MEMORIES EXPIRE AT DIFFERENT RATES: CALIBRATING AGENT MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Agent memory can retrieve the right record, place it in time and revise a belief once a contradiction arrives. Yet none of these abilities tells whether a retrieved observation is still true before a contradiction appears. Recency heuristics discount age alone, but once memory types expire at different rates no function of age is calibrated for all types. We formalize the missing quantity as **current validity**, the probability that a stored proposition was correct when written and has not since been superseded, and build its estimator and decision rule together. Transition-Calibrated Memory (TCM) reads a memory's transition context into a discrete-time invalidation hazard and integrates it over elapsed time into a probability. A cost-optimal rule then prices that probability into using or verifying the memory. We evaluate TCM on human re-query judgments, two intensive-care cohorts and an exact-event simulator. Conditioning on transition context is the robust gain: it cuts the worst group-calibration gap of age-only rules from 0.358 to 0.197 on TicToc, survives standard recalibration and largely disappears when the context is permuted. Beyond calibration, current validity also changes what an agent does. Routed through one cost-sensitive rule at a matched 20% verification budget, TCM earns 0.05 to 0.09 more harm-weighted utility than the strongest baseline on both clinical cohorts in every seed. By contrast, prompted and post-trained agents deciding alone verify out of proportion to risk, over-verifying eight probes in ten or acting on stale memories in three. Survival factorization is only one estimator of current validity, and its added value is regime-dependent, which separates the robust benefit of transition conditioning from the inductive bias used to estimate it.

1 INTRODUCTION

Persistent memory lets an agent carry information across turns and sessions (Park et al., 2023; Packer et al., 2023), but a stored record is useful only while the record remains applicable. A flight status can change within minutes, a treatment goal may change after a new diagnosis, and an allergy record should not be discounted merely because time has passed. Figure 1 makes the difficulty concrete with two records from TicToc, a benchmark of human re-query judgments. Both were stored one hour before the agent is asked about them, and both came from equally reliable tools. One holds job postings and is asked which pays most, while the other holds a laundry order whose reported waiting time was one hour and is asked how much longer it will take. Annotators judged the first safe to reuse and the second not. An age-only rule sees only that an hour has passed and assigns both the same confidence. Transition context separates the two beliefs, because job postings and laundry queues are invalidated at very different rates.

Research on agent memory has answered which evidence to retrieve from a long history (Maharana et al., 2024; Wu et al., 2025), which time a query refers to (Du et al., 2026; Su et al., 2026), and how to revise a stored belief once new evidence arrives (Sun et al., 2026; Liao et al., 2026). However, one question these components leave open governs whether acting on a memory is safe at all: **how strongly should an agent trust a previously correct observation when no explicit contradiction has arrived?** Semantic relevance favors the old record, and a recency heuristic (Park et al., 2023) can only apply a generic penalty. Neither approach uses the fact that job postings, laundry queues, temperatures and allergies change at very different rates. Verbalized confidence (Tian et al., 2023; Xiong et al., 2024) does not resolve this either, because what needs calibrating is an unobserved

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.

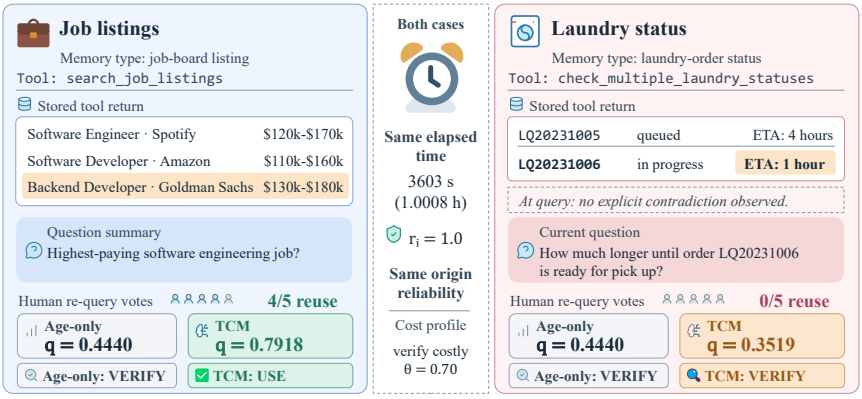

Figure 1: **Equal age does not imply equal re-query judgments.** Two TicToc records from the entity-level test split (seed 3) share elapsed time (3,603 s) and origin reliability ($r_i = 1$) but receive 4/5 and 0/5 human votes for reuse. Both receive $q = 0.4440$ from the flexible age-only hazard. TCM assigns 0.7918 to the job-board listing and 0.3519 to the laundry-order status, which gives USE and VERIFY under the *verify costly* profile (threshold $\theta = 0.70$, Appendix O). Labels are human re-query judgments, not observed transitions. One pair shows the limit of age-only confidence and does not rank the conditioned estimators (Sections 5.2 to 5.5).

state transition rather than the correctness of an answer. Getting it wrong is costly in both directions: excessive trust produces a stale action, excessive distrust repeated tests and delay.

A natural response is to fit a better decay curve, but that cannot succeed, for a structural reason rather than an empirical one. Once two memory types differ in survival probability at the same age, any predictor that sees only age gives them one confidence and is miscalibrated for at least one (Proposition 1). Current validity has to be modeled, not tuned.

To address this, we treat **current validity** as an explicit belief the policy carries about each stored memory, and propose Transition-Calibrated Memory (**TCM**)[1] to estimate current validity. In this sense a memory expires when the transition process behind it ends the validity episode, not when its text is discarded. Every entry keeps its proposition, observation time, type and provenance. From that context TCM predicts a conditional hazard of invalidation, then integrates the hazard over elapsed intervals into a current-validity probability. Fresh evidence revises that probability, and a cost-optimal rule then chooses between using the memory and verifying it. A conditional estimator could improve for either of two reasons, so we separate them in the evaluation. We ask first whether conditioning on the transition context beats age-only and type-aware rules. Only then do we ask whether the survival factorization adds anything over a direct predictor on the same context, which we report beside it as TCM-D.

Three findings follow. Transition conditioning is the robust gain: standard recalibration does not reproduce it, and permuting the context largely removes it. Current validity is a different target from event-time modeling on the simulator, and the factorization adds most when supervision includes censored episodes. Decisions follow the estimate, and an age-only rule stays the better-calibrated fallback under type or time shift.

2 Related Work

Agent memory and its temporal failures. Early language-agent memories stored interaction streams or compressed episodes and retrieved them by relevance (Park et al., 2023; Packer et al., 2023). Evaluation then moved to multi-session recall and temporal reasoning under knowledge updates. LoCoMo and LongMemEval showed that retrieval alone mishandles questions whose answer changed between sessions (Maharana et al., 2024; Wu et al., 2025). A second generation put memory operations inside the policy. AgeMem learns when to store, retrieve, update and discard

[1]Code: https://anonymous.4open.science/r/transition-calibrated-memory-D23D

(Yu et al., 2026). Memory-T1 aligns evidence with the time scope of a query (Du et al., 2026), Temporal Semantic Memory represents durative events (Su et al., 2026) and PAMU tracks evolving preferences (Sun et al., 2026). Several benchmarks isolate the stale case directly. TicToc measures whether elapsed time changes an agent's decision to re-query a tool (Cheng et al., 2026), and STALE asks whether later evidence implicitly invalidates an earlier fact (Chao et al., 2026). Memora credits an agent for withholding information that has since been deleted or updated (Uddin et al., 2026). BeliefMem keeps competing conclusions with probabilities so that a contradiction can be resolved rather than overwritten (Liao et al., 2026). Existing systems model when an event happened or which of two conflicting records to keep, and the re-query benchmarks measure agents that decide validity by prompt or by post-training. None estimates how the validity of a single record survives elapsed time and transition context. TCM supplies that estimate and routes it to a decision.

Calibration and time-to-event estimation. Accurate neural networks can be badly miscalibrated, which motivated post-hoc scaling and evaluation by proper scoring rules (Guo et al., 2017; Gneiting & Raftery, 2007). Language-model calibration elicits confidence from token probabilities, sampled answers or verbal reports (Tian et al., 2023; Xiong et al., 2024). However, the object scored there is a final answer at one moment, not a stored proposition whose truth can lapse. T^3 reduces belief deviation while an agent actively seeks information (Zou et al., 2026), whereas we calibrate the belief about one stored observation and price re-verification against it. Survival analysis models the time until a state changes, from proportional hazards (Cox, 1972) to discrete-time hazards parameterized by neural networks and trained with right-censored records (Gensheimer & Narasimhan, 2019). Clinical temporal reasoning has been benchmarked on timestamped histories (Cui et al., 2025), drawn from irregularly sampled real records (Johnson et al., 2023) and from simulators that expose exact transitions (Walonoski et al., 2018). None of this work evaluates whether a stored measurement's validity stays calibrated over elapsed time. We use the discrete-time hazard for heterogeneous time-to-change processes and evaluate its probabilities as validity forecasts across elapsed-time strata and memory types.

3 Problem Setting

An agent takes actions a_t at irregular decision times $t = 0, 1, \ldots, H$, between which a latent state s_t and current goal g_t evolve according to $(s_{t+1}, g_{t+1}) \sim P(\cdot \mid s_t, g_t, a_t, \Delta t_t)$, where Δt_t is the elapsed wall-clock or simulated time. At each step the agent observes $o_t \sim O(\cdot \mid s_t, g_t)$ (Kaelbling et al., 1998) and maintains a persistent memory $\mathcal{M}_t$. A policy-visible history holds timestamped observations, actions and retrieved memories, while protected state and future records never enter the prompt.

A memory item is $m_i = (\phi_i, y_i, t_i, k_i, p_i)$: proposition ϕ_i names a state or goal attribute whose value y_i was observed at time t_i, type k_i identifies the transition family, and p_i records provenance. Let $C_i = 1$ when the observation correctly represented the attribute at t_i, so that source reliability is $r_i = P(C_i = 1 \mid p_i, o_{t_i})$. A reliable origin does not make the proposition current.

Definition 1 (Current validity). *Origin C_i, correct when written. Current $V_{i,t}$, still true now.*

Let $A_{i,t} = 1$ when no transition relevant to ϕ_i has superseded memory m_i during $(t_i, t]$. Current validity is $V_{i,t} = C_i A_{i,t}$, which requires both a correct origin observation and a surviving validity episode. Current-validity confidence is

$$q_{i,t} = P(V_{i,t} = 1 \mid \mathcal{H}_t), \tag{1}$$

where $\mathcal{H}_t$ contains only information available to the deployed agent.

Episode survival treats a later coincidental return to y_i as a new episode, preventing an old observation from becoming trusted again without a fresh measurement.

Calibration drift. A predictor is group-calibrated (Dawid, 1982; Kleinberg et al., 2017; Hébert-Johnson et al., 2018) at age τ for type k when

$$P(V_{i,t} = 1 \mid q_{i,t} = q, k_i = k, t - t_i = \tau) = q. \tag{2}$$

We ask whether common recency scores violate Equation 2 after transition rates change. Our claim concerns confidence in the memory, not in the final language-model answer.

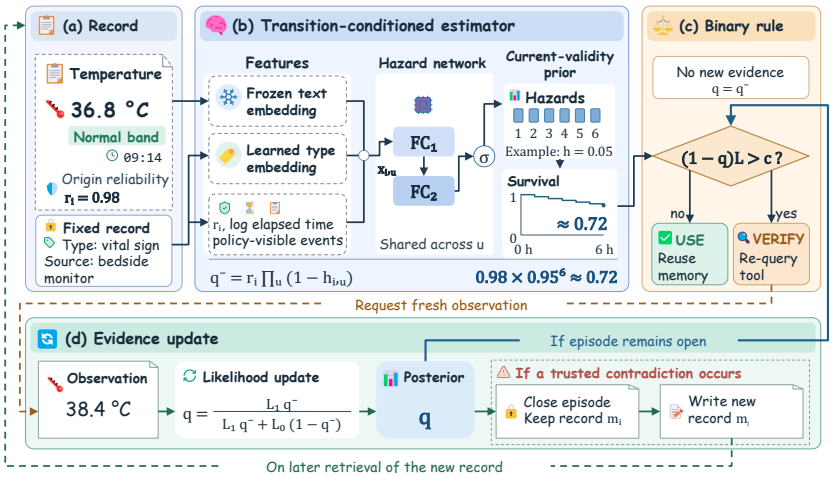

Figure 2: **Transition-Calibrated Memory.** (a) An immutable record keeps the proposition and its observation time together with transition type, provenance and origin reliability. (b) Text and type embeddings form $x_{i,u}$ with provenance reliability, log elapsed time and policy-visible events. A shared hazard network yields interval hazards whose survival product scaled by r_i is the prior $q^-_{i,t}$. Six equal hazards and $0.98 \times 0.95^6 \approx 0.72$ serve as an illustration, not as an assumption of constant hazard. (c) Without new evidence $q_{i,t} = q^-_{i,t}$, and the binary case of Proposition 2 verifies exactly when $(1 - q)L > c$. Equation 6 retains ABSTAIN. (d) Fresh evidence updates the belief by a likelihood ratio, and a trusted contradiction closes the episode and creates a new memory without rewriting the record.

Proposition 1 (Heterogeneous-hazard impossibility)**.** *Suppose memory types k_1 and k_2 have the same origin reliability $r > 0$, and let $S_k(\tau) = P(A_{i,t} = 1 \mid C_i = 1, k_i = k, t - t_i = \tau)$ denote the survival probability of type k at age τ. Suppose $S_{k_1}(\tau) \neq S_{k_2}(\tau)$ at some age τ. Any predictor $f(\tau)$ that uses age alone assigns the same confidence to both groups and cannot be calibrated for both at τ.*

Appendix A gives the proof. Proposition 1 motivates conditioning current validity on transition context rather than on age.

