WHEN WILL A MODEL'S KNOWLEDGE EXPIRE?

Anonymous authors
Paper under double-blind review

ABSTRACT

A language model may answer a factual query correctly at deployment and later become wrong even though its parameters have not changed, simply because the world has. We study how long such currently correct knowledge can be expected to remain valid. We cast this as a survival problem in calendar time, with left truncation because many facts have already been valid for some time at the evaluation date, interval censoring when the change time is known only within a window, and right censoring when no change is observed. We introduce WIKILIFETIME, a rolling-origin benchmark containing only facts that are both valid in the world and recoverable from a point-in-time language model at the anchor date. We propose EXPIRE (*EXpiration Prediction from Internal Representations and Event histories*), which combines relation-specific duration patterns, entity-level histories, frozen hidden states, and calendar drift in a low-rank time-varying hazard model. The analysis separates model fitting, tuning, recalibration, retrieval-risk certification, and final testing chronologically. We derive estimation and calibration guarantees, characterize the limits of relation-recency rules, and obtain finite-sample control of missed stale answers for retrieval decisions. Numerical studies show that EXPIRE improves remaining-lifetime prediction, within-relation discrimination, and calibration, while reducing missed stale answers in retrieval decisions.

1 INTRODUCTION

Large language models (LLMs) store significant factual knowledge in their parameters (Petroni et al., 2019; Roberts et al., 2020). A frozen model may correctly answer who leads an organization, which team an athlete represents, or which office a politician holds at its cutoff, and later return exactly the same answer after the world has changed. This creates a deployment failure that ordinary factuality evaluation does not capture: correctness at one time says little about how long that correctness will persist. Because the model remains unchanged while the underlying world evolves, a once-correct answer can become stale even though the model continues to reproduce it reliably. The model has not forgotten the fact; rather, the fact is no longer valid for the present-tense query.

A substantial literature has studied temporal factuality and the updating of language-model knowledge. Temporal language modeling, continually updated question answering, and time-sensitive benchmarks represent or evaluate facts across time (Lazaridou et al., 2021; Saxena et al., 2021; Zhang and Choi, 2021; Dhingra et al., 2022; Kasai et al., 2023; Meem et al., 2024; Vu et al., 2024), while TemporalWiki and continual-knowledge-learning benchmarks study the acquisition of updated facts (Jang et al., 2022a;b). Other work predicts which facts or questions are likely to change (Fierro et al., 2024; Piryani et al., 2026), edits factual associations after an update has been observed (De Cao et al., 2021; Mitchell et al., 2022; Meng et al., 2022; 2023; Yao et al., 2023; Yin et al., 2024), or supplements model parameters with refreshable external information (Guu et al., 2020; Khandelwal et al., 2020; Lewis et al., 2020; Izacard et al., 2023). Selective-retrieval methods further decide when such external evidence should be consulted (Jiang et al., 2023; Mallen et al., 2023; Asai et al., 2024; Jeong et al., 2024; Labruna et al., 2025; Guan et al., 2026), and conflict benchmarks examine whether current evidence can override outdated model knowledge (Nakshatri et al., 2025; Wallat et al., 2026). These lines of work address temporal sensitivity, updating, conflict, or retrieval. They do not directly estimate, for prospective deployment-time trust decisions, how much longer a fact that is both valid and correctly recoverable at a given anchor is likely to remain valid before any factual change is observed.

For a factual triple $f_i = (s_i, r_i, o_i)$ whose current validity spell begins at B_i, let E_i denote its unknown expiration time and $\tau_i \in [B_i, E_i)$ a historical anchor. With spell age $A_i = \tau_i - B_i$, our target is

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.

$$S_i^{\mathrm{res}}(u) = \mathbb{P}(E_i - \tau_i > u \mid E_i > \tau_i, \mathcal{H}_{\tau_i}), \tag{1.1}$$

where $\mathcal{H}_{\tau_i}$ contains only information available by the anchor. For example, $S_i^{\mathrm{res}}(12) = 0.82$ means that a fact correctly recoverable at τ_i is estimated to have an 82% chance of remaining valid for at least another year. The target is instance-specific: two facts sharing the same relation can have different remaining lifetimes because their spell ages, event histories, and semantic contexts differ. The resulting observation process has the standard features of survival data. Facts already valid at the anchor enter under left truncation, a transition may be observed only within an interval, and facts that remain valid through the end of follow-up are right censored (Andersen and Gill, 1982; Tsai et al., 1987; Andersen et al., 1993; Kalbfleisch and Prentice, 2002). Recent work has also used survival methods to study epoch-indexed factual acquisition during fine-tuning; here the event of interest is instead calendar-time invalidation of a fact that is already correct at the anchor (Hier and Obafemi-Ajayi, 2026).

We introduce WIKILIFETIME, a rolling-origin benchmark in which every primary test fact is both valid in the world and correctly recoverable from a temporally admissible LM at the anchor. We then propose EXPIRE, a discrete-time expiration model that combines relation-specific duration structure, entity-level event history, smooth calendar drift, and a low-rank time-varying semantic hazard derived from frozen hidden states. Before the semantic component is fitted, hidden states are centered within relation using training data, so improvements from the representation pathway are evaluated primarily through within-relation variation, not through broad differences in turnover across relation types. Predicted hazards are subsequently recalibrated using only a separate chronological block, with calibration assessed by censoring-aware probability criteria (Graf et al., 1999; Guo et al., 2017). A further held-out block converts expiration probabilities into retrieval decisions with finite-sample control of missed stale answers under the stated exchangeability conditions (Angelopoulos et al., 2024; 2025).

The paper makes four contributions. First, we formulate the remaining lifetime of currently valid and correctly recoverable knowledge as a prospective survival target under left truncation and interval or right censoring. Second, WIKILIFETIME provides a rolling-origin evaluation design with separate chronological samples for fitting, tuning, recalibration, risk certification, and final testing. Third, EXPIRE combines duration, event-history, and frozen representation information in a nested model whose incremental semantic contribution can be examined through same-relation comparisons. Fourth, we establish identification of the survival target, an oracle inequality for the censored hazard estimator, a separation result for relation-recency rules, stability and calibration bounds for the resulting survival forecasts, and finite-sample control of missed stale answers. Monte Carlo simulations isolate the effects of censoring, semantic heterogeneity, and calendar drift. A historical evaluation on WIKILIFETIME shows that EXPIRE improves remaining-lifetime prediction, within-relation discrimination, and calibration, while reducing missed stale answers in retrieval decisions across diverse relation types.

2 PROSPECTIVE KNOWLEDGE EXPIRATION

For subject s and relation r, let $O_{sr}(t)$ denote the externally valid object at calendar time t for a relation that admits a single current object. For a fact $f_i = (s_i, r_i, o_i)$, a spell is a maximal interval $[B_i, E_i)$ over which $O_{s_i r_i}(t) = o_i$. Thus B_i is the calendar time at which the spell begins and E_i is the time at which its present-tense validity ends. We focus on relation families for which replacement or termination of the current object has a well-defined interpretation. Relations that permit several concurrent objects are excluded from the primary benchmark. Expiration therefore refers only to the end of present-tense validity; the corresponding historical statement remains true.

Let M_τ denote a frozen LM whose documented data cutoff is no later than calendar time τ. For each fact f_i that is valid at anchor τ_i, we construct J fixed paraphrases $q_i^{(1)}, \ldots, q_i^{(J)}$ and an accepted alias set $\mathcal{A}(o_i)$. Let $\mathrm{Ans}(M, q)$ denote the normalized short answer returned by model M to prompt q. For a fixed agreement threshold $k_0 \in \{1, \ldots, J\}$, define

$$K_i(M_{\tau_i}) = \mathbf{1}\left\{ \sum_{j=1}^{J} \mathbf{1}\left\{ \mathrm{Ans}(M_{\tau_i}, q_i^{(j)}) \in \mathcal{A}(o_i) \right\} \geq k_0 \right\}. \tag{2.1}$$

The primary benchmark uses $J = 3$ paraphrases and requires agreement on at least $k_0 = 2$ of them. Three prompts are the smallest collection permitting a majority rule, while the two-of-three criterion reduces sensitivity to prompt-specific failures without requiring unanimity. The construction itself is not tied to this choice, and sensitivity to alternative agreement thresholds is reported in Table E.12.

The primary WIKILIFETIME population at anchor τ is $\mathcal{P}_\tau = \{i : \tau \in [B_i, E_i), K_i(M_\tau) = 1\}$. Thus every retained fact is both valid in the world and correctly recoverable from the point-in-time LM at the anchor. Facts that the model does not recover correctly are outside the target population and are not treated as expired model knowledge.

Let $T_i = E_i - B_i$ denote total spell lifetime and $A_i = \tau_i - B_i$ its age at the anchor. Let X_i collect all predictors available by τ_i, including relation and calendar information, spell history, model-confidence diagnostics, and the frozen representation; Section 3 gives the concrete decomposition. By construction, X_i contains no post-anchor information. Writing $S^0(t \mid X) = \mathbb{P}(T > t \mid X)$ for total-spell survival, the remaining-survival target for a fact that has already survived to age A_i is

$$S^{\text{res}}(u \mid A_i, X_i) = \frac{S^0(A_i + u \mid X_i)}{S^0(A_i \mid X_i)}. \tag{2.2}$$

The denominator is the left-truncation adjustment induced by conditioning on survival to the anchor. Ignoring it treats a spell that has already survived for A_i units of time as though it had just begun and can distort remaining-lifetime predictions when the hazard depends on spell age.

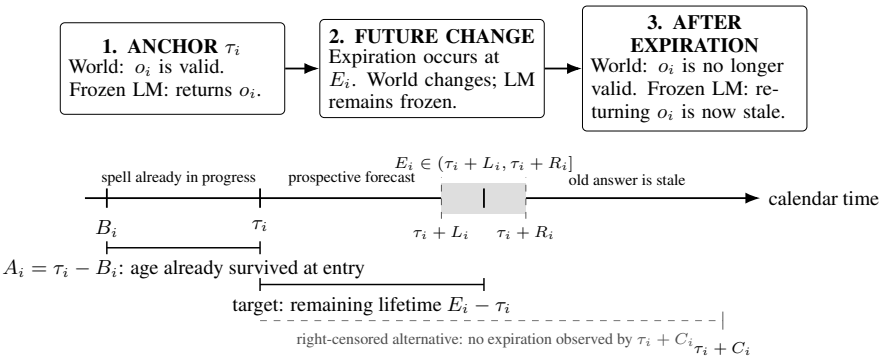


Figure 1: **Prospective knowledge expiration.** At anchor τ_i, a retained fact is both valid in the world and correctly recoverable from the frozen point-in-time LM. Its current spell began earlier at B_i, so entry occurs after survival to age $A_i = \tau_i - B_i$. The target is the remaining lifetime $E_i - \tau_i$. When expiration is interval censored, L_i and R_i denote post-anchor durations, so the corresponding calendar interval is $(\tau_i + L_i, \tau_i + R_i]$; under right censoring, no expiration is observed by $\tau_i + C_i$.

Expiration need not be observed exactly. Define the remaining lifetime $\tilde{T}_i = T_i - A_i = E_i - \tau_i$. If expiration is observed only within a post-anchor interval, let $(L_i, R_i]$ satisfy $L_i < \tilde{T}_i \le R_i, R_i < \infty$. Thus L_i and R_i are elapsed times measured from the anchor, not absolute calendar dates. Exact monthly observation is the special case $(L_i, R_i] = (m-1, m]$. Let C_i denote the administrative follow-up duration after the anchor. If no expiration is observed by $\tau_i + C_i$, the record is right censored and $\tilde{T}_i > C_i$. We write $\delta_i = \mathbf{1}\{R_i < \infty\}$ for the observed-expiration indicator and, for notational uniformity, set $R_i = \infty$ when $\delta_i = 0$. The observed record is $W_i = (A_i, X_i, L_i, R_i, C_i, \delta_i)$. For a right-censored record, L_i is unused. The likelihood conditions on survival to the anchor and uses the observed interval directly and without replacing it by an imputed event time. Figure 1 summarizes the prospective forecasting setup. Note that all information used at a historical origin τ must be reconstructible from the database state and frozen LM available by that origin. We therefore use chronologically separated blocks $\mathcal{D}^{\text{tr}}$, $\mathcal{D}^{\text{val}}$, $\mathcal{D}^{\text{recal}}$, $\mathcal{D}^{\text{cert}}$, and $\mathcal{D}^{\text{test}}$, with distinct roles for each block. Candidate hazard models are fitted on $\mathcal{D}^{\text{tr}}$ and selected using $\mathcal{D}^{\text{val}}$. After selection, the hazard model is refitted on $\mathcal{D}^{\text{tr}} \cup \mathcal{D}^{\text{val}}$; recalibration uses only $\mathcal{D}^{\text{recal}}$, retrieval-risk certification uses only $\mathcal{D}^{\text{cert}}$, and $\mathcal{D}^{\text{test}}$ remains untouched until final evaluation. Spell and subject identifiers are grouped when the blocks are formed so that observations from the same spell or subject cannot cross the relevant split boundaries within a rolling-origin experiment. Appendix C gives the complete benchmark construction.

3 EXPIRE: EVENT-HISTORY AND REPRESENTATION HAZARDS

We model the remaining lifetime in Section 2 on a monthly grid. For a retained fact at anchor τ_i, let $h = 1, \ldots, M$ index months after the anchor and define

$$\lambda_i(h) = \mathbb{P}\left(h - 1 < \tilde{T}_i \le h \mid \tilde{T}_i > h - 1, A_i, X_i\right) = \text{sigm}\{f_i(h)\}, \tag{3.1}$$

where $\tilde{T}_i = E_i - \tau_i$, $\mathrm{sigm}(v) = (1 + e^{-v})^{-1}$, and M is the maximum forecast horizon. The corresponding remaining-survival curve is $S_i^{\mathrm{res}}(u) = \prod_{h=1}^{u}\{1 - \lambda_i(h)\},\ u = 1, \ldots, M$. Thus the hazard is defined directly on the post-anchor time scale, conditional on the spell having survived to age A_i. The observed durations L_i, R_i, and C_i are on the same monthly grid. The EXPIRE hazard logit is

$$f_i(h) = \alpha_0 + B_a(A_i + h)^\top \theta_{r_i} + B_c(\tau_i + h)^\top \xi_{r_i} + \beta^\top x_i + \left\{U^\top \tilde{z}_i\right\}^\top \left\{V^\top B_s(h)\right\}, \tag{3.2}$$

where α_0 is a global intercept. The training-centered spline bases $B_a \in \mathbb{R}^{K_a}$, $B_c \in \mathbb{R}^{K_c}$, and $B_s \in \mathbb{R}^{K_s}$ contain no intercept term and represent total spell age $A_i + h$, calendar time $\tau_i + h$, and forecast horizon h, respectively. Calendar time is encoded on the benchmark's common monthly index; spline knots are determined from $\mathcal{D}^{\mathrm{tr}}$ and then fixed. The vectors $\theta_r \in \mathbb{R}^{K_a}$ and $\xi_r \in \mathbb{R}^{K_c}$ are relation-specific age and calendar coefficients. The vector $x_i \in \mathbb{R}^p$ contains pre-anchor event-history and model-confidence features, including previous-spell summaries for (s_i, r_i), relation-level summaries of completed spells available by τ_i, entity type, answer-confidence diagnostics, and prompt agreement. No component of x_i uses post-anchor information.

The final term in (3.2) allows the contribution of the frozen LM representation to vary with forecast horizon. For each of the J fixed prompts, we extract the hidden state at the final input token from frozen layer ℓ, average across prompts, and ℓ_2-normalize the result to obtain $z_i \in \mathbb{R}^{d_z}$. Because z_i can strongly encode relation identity, we center representations within relation using training data only

$$\hat{\mu}_r = \frac{1}{n_r} \sum_{i \in \mathcal{D}^{\mathrm{tr}}: r_i = r} z_i \text{ and } \tilde{z}_i = P(z_i - \hat{\mu}_{r_i}), \tag{3.3}$$

where $n_r = |\{i \in \mathcal{D}^{\mathrm{tr}} : r_i = r\}|$ and $P \in \mathbb{R}^{d_p \times d_z}$ contains the leading d_p principal directions of the relation-centered representations in $\mathcal{D}^{\mathrm{tr}}$. The projection is fixed when the candidate configuration is evaluated on later chronological blocks. Matrices $U \in \mathbb{R}^{d_p \times q}$ and $V \in \mathbb{R}^{K_s \times q}$ form a rank-q interaction between representation and forecast horizon. The analysis is restricted to relations represented in $\mathcal{D}^{\mathrm{tr}}$. The specification is nested: $U = 0$ removes the LM representation, $\beta = U = 0$ further removes entity-level history, and additionally constraining the age and calendar effects to be common across relations yields a global duration model. These restrictions are used directly in the ablation experiments.

For a candidate hazard logit f, we write

$$\lambda_f(h \mid A, X) = \mathrm{sigm}\{f(A, X, h)\} \text{ and } S_f^{\mathrm{res}}(u \mid A, X) = \prod_{h=1}^{u}\{1 - \lambda_f(h \mid A, X)\},$$

with $S_f^{\mathrm{res}}(0 \mid A, X) = 1$. Using the observed record W_i from Section 2, its likelihood contribution is

$$p_f(W_i) = \begin{cases} S_f^{\mathrm{res}}(L_i \mid A_i, X_i) - S_f^{\mathrm{res}}(R_i \mid A_i, X_i), & \delta_i = 1, \\ S_f^{\mathrm{res}}(C_i \mid A_i, X_i), & \delta_i = 0. \end{cases} \tag{3.4}$$

This likelihood conditions on survival to the anchor and uses the observed censoring information directly, avoiding midpoint imputation. The hazard parameters are estimated on $\mathcal{D}^{\mathrm{tr}}$ by minimizing

$$\begin{aligned} \hat{Q}(f) = &- \frac{1}{n_{\mathrm{tr}}} \sum_{i \in \mathcal{D}^{\mathrm{tr}}} \log p_f(W_i) + \lambda_{\mathrm{rel}} \sum_{r=1}^{R_{\mathrm{rel}}} \|\theta_r - \bar{\theta}\|_2^2 \\ &+ \lambda_{\mathrm{drift}} \sum_{r=1}^{R_{\mathrm{rel}}} \|D_2 \xi_r\|_2^2 + \lambda_U \|U\|_F^2 + \lambda_V \|V\|_F^2 + \lambda_1 \|\beta\|_1, \end{aligned} \tag{3.5}$$

where $n_{\mathrm{tr}} = |\mathcal{D}^{\mathrm{tr}}|$, R_{rel} is the number of modeled relations, $\bar{\theta} = R_{\mathrm{rel}}^{-1} \sum_{r=1}^{R_{\mathrm{rel}}} \theta_r$, and D_2 is the second-difference operator applied to the calendar coefficients. The nonnegative tuning parameters $\lambda_{\mathrm{rel}}, \lambda_{\mathrm{drift}}, \lambda_U, \lambda_V$, and λ_1 control relation-age shrinkage, calendar smoothness, ridge regularization of U and V, and sparsity of β, respectively.

Accurate lifetime ranking does not ensure accurate probabilities. After model selection and refitting, let $\hat{\lambda}_i(h)$ denote the fitted hazard and $\hat{S}_i^{\mathrm{res}}(u) = \prod_{h=1}^{u}\{1 - \hat{\lambda}_i(h)\}$. Using only $\mathcal{D}^{\mathrm{recal}}$, we fit

$$\lambda_i^{\mathrm{cal}}(h) = \mathrm{sigm}\left[a\ \mathrm{logit}\{\hat{\lambda}_i(h)\} + b(h)\right], \qquad a \geq 0, \tag{3.6}$$

where a is a calibration slope and $b(h)$ is a low-dimensional cubic-spline offset. The recalibration pa-

rameters are estimated with the interval-censored likelihood in (3.4), holding the hazard model fixed. The calibrated survival estimate is $\hat{S}_i^{\mathrm{cal}}(u) = \prod_{h=1}^{u}\{1 - \lambda_i^{\mathrm{cal}}(h)\}$. It is non-increasing in u by construction, while $a \geq 0$ prevents the scalar calibration term from reversing the fitted hazard ordering at a fixed horizon. Appendix B gives the recalibration spline specification.

At deployment horizon u, let $\hat{p}_i(u) = 1 - \hat{S}_i^{\mathrm{cal}}(u)$ be the predicted probability of expiration within u months. For $\rho \in [0, 1]$, define $d_\rho(X_i; u) = \mathbf{1}\{\hat{p}_i(u) \geq 1 - \rho\}$, where $d_\rho = 1$ denotes retrieval, so larger ρ retrieves more often. The missed-stale loss is $\mathcal{L}_i^{\mathrm{ms}}(\rho; u) = \mathbf{1}\{d_\rho(X_i; u) = 0\}\mathbf{1}\{\tilde{T}_i \leq u\}$, which records an expiration within the deployment horizon that is not followed by retrieval. Let $\mathcal{D}^{\mathrm{cert}}$ contain n_{cert} examples whose expiration status is fully observed through horizon u, and define $\hat{R}_{\mathrm{cert}}(\rho; u) = n_{\mathrm{cert}}^{-1} \sum_{i \in \mathcal{D}^{\mathrm{cert}}} \mathcal{L}_i^{\mathrm{ms}}(\rho; u)$. For a prespecified target risk $\alpha \in (0, 1)$, conformal risk control (Bates et al., 2021; Angelopoulos et al., 2024) selects

$$\hat{\rho}_u = \inf\left\{\rho : \frac{n_{\mathrm{cert}}}{n_{\mathrm{cert}} + 1}\hat{R}_{\mathrm{cert}}(\rho; u) + \frac{1}{n_{\mathrm{cert}} + 1} \leq \alpha\right\}. \tag{3.7}$$

We set $d_1 = 1$ so that the always-retrieve endpoint is available whenever the target is feasible. Certification leaves the lifetime model unchanged and selects a retrieval rule with the finite-sample guarantee established in Theorem 4.6. The hazard-model tuning configuration is $\gamma = (\ell, d_p, q, K_a, K_c, K_s, \lambda_{\mathrm{rel}}, \lambda_{\mathrm{drift}}, \lambda_U, \lambda_V, \lambda_1)$, where ℓ is the frozen LM layer, d_p the retained representation dimension, q the semantic interaction rank, and K_a, K_c, K_s the spline dimensions; the remaining entries are the penalty parameters defined above. All data-dependent preprocessing for a candidate configuration is constructed from $\mathcal{D}^{\mathrm{tr}}$ only. Prespecified configuration-epoch candidates are generated from $\mathcal{D}^{\mathrm{tr}}$ and selected on $\mathcal{D}^{\mathrm{val}}$ by held-out interval-censored negative log likelihood based on (3.4). The selected configuration is refitted on $\mathcal{D}^{\mathrm{tr}} \cup \mathcal{D}^{\mathrm{val}}$ for the selected number of epochs, the recalibrator uses only $\mathcal{D}^{\mathrm{recal}}$, and $\hat{\rho}_u$ is selected only on $\mathcal{D}^{\mathrm{cert}}$. Appendix B gives the candidate pool, finite-family construction, and deterministic tie-breaker. Figure 2 summarizes the forecasting pipeline and data separation, and Algorithm 1 gives the complete estimation procedure.

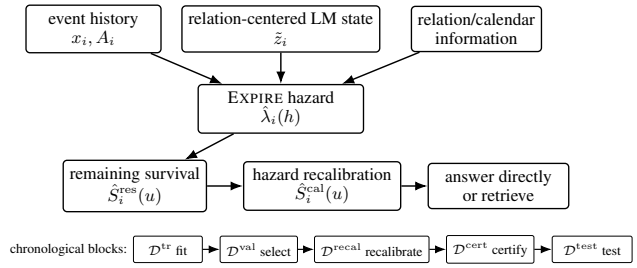


Figure 2: **EXPIRE pipeline and chronological data separation.** Pre-anchor event history, relation and calendar structure, and the relation-centered frozen representation determine the expiration hazard. The hazard yields a remaining-survival forecast, which is recalibrated before the direct-answer or retrieval decision. The lower strip shows the chronologically separated samples.

Algorithm 1 EXPIRE estimation and deployment

Require: Historical spells, point-in-time checkpoints, chronological blocks $\mathcal{D}^{\mathrm{tr}}, \mathcal{D}^{\mathrm{val}}, \mathcal{D}^{\mathrm{recal}}, \mathcal{D}^{\mathrm{cert}}$, forecast horizon u, and target risk α.

1: Reconstruct validity spells and anchor ages; retain facts satisfying the point-in-time knowledge rule (2.1).
2: Using information available by each anchor only, compute event-history features, relation summaries, confidence diagnostics, and frozen hidden states.
3: For each candidate representation configuration, use $\mathcal{D}^{\mathrm{tr}}$ only to compute relation centroids and principal directions and construct $\tilde{z}_i$ by (3.3).
4: For each candidate γ, fit (3.5) on $\mathcal{D}^{\mathrm{tr}}$ and score its prespecified training checkpoints on $\mathcal{D}^{\mathrm{val}}$ by interval-censored negative log likelihood.
5: Select $(\hat{\gamma}, \hat{e})$, refit the hazard model on $\mathcal{D}^{\mathrm{tr}} \cup \mathcal{D}^{\mathrm{val}}$ for exactly $\hat{e}$ epochs, and freeze its parameters.
6: Fit the recalibrator (3.6) on $\mathcal{D}^{\mathrm{recal}}$ and freeze it.
7: On $\mathcal{D}^{\mathrm{cert}}$, select retrieval aggressiveness $\hat{\rho}_u$ using (3.7).
8: For a new eligible fact, return $\hat{S}^{\mathrm{cal}}(\cdot)$ and retrieve iff $d_{\hat{\rho}_u}(X; u) = 1$.

