Do LLMs Have a Sense of Time? Zero-Shot Survival Curve Prediction with Frontier Language Models

Emma Chen YINGCHEN@G.HARVARD.EDU
Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
Harvard John A. Paulson School of Engineering and Applied Sciences, Cambridge, MA, USA

Grace Chang Yuan
Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA

Christine Yang Zhou
Division of Pulmonary, Critical Care, and Sleep Medicine, University of Cincinnati, Cincinnati, OH, USA

David A. Kim
Department of Emergency Medicine, Stanford University School of Medicine, Stanford, CA, USA

Pranav Rajpurkar PRANAV_RAJPURKAR@HMS.HARVARD.EDU
Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

Abstract

Survival curve prediction is an underexplored formulation for time-to-event prediction with large language models (LLMs). We benchmark four frontier LLMs on zero-shot survival curve prediction for emergency department (ED) revisit and hospital readmission, comparing against regression and ordinal classification as alternative prediction formulations. Survival prediction is the top-performing formulation for every model on Readmission and three of four on ED Revisit, with Harrell's C-index ranging from 0.703–0.732 and 0.713–0.802 respectively. Frontier LLMs can produce largely well-formed survival curves zero-shot, with all predictions being valid probabilities and monotonicity holds on 99–100% of predicted curves without post-hoc correction; the initial value condition $S(0) = 1$ is satisfied exactly by two of four models on both tasks and by a third model on Readmission, and approximately otherwise. Beyond structural validity, the best LLMs approach a supervised Random Survival Forest baseline (C-index 0.732 vs. 0.741) on Readmission with no training data. Compared to direct binary prompting, survival-derived probabilities are substantially better calibrated at every horizon tested, with ECE reductions of up to 7.5× at short horizons. Structured multi-agent deliberation achieved C-indices of 0.734 for Readmission and 0.790 for ED Revisit; it significantly improved on the weakest baseline for ED Revisit ($\Delta = +0.031$, two-sided $p = 0.049$ for GPT-5) but did not consistently outperform the strongest individual model. These results establish survival curve prediction as the preferred formulation for LLM-based clinical time-to-event prediction.

1. Introduction

Large language models (LLMs) have emerged as a promising tool for clinical risk prediction (Oh et al., 2024; Han et al., 2023), yet their application to coherent survival curve prediction remains underexplored. Prior LLM survival prompting either reduces to single-horizon

binary classification or queries horizons independently without enforcing survival function coherence (Oh et al., 2024; Jeanselme et al., 2024), overlooking the richer temporal structure that survival analysis captures. The probability that a patient remains event-free at multiple timepoints enables risk to be tracked continuously rather than collapsed to a single yes or no decision.

The stakes are highest for tasks where timing determines the clinical response. ED revisit and hospital readmission are two of the most consequential and costly outcomes in healthcare (Kansagara et al., 2011), yet interventions such as follow-up calls, care coordination, and discharge planning are only effective if deployed at the right moment. A binary readmission flag tells a clinician little about whether to act today or in two weeks.

Our prior work (Chen et al., 2026) evaluated frontier LLMs for 30-day ED revisit prediction and found that no LLM strategy (direct prompting, retrieval-augmented generation, or embedding-based approaches) exceeded a LightGBM (Ke et al., 2017) baseline trained on structured EHR features. This leaves open whether the binary formulation itself is a limiting factor, or whether the underperformance reflects a more fundamental capability gap.

