiv:2607.01935*, 2026. doi: 10.48550/arXiv.2607.01935. URL https://arxiv.org/abs/2607.01935.

Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. MemBench: Towards more comprehensive evaluation on the memory of LLM-based agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19336–19352. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.989. URL https://aclanthology.org/2025.findings-acl.989/.

Zhen Tao, Jinxiang Zhao, Peng Liu, Dinghao Xi, Yanfang Chen, Wei Xu, and Zhiyu Li. MemConflict: Evaluating long-term memory systems under memory conflicts. *arXiv preprint arXiv:2605.20926*, 2026. doi: 10.48550/arXiv.2605.20926. URL https://arxiv.org/abs/2605.20926.

Mohammad Tavakoli, Alireza Salemi, Carrie Ye, Mohamed Abdalla, Hamed Zamani, and J. Ross Mitchell. Beyond a million tokens: Benchmarking and enhancing long-term memory in LLMs. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=y59hf5lrMn.

Md Nayem Uddin, Kumar Shubham, Eduardo Blanco, Chitta Baral, and Gengyu Wang. From recall to forgetting: Benchmarking long-term memory for personalized agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26814–26841. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1337. URL https://aclanthology.org/2026.findings-acl.1337/.

Arseniy Varlamov, Rishat Zinnatullin, Elisei Rykov, Alexander Panchenko, and Ilseyar Alimova. MemToC: Benchmarking memory-tool conflict resolution in large language models. *arXiv preprint arXiv:2608.26295*, 2026. URL https://arxiv.org/abs/2608.26295.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=pZiyCaVuti.

Han Zhang, Zihao Tang, Xin Yu, Xiao Liu, Yeyun Gong, Haizhen Huang, Yan Lu, Weiwei Deng, Feng Sun, Qi Zhang, and Hanfang Yang. Beyond static dialogues: Benchmarking realistic, heterogeneous, and evolving long-term memory. *arXiv preprint arXiv:2605.31086*, 2026. URL https://arxiv.org/abs/2605.31086v2. Accepted to Findings of EMNLP 2026.

Yan Zhou, Yue Ouyang, Kaiyang Zheng, and Suncheng Xiang. TEPA: Revoking stale memories for conflict-robust language agents. *arXiv preprint arXiv:2608.07429*, 2026. doi: 10.48550/arXiv.2608.07429. URL https://arxiv.org/abs/2608.07429.

A Formal Task, Time Semantics, and Cost Matrix

A.1 Online revision task

For one fact key, let $x(\tau) \in \{0, 1\}$ denote the evaluator-only latent state at logical time τ. The controller receives a strictly time-ordered event sequence indexed by i,

$$e_i = (\tau_i, y_i, s_i, \rho_i), \qquad y_i \in \{0, 1\}, \quad \rho_i \in (0, 1), \tag{7}$$

where s_i is a source class and ρ_i is an application-supplied reliability score. The reference Bayesian controller interprets ρ_i as $\Pr(y_i = x(\tau_i))$; this is an uncalibrated modeling assumption. The public event contains neither $x(\tau_i)$ nor scenario, expected answer, future-event, or provenance labels. Reliabilities below 0.5 are canonically represented by reversing the observation, $(y_i, \rho_i) \mapsto (1 - y_i, 1 - \rho_i)$, before controller input. This makes the released benchmark's reliability field evidence strength rather than an implicit adversarial-source flag.

After initialization, a conflicting candidate can be SWITCHed into the active memory, the incumbent can be KEEPed, or the candidate can be QUARANTINEd. Initialization is a separate action: an initialization mistake is recorded, but is not charged as false supersession (FS) or missed update (MU) in the frozen v1 risk numerator. For a post-initialization decision, FS occurs when a switch displaces an incumbent that agrees with the current oracle; MU occurs when KEEP leaves the active value inconsistent with the oracle. Every post-initialization observation is eligible: a same-valued observation is coded KEEP and is MU when the persistently wrong incumbent remains wrong. Every quarantine contributes one Q event.

A.2 Logical time and versioned retrieval

Time is discrete and need not be contiguous. Events must have strictly increasing integer steps. If an event and query share a step, the event is processed first; queries do not themselves advance time. A stored version is valid on the half-open interval $[\tau_{\text{from}}, \tau_{\text{to}})$, so a query at the closing boundary is answered by the next version. Current queries read the active state. Historical queries search active and archived *versions*; the reference controllers do not replay raw events at query time. A target before the first observation is answered by abstention.

A.3 EXPECTED-RISK ACTION

Let $q_i = \Pr(x(\tau_i) = y_i \mid e_{1:i})$ be the estimator's posterior probability that the incoming candidate is correct. With application costs $(C^{\mathrm{FS}}, C^{\mathrm{MU}}, C^{\mathrm{Q}})$, the plug-in risks are

$$R_{\mathrm{switch}}(q_i) = (1 - q_i)C^{\mathrm{FS}}, \tag{8}$$

$$R_{\mathrm{keep}}(q_i) = q_i C^{\mathrm{MU}}, \tag{9}$$

$$R_{\mathrm{Q}}(q_i) = C^{\mathrm{Q}}. \tag{10}$$

The policy chooses the minimum and resolves numerical ties in the fixed order KEEP>QUARANTINE>SWITCH. This convention favors the more reversible action and makes boundary behavior deterministic; it is a design choice, not a claim that these actions have a universal ethical order. The policy is optimal only for this one-step loss; delay, repeated unanswered queries, and review workload are not modeled. The three frozen profiles and their induced decision regions are shown in Table 3. They are author-chosen diagnostic stress conditions, not stakeholder-elicited deployment utilities, and are not learned from held-out outcomes.

Table 3: Frozen cost profiles and exact action regions. At equality the stated tie order applies.

Profile	C^{FS}	C^{MU}	C^{Q}	Keep	Quarantine	Switch
Symmetric	1	1	0.5	$q \leq .50$	never selected	$q > .50$
FS-critical	20	2	1	$q \leq .50$	$.50 < q \leq .95$	$q > .95$
MU-critical	2	20	1	$q \leq .05$	$.05 < q \leq .50$	$q > .50$

Evidence-identical risk clones copy the oracle, events, timestamps, source classes, reliabilities, and queries and change only the cost profile and derived identifiers. Therefore a deterministic estimator that keeps risk out of belief formation must satisfy $p_i^{(r)} = p_i^{(r')}$ for every profile pair, while its actions may differ. This is a controlled intervention on the loss, not three independent samples.