4 THEORETICAL PROPERTIES

The theoretical analysis separates lifetime estimation from the optional retrieval certificate. Let $\mathcal{F}$ be a bounded class of post-anchor hazard logits $f(A, X, h)$, let $\ell_f(W) = -\log p_f(W)$ denote the observed-data negative log likelihood in (3.4), and define $\mathcal{R}(f) = \mathbb{E}\{\ell_f(W)\}$. A superscript $\star$ denotes a true population quantity, while $f^\star_{\mathcal{F}} \in \arg\min_{f\in\mathcal{F}} \mathcal{R}(f)$ is a best-in-class hazard logit. We assume conditionally non-informative coarsening given anchor information, positivity of the inspection patterns needed to distinguish supported grid cells, bounded logits, and observed-interval probabilities bounded away from zero. When a subject contributes multiple records, subject clusters are independent with uniformly bounded cluster size; n below denotes the number of independent clusters. Empirical-process statements are conditional on the anchor-side covariates and training-derived preprocessing maps. Appendix A states these conditions formally and gives the cluster-level formulation.

Proposition 4.1. *Under Assumptions 1-2, suppose the inspection family separates every supported post-anchor grid cell with positive probability. Then the mapping from the residual-survival distribution to the conditional distribution of the observed record W is injective, and the population observed-data likelihood is uniquely minimized by the true residual-survival distribution on the supported grid. Consequently,*

$$S^{\mathrm{res},\star}(u \mid A, X) = \frac{S^{0,\star}(A + u \mid X)}{S^{0,\star}(A \mid X)}$$

is identified at every supported horizon u.

Thus the target is identified from the left-truncated, coarsened observation process itself; midpoint imputation and complete-case filtering are unnecessary. To match the estimator in Section 3, let ϑ denote a parameterization of $f_\vartheta \in \mathcal{F}$ and write $\mathrm{Pen}_\lambda(\vartheta)$ for the nonnegative penalty terms in (3.5). Define the induced loss class $\mathcal{L}_{\mathcal{F}} = \{W \mapsto \ell_f(W) : f \in \mathcal{F}\}$.

Theorem 4.2. *Suppose Assumptions 1-3 hold and $\hat\vartheta$ is an $\varepsilon_{\mathrm{opt}}$-approximate minimizer of the penalized empirical objective* (3.5)*. Let $\mathfrak{R}_n(\mathcal{L}_{\mathcal{F}})$ denote the cluster-level empirical Rademacher complexity of the induced observed-data loss class. Then, for every $\eta \in (0, 1)$, with probability at least $1 - \eta$,*

$$\mathcal{R}(f_{\hat\vartheta}) \le \inf_{\vartheta : f_\vartheta \in \mathcal{F}} \{\mathcal{R}(f_\vartheta) + \mathrm{Pen}_\lambda(\vartheta)\} + C_1 \mathfrak{R}_n(\mathcal{L}_{\mathcal{F}}) + C_2 \sqrt{\frac{\log(2/\eta)}{n}} + \varepsilon_{\mathrm{opt}},$$

where $C_1, C_2 < \infty$ depend only on the boundedness and positivity constants, maximum horizon, and cluster-size bound. For the rank-q specification (3.2)*, bounded spline, representation, and parameter norms imply*

$$\mathfrak{R}_n(\mathcal{L}_{\mathcal{F}}) \lesssim \sqrt{\frac{q(d_p + K_s) + R_{\mathrm{rel}}(K_a + K_c) + p + 1}{n}},$$

where $\lesssim$ absorbs bounded-norm, horizon, interval-positivity, and logarithmic factors. Thus the interaction contributes through $q(d_p + K_s)$, not through a $d_p \times M$ horizon-specific coefficient table.

For any best-in-class parameterization $\vartheta^\star_{\mathcal{F}}$, Theorem 4.2 also bounds $\mathcal{R}(f_{\hat\vartheta}) - \mathcal{R}(f^\star_{\mathcal{F}})$ by the stochastic terms, optimization error, and $\mathrm{Pen}_\lambda(\vartheta^\star_{\mathcal{F}})$. The displayed result concerns a fixed candidate class; Appendix A gives the finite-family extension for the validation-selected configuration-epoch candidates in Section 3. Relation and recency information alone cannot distinguish facts with the same observed turnover profile but different remaining lifetimes. The next result isolates this limitation.

Theorem 4.3. *We fix a horizon u and a stratum $G = g$, where G contains the relation identity and recency variables available to a relation-recency rule. Suppose facts in this stratum have a binary subtype $Z \in \{0, 1\}$ with $\mathbb{P}(Z = 1 \mid G = g) = 1/2$ and $S^\star(u \mid G = g, Z = 0) = s_0$, $S^\star(u \mid G = g, Z = 1) = s_1$, where $s_0 \neq s_1$. Under squared loss for $Y_u = \mathbf{1}\{\tilde T > u\}$, the minimum conditional Brier risk among G-measurable predictors exceeds the subtype-aware Bayes risk by $(s_1 - s_0)^2/4$. If the relation-centered representation permits a classifier $\hat Z$ with $\mathbb{P}(\hat Z \neq Z \mid G = g) \to 0$ and the predictor class contains consistently estimable subtype-specific survival probabilities, then a representation-aware predictor attains the subtype-aware Bayes risk asymptotically.*

Theorem 4.3 does not assume that every frozen LM contains such information. It characterizes the gain available when the residual representation carries within-stratum lifetime signal; the matched-relation and representation ablations in Section 5 examine this premise empirically. The direct post-anchor parameterization in (3.1) also yields a simple stability result.

Lemma 4.4. *For two hazard logits f and g, write $f_h = f(A, X, h)$ and $g_h = g(A, X, h)$. Then, for every integer $u \le M$,*

$$\left|S_f^{\mathrm{res}}(u \mid A, X) - S_g^{\mathrm{res}}(u \mid A, X)\right| \le \sum_{h=1}^{u} |\lambda_f(h \mid A, X) - \lambda_g(h \mid A, X)| \le \frac{1}{4}\sum_{h=1}^{u} |f_h - g_h|.$$

Thus cumulative hazard-logit error directly controls the remaining-survival quantity reported by the model, without an additional pre-anchor stability term.

Proposition 4.5. *We fix a horizon u and let $Y_u = \mathbf{1}\{\tilde{T} > u\}$, $S_u^\star = S^{\mathrm{res},\star}(u \mid A, X) = \mathbb{E}(Y_u \mid A, X)$. For any measurable survival estimate $\hat{S}_u \in [0, 1]$, define $\mathrm{CE}_u(\hat{S}) = \mathbb{E}|\mathbb{E}(Y_u \mid \hat{S}_u) - \hat{S}_u|$. Then $\mathrm{CE}_u(\hat{S}) \le \mathbb{E}|S_u^\star - \hat{S}_u|$.*

Hence remaining-survival accuracy also controls horizon-specific validity calibration. In particular, the proposition applies to the calibrated estimate $\hat{S}^{\mathrm{cal}}$ in Section 3 for the same horizon and covariates. The optional certification stage controls a different quantity; the probability of answering directly when the fact expires within the certified deployment horizon.

Theorem 4.6. *We fix a horizon u and condition on all fitting, model-selection, and recalibration data. Suppose the n_{cert} certificate examples and one future example are exchangeable at the certified deployment regime and their indicators $\mathbf{1}\{\tilde{T}_i \le u\}$ are fully observed. Let $\mathcal{L}_i^{\mathrm{ms}}(\rho; u)$ be the bounded, non-increasing loss in Section 3. If $\alpha \ge 1/(n_{\mathrm{cert}} + 1)$, then the rule $\hat{\rho}_u$ in (3.7), with $d_1 = 1$, satisfies*

$$\mathbb{E}\left[\mathcal{L}^{\mathrm{ms}}_{n_{\mathrm{cert}}+1}(\hat{\rho}_u; u)\right] \le \alpha.$$

The guarantee is local to the certified deployment regime; it does not assert exchangeability across arbitrary future calendar drift beyond the observed certification time window.

Finally, consider a separate cost-based retrieval rule for horizon-specific deployment under explicit costs. Suppose answering directly after expiration costs $c_E > 0$, retrieval costs $c_R \ge 0$, and retrieval is correct in the idealized decision model. Let $p^\star(u) = 1 - S_u^\star$ and define $\mathrm{Risk}(d) = \mathbb{E}[d(X)c_R + \{1 - d(X)\}c_E p^\star(u)]$. The Bayes rule retrieves iff $p^\star(u) > c_R/c_E$.

Corollary 4.7. *Let $d^\star(X) = \mathbf{1}\{p^\star(u) > c_R/c_E\}$ and $\hat{d}(X) = \mathbf{1}\{\hat{p}(u) > c_R/c_E\}$, where $\hat{p}(u) = 1 - \hat{S}(u)$. Then*

$$\mathrm{Risk}(\hat{d}) - \mathrm{Risk}(d^\star) \le c_E \mathbb{E}|\hat{p} - p^\star| = c_E \mathbb{E}|\hat{S} - S^\star|.$$

This cost-based rule is distinct from the certified rule $d_{\hat{\rho}_u}$. Corollary 4.7 shows that, when a cost threshold is used, the same survival error that controls probabilistic validity also bounds excess retrieval cost. This separates predictive accuracy from deployment-time decision quality. Extensions allowing retrieval failure and query-dependent retrieval cost are given in Appendix F.

5 Numerical Studies

We evaluate EXPIRE in two complementary settings. Monte Carlo simulations isolate the effects of interval censoring, within-relation semantic heterogeneity, and calendar drift under controlled data-generating mechanisms. The empirical application then evaluates prospective expiration forecasting and retrieval decisions on WIKILIFETIME using strictly chronological point-in-time information.

5.1 Monte Carlo Simulations

The simulation study isolates interval censoring, within-relation semantic signal, and calendar drift. We use $R_{\mathrm{rel}} = 12$, maximum forecast horizon $M = 24$, $A_i \sim \mathrm{Unif}\{1, \ldots, 48\}$, observed history features, and an eight-dimensional pseudo-LM state. Following discrete-time survival modeling (Gensheimer and Narasimhan, 2019; Kvamme and Borgan, 2021), data are generated from

$$\mathrm{logit}\, \lambda_i(h) = \alpha_{r_i} + \beta_U U_i + 0.55 V_i + 0.42 H_i + 0.02(A_i + h) + 0.12 \sin(h/4) + D_{r_i}(\tau_i + h), \quad (5.1)$$

where U_i is a within-relation semantic state, V_i is latent history, H_i is its observed history summary, and D_r is a relation-specific calendar effect. The first two pseudo-LM coordinates encode U_i. Designs S0-S3 use $(\beta_U, D_r) = (0, 0), (0.45, 0), (0.75, 0), (0.75, D_r^{\mathrm{drift}})$, respectively, while the

history mechanism is held fixed. Appendix D gives the complete data-generating distributions and calendar-drift specification. Expiration is observed through coarsened inspection intervals together with random administrative censoring, with censoring times drawn from months 8-24 and approximately 35% right censoring in the primary design. Each design has 500 independent replications with $(n_{\mathrm{tr}}, n_{\mathrm{val}}, n_{\mathrm{test}}) = (5000, 1000, 3000)$. These three blocks evaluate lifetime forecasting itself; recalibration and risk certification are omitted from the controlled study and examined separately in the empirical application. Relation only, metadata/history, representation only, and full EXPIRE provide structured ablations. Interval-censored proportional-hazards, survival-forest, and neural-survival comparisons are reported in Appendix D. All methods receive the same coarsened observations, and no baseline is given the latent expiration time or a midpoint-imputed event time.

Table 1: **Lifetime forecasting.** Entries are Monte Carlo means over 500 independent replications, with Monte Carlo standard errors (SEs) in parentheses. IBS is coarsening-aware integrated Brier score; pair accuracy is the S2 same-relation ranking metric.

Method	S0 IBS ↓	S1 IBS ↓	S2 IBS ↓	S3 IBS ↓	S2 pair acc. ↑
Relation only	.214 (.0012)	.220 (.0013)	.233 (.0014)	.241 (.0016)	.500 (.0048)
Metadata/history	.195 (.0010)	.211 (.0011)	.220 (.0012)	.226 (.0013)	.682 (.0042)
Representation only	.229 (.0013)	.218 (.0012)	.213 (.0011)	.221 (.0013)	.745 (.0039)
EXPIRE	.193 (.0009)	.201 (.0010)	.200 (.0010)	.206 (.0011)	.799 (.0035)

Table 1 separates the information sources targeted by the method. In S0, where the pseudo-LM state contains no lifetime signal, the full model matches the metadata/history specification, indicating that the representation pathway does not generate a meaningful spurious gain. As within-relation semantic information strengthens in S1 and S2, the advantage of EXPIRE increases, while the representation-only specification also becomes more competitive. Under S3, EXPIRE retains the smallest IBS after relation-specific calendar drift is introduced. The S2 matched-pair result is particularly informative; relation-only prediction remains at chance, whereas EXPIRE reaches pair accuracy .799, consistent with recovery of within-relation lifetime heterogeneity. Figure 3 examines recovery of the post-anchor hazard and sensitivity to increasingly coarse expiration intervals. The diagnostics reinforce the table-level comparison. Hazard recovery remains close to the data-generating trajectory over the forecast horizon, while coarser inspection reduces the information available to every method. The degradation is gradual, never becoming abrupt, and EXPIRE retains an advantage across the examined interval widths.

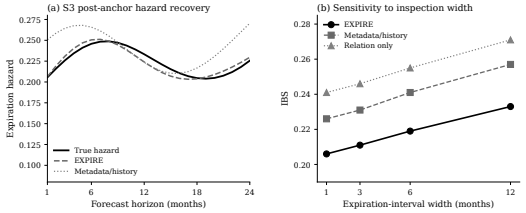


Figure 3: **Monte Carlo diagnostics.** (a) Post-anchor expiration-hazard recovery in S3. (b) Lifetime-prediction error as the observed expiration interval is widened. Replication uncertainty and the complete interval-width sensitivity analysis are reported in Appendix D.

5.2 EMPIRICAL APPLICATION

We next evaluate prospective expiration on WIKILIFETIME. Using Wikidata histories (Vrandečić and Krötzsch, 2014) and point-in-time LMs (Kelly et al., 2026), we retain facts that are valid and correctly recoverable at each historical anchor. January and July origins through 2020 form $\mathcal{D}^{\mathrm{tr}}$, 2021 forms $\mathcal{D}^{\mathrm{val}}$, January 2022 forms $\mathcal{D}^{\mathrm{recal}}$, July 2022 forms $\mathcal{D}^{\mathrm{cert}}$, and 2023 is the untouched $\mathcal{D}^{\mathrm{test}}$ block. Spell and subject grouping follows Section 2. OLMo provides cutoff-documented sensitivity analysis (Groeneveld et al., 2024); block counts and censoring summaries are in Appendix C. Baselines include global and relation-specific Turnbull estimators (Turnbull, 1976), interval-censored proportional hazards and survival forests, discrete-time neural survival, mutability/recency labels (Fierro et al., 2024; Piryani et al., 2026), LM-verbalized probabilities, a relation-centered survival probe, and EXPIRE-History. A post-anchor checkpoint is used only for leakage diagnosis. All survival methods receive the same observed intervals and censoring, without latent or midpoint-imputed event times; implementations and tuning are in Appendix E. IBS, horizon-specific Brier scores, and calibration use the coarsening-aware estimators in Appendix E, while NLL uses (3.4).

Table 2: **WIKILIFETIME application.** Entries are test-set estimates with subject-cluster bootstrap SEs in parentheses. NLL is observed-data negative log likelihood; ECE is expected calibration error; BS is Brier score; pair accuracy measures matched within-relation ranking.

Method	IBS ↓	NLL ↓	12m ECE ↓	Pair acc. ↑	24m BS ↓
Global Turnbull	.128 (.0038)	.439 (.0107)	.052 (.0046)	.500 (.0098)	.184 (.0051)
Relation Turnbull	.117 (.0035)	.405 (.0100)	.040 (.0041)	.500 (.0098)	.166 (.0048)
IC proportional hazards	.108 (.0032)	.379 (.0093)	.032 (.0036)	.584 (.0091)	.154 (.0044)
IC survival forest	.104 (.0031)	.368 (.0090)	.030 (.0034)	.607 (.0089)	.149 (.0042)
Discrete-time neural survival	.102 (.0030)	.362 (.0088)	.029 (.0033)	.620 (.0087)	.146 (.0041)
Mutability / recency class	.114 (.0034)	.398 (.0098)	.044 (.0042)	.500 (.0098)	.162 (.0047)
LM verbalized probability	.121 (.0036)	.418 (.0102)	.065 (.0050)	.574 (.0092)	.173 (.0049)
LM residual survival probe	.109 (.0033)	.384 (.0094)	.041 (.0040)	.641 (.0085)	.157 (.0044)
EXPIRE-History	.099 (.0029)	.352 (.0085)	.023 (.0030)	.636 (.0086)	.141 (.0039)
EXPIRE	.092 (.0027)	.331 (.0081)	.018 (.0026)	.699 (.0078)	.131 (.0036)

Table 2 shows consistent gains across likelihood, calibration, and ranking criteria. Relation-specific survival improves over the global baseline, while event-history and flexible survival models yield further gains. Mutability/recency labels and the residual LM probe remain weaker on survival-probability metrics. Relative to EXPIRE-History, full EXPIRE reduces IBS from .099 to .092 and NLL from .352 to .331, with 12-month ECE .018, 24-month Brier score .131, and matched pair accuracy .699. The matched-pair diagnostic controls the main alternative explanation that the LM representation merely identifies fast-changing relations: facts expiring within 12 months are matched to controls from the same relation, anchor quarter, and spell-age decile that remain valid beyond 24 months. Figure 4 then summarizes fact-level survival, calibration, within-relation discrimination, and retrieval utility. Alternative matching, layerwise probes, cutoff transfer, and leakage checks are reported in Appendix E.2.

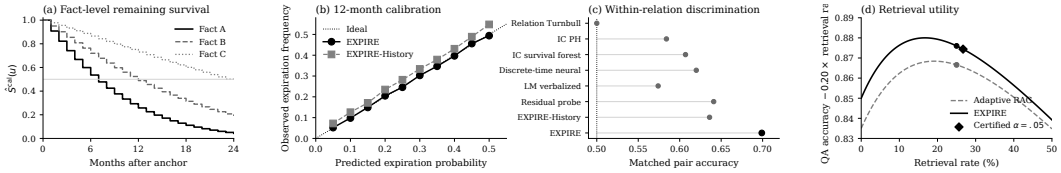


Figure 4: **Empirical Diagnostics**

Finally, we convert calibrated expiration probabilities into retrieval decisions. Budget-matched policies share the same retrieval backend and differ only in the trigger. As shown in Table 3, at a 25% retrieval rate, EXPIRE increases QA accuracy from .916 for adaptive RAG (Jeong et al., 2024) to .926 and reduces the missed-stale rate from .061 to .049. The separately certified $\alpha = .05$ rule retrieves on 26.7% of test queries, attains QA accuracy .929, and has missed-stale rate .047. The certified rule therefore improves protection without excessive additional retrieval. Retrieval rises by only 1.7 percentage points. This confirms that improved expiration ranking translates into safer answering. The budget-matched comparison isolates score quality, whereas the certified row evaluates the risk-control layer in Theorem 4.6; its threshold is selected without access to the test block. Appendix E.3 reports the full accuracy-retrieval frontier, alternative risk levels, and imperfect-retrieval sensitivity.

Table 3: **Retrieval-policy evaluation.** Budget-matched policies retrieve on exactly 25% of queries; the certified policy uses the threshold selected only on $\mathcal{D}^{\text{cert}}$.

Policy	Retrieval rate ↓	QA accuracy ↑	Missed stale ↓
Adaptive RAG, budget matched	25.0% (–)	.916 (.0037)	.061 (.0034)
EXPIRE, budget matched	25.0% (–)	.926 (.0034)	.049 (.0030)
EXPIRE, certified ($\alpha = .05$)	26.7% (0.9%)	.929 (.0033)	.047 (.0029)

6 CONCLUDING REMARKS

A frozen LM can answer a fact correctly at an anchor and later become wrong because the world changes. We formulate this as prospective remaining-lifetime prediction under left truncation and interval or right censoring, introduce WIKILIFETIME for rolling-origin evaluation, and propose EXPIRE to combine relation-specific turnover, event history, calendar drift, and relation-centered frozen representations. The theory establishes identification, estimation and calibration guarantees, and risk-controlled retrieval, while simulations and the empirical application show gains in lifetime prediction, within-relation discrimination, and missed-stale control, clarifying how long currently correct knowledge can be trusted before external verification becomes necessary.

AI USE STATEMENT

Generative AI was used for language refinement and editorial polishing. The research design, methodological development, and theoretical derivations were carried out independently by the authors. All AI-assisted edits were subsequently checked and revised by the authors for correctness, clarity, and proper attribution. The authors retain full responsibility for the manuscript and every claim it contains.

REPRODUCIBILITY STATEMENT

The appendix specifies the assumptions and complete proofs, stable interval-censored likelihood, optimization and tuning, benchmark construction and chronology design, simulation data-generating processes, empirical metrics, falsification tests, retrieval evaluation, and ablation studies. To preserve anonymity during double-blind review, code for WIKILIFETIME construction, EXPIRE estimation and recalibration, and reproduction of all figures and tables will be released after peer review.

ACKNOWLEDGMENTS

The authors gratefully acknowledge support from XXX. The authors also thank seminar and conference participants at XXX for helpful comments and discussions.

REFERENCES

Andersen, P. K. and Gill, R. D. (1982). Cox's Regression Model for Counting Processes: A Large Sample Study. *Annals of Statistics*, 10(4), 1100-1120.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., and Schuster, T. (2024). Conformal Risk Control. *International Conference on Learning Representations*.

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., and Lei, L. (2025). Learn Then Test: Calibrating Predictive Algorithms to Achieve Risk Control. *Annals of Applied Statistics*, 19(2), 1641-1662.

Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. (2024). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. *International Conference on Learning Representations*.

Bates, S., Angelopoulos, A. N., Lei, L., Malik, J., and Jordan, M. I. (2021). Distribution-Free, Risk-Controlling Prediction Sets. *Journal of the ACM*, 68(6), 43.

De Cao, N., Aziz, W., and Titov, I. (2021). Editing Factual Knowledge in Language Models. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6491-6506.

Dhingra, B., Cole, J. R., Eisenschlos, J. M., Gillick, D., Eisenstein, J., and Cohen, W. W. (2022). Time-Aware Language Models as Temporal Knowledge Bases. *Transactions of the Association for Computational Linguistics*, 10, 257-273.

Fierro, C., Garneau, N., Bugliarello, E., Kementchedjhieva, Y., and Søgaard, A. (2024). MuLan: A Study of Fact Mutability in Language Models. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 762-771.

Gensheimer, M. F. and Narasimhan, B. (2019). A Scalable Discrete-Time Survival Model for Neural Networks. *PeerJ*, 7, e6257.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and Comparison of Prognostic Classification Schemes for Survival Data. *Statistics in Medicine*, 18, 2529–2545.

Groeneveld, D., Beltagy, I., Walsh, E., Bhagia, A., Kinney, R., Tafjord, O., Jha, A., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., Khot, T., Merrill, W., Morrison, J., Muennighoff, N., Naik, A., Nam, C., Peters, M., Pyatkin, V., Ravichander, A., Schwenk, D., Shah, S., Smith, W., Strubell, E., Subramani, N., Wortsman, M., Dasigi, P., Lambert, N., Richardson, K., Zettlemoyer, L., Dodge, J., Lo, K., Soldaini, L., Smith, N., and Hajishirzi, H. (2024). OLMo: Accelerating the Science of Language Models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 1, 15789-15809.

Guan, X., Zeng, J., Meng, F., Xin, C., Lu, Y., Lin, H., Han, X., Sun, L., and Zhou, J. (2026). DeepRAG: Thinking to Retrieve Step by Step for Large Language Models. *International Conference on Learning Representations*.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. *Proceedings of the 34th International Conference on Machine Learning*, 70, 1321-1330.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M. (2020). Retrieval-Augmented Language Model Pre-Training. *Proceedings of the 37th International Conference on Machine Learning*, 119, 3929-3938.

Hier, D. B. and Obafemi-Ajayi, T. (2026). Reframing Ontology Fact Acquisition During Large Language Model Fine-Tuning as a Time-to-Event Process. *Frontiers in Artificial Intelligence*, 9, 1805890.

Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S., and Grave, E. (2023). Atlas: Few-Shot Learning with Retrieval Augmented Language Models. *Journal of Machine Learning Research*, 24, 1-43.

Jang, J., Ye, S., Lee, C., Yang, S., Shin, J., Han, J., Kim, G., and Seo, M. (2022). TemporalWiki: A Lifelong Benchmark for Training and Evaluating Ever-Evolving Language Models. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 6237-6250.

Jang, J., Ye, S., Yang, S., Shin, J., Han, J., Kim, G., Choi, S. J., and Seo, M. (2022). Towards Continual Knowledge Learning of Language Models. *International Conference on Learning Representations*.

Jeong, S., Baek, J., Cho, S., Hwang, S. J., and Park, J. (2024). Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 7036-7050.