We ask whether LLMs can produce valid survival curves zero-shot, and whether survival curve prediction is the best formulation for these tasks. We benchmark four frontier LLMs: GPT-5, Claude Sonnet 4.5, Gemini 3.1 Pro, and Llama 4 Maverick on ED revisit and hospital readmission. For each task, we compared three prediction formulations: **(1) Regression** asks the model to predict a continuous duration directly; **(2) Ordinal classification** asks the model to assign the patient to a categorical time bin, where bin boundaries are derived from training-set percentiles; **(3) Survival curve prediction** asks the model to predict $S(t) = P(T > t)$ at a set of clinically relevant timepoints, producing a coherent probability sequence over the full observation horizon. All three formulations are evaluated using concordance index as a common discriminative metric, enabling direct comparison.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our results yield four generalizable insights. First, frontier LLMs can produce structurally valid survival curves zero-shot, suggesting that survival analysis is a viable paradigm for clinical time-to-event prediction. Second, survival curve prediction is the preferred formulation for such tasks. It outperforms regression and ordinal classification in seven of eight model–task settings, despite being a harder problem and despite classification having explicit access to training-set event-time distributions. Third, when binary probability estimates drive clinical decisions, survival-derived risk scores are substantially better calibrated than direct binary prompting, particularly at short horizons where binary prompts systematically overestimate low base-rate event probabilities. Fourth, multi-agent deliberation among mixed-vendor LLMs is competitive with individual-model prediction and can surface complementary clinical reasoning, but does not consistently outperform the strongest individual model.

2. Related Work

Prior work applying LLMs to time-to-event prediction in clinical settings has grown substantially since 2022, with oncology representing the most studied domain. Studies have used

GPT-4 and LLaMA-based models to predict survival outcomes in hepatocellular carcinoma (Kavak et al., 2025), non-small cell lung and colon cancer (Kiermeyer et al., 2025), and patients undergoing radiotherapy (Park et al., 2025), alongside work in cardiovascular risk prediction (Han et al., 2024) and sepsis mortality (Oh et al., 2024). Despite this breadth, the dominant paradigm reduces survival prediction to binary classification at a fixed time horizon rather than modeling the full survival distribution. For example, Jeanselme et al. (2024) proposed querying each time horizon independently with a binary probability question on TCGA cancer data, asking the estimated probability that a patient will die within the next h years, repeated separately for each horizon h. This framing sidesteps the core statistical challenge of right censoring: patients whose event time is unobserved after a known follow-up period are either excluded or misclassified, introducing selection bias and discarding the partial information they provide. As a result, these methods cannot produce survival curves, dynamic risk estimates, or well-calibrated probability predictions across the time horizon, and their outputs are not directly comparable to standard survival analysis metrics such as Harrell's concordance index or integrated Brier score.

A parallel line of work treats LLMs as feature extractors or uses transformer-based architectures as encoders that feed into downstream survival models, thereby preserving statistical rigor in the prediction head. Park et al. (2025) used LLaMA-3-70B to structure unstructured EHR notes into discrete prognostic features subsequently ingested by Cox proportional hazards and Random Survival Forest models, improving the concordance index from 0.779 to 0.842 on external validation. Kiermeyer et al. (2025) demonstrated analogous gains in cancer survival by extracting comorbidities and functional status from clinical notes zero-shot and integrating them into survival pipelines. On the architecture side, BERTSurv (Zhao et al., 2021) fine-tuned Clinical BERT with a Cox partial log-likelihood loss, directly connecting a pretrained language model to a censoring-aware training objective. MOTOR (Steinberg et al., 2023) extended this further by pretraining a transformer foundation model on 55 million patient records using a piecewise exponential survival objective, achieving strong label efficiency and cross-site generalization. Most recently, the tabular foundation model framework of Kim et al. (2026) showed that discretizing event times and treating censored bins as missing labels enables in-context survival prediction with theoretical guarantees under standard censoring assumptions.

In contrast to both lines of work, we ask whether frontier LLMs can produce valid survival curves zero-shot, without finetuning, feature extraction pipelines, or access to training data, as well as whether this formulation outperforms other alternatives for clinical time-to-event prediction.

3. Methods

3.1. Clinical Prediction Tasks

Time to ED Revisit. Predicts days until ED return for discharged patients. Patients without revisit within 60 days are right-censored. Features are extracted *at ED discharge*, including labs, diagnosis codes, and length of stay.