A.4 METRICS AND DENOMINATORS

For additive counts within one compatible reporting protocol,

$$\mathrm{NRC} = \frac{N_{\mathrm{FS}}C^{\mathrm{FS}} + N_{\mathrm{MU}}C^{\mathrm{MU}} + N_{\mathrm{Q}}C^{\mathrm{Q}}}{N_{\mathrm{decisions}}}. \tag{11}$$

Counts and decision denominators are summed before division. NRC is a cost per eligible decision, can exceed one, and is not an accuracy. Canonical and Track A denominators include initialization; Track B excludes initialization, although a pre-initialization quarantine would remain eligible. Absolute NRC must consequently not be compared or pooled across these protocols. The zero-cost initialization dilutes canonical/Track A absolute values, but the shared within-track denominator leaves paired rankings and signs unchanged. NRC does not include current- or historical-query error.

For queries, coverage is $N_{\mathrm{answered}}/N_{\mathrm{queries}}$, selective accuracy is $N_{\mathrm{correct}}/N_{\mathrm{answered}}$, and unconditional accuracy is $N_{\mathrm{correct}}/N_{\mathrm{queries}}$. Selective accuracy is undefined when nothing is answered. Belief quality is reported as Brier score $n^{-1}\sum_i (p_i - x_i)^2$ and negative log likelihood; only the NLL scorer clips probabilities to 10^{-12} for finite logs. The controller beliefs themselves are not clipped for scoring.

B BENCHMARK GENERATION AND VALIDATION

B.1 CANONICAL CONTROLLED SUITE

The canonical held-out suite is generated deterministically from split, scenario, replicate, and base seed. It contains 120 independent bases—20 for each of the six scenarios in Table 4—and three

evidence-identical cost clones per base. Before cloning, the suite has 749 observations and 936 queries (545 current and 391 historical); individual bases contain 3–15 observations and 4–13 queries. Thus each controller processes 360 streams, 2,247 observations, and 2,808 queries.

Generation is hybrid rather than unconstrained simulation. Scenario-defining landmarks and required queries are forced, while background observations are sampled under the frozen configuration. Replicates cover four difficulty layers (easy, medium, hard, and temporal stress), both initial states, two scenario subtypes, two source orders, and three duration buckets. These factors are balanced to expose mechanisms; they are not estimates of real deployment prevalence.

Table 4: The six controlled state-revision scenarios.

Scenario	Controlled phenomenon
Clean update	A genuine persistent change with truthful landmark evidence.
False update	Convincing false observations while the latent state remains stable.
Temporary change	A latent $A \to B \to A$ trajectory.
Retraction	Correct and false retraction subtypes.
Recurrent state	Repeated return to a previously valid state.
Conflicting sources	Ordered source disagreement under stable- and changed-state subtypes.

Queries occupy six roles: initialization, immediate-landmark, delayed-landmark, phase, boundary, and final. Every required state-change landmark receives an immediate current query. Historical targets cover formal version intervals and their boundaries. Oracle timelines are contiguous and half open; expected answers are generated from that oracle and kept outside controller-facing records.

Production validation checks schema and generator versions, scenario and replicate cardinalities, binary/range constraints, strictly increasing event steps, oracle continuity, query correctness, landmark coverage, exact clone identity apart from costs/IDs, and manifest round trips. The frozen held-out suite SHA-256 is `cd431165353be4684c6a3ac648f1564bde0f48e64685e94e6ccc428d8e30de97`; the base seed is 20260829 and the generation-configuration digest is `e1508bc1b5f25479a88fcb9b1c4d64a679e6ef6756a6bc9eb69e0418a697a789`. The deterministic release archive has SHA-256 `a5240f25d35428726e081174741c0fd2ade6a40fb76d9f33e68e254302583ccc`.

B.2 Development separation

The 24-base development diagnostic (seed 7), the 120-base inspected pilot (seed 20260215), and the 120-base held-out final are distinct. Pilot and held-out values are never pooled. The final suite was executed once, frozen, and subsequently used only for read-only result packaging and audit. The three risk clones of one base always remain in the same statistical cluster.

B.3 Independently authored external suite

The external suite was authored after method freeze by a contributor who did not develop Track A/B and was then reviewed by a different human validator. It contains 60 synthetic bases (10 per scenario), 260 raw-language sessions with corresponding gold structured events, and 310 base queries (180 current, 130 historical). Ten application domains contribute six bases each. Three cost clones yield 180 evaluated streams. Track A consumes the gold events; Track B consumes the same cases as raw sessions. They are complementary input views, not independent dataset replications.

The external file passed structural, raw-content, query-diversity, oracle, provenance, and semantic audits before sealing. Exact overlap checks against development data returned zero matches for contributor IDs, event-sequence IDs, fact keys, identifiers, query text, and raw-session text. The public suite digest is `c7d1f80c4ce061cae6d6aae0f5a21570340f1dbedcda14a71c89f3f70c6c45f2`; the seal-manifest digest is `27005b292c4203ee0cf3d950da7f275b70aeb19a05a10076b83e60b548cc9ea1`. "External" denotes external authorship and a sealed held-out process; the cases remain synthetic and are not real-world events.

B.4 SOURCE RELIABILITY IN THE EXTERNAL SUITE

The frozen source-class mapping is: official release 0.97, official documentation 0.93, maintainer statement 0.85, verified user 0.75, ordinary issue 0.65, rumor 0.55. Unknown classes are rejected. In Track B, extraction confidence is not substituted for this application reliability: the extractor must identify the source class, after which the frozen mapping is applied. These author-fixed scores are diagnostic conventions, not empirical source calibrations; v1 neither estimates nor perturbs them.

C CONTROLLER EQUATIONS AND HYPERPARAMETERS

C.1 VERSIONED LAST-WRITE-WINS

LWW installs every new observation as the active value and closes the previous version at the incoming event step. Reliability is exposed as confidence but does not affect the action. Same-valued observations create a fresh version snapshot, preserving the confidence and source evidence current at that time. Historical queries search the half-open active/archive timeline; no event replay is used. LWW ignores the cost profile.

C.2 CUMULATIVE RELIABILITY-WEIGHTED VOTE

For observation (y_i, ρ_i), WV assigns one unit of mass,

$$w_i(1) = \begin{cases} \rho_i & y_i = 1 \\ 1 - \rho_i & y_i = 0, \end{cases} \qquad w_i(0) = \begin{cases} \rho_i & y_i = 0 \\ 1 - \rho_i & y_i = 1. \end{cases} \tag{12}$$

With $S_i(v) = \sum_{j \le i} w_j(v)$, $p_i = S_i(1)/(S_i(0) + S_i(1))$. The equivalent signed margin is $M_i = \sum_{j \le i}(2y_j - 1)(2\rho_j - 1)$. Plain WV chooses the sign of this margin; ties retain the incumbent, or use the first observation when no incumbent exists. It has no transition model, temporal decay, window, learned parameter, or source-correlation correction. Policy WV uses the same belief and passes candidate probability $q_i = p_i$ when $y_i = 1$ and $q_i = 1 - p_i$ when $y_i = 0$ to either the fixed or expected-risk policy.