Jiang, Z., Xu, F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., and Neubig, G. (2023). Active Retrieval Augmented Generation. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7969-7992.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. 2nd edition, Wiley.

Kasai, J., Sakaguchi, K., Takahashi, Y., Le Bras, R., Asai, A., Yu, X., Radev, D., Smith, N., Choi, Y., and Inui, K. (2023). RealTime QA: What's the Answer Right Now?. *Advances in Neural Information Processing Systems*, 36.

Kelly, B. T., Malamud, S., Schwab, J., and Xu, T. A. (2026). Scaling Point-in-Time Language Models. National Bureau of Economic Research, Working Papers 35247.

Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., and Lewis, M. (2020). Generalization through Memorization: Nearest Neighbor Language Models. *International Conference on Learning Representations*.

Kvamme, H. and Borgan, Ø. (2021). Continuous and Discrete-Time Survival Prediction with Neural Networks. *Lifetime Data Analysis*, 27, 710-736.

Labruna, T., Campos, J. A., and Azkune, G. (2025). When to Retrieve: Teaching LLMs to Utilize Information Retrieval Effectively. *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing–Natural Language Processing in the Generative AI Era*, 623-632.

Lazaridou, A., Kuncoro, A., Gribovskaya, E., Agrawal, D., Liška, A., Terzi, T., Gimenez, M., de Masson d'Autume, C., Kočiský, T., Ruder, S., Yogatama, D., Cao, K., Young, S., and Blunsom, P. (2021). Mind the Gap: Assessing Temporal Generalization in Neural Language Models. *Advances in Neural Information Processing Systems*, 34.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33.

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., and Hajishirzi, H. (2023). When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 1, 9802-9822.

Meem, J., Rashid, M., Dong, Y., and Hristidis, V. (2024). PAT-Questions: A Self-Updating Benchmark for Present-Anchored Temporal Question-Answering. *Findings of the Association for Computational Linguistics: ACL 2024*, 13129-13148.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. (2022). Locating and Editing Factual Associations in GPT. *Advances in Neural Information Processing Systems*, 35.

Meng, K., Sen Sharma, A., Andonian, A. J., Belinkov, Y., and Bau, D. (2023). Mass-Editing Memory in a Transformer. *International Conference on Learning Representations*.

Mitchell, E., Lin, C., Bosselut, A., Finn, C., and Manning, C. D. (2022). Fast Model Editing at Scale. *International Conference on Learning Representations*.

Nakshatri, N. S., Roy, S., Arivazhagan, M. G., Zhou, H., Kumar, V. B., and Gangadharaiah, R. (2025). When Facts Change: Probing LLMs on Evolving Knowledge with evolveQA. *arXiv:2510.19172v2*.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. (2019). Language Models as Knowledge Bases?. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2463-2473.

Piryani, B., Mert, Z., and Jatowt, A. (2026). How Often Do Answers Change? Estimating Recency Requirements in Question Answering. *arXiv:2603.16544v1*.

Roberts, A., Raffel, C., and Shazeer, N. (2020). How Much Knowledge Can You Pack Into the Parameters of a Language Model? *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 5418-5426.

Saxena, A., Chakrabarti, S., and Talukdar, P. (2021). Question Answering over Temporal Knowledge Graphs. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 6663-6676.

Tsai, W.-Y., Jewell, N. P., and Wang, M.-C. (1987). A Note on the Product-Limit Estimator under Right Censoring and Left Truncation. *Biometrika*, 74(4), 883-886.

Turnbull, B. W. (1976). The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data. *Journal of the Royal Statistical Society Series B*, 38(3), 290-295.

Vrandečić, D. and Krötzsch, M. (2014). Wikidata: A Free Collaborative Knowledgebase. *Communications of the ACM*, 57(10), 78-85.

Vu, T., Iyyer, M., Wang, X., Constant, N., Wei, J., Wei, J., Tar, C., Sung, Y.-H., Zhou, D., Le, Q., and Luong, T. (2024). FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation. *Findings of the Association for Computational Linguistics: ACL 2024*, 13697-13720.

Wallat, J., Nejdl, W., and Sikdar, S. (2026). When Facts Change: Temporal Knowledge Conflict Resolution in LLMs. *Findings of the Association for Computational Linguistics: ACL 2026*, 2154-2184.

Yao, Y., Wang, P., Tian, B., Cheng, S., Li, Z., Deng, S., Chen, H., and Zhang, N. (2023). Editing Large Language Models: Problems, Methods, and Opportunities. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10222-10240.

Yin, X., Jiang, J., Yang, L., and Wan, X. (2024). History Matters: Temporal Knowledge Editing in Large Language Model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17), 19413-19421.

Zhang, M. and Choi, E. (2021). SituatedQA: Incorporating Extra-Linguistic Contexts into QA. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7371-7387.

APPENDIX

Appendix A states the observation model, assumptions, and complete proofs; Appendix B gives estimation and implementation details; Appendix C documents WIKILIFETIME construction; Appendix D gives the complete simulation design and additional results; Appendix E collects empirical metrics, robustness analyses, retrieval evaluation, and ablations; and Appendix F gives model extensions and additional diagnostics. All theoretical arguments use the post-anchor observed record $W = (A, X, L, R, C, \delta)$, and every data-dependent operation is restricted to information available in its designated chronological block.

A THEORETICAL DETAILS AND PROOFS

This part gives the assumptions and complete proofs for Section 4. We use exactly the post-anchor formulation of Sections 2-3. For a retained factual spell, the remaining lifetime is $\tilde{T} = E - \tau$, the anchor age is $A = \tau - B$, and the observed record is $W = (A, X, L, R, C, \delta)$. Here $(L, R]$ is the observed post-anchor expiration interval when $\delta = 1$, while C is the post-anchor administrative follow-up duration when $\delta = 0$. Thus L, R, and C are elapsed durations after the anchor, not calendar dates. For a hazard logit f, we write $\lambda_f(h \mid A, X) = \mathrm{sigm}\{f(A, X, h)\},\ h = 1, \ldots, M$, where $\mathrm{sigm}(v) = (1 + e^{-v})^{-1}$, and define

$$S_f^{\mathrm{res}}(u \mid A, X) = \prod_{h=1}^{u} \{1 - \lambda_f(h \mid A, X)\}, \quad S_f^{\mathrm{res}}(0 \mid A, X) = 1.$$

The observed-data likelihood used throughout the paper is therefore

$$p_f(W \mid A, X) = \begin{cases} S_f^{\mathrm{res}}(L \mid A, X) - S_f^{\mathrm{res}}(R \mid A, X), & \delta = 1, \\ S_f^{\mathrm{res}}(C \mid A, X), & \delta = 0. \end{cases}$$

Left truncation has not disappeared from this formulation. It is incorporated by conditioning on the fact having survived to the anchor. Indeed,

$$S^{\mathrm{res}}(u \mid A, X) = \mathbb{P}(T > A + u \mid T > A, X) = \frac{S^0(A + u \mid X)}{S^0(A \mid X)}.$$

Consequently, for an observed expiration interval,

$$\mathbb{P}(L < \tilde{T} \le R \mid A, X) = \frac{S^0(A + L \mid X) - S^0(A + R \mid X)}{S^0(A \mid X)}.$$

Parameterizing the model directly in post-anchor time simply absorbs this conditioning into S^{res} and avoids carrying the same denominator through every likelihood expression.

A.1 OBSERVATION MODEL, ASSUMPTIONS, AND IDENTIFICATION

We fix (A, X) and the maximum forecast horizon M. On the monthly grid, define the residual-time cell probabilities $\pi_m(A, X) = \mathbb{P}(m - 1 < \tilde{T} \le m \mid A, X),\ m = 1, \ldots, M$, together with the right-tail probability $\pi_\infty(A, X) = \mathbb{P}(\tilde{T} > M \mid A, X)$. We write $\boldsymbol{\pi}(A, X) = \{\pi_1(A, X), \ldots, \pi_M(A, X), \pi_\infty(A, X)\}^\top$. The remaining-survival curve is determined uniquely by this vector

$$S^{\mathrm{res}}(u \mid A, X) = \sum_{m=u+1}^{M} \pi_m(A, X) + \pi_\infty(A, X), \quad u = 0, \ldots, M.$$

The inspection process need not reveal the cell containing $\tilde{T}$ exactly. It may instead reveal a union of adjacent cells, as with an interval $(L, R]$, or only a tail event, as with right censoring. The following assumptions state precisely what is needed to recover the underlying residual-time distribution and to obtain the finite-sample estimation bounds in Section 4.

Assumption 1. *Conditional on (A, X), the mechanism determining inspection intervals and administrative censoring is coarsening at random. Specifically, if an observable inspection outcome w is compatible with two residual event-time cells m and m', then*

$$\mathbb{P}(\Gamma = w \mid \tilde{T} \in m, A, X) = \mathbb{P}(\Gamma = w \mid \tilde{T} \in m', A, X),$$

where Γ denotes the inspection mechanism. The common probability may depend on (A, X) and on w, but not on which compatible residual-time cell contains the event.

Assumption 1 does not require inspection times to be independent of the observed history. Revision frequency, relation type, spell age, calendar date, and other pre-anchor variables may all affect how often a fact is inspected, provided those variables are included in the conditioning information. What the assumption excludes is residual dependence on the unobserved expiration time after conditioning on the available anchor-side information. For example, it would be violated if an unrecorded signal of an imminent replacement systematically caused more frequent observation even after conditioning on X. The observation-time dilation, exact-qualifier, and snapshot-only analyses in Appendix E.2 examine sensitivity to departures from an overly precise observation model.

Assumption 2. *For every (A, X) in the analyzed support, the inspection outcomes occurring with positive probability separate all supported residual-time cells. Moreover, there exist constants $s_{\min} > 0$ and $\pi_{\min} > 0$ such that $S^0(A \mid X) \geq s_{\min}$, and every observed interval or censoring event used in likelihood evaluation has conditional probability at least $\pi_{\min}$.*

The separation requirement is an identifiability condition, not simply a requirement that every expiration be observed exactly. To make it explicit, let $G(A, X)$ be the matrix whose rows correspond to observable inspection outcomes and whose columns correspond to the supported residual-time cells. The entry in row w and column m equals one when cell m is compatible with outcome w and zero otherwise. Separation means that $G(A, X)$ has full column rank after unsupported cells are removed. If two adjacent months can never be distinguished by any inspection pattern, they should be treated as one supported cell; the theory then applies to that coarsened grid.

The lower bound $S^0(A \mid X) \geq s_{\min}$ guarantees that entry at the historical anchor occurs on a region having nonnegligible probability. The bound $\pi_{\min} > 0$ keeps the logarithmic likelihood away from singular interval probabilities and is needed for the uniform Lipschitz and concentration arguments below. These lower bounds are imposed only on the reported analysis support. They do not require every conceivable relation, age, or interval pattern to occur with positive probability.

Assumption 3. *There is a finite B_f such that $|f(A, X, h)| \leq B_f$ for every $f \in \mathcal{F}$, every supported (A, X), and $h = 1, \ldots, M$. The spline bases, event-history vectors, projected representations, and admissible parameter radii are uniformly bounded. Sampling units are independent subject clusters $\mathcal{C}_1, \ldots, \mathcal{C}_n$. Records within a subject cluster may be arbitrarily dependent, but cluster size is bounded by a constant $B_{\mathrm{cl}} < \infty$ independent of n. All records from one subject remain in the same chronological block. Empirical-process arguments are conditional on the training-derived relation centroids, representation projection, spline knots, and other preprocessing maps.*

The boundedness condition controls the stochastic complexity of the fitted hazard class; it does not require the true hazard to belong to that class. Misspecification is allowed and appears through the best-in-class approximation error below. In implementation, the condition corresponds to finite parameter-search ranges, finite forecast horizon, normalized hidden states, and bounded spline designs.

The cluster condition is needed only for the finite-sample estimation statements, not for identification. It explicitly allows repeated anchors or multiple factual records from the same subject without pretending that those records are independent. The effective sample size in the theorem is the number n of independent subject clusters. We now prove Proposition 4.1.

Proof of Proposition 4.1. We fix a supported value of (A, X). Because the proof is conditional on (A, X), we suppress these arguments until the final step. Let $\boldsymbol{\pi}^\star = (\pi_1^\star, \ldots, \pi_M^\star, \pi_\infty^\star)^\top$ denote the true residual-time cell-probability vector and let $\boldsymbol{\pi}$ be any candidate vector on the same supported simplex. Consider an observable inspection outcome w. Such an outcome may be an interval $(L, R]$, a right-censoring event beyond C, or any other admissible coarsening pattern generated by the benchmark. Let $g_w^\top \boldsymbol{\pi}$ denote the probability, under the candidate residual-time distribution, that $\tilde{T}$ lies in one of the cells compatible with w. Here g_w is the corresponding row of the incidence matrix G.

Under Assumption 1, the probability that the inspection mechanism reports w, conditional on the event lying in any compatible cell, is a common quantity that we denote by q_w. Therefore the complete conditional probability of observing outcome w is $p_w(\boldsymbol{\pi}) = q_w \, g_w^\top \boldsymbol{\pi}$. For every supported observable outcome, $q_w > 0$. The true observed-outcome probability is correspondingly $p_w^\star = q_w \, g_w^\top \boldsymbol{\pi}^\star$. The part of the population negative log likelihood that depends on the candidate residual-time distribution is the

cross entropy

$$\mathcal{L}(\boldsymbol{\pi}) = -\sum_{w} p^{\star}_{w} \log p_{w}(\boldsymbol{\pi}),$$

where the sum ranges over observable outcomes having positive true probability. At the true vector,

$$\mathcal{L}(\boldsymbol{\pi}^{\star}) = -\sum_{w} p^{\star}_{w} \log p^{\star}_{w}.$$

Subtracting these two quantities gives

$$\mathcal{L}(\boldsymbol{\pi}) - \mathcal{L}(\boldsymbol{\pi}^{\star}) = \sum_{w} p^{\star}_{w} \log \frac{p^{\star}_{w}}{p_{w}(\boldsymbol{\pi})} = \left(\mathbf{p}^{\star} \,\|\, \mathbf{p}(\boldsymbol{\pi})\right),$$

where $\mathbf{p}^{\star} = (p^{\star}_{w})_{w}$ and $\mathbf{p}(\boldsymbol{\pi}) = \{p_{w}(\boldsymbol{\pi})\}_{w}$.

The Kullback-Leibler divergence is nonnegative, so $\mathcal{L}(\boldsymbol{\pi}) \geq \mathcal{L}(\boldsymbol{\pi}^{\star})$. Equality holds only if $p_{w}(\boldsymbol{\pi}) = p^{\star}_{w}$ for every supported observable outcome w. Since $q_{w} > 0$, equality is equivalent to $g_{w}^{\top}\boldsymbol{\pi} = g_{w}^{\top}\boldsymbol{\pi}^{\star}$ for every such w, or, in matrix form, $G\boldsymbol{\pi} = G\boldsymbol{\pi}^{\star}$. Assumption 2 states that G has full column rank on the supported residual-time subspace. Hence its null space on that subspace is trivial, and $G(\boldsymbol{\pi} - \boldsymbol{\pi}^{\star}) = 0$ implies $\boldsymbol{\pi} = \boldsymbol{\pi}^{\star}$. Thus the conditional observed-data likelihood is uniquely minimized by the true residual-time cell probabilities. The survival probability at horizon u is a deterministic tail sum of these identified probabilities

$$S^{\mathrm{res}}(u \mid A, X) = \sum_{m=u+1}^{M} \pi_{m}(A, X) + \pi_{\infty}(A, X).$$

It follows immediately that $S^{\mathrm{res}}(u \mid A, X)$ is identified at every supported horizon.

Finally, because $\tilde{T} = T - A$, conditioning on entry at the anchor gives

$$S^{\mathrm{res}}(u \mid A, X) = \mathbb{P}(\tilde{T} > u \mid A, X) = \mathbb{P}(T > A + u \mid T > A, X) = \frac{S^{0}(A + u \mid X)}{S^{0}(A \mid X)}.$$

The denominator is well defined by Assumption 2. This completes the identification argument.

□

Before proving the oracle inequality, we record the regularity property of the observed-data loss that will be used repeatedly.

Lemma A.1. *Under Assumptions 2-3, there exist finite constants L_{ℓ} and B_{ℓ}, depending only on M, B_{f}, and $\pi_{\min}$, such that for every supported record W and every $f, g \in \mathcal{F}$,*

$$|\ell_{f}(W) - \ell_{g}(W)| \leq L_{\ell} \,\|(f_{1}, \ldots, f_{M}) - (g_{1}, \ldots, g_{M})\|_{2} \text{ and } 0 \leq \ell_{f}(W) \leq B_{\ell},$$

where $f_{h} = f(A, X, h)$ and $g_{h} = g(A, X, h)$.

Proof. We fix a supported record W and two candidates $f, g \in \mathcal{F}$. We write $f_{h} = f(A, X, h)$, $g_{h} = g(A, X, h)$, and abbreviate $\lambda_{f,h} = \mathrm{sigm}(f_{h})$, $\lambda_{g,h} = \mathrm{sigm}(g_{h})$. Also define $S_{f}(u) = \prod_{h=1}^{u}(1 - \lambda_{f,h})$ and $S_{g}(u) = \prod_{h=1}^{u}(1 - \lambda_{g,h})$, with $S_{f}(0) = S_{g}(0) = 1$. We use $\mathrm{softplus}(v) = \log(1 + e^{v})$.

We first record a bound for the difference between two survival probabilities. For any $u \leq M$, the telescoping identity for products gives

$$S_{f}(u) - S_{g}(u) = \sum_{k=1}^{u} \big(\lambda_{g,k} - \lambda_{f,k}\big) \prod_{h<k} (1 - \lambda_{f,h}) \prod_{h>k} (1 - \lambda_{g,h}).$$

All product factors lie in $[0, 1]$, and the sigmoid function is $1/4$-Lipschitz because $\sup_{v \in \mathbb{R}} \mathrm{sigm}(v)\{1 - \mathrm{sigm}(v)\} = \frac{1}{4}$. Consequently,

$$|S_{f}(u) - S_{g}(u)| \leq \sum_{k=1}^{u} |\lambda_{f,k} - \lambda_{g,k}| \leq \frac{1}{4} \sum_{k=1}^{u} |f_{k} - g_{k}|.$$

Suppose first that $\delta = 1$. The two observed interval probabilities are $p_{f}(W) = S_{f}(L) - S_{f}(R)$ and

$p_g(W) = S_g(L) - S_g(R)$. Therefore,

$$|p_f(W) - p_g(W)| \le |S_f(L) - S_g(L)| + |S_f(R) - S_g(R)| \le \frac{1}{4}\sum_{h=1}^{L}|f_h - g_h| + \frac{1}{4}\sum_{h=1}^{R}|f_h - g_h|$$
$$\le \frac{1}{2}\sum_{h=1}^{M}|f_h - g_h| \le \frac{\sqrt{M}}{2}\,\|(f_1,\dots,f_M) - (g_1,\dots,g_M)\|_2\,.$$

By Assumption 2, every supported observed interval probability entering the likelihood is bounded below by $\pi_{\min}$. Since the derivative of $-\log x$ has absolute value at most $1/\pi_{\min}$ on $[\pi_{\min}, 1]$,

$$|\ell_f(W) - \ell_g(W)| = |-\log p_f(W) + \log p_g(W)|$$
$$\le \frac{1}{\pi_{\min}}|p_f(W) - p_g(W)| \le \frac{\sqrt{M}}{2\pi_{\min}}\,\|(f_1,\dots,f_M) - (g_1,\dots,g_M)\|_2\,.$$

Now suppose that $\delta = 0$. Then $p_f(W) = S_f(C)$ and hence

$$\ell_f(W) = -\log S_f(C) = -\sum_{h=1}^{C}\log\{1 - \mathrm{sigm}(f_h)\} = \sum_{h=1}^{C}\mathrm{softplus}(f_h).$$

Because $\frac{d}{dv}\,\mathrm{softplus}(v) = \mathrm{sigm}(v) \in [0,1]$, the softplus function is 1-Lipschitz. Therefore,

$$|\ell_f(W) - \ell_g(W)| \le \sum_{h=1}^{C}|f_h - g_h| \le \sqrt{M}\,\|(f_1,\dots,f_M) - (g_1,\dots,g_M)\|_2\,.$$

Combining the interval-censored and right-censored cases, we may take $L_\ell = \sqrt{M}\max\{1, \frac{1}{2\pi_{\min}}\}$.

It remains to establish a uniform upper bound on the loss. Since every likelihood contribution is a probability, $\ell_f(W) \ge 0$. If $\delta = 1$, Assumption 2 gives $p_f(W) \ge \pi_{\min}$, and therefore

$$\ell_f(W) = -\log p_f(W) \le \log\frac{1}{\pi_{\min}}.$$

If $\delta = 0$, Assumption 3 gives $f_h \le B_f$ for every h, and softplus is increasing. Hence

$$\ell_f(W) = \sum_{h=1}^{C}\mathrm{softplus}(f_h) \le M\ \mathrm{softplus}(B_f).$$

Thus we may take

$$B_\ell = \max\left\{\log\frac{1}{\pi_{\min}},\ M\ \mathrm{softplus}(B_f)\right\}.$$

Both L_ℓ and B_ℓ depend only on M, B_f, and $\pi_{\min}$ and are finite under Assumptions 2-3. This proves the result.

□

A.2 Oracle Inequality, Validation Selection, and Separation

We next prove Theorem 4.2. Because the empirical study may contain several records from one subject, we write the proof directly at the cluster level and avoid first pretending that every row is independent. Let $\mathcal{C}_j = \{W_{j1},\dots,W_{jm_j}\}$, $j = 1,\dots,n$, denote the independent subject clusters and let $N = \sum_{j=1}^{n} m_j$ be the total number of factual records. Conditional on the cluster sizes, define $w_j = m_j/N$ and the within-cluster average loss $\bar{\ell}_f(\mathcal{C}_j) = m_j^{-1}\sum_{k=1}^{m_j}\ell_f(W_{jk})$. Then the ordinary record-average empirical loss used by the estimator can be written exactly as $\hat{\mathcal{R}}_n(f) = \sum_{j=1}^{n} w_j\bar{\ell}_f(\mathcal{C}_j)$. Because $m_j \le B_{\mathrm{cl}}$ and $N \ge n$,

$$\max_j w_j \le \frac{B_{\mathrm{cl}}}{n} \text{ and } \sum_{j=1}^{n} w_j^2 \le \frac{B_{\mathrm{cl}}}{n}.$$

Thus clustering changes constants but not the $n^{-1/2}$ stochastic rate.

For the induced cluster-loss class, define the weighted empirical Rademacher complexity

$$\mathfrak{R}_n(\mathcal{L}_\mathcal{F}) = \mathbb{E}_\varepsilon \left[\sup_{f\in\mathcal{F}} \sum_{j=1}^{n} w_j \varepsilon_j \bar{\ell}_f(\mathcal{C}_j) \,\middle|\, \mathcal{C}_1, \ldots, \mathcal{C}_n \right],$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent Rademacher signs. Let ϑ parameterize $f_\vartheta \in \mathcal{F}$, and let $\mathcal{P}_\lambda(\vartheta)$ denote the nonnegative penalty appearing in the training objective of Section 3. An $\varepsilon_{\mathrm{opt}}$-approximate empirical minimizer satisfies

$$\hat{\mathcal{R}}_n(f_{\hat{\vartheta}}) + \mathcal{P}_\lambda(\hat{\vartheta}) \le \inf_{\vartheta : f_\vartheta \in \mathcal{F}} \left\{ \hat{\mathcal{R}}_n(f_\vartheta) + \mathcal{P}_\lambda(\vartheta) \right\} + \varepsilon_{\mathrm{opt}}.$$

Proof of Theorem 4.2. We write $\Delta_n = \sup_{f\in\mathcal{F}} |\mathcal{R}(f) - \hat{\mathcal{R}}_n(f)|$. For any admissible parameter value ϑ, $\mathcal{R}(f_{\hat{\vartheta}}) \le \hat{\mathcal{R}}_n(f_{\hat{\vartheta}}) + \Delta_n$. Approximate optimality implies

$$\hat{\mathcal{R}}_n(f_{\hat{\vartheta}}) + \mathcal{P}_\lambda(\hat{\vartheta}) \le \hat{\mathcal{R}}_n(f_\vartheta) + \mathcal{P}_\lambda(\vartheta) + \varepsilon_{\mathrm{opt}}.$$

Therefore

$$\mathcal{R}(f_{\hat{\vartheta}}) \le \hat{\mathcal{R}}_n(f_\vartheta) + \mathcal{P}_\lambda(\vartheta) - \mathcal{P}_\lambda(\hat{\vartheta}) + \varepsilon_{\mathrm{opt}} + \Delta_n \le \hat{\mathcal{R}}_n(f_\vartheta) + \mathcal{P}_\lambda(\vartheta) + \varepsilon_{\mathrm{opt}} + \Delta_n,$$

because the penalty is nonnegative. Applying the definition of Δ_n once more yields

$$\mathcal{R}(f_{\hat{\vartheta}}) \le \mathcal{R}(f_\vartheta) + \mathcal{P}_\lambda(\vartheta) + 2\Delta_n + \varepsilon_{\mathrm{opt}}.$$

Thus the remaining task is to control Δ_n uniformly over $\mathcal{F}$.