4 Transition-Calibrated Memory

By Proposition 1 an age-only rule cannot represent two memories of the same age whose processes differ. Each record already carries that process in its type and provenance and in the events observed since writing. TCM has three parts, drawn in Figure 2. A *transition-conditioned hazard* reads the immutable record and policy-visible inputs (a, b). A *survival product* scaled by source reliability integrates the hazard over elapsed intervals into a current-validity probability (Section 4.1). A *cost-optimal rule* then prices that probability into a decision (c, Section 4.2). Fresh evidence revises the belief rather than the record (d).

4.1 Conditional validity survival

Two design choices follow from the estimand itself. First, we model origin correctness and later invalidation separately. A wrong origin calls for re-observation and a lapsed episode for re-verification, and one confidence score cannot tell the two remedies apart. We therefore fix source reliability r_i by provenance, and the hazard model estimates only survival given a correct origin. Second, we represent validity as survival over elapsed-time intervals rather than as a function of age, because an episode ends at most once and at a moment that is usually unobserved. A per-interval hazard makes the estimate a product of survivals that is monotone in age for fixed context. A hazard model also accepts right-censored episodes and exact event times where the data provide them. Where the data

give only validity at the next probe, it can still be trained. TCM thus takes the discrete-time survival form of Singer & Willett (1993) with a neural hazard (Gensheimer & Narasimhan, 2019) conditioned on memory context. Let $u = 1, \ldots, U_{i,t}$ index elapsed-time intervals after memory creation. Each input $x_{i,u}$ collects type and provenance together with elapsed time and policy-visible events since t_i. A shared hazard network gives interval hazards and prior validity,

$$h_{i,u} = P(A_{i,u} = 0 \mid A_{i,u-1} = 1, C_i = 1, x_{i,u}) = \sigma(f_\psi(x_{i,u})), \qquad q^-_{i,t} = r_i \prod_{u=1}^{U_{i,t}} (1 - h_{i,u}), \quad (3)$$

the computation Figure 2(b) draws. Since $x_{i,u}$ holds only policy-visible events, no future event reaches the estimator.

Two data regimes train TCM differently, and that difference matters for reading our results. Our simulator records the transition that ends an episode. An uncensored sequence has $d_{i,u} = 1$ in the first interval containing it, where observation ends, and a right-censored sequence has every $d_{i,u} = 0$. With U_i the number of observed intervals for memory i, the discrete survival likelihood is

$$\mathcal{L}_{\text{surv}}(\psi) = -\sum_i \sum_{u=1}^{U_i} \left[d_{i,u} \log h_{i,u} + (1 - d_{i,u}) \log(1 - h_{i,u}) \right]. \quad (4)$$

Real datasets record no transition time, only the validity $v_{i,t}$ of the memory at the next probe. Under current-status supervision the survival term reduces to a Bernoulli likelihood over probe labels (Appendix H), and Equation 4 as written is exercised only on the simulator. Probe labels also enter through a proper score (Brier, 1950),

$$\mathcal{L}_{\text{cal}}(\psi) = \mathbb{E}_{(i,t)} \left(q_{i,t} - v_{i,t} \right)^2, \quad (5)$$

whose expectation is minimized only by the true conditional probability (Gneiting & Raftery, 2007), so that a well-fitted hazard is also a calibrated one. We minimize $\mathcal{L}_{\text{surv}} + \alpha \mathcal{L}_{\text{cal}}$, with α fixed on development data.

A direct variant. Under current-status supervision the same context can also feed a *direct* predictor $q^-_{i,t} = r_i \, \sigma(g_\psi(x_i, \log(t - t_i)))$ without the factorization. We call it TCM-D and train it with the same losses. TCM-D cannot consume censored episodes or event times and is not monotone in age. Every estimator, controls included, makes its open choices on the development split from a fixed grid (Appendix C). TCM and TCM-D choose their conditioning blocks there, and every hazard model chooses whether the last interval is charged in full or by the log-time spent in it.

Evidence update. Fresh evidence revises the prior by the ratio L_1/L_0, where L_1 and L_0 score a new observation under a valid and an invalid memory respectively (Kaelbling et al., 1998), as Figure 2(d) shows. A trusted measurement that contradicts m_i closes its validity episode and creates a new memory without deleting the audit trail. Appendix C gives the likelihood models used in each environment.

4.2 Acting on current validity

Whether a memory should be verified is a value-of-information question (Howard, 1966). Two quantities decide it: the probability $1 - q_{i,t}$ that the memory has expired, and the cost of acting on an expired memory relative to checking it. At each memory-dependent decision the policy chooses $d_t \in \{\text{USE}, \text{VERIFY}, \text{ABSTAIN}\}$ before the task action. Verification invokes the environment's observation tool and abstention returns a request for clarification or deferral. Each step earns the reward

$$R_t = R_t^{\text{task}} - c_{\text{ver}} \mathbf{1}[d_t = \text{VERIFY}] - c_{\text{abs}} \mathbf{1}[d_t = \text{ABSTAIN}] - c_{\text{stale}} \mathbf{1}[d_t = \text{USE}, V_{i,t} = 0], \quad (6)$$

with costs fixed by the environment or a predefined risk tier.

Proposition 2 (One-step verification rule)**.** *Assume verification reveals validity without error and does not alter the task state. Let stale use incur loss $L > 0$ and verification cost $c > 0$. For a memory with current validity q, verification has lower expected loss than immediate use exactly when $(1 - q)L > c$.*

Table 1: Current-validity calibration: Brier score overall and on the worst memory-type group, over five seeds and fifteen for MIMIC-IV. Bold marks the best value in each column. TCM-D recalibrated adds the post-hoc calibrator chosen on development data (Table 9).

	PhysioNet-2012		TicToc		MIMIC-IV	
Method	overall	worst group	overall	worst group	overall	worst group
Age-only rules						
Global exponential decay	0.1923	0.2919	0.2931	0.5351	0.1961	0.3691
Flexible age-only hazard	0.1649	0.2567	0.1763	0.2775	0.1497	0.2886
Type-aware rules						
Typed constant decay	0.1717	0.2762	0.1617	0.3748	0.1875	0.3532
Flexible typed-age hazard	0.1588	0.2375	0.1368	**0.2028**	0.1415	**0.2371**
Transition-calibrated memory (ours)						
TCM (ours)	**0.1477**	**0.2256**	0.1368	0.2100	0.1233	0.2391
TCM-D (ours)	0.1485	0.2260	**0.1359**	0.2049	0.1231	0.2377
TCM-D recalibrated (ours)	0.1484	0.2258	0.1364	0.2113	**0.1230**	0.2376

Decision rule. A cost-optimal action given $q_{i,t}$ maximizes expected utility over the three choices of Equation 6. Proposition 2 is its binary case (Figure 2c) and coincides with the cost-sensitive threshold of Elkan (2001) with the reject option of Chow (1970). Each cost profile therefore implies a threshold θ on q (Appendix O). Our rule has no free parameters, so a utility difference between estimators routed through it is a difference in the estimate. Under a verification budget B the B decisions of largest expected gain are verified, so the *ranking* induced by q decides which memories are checked. Without a budget the rule compares $(1-q)L$ with c, so the numerical value of q is what is paid for. Where verification is noisy or alters the task state, the proposition's assumptions fail and a learned policy could do better (Appendix C).

5 Experiments

5.1 Setup

We evaluate on three real datasets and one simulator (Appendix L). TicToc supplies timestamped tool-use trajectories in which annotators judge whether a stored observation should be re-queried under a controlled elapsed-time intervention (Cheng et al., 2026), and we read those votes as a human proxy for current validity rather than as observed changes in the world. PhysioNet-2012 (Silva et al., 2012; Goldberger et al., 2000) supplies intensive-care trajectories in which a stored measurement is probed at the next measurement of the same concept and is current if it stays in the same predefined clinical band, with a concept's final measurement as censoring. MIMIC-IV (Johnson et al., 2023) is a larger clinical cohort from a later database release. All three real datasets carry interval-censored current-status labels only. A simulator (Walonoski et al., 2018) adds exact transition times with right censoring and is the one setting that exercises Equation 4 as written. Splits are by episode or patient, and real data serve retrospective calibration only.

We compare two age-only rules (a tuned global exponential and a flexible age-only hazard) and two type-aware rules (typed constant decay and a flexible typed-age hazard) with TCM and its direct variant TCM-D (Section 4.1). Post-hoc calibrators and survival models enter in Sections 5.3 and 5.4. Scores are overall and worst-group Brier, and Appendix J audits every baseline, Memory-T1 included.

5.2 Can current validity be estimated?

Table 1 gives the main result on the three real target distributions. Every flexible estimator that conditions on type or context beats both age-only rules on both metrics in every study, as Proposition 1 predicts once types expire at different rates. A tuned global exponential, the recency rule in common use, is worse than even a constant base rate on PhysioNet-2012 and TicToc (Appendix I): one decay imposed on every memory is worse than none. TCM or TCM-D has the lowest overall Brier in every study, and the two differ by less than 0.001 in overall Brier everywhere ($p = 0.34$ on MIMIC-IV), as expected when both fit the same likelihood. On real data the gain therefore comes from conditioning

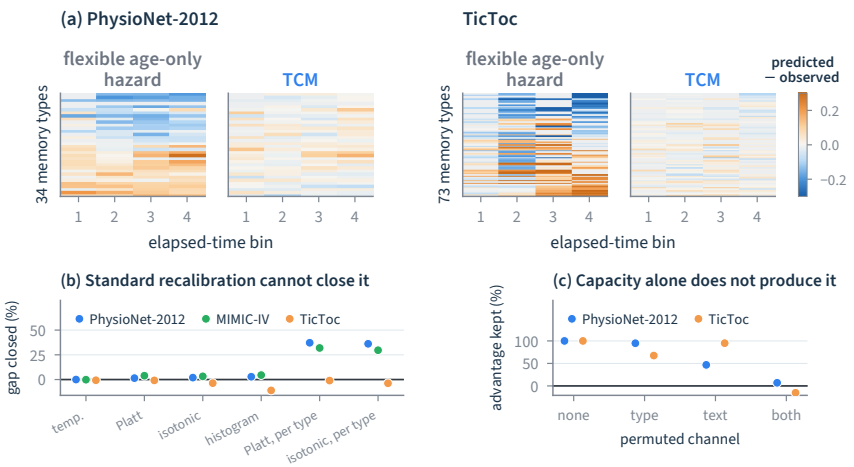

Figure 3: **Age-only error is structured, and neither recalibration nor capacity removes it.** (a) Predicted minus observed validity of the flexible age-only hazard and TCM, by memory type and elapsed time. (b) Share of the age-only gap to TCM-D closed by each calibrator, which moves TCM-D itself by at most 0.0014 Brier. (c) Advantage kept when a channel is permuted. Five seeds, fifteen for MIMIC-IV in (b) and three in (c).

on transition context, not from the survival factorization. How much context adds beyond the type label depends on the domain. On the clinical cohorts the proposition text carries information the type label does not. On MIMIC-IV, TCM therefore beats the flexible typed-age hazard by 0.018 and typed constant decay by 0.064 in all fifteen seeds ($p < 10^{-4}$). On TicToc the typed-age hazard ties TCM, because type and elapsed time carry nearly all the signal. Per-type curves also keep a slightly better worst group than TCM-D on TicToc and MIMIC-IV (0.002 and 0.0006), so a shared estimator trades a little of its rarest group for average fit.

5.3 Why does transition context matter?

Figure 3 tests three explanations of the gain. Age-only error is structured, not noisy. We find that the flexible age-only hazard gives nearly the same confidence to groups whose observed validity ranges from near zero to near one. As time passes it distrusts durable types and over-trusts volatile ones more (Figure 3a), as Proposition 1 predicts. More data cannot remove a bias of this kind, and conditioning removes much of it, cutting the maximum group calibration gap from 0.358 to 0.197 on TicToc. PhysioNet-2012 and TicToc differ in how validity depends on context. On PhysioNet-2012 the error is mostly a per-concept level, and TCM-D gains almost nothing from elapsed time once it knows the context (0.1503 Brier without it, 0.1499 with). On TicToc neither suffices alone: context without time (0.2352) is worse than time without context (0.1763), and only the two together reach 0.1357 (Appendix I). In these data clinical memories differ mainly in how often they hold and tool returns in how fast they expire, and a conditional hazard covers both.

Standard recalibration does not reproduce the gain (Figure 3b, Table 10). We fit temperature (Guo et al., 2017), Platt (Platt, 2000), isotonic (Zadrozny & Elkan, 2002) and histogram (Zadrozny & Elkan, 2001) calibration on a development half disjoint from early stopping (Park et al., 2025). Global maps close at most 5% of the age-only hazard's gap to TCM-D, and even per-type maps close under half of it. A monotone map of one score cannot separate groups that share that score, so what is missing is context, not scale. Applied to TCM-D, the same maps move it by at most 0.0014 Brier, so its probabilities are already calibrated.

Capacity alone does not produce it (Figure 3c). Permuting the conditioning context across memories before training (Breiman, 2001), with elapsed time and labels untouched, keeps the model's capacity and destroys only the link between a memory and its context. Permuting one channel keeps 47% to 95% of the advantage because type and text overlap, the text carrying more of it on PhysioNet-2012 and the type on TicToc. Permuting both leaves 7% and −15% (Appendix G). We conclude that the

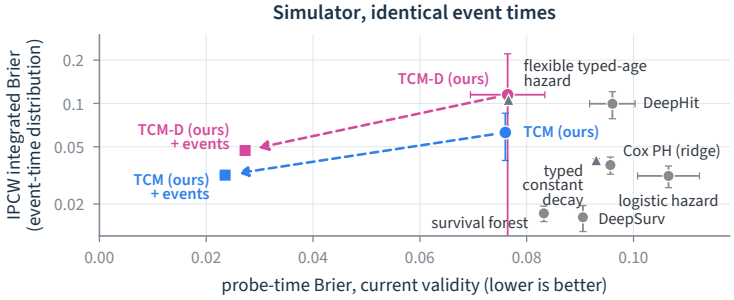

Figure 4: **Current validity and event time are best estimated by different models.** Simulator, identical event times, five seeds and four for TCM-D + events. Probe-time Brier against IPCW integrated Brier, with dashed arrows adding policy-visible events to the static inputs.

gain comes from transition information that neither recalibration nor capacity supplies, a necessity argument rather than a causal identification.