C.3 BINARY BAYESIAN FILTER

The filter maintains $p_i = \Pr(x(\tau_i) = 1 \mid e_{1:i})$. With per-step symmetric hazard h and gap $\Delta_i = \tau_i - \tau_{i-1}$, the probability of an odd number of flips is

$$a_{\Delta_i} = \frac{1 - (1 - 2h)^{\Delta_i}}{2}, \qquad p_i^- = \tfrac{1}{2} + (p_{i-1} - \tfrac{1}{2})(1 - 2h)^{\Delta_i}. \tag{13}$$

The observation update is implemented in uncapped log odds,

$$\ell_i = \operatorname{logit}(p_i^-) + (2y_i - 1) \log \frac{\rho_i}{1 - \rho_i}, \qquad p_i = \sigma(\ell_i). \tag{14}$$

The implementation uses stable log-add-exp transitions and converts to probability only for reporting. The first observation receives the initial prior without a transition from absolute time, ensuring time-translation invariance. Static BF sets $h = 0$; dynamic BF uses the preregistered $h = 0.05$. Both use prior $p_0 = 0.5$. Candidate belief is again p_i for candidate 1 and $1 - p_i$ for candidate 0.

C.4 DECISION LAYERS AND CONFIGURATIONS

The fixed three-way policy uses $q \le 1/3$ for KEEP, $1/3 < q < 2/3$ for QUARANTINE, and $q \ge 2/3$ for SWITCH; it ignores the application cost profile. The expected-risk policy uses the equations in Appendix A. A no-Q ablation minimizes only switch and keep risk. All floating comparisons use relative tolerance 10^{-12} and absolute tolerance 10^{-15}.

The label RA-SBF is shorthand for dynamic BF plus the expected-risk policy and candidate quarantine. It is a transparent reference configuration, not a claim of a new Bayesian filtering theorem. Table 5 lists every canonical configuration.

Table 5: Canonical configurations and the components varied.

Configuration	Estimator	Policy	Quarantine
LWW	latest observation	overwrite	no
Plain WV	cumulative WV	vote winner	no
WV + fixed	cumulative WV	$1/3, 2/3$	yes
WV + expected risk	cumulative WV	application costs	yes
WV + expected risk, no-Q	cumulative WV	application costs	no
Static BF + fixed	BF, $h = 0$	$1/3, 2/3$	yes
Static BF + expected risk	BF, $h = 0$	application costs	yes
Dynamic BF + fixed	BF, $h = .05$	$1/3, 2/3$	yes
Dynamic BF + expected risk	BF, $h = .05$	application costs	yes
Dynamic BF + expected risk, no-Q	BF, $h = .05$	application costs	no

C.5 Candidate and archive lifecycle

Accepted values produce formal active versions and close prior versions into the archive. Quarantined candidates are tracked separately from the active timeline. On a confirming observation the active version is retained and any pending conflicting candidate is rejected. On a conflict, SWITCH accepts that value and closes the incumbent; KEEP rejects that candidate; and QUARANTINE creates a pending candidate, refreshes the same pending value, or marks a different pending value superseded before creating its replacement. A later switch to a pending value records acceptance and resolution delay. The implementation records creation, acceptance, rejection, pending status, resolution delay, and archive size. Optional expiration and reactivation categories exist, but neither they nor lifecycle supersession were activated in canonical held-out results; their counts are zero (Appendix E). We therefore make no empirical claim about a complete retention/deletion state machine.

D Statistical Protocol

D.1 Estimand and resampling

For each contrast, the point estimand is the difference between pooled cost per decision for configurations A and B under the benchmark's precommitted equal-profile/equal-scenario sample, $\Delta = \mathrm{NRC}(A) - \mathrm{NRC}(B)$. Negative values favor A. A bootstrap replicate samples base IDs with replacement, brings all three risk clones of each sampled base together, and uses identical sampled IDs for both methods. NRC is recomputed from pooled additive costs and denominators within the replicate. We report the 2.5th and 97.5th percentiles from 5,000 replicates. We use an unstratified percentile cluster bootstrap without BCa correction or multiplicity adjustment. The resulting intervals quantify empirical-base variation while allowing scenario composition to vary; we interpret them individually as unadjusted compatibility intervals. An interval containing zero is classified as unresolved.

The independent unit is the base stream: events, queries, and risk clones are nested observations. Canonical resampling uses $n = 120$ bases; each external track uses $n = 60$. Track A and Track B reuse the same external bases, so their intervals must not be interpreted as two independent replications. The external bootstrap seed is 20260904.

D.2 Analysis hierarchy

The canonical 12 pooled contrasts in Table 7 were precommitted. Six central contrasts stratified by the three risk profiles are secondary heterogeneity analyses. Precommitted pooled contrasts form the analysis hierarchy; scenario, query-role, lifecycle, robustness, and pilot-transfer summaries characterize heterogeneity. The external protocol precommitted its own overall paired contrasts; scenario and profile breakdowns remain descriptive.

D.3 WHY POOLED COUNTS ARE REQUIRED

Averaging per-stream ratios gives a short stream the same weight as a long stream and can disagree with event-level cost. We therefore sum FS, MU, Q, cost, and eligible-decision counts before division. The P4 exporter makes these additive fields and the denominator protocol mandatory. The legacy canonical CSV retained ratios rather than all counts; a read-only audit recovered the aggregate counts without rerunning controllers. Complete per-method event traces are unavailable for that legacy run; P5-compliant extensions preserve traces and additive counts as standard outputs.

D.4 INITIALIZATION-DENOMINATOR SENSITIVITY

The frozen canonical and Track A protocols include one zero-cost initialization per clone-stream, whereas Track B begins after initialization. A read-only arithmetic sensitivity removes those initializations from the two structured tracks without changing any prediction or cost numerator. Canonical decisions change from 2,247 to 1,887, rescaling every canonical NRC by $2247/1887 = 1.191$: the main controller changes from $.676$ to $.804$, LWW from 1.331 to 1.585, and their point difference from $-.655$ to $-.780$. Track A decisions change from 780 to 600: the main controller changes from $.553$ to $.718$, LWW from 1.150 to 1.495, and their point difference from $-.597$ to $-.777$. Because the same positive constant rescales every method within a track, rankings and contrast signs are algebraically invariant. Base-level canonical costs were not retained, so this sensitivity reports exact pooled point estimates rather than reconstructed paired intervals.

D.5 INTERPRETATION OF BELIEF AND EXTRACTION METRICS

Brier and NLL are computed once per base event trajectory and are not treated as tripled independent evidence across the three cost clones. In Track B, exact extraction requires simultaneous equality of value, source class, and logical time. Extraction metrics diagnose the automatic front end; system NRC evaluates the end-to-end decision path. A high extraction F1 cannot by itself establish equivalence to gold extraction, which is assessed by the paired NRC contrast.