We introduce an independent ghost sample $\mathcal{C}'_1, \ldots, \mathcal{C}'_n$ having the same conditional cluster distributions and the same fixed cluster weights. By Jensen's inequality,

$$\mathbb{E}\Delta_n = \mathbb{E} \sup_{f\in\mathcal{F}} \left| \mathbb{E}_{\mathcal{C}'} \sum_{j=1}^{n} w_j \{ \bar{\ell}_f(\mathcal{C}'_j) - \bar{\ell}_f(\mathcal{C}_j) \} \right| \le \mathbb{E} \sup_{f\in\mathcal{F}} \left| \sum_{j=1}^{n} w_j \{ \bar{\ell}_f(\mathcal{C}'_j) - \bar{\ell}_f(\mathcal{C}_j) \} \right|.$$

Because each pair $(\mathcal{C}_j, \mathcal{C}'_j)$ is exchangeable, multiplying its difference by an independent Rademacher sign does not change its distribution. Consequently,

$$\begin{aligned} \mathbb{E}\Delta_n &\le \mathbb{E} \sup_{f\in\mathcal{F}} \left| \sum_{j=1}^{n} w_j \varepsilon_j \{ \bar{\ell}_f(\mathcal{C}'_j) - \bar{\ell}_f(\mathcal{C}_j) \} \right| \\ &\le \mathbb{E} \sup_{f\in\mathcal{F}} \left| \sum_{j=1}^{n} w_j \varepsilon_j \bar{\ell}_f(\mathcal{C}'_j) \right| + \mathbb{E} \sup_{f\in\mathcal{F}} \left| \sum_{j=1}^{n} w_j \varepsilon_j \bar{\ell}_f(\mathcal{C}_j) \right| = 2\,\mathbb{E}\{\mathfrak{R}_n(\mathcal{L}_\mathcal{F})\}. \end{aligned}$$

We also need a high-probability, not an expectation, bound. By Lemma A.1, $0 \le \bar{\ell}_f(\mathcal{C}_j) \le B_\ell$. If one independent cluster is replaced while the cluster sizes and hence the weights are held fixed, the value of $\hat{\mathcal{R}}_n(f)$ changes by at most $w_j B_\ell$ for every f. Therefore the supremum Δ_n changes by at most the same order. The bounded-differences inequality gives, with probability at least $1 - \eta/2$,

$$\Delta_n \le \mathbb{E}\Delta_n + B_\ell \sqrt{\frac{\log(2/\eta)}{2} \sum_{j=1}^{n} w_j^2}.$$

Using $\sum_j w_j^2 \le B_{\mathrm{cl}}/n$ gives

$$\Delta_n \le \mathbb{E}\Delta_n + B_\ell \sqrt{\frac{B_{\mathrm{cl}} \log(2/\eta)}{2n}}.$$

The empirical Rademacher complexity itself is also a bounded-difference function of the independent clusters. Replacing $\mathcal{C}_j$ changes it by at most a constant multiple of $w_j B_\ell$. Applying the same argument and combining the two deviation events shows that, with probability at least $1 - \eta$,

$$\Delta_n \le C_0 \mathfrak{R}_n(\mathcal{L}_\mathcal{F}) + C'_0 \sqrt{\frac{\log(2/\eta)}{n}},$$

where C_0, C_0' depend only on B_ℓ and the cluster-size bound. Substituting this uniform deviation bound into the basic inequality above gives

$$\mathcal{R}(f_{\hat{\vartheta}}) \le \mathcal{R}(f_\vartheta) + \mathcal{P}_\lambda(\vartheta) + C_1\mathfrak{R}_n(\mathcal{L}_\mathcal{F}) + C_2\sqrt{\frac{\log(2/\eta)}{n}} + \varepsilon_{\mathrm{opt}}.$$

Because this inequality holds for every admissible ϑ, taking the infimum proves the first part of Theorem 4.2.

We now derive the complexity specialization for the EXPIRE class. Conditional on all preprocessing learned from the training block, the hazard logit consists of a global intercept, relation-specific age and calendar spline terms, the p-dimensional event-history term, and the semantic interaction $(U^\top \tilde{z})^\top (V^\top B_s(h))$. The numbers of free coordinates in these blocks are respectively $1,\ R_{\mathrm{rel}}K_a,\ R_{\mathrm{rel}}K_c,\ p,\ d_p q,\ K_s q$. Hence the total parameter count in this factorized representation is

$$D = 1 + R_{\mathrm{rel}}(K_a + K_c) + p + q(d_p + K_s).$$

Under Assumption 3, each coefficient block lies in a bounded Euclidean or Frobenius ball. A d-dimensional Euclidean ball of radius R admits an ϵ-net with cardinality no larger than $(1 + 2R/\epsilon)^d$. Applying this construction independently to the intercept, relation spline, history, U, and V blocks and multiplying the resulting covering numbers gives a logarithmic parameter-space covering bound of order $D\log(C/\epsilon)$ for a constant C depending only on the parameter radii.

It remains to verify that closeness in the factor parameters implies closeness of the hazard logit. The linear and spline components are Lipschitz in their coefficient vectors because their feature vectors are uniformly bounded. For the semantic term, let $\phi_{U,V}(\tilde{z}, h) = (U^\top \tilde{z})^\top (V^\top B_s(h))$. For two parameter pairs (U, V) and (U', V'),

$$\begin{aligned}|\phi_{U,V}(\tilde{z},h) - \phi_{U',V'}(\tilde{z},h)| &\le \left|\{(U-U')^\top \tilde{z}\}^\top V^\top B_s(h)\right| + \left|(U'^\top \tilde{z})^\top \{(V-V')^\top B_s(h)\}\right| \\ &\le \|\tilde{z}\|_2 \|B_s(h)\|_2 \left\{\|U-U'\|_F \|V\|_F + \|U'\|_F \|V-V'\|_F\right\}.\end{aligned}$$

All factors multiplying the parameter differences are uniformly bounded by Assumption 3. Therefore the parameter-space net induces a uniform net for the hazard-logit class with the same effective dimension D, up to constants. Lemma A.1 transfers this covering bound from hazard logits to the observed-data loss class. Integrating the resulting entropy bound over the covering radius gives

$$\mathfrak{R}_n(\mathcal{L}_\mathcal{F}) \le C\sqrt{\frac{D\,\log(C'n)}{n}}.$$

The logarithmic term, the bounded parameter radii, feature norms, cluster-size bound, forecast horizon, and interval-positivity constants are exactly the quantities absorbed by the $\lesssim$ notation in the main paper. Hence

$$\mathfrak{R}_n(\mathcal{L}_\mathcal{F}) \lesssim \sqrt{\frac{q(d_p + K_s) + R_{\mathrm{rel}}(K_a + K_c) + p + 1}{n}}.$$

This proves the complexity statement and completes the proof of Theorem 4.2.

□

The theorem is best-in-class. If the true hazard $f^\star$ does not belong to $\mathcal{F}$, let $f^\star_\mathcal{F}$ minimize population risk over $\mathcal{F}$ and define the approximation error $\mathcal{A}(\mathcal{F}) = \mathcal{R}(f^\star_\mathcal{F}) - \mathcal{R}(f^\star)$. Applying Theorem 4.2 to a parameterization $\vartheta^\star_\mathcal{F}$ of $f^\star_\mathcal{F}$ gives

$$\mathcal{R}(f_{\hat{\vartheta}}) - \mathcal{R}(f^\star) \le \mathcal{A}(\mathcal{F}) + \mathcal{P}_\lambda(\vartheta^\star_\mathcal{F}) + C_1\mathfrak{R}_n(\mathcal{L}_\mathcal{F}) + C_2\sqrt{\frac{\log(2/\eta)}{n}} + \varepsilon_{\mathrm{opt}}.$$

Thus the low-rank semantic pathway trades approximation error against finite-sample complexity and does not assume that the expiration hazard has exactly the EXPIRE form. The main analysis chooses among a finite collection of prespecified configuration-epoch predictors on the chronologically later validation block. The following result makes explicit what this selection step guarantees.

Proposition A.2. *Let $\hat{f}_1, \dots, \hat{f}_K$ be candidate predictors constructed without using $\mathcal{D}^{\mathrm{val}}$, and suppose their observed-data losses are bounded by B_ℓ. Conditional on the fitted candidates, let $\hat{k} \in$*

$\arg\min_{1\le k\le K} \hat{\mathcal{R}}_{\mathrm{val}}(\hat{f}_k)$ *be the candidate checkpoint selected by held-out interval-censored negative log likelihood. If the validation subject clusters are independent, then with probability at least* $1-\eta$*,*

$$\mathcal{R}_{\mathrm{val}}(\hat{f}_{\hat{k}}) \le \min_{1\le k\le K} \mathcal{R}_{\mathrm{val}}(\hat{f}_k) + CB_\ell \sqrt{\frac{\log(2K/\eta)}{n_{\mathrm{val}}}},$$

where C *depends only on the validation cluster-size bound.*

Proof. Condition on the training data and hence on $\hat{f}_1, \dots, \hat{f}_K$. For each fixed candidate k, the validation loss is an average over independent subject clusters and is bounded by B_ℓ. The bounded-differences inequality therefore gives

$$\mathbb{P}\left(\left|\hat{\mathcal{R}}_{\mathrm{val}}(\hat{f}_k) - \mathcal{R}_{\mathrm{val}}(\hat{f}_k)\right| > t\right) \le 2\exp\left\{-\frac{c\, n_{\mathrm{val}} t^2}{B_\ell^2}\right\},$$

where $c > 0$ depends only on the bounded cluster size.

Applying the union bound over the K candidates shows that, with probability at least $1-\eta$,

$$\max_{1\le k\le K}\left|\hat{\mathcal{R}}_{\mathrm{val}}(\hat{f}_k) - \mathcal{R}_{\mathrm{val}}(\hat{f}_k)\right| \le C_0 B_\ell \sqrt{\frac{\log(2K/\eta)}{n_{\mathrm{val}}}}.$$

Let $k^\star \in \arg\min_k \mathcal{R}_{\mathrm{val}}(\hat{f}_k)$. On this simultaneous event,

$$\begin{aligned}\mathcal{R}_{\mathrm{val}}(\hat{f}_{\hat{k}}) &\le \hat{\mathcal{R}}_{\mathrm{val}}(\hat{f}_{\hat{k}}) + C_0 B_\ell \sqrt{\frac{\log(2K/\eta)}{n_{\mathrm{val}}}} \\ &\le \hat{\mathcal{R}}_{\mathrm{val}}(\hat{f}_{k^\star}) + C_0 B_\ell \sqrt{\frac{\log(2K/\eta)}{n_{\mathrm{val}}}} \le \mathcal{R}_{\mathrm{val}}(\hat{f}_{k^\star}) + 2C_0 B_\ell \sqrt{\frac{\log(2K/\eta)}{n_{\mathrm{val}}}}.\end{aligned}$$

Absorbing the factor two into C proves the result.

□

This proposition does not assume that chronological distributions are invariant across calendar time. After the candidate class is selected, the hazard model is refitted using $\mathcal{D}^{\mathrm{tr}} \cup \mathcal{D}^{\mathrm{val}}$ for the epoch count selected during the original training-validation fit; neither $\mathcal{D}^{\mathrm{recal}}$ nor any later block is used for early stopping or hazard-model selection. We next prove the relation-recency separation result.

Proof of Theorem 4.3. We fix the forecast horizon u and condition on the stratum $G = g$, where G contains exactly the relation identity and recency information available to the class of relation-recency rules considered in the theorem. Define $Y_u = \mathbf{1}\{\tilde{T} > u\}$. By assumption,

$$\mathbb{P}(Z = 0 \mid G = g) = \mathbb{P}(Z = 1 \mid G = g) = \frac{1}{2},$$

and

$$\mathbb{E}(Y_u \mid G = g, Z = 0) = s_0, \quad \mathbb{E}(Y_u \mid G = g, Z = 1) = s_1.$$

Any predictor measurable only with respect to G is constant within the fixed stratum $G = g$. Let its value be c. Conditional squared-error risk is $\mathbb{E}\{(Y_u - c)^2 \mid G = g\}$. The minimizer of squared loss is the conditional mean,

$$c^\star = \mathbb{E}(Y_u \mid G = g) = \frac{s_0 + s_1}{2}.$$

A predictor that additionally observes the subtype Z uses the Bayes conditional mean

$$s_Z = \begin{cases} s_0, & Z = 0, \\ s_1, & Z = 1. \end{cases}$$

For squared loss, the conditional-mean decomposition gives

$$\mathbb{E}\{(Y_u - c^\star)^2 \mid G = g\} = \mathbb{E}\{(Y_u - s_Z)^2 \mid G = g\} + \mathbb{E}\{(s_Z - c^\star)^2 \mid G = g\}.$$

To verify the identity directly, we write $Y_u - c^\star = (Y_u - s_Z) + (s_Z - c^\star)$ and square. The cross term has conditional expectation zero because $\mathbb{E}(Y_u - s_Z \mid G, Z) = 0$. The excess risk is therefore

$$\mathbb{E}\{(s_Z - c^\star)^2 \mid G = g\} = \frac{1}{2}\left(s_0 - \frac{s_0 + s_1}{2}\right)^2 + \frac{1}{2}\left(s_1 - \frac{s_0 + s_1}{2}\right)^2 = \frac{(s_1 - s_0)^2}{4}.$$

Because $s_0 \neq s_1$, this gap is strictly positive and does not disappear with sample size for a predictor restricted to G.

For the second part of the theorem, let $\hat{Z}$ be a representation-based subtype classifier satisfying $\mathbb{P}(\hat{Z} \neq Z \mid G = g) \to 0$, and let $\hat{s}_0, \hat{s}_1$ consistently estimate the two subtype-specific survival probabilities. Consider $\hat{s}_{\hat{Z}}$. For every observation,

$$|\hat{s}_{\hat{Z}} - s_Z| \leq \max_{z \in \{0,1\}} |\hat{s}_z - s_z| + \mathbf{1}\{\hat{Z} \neq Z\}.$$

Squaring and using $(a + b)^2 \leq 2a^2 + 2b^2$ gives

$$\mathbb{E}\{(\hat{s}_{\hat{Z}} - s_Z)^2 \mid G = g\} \leq 2\mathbb{E}\left[\max_z |\hat{s}_z - s_z|^2 \mid G = g\right] + 2\mathbb{P}(\hat{Z} \neq Z \mid G = g).$$

Both terms converge to zero by the stated consistency conditions.

Finally, for any predictor $q(G, Z)$, the squared-loss excess risk relative to the subtype-aware Bayes rule satisfies

$$\mathbb{E}\{(Y_u - q)^2 - (Y_u - s_Z)^2 \mid G = g\} = \mathbb{E}\{(q - s_Z)^2 \mid G = g\}.$$

Taking $q = \hat{s}_{\hat{Z}}$ shows that the representation-aware predictor's risk converges to the subtype-aware Bayes risk. The relation-recency gap remains $(s_1 - s_0)^2/4$, whereas the additional error associated with the residual representation vanishes.

□

A.3 Stability, Calibration, Certification, and Decision Risk

The remaining results connect the fitted hazard to the quantities used for probabilistic reporting and deployment.

Proof of Lemma 4.4. For arbitrary sequences $a_1, \ldots, a_u$ and $b_1, \ldots, b_u$, the difference between two products admits the telescoping representation

$$\prod_{h=1}^{u} a_h - \prod_{h=1}^{u} b_h = \sum_{k=1}^{u} \left(\prod_{h<k} a_h\right)(a_k - b_k)\left(\prod_{h>k} b_h\right).$$

This identity follows by replacing the factors one at a time: first replace b_1 by a_1, then b_2 by a_2, and continue through u. If all $a_h, b_h \in [0, 1]$, every product multiplying $(a_k - b_k)$ has absolute value at most one. Hence

$$\left|\prod_{h=1}^{u} a_h - \prod_{h=1}^{u} b_h\right| \leq \sum_{k=1}^{u} |a_k - b_k|.$$

Take $a_h = 1 - \lambda_f(h \mid A, X)$ and $b_h = 1 - \lambda_g(h \mid A, X)$. By the post-anchor product representation, $\prod_{h=1}^{u} a_h = S_f^{\mathrm{res}}(u \mid A, X)$ and analogously for g. Therefore

$$\left|S_f^{\mathrm{res}}(u \mid A, X) - S_g^{\mathrm{res}}(u \mid A, X)\right| \leq \sum_{h=1}^{u} \left|\lambda_f(h \mid A, X) - \lambda_g(h \mid A, X)\right|.$$

The derivative of the logistic function is $\mathrm{sigm}(v)\{1 - \mathrm{sigm}(v)\}$, whose maximum over $v \in \mathbb{R}$ is $1/4$. By the mean-value theorem,

$$|\,\mathrm{sigm}(x) - \mathrm{sigm}(y)| \leq \frac{1}{4}|x - y|.$$

Applying this inequality at each horizon gives

$$\left|S_f^{\mathrm{res}}(u \mid A, X) - S_g^{\mathrm{res}}(u \mid A, X)\right| \leq \frac{1}{4}\sum_{h=1}^{u} |f(A, X, h) - g(A, X, h)|.$$

This completes the proof.

□

Proof of Proposition 4.5. We fix the horizon u and define $Y_u = \mathbf{1}\{\tilde{T} > u\}$ and $S_u^\star = S^{\mathrm{res},\star}(u \mid A, X)$. By definition of the true conditional survival probability, $S_u^\star = \mathbb{E}(Y_u \mid A, X)$. Condition on the data used to fit the predictor. For a fresh evaluation record, $\hat{S}_u$ is then a measurable function of (A, X). Applying the tower property,

$$\mathbb{E}(Y_u \mid \hat{S}_u) = \mathbb{E}\{\mathbb{E}(Y_u \mid A, X) \mid \hat{S}_u\} = \mathbb{E}(S_u^\star \mid \hat{S}_u).$$

Hence

$$\mathrm{CE}_u(\hat{S}) = \mathbb{E}\left|\mathbb{E}(Y_u \mid \hat{S}_u) - \hat{S}_u\right| = \mathbb{E}\left|\mathbb{E}(S_u^\star - \hat{S}_u \mid \hat{S}_u)\right|.$$

Conditional Jensen's inequality gives

$$\left|\mathbb{E}(S_u^\star - \hat{S}_u \mid \hat{S}_u)\right| \le \mathbb{E}(|S_u^\star - \hat{S}_u| \mid \hat{S}_u).$$

Taking expectation once more yields $\mathrm{CE}_u(\hat{S}) \le \mathbb{E}|S_u^\star - \hat{S}_u|$. Thus horizon-specific calibration error cannot exceed the corresponding L_1 survival-probability error.

□

We next prove the finite-sample retrieval certificate in Theorem 4.6. For fixed u, let

$$L_i^{\mathrm{ms}}(\rho; u) = \mathbf{1}\{d_\rho(X_i; u) = 0\}\mathbf{1}\{\tilde{T}_i \le u\}.$$

With the threshold convention

$$d_\rho(X_i; u) = \mathbf{1}\{\hat{p}_i(u) \ge 1 - \rho\},$$

increasing ρ can only change a direct answer into retrieval. Thus $L_i^{\mathrm{ms}}(\rho; u)$ is non-increasing and right-continuous in ρ. It also lies in $[0, 1]$.

Proof of Theorem 4.6. Condition on all data and random choices used for model fitting, model selection, refitting, and recalibration. After this conditioning, the expiration score $\hat{p}(\cdot)$ is fixed. The only remaining random quantities are the certificate examples and the future example. Let $n = n_{\mathrm{cert}}$. Consider $n+1$ exchangeable examples indexed by $1, \ldots, n+1$. For each i, define the threshold obtained by treating example i as the future observation and using the other n examples for certification

$$\hat{\rho}_{-i} = \inf\left\{\rho : \frac{\sum_{j \ne i} L_j^{\mathrm{ms}}(\rho; u) + 1}{n+1} \le \alpha\right\}.$$

The feasible set is nonempty. At $\rho = 1$, the rule always retrieves because $\hat{p}_i(u) \ge 0 = 1 - 1$, so $L_j^{\mathrm{ms}}(1; u) = 0$ for every j. Therefore the quantity inside the selector equals $1/(n+1)$ at $\rho = 1$. The assumption $\alpha \ge \frac{1}{n+1}$ guarantees feasibility.

Now define the threshold that would be obtained if all $n+1$ losses were available without the finite-sample correction

$$\rho^\dagger = \inf\left\{\rho : \frac{1}{n+1}\sum_{j=1}^{n+1} L_j^{\mathrm{ms}}(\rho; u) \le \alpha\right\}.$$

For any fixed i, the defining property of $\hat{\rho}_{-i}$ gives

$$\frac{\sum_{j \ne i} L_j^{\mathrm{ms}}(\hat{\rho}_{-i}; u) + 1}{n+1} \le \alpha.$$

Since $L_i^{\mathrm{ms}}(\hat{\rho}_{-i}; u) \le 1$, we also have

$$\frac{1}{n+1}\sum_{j=1}^{n+1} L_j^{\mathrm{ms}}(\hat{\rho}_{-i}; u) \le \alpha.$$

Thus $\hat{\rho}_{-i}$ belongs to the feasible set defining $\rho^\dagger$. Because each loss is non-increasing in ρ, that feasible set is an interval extending to one, and therefore $\hat{\rho}_{-i} \ge \rho^\dagger$. Monotonicity of the loss then gives $L_i^{\mathrm{ms}}(\hat{\rho}_{-i}; u) \le L_i^{\mathrm{ms}}(\rho^\dagger; u)$. Averaging over all $n+1$ possible held-out indices,

$$\frac{1}{n+1}\sum_{i=1}^{n+1} L_i^{\mathrm{ms}}(\hat{\rho}_{-i}; u) \le \frac{1}{n+1}\sum_{i=1}^{n+1} L_i^{\mathrm{ms}}(\rho^\dagger; u) \le \alpha.$$

The construction of $\hat{\rho}_{-i}$ is permutation equivariant; relabeling the $n+1$ examples merely relabels the corresponding leave-one-out thresholds. Exchangeability therefore implies that every term $L_i^{\mathrm{ms}}(\hat{\rho}_{-i}; u)$ has the same expectation. Consequently,

$$\mathbb{E}L_{n+1}^{\mathrm{ms}}(\hat{\rho}_{-(n+1)}; u) = \mathbb{E}\left[\frac{1}{n+1}\sum_{i=1}^{n+1} L_i^{\mathrm{ms}}(\hat{\rho}_{-i}; u)\right] \leq \alpha.$$

When $i = n+1$, the other n observations are exactly the certificate sample. Hence $\hat{\rho}_{-(n+1)} = \hat{\rho}_u$ defined by the certification rule in Section 3. Therefore

$$\mathbb{E}\left[L_{n_{\mathrm{cert}}+1}^{\mathrm{ms}}(\hat{\rho}_u; u)\right] \leq \alpha.$$

All expectations above are conditional on the fitting, selection, and recalibration stages. Since none of those stages uses $\mathcal{D}^{\mathrm{cert}}$, removing that conditioning preserves the same inequality. The argument does not require exchangeability between the training periods and the certificate period; it requires exchangeability only between the certificate examples and the future example in the deployment regime being certified. This is why Theorem 4.6 does not claim protection under arbitrary later calendar drift.

□

Proof of Corollary 4.7. We fix X and suppress the horizon u from the notation. The conditional expected loss of retrieving is c_R, while the conditional expected loss of answering directly is $c_E p^\star$. Thus retrieval is optimal when $c_R < c_E p^\star$, or equivalently when $p^\star > \tau,\ \tau = c_R/c_E$. The Bayes decision and plug-in decision are therefore $d^\star = \mathbf{1}\{p^\star > \tau\}$ and $\hat{d} = \mathbf{1}\{\hat{p} > \tau\}$. If $\hat{d} = d^\star$, the conditional excess decision risk is zero. Suppose instead that the two rules disagree. Then $p^\star$ and $\hat{p}$ lie on opposite sides of τ. Hence the threshold τ lies between them and $|p^\star - \tau| \leq |\hat{p} - p^\star|$. The absolute conditional cost difference between retrieval and direct answering is $|c_R - c_E p^\star| = c_E|p^\star - \tau|$. When the plug-in rule chooses the wrong action, its conditional excess loss is exactly this quantity. Therefore, whether or not the two rules disagree,

$$\mathrm{Risk}(\hat{d} \mid X) - \mathrm{Risk}(d^\star \mid X) \leq c_E|\hat{p} - p^\star|.$$

Taking expectation over X gives

$$\mathrm{Risk}(\hat{d}) - \mathrm{Risk}(d^\star) \leq c_E\mathbb{E}|\hat{p} - p^\star|.$$

Finally, $\hat{p} = 1 - \hat{S}$ and $p^\star = 1 - S^\star$, so $|\hat{p} - p^\star| = |\hat{S} - S^\star|$. Hence

$$\mathrm{Risk}(\hat{d}) - \mathrm{Risk}(d^\star) \leq c_E\mathbb{E}|\hat{S} - S^\star|,$$

which proves Corollary 4.7.

□

B Estimation and Implementation Details

This part gives the numerical details needed to reproduce the estimation procedure in Section 3. The implementation follows the post-anchor formulation throughout: the model is evaluated on forecast months $h = 1, \ldots, M$, interval endpoints L_i and R_i and censoring duration C_i are measured from the anchor, and no expiration time is imputed inside an observed interval. All representation preprocessing and spline construction used during model selection are learned from $\mathcal{D}^{\mathrm{tr}}$ only. The chronologically later blocks are used only for the purposes assigned to them in Algorithm 1.