Time to Hospital Readmission. Predicts days from discharge to readmission. Patients without readmission within 365 days are right-censored. Features are extracted *at hospital*

discharge, including the complete admission record. Patients who died during hospitalization and same-day readmissions are excluded.

3.2. Prediction Formulations and Label Generation

Regression. Direct prediction of the continuous outcome variable (days). We use the observed time-to-event for patients who experienced the event. LLMs are asked to predict duration for all patients or -1 if no event is expected.

Classification. Prediction of categorical time bins. We compute decile thresholds from events occurring *within the observation window* only (60 days for ED Revisit, 365 days for Readmission), and add an 11th "No Event" class for patients who did not experience the event within the cutoff. This includes both right-censored patients and patients whose event occurred after the cutoff.

Survival. Prediction of the survival function $S(t) = P(T > t)$ at discrete time points (Table 1). Each observation includes an event indicator $\delta_i \in \{0, 1\}$ where $\delta_i = 1$ indicates the event was observed and $\delta_i = 0$ indicates right-censoring. Patients who do not revisit/readmit within the observation window (60 or 365 days) are censored with $\delta_i = 0$ and T_i set to the censoring threshold.

Table 1: Survival analysis time points for each task.

Task	Time Points
ED Revisit	1, 2, 3, 7, 14, 30, 60 days
Readmission	7, 14, 30, 60, 90, 180, 365 days

3.3. LLM Prompting Strategy

We construct structured natural language prompts that present clinical information in a format similar to clinical documentation. Each prompt contains: (1) **System context:** Task-specific framing establishing the clinical scenario (e.g., "You are an emergency department physician..."); (2) **Clinical data sections:** Features organized by category (demographics, clinical characteristics, laboratory results, medical history); and (3) **Prediction instructions:** Explicit request for a prediction type with structured output format, with each of the three prediction types queried via a separate, independent model call. All predicted probabilities are read directly from the numeric values the model writes in its JSON response; we do not use token logits or any internal model probabilities. (Appendix D)

3.4. Model Configuration and Inference

We evaluated four large language models, three proprietary and one open-weight. These models include GPT-5 (gpt-5-2025-08-07), Claude Sonnet 4.5 (claude-sonnet-4-5-20250929), Gemini 3.1 Pro (gemini-3.1-pro-preview), and Llama 4 Maverick (Llama-4-Maverick-17B-128E-Instruct-FP8). Unless stated otherwise, all main-text results use greedy decoding at temperature 0 with `top_p`=1. GPT-5 exposes no temperature control and is queried at its fixed default (temperature 1). Temperature-sensitivity analysis can be found in Appendix J.

3.5. Baseline Models

We compare LLMs against three baselines drawn from the same underlying patient records: LightGBM (Ke et al., 2017) regressors for the regression formulation, LightGBM classifiers for classification, and Random Survival Forests (RSF) (Ishwaran et al., 2008) for survival. We also evaluated two additional survival baselines, Cox Proportional Hazards regression (CoxPH) (Cox, 1972) and DeepSurv neural networks (Katzman et al., 2018), and reported the results in Appendix L. All baseline model hyperparameters were tuned using Optuna (Akiba et al., 2019) with 20 trials per model, optimizing MAE for regression, log-loss for classification, and C-index for survival models (Appendix C). Numeric features are standardized (z-score normalization); missing numeric values are imputed with the training-set median before standardization, and the training medians and scaling statistics are applied unchanged to the validation and test splits. Categorical features are one-hot encoded with an explicit missing level, chief complaints are encoded via multi-hot encoding over the 50 most frequent complaints, and clinical text fields (diagnoses, medications, procedures) are converted to structured features using domain-specific encodings (e.g., CCS diagnosis categories, medication drug classes, procedure type indicators). Structured features derived from records that are absent for a given patient (for example a laboratory test that was never drawn) are encoded as zero prior to standardization and, for the laboratory panels, are accompanied by an explicit per-test missingness indicator so that an unmeasured test remains distinguishable from a measured value. All vocabularies and encoders (top-k complaint and CCS category lists, one-hot levels, imputation medians, and scaling statistics) are fit on the training split only. (Appendix B)