E CANONICAL FULL RESULTS

E.1 ALL CONFIGURATIONS AND PRECOMMITTED CONTRASTS

Table 6 reports all ten frozen configurations. Accuracy is selective accuracy; coverage is reported separately. Historical queries are always answered by the formal active/archive timeline in these structured controllers.

Table 6: Canonical held-out overall results (120 bases, 360 clones).

Method	NRC↓	FS	MU	Q	Cur. cov.	Cur. sel.	Hist. acc.	Brier↓	NLL↓
Dynamic BF + expected risk (main)	.676	.063	.069	.167	.825	.809	.824	.148	.464
WV + expected risk	.741	.037	.135	.191	.809	.745	.728	.186	.552
Dynamic BF + expected risk, no-Q	.746	.074	.165	.000	1.000	.728	.844	.148	.464
Static BF + expected risk	.779	.039	.139	.155	.829	.742	.728	.180	.554
WV + expected risk, no-Q	.795	.060	.231	.000	1.000	.692	.795	.186	.552
WV + fixed policy	.797	.005	.045	.490	.516	.893	.652	.186	.552
Dynamic BF + fixed policy	.906	.044	.047	.252	.703	.843	.729	.148	.464
Static BF + fixed policy	1.296	.028	.116	.228	.725	.767	.660	.180	.554
LWW	1.331	.124	.049	.000	1.000	.794	.969	.154	.490
Plain WV	1.914	.056	.194	.000	1.000	.695	.767	.186	.552

E.2 RISK-PROFILE HETEROGENEITY

Table 8 contains all 18 prespecified profile-stratified comparisons. These intervals are secondary and unadjusted.

The pooled main–LWW effect is therefore driven principally by the FS-critical profile. Symmetric-cost results favor the main configuration over the WV policy variants but not over LWW, and every central MU-critical interval contains zero.

Table 7: All 12 precommitted canonical paired contrasts. $\Delta = A - B$; negative values favor A.

Contrast (A vs. B)	Δ NRC	95% paired CI	Reading
Main vs. LWW	-.655	[-.833, -.482]	A lower
Main vs. risk-aware WV	-.065	[-.191, .072]	inconclusive
Main vs. fixed WV	-.122	[-.300, .083]	inconclusive
WV risk vs. fixed	-.057	[-.188, .095]	inconclusive
Static BF risk vs. fixed	-.516	[-.689, -.346]	A lower
Dynamic BF risk vs. fixed	-.231	[-.379, -.082]	A lower
Static BF vs. WV, fixed	.498	[.306, .704]	B lower
Static BF vs. WV, risk	.039	[-.006, .091]	inconclusive
Dynamic vs. static BF, fixed	-.389	[-.585, -.196]	A lower
Dynamic vs. static BF, risk	-.104	[-.243, .038]	inconclusive
WV quarantine effect	-.055	[-.178, .066]	inconclusive
BF quarantine effect	-.070	[-.131, -.011]	A lower

Table 8: Canonical profile-stratified paired contrasts.

Profile	Contrast	Δ NRC	95% CI	Excludes 0
Symmetric	Main–LWW	.008	[-.015, .030]	no
	Main–risk WV	-.068	[-.100, -.034]	yes
	Main–fixed WV	-.114	[-.156, -.068]	yes
	Main risk–fixed	-.035	[-.060, -.011]	yes
	Main dynamic–static	-.071	[-.102, -.037]	yes
	Main Q effect	.000	[.000, .000]	no
FS-critical	Main–LWW	-1.957	[-2.440, -1.488]	yes
	Main–risk WV	-.059	[-.139, .043]	no
	Main–fixed WV	-.063	[-.164, .036]	no
	Main risk–fixed	-.602	[-.921, -.319]	yes
	Main dynamic–static	-.044	[-.125, .020]	no
	Main Q effect	-.195	[-.312, -.092]	yes
MU-critical	Main–LWW	-.016	[-.176, .139]	no
	Main–risk WV	-.068	[-.388, .275]	no
	Main–fixed WV	-.188	[-.689, .364]	no
	Main risk–fixed	-.055	[-.418, .357]	no
	Main dynamic–static	-.196	[-.560, .171]	no
	Main Q effect	-.016	[-.176, .139]	no

E.3 SCENARIO MACRO RESULTS

Table 9 reports descriptive macro NRC for the six most interpretable methods. Bold values mark the lowest point estimate in each scenario; no scenario-level multiplicity correction is applied.

Table 9: Canonical scenario-macro NRC.

Scenario	LWW	WV fixed	WV risk	Static BF risk	Main	Main no-Q
Clean update	**.000**	.690	.429	.484	.238	.286
Conflicting sources	1.704	.502	.313	**.290**	.306	.343
False update	3.324	**.415**	1.140	1.291	1.834	2.068
Recurrent state	.697	1.162	.746	.702	**.322**	.487
Retraction	2.232	1.384	.979	.979	.755	**.616**
Temporary change	.697	.836	.924	.982	.600	**.552**

These cells resolve the pooled result into distinct revision regimes. LWW reacts quickly to clean evidence; cumulative voting strongly resists the isolated false-update regime; the dynamic model is most useful when state can recur. Those explanations are mechanistic interpretations consistent with the design, not separate causal estimates.

E.4 QUERY-ROLE AND RISK-CLONE DIAGNOSTICS

Table 10 gives the complete query-role breakdown for the main configuration. The role-level table for all ten configurations is retained in the release as `query_role_metrics.csv`. Historical coverage is one, whereas current coverage falls at immediate landmarks under quarantine. Final-current selective accuracy is higher than its unconditional accuracy because 57 of 360 final queries are not answered.

Table 10: Canonical main-configuration query roles.

Role	Type	Total	Ans.	Correct	Cov.	Sel. acc.	Uncond. acc.
Boundary	historical	825	825	618	1.000	.749	.749
Final	current	360	303	257	.842	.848	.714
Initialization	current	360	360	360	1.000	1.000	1.000
Landmark delayed	current	711	567	413	.797	.728	.581
Landmark immediate	current	759	582	396	.767	.680	.522
Phase	historical	780	780	633	1.000	.812	.812

Table 11: Risk-clone invariance and action divergence.

Estimator	Profile pair	Max belief diff.	Action divergence
WV risk	Sym.–FS	0	.417
WV risk	Sym.–MU	0	.215
WV risk	FS–MU	0	.632
Dynamic BF risk	Sym.–FS	0	.563
Dynamic BF risk	Sym.–MU	0	.072
Dynamic BF risk	FS–MU	0	.636

E.5 LIFECYCLE, ROBUSTNESS, AND PILOT CHECKS

Table 12 reports the lifecycle counters. No-Q methods do not instantiate candidates. The zero expiration, supersession, and reactivation columns are evidence that those lifecycle paths were not exercised, not evidence that the paths are unnecessary.

Across the 12 primary pilot and held-out effects, Pearson correlation was .975 and 11 of 12 effect directions agreed. Ten pilot intervals but only six held-out intervals excluded zero; this check supports

Table 12: Canonical candidate-lifecycle diagnostics.