B.1 Stable Likelihood and Parameterization

For an interval-censored record, the likelihood contribution in Section 3 is $p_f(W_i) = S_{f,i}^{\mathrm{res}}(L_i) - S_{f,i}^{\mathrm{res}}(R_i)$. Direct subtraction is numerically unstable when both survival probabilities are small or nearly equal. We thus work entirely on the log scale. Set $a_i = \log S_{f,i}^{\mathrm{res}}(L_i)$ and $b_i = \log S_{f,i}^{\mathrm{res}}(R_i)$. Because survival is non-increasing and $L_i < R_i$, we have $a_i \geq b_i$, and

$$\log p_f(W_i) = a_i + \log\{1 - \exp(b_i - a_i)\}.$$

The second term is evaluated by a stable `log1mexp` operation. For $x \le 0$, we use

$$\operatorname{log1mexp}(x) = \begin{cases} \log\{-\operatorname{expm1}(x)\}, & -\log 2 \le x \le 0, \\ \operatorname{log1p}\{-\exp(x)\}, & x < -\log 2. \end{cases}$$

This avoids cancellation when x is close to zero. Remaining log survival is itself accumulated without explicit products. Writing $f_i(h) = f(A_i, X_i, h)$,

$$\log S^{\mathrm{res}}_{f,i}(u) = \sum_{h=1}^{u} \log\left[1 - \operatorname{sigm}\{f_i(h)\}\right] = -\sum_{h=1}^{u} \operatorname{softplus}\{f_i(h)\}.$$

Thus an interval-censored loss requires only two cumulative log-survival values. For a right-censored record, $-\log p_f(W_i) = -\log S^{\mathrm{res}}_{f,i}(C_i)$, which is obtained from the same cumulative sum. No product of many near-one probabilities is ever formed explicitly.

The modeling horizon is M. Every event interval entering the primary likelihood satisfies $0 \le L_i < R_i \le M$. If an observed interval extends beyond M, only the survival information identified inside the modeled window is retained: an interval with $L_i < M < R_i$ contributes right censoring at L_i, whereas a fact known to survive beyond M contributes right censoring at M. This convention prevents information beyond the declared forecast horizon from entering an M-month model and does not assign an artificial event time at the boundary. The hazard specification uses three cubic B-spline bases. The default dimensions are $K_a = 6,\ K_c = 6,\ K_s = 5$, for total spell age $A_i + h$, calendar time $\tau_i + h$, and post-anchor forecast horizon h, respectively. Spline knots are determined from the training design grid and then frozen. The columns of B_a, B_c, and B_s are centered using $\mathcal{D}^{\mathrm{tr}}$ and contain no intercept term, so the global intercept α_0 remains separately interpretable. The same basis functions, knots, centering constants, and calendar-month encoding are used on every later chronological block.

The relation-age penalty $\lambda_{\mathrm{rel}} \sum_{r=1}^{R_{\mathrm{rel}}} \|\theta_r - \bar{\theta}\|_2^2,\ \bar{\theta} = R_{\mathrm{rel}}^{-1} \sum_{r=1}^{R_{\mathrm{rel}}} \theta_r$, shrinks relation-specific age profiles towards their common mean, not directly towards zero. This preserves a shared duration pattern when a relation has limited data while still permitting relation-specific departures. Calendar coefficients are regularized through $\lambda_{\mathrm{drift}} \sum_{r=1}^{R_{\mathrm{rel}}} \|D_2 \xi_r\|_2^2$, where D_2 is the second-difference operator. The penalty discourages unnecessarily irregular month-to-month curvature but does not force the calendar effect to be constant or linear. The linear coefficient β on x_i receives an ℓ_1 penalty. The vector x_i contains the pre-anchor event-history and model-confidence features defined in Section 3, including previous-spell and relation summaries, entity type, answer-confidence diagnostics, and prompt agreement. The penalty permits irrelevant or redundant components of this combined feature vector to be removed without changing the representation pathway.

The semantic matrices receive Frobenius penalties, $\lambda_U \|U\|_F^2 + \lambda_V \|V\|_F^2$. Only the induced bilinear interaction matters for prediction

$$(U^\top \tilde{z}_i)^\top (V^\top B_s(h)) = \tilde{z}_i^\top U V^\top B_s(h).$$

Accordingly, the individual factors U and V are not claimed to be uniquely identified. Their factorization is a low-rank parameterization of the horizon-varying semantic coefficient surface, and the two Frobenius penalties control otherwise arbitrary rescaling between the factors. For every candidate representation configuration, the relation centroid $\hat{\mu}_r$ and projection P are estimated from $\mathcal{D}^{\mathrm{tr}}$ only. The analyzed default takes P to be the first d_p principal directions of the relation-centered training representations. The map is not updated while the candidate is evaluated on $\mathcal{D}^{\mathrm{val}}$ and remains fixed after model selection. This convention ensures that apparent representation gains cannot arise because validation, recalibration, certificate, or test examples influenced the residualization or projection.

B.2 Optimization, Tuning, Refitting, and Recalibration

Optimization is performed at the factual-record level. A minibatch contains complete records $(A_i, X_i, L_i, R_i, C_i, \delta_i)$ and excludes expanded person-month rows. For each record define the finite computational horizon

$$H_i = \begin{cases} R_i, & \delta_i = 1, \\ C_i, & \delta_i = 0. \end{cases}$$

The implementation computes the complete vector $\{f_i(1), \ldots, f_i(H_i)\}$ in one forward pass. At forecast month h, the three temporal bases are evaluated at $B_a(A_i + h),\ B_c(\tau_i + h),\ B_s(h)$, exactly as in

the hazard specification in Section 3. In particular, $R_i = \infty$ for a right-censored record is never passed to the network or spline routines; its finite computational endpoint is C_i. We minimize the penalized average negative observed-data likelihood in Section 3. AdamW is used as the numerical optimizer, with decoupled weight decay set to zero so that all regularization is represented explicitly by the penalties appearing in the stated objective. The default learning rate is 10^{-3} for the intercept, relation, calendar, and history parameters and 3×10^{-4} for U and V. We use cosine learning-rate decay, minibatch gradient-norm clipping at 5, and at most 300 epochs.

The optimization path for a candidate is determined by $\mathcal{D}^{\mathrm{tr}}$ only. Validation loss is not used to alter gradient updates or to terminate one candidate earlier than another. Each candidate is therefore associated with a sequence of training-data checkpoints $f_{\gamma,1}, \ldots, f_{\gamma,E_{\max}}$, $E_{\max} = 300$. If the penalized training objective changes by less than 10^{-6} in relative value for 25 consecutive epochs, the remaining checkpoints are defined to equal the converged checkpoint. This stopping criterion depends only on $\mathcal{D}^{\mathrm{tr}}$. Initialization is also fixed independently of all later blocks. The global intercept is initialized from the pooled training event frequency; age, calendar, and history coefficients start at zero; and entries of U and V are initialized from a zero-mean Gaussian distribution with standard deviation 0.01. The random seed is a deterministic function of the candidate identifier and is stored with the run. Thus rerunning the same candidate reproduces the same optimization path before validation is opened. Table B.1 gives the candidate-value pool from which the prespecified tuning family is constructed.

Table B.1: **Candidate-value pool for primary tuning.** The finite evaluated candidate family is constructed from these values before $\mathcal{D}^{\mathrm{val}}$ is accessed and is then held fixed throughout validation.

Component	Candidate values or rule
LM layer ℓ	final layer and four approximately equally spaced intermediate layers
Projection dimension d_p	$\{32, 64, 128\}$
Semantic interaction rank q	$\{1, 2, 4\}$
Spline dimensions (K_a, K_c, K_s)	$(5,5,4)$, $(6,6,5)$, $(8,8,6)$
λ_{rel}	$\{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$
λ_{drift}	$\{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$
λ_U	$\{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$
λ_V	$\{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$
λ_1	$\{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$
Training checkpoint	epochs $1, \ldots, 300$ along the $\mathcal{D}^{\mathrm{tr}}$-only optimization path
Primary validation criterion	smallest per-record interval-censored NLL on $\mathcal{D}^{\mathrm{val}}$

The Cartesian product of the values in Table B.1 contains $421{,}875$ structural and regularization configurations, so the primary experiment uses a prespecified space-filling subset and avoids fitting the entire product. Specifically, before $\mathcal{D}^{\mathrm{val}}$ is accessed, every coordinate of the Cartesian pool is replaced by its ordinal rank among the admissible values for that coordinate and rescaled to $[0, 1]$. Starting from the reference configuration $\gamma_0 = (\ell_{\mathrm{mid}},\ 64,\ 2,\ 6, 6, 5,\ 10^{-3}, 10^{-3}, 10^{-3}, 10^{-3}, 10^{-3})$, where ℓ_{mid} is the middle of the five candidate LM layers, we construct a deterministic maximin design. At each step, among all configurations not yet selected, we add the configuration whose rescaled coordinate vector has the largest minimum Euclidean distance to the configurations already selected. Ties are resolved lexicographically by the stored configuration identifier. The procedure stops after 256 distinct configurations have been selected. The resulting 256 configuration identifiers are fixed and stored before any validation loss is computed. The resulting evaluated family is therefore fixed in full before $\mathcal{D}^{\mathrm{val}}$ is accessed. No configuration is added, removed, replaced, or refined in response to validation performance. Each of the 256 configurations is fitted using $\mathcal{D}^{\mathrm{tr}}$ only and yields the training-data checkpoint sequence described above. Every associated checkpoint is then scored on $\mathcal{D}^{\mathrm{val}}$ using the same per-record observed-data NLL. With at most 300 checkpoints per configuration, the validation comparison therefore contains at most $256\times300 = 76{,}800$ prespecified configuration-epoch candidates. Selection is joint over configuration and epoch. Neither IBS nor calibration error enters this step; those quantities are reserved for evaluation after the hazard model has been selected.

If two candidate checkpoints have validation NLLs within 10^{-4} per factual record, ties are broken lexicographically by

$$(q,\ d_p,\ K_a + K_c + K_s,\ -\lambda_{\mathrm{rel}},\ -\lambda_{\mathrm{drift}},\ -\lambda_U,\ -\lambda_V,\ -\lambda_1,\ \mathrm{id})\,,$$

where id is the fixed unique identifier of the candidate checkpoint, including its training epoch. Thus the rule first favors lower semantic and spline complexity, then stronger regularization in the stated

fixed order, and finally resolves any remaining tie without consulting an additional performance criterion or any later chronological block. After selecting $(\hat{\gamma}, \hat{e})$, we refit the hazard parameters on $\mathcal{D}^{\mathrm{tr}} \cup \mathcal{D}^{\mathrm{val}}$ for exactly $\hat{e}$ epochs. The refit starts from a fresh initialization generated by the same deterministic initialization rule. The relation centroids, PCA projection, spline knots, centering constants, and other preprocessing maps learned from $\mathcal{D}^{\mathrm{tr}}$ remain fixed. In particular, validation representations are not used to redefine the relation centroids or principal directions during the refit. No observation from $\mathcal{D}^{\mathrm{recal}}$, $\mathcal{D}^{\mathrm{cert}}$, or $\mathcal{D}^{\mathrm{test}}$ is used for hazard-model early stopping or parameter selection.

The fitted hazard is then frozen before recalibration begins. Using only $\mathcal{D}^{\mathrm{recal}}$, we fit

$$\lambda_i^{\mathrm{cal}}(h) = \mathrm{sigm}\Big[a\ \mathrm{logit}\{\hat{\lambda}_i(h)\} + b(h)\Big], \quad a \geq 0.$$

The scalar a is optimized subject directly to the constraint $a \geq 0$ using a projected update after each optimizer step. This preserves exactly the parameter space in Section 3, including the boundary case $a = 0$. The function $b(h)$ is a five-dimensional cubic B-spline on the fixed forecast grid $h = 1, \ldots, M$. Its basis is determined before recalibration and is common to all facts. The only smoothness penalty is $\lambda_b \|D_2 b\|_2^2,\ \lambda_b \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. The value of λ_b is selected by five-fold subject-cluster cross-validation entirely within $\mathcal{D}^{\mathrm{recal}}$. All records from one subject are assigned to the same fold. For each candidate λ_b, the calibrator is fitted on four folds and evaluated on the fifth with the same interval-censored observed-data NLL used for the hazard model. After selection, (a, b) is refitted on all of $\mathcal{D}^{\mathrm{recal}}$. This inner recalibration search does not revisit $\hat{\gamma}$, $\hat{e}$, the representation layer, or any hazard parameter. The resulting calibrated survival curve is frozen before $\mathcal{D}^{\mathrm{cert}}$ is accessed. Algorithm B.1 records the numerical operations for one training minibatch.

Algorithm B.1 Stable minibatch update for the EXPIRE hazard

Require: A minibatch of records $(A_i, X_i, L_i, R_i, C_i, \delta_i)$ and fixed training-derived preprocessing.

for each record i in the minibatch **do**
  Set $H_i = R_i$ if $\delta_i = 1$ and $H_i = C_i$ otherwise.
  Compute $f_i(h)$ for $h = 1, \ldots, H_i$ using $B_a(A_i + h)$, $B_c(\tau_i + h)$, and $B_s(h)$.
  Set $\log S_i^{\mathrm{res}}(0) = 0$.
  Form $\log S_i^{\mathrm{res}}(u) = -\sum_{h=1}^{u} \mathrm{softplus}\{f_i(h)\},\ u = 1, \ldots, H_i$.
  if $\delta_i = 1$ **then**
    Set $a_i = \log S_i^{\mathrm{res}}(L_i)$ and $b_i = \log S_i^{\mathrm{res}}(R_i)$.
    Add $-a_i - \mathrm{log1mexp}(b_i - a_i)$ to the negative log likelihood.
  else
    Add $-\log S_i^{\mathrm{res}}(C_i)$ to the negative log likelihood.
  end if
end for
Average the likelihood contributions over the minibatch.
Add the relation-shrinkage, calendar-roughness, semantic-Frobenius, and linear-feature sparsity penalties.
Backpropagate and clip the gradient norm at 5.
Apply the AdamW update with decoupled weight decay disabled.
Apply the proximal soft-thresholding step to β for the ℓ_1 penalty.

The proximal update is applied after the smooth optimizer step. If β_j^+ denotes the coordinate after the AdamW update and η_t the current learning rate, then

$$\beta_j \leftarrow \mathrm{sign}(\beta_j^+)\max\{|\beta_j^+| - \eta_t \lambda_1, 0\}.$$

Thus the ℓ_1 term is implemented by an explicit proximal-style soft-thresholding step and not by ordinary weight decay. Any numerical discrepancy from exact minimization is absorbed by the optimization-error term $\varepsilon_{\mathrm{opt}}$ in Theorem 4.2.

With hidden states cached, evaluating one factual record through horizon M requires $O(p + qd_p + M\{K_a + K_c + qK_s\})$ arithmetic operations. The relation-specific age and calendar coefficients are direct table lookups, so their relation index does not introduce a factor of R_{rel} in per-record inference. The semantic computation first forms $U^\top \tilde{z}_i$ and then evaluates its interaction with $V^\top B_s(h)$ across the forecast grid. Memory for the survival model is small relative to the cached LM representations. No additional LM decoding is required to produce an expiration curve once the representation of the factual query has been obtained. In an online setting, the hidden state can be collected during the

frozen prompt forward passes used for factual querying. Subsequent lifetime prediction therefore requires only relation/history feature construction and the comparatively small EXPIRE hazard model. Retrieval itself is invoked only after the separate deployment rule in Section 3 has been applied. Each reported run records the source-snapshot identifier, point-in-time model identifier and hash, prompt-template hash, representation layer, relation centroids, PCA map, spline knots, candidate configuration, selected epoch, random seed, optimizer settings, recalibration configuration, and serialized predictions. Predictions for a chronological block are written before its outcomes are made available to any later stage of the pipeline. The implementation also performs three numerical checks before an experiment is accepted. First, on moderately scaled synthetic examples, the stable interval log likelihood is compared with direct subtraction and must agree to floating-point tolerance whenever the latter is well conditioned. Second, automatic-differentiation gradients of interval-censored and right-censored losses are compared with finite-difference derivatives on randomly generated small batches. Third, survival curves are checked mechanically for

$$1 = \hat{S}_i^{\mathrm{res}}(0) \geq \hat{S}_i^{\mathrm{res}}(1) \geq \cdots \geq \hat{S}_i^{\mathrm{res}}(M) \geq 0,$$

both before and after recalibration. These checks detect numerical errors without using validation, certificate, or test outcomes.

C WIKILIFETIME CONSTRUCTION

This part gives the complete construction protocol for WIKILIFETIME used in the empirical study. The benchmark is prospective by construction. For a row anchored at calendar time τ_i, the factual state, language-model output, historical summaries, metadata, and frozen representation must all be recoverable using information available no later than τ_i. Later revisions are used only to determine the eventual post-anchor outcome $(L_i, R_i, C_i, \delta_i)$ and are never allowed to alter the anchor-side predictors X_i, the known-fact decision, the prompt family, or the frozen representation. The primary build is `WikiLifetime-v1.0-2025-12-31`. It uses dated Wikidata revision histories frozen on December 31, 2025, together with point-in-time language models whose documented training-data cutoffs precede their assigned historical anchors. Historical origins occur on January 1 and July 1. The primary forecast horizon is $M = 24$ months, so follow-up through December 31, 2025 supports the complete 24-month evaluation of both 2023 test origins.

Every retained benchmark row satisfies two conditions at its anchor; the fact is currently valid in the external source history, and the point-in-time language model recovers its current object under the fixed two-of-three prompt rule in Section 2. Consequently, WIKILIFETIME measures the remaining lifetime of knowledge already possessed by the frozen model and avoids mixing knowledge acquisition with subsequent expiration.

C.1 HISTORICAL CONSTRUCTION AND ANCHOR-SIDE INFORMATION

The source manifest for the primary build contains the revision-history release `wd-history-2025-12-31-v1`, the entity and alias release `wd-entities-2025-12-31-v1`, and the benchmark freeze timestamp `2025-12-31T23:59:59Z`. Revision timestamps are retained at their native resolution during factual-spell reconstruction and are then mapped to the benchmark's common monthly time scale. No revision dated after the historical anchor contributes to the row's anchor-side information. The source state is reconstructed independently at every historical origin. This is important because reading the final 2025 database and then retrospectively assigning its labels, aliases, or completed spell durations to earlier years would expose information that was unavailable at the time of prediction. For each subject-relation pair (s, r), the dated history is converted into a sequence of maximal nonoverlapping factual spells $(o_1, [B_1, E_1)), \ldots, (o_J, [B_J, E_J))$, where o_j is the externally valid current object during $[B_j, E_j)$. The spell begins when o_j becomes the current object and ends when the present-tense relation no longer supports o_j. Expiration therefore means the end of present-tense validity. The corresponding historical statement remains true after E_j.

Explicit start and end qualifiers are used whenever they determine a validity boundary without contradiction. When an exact boundary is not available, let t_i^- be the latest admissible source observation still supporting the old object and let t_i^+ be the earliest admissible observation supporting the replacement. The source history then identifies $t_i^- < E_i \leq t_i^+$. For a row anchored at τ_i, this becomes the post-anchor interval $(L_i, R_i]$ used in Sections 2-3. The event time is never replaced by $(L_i + R_i)/2$ or

by either endpoint. A newly observed object is considered a genuine replacement only when it persists for at least 30 days or is supported by an explicit temporal qualifier. The persistence requirement prevents transient edits and rapid reverts from being interpreted as factual expiration. The robustness analysis repeats the construction with persistence thresholds of 7 and 90 days. Label-only changes, changes of capitalization or punctuation, identifier redirects, duplicated statements, entity merges preserving the same referent, and edits identified as vandalism followed by reversion are not expiration events. The primary benchmark retains event intervals having width at most 180 days. Wider intervals are not midpoint-imputed; they are moved to the coarse-observation sensitivity set. This cutoff is fixed before model evaluation and is applied identically to every method.

If no supported replacement is observed before the end of available follow-up, the row is right censored. On the 24-month analysis horizon,

$$C_i = \min\{24, F_i - \tau_i\},$$

where F_i is the last admissible follow-up date expressed in months after the anchor. If the spell is still valid at $\tau_i + C_i$, the observation states only $\tilde{T}_i > C_i$. An expiration interval extending beyond month 24 is handled as described in Appendix B.1; no artificial event is placed at the 24-month boundary.

Relation eligibility is determined without using validation, recalibration, certificate, or test outcomes. The primary build contains 20 operational single-current-object relations grouped into six families. A relation enters the primary inventory only when all of the following conditions are satisfied using information available through December 31, 2020:

1. the relation has a well-defined current-object interpretation;
2. historical values can be arranged into nonoverlapping dated spells;
3. at least 500 eligible training-origin factual rows are available;
4. at least 250 completed historical transitions are available before the end of the training period;
5. at least 100 distinct training subjects are represented; and
6. canonical entity identifiers and anchor-time aliases permit unambiguous short-answer matching.

The numerical thresholds are minimum-support filters and not outcome-dependent model-selection rules. Once the 20-relation inventory is formed, no relation is added or removed because of subsequent predictive performance. The relation-family inventory is summarized in Table C.1. Exact source relation identifiers and the single-valued schema filters are stored in the released benchmark manifest.

Table C.1: **Primary relation-family inventory.** The inventory is fixed before validation and test evaluation. "Relations" denotes operational single-current-object relations after schema filtering.

Relation family	Relations	Training completed spells	Test records
Corporate leadership	3	9,684	5,842
Political and public office	4	8,917	4,916
Sports-team affiliation	3	13,406	7,388
Coach / manager	3	6,238	3,276
Employment / organizational membership	4	10,751	6,104
Low-turnover current-object controls	3	7,963	8,452
Total	20	56,959	35,978

Relations admitting several simultaneously valid current objects are excluded from the primary task. The restriction avoids an ambiguity that is orthogonal to the survival problem; for a genuinely set-valued relation, we must first specify whether the event is removal of one object, replacement of the full set, or any change to the set. The corresponding extension is discussed in Appendix F.

At historical origin τ, a candidate fact must first satisfy $\tau \in [B_i, E_i)$, so the object is externally valid before the language model is queried. The model-side criterion then follows Section 2 exactly. The primary benchmark uses $J = 3$ fixed prompts and $k_0 = 2$. The prompt family is frozen before benchmark evaluation:

1. `Question: What is the current [RELATION] of [SUBJECT]? Answer with the entity name only.`

2. `[SUBJECT] --- current [RELATION]:`
3. `Which entity is the current [RELATION] for [SUBJECT]? Answer with one entity only.`

Relation-specific grammatical substitutions are deterministic and are specified in the prompt manifest; they do not introduce dates, future replacements, or outcome information. Generation is deterministic. The primary configuration uses greedy decoding, temperature 0, no sampling, a maximum of 24 new tokens, and termination at the first newline, end-of-sequence token, or completed short entity span. The same configuration is used at every historical origin.

Let $\mathcal{A}_{\tau_i}(o_i)$ denote the alias set for object o_i that is admissible at the anchor. A response is normalized by Unicode NFC normalization, whitespace collapse, case folding when appropriate for the entity type, removal of non-semantic terminal punctuation and leading articles, and exact resolution against $\mathcal{A}_{\tau_i}(o_i)$. An alias first introduced after τ_i cannot make an earlier response correct. The retained population is therefore

$$\mathcal{P}_\tau = \left\{ i : \tau \in [B_i, E_i),\ \sum_{j=1}^{3} \mathbf{1}\left\{ \mathrm{Ans}(M_\tau, q_i^{(j)}) \in \mathcal{A}_\tau(o_i) \right\} \geq 2 \right\}.$$

A fact that is externally valid but not recovered by the frozen model is excluded from $\mathcal{P}_\tau$ and is not labeled as expired model knowledge. The primary cohort is always defined by generated answers. Token probabilities do not replace the two-of-three rule for base or instruction-tuned checkpoints. Where token-level probabilities are available, they are used only as anchor-side confidence variables or in a separately labeled sensitivity analysis. All primary checkpoints belong to the same 7B-parameter point-in-time model family and use the same tokenizer and architecture. Each January origin is paired with the checkpoint ending on December 31 of the preceding year; each July origin is paired with the checkpoint ending on June 30 of the same year. Thus the documented cutoff always precedes the anchor. Table C.2 gives the complete assignment.

Table C.2: **Point-in-time checkpoint assignment in the primary build.** Every documented data cutoff is no later than its historical anchor.

Anchor	Checkpoint ID	Parameters	Documented data cutoff
January 2018	`pitlm-7b-2017-12`	7B	Dec. 31, 2017
July 2018	`pitlm-7b-2018-06`	7B	Jun. 30, 2018
January 2019	`pitlm-7b-2018-12`	7B	Dec. 31, 2018
July 2019	`pitlm-7b-2019-06`	7B	Jun. 30, 2019
January 2020	`pitlm-7b-2019-12`	7B	Dec. 31, 2019
July 2020	`pitlm-7b-2020-06`	7B	Jun. 30, 2020
January 2021	`pitlm-7b-2020-12`	7B	Dec. 31, 2020
July 2021	`pitlm-7b-2021-06`	7B	Jun. 30, 2021
January 2022	`pitlm-7b-2021-12`	7B	Dec. 31, 2021
July 2022	`pitlm-7b-2022-06`	7B	Jun. 30, 2022
January 2023	`pitlm-7b-2022-12`	7B	Dec. 31, 2022
July 2023	`pitlm-7b-2023-06`	7B	Jun. 30, 2023

A checkpoint with a later cutoff is never an admissible primary predictor. The future-contaminated checkpoint examined in Appendix E.2 violates this rule and is used only to measure the magnitude of look-ahead bias. The OLMo analysis in the empirical sensitivity study follows the same admissibility rule; an OLMo checkpoint is used only at origins for which its documented cutoff is no later than the anchor. Those experiments are performed on the intersection cohort shared by the compared checkpoints and do not redefine the primary WIKILIFETIME population.