5.4 What does survival structure buy?

We use the simulator to ask what survival structure buys. Every model in Figure 4 reads the same static inputs and event times (Appendix K). We compare Cox proportional hazards (Cox PH) with a ridge penalty chosen on development episodes (Cox, 1972), a discrete-time logistic hazard (Singer & Willett, 1993) and a random survival forest (Ishwaran et al., 2008). DeepHit (Lee et al., 2018), DeepSurv (Katzman et al., 2018) and the type-aware rules complete the set of comparators for TCM and TCM-D.

On this simulator, the best current-validity estimate and the best event-time distribution come from different models (Figure 4). TCM and TCM-D reach a lower probe-time Brier than every survival-literature model in Table 2 (0.0760 and 0.0764 against 0.083), while the survival forest and DeepSurv fit the event-time distribution best. Event-time models are fitted to the whole time-to-event distribution, whereas current validity is scored at the moments memories are used, and the two objectives reward different models. From static inputs the factorization adds little, TCM-D being as accurate as TCM. Factorization pays off when supervision is incomplete. With policy-visible events and censored episodes, which only a hazard model can consume, TCM reaches 0.0235 at a concordance of 0.94. TCM also beats TCM-D overall and on the worst group in all four paired seeds (mean 0.42 against 1.00). Neither result is an artifact of training size (Table 3). We refit every model on the survival forest's 60 000 training episodes. TCM keeps the lowest probe-time Brier against the forest, DeepSurv and the logistic hazard in all five seeds ($p < 10^{-3}$). On static inputs TCM now also leads TCM-D (0.0756 against 0.0825 in four of five seeds, worst group 0.26 against 1.00 in all five), so with less training data the factorization helps even without events (Appendix K). NeuralSurv (Monod et al., 2025) fits only at the 125 episodes of its authors' experiments and is compared there (Appendix K).

Table 2: Survival models on the simulator, probe-time Brier and IPCW integrated Brier, five seeds and four for TCM-D + events. Below the rule both also see policy-visible events. Bold marks each block's best.

Method	Brier ↓	IBS ↓
Cox PH (ridge)	0.0957	0.0373
Logistic hazard	0.1066	0.0314
Survival forest	0.0832	0.0172
DeepSurv	0.0905	**0.0162**
TCM-D (ours)	0.0764	0.1150
TCM (ours)	**0.0760**	0.0628
TCM-D (ours) + events	0.0273	0.0471
TCM (ours) + events	**0.0235**	**0.0317**

Table 3: Survival models and ours refitted on the same 60 000 training episodes, static inputs, five seeds.

Method	Brier ↓	IBS ↓
TCM (ours)	**0.0756**	0.0628
TCM-D (ours)	0.0825	0.1592
Survival forest	0.0832	**0.0172**
DeepSurv	0.0911	0.0181
Logistic hazard	0.1068	0.0312

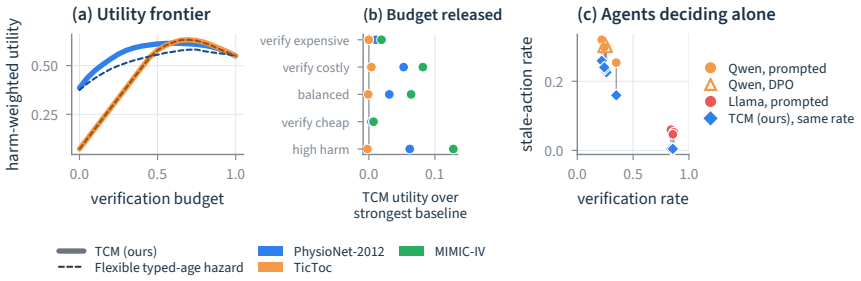

Figure 5: **Better validity estimates make better verification decisions, and agents deciding alone do not.** (a) Utility against budget, balanced profile. (b) TCM over the strongest baseline with the budget released. (c) Stale actions against verification rate, with TCM at each agent's rate. Five seeds, fifteen for MIMIC-IV.

5.5 Do better probabilities lead to better actions?

TCM leads the matched-budget frontier on both clinical cohorts under one decision rule. At a 20% budget it improves harm-weighted utility over the strongest baseline by $+0.052$ on PhysioNet-2012 (Figure 5a) and $+0.093$ on MIMIC-IV in every seed. With the budget released the gain stays positive at every cost profile and is largest when stale actions are costly ($+0.062$ and $+0.128$ under high harm, Table 21). When verifying is nearly free or prohibitive, every rule verifies or uses everything and calibration matters less. Decisions follow the estimate. On both clinical cohorts the lower average Brier among the three leading estimators earns the higher utility. On TicToc the three differ by under 0.001 Brier and tie in balanced-profile utility within 0.002. One estimate also serves every cost profile. A policy trained end to end for one profile gains at most 0.018 on average at its own costs and loses 0.13 to 0.17 on average at the others (Appendix Q), whereas q is fitted once. Current validity is a decision variable, not a decision.

Agents deciding alone do not verify in proportion to risk (Figure 5c). Qwen agents (Yang et al., 2025) under-verify, re-querying in 22% to 35% of decisions and acting on stale memories in up to 32%. Llama (Grattafiori et al., 2024) over-verifies and re-queries in over eight of ten. Timestamps or TicToc's re-query rule in the prompt shift these rates but not their direction, and preference optimization (Rafailov et al., 2023) on TicToc's re-query pairs changes no decision. Model family, not memory risk, sets how often an agent verifies. At each agent's verification rate TCM acts on stale memories less often, 0.24 against 0.30 for Qwen and 0.004 against 0.05 for Llama, so the gain comes from which memories are verified. Abstention cuts the Qwen rate to 0.09 (Appendix P).

6 Conclusion

We identified **current validity** as the confidence target a persistent memory needs and introduced TCM, which estimates it from transition context for a cost-optimal rule. Conditioning on transition context nearly halves the worst group-calibration gap of age-only rules on TicToc, survives recalibration and largely vanishes when the context is shuffled. Current validity differs from event time as a target, and decisions follow the estimate. Transition conditioning is the robust contribution. Survival factorization, by contrast, is one estimator of current validity whose added value depends on the supervision regime. TCM's scope is bounded, however. On transition families and elapsed times unseen in training, an age-only hazard that holds its last value is better calibrated than learned context. A label-free training-support test flags every such probe and under 0.1% of in-distribution ones, and falling back to the age-only hazard on flagged probes cuts worst-group Brier by at least a third (Appendix M).

Three assumptions bound these results (Appendix L). On real data validity is observed only at the next measurement or annotator judgment, and clinical probe times are informative because clinicians order tests when they suspect change. Episode survival treats a return to an earlier value as a new episode, which is conservative for reversible processes. We also assume a type schema that stays fixed between training and deployment.

AI USE STATEMENT

The research idea, the problem formulation and the study design are the authors' own. We also ran and supervised every experiment ourselves. Within this work, we used generative AI tools as assistants for feedback on hypotheses and experimental design. Generative AI tools also helped formulate the mathematical claims and write their proofs, which we checked step by step. AI assistance further covered implementing methods and data processing to our specification and drafting interpretations of results, which we checked against the data. We have not used generative AI tools to generate synthetic data sets or to assist with translation, and qualitative or thematic data analysis is not applicable to this work. Additionally, we used generative AI tools to edit code and figures and to search and summarize literature. Drafting and polishing the paper and formatting references were also AI-assisted. We reviewed all AI-assisted work and checked every cited source by hand. Every reported number, proof, table and figure was verified by the authors. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The clinical analysis is retrospective and does not produce treatment recommendations. MIMIC-IV remains under its credentialed PhysioNet agreement. No protected record, derived row, or patient-level output is redistributed. Synthea is used only for controlled validation against known transitions, and results from synthetic patients are reported separately from real records.

REPRODUCIBILITY STATEMENT

Section 5 and Appendices B and L specify validity labels and censoring rules. Appendix L fixes data splits and statistical tests. Losses are in Appendix C and decision costs in Appendix O. Appendices J and K give baseline adaptations. Appendix F describes how each run is recorded. Our anonymous repository, linked in the introduction, holds the code for every reported run.

REFERENCES

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

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

Yize Cheng, Arshia Soltani Moakhar, Chenrui Fan, Parsa Hosseini, Kazem Faghih, Zahra Sodagar, Wenxiao Wang, and Soheil Feizi. Your LLM agents are temporally blind: The misalignment between tool use decisions and human time perception. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 37082–37104. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.FINDINGS-ACL.1848. URL https://doi.org/10.18653/v1/2026.findings-acl.1848.

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

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Hejie Cui, Alyssa Unell, Bowen Chen, Jason Alan Fries, Emily Alsentzer, Sanmi Koyejo, and Nigam H. Shah. TIMER: Temporal instruction modeling and evaluation for longitudinal clinical records. *npj Digit. Medicine*, 8(1), 2025. doi: 10.1038/S41746-025-01965-9. URL https://doi.org/10.1038/s41746-025-01965-9.

Cameron Davidson-Pilon. lifelines: survival analysis in Python. *J. Open Source Softw.*, 4(40):1317, 2019. doi: 10.21105/JOSS.01317. URL https://doi.org/10.21105/joss.01317.

A Philip Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77 (379):605–610, 1982.

Yiming Du, Baojun Wang, Yifan Xiang, Zhaowei Wang, Wenyu Huang, Boyang Xue, Bin Liang, Xingshan Zeng, Fei Mi, Haoli Bai, Lifeng Shang, Jeff Z. Pan, Yuxin Jiang, and Kam-Fai Wong. Memory-T1: Reinforcement learning for temporal reasoning in multi-session agents. In *International Conference on Learning Representations*, 2026.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993. doi: 10.1201/9780429246593.

Charles Elkan. The foundations of cost-sensitive learning. In Bernhard Nebel (ed.), *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, IJCAI 2001, Seattle, Washington, USA, August 4-10, 2001*, pp. 973–978. Morgan Kaufmann, 2001.

Michael F Gensheimer and Balasubramanian Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257, 2019.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545, 1999.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL http://proceedings.mlr.press/v70/guo17a.html.

Humza Haider, Bret Hoehn, Sarah Davis, and Russell Greiner. Effective ways to build and evaluate individual survival distributions. *J. Mach. Learn. Res.*, 21:85:1–85:63, 2020. URL https://jmlr.org/papers/v21/18-772.html.

Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546, 1982.

Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1939–1948. PMLR, 2018. URL http://proceedings.mlr.press/v80/hebert-johnson18a.html.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Ronald A Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966.

Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3), 2008. doi: 10.1214/08-AOAS169.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, 2018.

Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In Christos H. Papadimitriou (ed.), *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, Berkeley, CA, USA, January 9-11, 2017*, volume 67 of *LIPIcs*, pp. 43:1–43:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017. doi: 10.4230/LIPICS.ITCS.2017.43. URL https://doi.org/10.4230/LIPIcs.ITCS.2017.43.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and Cox regression. *J. Mach. Learn. Res.*, 20:129:1–129:30, 2019. URL https://jmlr.org/papers/v20/18-424.html.

Changhee Lee, William R. Zame, Jinsung Yoon, and Mihaela van der Schaar. DeepHit: A deep learning approach to survival analysis with competing risks. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 2314–2321. AAAI Press, 2018. doi: 10.1609/AAAI.V32I1.11842. URL https://doi.org/10.1609/aaai.v32i1.11842.

Junfeng Liao, Qizhou Wang, Jianing Zhu, Bo Du, Rui Yan, and Xiuying Chen. Belief memory: Agent memory under partial observability. *CoRR*, abs/2605.05583, 2026. doi: 10.48550/ARXIV.2605.05583. URL https://doi.org/10.48550/arXiv.2605.05583.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 13851–13870. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.747. URL https://doi.org/10.18653/v1/2024.acl-long.747.

Mélodie Monod, Alessandro Micheli, and Samir Bhatt. NeuralSurv: Deep survival analysis with Bayesian uncertainty quantification. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/4985e71a19eb3c0e1b31657f8d9dec9e-Abstract-Conference.html.

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

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In Sean Follmer, Jeff Han, Jürgen Steimle, and Nathalie Henry Riche (eds.), *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, pp. 2:1–2:22. ACM, 2023. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Young-Jin Park, Kristjan Greenewald, Kaveh Alimohammadi, Hao Wang, and Navid Azizan. Know what you don't know: Uncertainty calibration of process reward models. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/37664246a1e07e212ddacea6e5a523f2-Abstract-Conference.html.

John C. Platt. Probabilities for SV machines. In Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans (eds.), *Advances in Large-Margin Classifiers*, pp. 61–74. MIT Press, 2000. doi: 10.7551/mitpress/1113.003.0008.

Sebastian Pölsterl. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *J. Mach. Learn. Res.*, 21:212:1–212:6, 2020. URL https://jmlr.org/papers/v21/20-729.html.

Shiang Qi, Weijie Sun, and Russell Greiner. SurvivalEVAL: A comprehensive open-source Python package for evaluating individual survival distributions. In Christopher W. Geib and Ron P. A. Petrick (eds.), *Proceedings of the 2023 AAAI Fall Symposia, Arlington, Virginia, USA, October 25-27, 2023*, pp. 453–457. AAAI Press, 2024. doi: 10.1609/AAAISS.V2I1.27713. URL https://doi.org/10.1609/aaaiss.v2i1.27713.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 3982–3992. Association for Computational Linguistics, 2019. doi: 10.18653/V1/D19-1410. URL https://doi.org/10.18653/v1/D19-1410.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/arXiv.2402.03300.

Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. Predicting in-hospital mortality of ICU patients: The PhysioNet/Computing in Cardiology challenge 2012. In *2012 Computing in Cardiology*, pp. 245–248. IEEE, 2012.

Judith D Singer and John B Willett. It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics*, 18(2):155–195, 1993.

Miao Su, Yucan Guo, Zhongni Hou, Long Bai, Zixuan Li, Yufei Zhang, Guojun Yin, Wei Lin, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. Beyond dialogue time: Temporal semantic memory for personalized LLM agents. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 29935–29951. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.FINDINGS-ACL.1496. URL https://doi.org/10.18653/v1/2026.findings-acl.1496.

Haoran Sun, Zekun Zhang, and Shaoning Zeng. Preference-aware memory update for long-term LLM agents. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 783–793. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.FINDINGS-ACL.38. URL https://doi.org/10.18653/v1/2026.findings-acl.38.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 5433–5442. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.330. URL https://doi.org/10.18653/v1/2023.emnlp-main.330.

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.

Jason A. Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *J. Am. Medical Informatics Assoc.*, 25(3):230–238, 2018. doi: 10.1093/JAMIA/OCX079. URL https://doi.org/10.1093/jamia/ocx079.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pp. 38–45. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-DEMOS.6. URL https://doi.org/10.18653/v1/2020.emnlp-demos.6.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=pZiyCaVuti.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yi Yu, Liuyi Yao, Yuexiang Xie, Qingquan Tan, Jiaqi Feng, Yaliang Li, and Libing Wu. Agentic memory: Learning unified long-term and short-term memory management for large language model agents. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.),

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026, pp. 21457–21483. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.981. URL https://doi.org/10.18653/v1/2026.acl-long.981.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In Carla E. Brodley and Andrea Pohoreckyj Danyluk (eds.), *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pp. 609–616. Morgan Kaufmann, 2001.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada*, pp. 694–699. ACM, 2002. doi: 10.1145/775047.775151. URL https://doi.org/10.1145/775047.775151.

Deyu Zou, Yongqiang Chen, Jianxiang Wang, Garry Yang, Mufei Li, Qing Da, James Cheng, Pan Li, and Yu Gong. Reducing belief deviation in reinforcement learning for active reasoning of LLM agents. In *International Conference on Learning Representations*, 2026.

A PROOFS

Proof of Proposition 1. At fixed age τ, an age-only predictor emits one value $f(\tau)$ for both types. Calibration for k_1 requires $f(\tau) = rS_{k_1}(\tau)$. Calibration for k_2 requires $f(\tau) = rS_{k_2}(\tau)$. Under the assumed inequality and $r > 0$, both equalities cannot hold. At least one type has nonzero absolute calibration error. □

Proof of Proposition 2. Immediate use has expected loss $(1 - q)L$. Perfect verification has loss c and reveals the state before use. Verification is strictly preferred when its loss is lower, which gives $c < (1 - q)L$. Equality permits either action. □

B VALIDITY-LABEL CONSTRUCTION

On the simulator the generator's event log stays with the evaluator. A memory episode begins when a policy-visible observation is written and ends at the first logged event that changes the proposition's referenced attribute. We assign the discrete hazard interval from the event timestamp and ignore events that do not affect the proposition. Appendix L gives the re-query and PhysioNet-2012 labels.

For MIMIC-IV, labels are interval-censored. A structured measurement creates a memory after unit normalization and plausibility filtering. Each interval closes at the next recorded measurement of the same concept. A change event occurs when the new value crosses a predefined clinical bin or changes the relevant categorical state. No next measurement means right censoring. We do not use notes to infer hidden changes in the primary analysis.

C MODEL INPUTS AND TRAINING SEQUENCE

Fresh evidence revises the survival prior through an explicit likelihood step. An evidence model supplies $L_1 = P(o_t \mid V_{i,t} = 1)$ and $L_0 = P(o_t \mid V_{i,t} = 0)$, and the posterior current validity is

$$q_{i,t} = \frac{L_1 q_{i,t}^-}{L_1 q_{i,t}^- + L_0(1 - q_{i,t}^-)}. \tag{7}$$

Executable environments compute L_1 and L_0 from exact observation-noise models, whereas text settings use a frozen evidence classifier trained only on development annotations.

Every estimator comparison we report routes through the closed-form rule of Proposition 2, which is optimal for the one-step problem the environments here pose. An environment with genuine

Table 4: Development-split choices switched off one at a time, five seeds. Bold marks the best value in each column.

	PhysioNet-2012		TicToc	
Estimator	overall	worst group	overall	worst group
TCM (ours)	**0.1477**	**0.2256**	0.1368	0.2100
discrete exposure only	**0.1477**	**0.2256**	0.1408	0.2128
TCM-D (ours)	0.1485	0.2260	0.1359	0.2049
all features	0.1499	0.2259	**0.1357**	**0.2046**

sequential coupling would instead call for a learned policy under a KL penalty against a reference policy (Schulman et al., 2017; Shao et al., 2024),

$$\max_{\theta} \ \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{H} \gamma^t R_t\right] - \beta\, \mathbb{E}_{\mathcal{H}_t} D_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\mathrm{ref}})\,. \tag{8}$$

In such a setting validity labels would shape training rewards and never enter the deployed prompt.

TCM's hazard input contains a frozen text embedding of the proposition from `all-MiniLM-L6-v2` (Reimers & Gurevych, 2019; Wang et al., 2020) and a learned type embedding. Further inputs are provenance reliability, log elapsed time and policy-visible events since creation. A two-layer multilayer perceptron forms the hazard network. Every run uses a hidden width of 128. A monotone time constraint is enforced by the survival product rather than a special network architecture.

Choices made on the development split. We fit each estimator once for every candidate in a fixed grid. We report the candidate with the lowest development loss, which is the training objective evaluated on development probes. TCM and TCM-D share five candidate sets of conditioning blocks. One candidate uses all blocks. Four others drop policy-visible events, memory type or text, or keep only type and provenance. Every hazard model, TCM and the two flexible age-only controls alike, also chooses its exposure. With discrete exposure a probe in interval U is charged that interval's whole hazard, as in Equation 3. With continuous exposure the final factor becomes $(1-h_{i,U})^{\phi}$, where ϕ is the share of the interval's log-time that has elapsed. Continuous exposure makes q^{-} continuous in elapsed time and lets the last interval's rate continue past the grid. For TCM, development data choose discrete exposure in all twenty-five clinical runs including shift splits, where validity is mostly a type-level offset that the first interval absorbs. On TicToc they choose continuous exposure in all fifteen runs, shift splits included. TCM-D takes no exposure because it has no intervals. Table 4 switches each choice off.

Training fits the hazard network and the evidence model with Equations 4 and 5, and every result we report uses them with the closed-form decision rule.

All methods share the base retriever and its context budget. TCM does not add a new candidate selector. Each retrieved memory receives a validity probability, and the policy-visible record contains the probability with its timestamp and provenance tier. Only the evaluator sees the raw protected label.

D Metrics and aggregation

For N validity probes, Brier score (Brier, 1950) is

$$\mathrm{BS} = \frac{1}{N}\sum_{n=1}^{N}(q_n - v_n)^2. \tag{9}$$

Time-stratified Brier score uses log-time bins defined on training data. Worst-group Brier takes the maximum over memory types. Reliability diagrams display equal-mass bins but do not determine the primary statistic.

Harm-weighted utility sums task reward and the costs in Equation 6. We report the full verification-cost sweep. A method dominates another only when its paired confidence interval excludes zero at a shared budget. Stopping without a stale action does not count as task success.

Table 5: Context-permutation control. Advantage is $\mathrm{Brier}(\text{age-only}) - \mathrm{Brier}(\text{TCM, all features})$, three seeds, mean±sd. Retained is that advantage as a percentage of its unpermuted value.

Permutation	PhysioNet-2012		TicToc	
	Advantage	Retained	Advantage	Retained
none (unpermuted)	+0.0152±0.0001	100%	+0.0394±0.0029	100%
memory type + provenance	+0.0144±0.0001	95%	+0.0266±0.0016	67%
proposition text	+0.0071±0.0001	47%	+0.0374±0.0036	95%
both channels	+0.0011±0.0000	7%	−0.0060±0.0026	−15%

Table 6: Replication on a third patient cohort that was never used for fitting, selection or tuning, three seeds. Bold marks the lowest Brier.

Method	Brier on held-out cohort ↓
Flexible age-only hazard	0.1654
Flexible typed-age hazard	0.1592
TCM-D (ours)	**0.1489**
TCM, all features	0.1504

E Clinical scope and leakage controls

MIMIC-IV access requires credentialing and a data-use agreement. Raw rows and patient identifiers remain outside the repository. We split by subject before concept extraction. Normalization statistics, bin boundaries, and censoring policies use training subjects only. No discharge summary or future measurement enters a policy prompt.

Our medical analysis answers one narrow question: whether a stored structured observation remains calibrated at the next recorded measurement. Diagnosis, treatment selection and clinical outcomes are outside its scope. Any subgroup analysis requires sufficient sample size and is reported with uncertainty rather than a deployment recommendation.

F Reproducibility

Each run records its configuration and data splits with the seeds and model revisions it used. Prompts, hardware, cost profiles and censoring rules are recorded with them. Every table and figure in the paper is generated from these records rather than transcribed. We report null results alongside positive ones in Section 5.

G Context-permutation control

Our control permutes the conditioning context across memories within a split. Elapsed time, origin reliability and the validity label are never altered, so only the association between a memory and its transition context is destroyed. We report three permutations: memory type and provenance jointly, proposition text, and both together. Each is run for three seeds on TicToc and on PhysioNet-2012. For each we report the retained advantage over the flexible age-only hazard as a fraction of its unpermuted value, where the advantage is $\mathrm{Brier}(\text{age-only}) - \mathrm{Brier}(\text{TCM, all features})$.

A conditional estimator that merely exploited extra capacity would retain its advantage under permutation, because capacity is unchanged. An estimator that uses the conditioning information should lose it. Retention near zero under the joint permutation therefore supports the mechanism interpretation. Single-channel retention above zero reflects redundancy between explicit type metadata and proposition semantics rather than a failure of the control.

Table 5 reports the matrix. Which channel carries the information differs by domain: on PhysioNet-2012 the proposition text is load-bearing and the type identifier is nearly redundant, while on TicToc the reverse holds. Only destroying both removes nearly all of the advantage. Table 6 repeats the comparison on a third patient cohort held out from fitting and selection. TCM-D reaches 0.1489

Table 7: TCM-D with its time inputs removed, against the age-only rule that sees time alone, five seeds. Bold marks the best value in each column.

	PhysioNet-2012		TicToc	
Inputs	overall	worst group	overall	worst group
Flexible age-only hazard	0.1649	0.2567	0.1763	0.2775
TCM-D, context only	0.1497	0.2256	0.2349	0.2781
TCM-D, no elapsed time	0.1503	**0.2254**	0.2352	0.2777
TCM-D, no policy-visible events	0.1486	0.2258	0.1359	0.2059
TCM-D, all features	0.1499	0.2259	**0.1357**	**0.2046**
TCM (ours)	**0.1477**	0.2256	0.1368	0.2100

Brier and TCM with every feature block 0.1504, against 0.1592 for the flexible typed-age hazard and 0.1654 for the age-only hazard, so the ordering of Table 1 holds on patients no choice was made on.

H Loss weighting across supervision regimes

Under current-status supervision the survival term is the Bernoulli likelihood of $q_{i,t}$, which shares the proper-scoring term's minimizer, so a single weight balances them. Under exact-event supervision the survival term sums over every elapsed interval of an episode and grows with episode length, while the proper-scoring term remains an average over probes. On the controlled clinical simulator the two terms differ by more than an order of magnitude at the weight fixed on the current-status data, so that weight implies a different trade-off than intended. Loss weights therefore do not transfer across supervision regimes, which shows that objective balancing depends on the form of temporal supervision. We instead select the weight on development trajectories and evaluate it on a separately generated cohort. Our loss definition itself is unchanged, and the weight carries the balancing.

I Calibration and ablations in full

What elapsed time adds once context is known. Feature ablations remove context and keep time. Table 7 does the reverse. We refit TCM-D with every conditioning block on the splits and settings of the main runs, once without its elapsed-time input and once without policy-visible events, which carry time indirectly. A third refit drops both. Both studies separate cleanly. On PhysioNet-2012 context alone reaches 0.1497 Brier and adding elapsed time moves TCM-D by at most 0.0011. We therefore read the gain over the age-only rule there as a type-level offset. Dropping the event block, which development data choose for TCM and TCM-D, is worth a further 0.001 to 0.002. On TicToc context alone is worse than age alone, 0.2349 against 0.1763. Only the two together reach 0.1357, so validity there depends on how long a given kind of record has been held rather than on its kind or its age separately.

Table 8 lists every method on PhysioNet-2012 and TicToc, including the elapsed-time extrapolation splits and the feature ablations. Removing the proposition text costs TCM 0.010 Brier on PhysioNet-2012 and nothing on TicToc. With elapsed time as its only input TCM falls back to the age-only hazard on both (0.1646 against 0.1649 and 0.1770 against 0.1763).

Table 8: Every method on PhysioNet-2012 and TicToc, including the elapsed-time extrapolation splits and the feature ablations, five seeds. Bold marks the best value per column within each study.