Config.	Created	Accept	Reject	Expire	Supersede	Reactivate	Pending	Mean delay	Formal archive
WV fixed	510	33	324	0	0	0	153	8.227	786
WV risk	241	51	118	0	0	0	72	6.036	1458
Static BF fixed	429	171	198	0	0	0	60	5.569	1374
Static BF risk	240	77	113	0	0	0	50	6.421	1539
Dynamic BF fixed	462	189	192	0	0	0	81	5.165	1320
Dynamic BF risk	211	44	110	0	0	0	57	5.773	1511
WV risk no-Q	0	0	0	0	0	0	0	–	1887
Dynamic BF risk no-Q	0	0	0	0	0	0	0	–	1887

directional stability, not reuse of the pilot as additional test evidence. Across 13 predefined robustness levels, every main–LWW point estimate was negative, but individual strata were not powered or multiplicity-adjusted for separate claims.

F Independently Authored Track A: Structured Inputs

Track A applies the frozen structured controllers directly to the independent author's gold events. It therefore tests transfer of the belief/action mechanism without natural-language extraction error. Each method has 780 eligible decisions under the all-update-opportunities denominator.

Table 13: Track A overall results (60 bases, 180 clones).

Method	NRC	FS	MU	Q	Cur. cov.	Cur. sel.	Hist. cov.	Hist. sel.	Brier
LWW	1.150	.115	.035	.000	1.000	.783	.923	.850	.132
WV + fixed	.727	.000	.054	.377	.511	.848	.923	.708	.167
WV + expected risk	.614	.010	.162	.147	.806	.701	.923	.714	.167
Dynamic BF + fixed	.950	.042	.062	.185	.733	.795	.923	.717	.134
Dynamic BF + expected risk (main)	.553	.036	.096	.146	.809	.764	.923	.769	.134
Dynamic BF + expected risk, no-Q	**.550**	.056	.147	.000	1.000	.711	.923	.786	.134

The main and no-Q point estimates differ by .003; the paired interval ($[-.102, .107]$) leaves the direction unresolved. We therefore characterize quarantine through its coverage and duration effects rather than a pooled Track A NRC contrast.

Table 14: Track A precommitted overall paired contrasts.

Contrast	Δ NRC	95% CI	Role
Main–LWW	-.597	[-.885, -.335]	primary
Main–WV risk	-.062	[-.231, .098]	primary estimation
Main–WV fixed	-.174	[-.387, .033]	primary estimation
BF risk–fixed	-.397	[-.588, -.219]	primary
Main–no-Q	.003	[-.102, .107]	exploratory

F.1 Profile and scenario diagnostics

Table 15 shows all method-by-profile point estimates. The winner changes in every profile: LWW is lowest under symmetric costs, fixed WV under FS-critical costs, and LWW and no-Q tie under MU-critical costs.

Two preregistered diagnostic strata received paired intervals. On recurrent state, main–LWW was .353 [.267, .440], main–WV risk was $-.187$ [-.273, -.100], and main–WV fixed was $-.420$ [-.757, -.160]. On false update, the corresponding values were -1.708 [-1.792, -1.617], .392 [.025, .875], and .725 [.167, 1.300]. These reversals show that the pooled contrast averages distinct revision regimes.

Table 15: Track A NRC by risk profile. Cells are descriptive point estimates.

Profile	LWW	WV fixed	WV risk	BF fixed	Main	Main no-Q
Symmetric	**.150**	.242	.254	.196	.196	.196
FS-critical	2.377	**.485**	.562	1.154	.492	.531
MU-critical	**.923**	1.454	1.027	1.500	.969	**.923**

Table 16: Track A NRC by scenario, pooled over the three cost profiles.

Scenario	LWW	WV fixed	WV risk	BF fixed	Main	Main no-Q
Clean update	**.000**	.417	.517	.146	.183	.283
Conflicting sources	2.875	2.146	1.608	2.404	**1.333**	1.342
False update	2.683	**.250**	.583	1.700	.975	.925
Recurrent state	**.000**	.773	.540	.490	.353	.313
Retraction	1.917	.188	**.083**	.679	.183	.200
Temporary change	**.000**	.607	.420	.507	.380	.347

F.2 Clone mechanism and transfer

For WV, action-divergence rates for symmetric–FS, symmetric–MU, and FS–MU clones were .185, .304, and .488; for dynamic BF they were .369, .169, and .538. Maximum paired belief difference was zero in all six checks over 260 events, confirming the intended loss-only intervention on separately authored streams.

Across the 18 scenario-by-profile transfer cells, canonical-to-external sign agreement/Spearman correlation was 12/18 and .375 for main–LWW, 13/18 and .565 for main–risk WV, 12/18 and .484 for main–fixed WV, 13/18 and .633 for BF risk–fixed, and 13/18 and .368 for the BF quarantine effect. These are descriptive transfer diagnostics. They indicate partial structural transfer, not distributional equivalence between the generators.

G Independently Authored Track B: Raw-Language End-to-End Evaluation

G.1 Systems and frozen runtime

All deployable Track B systems receive the same causally available raw sessions, public binary labels, and queries. The RA-SBF decision layer additionally receives the declared numeric loss vector required by the task; full-context, naive overwrite, and Mem0 remain in their frozen native, cost-agnostic forms, with costs used only by the evaluator. No system receives the scenario, oracle, expected answer, provenance, or readable profile name. This design estimates the incremental value of adding benchmark-aligned risk control to frozen native baselines; equal-information action selection is analyzed in Appendix K. The evaluated systems are:

- **Full context**: the frozen language model answers from all raw sessions causally available at the query step.
- **Naive overwrite**: the automatic extractor is shared with RA-SBF auto, but the newest extracted observation overwrites memory; historical queries abstain by design.
- **Mem0 OSS**: Mem0's native inferred-memory workflow with an isolated namespace per run, system, base, and risk clone.
- **RA-SBF gold**: evaluator-owned structured extraction followed by the frozen decision layer. This is an extraction reference, not a deployable system.
- **RA-SBF auto**: frozen automatic value/source/time extraction followed by the same decision layer.

The frozen runtime used Python 3.12.10, Ollama 0.32.15, `qwen3:8b`, `embeddinggemma:300m` (768 dimensions), Mem0 OSS 2.0.20, Qdrant Client 1.19.0, and the Ollama Python package 0.6.2. Temperature was zero, seed 20260904, thinking disabled, maximum output 512 tokens, timeout 120 seconds, and maximum attempts three. Mem0 used top-$k = 20$, threshold .1, native inference enabled, telemetry disabled, and neither BM25 hybrid search nor reranking. State was isolated in 180 namespaces; the contamination count was zero.