Every component of X_i is reconstructed as if the source database had been frozen at τ_i. A spell that is known from the final 2025 history to end after τ_i remains active and censored when features are constructed at τ_i. For example, suppose a spell is active at a July 2019 anchor and later ends in February 2021. Its 2021 completion may determine a prospective outcome, but the February 2021 duration cannot appear in any July 2019 history variable. At that anchor the feature builder knows only that the spell has survived through its then-current age. Table C.3 gives the complete feature groups supplied to the primary model. For the relation-level survival summaries, only outcomes available by the anchor are supplied to the interval-censored estimator. Completed historical spells contribute their

then-observed intervals, while spells still active at τ_i are censored at their anchor-time ages. No final-database completion time is inserted retrospectively. Continuous history variables are transformed and standardized using statistics learned from $\mathcal{D}^{\mathrm{tr}}$ only. Missing history is represented explicitly. For example, a fact with no completed previous spell receives the prespecified no-history indicators instead of a future duration or a value imputed from later data.

Table C.3: **Anchor-side feature groups.** Every quantity is reconstructible using information available no later than the historical anchor.

Feature group	Construction at τ_i
Spell age	$A_i = \tau_i - B_i$; anchor month; current spell duration through the anchor.
Subject-relation history	Number of completed previous transitions; lower and upper endpoints of the most recent completed duration interval; width of that interval; time since the latest completed transition; numbers of transitions in the previous 12, 24, and 60 months.
Relation history	Anchor-time interval-censored survival summaries for relation r_i; estimated survival at 6, 12, and 24 months; censoring-aware median and interquartile lifetime when estimable; transition counts during the previous 12 and 24 months.
Static / historical metadata	Entity type and relation-schema indicators from the latest admissible source state at or before τ_i.
Model-confidence diagnostics	Number of successful prompts in $\{2, 3\}$; agreement pattern across the three prompts; mean normalized log probability of the generated accepted alias when available; margin between the current object and relation-matched alternatives.
Frozen LM representation	Final-input-token hidden states from the same three fixed prompts, averaged and then ℓ_2-normalized. Relation centering and PCA are learned from $\mathcal{D}^{\mathrm{tr}}$ only.

The representation extraction rule is exactly the one used in Section 3. For prompt j and candidate layer ℓ, let $h_{ij}^{(\ell)}$ be the hidden state at the final input token before answer generation. All three fixed prompts are used. We first form

$$\bar{h}_i^{(\ell)} = \frac{1}{3} \sum_{j=1}^{3} h_{ij}^{(\ell)}$$

and then normalize

$$z_i^{(\ell)} = \frac{\bar{h}_i^{(\ell)}}{\|\bar{h}_i^{(\ell)}\|_2}.$$

The representation is not averaged only over prompts that answered correctly. Cohort membership and representation extraction are separate: the two-of-three rule determines whether the fact enters $\mathcal{P}_\tau$, whereas all three prompt states determine z_i. This prevents prompt-level success from changing the representation definition from fact to fact. For every candidate layer, relation centroids and PCA directions are estimated from $\mathcal{D}^{\mathrm{tr}}$ only. The centering and projection maps are then frozen when that candidate is evaluated on $\mathcal{D}^{\mathrm{val}}$. Neither $\mathcal{D}^{\mathrm{recal}}$, $\mathcal{D}^{\mathrm{cert}}$, nor $\mathcal{D}^{\mathrm{test}}$ contributes to the representation map. Layer selection itself occurs only through the validation procedure in Appendix B.2.

C.2 Rolling-Origin Blocks and Composition

The primary empirical experiment uses 12 semiannual historical origins from January 2018 through July 2023. Their statistical roles are fixed before any final-test outcome is inspected, $\mathcal{D}^{\mathrm{tr}} \longrightarrow \mathcal{D}^{\mathrm{val}} \longrightarrow \mathcal{D}^{\mathrm{recal}} \longrightarrow \mathcal{D}^{\mathrm{cert}} \longrightarrow \mathcal{D}^{\mathrm{test}}$. Table C.4 gives the exact chronology. The outcome-reveal date is not part of the predictor information set. For example, a July 2023 test row is represented using only information available on July 1, 2023 even though source revisions through June 30, 2025 may subsequently determine whether that fact expired within 24 months.

Repeated anchors create a potential leakage channel because one factual spell may remain active across several origins and one subject may contribute several relations. The benchmark therefore retains both subject and spell identifiers throughout construction. After chronological eligibility is determined, subject clusters are made disjoint across statistical blocks. The implementation applies the priority $\mathcal{D}^{\mathrm{test}} \succ \mathcal{D}^{\mathrm{cert}} \succ \mathcal{D}^{\mathrm{recal}} \succ \mathcal{D}^{\mathrm{val}} \succ \mathcal{D}^{\mathrm{tr}}$. If one subject has eligible rows in more than one block,

the subject is assigned to the latest block in this priority ordering and all of its rows in earlier blocks are removed. All spells belonging to that subject inherit the same block assignment. This operation uses only subject identifiers and anchor dates, not outcomes or model scores. It ensures that no subject or factual spell appears in more than one of the fitting, validation, recalibration, certification, or final test samples. Within a block, a subject may contribute more than one eligible fact, and those records are kept together as a cluster for bootstrap and finite-sample accounting. The deterministic assignment removes 7.6% of otherwise eligible rows in the primary build. The same fixed subject map is reused for every method, so comparisons are performed on identical samples.

Table C.4: **Chronological roles in the primary WIKILIFETIME experiment.** A later block is never used for an earlier statistical task.

Block	Historical origins	Permitted use
$\mathcal{D}^{\mathrm{tr}}$	Jan./Jul. 2018-2020	Fit candidate hazards; construct relation centroids and PCA maps; determine spline knots, feature transformations, and training normalizations.
$\mathcal{D}^{\mathrm{val}}$	Jan./Jul. 2021	Select LM layer, d_p, q, spline dimensions, penalties, and training checkpoint using observed-data NLL only.
$\mathcal{D}^{\mathrm{recal}}$	Jan. 2022	Fit the calibration slope and horizon-dependent calibration offset only.
$\mathcal{D}^{\mathrm{cert}}$	Jul. 2022	Select the retrieval-risk threshold for each prespecified deployment horizon and target risk; no lifetime-model parameter is changed.
$\mathcal{D}^{\mathrm{test}}$	Jan./Jul. 2023	Final prospective evaluation only; no fitting, tuning, recalibration, or threshold adaptation.
Outcome follow-up	Through Dec. 31, 2025	Reveal expiration intervals or right-censoring information for supported horizons through 24 months.

After all validity, known-fact, relation-support, interval-width, and subject-disjointness filters, the primary build contains 237,937 factual rows from 141,911 distinct subject clusters. Table C.5 reports the complete block composition. The 12-month and 24-month columns give the observed proportion of rows whose current fact expires within the corresponding horizon. "24m right censored" includes facts remaining valid beyond the 24-month analysis window. The mild decline in observed turnover towards the later chronological blocks is one reason random train-test splitting would be inappropriate; it would partially erase the calendar shift that the proposed evaluation is intended to preserve.

Table C.5: **WIKILIFETIME composition after all primary eligibility and chronology filters.** Counts refer to the records actually entering each statistical block.

Block	Facts	Subjects	12m expire	24m expire	24m right cens.
$\mathcal{D}^{\mathrm{tr}}$	124,632	68,417	17.0%	30.7%	69.3%
$\mathcal{D}^{\mathrm{val}}$	39,846	24,932	16.4%	29.9%	70.1%
$\mathcal{D}^{\mathrm{recal}}$	18,964	13,112	16.0%	29.2%	70.8%
$\mathcal{D}^{\mathrm{cert}}$	18,517	12,846	15.8%	28.9%	71.1%
$\mathcal{D}^{\mathrm{test}}$	35,978	22,604	15.6%	28.5%	71.5%
Total	237,937	141,911	16.5%	30.0%	70.0%

Table C.6: **Observed expiration-interval resolution in the primary 2023 test block.** Percentages are conditional on an expiration observed within the 24-month analysis horizon.

Observed interval width	Number of events	Percentage
1 month	4,428	43.2%
2-3 months	3,557	34.7%
4-6 months	1,753	17.1%
7-12 months	513	5.0%
Total observed within 24m	10,251	100.0%

Among observed test expirations within 24 months, 43.2% are localized to a one-month interval, 34.7% have interval width 2-3 months, 17.1% have width 4-6 months, and 5.0% have width 7-12 months. The median observed event-interval width is 2 months, the mean is 2.7 months, and the

90th percentile is 6 months. No event interval wider than 180 days enters the primary analysis. These summaries are given in Table C.6 for clarity. The event count in the table corresponds to 28.5% of the 35,978 test records up to rounding. The remaining 25,727 rows are observed to remain valid through month 24 and are treated as right censored at that horizon.

Table C.7 gives the prespecified relation-family composition of the untouched 2023 test block. This table is computed after all eligibility and subject-disjointness filters. The large difference in turnover across relation families motivates the relation-specific duration component of EXPIRE, but it also creates a potentially easy shortcut. The matched same-relation analysis in Appendix E.1 and the representation falsifications in Appendix E.2 are therefore designed specifically to test whether the semantic pathway contains information beyond relation identity and coarse recency.

Table C.7: **Primary 2023 test composition by relation family.** The test block contains 35,978 eligible factual records.

Relation family	Records	12m expiration	24m expiration
Corporate leadership	5,842	18.7%	34.5%
Political and public office	4,916	15.2%	29.7%
Sports-team affiliation	7,388	24.5%	42.1%
Coach / manager	3,276	27.8%	46.6%
Employment / organizational membership	6,104	12.6%	24.3%
Low-turnover current-object controls	8,452	3.4%	7.8%
Total	35,978	15.6%	28.5%

A horizon-specific outcome is used only when the observed coarsening set supports that horizon. Complete 24-month administrative follow-up is available for both 2023 test origins under the December 2025 freeze. For earlier blocks, follow-up is also truncated to 24 months so that all methods are fit and evaluated on the same maximum horizon. A record is never counted as "no expiration" merely because the source freeze occurs before the required horizon. In such a case the row would be right censored at its actual C_i, and the horizon-specific probability metric would use the coarsening-aware procedure in Appendix E.1.

D SIMULATION DESIGN AND ADDITIONAL RESULTS

This part gives the complete data-generating mechanisms, observation process, tuning rules, baseline specifications, and additional results for the controlled experiments in Section 5. The simulations are designed around the post-anchor estimand in Sections 2-3 and not around an uncensored regression surrogate. Each generated fact has already survived to its anchor age, its expiration occurs prospectively after the anchor, and the fitted methods receive only an interval-censored or right-censored record. The latent event time is retained by the simulator for diagnostics but is never provided to a fitted method. The four designs separate three sources of predictive information. S0 contains relation, duration, and history information but removes the residual semantic signal from the pseudo-LM state. S1 and S2 introduce increasing within-relation semantic information while leaving the history mechanism unchanged. S3 retains the S2 semantic signal and adds relation-specific calendar drift. Consequently, a representation gain in S0 is spurious by construction, a gain in S1-S2 measures recovery of genuine within-relation heterogeneity, and S3 additionally tests whether the method can accommodate changing turnover across historical origins.

D.1 DATA-GENERATING MECHANISM AND OBSERVATION PROCESS

Each design is repeated independently 500 times. In every replication, $n_{\mathrm{tr}} = 5000$, $n_{\mathrm{val}} = 1000$, $n_{\mathrm{test}} = 3000$. The simulation intentionally omits separate recalibration and certification blocks because the controlled study evaluates lifetime estimation itself; recalibration and retrieval-risk certification are examined in the empirical application. Calendar time is measured in months from a notional January 2018 origin. Training anchors are sampled uniformly from $\{0, 6, 12, 18, 24, 30\}$, validation anchors from $\{36, 42\}$, and test anchors from $\{60, 66\}$. Thus the three samples mimic January/July origins through 2020, validation origins in 2021, and test origins in 2023, with the intermediate 2022 period left unused in the controlled experiment. This chronological gap prevents S3 from reducing to an ordinary random-split interpolation problem.

Within each block, relation identity is sampled independently as $r_i \sim \mathrm{Uniform}\{1, \ldots, 12\}$. The 12 relation baseline logits are the equally spaced values on $[-4.5, -2.7]$ and are randomly permuted across relation labels once at the start of each replication. Anchor age is generated independently as $A_i \sim \mathrm{Uniform}\{1, \ldots, 48\}$. The implied spell start is $B_i = \tau_i - A_i$. Because the target distribution is conditional on survival to the anchor, the simulation generates the remaining lifetime directly. No artificial pre-anchor failure time is generated and discarded. Replication b uses master seed $2027000 + b$, $b = 1, \ldots, 500$. All train/validation/test data and all method-specific initialization seeds are deterministic descendants of that master seed.

For each fact, generate mutually independent

$$U_i, V_i, \varepsilon_{i1}, \ldots, \varepsilon_{i8}, \varepsilon_{iD}, \varepsilon_{iQ} \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, 1).$$

The scalar U_i determines the residual semantic signal. The scalar V_i represents latent subject-history heterogeneity that is only partially observed through noisy history variables. Three observed history variables are generated. The previous-transition count is

$$N_i \sim \mathrm{Poisson}\left[\exp\left\{-0.25 + 0.03(r_i - 1) + 0.35\max(V_i, 0)\right\}\right].$$

The most recent completed spell-duration proxy is

$$D_i = \exp\{2.3 - 0.18V_i + 0.35\varepsilon_{iD}\},$$

and an auxiliary entity-history score is

$$Q_i = 0.25V_i + 0.8\varepsilon_{iQ}.$$

The standardized history summary entering the data-generating hazard is

$$H_i = \frac{0.45\log(1 + N_i) - 0.35\log D_i + 0.20Q_i + 0.5314}{0.3382}.$$

The centering and scaling constants are fixed population-level design constants obtained before the 500 Monte Carlo replications. They are not re-estimated from test data. The fitted metadata/history methods receive the observed variables (N_i, D_i, Q_i, H_i) but never receive V_i. The direct $0.55V_i$ term in the event hazard therefore leaves realistic residual outcome variation even when the observed history variables are modeled correctly.

The eight-dimensional pseudo-LM state is

$$z_{i1} = 0.90U_i + 0.45\varepsilon_{i1}, \quad z_{i2} = 0.55U_i + 0.80\varepsilon_{i2}, \quad z_{ij} = \varepsilon_{ij},\ j = 3, \ldots, 8.$$

Thus only the first two coordinates contain information about the latent semantic factor, with the remaining six coordinates acting as representation noise. Exactly as in the empirical method, the pseudo-LM state is centered within relation using the training block only. Candidate PCA projections are also constructed from the training block only. The projection dimension is selected from $d_p \in \{2, 4, 8\}$. This construction is useful for S0; the representation has the same marginal distribution as in the semantic designs, but U_i has coefficient zero in the event hazard. A method cannot therefore succeed in S0 simply because the representation distribution has been made easier.

For forecast month $h = 1, \ldots, 24$, the true conditional expiration hazard is

$$\lambda_i^\star(h) = \mathrm{sigm}\{\eta_i(h)\},$$

with

$$\eta_i(h) = \alpha_{r_i} + \beta_U U_i + 0.55V_i + 0.42H_i + 0.02(A_i + h) + 0.12\sin(h/4) + D_{r_i}(\tau_i + h).$$

The total-age term $A_i + h$ makes the remaining lifetime depend on how long the current spell has already survived. This is the attained-age information induced by left-truncated entry. The four designs are shown as follows.

Design	(β_U, D_r)	Purpose
S0	$(0, 0)$	no residual semantic signal
S1	$(0.45, 0)$	moderate residual semantic signal
S2	$(0.75, 0)$	strong residual semantic signal
S3	$(0.75, D_r^{\mathrm{drift}})$	strong semantic signal plus calendar drift

The S3 drift function is

$$D_r^{\text{drift}}(t) = \begin{cases} 0.40(t-36)/36, & r = 0 \pmod 3, \\ -0.25(t-36)/36, & r = 1 \pmod 3, \\ 0, & r = 2 \pmod 3. \end{cases}$$

Thus one third of the relations become progressively more volatile, one third become progressively less volatile, and one third have no calendar shift. The drift is centered around the validation period and changes only the relation-specific calendar component; the distributions of U_i, V_i, and the pseudo-LM state are unchanged. Conditional on being event-free through month $h-1$, an independent uniform draw determines whether expiration occurs during month h with probability $\lambda_i^\star(h)$. The latent remaining lifetime $\tilde{T}_i$ is the first month in which an event occurs. If no event occurs through month 24, we record only $\tilde{T}_i > 24$. The corresponding calendar endpoint is $E_i = \tau_i + \tilde{T}_i$ when the event occurs within the modeled horizon.

Administrative censoring is independent of the event process and all covariates. We draw

$$Z_i^C \sim \text{Bernoulli}(0.25).$$

If $Z_i^C = 1$, draw

$$C_i \sim \text{DiscreteUniform}\{8, \ldots, 24\};$$

otherwise set $C_i = 24$. Hence approximately 23.5% of all records receive an administrative endpoint strictly before month 24, while the remainder are followed through the full forecast horizon. The observed-expiration indicator is $\delta_i = \mathbf{1}\{\tilde{T}_i \le C_i\}$. When $\delta_i = 0$, the fitted method sees only $\tilde{T}_i > C_i$ and, consistently with Section 2, we set $R_i = \infty$. This censoring mechanism was fixed at the design stage because it produces approximately 35% realized right censoring after accounting jointly for events that naturally survive beyond month 24 and earlier administrative loss of follow-up. No design-specific censoring parameter is tuned after outcomes are generated. The realized fractions remain close to the intended 35% across all four designs (Table D.1), so differences in predictive difficulty across S0-S3 are not driven by materially different censoring rates.

Table D.1: **Realized censoring in the primary simulation design.** Values are Monte Carlo summaries. "Early administrative endpoint" denotes $C_i < 24$; realized right censoring additionally includes facts that remain event-free through their available follow-up.

Design	Early administrative endpoint	Realized right censoring
S0	23.5%	34.4%
S1	23.5%	34.9%
S2	23.5%	35.6%
S3	23.5%	35.4%

When an expiration occurs before administrative censoring, its exact latent month is not necessarily revealed. In the primary observation mechanism, draw an inspection spacing G_i, independently of the event time and covariates, according to

$$\mathbb{P}(G_i = 1) = 0.45, \quad \mathbb{P}(G_i = 3) = 0.40, \quad \mathbb{P}(G_i = 6) = 0.15.$$

For an observed event at latent month $\tilde{T}_i$, define

$$L_i = G_i \left\lfloor \frac{\tilde{T}_i - 1}{G_i} \right\rfloor \text{ and } R_i = \min\left\{ G_i \left\lceil \frac{\tilde{T}_i}{G_i} \right\rceil, C_i \right\}.$$

Because $\tilde{T}_i \le C_i$ whenever $\delta_i = 1$, this construction always satisfies $L_i < \tilde{T}_i \le R_i$. The resulting observed record is exactly $W_i = (A_i, X_i, L_i, R_i, C_i, \delta_i)$ from Sections 2-3. The mixed primary scheme avoids making all intervals artificially identical while preserving a transparent coarsening process. The interval-width sensitivity replaces this mixture by fixed spacings $G_i \equiv 1,\ 3,\ 6,$ or 12. All methods receive the same coarsened record within a replication.

D.2 Estimation, Evaluation, and Additional Results

The four focal methods in Table 1 are fitted to exactly the same training and validation observations. *Relation only* receives relation identity, calendar origin, and forecast horizon. It contains no subject-

level history, pseudo-LM state, or anchor-age variable. Its purpose is to measure what can be predicted from relation-specific turnover alone. *Metadata/history* adds A_i, the observed history variables (N_i, D_i, Q_i, H_i), and the age/calendar spline structure, but sets the semantic interaction to zero. *Representation only* uses relation, calendar, anchor age, and the relation-centered pseudo-LM representation, but excludes the observed history variables. EXPIRE uses all of the relation, age, calendar, observed-history, and representation components. For the three methods using the EXPIRE parameterization, the simulation grid is

$$d_p \in \{2, 4, 8\}, \quad q \in \{1, 2, 4\}, \quad (K_a, K_c, K_s) \in \{(4, 4, 4), (6, 6, 5), (8, 8, 6)\},$$

and

$$\lambda_{\text{rel}}, \lambda_{\text{drift}}, \lambda_U, \lambda_V, \lambda_1 \in \{0, 10^{-3}, 10^{-2}, 10^{-1}\},$$

with irrelevant penalty coordinates omitted for the corresponding ablation. The maximum number of optimization epochs is 200 in the simulation. All candidate preprocessing maps are constructed from the replication's training block only. Candidate configurations and training checkpoints are selected by the smallest observed-data interval-censored NLL on the validation block. IBS, ECE, pair accuracy, and test outcomes are not used for model selection. The selected configuration is refitted on $\mathcal{D}^{\text{tr}} \cup \mathcal{D}^{\text{val}}$ for the selected number of epochs before test evaluation.

Three flexible survival baselines are included in the appendix comparison. The interval-censored proportional-hazards model uses relation strata and linear terms for age, calendar position, the observed history variables, and the projected pseudo-LM coordinates. Its ridge parameter is selected from $\{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. Its likelihood is computed from the observed event intervals and censoring times; no event midpoint is created. The interval-censored survival forest uses 500 trees. The candidate minimum terminal-node sizes are $\{10, 25, 50\}$, and the candidate numbers of variables considered at a split are $\{\lceil\sqrt{p}\rceil, \lceil p/3\rceil, \lceil p/2\rceil\}$, where p denotes the number of available scalar predictors after representation projection. Candidate forest survival distributions are scored on validation data using the same observed interval likelihood. The discrete-time neural survival baseline receives relation indicators, anchor age, calendar position, all observed history variables, and the projected pseudo-LM state. It predicts 24 monthly hazards using two hidden layers. Hidden width is selected from $\{32, 64\}$, dropout from $\{0, 0.1, 0.2\}$, and ℓ_2 weight decay from $\{0, 10^{-4}, 10^{-3}\}$. The learning rate is 10^{-3}, the maximum epoch count is 200, and the network is trained directly with the same interval/right-censored observed-data likelihood. Its configuration and training checkpoint are selected by validation NLL. Thus all appendix baselines receive the same information and the same coarsened outcomes as EXPIRE. Differences in performance cannot be attributed to one method receiving an exact latent event time.

Primary predictive accuracy is evaluated at $\mathcal{U} = \{3, 6, 12, 18, 24\}$. The integrated Brier score is the arithmetic mean of the five horizon-specific Brier scores. The primary reported IBS uses the same coarsening-aware evaluation construction as the empirical application, applied only to the observed test records. Because the simulator retains complete latent event times, we additionally compute the complete-data oracle IBS as a diagnostic; the two agree within Monte Carlo error under the correct observation model. Latent event times are not used to tune any method. Twelve-month calibration error is computed using ten equal-frequency bins of predicted expiration probability. Bin boundaries are determined separately within each test replication from predicted values only. The observed expiration frequency within a bin is estimated using the same coarsening-aware horizon procedure as in the empirical study. For S2, the within-relation ranking diagnostic matches an event fact that expires within 12 months to a control from the same relation and the same calendar origin that survives beyond month 24. Each control is used at most once. If several controls are available, the one with the closest anchor age is chosen, with deterministic random-seed tie breaking. Pair accuracy assigns score one when the event fact receives a larger predicted 12-month expiration probability, zero when the ordering is reversed, and one half to an exact tie. Because relation-only prediction is identical for two facts having the same relation and anchor, its population value on this diagnostic is exactly one half. The metric therefore directly tests instance-level discrimination within relation and not merely recovery of relation-wide turnover. For the auxiliary budget-retrieval diagnostic, test facts are ranked by predicted 12-month expiration probability and the highest-risk 25% are retrieved. Missed stale is the fraction of all test facts that expire within 12 months but fall outside that retrieved set. This experiment is a fixed-budget ranking diagnostic only; it does not use recalibration or the finite-sample certification procedure from Section 3.

For any scalar metric M, let M_b be its value in replication $b = 1, \ldots, 500$. We report

$$\bar{M} = \frac{1}{500} \sum_{b=1}^{500} M_b,$$

the empirical Monte Carlo standard deviation

$$\mathrm{SD}(M) = \left\{ \frac{1}{499} \sum_{b=1}^{500} (M_b - \bar{M})^2 \right\}^{1/2},$$

and the Monte Carlo standard error

$$\mathrm{MCSE}(M) = \frac{\mathrm{SD}(M)}{\sqrt{500}}.$$

The parentheses in the main simulation table contain MCSEs, not within-replication sampling standard errors. Several features of Table D.2 are diagnostic of the intended design. In S0, the representation-only specification is worse than the history model and full EXPIRE differs from metadata/history only modestly. The pseudo-LM state therefore does not create a substantial gain when its latent semantic factor has zero coefficient in the event mechanism. In S1 and S2, representation-only prediction improves as β_U increases, and the full method benefits further from combining semantic and history information. Under S3, the relative advantage persists after relation-specific calendar drift is introduced. The comparison against the three flexible survival baselines is also informative. Their access to the same pseudo-LM coordinates rules out an interpretation in which the gain arises simply because EXPIRE is the only method given semantic predictors. The improvement instead reflects the structured relation, age, calendar, and horizon-dependent representation parameterization.