Method	Brier ↓	Worst-group Brier ↓	Max group gap ↓
PhysioNet-2012, 5 seeds			
Constant base rate	0.1650	0.2690	0.178
Global exponential decay	0.1923	0.2919	0.322
Typed constant decay	0.1717	0.2762	**0.062**
Flexible age-only hazard	0.1649	0.2567	0.185
Flexible typed-age hazard	0.1588	0.2375	0.096
TCM-D (ours)	0.1485	0.2260	0.119
TCM-D, all features	0.1499	0.2259	0.131
TCM (ours)	**0.1477**	0.2256	0.158
TCM, discrete exposure	0.1477	0.2256	0.158

Method	Brier ↓	Worst-group Brier ↓	Max group gap ↓
TCM, all features	0.1497	**0.2253**	0.129
TCM without the Brier term	0.1502	0.2258	0.095
TCM without proposition text	0.1578	0.2371	0.076
TCM without memory type	0.1483	0.2261	0.063
TCM without policy-visible events	0.1477	0.2256	0.158
TCM with elapsed time only	0.1646	0.2613	0.167
PhysioNet-2012, elapsed-time shift, 5 seeds			
Constant base rate	0.1837	0.2852	0.208
Global exponential decay	0.6625	0.8335	0.877
Typed constant decay	0.5083	0.7534	0.808
Flexible age-only hazard	0.1847	**0.2576**	0.220
Flexible typed-age hazard	**0.1809**	0.2586	0.207
TCM-D (ours)	0.1828	0.2763	0.225
TCM-D, all features	0.1821	0.2636	0.191
TCM (ours)	0.1870	0.3896	0.383
TCM, discrete exposure	0.1870	0.3896	0.383
TCM, all features	0.1855	0.2659	0.275
TCM without the Brier term	0.1939	0.3966	0.424
TCM without proposition text	0.1852	0.2717	0.221
TCM without memory type	0.1883	0.3036	0.247
TCM without policy-visible events	0.1870	0.3896	0.383
TCM with elapsed time only	0.1827	0.2671	**0.188**
TicToc, 5 seeds			
Constant base rate	0.2449	0.2794	0.279
Global exponential decay	0.2931	0.5351	0.558
Typed constant decay	0.1617	0.3748	0.362
Flexible age-only hazard	0.1763	0.2775	0.358
Flexible typed-age hazard	0.1368	**0.2028**	0.207
TCM-D (ours)	0.1359	0.2049	0.192
TCM-D, all features	**0.1357**	0.2046	**0.184**
TCM (ours)	0.1368	0.2100	0.197
TCM, discrete exposure	0.1408	0.2128	0.190
TCM, all features	0.1371	0.2075	0.187
TCM without the Brier term	0.1365	0.2123	0.198
TCM without proposition text	0.1366	0.2087	0.191
TCM without memory type	0.1378	0.2093	0.189
TCM without policy-visible events	0.1368	0.2058	0.185
TCM with elapsed time only	0.1770	0.2744	0.349
TicToc, elapsed-time shift, 5 seeds			
Constant base rate	0.4327	0.6475	0.805
Global exponential decay	0.3524	0.7452	0.745
Typed constant decay	0.3524	0.7452	0.745
Flexible age-only hazard	0.3205	0.6900	0.708
Flexible typed-age hazard	0.3184	0.6935	0.711
TCM-D (ours)	0.2654	0.6076	0.645
TCM-D, all features	**0.2446**	**0.5535**	**0.596**
TCM (ours)	0.2601	0.6139	0.637
TCM, discrete exposure	0.3384	0.6475	0.713
TCM, all features	0.2800	0.6650	0.685
TCM without the Brier term	0.2733	0.6654	0.673
TCM without proposition text	0.2780	0.6805	0.707
TCM without memory type	0.2614	0.6432	0.673
TCM without policy-visible events	0.2688	0.6437	0.663
TCM with elapsed time only	0.3101	0.6404	0.662
TicToc, unseen transition families, 5 seeds			
Constant base rate	0.2457	0.2781	**0.230**
Global exponential decay	0.3135	0.5104	0.532
Typed constant decay	0.3135	0.5100	0.532
Flexible age-only hazard	0.1774	**0.2612**	0.293
Flexible typed-age hazard	0.2030	0.3252	0.390
TCM-D (ours)	0.1837	0.3135	0.382
TCM-D, all features	0.1855	0.3019	0.378
TCM (ours)	0.1917	0.4115	0.461
TCM, discrete exposure	0.1928	0.4007	0.449
TCM, all features	0.2004	0.4382	0.491
TCM without the Brier term	0.1988	0.4473	0.511
TCM without proposition text	0.2002	0.4320	0.480
TCM without memory type	**0.1604**	0.2836	0.320
TCM without policy-visible events	0.2038	0.4583	0.510
TCM with elapsed time only	0.1774	0.2651	0.303

Figure 6 sets the estimators of Table 1 side by side and adds TicToc's unseen transition families. In distribution the conditioned estimators and the flexible typed-age hazard stay close, and typed constant decay trails on every study. On unseen families typed constant decay degrades sharply to

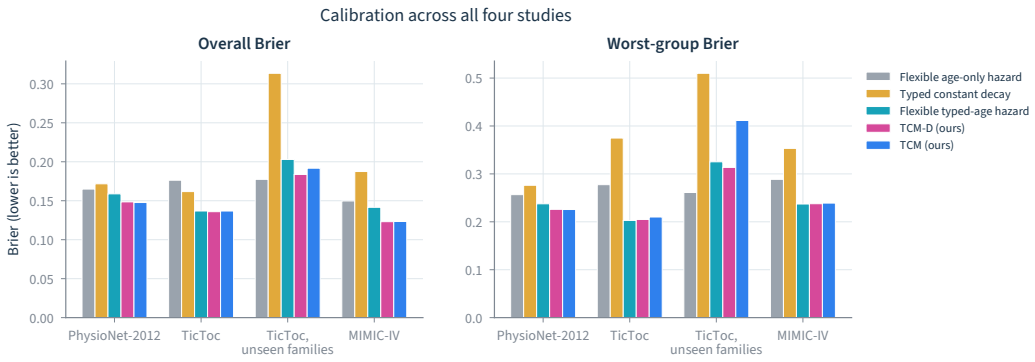

Figure 6: Overall and worst-group calibration across all four studies. Typed constant decay degrades sharply where memory types do not transfer, whereas the flexible and conditioned estimators stay close together.

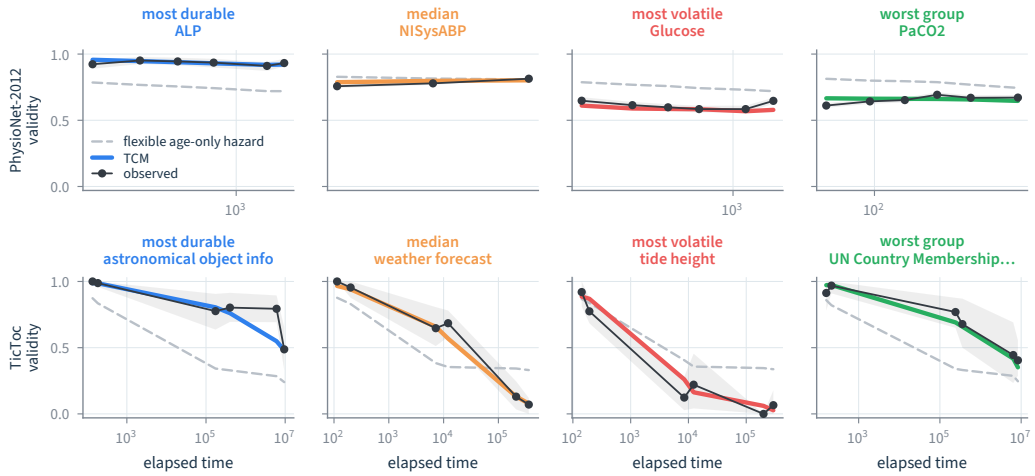

Figure 7: Validity against elapsed time on PhysioNet-2012 and TicToc. Fixed rules pick the most durable, median and most volatile memory type and the one TCM fits worst. On PhysioNet-2012 a stored measurement is probed at the next measurement of the same concept, so the observed elapsed range is short and validity is nearly flat across it. On that dataset the flexible age-only hazard errs in *level*, sitting above observed validity on glucose (0.75 against 0.61) and below it on ALP (0.75 against 0.93). TCM removes that offset. On TicToc the observed curves differ in *shape* as well. TCM follows them, predicting 0.76 on the most durable type where 0.80 is observed and 0.37 on the most volatile where 0.34 is, while the age-only hazard gives 0.48 and 0.52. Intervals are entity-level bootstraps over five pooled seeds.

a worst group of 0.51 because it has no rate for a type it never saw. Only there does the flexible age-only hazard reach the best worst group (0.26), one boundary Section 6 states.

Figure 7 follows four memory types per dataset over elapsed time: the most durable, the median, the most volatile and the one TCM fits worst. On PhysioNet-2012 the age-only hazard errs in level, above observed validity on glucose and below it on ALP, and TCM removes that offset. On TicToc the curves also differ in shape. TCM follows each of them, whereas the age-only hazard draws one curve for every type.

Figure 8 draws the group calibration of Figure 3 as predicted against observed validity, one point per memory type and elapsed-time stratum. Age-only predictions place nearly every group in one narrow band although observed validity runs from near zero to one, the miscalibration Proposition 1 predicts. TCM spreads its predictions along the diagonal and cuts the maximum group gap from

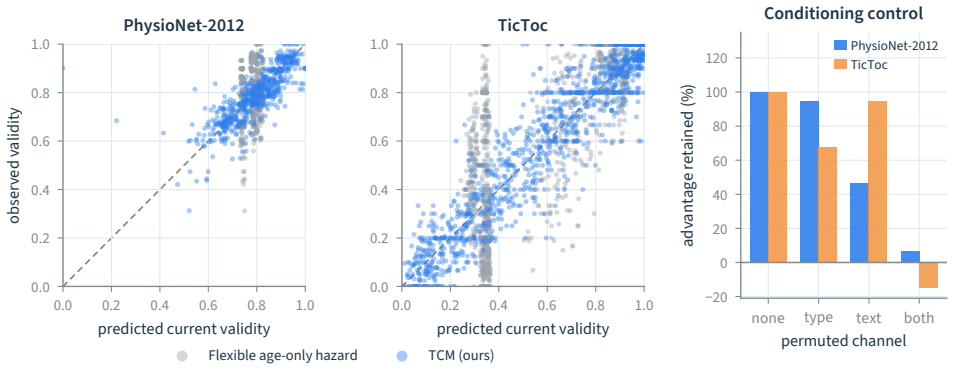

Figure 8: Group calibration of Figure 3, drawn as predicted against observed validity with one point per elapsed-time stratum and memory type. Predictions of the flexible age-only hazard collapse onto a narrow band while observed validity spans the full range. Conditioning cuts the maximum group-calibration gap from 0.185 to 0.158 on PhysioNet-2012 and from 0.358 to 0.197 on TicToc. Right, the advantage retained when a conditioning channel is permuted (Table 5).

0.185 to 0.158 on PhysioNet-2012 and from 0.358 to 0.197 on TicToc. Figure 8's right panel repeats the permutation control of Table 5.

J BASELINE APPLICABILITY

Retrieval and current-validity estimation are orthogonal axes. A retriever chooses which records reach the policy, and a validity estimator says whether a chosen record still holds. Every current-validity comparison here probes a given stored record, so it varies only the second axis and the baseline set spans that axis. Where a published system answers a different question than the use, verify or abstain decision, we state its status below rather than force it into Table 1.

- **Full context** and **semantic retrieval** vary the first axis and are held fixed here: every estimator sees the same retrieved record.
- **Tuned global exponential decay** and **typed constant decay**: the primary decay comparators, fitted by exhaustive one-dimensional search rather than gradient descent, since the exponent saturates on elapsed times spanning six orders of magnitude.
- **TicToc DPO** (Rafailov et al., 2023): the direct post-training baseline for the re-query decision, trained and evaluated with the released scripts and reported in Figure 5 and Table 20. DPO is the sharpest external control because it optimizes the decision directly rather than through a validity belief.
- **Memory-T1 / Time-Dialog**: a time-aware evidence selector for questions about closed past intervals. We score its three released Time-Dialog outputs with the authors' metric. Full context, BM25 retrieval and BM25 with local inference reach exact-match accuracy of 0.250, 0.120 and 0.320 over 200 questions. Time-Dialog questions ask what happened in a past window, so the task is retention rather than current validity. We do not re-train the method itself.
- **BeliefMem**: conceptually the closest alternative, because it maintains a probabilistic belief over competing conclusions whereas we estimate the probability that a stored proposition remains current. No official implementation was identified during the resource survey, so a comparison would require an independent reproduction. BeliefMem is the most informative comparison remaining for future work.
- **Temporal Semantic Memory** and **PAMU**: no official implementation was identified, and their estimands differ from current validity. Temporal Semantic Memory targets durative temporal retrieval and PAMU targets preference update, neither of which is a use, verify or abstain decision over a stored proposition whose referenced state may have moved.

Table 9: Every base estimator under every calibrator: Brier overall and on the worst group, and utility at a matched 20% verification budget, over five seeds and fifteen for MIMIC-IV. Per-type maps are our group-aware null, not a published method. Bold marks the best unrounded value in each column.