The execution host was a Dell 14 Plus DB14250 with an Intel Core Ultra 9 288V (8 cores/8 logical processors) and 32 GB physical memory. Although the host exposed an Intel Arc 140V integrated GPU, Ollama reported CPU execution and no NVIDIA runtime was available. The Qwen model was GGUF Q4_K_M (8.2B) and the embedding model was BF16 (307.58M). The recovery lock records an end-to-end execution window of 50 h 47 min 44 s for 3,680 model calls and 900 system–stream evaluations (72.45 calls/h and 17.72 evaluations/h). This wall time includes orchestration, persistence, Mem0/Ollama overhead, and retry handling; it is not a per-call inference latency benchmark.

The frozen LLM digest is `500a1f067a9f782620b40bee6f7b0c89e17ae61f686b92c24933e4ca4b2b8b41`; the embedding-model digest is `85462619ee721b466c5927d109d4cb765861907d5417b9109caebc4e614679f1`. The complete runtime-configuration digest is `12e57fd8bc42ca7d008e606e4890c767a79dee4523feac3fc18f20d09c7e97eb` and the Mem0-configuration digest is `b1eba04d8172dfec5f1598e213e24ced90eab2fdf0e3d77a166388fdba0f4c29`. The exact system and extraction prompts are retained in the frozen source and bound by the joint-method manifest; they were not edited after sealing.

G.2 Overall risk and paired comparisons

Track B uses 600 post-initialization decision opportunities per system. Table 17 co-reports risk, coverage, and selective accuracy. Naive overwrite's historical coverage is zero by construction, so historical selective accuracy is undefined rather than zero.

Table 17: Track B incremental-intervention results; only RA-SBF receives numeric costs, and NRC excludes query error.

System	NRC	FS	MU	Q	Cur. cov.	Cur. sel.	Hist. cov.	Hist. sel.
Full context	1.653	.130	.088	.000	.996	.762	1.000	.754
Mem0 OSS	2.196	.105	.183	.012	.991	.680	1.000	.664
Naive overwrite	1.725	.145	.080	.000	1.000	.750	.000	–
RA-SBF auto	.955	.047	.148	.178	.819	.735	.923	.742
RA-SBF gold reference	**.718**	.047	.125	.190	.809	.764	.923	.769

Table 18: Track B intervention contrasts; negative favors RA-SBF, the sole cost-conditioned system.

Comparison	Δ NRC	95% CI	Role
Auto–naive overwrite	-.770	[-1.133, -.442]	primary
Auto–full context	-.698	[-1.108, -.313]	primary estimation
Auto–Mem0 OSS	-1.241	[-1.854, -.662]	primary estimation
Auto–gold reference	.237	[-.061, .556]	exploratory extraction gap

The first three intervals exclude zero under the frozen model, prompts, runtime, Mem0 configuration, and synthetic suite. The auto–gold interval spans zero, leaving the extraction contribution unresolved at this sample size.

G.3 Profile and scenario diagnostics

The profile and scenario map localizes the end-to-end effect. RA-SBF's largest NRC reductions occur under FS-critical costs and on conflicting-source, false-update, and retraction streams; full context leads on clean and temporary-change point estimates, while gold extraction leads on recurrent state. Risk, coverage, and QA therefore identify distinct operating properties of the systems.

Table 19: Track B NRC by risk profile.

Profile	Full context	Mem0	Overwrite	Auto	Gold
Symmetric	**.220**	.308	.225	.280	.255
FS-critical	2.780	2.355	3.060	**.640**	**.640**
MU-critical	1.960	3.925	1.890	1.945	**1.260**

Table 20: Track B NRC by scenario. Cells are descriptive.

Scenario	Full context	Mem0	Overwrite	Auto	Gold
Clean update	**.000**	.017	.511	.700	.244
Conflicting sources	3.578	3.044	3.578	**1.778**	**1.778**
False update	3.578	4.361	3.578	**1.300**	**1.300**
Recurrent state	.958	1.204	.575	.992	**.442**
Retraction	2.300	2.833	2.556	**.244**	**.244**
Temporary change	**.217**	2.083	.383	.767	.475

G.4 Extraction and call audit

The extractor emitted 260 predictions for 260 gold events. Event detection, source class, and temporal-boundary precision/recall/F1 were all 1.000. Value and simultaneous exact-match precision/recall/F1 were .962, corresponding to 10 value mismatches. The errors occurred in clean update (2), conflicting sources (3), recurrent state (3), and temporary change (2), with none in false update or retraction. All ten involved comparative, negated, or transition wording that required binding a surface value to the case-specific binary mapping; Table 25 shows a deterministically selected example. This descriptive pattern was not used to alter the extractor.

The run issued 3,680 model calls. Transport completion was 1.000 and strict schema-parse completion was .997. Eleven answer calls exhausted three retries after request-ID mismatches: nine Mem0 and two full-context calls. They were retained as abstentions rather than deleted or rerun. The 260 extraction calls all completed successfully; 33 calls in total required more than one attempt. The preregistered completion gate was at least .95 with exactly zero namespace contamination, and both conditions were met. Raw responses, prompt and request digests, error lists, preserved runtime state, and the full call audit are part of the frozen artifact.

H Seal, Role Separation, and Operational Recovery

H.1 External-test separation and seal

The synthetic external suite was authored after method freeze by a contributor who did not develop Track A/B and was reviewed by a different human validator. The author attested that cases were created without candidate outputs or test-time tuning; the validator attested to role independence, human semantic review, and absence of candidate-output inspection. AI tools assisted with checklists and consistency checks, while the humans retained responsibility. Legal identities and signatures are held privately to preserve anonymity.

The final test file was sealed on 2026-09-15 before execution. The seal and joint-method freeze bind the data, schema, hypotheses, controller/runtime configurations, metrics, and runner. The accessible artifact contains the signed role templates, manifests, detached SHA-256 records, and verification commands. These records establish object identity and execution order; semantic review, contract tests, and the reported analyses address correctness and validity.

H.2 Disclosed launch failure and recovery

The original joint runner completed Track A, then failed Track B's readiness check because it had been launched with the system interpreter rather than the frozen Track B environment. The failure occurred before Track B backend initialization, model calls, Mem0 state, staging data, or output artifacts. Track A results were not inspected and Track A was not rerun.

Before recovery, the methodology signatory and independent witness attested that no method, prompt, model, runtime configuration, evaluator, metric, hypothesis, test item, label, or seal had changed. A versioned operational amendment then preserved Track A unchanged and executed only the previously uninitialized Track B in a fresh namespace under the frozen runtime. Recovery completed on 2026-09-17 UTC. We therefore report Track A unchanged and Track B as a witnessed operational recovery, not as one uninterrupted joint launch. Full timestamps and digests remain in the executable artifact rather than the paper.