Table D.2: **Controlled simulation results across S0-S3.** Entries are Monte Carlo means over 500 independent replications; parentheses contain MCSEs. IBS is the coarsening-aware integrated Brier score and ECE is the 12-month calibration error.

	S0		S1		S2		S3	
Method	IBS	12m ECE	IBS	12m ECE	IBS	12m ECE	IBS	12m ECE
Relation only	.214 (.0012)	.049 (.0018)	.220 (.0013)	.053 (.0019)	.233 (.0014)	.061 (.0020)	.241 (.0016)	.069 (.0022)
Metadata/history	.195 (.0010)	.031 (.0014)	.211 (.0011)	.037 (.0015)	.220 (.0012)	.040 (.0016)	.226 (.0013)	.046 (.0017)
Representation only	.229 (.0013)	.057 (.0020)	.218 (.0012)	.042 (.0017)	.213 (.0011)	.036 (.0015)	.221 (.0013)	.044 (.0017)
IC proportional hazards	.207 (.0011)	.039 (.0016)	.216 (.0012)	.043 (.0017)	.221 (.0012)	.045 (.0018)	.228 (.0014)	.052 (.0020)
IC survival forest	.202 (.0010)	.035 (.0015)	.211 (.0011)	.039 (.0016)	.213 (.0011)	.040 (.0016)	.219 (.0012)	.046 (.0018)
Discrete-time neural survival	.198 (.0010)	.034 (.0014)	.207 (.0010)	.036 (.0015)	.209 (.0010)	.035 (.0015)	.214 (.0011)	.041 (.0017)
EXPIRE	.193 (.0009)	.028 (.0013)	.201 (.0010)	.030 (.0014)	.200 (.0010)	.026 (.0013)	.206 (.0011)	.031 (.0014)

Table D.3: **S2 within-relation pair accuracy.** Pairs contain one fact expiring within 12 months and one same-relation, same-origin fact surviving beyond 24 months. Parentheses contain MCSEs.

Method	Pair accuracy
Relation only	.500 (.0048)
Metadata/history	.682 (.0042)
Representation only	.745 (.0039)
IC proportional hazards	.655 (.0044)
IC survival forest	.703 (.0041)
Discrete-time neural survival	.724 (.0040)
EXPIRE	.799 (.0035)

The relation-only value of .500 in Table D.3 is expected because two members of a pair have the same relation and anchor. Metadata/history separates some pairs through age and observed history. The representation-only model performs better because U_i is genuinely associated with lifetime in S2, while the full model obtains the strongest ranking by combining both sources.

Table D.4: **Monte Carlo standard deviations of IBS.** Each value is the empirical SD across 500 replications.

Method	S0	S1	S2	S3
Relation only	.0268	.0291	.0313	.0358
Metadata/history	.0224	.0246	.0268	.0291
Representation only	.0291	.0268	.0246	.0291
EXPIRE	.0201	.0224	.0224	.0246

The Monte Carlo standard deviations in Table D.4 confirm that the mean comparisons are not driven by a small number of unusually favorable replications. EXPIRE has the smallest between-replication variability in every design, with SD increasing only from .0201 in S0 to .0246 in S3 as calendar drift makes the prediction problem harder. Variability for the relation-only specification increases more substantially, from .0268 to .0358, reflecting its inability to explain the additional subject-level and time-varying heterogeneity. The representation-only model becomes more stable from S0 to S2 as the semantic signal strengthens, but its variability rises again under S3 when calendar drift is introduced.

Figure D.1 provides two diagnostics not visible from IBS alone. The first panel groups S3 test facts by predicted 12-month survival. The full method remains close to the diagonal, whereas omitting semantic information or both semantic and history information produces larger systematic deviations. The second panel averages the true and fitted post-anchor hazards across S3 test facts. EXPIRE follows both the early-horizon increase and the later decrease induced jointly by the common horizon term and relation-specific calendar drift. These diagnostics serve different purposes. Calibration checks whether the estimated survival probabilities have the correct absolute scale, while hazard recovery checks whether a similar integrated score might conceal an incorrect temporal shape. The agreement of EXPIRE on both criteria supports the interpretation of the primary IBS comparison as genuine lifetime recovery and not only ranking improvement.

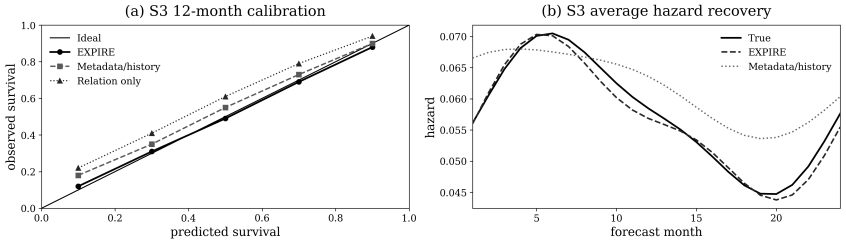


Figure D.1: **Additional controlled-experiment diagnostics.** (a) Twelve-month calibration in S3. The diagonal is perfect calibration. (b) Average post-anchor hazard recovery in S3. All fitted curves are obtained without access to latent event times.

To isolate loss of information from interval coarsening, the S2 design is rerun after replacing the mixed primary inspection spacing by a common spacing of 1, 3, 6, or 12 months. Every other component of the DGP, administrative censoring process, sample split, and tuning procedure is held fixed. As shown in Table D.5, all methods deteriorate as the observation interval widens because less information about the latent event time is available. The change is gradual: for EXPIRE, IBS rises from .196 under monthly resolution to .206 under three-month intervals and .235 under 12-month intervals. Its advantage over the history and relation baselines remains present throughout the examined range. These are the numerical values underlying the interval-width diagnostic in Figure 3.

Table D.5: **S2 sensitivity to expiration-interval width.** Entries are IBS means; parentheses contain MCSEs. A one-month interval is the exact monthly observation limit.

Method	1 month	3 months	6 months	12 months
Relation only	.225 (.0013)	.236 (.0014)	.250 (.0015)	.271 (.0017)
Metadata/history	.206 (.0011)	.218 (.0012)	.233 (.0013)	.251 (.0015)
EXPIRE	.196 (.0009)	.206 (.0010)	.218 (.0011)	.235 (.0013)

Table D.6: **S2 observation and model-design sensitivity.** The missed-stale column uses the fixed 25% retrieval-budget diagnostic, not the certified retrieval rule.

Variant	IBS	Pair acc.	12m ECE	Missed stale
Full EXPIRE	.200	.799	.026	.083
3-month intervals only	.206	.781	.031	.091
Ignore attained age ($A_i = 0$)	.214	.744	.043	.104
Permuted residual LM states	.212	.691	.039	.101
Random projection	.209	.732	.035	.096
Midpoint imputation	.208	.751	.034	.094

Table D.6 reports a more stringent S2 sensitivity set. The first row is the full primary design with its mixed interval-resolution mechanism. "3-month intervals only" forces every observed expiration onto

the three-month inspection grid. "Ignore attained age" sets $A_i = 0$ at fitting time while leaving the data-generating hazard unchanged. This mimics treating every retained spell as though it had just begun at the anchor and therefore removes the attained-age information induced by left-truncated entry. "Permuted residual LM states" permutes relation-centered representations within relation and historical origin. "Random projection" replaces the training PCA map by an independent orthogonal projection of the same selected dimension. "Midpoint imputation" replaces each observed interval by its midpoint before fitting and is included only to quantify the cost of violating the interval-likelihood principle. The sensitivity pattern distinguishes several failure modes. Forcing all events to a coarser inspection grid primarily reduces temporal information. Removing attained age is more damaging because the true hazard depends on $A_i + h$. Permuting residual representations removes the within-relation semantic signal while preserving relation and calendar structure. A random subspace retains some semantic information but does not target the dominant representation directions. Finally, midpoint imputation is less damaging than removing an entire information source but is uniformly inferior to using the observed interval directly.

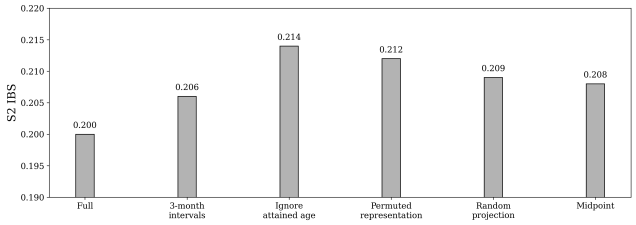


Figure D.2: **S2 sensitivity to observation and model misspecification.** Bars show IBS for the full primary specification, forced three-month intervals, removal of attained-age information, within-relation representation permutation, a random projection of the same dimension, and midpoint imputation.

Figure D.2 makes clear that the strongest degradation is not caused merely by changing the numerical representation basis. Removing attained age is worse, while destroying within-relation semantic information also removes a substantial fraction of the full model's gain.

E Additional Empirical Evaluation

This part gives the empirical scoring rules, baseline implementations, matched-relation construction, falsification and robustness analyses, and retrieval evaluation used to support Section 5.2. The chronology from Appendix C is maintained throughout, $\mathcal{D}^{\mathrm{tr}} \longrightarrow \mathcal{D}^{\mathrm{val}} \longrightarrow \mathcal{D}^{\mathrm{recal}} \longrightarrow \mathcal{D}^{\mathrm{cert}} \longrightarrow \mathcal{D}^{\mathrm{test}}$. No test outcome is used for fitting, model selection, recalibration, retrieval threshold selection, representation preprocessing, or baseline tuning. All probabilistic methods are evaluated on the same factual records and the same observed intervals. An interval-censored event is never replaced by a latent event time or midpoint in the primary analysis. Unless otherwise stated, uncertainty in the empirical tables is quantified by 1,000 subject-cluster bootstrap replicates of the untouched test block, holding all training, validation, recalibration, and certificate fits fixed. The bootstrap therefore measures uncertainty of the reported test-set performance and does not refit the complete development pipeline.

E.1 Metrics, Matching, and Baseline Details

For any fitted survival model, NLL is the test average of the post-anchor observed-data loss from Section 3

$$-\log p_f(W_i) = \begin{cases} -\log\{\hat{S}_i^{\mathrm{res}}(L_i) - \hat{S}_i^{\mathrm{res}}(R_i)\}, & \delta_i = 1, \\ -\log \hat{S}_i^{\mathrm{res}}(C_i), & \delta_i = 0. \end{cases}$$

The numerically stable implementation in Appendix B.1 is used for all methods that expose monthly survival probabilities. This metric uses the observed interval directly and therefore requires no horizon-resolution assumption. Let $Y_i(u) = \mathbf{1}\{\tilde{T}_i > u\}$ be the indicator that the anchor object remains valid beyond horizon u. Under interval censoring, $Y_i(u)$ is not always directly observed. For example, an event interval satisfying $L_i < u < R_i$ straddles the horizon. To avoid treating such records as either events or survivors, the primary Brier and calibration calculations use a method-independent interval-conditional pseudo-outcome. A five-fold subject-cluster cross-fitting scheme is run within the test sample solely for evaluation. On four folds we fit a relation-stratified interval-censored reference survival distribution using relation, anchor quarter, and spell-age quintile. No candidate model prediction

enters this reference fit. The held-out fold then receives $\hat{\mu}_i(u) = \hat{\Pr}\{Y_i(u) = 1 \mid W_i, Z_i\}$, where Z_i denotes the reference-evaluation stratum.

For a horizon-resolved record, $\hat{\mu}_i(u)$ equals the observed indicator. Specifically,

$$\hat{\mu}_i(u) = \begin{cases} 0, & \delta_i = 1,\ R_i \le u, \\ 1, & \delta_i = 1,\ L_i \ge u, \\ 1, & \delta_i = 0,\ C_i \ge u. \end{cases}$$

For an interval-censored record with $L_i < u < R_i$,

$$\hat{\mu}_i(u) = \frac{\hat{S}_{\mathrm{ref}}(u \mid Z_i) - \hat{S}_{\mathrm{ref}}(R_i \mid Z_i)}{\hat{S}_{\mathrm{ref}}(L_i \mid Z_i) - \hat{S}_{\mathrm{ref}}(R_i \mid Z_i)},$$

and for a right-censored record with $C_i < u$,

$$\hat{\mu}_i(u) = \frac{\hat{S}_{\mathrm{ref}}(u \mid Z_i)}{\hat{S}_{\mathrm{ref}}(C_i \mid Z_i)}.$$

Ratios are evaluated only on strata with positive estimated denominator; a sparse stratum is deterministically pooled to the corresponding relation and anchor-year stratum, then to relation alone if necessary. The pooling rule is fixed before model comparison.

For a candidate survival prediction $\hat{S}_i(u)$, the coarsening-aware Brier score is

$$\hat{\mathrm{BS}}(u) = \frac{1}{n_{\mathrm{test}}} \sum_{i \in \mathcal{D}^{\mathrm{test}}} \left[\hat{\mu}_i(u)\{1 - \hat{S}_i(u)\}^2 + \{1 - \hat{\mu}_i(u)\}\hat{S}_i(u)^2 \right].$$

This is the conditional expected squared error of the unobserved binary horizon status given the observed coarsening record and the method-independent reference distribution. The integrated Brier score is the arithmetic mean over $\mathcal{U} = \{3, 6, 12, 18, 24\}$. As a sensitivity check, we also compute an inverse-probability-weighted resolved-outcome Brier score. Let $Q_i(u) = 1$ when $Y_i(u)$ is directly resolved from $(L_i, R_i, C_i, \delta_i)$ and zero otherwise. That diagnostic requires the additional horizon-level condition $Q_i(u) \perp Y_i(u) \mid Z_i$, which is distinct from the coarsening condition used for identification in Appendix A. The IPCW (Inverse Probability of Censoring Weighting) results are numerically similar to the primary interval-conditional scores and are not used for model selection.

For 12-month calibration, predicted expiration probabilities $\hat{p}_i(12) = 1 - \hat{S}_i(12)$ are divided into ten equal-frequency bins. If $\mathcal{B}_k$ is bin k, its coarsening-aware observed expiration frequency isInverse Probability of Censoring Weighting

$$\hat{e}_k = 1 - \frac{1}{|\mathcal{B}_k|} \sum_{i \in \mathcal{B}_k} \hat{\mu}_i(12),$$

and its mean prediction is

$$\hat{p}_k = \frac{1}{|\mathcal{B}_k|} \sum_{i \in \mathcal{B}_k} \hat{p}_i(12).$$

The reported expected calibration error is

$$\hat{\mathrm{ECE}}_{12} = \sum_{k=1}^{10} \frac{|\mathcal{B}_k|}{n_{\mathrm{test}}} |\hat{e}_k - \hat{p}_k|.$$

The same pseudo-outcomes are used for the calibration plots, so a horizon-straddling interval is never silently assigned to one side of the 12-month cutoff.Inverse Probability of Censoring Weighting

The matched analysis is designed to remove the easiest alternative explanation for representation gains: that the LM state merely identifies relations that turn over quickly. An eligible event fact must have a 12-month status that is unambiguously resolved as expired, $R_i \le 12$. An eligible control must be unambiguously known to remain valid beyond 24 months. This holds when either the record is right censored with $C_i \ge 24$ or an observed event interval satisfies $L_i \ge 24$. Each event fact is matched one-to-one, without replacement, to a control having the same relation, the same anchor quarter, and the

same spell-age decile. If several controls satisfy these conditions, the control with the smallest absolute difference in attained age is selected; remaining ties are resolved by a deterministic hash of the subject identifiers. Event facts without an eligible control remain in all global metrics but are omitted from the pair analysis. For the resulting matched set $\mathcal{M}$, pair accuracy is

$$\hat{A}_{\text{pair}} = \frac{1}{|\mathcal{M}|} \sum_{(i,j)\in\mathcal{M}} \left[\mathbf{1}\{\hat{p}_i(12) > \hat{p}_j(12)\} + \frac{1}{2}\mathbf{1}\{\hat{p}_i(12) = \hat{p}_j(12)\} \right],$$

where i is the event fact and j its control. Subject-cluster bootstrap replicates reconstruct the eligible event and control pools and rerun the matching rule, thereby propagating uncertainty from the matching step. The primary matched sample contains 3,842 event-control pairs. Alternative matching definitions give the results in Table E.1. The representation advantage is stable to matching ratio, calendar caliper, and age-matching resolution. The exact-qualifier subset is smaller but gives a similar pair-accuracy gap, indicating that the primary result is not driven by broad snapshot-derived event intervals.

Table E.1: **Sensitivity of the within-relation matched analysis.** Every specification preserves exact relation matching and excludes horizon-unresolved event/control statuses. Parentheses contain subject-cluster bootstrap SEs.

Matching specification	Pairs	EXPIRE pair acc.	EXPIRE-History pair acc.
Primary: quarter + age decile, 1:1	3,842	.699 (.0078)	.636 (.0086)
Quarter + age decile, 1:3	3,711	.696 (.0075)	.634 (.0083)
Anchor within 45 days + age decile	3,605	.697 (.0080)	.635 (.0088)
Quarter + age quintile	4,216	.691 (.0076)	.632 (.0084)
Quarter + attained age within 3 months	3,146	.704 (.0085)	.638 (.0093)
Exact-validity-qualifier subset	1,784	.701 (.0112)	.639 (.0120)

Table E.2: **Empirical baseline implementations.** All survival models use the same 24-month horizon and the same observed interval/right-censoring records.

Method	Predictor / fitting rule	Validation choices
Global Turnbull	One nonparametric interval-censored survival curve, applied prospectively after conditioning on anchor entry.	No structural tuning.
Relation Turnbull	Separate interval-censored survival curve for each modeled relation; sparse relations are pooled only by the fixed support rule from Appendix C.	No structural tuning.
IC proportional hazards	Relation strata with attained age, calendar position, entity/history features, and projected relation-centered LM coordinates.	Ridge penalty $\{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$.
IC survival forest	Interval-censored survival forest using attained age, calendar position, relation, history features, and projected relation-centered LM coordinates.	500 trees; terminal node size $\{25, 50, 100\}$; mtry $\{\lceil\sqrt{p}\rceil, \lceil p/3\rceil, \lceil p/2\rceil\}$.
Discrete-time neural survival	24 monthly hazards from relation embedding, age/calendar features, history, and projected relation-centered LM coordinates.	Two hidden layers; width $\{64, 128\}$; dropout $\{0, .1, .2\}$; weight decay $\{0, 10^{-4}, 10^{-3}\}$.
Mutability / recency class	Fixed relation-mutability category and pre-anchor recency summaries are mapped to a monthly hazard by a monotone discrete-time calibration model. No LM representation is used.	Spline dimension $\{4, 6\}$; ridge $\{0, 10^{-3}, 10^{-2}, 10^{-1}\}$.
LM verbalized probability	At the anchor-time checkpoint, five fixed prompts request expiration probability by 3, 6, 12, 18, and 24 months. The five values are projected onto the monotone cone and converted to a monthly survival curve by linear interpolation of cumulative hazard.	No hazard-model tuning; only the common $\mathcal{D}^{\text{recal}}$ recalibration.
LM residual survival probe	Linear discrete-time hazard using relation intercepts, age/calendar splines, and the relation-centered frozen state. It excludes event history and the horizon-varying low-rank semantic interaction.	$d_p \in \{32, 64, 128\}$; ridge $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$.
EXPIRE-History	Full EXPIRE age, calendar, relation, and event-history specification with $U = 0$.	Corresponding subset of the grid in Appendix B.2.
EXPIRE	Full model in Section 3.	Grid and deterministic tie-breaking in Appendix B.2.

To make probability comparisons fair, every method producing a survival curve receives the same chronological opportunity for recalibration. Its structural parameters are fitted on $\mathcal{D}^{\text{tr}}$, selected on $\mathcal{D}^{\text{val}}$, and refitted on $\mathcal{D}^{\text{tr}} \cup \mathcal{D}^{\text{val}}$. Holding that fitted predictor fixed, the hazard-scale recalibrator from Section 3 is then fit on $\mathcal{D}^{\text{recal}}$ using the same candidate offset-smoothness grid described in Appendix B.2. Neither pair accuracy nor any other rank-only metric is used to choose the calibrator. Methods that are already well calibrated can select a near-identity transformation. An explicit "no

recalibration" ablation is reported below. Table E.2 gives the complete baseline implementations. All tuning criteria are held-out observed-data NLL on $\mathcal{D}^{\mathrm{val}}$ unless the method has no fitted hyperparameter. No baseline uses a test-derived event time, event midpoint, representation projection, or threshold.

For LM verbalization, prompts contain only the current anchor-side fact and a prospective horizon; they do not contain future source revisions or the eventual replacement. If the five raw verbalized probabilities violate monotonicity in horizon, isotonic projection is applied mechanically before any recalibration. This operation uses no outcome labels. The main empirical results in Table 2 use these definitions. In particular, full EXPIRE attains

$$\mathrm{IBS} = .092, \quad \mathrm{NLL} = .331, \quad \mathrm{ECE}_{12} = .018, \quad \hat{A}_{\mathrm{pair}} = .699, \quad \mathrm{BS}_{24} = .131,$$

compared with $.099$, $.352$, $.023$, $.636$, $.141$ for EXPIRE-History. The semantic contribution is therefore visible not only in global probability scores but also in the same-relation ranking analysis.

E.2 FALSIFICATION, ROBUSTNESS, AND GENERALIZATION

We use four falsification conditions to distinguish instance-level semantic information from relation identity or temporal leakage. In the *within-relation permutation* condition, relation-centered states $\tilde{z}_i$ are permuted among facts sharing the same relation and anchor quarter. Relation, age, calendar, and event-history information are left unchanged, but the state can no longer identify which fact within the stratum has which future lifetime. In the *relation-centroid only* condition, the semantic pathway receives only a training-derived relation centroid. Because relation-specific duration and calendar components are already in the hazard, any additional improvement from this pathway can arise only from redundant relation identity, not within-relation information. In the *random subspace* condition, the PCA projection is replaced by a training-independent random orthogonal map having exactly the selected dimension. Relation centering is retained. Finally, the *future-contaminated checkpoint* intentionally violates the point-in-time cutoff rule by extracting the representation from a checkpoint whose documented data cutoff is after the historical anchor. All other features, outcomes, and splits are unchanged. This condition is a leakage diagnostic and is never treated as a valid forecasting baseline.

Table E.3: **Representation falsification and leakage diagnostics.** The future-contaminated condition is invalid and is reported only to quantify look-ahead sensitivity.

Condition	IBS	12m ECE	Pair acc.	24m BS
Full EXPIRE	.092	.018	.699	.131
Within-relation permutation	.099	.024	.639	.141
Relation-centroid only	.099	.025	.636	.142
Random subspace	.096	.022	.674	.137
Future-contaminated checkpoint	.081	.014	.762	.116

The results are shown in Table E.3. Destroying the fact-to-state correspondence removes most of the pair-ranking advantage while leaving relation and calendar information intact. The relation-centroid condition is essentially indistinguishable from the history-only specification. Conversely, the contaminated future checkpoint performs implausibly well, directly illustrating why the point-in-time restriction in Appendix C is necessary.

Table E.4: **Observation-time robustness.** All variants refit the full pipeline using the modified event intervals.

Outcome construction	IBS	NLL	12m ECE	Pair acc.
Primary intervals	.092	.331	.018	.699
Dilated by 30 days	.094	.338	.020	.693
Dilated by 60 days	.096	.347	.023	.685
Dilated by 90 days	.099	.359	.028	.671
Exact validity qualifiers only	.094	.336	.020	.701
Revision/snapshot intervals only	.096	.348	.024	.688

Knowledge-base revisions need not coincide exactly with real-world change times. To test sensitivity to that uncertainty, each event interval is dilated symmetrically by 30, 60, or 90 days, truncated at zero on the left and at the neighboring spell boundary on the right. The fitted methods then receive the dilated interval; no midpoint is introduced. We separately restrict evaluation to facts whose endpoints

are supported by explicit validity qualifiers and to facts whose endpoints are obtained from revision or snapshot brackets. As shown in Table E.4, performance degrades gradually as the observation interval is widened without abruptly collapsing when temporal precision is reduced. Pair ranking is more stable than absolute calibration, which is the expected pattern when the ordering of lifetimes is preserved but their exact temporal localization is blurred.

Table E.5 changes one component of the primary pipeline at a time. "Ignore attained age" sets the anchor age input to zero during fitting and prediction; it does not alter the post-anchor likelihood. "Midpoint imputation" is the only row that replaces each event interval by a single event time and is included solely as a misspecification diagnostic. The largest probability-score deterioration comes from removing relation-specific duration structure or attained age, whereas deleting the LM state most clearly reduces within-relation pair accuracy. Raw, non-residualized states recover part of the semantic gain but calibrate less well, supporting the use of relation centering. Recalibration has little effect on ranking, as expected, but materially reduces ECE.

Table E.5: **Empirical component ablations.** Missed stale is evaluated at the 12-month horizon. The retrieval threshold for each row is selected from $\mathcal{D}^{\text{cert}}$ using row's own calibrated expiration scores.