	PhysioNet-2012			TicToc			MIMIC-IV		
Method	Brier	worst	utility	Brier	worst	utility	Brier	worst	utility
Flexible age-only hazard, uncalibrated	0.1649	0.2566	+0.4305	0.1763	0.2774	+0.1924	0.1497	0.2887	+0.4661
Flexible age-only hazard, temperature scaling	0.1649	0.2567	+0.4305	0.1766	0.2783	+0.1924	0.1497	0.2886	+0.4661
Flexible age-only hazard, Platt scaling	0.1647	0.2620	+0.4305	0.1766	0.2767	+0.1924	0.1487	0.2849	+0.4933
Flexible age-only hazard, isotonic regression	0.1646	0.2609	+0.4303	0.1778	0.2864	+0.1901	0.1488	0.2862	+0.4727
Flexible age-only hazard, histogram binning	0.1645	0.2615	+0.4300	0.1807	0.2823	+0.1750	0.1485	0.2836	+0.4908
Flexible age-only hazard, Platt, per memory type	0.1588	0.2380	+0.4818	0.1767	0.2767	+0.1921	0.1412	0.2404	+0.5568
Flexible age-only hazard, isotonic, per memory type	0.1590	0.2381	+0.4802	0.1778	0.2864	+0.1897	0.1418	0.2433	+0.5508
TCM-D (ours), uncalibrated	0.1485	0.2258	+0.5310	**0.1361**	**0.2099**	+0.2793	0.1230	**0.2372**	**+0.6476**
TCM-D (ours), temperature scaling	0.1485	0.2257	+0.5310	0.1361	0.2110	+0.2793	**0.1230**	0.2372	+0.6476
TCM-D (ours), Platt scaling	0.1485	0.2257	+0.5310	0.1364	0.2100	+0.2812	0.1230	0.2372	+0.6476
TCM-D (ours), isotonic regression	0.1484	0.2258	+0.5311	0.1370	0.2130	+0.2806	0.1231	0.2380	+0.6473
TCM-D (ours), histogram binning	0.1491	0.2262	+0.5311	0.1373	0.2109	+0.2778	0.1233	0.2380	+0.6468
TCM-D (ours), Platt, per memory type	0.1479	**0.2254**	+0.5329	0.1365	0.2100	+0.2806	0.1231	0.2395	+0.6460
TCM-D (ours), isotonic, per memory type	0.1481	0.2258	+0.5332	0.1372	0.2130	+0.2803	0.1244	0.2466	+0.6442
TCM (ours), uncalibrated	0.1477	0.2256	+0.5367	0.1366	0.2105	**+0.2815**	0.1236	0.2394	+0.6445
TCM (ours), temperature scaling	0.1475	0.2257	+0.5367	0.1367	0.2107	+0.2815	0.1235	0.2395	+0.6445
TCM (ours), Platt scaling	0.1475	0.2258	+0.5367	0.1366	0.2137	+0.2807	0.1235	0.2395	+0.6445
TCM (ours), isotonic regression	0.1473	0.2258	+0.5368	0.1374	0.2170	+0.2784	0.1235	0.2392	+0.6448
TCM (ours), histogram binning	0.1481	0.2298	+0.5367	0.1377	0.2153	+0.2760	0.1239	0.2407	+0.6436
TCM (ours), Platt, per memory type	0.1467	0.2255	+0.5396	0.1367	0.2137	+0.2805	0.1234	0.2393	+0.6443
TCM (ours), isotonic, per memory type	**0.1464**	0.2266	**+0.5416**	0.1374	0.2170	+0.2784	0.1243	0.2453	+0.6448

Adapting either to this interface would report a variant under that method's name rather than the method itself.

Every independent reproduction is labeled as such and never described as an official result.

K LITERATURE BASELINES IN FULL

Post-hoc calibration. We rebuild each finished run from its recorded configuration. A rebuild is accepted only if its test arrays match the saved ones element by element. Both deterministic grid-search baselines must also reproduce the saved predictions within an absolute tolerance of 10^{-6}, which pins the training and development splits as well. We then halve the development entities at random. One half early-stops the refitted base estimators, the other fits the calibrators, and every estimator keeps the configuration its original run selected. Every refit reproduces its original test Brier within 0.002, against an enforced tolerance of 0.01, and overall-Brier seed means agree within 0.0005. Temperature, Platt and isotonic calibration are scikit-learn's own implementations applied to the frozen score, and histogram binning uses fifteen equal-mass bins. Per-type calibrators need 100 development observations with both outcomes, and a type below that falls back to the global map. On TicToc no type qualifies in four of five seeds and two in the fifth, so the per-type maps coincide with the global ones there.

Recalibration does not separate TCM and TCM-D either. Against the best-recalibrated TCM-D, refitted TCM is 0.0003 Brier ahead on PhysioNet-2012 ($p = 0.16$), 0.0005 behind on TicToc ($p = 0.37$) and 0.0006 behind on MIMIC-IV ($p = 0.09$). TCM itself gains at most 0.0013 on average from recalibration. Table 9 lists every base estimator under every calibrator, and Table 10 condenses what the calibrators do. Global maps close at most 4.5% of the age-only hazard's gap to TCM-D. Per-type maps close about a third of it on PhysioNet-2012 and MIMIC-IV (37% and 32%) and none on TicToc, and no calibrator moves TCM-D by more than 0.0014 Brier. Table 10's bottom block repeats Table 5.

Survival baselines. We fit every library survival model on the same 191 covariates: reference-coded memory type and provenance and the 64-dimensional text embedding. Pivoted QR on the training episodes reduces them to full rank, and durations are in days. TCM and the type-aware rules read the same type, provenance and text through their own feature path. We use the lifelines implementation of Cox proportional hazards (Davidson-Pilon, 2019). Development-patient episodes choose its ridge penalty from $\{0, 0.001, 0.01, 0.1\}$ by partial log-likelihood. Without a penalty, lifelines halts with a matrix-inversion error in every seed. Development selects 0.1, the largest value in the grid, in all

Table 10: Two controls on the conditioning gain, five seeds and fifteen for MIMIC-IV. Top: what each calibrator does to TCM-D, and the share of the age-only rule's gap to it that the calibrators close. All calibrators are fitted on a development half disjoint from early stopping, with per-type maps as our group-aware null. Bottom: the advantage left when a conditioning channel is permuted, three seeds.

	PhysioNet-2012		TicToc		MIMIC-IV	
Calibrator	ΔBrier	gap closed (%)	ΔBrier	gap closed (%)	ΔBrier	gap closed (%)
Recalibrating TCM-D, and the age-only gap it closes						
temp.	0.0000	0.0	0.0000	−0.7	0.0000	−0.1
Platt	0.0000	1.4	+0.0003	−0.9	0.0000	3.9
isotonic	−0.0002	2.0	+0.0009	−3.6	+0.0001	3.2
histogram	+0.0006	2.8	+0.0012	−11.0	+0.0003	4.5
Platt, per type	−0.0006	37.2	+0.0004	−1.0	+0.0001	31.9
isotonic, per type	−0.0004	36.1	+0.0011	−3.8	+0.0014	29.7
Advantage over age-only retained when the context is permuted (%)						
permuted type		95		67		–
permuted text		47		95		–
permuted both		7		−15		–

Table 11: Probe-time and survival scores on the simulator, five seeds. Every row is refitted in this study, and the last two rows also see policy-visible events. TCM-D + events is averaged over four seeds. In the fifth, its refit did not reproduce the original run (Brier 0.054 against 0.025) and is excluded.

Method	Brier ↓	Worst group ↓	IBS ↓	C-index ↑
Cox PH (ridge)	0.0957	0.5373	0.0373	0.780
Logistic hazard	0.1066	0.2139	0.0314	0.787
Survival forest	0.0832	0.8033	0.0172	0.792
DeepHit	0.0960	0.5993	0.0995	0.784
DeepSurv	0.0905	0.5666	**0.0162**	0.793
Flexible typed-age hazard	0.0766	**0.2032**	0.1059	0.763
TCM-D (ours)	0.0764	0.5779	0.1150	0.730
TCM (ours)	0.0760	0.2870	0.0628	0.780
TCM-D (ours) + events	0.0273	1.0000	0.0471	0.926
TCM (ours) + events	**0.0235**	0.4336	0.0317	**0.942**

Table 12: Survival-native scores on the simulator's test episodes, five seeds and four for TCM-D + events. Every row is refitted in this study.

Method	Event NLL ↓	IPCW IBS ↓	Harrell C ↑	D-cal. max dev. ↓	Max group gap ↓
Cox PH (ridge)	1.578	0.0373	0.780	0.057	0.562
Logistic hazard	1.529	0.0314	0.787	0.094	0.251
Survival forest	1.351	0.0172	0.792	0.061	0.896
DeepHit	2.102	0.0995	0.784	0.120	0.662
DeepSurv	1.557	**0.0162**	0.793	**0.049**	0.614
Typed constant decay	2.356	0.0403	0.767	0.197	0.229
Flexible typed-age hazard	5.236	0.1059	0.763	0.116	**0.208**
TCM-D (ours)	4.216	0.1150	0.730	0.150	0.638
TCM (ours)	2.773	0.0628	0.780	0.098	0.323
TCM-D (ours) + events	1.190	0.0471	0.926	0.279	1.000
TCM (ours) + events	**0.938**	0.0317	**0.942**	0.119	0.524

Table 13: Cox proportional hazards on full training data at every ridge penalty of the development grid, five seeds. Scores of unselected penalties are computed for this table only.

Training data	Penalty	Dev. log-lik. ↑	Brier ↓	Worst group ↓	IBS ↓	Selected
full	0 (5 failed)	–	–	–	–	0/5
	0.001	−7.655	0.1174	1.0000	0.1032	0/5
	0.01	−13.365	0.1244	0.7379	0.5283	0/5
	0.1	−7.069	0.0957	0.5373	0.0373	5/5

Table 14: Survival models and ours refitted on the same 60 000 training episodes the survival forest is fitted on, five seeds.

Method	Brier ↓	Worst group ↓	IBS ↓
TCM (ours)	**0.0756**	0.2553	0.0628
TCM-D (ours)	0.0825	1.0000	0.1592
Survival forest	0.0832	0.8033	**0.0172**
DeepSurv	0.0911	0.5960	0.0181
Logistic hazard	0.1068	**0.2101**	0.0312

five seeds (Table 13). Our discrete-time logistic hazard is an unpenalized logistic regression on the person-period table of TCM's own grid, converging in 62 to 75 iterations. We use the scikit-survival implementation of the random survival forest (Pölsterl, 2020) and fit it on 60 000 episodes sampled at random. A minimum leaf size of 50 bounds memory, since each leaf stores a survival curve over the 9 615 to 9 960 distinct times of its seed. DeepHit and DeepSurv (Katzman et al., 2018) are the pycox implementations (Kvamme et al., 2019) in the configuration of its reference examples, early-stopped on development-patient episodes after 19 to 54 epochs. NeuralSurv runs its authors' code, environment and real-data configuration on 125 training episodes, the subsample size of the authors' own real-data experiments. NeuralSurv's variational update grows with episodes times quadrature points times parameters and exceeds 96 GB of memory at 1 000 episodes. We compute its posterior-median survival curves once per distinct covariate row and interpolate them at each probe time on a 200-quantile grid. Test episodes (181 487) contain 515 distinct rows and their probes (543 465) contain 506. Matrix products run at full float32 precision, because the reduced TF32 precision that GPUs default to moves probe-time validity by up to 1.0 on this design. In every seed, the variational fit spends the authors' full budget of 200 coordinate-ascent iterations without reaching their convergence tolerance.

Probe-time scores use the paper's metrics on every test probe. We score survival curves on the test episodes. Discrete event negative log-likelihood uses TCM's grid over all test episodes. IPCW integrated Brier score (Graf et al., 1999), Harrell's concordance (Harrell et al., 1982) and D-calibration (Haider et al., 2020) come from SurvivalEVAL (Qi et al., 2024) on a fixed random subset of 10 000 episodes per seed. At that size every model's D-calibration test rejects uniformity ($p < 0.001$), so Table 12 reports the largest deviation of a bin from its uniform share of 0.1 instead.

Table 11 extends Table 2 with worst-group Brier and concordance. With policy-visible events TCM has the lowest probe-time Brier (0.0235) and the highest concordance (0.942), while DeepSurv and the survival forest keep the lowest integrated Brier. Table 12 scores the same curves as survival models. TCM with events has the lowest event negative log-likelihood (0.938), and DeepSurv has the smallest D-calibration deviation (0.049), so each family wins the score its objective targets. At the penalties development did not choose, Cox PH is worse on every score (Table 13).

Matched large-data training. Each survival-forest leaf stores a survival curve, so we fit the forest on 60 000 training episodes. Table 14 refits the comparators on exactly those episodes. TCM keeps the lowest probe-time Brier at 0.0756, and a survival-literature model keeps the lowest integrated Brier, now the survival forest at 0.0172. Hence the dissociation is not an artifact of the forest seeing fewer episodes than the models it is compared with.

Matched training data. Table 15 refits every model on the 125 training episodes NeuralSurv uses, drawn by the same seeded rule. Each model is then scored on the unchanged test episodes. We rebuild the time grid from the retained probes, and the survival forest uses scikit-survival's default minimum leaf size of three. Cox PH fails to converge at every penalty in all five seeds, and typed constant decay is not refitted at this size. TCM, TCM-D and the survival forest lie within 0.008 Brier of one another and of DeepSurv and NeuralSurv. TCM and the survival forest are indistinguishable ($p = 0.88$). Among these five, only the forest's lead over DeepSurv and NeuralSurv separates from zero ($p = 0.018$ and $p = 0.037$). In every seed, TCM leads the flexible typed-age hazard, the logistic hazard and DeepHit. On integrated Brier score the survival forest is lowest, 0.084 against 0.290 for TCM.

Table 15: Probe-time and survival scores with every model refitted on the same 125 training episodes per seed, five seeds. Cox PH fails to converge at every penalty and has no row.

Method	Brier ↓	Worst group ↓	IBS ↓	C-index ↑
Logistic hazard	0.2322	1.0000	0.1152	0.624
Survival forest	**0.1154**	0.9865	**0.0842**	**0.705**
DeepHit	0.2413	0.9177	0.1409	0.591
DeepSurv	0.1218	0.9999	0.0980	0.679
NeuralSurv	0.1231	0.9048	0.1210	0.626
Flexible typed-age hazard	0.1529	**0.6719**	0.1712	0.595
TCM-D (ours)	0.1208	0.9803	0.5749	0.583
TCM (ours)	0.1160	0.9988	0.2898	0.617

L FULL EXPERIMENTAL PROTOCOL

Records and labels. A memory is an immutable record of a proposition, its observed value and the time of observation. Each record also names its transition family as a memory type and carries a provenance tier. Two properties of this estimand bound what any result about it can mean. First, a survival label marks a memory as superseded even when the world later returns to the same value, which is conservative for reversible processes. Second, we assume the type schema is stable across training and deployment. On the re-query data the memory is the stored tool observation, and the label is an annotator's judgment of whether it should be re-queried. We score votes individually, so that a Brier score means the same thing as it does on binary clinical outcomes. On the clinical cohorts the memory is a measurement and the label is whether the next observed measurement of that concept remains in the same predefined band. Bands come from standard adult reference ranges fixed before any transition frequency was inspected, and sensitivity to that choice is reported below. Figure 1 displays two seed-3 instances from the paper's entity-level test split that share elapsed time and origin reliability. We selected the displayed pair post hoc for illustration and used it for no model, hyperparameter or threshold choice and no aggregate result.