I Benchmark Card and Extension Toolkit

I.1 Benchmark card

Intended use. NoisyUpdateBench v1 is a controlled diagnostic for online revision of a versioned binary fact under observation noise, latent change, supplied source reliability, and application-specific asymmetric costs. It supports estimator/policy ablations, FS/MU/Q diagnosis, risk-clone checks, current and historical retrieval, and CPU-friendly feasibility testing before expensive end-to-end experiments.

Out of scope. The benchmark is not a representative deployment sample, a leaderboard for general-purpose language models, a safety certification, or an estimator of real-world scenario prevalence. It does not cover disputed ground truth, multi-valued or compositional facts, correlated sources, learned reliabilities, privacy-sensitive personal memory, or human-review workload. Canonical held-out data must not be used for tuning. The independently authored suite is also synthetic, despite its process separation and domain diversity.

Data and bias. Both suites use complete evaluator-only oracles and deliberately balanced scenario/profile mixtures. Known reliabilities and symmetric binary transitions simplify real uncertainty. Forced scenario landmarks can favor methods whose assumptions resemble the generator; independent authorship, the raw-language track, strong WV baselines, and explicit failure strata reduce but do not eliminate this concern. Quarantine's scalar cost omits delay, reviewer workload, and downstream harm.

Ethics, privacy, and release. The current data are synthetic and contain no personal data. Numeric cost profiles are analytical inputs, not ethical values, and must be chosen with stakeholders before deployment. The benchmark must not be used to claim that a memory system is safe for medical, legal, financial, or other high-stakes use. The canonical v1 snapshot is immutable. Project-authored code is released under Apache License 2.0, while benchmark data and documentation are released under CC BY 4.0. The latter scope includes the independently authored external suite with its author's redistribution authorization. The license scope and standard license texts are distributed with the artifact; third-party dependencies remain under their own terms.

I.2 Extension toolkit

The artifact provides six versioned modules for extending the benchmark. P1 defines structured and raw adapters; P2 supplies deterministic parsing and scoring; P3 generates user-specified risk clones and verifies belief invariance; P4 exports additive FS/MU/Q/NRC and QA statistics with explicit denominator labels; P5 records provenance, licensing, splits, and trace-retention requirements; and P6 provides a privacy-aware, causally ordered schema for future enterprise domain packs. Contract tests use fictional inputs and establish software behavior rather than empirical performance.

We additionally provide a non-empirical integration pack that maps active, archived, and quarantined states to versioned event, vector, and graph storage semantics. Full schemas, manifests, digests,

test counts, and verification commands are included in the anonymous artifact. The toolkit and integration pack extend reproducibility and interoperability while leaving the frozen scientific results unchanged.

J Related-Work Scope Matrix

Tables 21–22 organize the closest literature by experimental object and isolate the joint intervention introduced here. Bibliographic metadata and claim-to-citation alignment for all 33 library entries, including all 24 cited entries, were human-checked.

Table 21: Benchmark scope and the diagnostic distinction studied here.

Work/group	Primary reported scope	Relationship to NoisyUpdateBench
LoCoMo (Maharana et al., 2024)	Long multi-session conversational memory	Broad long-horizon QA; our study isolates cost-sensitive revision.
LongMemEval (Wu et al., 2025)	Extraction, multi-session and temporal reasoning, updates, abstention	Broader system evaluation; our unit is the revision action under asymmetric costs.
MemBench; MemoryAgentBench (Tan et al., 2025; Hu et al., 2026)	Memory levels and incremental agent-memory operations	Broader capability coverage; our unit is a versioned revision decision.
BEAM; RHELM (Tavakoli et al., 2026; Zhang et al., 2026)	Architectural scale and heterogeneous memory tasks	Complementary breadth/realism; our suite is a controlled binary diagnostic.
Memora; STALE (Uddin et al., 2026; Chao et al., 2026)	Forgetting-aware evaluation; implicit stale-state resolution, premise resistance, and policy adaptation	Closely related update failure; our distinct variable is application loss with evidence held fixed.
MemConflict; MemTrace (Tao et al., 2026; Long et al., 2026)	Controlled conflicts; knowledge-point traces over memory age, question type, and evidence condition	Closely related conflict/trace diagnosis; our outcome taxonomy is FS/MU/Q plus paired costs.
A-TMA; StateMemBench; Supersede; TEPA (Shi et al., 2026; Fan et al., 2026; Patel, 2026; Zhou et al., 2026)	Temporal/state-centric updating and evolving memory	Establish state change and supersession; our benchmark adds asymmetric action costs and paired loss interventions.

Table 22: Risk-aware evaluation and memory-system scope.

Work/group	Primary reported scope	Relationship to NoisyUpdateBench
Reject option; selective classification (Chow, 1970; El-Yaniv & Wiener, 2010)	Abstention and risk–coverage trade-offs	Supplies decision-theoretic context; quarantine instantiates selective decisions in versioned revision.
PerMemSafe; MemRiskBench; MemSentry (An et al., 2026; Jiang et al., 2026; Roy & Basu, 2026)	Memory safety, risk taxonomies, or quarantine	Related risk-aware objectives; our diagnostic pairs identical evidence with FS/MU/Q cost interventions.
MemToC (Varlamov et al., 2026)	Post-tool-return arbitration between tool evidence and model parametric memory	Related asymmetric arbitration in parametric memory; our unit is revision of a persistent external store.
MemGPT; Mem0; Zep; APEX-MEM (Packer et al., 2023; Chhikara et al., 2025; Rasmussen et al., 2025; Banerjee et al., 2026)	End-to-end memory architectures combining extraction, storage, retrieval, and answering	Track B connects our decision layer to raw text and evaluates one frozen Mem0 configuration at the revision layer.
NoisyUpdateBench (ours)	Versioned current/historical revision with latent truth, FS/MU/Q costs, and evidence-identical risk clones	Controlled synthetic diagnosis separating belief estimation from cost-sensitive action, with independently authored and raw-text transfer.

Our novelty is conjunctive: versioned noisy streams with latent labels, explicit FS/MU/Q costs, and evidence-identical loss clones that audit belief–action separation. This creates a controlled bridge between temporal memory evaluation, selective decisions, and versioned memory systems.

K Development Sensitivity Analyses

These separately versioned studies were designed after the v1 freeze to map mechanisms suggested by the held-out results. Each package froze a method manifest before execution and used development inputs; the sealed suite remained untouched. We analyze them as sensitivity and mechanism evidence, separately from the v1 held-out estimates.

K.1 MATCHED COSTS AND THE COST SIMPLEX

MCB-1 replayed frozen development beliefs through the same expected-cost action wrapper. Table 23 shows an unresolved pooled contrast between cost-aware full-context and RA-SBF, while RA-SBF had lower NRC than the tested Mem0 belief source. Full-context responses varied on 2/51 state probes and Mem0 responses on 31/51; a frozen positional rule selected the first response, so a future confirmatory study requires runtime replicates and hierarchical uncertainty.