3.6. Evaluation Metrics

3.6.1. Harrell's Concordance Index.

We compare the three prediction formulations using Harrell's concordance index (C-index) (Harrell et al., 1982), which measures the probability that for a randomly selected pair of patients, the one with shorter predicted time actually experiences the event first:

$$C = \frac{\sum_{i,j} \mathbf{1}[T_i < T_j] \cdot \mathbf{1}[\hat{T}_i < \hat{T}_j] \cdot \delta_i}{\sum_{i,j} \mathbf{1}[T_i < T_j] \cdot \delta_i} \tag{1}$$

where T_i is the true event time, $\hat{T}_i$ is the predicted time, δ_i is the event indicator, and the sum is over all comparable pairs; tied predicted times count as half-concordant. Harrell's C-index handles censored observations and ranges from 0.5 (random) to 1.0 (perfect concordance).

For survival predictions, risk is quantified as the restricted mean survival time (RMST) (Royston and Parmar, 2013), computed via trapezoidal integration of the predicted survival curve $\hat{S}(t)$ across all queried timepoints, with higher RMST indicating lower risk. The integration begins at $t = 0$ with $\hat{S}(0) = 1$ by definition (the model's predicted $\hat{S}(0)$ is used only in the structural-validity analysis of Appendix E, not in the RMST risk score). For classification, the predicted ordinal bin rank serves as the risk score, with the NoEvent bin assigned the lowest risk. For regression, the predicted event time is used directly as the risk score; predictions of "no event within the observation window" (i.e., beyond the censoring boundary) are mapped to the boundary time and treated as a tied low-risk group.

Significance of pairwise C-index differences is assessed via a paired bootstrap test on the common patient intersection (N=1000), with unadjusted two-sided p-values reported.

3.6.2. Formulation-Specific Metrics

Mean Absolute Error (Regression). Computed on uncensored observations ($\delta_i = 1$) for which the model committed to a finite event time. Predictions of "no event within the observation window" (the -1 sentinel) are excluded from the average rather than scored as errors, so MAE is *conditional* on the model predicting an event time:

$$\text{MAE} = \frac{1}{|\mathcal{E}|} \sum_{i \in \mathcal{E}} |T_i - \hat{T}_i|, \qquad \mathcal{E} = \{\, i : \delta_i = 1 \ \wedge \ \hat{T}_i \neq -1 \,\}. \tag{2}$$

Because a model can lower its MAE simply by predicting "no event" on hard cases, we report MAE together with recall and the rate at which true events are wrongly predicted as "no event" in Appendix G (Table 10), so that abstention is not mistaken for accuracy.

Quadratic Weighted Kappa (Ordinal Classification). An ordinal agreement metric (Cohen, 1968) that penalizes classification errors by the squared distance between predicted and true bin labels:

$$\kappa = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}} \tag{3}$$

where O_{ij} is the observed confusion matrix, E_{ij} is the expected matrix under random agreement, and $w_{ij} = (i-j)^2/(K-1)^2$ are quadratic weights for K classes. QWK ranges from -1 to 1, with 0 representing chance agreement and 1 perfect agreement.

Integrated Brier Score (Survival Analysis). The Brier score (Brier, 1950) at time t measures the calibration of predicted survival probabilities:

$$\text{BS}(t) = \frac{1}{n} \sum_{i=1}^{n} \Big(\hat{S}_i(t) - \mathbf{1}[T_i > t] \Big)^2 \tag{4}$$

where $\hat{S}_i(t)$ is the predicted survival probability for individual i. To obtain a single summary measure over the evaluation time range, we compute the Integrated Brier Score (IBS):

$$\text{IBS} = \frac{1}{t_{\max} - t_{\min}} \int_{t_{\min}}^{t_{\max}} \text{BS}(t)\, dt \tag{5}$$

approximated using the trapezoidal rule over discrete time points. We assume administrative censoring (fixed 60- and 365-day windows) and report the unweighted Brier score rather than an IPCW-adjusted one. Lower values indicate better calibration.