Example 1: two clinical memories of the same age

	Alkaline phosphatase	**Serum glucose**
Stored record	ALP 24 IU/L, low band	Glucose 206 mg/dL, high band
Age at probe	255 min	251 min
Next measurement	same band, valid	band left, superseded
Flexible age-only hazard	0.78, VERIFY ✗	0.78, VERIFY ✓
TCM (ours)	0.96, USE ✓	0.35, VERIFY ✓

✓ agrees with the label, ✗ does not. Actions for q use the balanced threshold $\theta = 0.85$.

Example 1 shows two PhysioNet-2012 test records stored about four hours before their probe. Age-only confidence is 0.78 for both, so that rule verifies both. TCM separates a slowly drifting enzyme from serum glucose, reusing the first and verifying the second as the next measurements require. We chose this pair and that of Example 2 post hoc in the same way, for illustration only.

Splits and resampling. We split by episode or patient, never by admission or by individual probe. All elapsed-time variants of one scenario stay in the same fold, so the controlled intervention is never split across folds. Confidence intervals on differences come from a paired bootstrap that resamples entities (Efron & Tibshirani, 1993). Holm correction (Holm, 1979) applies to the three primary comparisons.

Training protocol. Feature standardization uses training statistics only. We select the conditioning configuration on the development split from a fixed candidate grid and likewise fix the loss weight on development data. Both are training protocol rather than method components, and are recorded here for reproducibility.

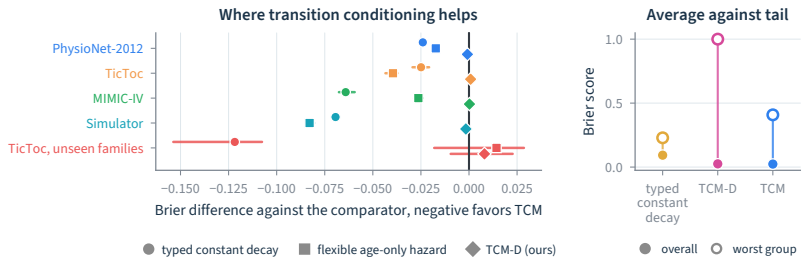

Figure 9: **Conditioning beats decay rules in overall Brier in every regime.** Left, paired Brier differences, negative favoring TCM. Right, average against tail. Five seeds, fifteen for MIMIC-IV.

Table 16: Freshly generated test cohort from the simulator, with the loss weight fixed on development trajectories, five seeds. TCM and TCM-D improve overall Brier over typed constant decay by a wide margin, and with censored episodes both are worse than it on the worst group. Bold marks the best value in each column and block.

Supervision	Method	Brier ↓	Worst-group Brier ↓
with censored episodes	Typed constant decay	0.0930	0.2288
	Flexible typed-age hazard	0.0766	**0.2032**
	TCM-D (ours)	0.0253	1.0000
	TCM (ours)	**0.0236**	0.4084
identical labeled data	Typed constant decay	0.0930	0.2288
	Flexible typed-age hazard	0.0751	0.1965
	TCM-D (ours)	**0.0239**	**0.1918**
	TCM (ours)	0.0241	0.2788

Informative observation. Clinicians order tests when they suspect change, so probe times in the clinical studies are not random. Our estimand is validity at the next observed measurement, not latent biological survival, and we do not claim otherwise.

M SCOPE STUDIES IN FULL

Figure 9 collects the in-distribution studies, the simulator and TicToc's unseen families. In the left panel TCM beats typed constant decay in every regime and the flexible age-only hazard in distribution, and it falls behind the age-only hazard only on TicToc's unseen families.

Exact event-time supervision. Our simulator provides exact transition times, with about a fifth of records right-censored. Only in this setting is the discrete survival likelihood exercised with real per-interval event indicators. We run the study twice. In one run survival models may additionally consume censored episodes, which a discriminative predictor structurally cannot use. In the other, every method sees identical labeled data. Matched data isolate the inductive structure, and the gap between the two conditions measures the value of consuming incomplete episodes at all.

Exact event times. Given identical labeled data, TCM and TCM-D are indistinguishable overall on the simulator's freshly generated test cohort (0.0241 for TCM against 0.0239), and TCM-D has the better worst group (0.192 against 0.279). In the original runs (Table 16) the censored episodes that only TCM can consume put TCM ahead by 0.0016 in four of five seeds, with a worst group of 0.41 against 1.00 in all five. Figure 9's right panel prices what flexibility costs. TCM improves overall Brier over typed constant decay (0.0236 against 0.0930) while its worst-group score is worse (0.4084 against 0.2288), so selecting on overall development score buys average fit at the cost of the tail.

We bound two failures here rather than explain them, the tail loss above and the time-shift degradation below. Two questions remain. Does the tail loss come from selecting on aggregate score or from

Table 17: Support-aware fallback, five seeds. Our support test fires on under 0.1% of in-distribution probes, leaving the estimate there unchanged to four decimals. In the three shifted regimes it fires on every probe and recovers the age-only result. A stricter first-to-ninety-ninth percentile band fires on 1.1% of PhysioNet-2012 probes and moves overall Brier from 0.1477 to 0.1480.

		Brier		Worst group		Utility
Study	fires on	TCM	fallback	TCM	fallback	TCM → fallback
PhysioNet-2012	0.000	0.1477	0.1477	0.2256	0.2256	$0.6141 \rightarrow 0.6141$
TicToc	0.000	0.1368	0.1368	0.2100	0.2100	$0.6335 \rightarrow 0.6335$
TicToc, unseen families	1.000	0.1917	**0.1777**	0.4115	**0.2602**	$0.5867 \rightarrow 0.5661$
PhysioNet-2012, time shift	1.000	0.1870	**0.1847**	0.3896	**0.2576**	$0.5301 \rightarrow$ **0.5500**
TicToc, time shift	1.000	0.2601	**0.2283**	0.6139	**0.3543**	$0.4435 \rightarrow$ **0.5500**

shared-parameter estimation as such? Does the time-shift degradation lie in the hazard network or in the feature scaling? Each needs its own control, which we leave to future work.

Support-aware fallback. Whatever their causes, both failures above share a condition the policy can detect: the question lies outside the support the estimator was trained on. We define support from the training split as the set of memory types seen and the range of elapsed times observed. Out-of-support probes go to the flexible age-only hazard with discrete exposure, which holds its last value past the trained range. Nothing is retrained. Each routed estimate is a per-probe choice between two prediction vectors. We constructed the rule after seeing the scope results and report it as post hoc. Table 17 reports the outcome. In distribution the test fires on under 0.1% of probes and leaves every score unchanged. In the three shifted regimes it fires on every probe and cuts worst-group Brier by a third or more, from 0.4115 to 0.2602 on unseen families. Utility rises under both time shifts (0.5301 to 0.5500 and 0.4435 to 0.5500) and falls on unseen families (0.5867 to 0.5661), where the fallback buys calibration with utility.

Elapsed-time extrapolation. Training on short elapsed gaps and evaluating on long ones separates extrapolation along the age axis from interpolation. Parametric decay rules extrapolate to zero validity far outside their fitted range and fail catastrophically, reaching Brier 0.66 and 0.51 on PhysioNet-2012. Every flexible model including ours degrades gracefully on those records and holds between 0.181 and 0.187. On TicToc the choice of exposure decides the outcome, because development probes lie inside the trained range and cannot see the tail. Development data there give every hazard model continuous exposure, which carries the last interval's rate past the grid. With it the pooled and typed age-only hazards reach 0.3205 and 0.3184, against 0.2283 and 0.2412 with discrete exposure, which holds the last value. TCM reaches 0.2601 with continuous exposure against 0.3384 with discrete, and TCM-D 0.2654. Our extrapolation study therefore bounds the parametric family and marks extrapolation along the age axis as the regime where no estimator here is reliable and the fallback above is the safer rule.

Sensitivity. Three sensitivities have a scientific reading and are reported in the main text or here. Elapsed-time bin resolution controls how finely the hazard can vary with age. Loss weight, which balances the survival and proper-scoring terms, is nearly inert under current-status supervision and decisive under event-time supervision. Verification budget and cost profile form the decision study's own sweep rather than a hyperparameter. Hazard network width, embedding width and the clinical band definition are reported for completeness, and leave the ordering of methods unchanged with one exception. On TicToc the widest hazard network is worse in each of ten seeds, a median of 0.1462 against 0.1379 at width 128 and no divergent seed, while TCM-D is unaffected by width. With a few thousand probes the widest hazard network overfits, and no narrower width and no clinical run shows it.

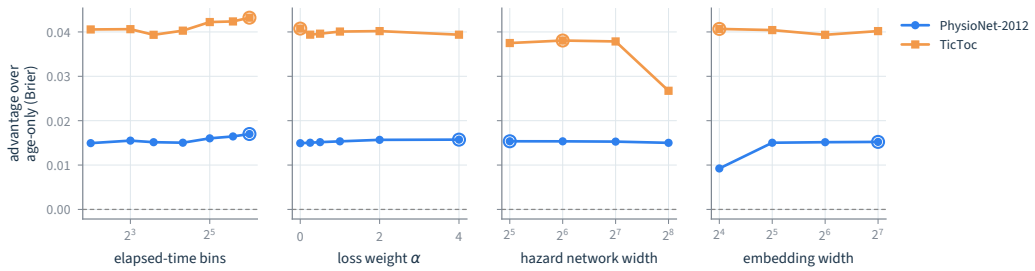

Figure 10: Sensitivity of the transition-conditioned gain. Curves show the advantage over the age-only hazard, which stays comparable when a swept factor changes the validity definition. Circles mark the best value of each sweep. Elapsed-time resolution and the loss weight are nearly inert under current-status supervision, which is why the weight fixed there transferred badly to exact event-time supervision.

Table 18: Sensitivity sweeps, three seeds per cell and ten for hazard network width. Our reported configuration uses $\alpha = 1$, twelve bins, hidden width 128 and embedding width 64 with the band label rule. Every cell uses all conditioning blocks, with exposure chosen on development.

Sweep	Value	PhysioNet-2012		TicToc	
		Brier	vs age-only	Brier	vs age-only
bins	4	0.1497	+0.0149	0.1372	+0.0406
	8	0.1492	+0.0155	0.1369	+0.0406
	12	0.1496	+0.0152	0.1379	+0.0394
	20	0.1501	+0.0150	0.1373	+0.0403
	32	0.1494	+0.0160	0.1363	+0.0422
	48	0.1496	+0.0164	0.1368	+0.0424
	64	0.1495	+0.0170	0.1371	+0.0432
alpha	0.0	0.1500	+0.0149	0.1367	+0.0408
	0.25	0.1498	+0.0150	0.1380	+0.0394
	0.5	0.1497	+0.0152	0.1377	+0.0396
	1.0	0.1495	+0.0153	0.1372	+0.0401
	2.0	0.1490	+0.0157	0.1369	+0.0402
	4.0	0.1489	+0.0157	0.1376	+0.0394
hidden	32	0.1496	+0.0154	0.1379	+0.0375
	64	0.1496	+0.0153	0.1374	+0.0381
	128	0.1497	+0.0153	0.1376	+0.0379
	256	0.1499	+0.0150	0.1487	+0.0267
textdim	16	0.1556	+0.0092	0.1366	+0.0407
	32	0.1498	+0.0150	0.1368	+0.0404
	64	0.1496	+0.0152	0.1379	+0.0394
	128	0.1496	+0.0152	0.1371	+0.0402
labelrule	absolute0.25	0.1874	+0.0334	–	–
	band0.10	0.1496	+0.0152	–	–
	relative0.05	0.1775	+0.0678	–	–
	relative0.10	0.1778	+0.0443	–	–
	relative0.20	0.1146	+0.0216	–	–

N SENSITIVITY SWEEPS

Table 18 and Figure 10 report every swept configuration. Absolute Brier is not comparable across label definitions, which change the base rate. We therefore report the advantage over the age-only hazard alongside it, and that advantage is the quantity to read for the label-rule sweep.

Our sweep supports three readings. Elapsed-time resolution is nearly inert on both datasets. With the continuous exposure that development data choose on TicToc, TCM moves by at most 0.0016 Brier between four and sixty-four bins, so the discretization that discrete exposure imposed there is gone. Loss weight is almost inert under current-status supervision across two orders of magnitude, which is exactly why the weight fixed there transferred so badly to exact event-time supervision. TCM's advantage over the age-only hazard also stays positive under every clinical validity definition tried, so our conclusion does not depend on the choice of band cut-points.

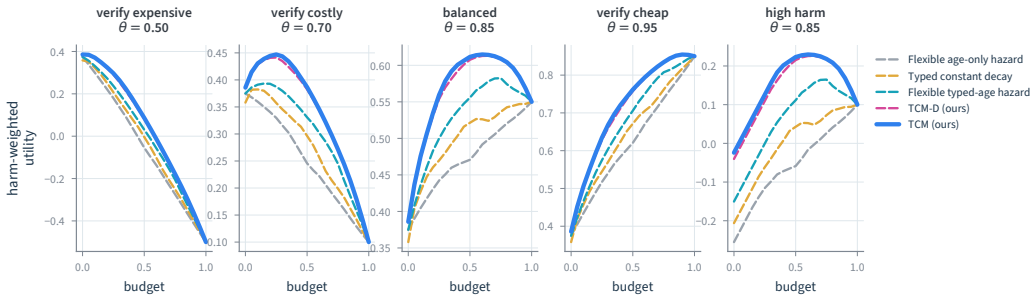

Figure 11: Utility against verification budget on PhysioNet-2012 under each cost profile, with the implied verification threshold θ above every panel. TCM and TCM-D lead every baseline under every profile, and what changes is how much a miscalibrated belief costs.