Table 23: Matched-cost development contrasts on 12 bases. Negative $\Delta\mathrm{NRC}$ favors the first-named system. Intervals are descriptive paired base-cluster intervals.

Contrast	ΔNRC	95% CI
Full-context expected − fixed	$-.325$	$[-.828, .028]$
Mem0 expected − fixed	.303	$[-.377, 1.073]$
RA-SBF − full-context expected	$-.154$	$[-.561, .099]$
RA-SBF − Mem0 expected	-2.393	$[-3.397, -1.390]$

CS-1 evaluated 741 strictly interior points of a normalized FS/MU/Q cost simplex. RA-SBF and cost-aware full-context uniquely minimized action loss on 50.20% and 48.31% of points and minimized-or-tied on 50.61% and 49.66%. RA-SBF had lower, equal, and higher action loss at 372, 3, and 366 points, respectively. This uniform geometric grid is a sensitivity surface, not a deployment prior or stakeholder-preference distribution.

K.2 MISSPECIFICATION STRESS TESTS

MS-1 regenerated 24 development bases and paired all conditions over the same bases and nested cost clones. Sharpening supplied reliability via logit scaling ($\alpha = 2$) worsened Bayesian expected-cost NRC by .236 $[.040, .463]$ relative to $\alpha = 1$; flattening it was inconclusive. Tested hazards from 0 to .5 produced no resolved action-cost change from the .05 reference, although belief scores differed. Deterministic asymmetric-noise conditions had adverse point estimates with intervals crossing zero.

The clearest violated assumption was correlated evidence. Repeating a same-source background observation three and five times worsened Bayesian NRC by .411 $[.053, .791]$ and .580 $[.095, 1.121]$. This is consistent with overcounting same-source evidence under an independent-observation likelihood and identifies a concrete failure mode for controllers that lack source-dependence modeling.

K.3 DURATION-AWARE QUARANTINE UTILITY

DQ-1 left v1 NRC unchanged and added time-integrated serving and state-integrity loss under a prespecified linear-cost, uniform-demand model. Relative to fixed Bayesian control, expected-cost control reduced serving loss by .533 $[-.984, -.144]$ and integrity loss by 1.088 $[-1.721, -.489]$; its restricted transition delay was .595 steps higher with an interval crossing zero $[-.387, 1.630]$. The matched WV expected-cost policy improved all three quantities. These results locate the duration trade-off of a rejection band under a linear-cost, uniform-demand model. Workload-calibrated duration weights and human-review costs remain targets for a deployment study.

K.4 MODEL AND ADAPTER DIAGNOSTICS

The original Qwen3-8B strict-adapter outputs were sealed as Q8-REF before a new wrapper-owned-metadata adapter was frozen. The wrapper supplies request and session IDs, source class, and logical step, leaving the model to infer event, value, and confidence. Qwen3-8B made zero semantic changes between adapters on all 51 sessions. Across seven cells, all 357 calls produced a valid response on their first attempt.

On this development set, the adapter removes structured-copy failures; semantic extraction determines the remaining downstream gap. Both 4B gaps exclude zero. Gemma3-4B has higher value F1 yet worse NRC than Qwen3-4B, showing that error location and action cost complement aggregate F1. Both families ran under one contract, motivating broader prompt and runtime studies.

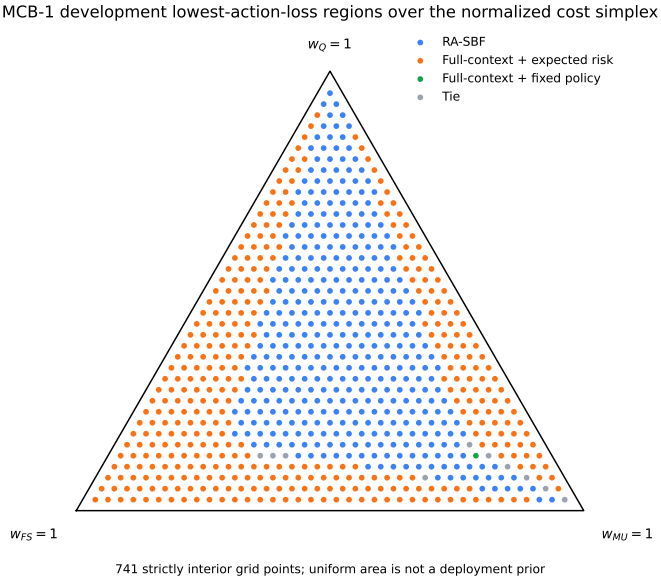


Figure 4: Development cost-simplex sensitivity surface. Each interior point uses the same evidence-derived beliefs and changes only the normalized FS/MU/Q cost vector. The near-even division between RA-SBF and cost-aware full-context beliefs makes the lower-action-loss source an explicit function of the application loss, rather than a global ranking. The uniform grid is geometric coverage, not a deployment prior.

Table 24: Frozen adapter-v2 development results. The three 4B temperature-zero seeds produced identical output fingerprints and are repeatability checks, not independent model samples.

Model	Event F1	Value F1	Auto NRC	Auto−gold NRC (95% CI)
Qwen3-8B bridge	1.000	.980	.197	.009 [.000, .029]
Qwen3-4B	.854	.809	.774	.585 [.126, 1.402]
Gemma3-4B	1.000	.863	1.077	.889 [.246, 1.576]

K.5 Deterministically selected qualitative cases

Table 25 links aggregate effects to concrete frozen records. To avoid favorable case selection, the four rows use fixed rules: the first high-confidence ($p \geq .9$) cross-profile action divergence in artifact order, the largest per-base burst-1 to burst-5 degradation, the longest BF-expected quarantine duration, and the first Track B extraction value mismatch in call order.

Table 25: Illustrative records selected by deterministic rules from frozen development or Track B artifacts.

Mechanism	Frozen record and interpretation
Cost changes action	On `extdev_temporary_change_01` at step 2, full-context belief is $p(x = 1) = .95$ in every clone. Symmetric and MU-critical costs switch to value 1; FS-critical costs quarantine the same candidate.
Correlated evidence	On `dev_temporary_change_base_000`, repeating the same-source observation five times raises BF-expected NRC from .500 to 4.083 and MU count from 1 to 13, the largest per-base burst degradation.
Quarantine duration	On the FS-critical temporary-change stream `base_015`, BF-expected quarantines for 29 of 47 elapsed steps, leaves one of two transitions unresolved, and accumulates 21 wrong-active steps.
Extraction reversal	For Orion Registry retention, the session says the revision "replaces 90 days with 30 days." With value 0 = 30 days and value 1 = 90 days, the frozen extractor returns 1 rather than gold 0. This is the first of ten value mismatches; event, source, and time fields remain correct.

Together, these studies show that action-loss ranking depends jointly on beliefs, costs, source dependence, deferral, and extraction; a confirmatory v2 should test longer, multi-valued streams under stakeholder-calibrated utility.