4. Cohort

4.1. Cohort Selection

We use two public clinical datasets with complementary characteristics. **MC-MED** (Kansal et al., 2025) contains emergency department visits from an academic medical center. We

restrict to discharged patients (excluding admissions, transfers, deaths) for ED Revisit prediction, yielding 70,246 visits of which 15,433 (22.0%) have an observed revisit within 60 days and 54,813 are censored. **MIMIC-IV** (Johnson et al., 2023, 2024) contains hospital admissions from Beth Israel Deaconess Medical Center. For Readmission, we exclude in-hospital deaths and same-day readmissions, following patients for 365 days post-discharge. The resulting cohort comprises 519,837 admissions from 217,948 patients, of which 222,695 (42.8%) have an observed readmission within 365 days and 297,142 are censored.

We use MC-MED's provided chronological train/validation/test splits, which enforce both temporal order and patient separation. For MIMIC-IV, we construct patient-level splits (70/10/20), ordering patients by their first admission and keeping all of a patient's admissions in the same split to prevent leakage across a patient's multiple admissions. LLM evaluations use a random subsample of 500 test patients per task to manage API costs; baselines are evaluated on the same subsample for fair comparison.

4.2. Feature Extraction

Features are extracted at the natural clinical decision point for each task to prevent leakage: ED discharge for ED Revisit and hospital discharge for Readmission. Table 5 in Appendix A summarizes feature availability by task.

LightGBM baselines and LLM prompts are built from the same underlying records (Table 6 in Appendix A). LightGBM uses fixed engineered representations: one-hot encoding for categoricals, multi-hot encoding over the top-50 diagnosis categories and a fixed set of therapeutic medication classes (16 for MC-MED, comprising 8 home and 8 ED classes; 10 for MIMIC), and 25 predefined labs summarized as structured features (most recent value, abnormal flag, and an explicit missingness indicator). These caps exist because a tabular model must commit to a fixed-width feature vector, and admitting every distinct code or lab would explode its dimensionality. The LLM faces no such constraint because it reads the record as natural language, so it receives all available data, including full diagnosis and procedure descriptions, all lab results with abnormal flags and aggregated statistics, and all medication names (plus during-visit vital signs with temporal aggregation for ED Revisit). We report this asymmetry transparently rather than restricting the LLM to the baselines' capped feature set: it favors neither approach uniformly, as LightGBM benefits from feature engineering tuned for gradient boosting while the LLM can exploit information the fixed encodings discard.

5. Results

5.1. LLM Performance on Clinical Time-to-Event Prediction

5.1.1. Survival prediction achieves the strongest discrimination across tasks and models

Before evaluating predictive performance, we verified that the zero-shot survival curves satisfy the formal conditions of a valid survival function (monotonicity, valid probability range, and correct initial value). All predicted values lie within $[0, 1]$ across all models and tasks. Claude, Gemini, and GPT-5 satisfy monotonicity on 100% of predictions; Llama satisfies it on 99.4% of ED Revisit predictions and 100% of Readmission predictions. The