Table 19: Utility under the unconstrained cost-optimal rule on PhysioNet-2012, five predefined cost profiles and five seeds. Implied verification thresholds are $\theta = 0.50$ for verify expensive and 0.70 for verify costly. Balanced uses 0.85, high harm 0.85 with $c_{\text{stale}} = 5$, and verify cheap 0.95. Figure 11 and the released-budget panel of Figure 5 use the same names and thresholds. Bold marks the highest utility under each profile.

Method	verify expensive	verify costly	balanced	verify cheap	high harm
Flexible age-only hazard	+0.3749	+0.3749	+0.5500	+0.8500	+0.1000
Flexible typed-age hazard	+0.3749	+0.3938	+0.5832	+0.8500	+0.1664
TCM-D (ours)	**+0.3859**	+0.4412	+0.6132	**+0.8548**	+0.2264
TCM (ours)	+0.3858	**+0.4465**	**+0.6141**	+0.8539	**+0.2282**

Hazard network width and embedding width change nothing of consequence, except that the widest hazard network overfits on TicToc and a sixteen-dimensional embedding costs 0.006 Brier on PhysioNet-2012. No sweep was used to select the reported configuration, and we report all of them as sensitivity after the fact.

O DECISION UTILITY ACROSS ALL COST PROFILES

Figure 11 and Table 19 gives the cost-optimal utility under all five profiles, which span implied verification thresholds θ from 0.50 to 0.95. TCM and TCM-D lead every baseline under every profile. What changes is how much a miscalibrated belief costs, which peaks at intermediate verification cost and grows with stale-use harm. Against the flexible typed-age hazard TCM gains 0.053 under verify costly and 0.062 under high harm but only 0.004 under verify cheap.

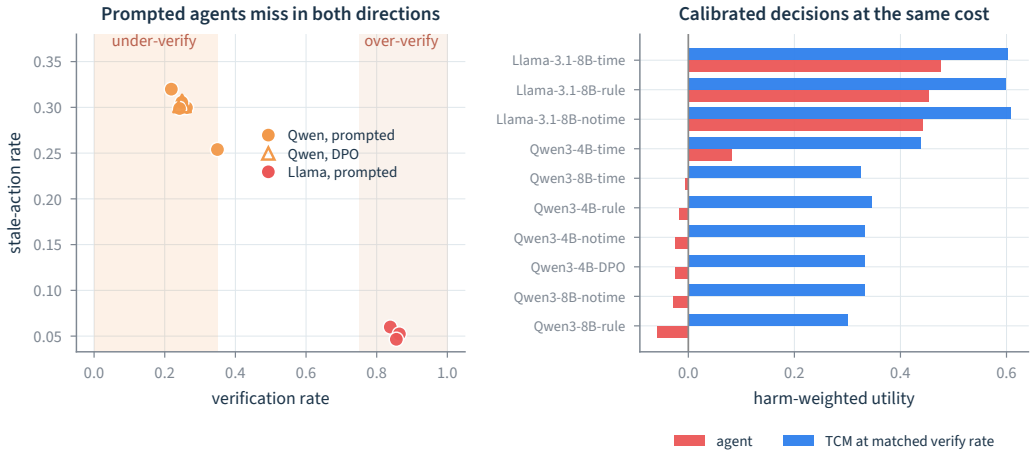

Figure 12: Left, verification rate against stale-action rate for every prompted agent, with the two shaded regions marking under- and over-verification. Right, harm-weighted utility of each agent against TCM evaluated at that agent's own verification rate. Post-training on the re-query preference moves neither.

P External agent baselines in full

Example 2: same age, four deciders

	Live game score	**Conservation status**
Stored observation	game: Warriors vs. Celtics, score: Warriors 75, Celtics 80, quarter: 3rd, time left: 6:05	species: Snow Leopard, status: Vulnerable
Pending request	If the Warriors score 10 more points from what they have, what will the new score difference be from the Celtics in the ongoing game?	What is the conservation status of Snow Leopard?
Age at probe	174 s	196 s
Annotator votes	0/5 reuse	5/5 reuse
Qwen3-8B, prompted	USE ✗	USE ✓
Llama-3.1-8B, prompted	VERIFY ✓	VERIFY ✗
Flexible age-only hazard	0.84, VERIFY ✓	0.83, VERIFY ✗
TCM (ours)	0.45, VERIFY ✓	0.98, USE ✓

✓ agrees with the label, ✗ does not. Actions for q use the balanced threshold $\theta = 0.85$.

Example 2 shows two re-query probes from the seed-3 test split, both about three minutes old. Qwen3-8B uses both records and Llama-3.1-8B verifies both, so neither agent's decision follows the record. Only TCM verifies the live score and reuses the conservation status, as the annotators judged. Actions for q follow the balanced threshold $\theta = 0.85$ rather than the matched rates of Table 20.

Figure 12 summarizes the agents. We run all ten agent baselines (Grattafiori et al., 2024; Yang et al., 2025) on the same decision problem through the benchmark's released evaluation code and prompt templates. Generation runs through `transformers` (Wolf et al., 2020) in place of the released vLLM backend, with every prompt and decoding setting unchanged. TCM is evaluated in the final column at the verification *rate* each agent realized, not merely at the same per-verification price, so no method wins by verifying more often. A prompted agent chooses between using the memory and re-querying, so the main text scores TCM over those two actions as well. TCM's three-action rule, which may also abstain, fills the column here. Against Qwen it abstains on 8.5% to 21.5% of probes and cuts its stale rate from 0.24 to 0.09. Against Llama, which verifies more than eight probes in ten, it never abstains and the two rules coincide. Post-training separates the response surface from

Table 20: External agents and the released post-trained re-query baseline on the same decision problem, run through that benchmark's own evaluation code under the balanced cost profile over five seeds. TCM appears in the final column at the verification rate each agent realized, so no method wins by verifying more often.

Agent	Prompt	Verify rate	Stale rate ↓	Utility ↑	TCM at matched verification rate ↑
Llama-3.1-8B-Instruct	no time cue	0.839	0.060	+0.4428	**+0.6081**
	re-query rule	0.864	0.052	+0.4539	**+0.5990**
	timestamps	0.856	0.047	+0.4754	**+0.6019**
Qwen3-4B	DPO	0.249	0.304	−0.0249	**+0.3333**
	no time cue	0.249	0.304	−0.0249	**+0.3333**
	re-query rule	0.261	0.300	−0.0169	**+0.3464**
	timestamps	0.349	0.254	+0.0811	**+0.4385**
Qwen3-8B	no time cue	0.248	0.305	−0.0282	**+0.3326**
	re-query rule	0.219	0.320	−0.0581	**+0.3015**
	timestamps	0.242	0.299	−0.0052	**+0.3259**

Table 21: Utility TCM and TCM-D gain over the strongest baseline once the verification budget is released, at every predefined cost profile, with the number of seeds in which the gain is positive. Strongest means highest mean utility at that profile. In all but two cells it is the flexible typed-age hazard, and in those two the constant base rate. Each implied threshold is the validity at which verifying and acting break even.

	PhysioNet-2012		TicToc		MIMIC-IV	
Threshold θ	TCM	TCM-D	TCM	TCM-D	TCM	TCM-D
0.50	$+0.011$ (5/5)	$+0.011$ (5/5)	0.000 (3/5)	0.000 (2/5)	$+0.019$ (15/15)	$+0.020$ (15/15)
0.70	$+0.053$ (5/5)	$+0.047$ (5/5)	$+0.004$ (4/5)	$+0.004$ (5/5)	$+0.082$ (15/15)	$+0.084$ (15/15)
0.85	$+0.031$ (5/5)	$+0.030$ (5/5)	-0.001 (2/5)	$+0.001$ (3/5)	$+0.064$ (15/15)	$+0.065$ (15/15)
0.95	$+0.004$ (5/5)	$+0.005$ (5/5)	$+0.007$ (5/5)	$+0.007$ (5/5)	$+0.007$ (15/15)	$+0.007$ (15/15)
0.85, high harm	$+0.062$ (5/5)	$+0.060$ (5/5)	-0.002 (2/5)	$+0.001$ (3/5)	$+0.128$ (15/15)	$+0.129$ (15/15)

the policy. Direct preference optimization on the released re-query pairs, run with the benchmark's own recipe, reaches a training reward accuracy of 0.57 to 0.61 and a reward margin near 0.03. We evaluate the full-weight checkpoint, which changes the wording of 558 of 5592 greedy responses. A margin that small moves no re-query decision, and tool-call counts stay identical to the base model on both splits.

Table 20 lists each agent and prompt. At every agent's own verification rate TCM earns more utility, by 0.13 to 0.36. Qwen3-4B after DPO matches its prompted counterpart in every column, so post-training changes no decision.

Q Decision utility in full

Table 21 gives the released-budget gains behind Figure 5b. On PhysioNet-2012 and MIMIC-IV both heads gain in every seed at every profile, most under high harm ($+0.062$ and $+0.128$ for TCM). On TicToc every gain lies within 0.007 of zero, the tie Section 5.5 reports.

A policy learned end to end. Our rule acts on an estimate of q. An alternative skips the probability and trains a policy for utility. We give that policy TCM-D's inputs and hidden layers with a three-action output layer and train it to maximize expected payoff under one cost profile. An entropy bonus is annealed to zero, and the greedy policy's development utility selects the epoch. A six-point grid over learning rate and entropy is chosen on development, a tuning budget the estimators of q never had. Without the entropy bonus the policy collapses to a single action in most profiles. Table 22 scores each policy at the costs it was trained on and at the other four. At its own costs the learned policy matches q plus the rule on TicToc and exceeds it on PhysioNet-2012 by 0.018 utility over TCM in all five seeds. A decision-focused objective spends its capacity near the threshold. Scored under costs it was not trained on, the policy loses 0.17 and 0.13 utility against TCM plus the rule, again in every seed. Meanwhile q is refitted for none of those costs. Only the estimate of q transfers.

Table 22: Utility under the unconstrained rule, five seeds. We score a policy trained end to end for one cost profile at those costs and, averaged, at the other four. Both are compared against the cost-optimal rule applied to TCM and to TCM-D. Bold marks the best value in each row and study.

	PhysioNet-2012				TicToc			
Threshold θ	policy, own	TCM	TCM-D	policy, other	policy, own	TCM	TCM-D	policy, other
0.50	**0.388**	0.386	0.386	0.007	0.069	0.072	**0.072**	−0.110
0.70	**0.461**	0.446	0.441	0.334	0.353	0.355	**0.355**	0.269
0.85	**0.636**	0.614	0.613	0.550	0.634	0.634	**0.635**	0.536
0.95	**0.859**	0.854	0.855	0.667	0.860	**0.860**	0.860	0.704
0.85, high harm	**0.275**	0.228	0.226	0.109	0.267	0.267	**0.270**	0.148
Mean	**0.524**	0.506	0.504	0.333	0.437	0.438	**0.439**	0.309

Table 23: Decision utility under the balanced cost profile ($\theta = 0.85$), five seeds. Left block: matched 20% verification budget (utility, stale use). Center: area under the utility-versus-budget frontier. Right block: the unconstrained cost-optimal rule (verification rate, stale use). Stale use is the fraction of all decisions that used an invalid memory.

	matched 20% budget		frontier	unconstrained rule	
Method	Utility ↑	Stale use ↓	AUC ↑	Verify rate	Stale use ↓
PhysioNet-2012					
Flexible age-only hazard	0.4305	0.1598	0.4736	1.000	**0.0000**
Flexible typed-age hazard	0.4828	0.1424	0.5288	0.742	0.0276
TCM-D (ours)	0.5309	0.1264	0.5665	0.626	0.0351
TCM (ours)	**0.5344**	**0.1252**	**0.5694**	0.612	0.0368
TicToc					
Flexible age-only hazard	0.1924	**0.0574**	0.4077	0.753	0.0272
Flexible typed-age hazard	0.2820	0.0888	0.4692	0.683	**0.0193**
TCM-D (ours)	**0.2821**	0.0919	**0.4704**	0.675	0.0203
TCM (ours)	0.2819	0.0924	0.4699	0.684	0.0196

One fit serves every cost profile, every budget and either action set, whereas a policy is tied to the costs it saw.

Table 23 reports the balanced profile in two execution settings side by side and labels each column by its setting. Under a matched 20% verification budget, the ranking induced by q decides which memories are verified. Memories outside the budget still choose between use and abstention by the value of q. Abstention occurs only on TicToc, in 23% to 36% of decisions. Under the unconstrained cost-optimal rule, the verification rate and stale use are those the rule itself selects from the numerical value of q. A verification rate of 1.000 there means the rule verifies every memory. Hence the same estimator can still show stale use under the budget. A degeneracy audit flags policies that collapse onto a single action. On PhysioNet-2012 under expensive verification, the flexible age-only and flexible typed-age hazards never verify and inherit the always-use anchor, while TCM and TCM-D keep a genuine mixture. At the 20% budget TCM leads on PhysioNet-2012 (0.5344 against 0.4828 for the flexible typed-age hazard), and on TicToc the three conditioned estimators lie within 0.0002 of one another. TCM's gain over the strongest baseline, the flexible typed-age hazard, is $+0.052$ on PhysioNet-2012 and $+0.093$ on MIMIC-IV at this budget. Both gains are positive in every seed, five of five and fifteen of fifteen.