Variant	IBS	12m ECE	Pair acc.	24m BS	Missed stale	Retrieval %
Full EXPIRE	.092	.018	.699	.131	.047	26.7
No LM representation	.099	.023	.636	.141	.057	27.1
Raw LM state; no relation residualization	.095	.027	.674	.136	.053	26.9
No event history	.098	.023	.663	.139	.055	27.0
Ignore attained age	.100	.025	.655	.143	.056	27.3
No calendar component	.096	.022	.684	.138	.052	26.8
Common duration, not relation-specific duration	.105	.035	.590	.151	.064	28.5
Midpoint imputation	.095	.022	.689	.136	.051	26.8
No hazard recalibration	.094	.044	.699	.134	.055	27.5
Independent horizon calibration	.093	.024	.695	.133	.053	27.0

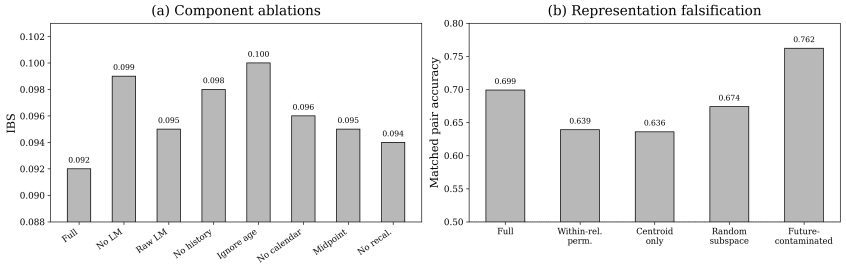


Figure E.1: **Empirical falsification and ablation summary.** (a) IBS after removing or altering individual components of the primary pipeline. (b) Matched within-relation pair accuracy under representation falsifications. The future-contaminated checkpoint is invalid and is included only as a leakage diagnostic.

The aggregate result is not driven by one high-turnover family. Table E.6 reports the primary test composition and relation-specific predictive diagnostics. Relation Turnbull is included to show how much improvement remains after conditioning on relation identity alone. The absolute improvement is largest in faster-changing families, but the low-turnover controls do not reverse the ordering. Pair accuracy remains above one half in every family for which enough matched pairs are available, which is stronger evidence than a global relation-frequency comparison.

Table E.6: **Relation-family diagnostics on the primary 2023 test block.**

Family	N	12m event	Relation Turnbull IBS	EXPIRE IBS	Pair acc.
Corporate leadership	5,842	18.7%	.135	.105	.725
Political and public office	4,916	15.2%	.121	.096	.713
Sports-team affiliation	7,388	24.5%	.163	.121	.740
Coach / manager	3,276	27.8%	.177	.130	.751
Employment / membership	6,104	12.6%	.110	.089	.686
Low-turnover current-object controls	8,452	3.4%	.045	.043	.624

The primary point-in-time checkpoint is complemented by cutoff-documented model-family sensitivities. Native known-fact cohorts differ across models, so two quantities are reported separately: the number of facts each model knows under the same two-of-three rule, and predictive performance on

the intersection cohort known by every compared model. The intersection contains 6,942 facts. For each model in Table E.7, candidate layers are evaluated with identical downstream capacity and the same training examples. Relative to the corresponding history-only specification, the IBS gains are .006, .010, and .014 for OLMo-1B, OLMo-7B, and the primary checkpoint, respectively. The representation effect therefore does not depend on a single layer or one model size.

Table E.7: **Model-size and layer diagnostics.** Performance columns use the common 6,942-fact intersection cohort; "native known" is the number of facts satisfying the model-specific known-fact rule before intersection.

Model	Native known	IBS	12m ECE	Pair acc.	Best layer
OLMo-1B	6,942	.101	.028	.654	11
OLMo-7B	8,518	.096	.023	.682	15
Primary point-in-time LM	9,104	.092	.018	.699	17

We also perform cutoff transfer by shifting every admissible LM checkpoint one historical release earlier while leaving the world-state anchors unchanged. The resulting cohort is smaller because the older model knows fewer current facts, but on the common eligible population EXPIRE has IBS .097 and pair accuracy .681. This is weaker than the primary .092/.699 result but remains better than the corresponding history-only values .103/.629. The reduction is consistent with staler internal representations and not with future-information leakage.

The benchmark estimates remaining lifetime for relations represented in $\mathcal{D}^{\mathrm{tr}}$, as stated in Section 3. We therefore label relation-held-out evaluation as a secondary generalization experiment and not as part of the primary claim. Relations are partitioned into train and held-out families before fitting. For a held-out relation, the learned relation-specific coefficients and centroid are unavailable. A small relation meta-predictor maps a frozen text embedding of the relation description to an age-profile coefficient and a representation centroid. It is trained only on relations available in the training family. All individual factual outcomes for the held-out relation remain unseen until evaluation. Across four held-out relation folds, the mean IBS is .112 for the transfer version of EXPIRE, compared with .120 for a history-only transfer model and .128 for a global duration model. Pair accuracies are .631, .588, and .500, respectively. These results are weaker than the in-inventory primary experiment, as expected, but show that the instance-level semantic signal is not limited to memorizing the fixed relation identifiers.

E.3 Retrieval Evaluation and Ablations

All gating policies use the same retrieval and answer-generation backend. The document index for a query at calendar time t contains only documents whose timestamps are no later than t. Gold benchmark object identifiers, spell endpoints, and explicit expiration annotations are not inserted as retrieval metadata. For each retrieved query, the backend retrieves the top 20 documents using the same dense index, reranks them with the same frozen cross-encoder, retains the top five passages, and supplies those passages to the same answer generator. A retrieval attempt that fails to return current supporting evidence is counted as a retrieval failure; it is not silently replaced by the gold answer. The primary retrieval horizon is 12 months. For a policy $d_i \in \{0, 1\}$, $d_i = 1$ denotes retrieval. Retrieval rate is $n_{\mathrm{test}}^{-1} \sum_i d_i$. Retrieval precision is the coarsening-aware fraction of retrieved facts that expire within 12 months, $\sum_i d_i\{1 - \hat{\mu}_i(12)\}/\sum_i d_i$. Missed stale is the coarsening-aware test estimate

$$\frac{1}{n_{\mathrm{test}}} \sum_i (1 - d_i)\{1 - \hat{\mu}_i(12)\}.$$

For certification itself, however, the threshold is fit only on certificate records whose 12-month expiration indicator is fully resolved, exactly as required by Theorem 4.6. The pseudo-outcome above is used for empirical test reporting, not for certificate construction. QA accuracy is the fraction of final generated answers judged correct under the same alias-based short-answer rule used by the benchmark. The utility shown in Figure 4 and below is

$$\text{Utility} = \text{QA accuracy} - 0.20 \times \text{retrieval rate},$$

where retrieval rate is expressed as a proportion in $[0, 1]$.

For the confidence policy, the 25% of test queries having the lowest anchor-time LM answer confidence are retrieved. Relation recency ranks facts by a training-estimated relation turnover score

together with attained age. Adaptive RAG uses its own fixed adaptive trigger but is thresholded to the same 25% retrieval budget. EXPIRE budget matching retrieves the 25% largest $\hat{p}_i(12)$ values. These budget-matched rows differ only in gating; retrieval documents, reranking, and answer generation are identical. The oracle row is not deployable. It ranks test facts by their eventual 12-month expiration status and is shown only as an upper-bound diagnostic. As shown in Table E.8, at the common 25% budget, expiration-aware gating raises QA accuracy by one percentage point relative to Adaptive RAG and reduces missed stale from .061 to .049. The certified rule retrieves 1.7 percentage points more often than the budget-matched EXPIRE policy and reduces missed stale to .047, matching the main-paper comparison.

Table E.8: **Complete expiration-aware retrieval comparison.** Budget-matched policies retrieve exactly 25% of queries. Utility is QA accuracy minus 0.20 times retrieval rate.

Policy	Retrieval	QA acc.	Retrieval precision	Utility	Missed stale
LM confidence	25.0%	.898	.292	.848	.083
Relation recency	25.0%	.907	.340	.857	.071
Adaptive RAG	25.0%	.916	.380	.866	.061
EXPIRE, budget matched	25.0%	.926	.428	.876	.049
EXPIRE, certified $\alpha = .05$	26.7%	.929	.408	.876	.047
Oracle expiration	25.0%	.941	.504	.891	.030

The certified rule is refit separately for each prespecified target risk and horizon. For every configuration, $\hat{\rho}_u$ is selected only from $\mathcal{D}^{\text{cert}}$ using fully resolved stale indicators. The resulting rule is then frozen before the 2023 test block is evaluated. Theorem 4.6 gives a finite-sample guarantee only under exchangeability of the certificate regime and the future deployment example. Because the primary 2023 test block is chronologically later, the following numbers in Table E.9 are reported as prospective temporal-transfer diagnostics and not as a claim that the theorem covers arbitrary calendar drift. The small excess over .010 in the first row is consistent with the fact that the 2023 block is a later calendar regime and not an exchangeable certificate draw. For the other targets, future missed-stale risk is at or below the nominal level. At fixed $\alpha = .05$, longer horizons require more retrieval, as shown in Table E.10.

Table E.9: **Prospective sensitivity of certified retrieval at the 12-month horizon.** Thresholds are selected only on $\mathcal{D}^{\text{cert}}$; reported risks are on the later 2023 test block.

Target α	Retrieval rate	QA accuracy	Test missed stale
.010	51.9%	.934	.012
.025	38.6%	.933	.025
.050	26.7%	.929	.047
.100	16.8%	.921	.088

Table E.10: **Certified retrieval across deployment horizons at $\alpha = .05$.**

Horizon	Retrieval rate	QA accuracy	Test missed stale
3 months	12.4%	.912	.031
6 months	18.7%	.920	.041
12 months	26.7%	.929	.047

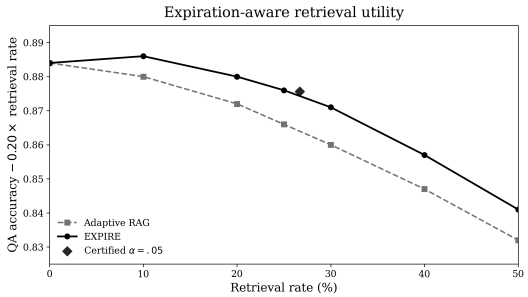


Figure E.2: **Accuracy-retrieval utility frontier.** Utility equals QA accuracy minus 0.20 times retrieval rate. The diamond is the separately certified $\alpha = .05$ operating point; the other curves use fixed retrieval budgets.

Figure E.2 compares the complete fixed-budget frontier and not only the 25% operating point. At each budget, EXPIRE and Adaptive RAG use the same retrieval backend and differ only in which queries are selected. The certified $\alpha = .05$ rule is shown at its data-selected retrieval rate. The advantage of EXPIRE is not confined to one manually selected budget. The utility values remain above Adaptive RAG from 10% through 50% retrieval, with the largest difference in the low-to-moderate retrieval range where selective verification is most useful.

EXPIRE determines when retrieval should be attempted; it does not guarantee that a retrieval system will locate adequate evidence. To separate those two issues, we degrade the retrieval backend after the gating decisions have already been frozen. With probability $1 - a_R$, the otherwise successful supporting passage is replaced by a relation-matched but non-supporting passage. Table E.11 shows that the gating decision and retrieval rate are unchanged. Missed stale is unchanged because it measures stale facts for which retrieval was not triggered. QA accuracy and utility decline as retrieval quality degrades, which is why the paper treats lifetime forecasting and retrieval quality as separate components of deployment risk.

Table E.11: **Sensitivity to imperfect retrieval for the certified $\alpha = .05$ policy.** The retrieval rate remains 26.7%; only backend support success is varied.

Supporting-evidence success a_R	QA accuracy	Utility	Missed stale
1.00	.929	.876	.047
.90	.921	.868	.047
.80	.913	.860	.047
.70	.904	.851	.047

Finally, the benchmark cohort is rebuilt under one-of-three, two-of-three, and three-of-three anchor-time agreement rules. Model fitting, chronological splits, and evaluation are then rerun from scratch for each cohort. Table E.12 shows that the stricter rule produces a smaller and somewhat easier population, but the relative conclusion is unchanged. The primary two-of-three definition retains significantly more facts than unanimity while reducing sensitivity to one prompt-specific failure.

Table E.12: **Sensitivity to the anchor-time known-fact rule.**

Known-fact rule	Test facts	IBS	12m ECE	Pair acc.
At least 1 of 3	44,816	.096	.022	.686
At least 2 of 3 (primary)	35,978	.092	.018	.699
3 of 3	27,604	.090	.017	.703

F Extensions and Additional Diagnostics

This part records extensions that are not required for the primary single-object, all-cause expiration task and gives additional diagnostics that complement Appendix E. The extensions preserve the central prospective distinction of the paper: eligibility and predictors are defined at the anchor, whereas later information is used only to reveal the future validity outcome. None of the results in Sections 2-5 requires the extensions below.

F.1 Model and Decision Extensions

The primary analysis treats the end of present-tense validity as one event because the deployment question is whether the currently correct answer remains safe to use. In some applications it may nevertheless be useful to distinguish why the spell ends. For example, a current officeholder may be replaced by another person, an organization may terminate a role without an immediate successor, or a team affiliation may end because the player moves to a new team. Let $K_i \in \{1, \dots, J\}$ denote the cause of expiration when the event occurs. For month h, define the cause-specific post-anchor hazards

$$\lambda_i^{(k)}(h) = \Pr\Big(h - 1 < \tilde{T}_i \le h,\ K_i = k \mid \tilde{T}_i > h - 1, A_i, X_i\Big).$$

A convenient multinomial parameterization is

$$\lambda_i^{(k)}(h) = \frac{\exp\{f_i^{(k)}(h)\}}{1 + \sum_{\ell=1}^{J} \exp\{f_i^{(\ell)}(h)\}}, \quad k = 1, \dots, J,$$

so the conditional probability of surviving month h is $1 - \sum_{k=1}^{J} \lambda_i^{(k)}(h)$. The overall remaining-survival curve is therefore

$$S_i^{\mathrm{res}}(u) = \prod_{h=1}^{u} \left\{ 1 - \sum_{k=1}^{J} \lambda_i^{(k)}(h) \right\}.$$

The relation, history, calendar, and semantic components in Section 3 can be shared across causes, with cause-specific intercepts or low-rank factors added when the data support them. If an observed interval $(L_i, R_i]$ ends with known cause k, its likelihood contribution is

$$\sum_{h=L_i+1}^{R_i} S_i^{\mathrm{res}}(h-1)\lambda_i^{(k)}(h).$$

If the cause is unknown, the likelihood sums the same expression over k. A right-censored record contributes $S_i^{\mathrm{res}}(C_i)$. Thus competing causes do not require midpoint imputation or any change to the benchmark chronology. The all-cause model in the main paper is recovered by summing the cause-specific hazards. We use that simpler target because retrieval should be triggered whenever the present answer is no longer valid, regardless of the mechanism producing the change.

The primary benchmark excludes relations that permit several distinct objects to be simultaneously correct. This avoids conflating two different survival targets. If the externally valid object is instead a set $O_{sr}(t)$, at least two natural targets are available. For one current member $o \in O_{sr}(\tau)$, define its object-specific remaining lifetime by

$$\tilde{T}_{sro} = \inf\{u > 0 : o \notin O_{sr}(\tau + u)\}.$$

This target asks how long one particular current answer remains acceptable. It reduces directly to the object-specific survival formulation of the main paper, with multiple records from the same subject-relation pair treated as a cluster. Alternatively, define the first set-change time

$$\tilde{T}_{sr}^{\mathrm{set}} = \inf\{u > 0 : O_{sr}(\tau + u) \neq O_{sr}(\tau)\}.$$

This target treats any addition or removal as expiration of the complete anchor-time answer set. The hazard and censoring machinery are unchanged, but the known-fact criterion must compare predicted and valid sets, not just one entity alias. For example, a fixed set-level threshold may be defined using exact set equality or a prespecified F1 requirement. These two targets are scientifically different. A relation can experience a set change while one particular member remains valid, and one member can expire while other members remain current. The primary WIKILIFETIME inventory therefore excludes genuinely concurrent relations and avoids choosing one of these definitions implicitly.

The monthly grid is natural for WIKILIFETIME because many source histories are only localized to intervals of weeks or months. If a future application contains sufficiently dense timestamps, the same model can be written in continuous post-anchor time. Let $\lambda_i(t)$ denote the conditional hazard t time units after the anchor. A continuous analogue of the Section 3 predictor may use $A_i + t$ for total attained age, $\tau_i + t$ for calendar time, and a smooth horizon-varying semantic interaction. Remaining survival is

$$S_i^{\mathrm{res}}(u) = \exp\left\{ -\int_0^u \lambda_i(t)\, dt \right\}.$$

An interval-censored event contributes $S_i^{\mathrm{res}}(L_i) - S_i^{\mathrm{res}}(R_i)$, and a right-censored record contributes $S_i^{\mathrm{res}}(C_i)$. The model is already conditional on survival to the anchor, so no additional pre-anchor likelihood factor is introduced. A piecewise-exponential hazard with one constant hazard in every calendar month gives a continuous-time implementation that approaches the discrete model used in the paper. The primary monthly specification is therefore a resolution choice, not a restriction of the underlying prospective survival formulation.

Corollary 4.7 considers an idealized cost rule in which retrieval is assumed to return a correct answer. This isolates the lifetime-forecasting component of the decision problem. In practice, retrieval can fail and its cost can depend on the query. We fix a deployment horizon u and write

$$p^\star(X; u) = \Pr(\tilde{T} \leq u \mid A, X)$$

for the true expiration probability. Let $c_E(X) > 0$ be the loss of answering directly after expiration, $c_R(X) \geq 0$ the direct cost of attempting retrieval, $a_R(X) \in [0, 1]$ the probability that retrieval returns

adequate current evidence, and $c_F(X) \geq 0$ the residual loss when retrieval fails. The expected loss of a retrieval attempt is

$$L_R(X) = c_R(X) + \{1 - a_R(X)\}c_F(X),$$

whereas the expected stale-answer loss from answering directly is $L_D(X) = c_E(X)p^\star(X;u)$. The Bayes rule retrieves precisely when $c_E(X)p^\star(X;u) > L_R(X)$. Equivalently, if $c_E(X)$ is positive,

$$p^\star(X;u) > \frac{L_R(X)}{c_E(X)}.$$

Thus retrieval quality changes the decision threshold but does not change the lifetime estimand.

Suppose c_E is known, $\hat{p}$ estimates $p^\star$, and $\hat{L}_R$ estimates L_R. Define

$$d^\star(X) = \mathbf{1}\{c_E(X)p^\star(X;u) > L_R(X)\} \text{ and } \hat{d}(X) = \mathbf{1}\{c_E(X)\hat{p}(X;u) > \hat{L}_R(X)\}.$$

At a covariate value where the two rules agree, their conditional excess decision loss is zero. If they disagree, the quantities $c_E p^\star - L_R$ and $c_E\hat{p} - \hat{L}_R$ have opposite signs or one is zero. Therefore

$$|c_E(X)p^\star(X;u) - L_R(X)| \leq c_E(X)\,|\hat{p}(X;u) - p^\star(X;u)| + \left|\hat{L}_R(X) - L_R(X)\right|.$$

Taking expectations yields

$$\mathrm{Risk}(\hat{d}) - \mathrm{Risk}(d^\star) \leq \mathbb{E}\left[c_E(X)|\hat{p}(X;u) - p^\star(X;u)| + |\hat{L}_R(X) - L_R(X)|\right].$$

This decomposition separates two error sources: estimating expiration risk and estimating retrieval quality. If $\hat{L}_R = \hat{c}_R + (1 - \hat{a}_R)\hat{c}_F$, and $c_F, \hat{c}_F \leq C_F$, then

$$|\hat{L}_R - L_R| \leq |\hat{c}_R - c_R| + C_F|\hat{a}_R - a_R| + |\hat{c}_F - c_F|.$$

Hence retrieval-cost, support-success, and failure-loss estimation errors enter additively in a simple upper bound. This extension is distinct from the certificate in Theorem 4.6. The theorem controls the missed-stale loss $\mathbf{1}\{d = 0\}\mathbf{1}\{\tilde{T} \leq u\}$, which asks whether retrieval was triggered when a stale answer would otherwise be used. It does not claim that the retrieval backend succeeds after being triggered. End-to-end answering safety would require a loss that also includes retrieval failure and, correspondingly, certificate data on that backend outcome. Appendix E.3 therefore reports retrieval-quality sensitivity separately and avoids attributing backend failures to the lifetime certificate.

F.2 Additional Empirical Diagnostics

The IBS values in Table 2 average prediction error over five deployment horizons. Table F.1 reports the horizon-specific Brier scores for representative methods. The row averages equal the corresponding IBS values in the main table. The improvement is not concentrated at one forecast horizon. Relative to EXPIRE-History, the Brier reduction for EXPIRE is present at 3, 6, 12, 18, and 24 months. The absolute gap is largest at the longer horizons, where more relation turnover has accumulated and instance-level lifetime heterogeneity has had more time to affect validity.

Table F.1: **Horizon-specific Brier scores on the primary 2023 test block.** The arithmetic mean across the five columns equals the reported IBS.

Method	3m BS	6m BS	12m BS	18m BS	24m BS
Global Turnbull	.076	.103	.130	.147	.184
Relation Turnbull	.069	.093	.118	.139	.166
EXPIRE-History	.059	.077	.098	.120	.141
EXPIRE	.052	.070	.091	.116	.131

Figure F.1 visualizes the relation-family results from Table E.6. Each horizontal segment joins the relation-specific Turnbull IBS to the corresponding EXPIRE IBS, so shorter positions indicate better prediction. The relative IBS reductions are 22.2% for corporate leadership, 20.7% for political/public office, 25.8% for sports-team affiliation, 26.6% for coach/manager relations, 19.1% for employment/membership, and 4.4% for the low-turnover controls. The smaller gain in the last family is expected because very few facts expire within the forecast window, leaving less avoidable Brier error.

The important comparison is not merely that fast-changing relations are easier to identify. Relation Turnbull already conditions completely on relation identity. The remaining improvement therefore reflects within-relation age, history, calendar, and representation information.

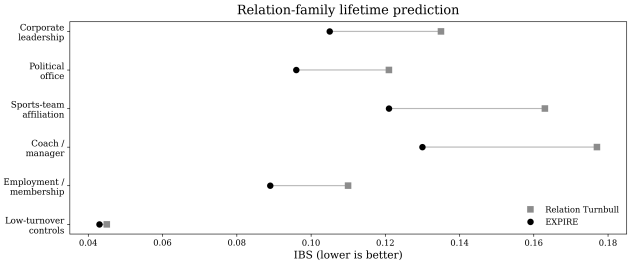


Figure F.1: **Relation-family lifetime prediction.** IBS for the relation-specific Turnbull baseline and EXPIRE on each prespecified test family.

The left panel of Figure F.2 summarizes the common-cohort model-size analysis from Appendix E.2. For the model-size analysis, history-only IBS is .107 for OLMo-1B, .106 for OLMo-7B, and .106 for the primary point-in-time model; adding the frozen representation reduces these values to .101, .096, and .092. The absolute semantic gain therefore increases from .006 to .010 to .014 across the three checkpoints. This pattern should not be interpreted as a universal monotonic law of model scale: the checkpoints differ in more than parameter count. It does show that the primary result is not produced only by one particular representation extractor.

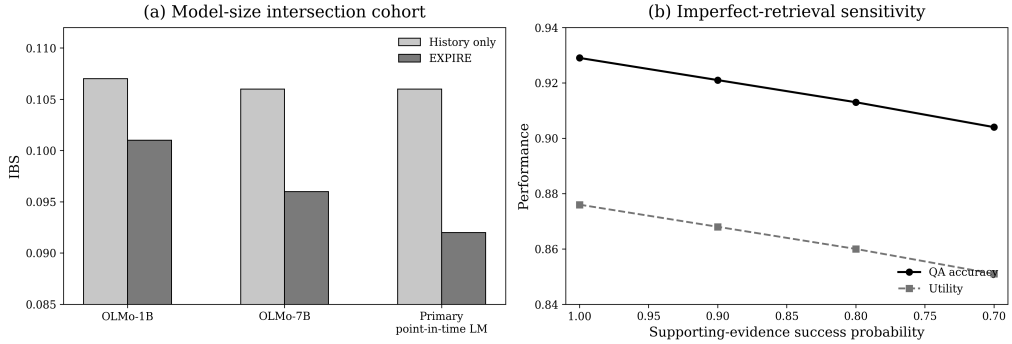


Figure F.2: **Model and deployment diagnostics.** (a) History-only and full-EXPIRE IBS on the common model-size intersection cohort. (b) QA accuracy and utility when the certified retrieval gate is held fixed but supporting-evidence success is degraded.

The right panel of Figure F.2 visualizes the imperfect-retrieval sensitivity in Appendix E.3. The certificate-selected gate is held fixed while the probability that the backend returns adequate current evidence is reduced from 1.0 to .7. As backend success falls from 1.00 to .70, QA accuracy decreases from .929 to .904 and utility from .876 to .851. The retrieval rate and missed-stale trigger risk remain unchanged because the gate itself is fixed. This is the empirical counterpart of the decision decomposition above: expiration prediction determines which facts are sent to retrieval, whereas backend reliability determines the benefit obtained after retrieval is attempted.

As an additional leakage diagnostic, we repeat the primary lifetime comparison after randomly repartitioning the same eligible rows into train/validation/test samples of the same sizes while preserving subject clusters. This experiment is not a legitimate estimate of future performance; it measures how much the task is simplified when chronological shift is removed. Under the random-split evaluation, EXPIRE obtains IBS .083 and 12-month ECE .014, compared with .092 and .018 under the prospective 2023 test. The relation Turnbull baseline changes from .117 to .108. Thus random splitting improves every method and compresses part of the calendar-shift difficulty. The primary paper therefore uses only rolling-origin results for substantive claims.