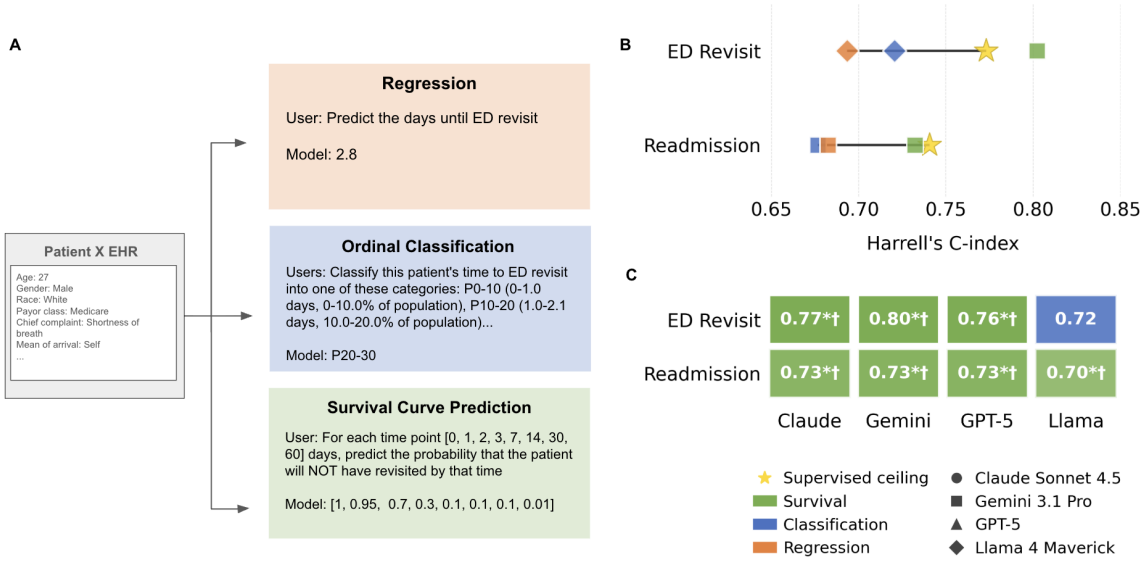

Figure 1: Overview of prediction formulations and model performance. (A) Three formulations for time-to-event prediction: regression predicts a continuous duration, ordinal classification assigns patients to categorical time bins, and survival curve prediction asks the model to produce a sequence over discrete timepoints, interpreted as $\hat{S}(t) = P(T > t)$ across the full observation horizon. (B) Best concordance index (C-index) achieved by any model for each prediction formulation and task; star markers indicate the best supervised baseline across all formulations for that task. (C) C-index of the best-performing formulation for each model–task combination.

initial value condition ($S(0) = 1$) is satisfied exactly by Claude and Llama on both tasks, and by Gemini on Readmission. Gemini (on ED Revisit) and GPT-5 (on both tasks) show a more systematic tendency to predict $S(0)$ slightly below 1 rather than exactly 1, as discussed in Appendix E.

Survival prediction achieves the strongest discrimination across tasks and models (Figure 1B; Table 2). On ED Revisit, survival C-index values range from 0.713 (95% CI: 0.665–0.757; Llama 4 Maverick) to 0.802 (95% CI: 0.761–0.843; Gemini 3.1 Pro), compared with 0.631–0.721 for classification and 0.639–0.693 for regression. On Readmission, survival again leads, ranging from 0.703 (95% CI: 0.668–0.739; Llama 4 Maverick) to 0.732 (95% CI: 0.700–0.766; Gemini 3.1 Pro), versus 0.609–0.677 for classification and 0.572–0.682 for regression. Survival is the top-performing formulation for every model on Readmission and for three of four models on ED Revisit, with statistically significant gains over both regression and classification in most cases ($p < 0.05$; Table 2).

Despite classification having structural access to bin boundaries derived from the training population, effectively encoding prior knowledge about the event-time distribution, it does not consistently dominate. Survival prediction matches or exceeds classification in

seven of eight model–task settings, which suggests that frontier LLMs already encode general intuitions about the temporal dynamics of clinical events such as ED revisit and hospital readmission.

Table 2: Harrell's C-index comparison across prediction types for censored tasks. Regression uses predicted time directly (predictions beyond the observation window are tied at the boundary); Classification uses ordinal bin rankings (NoEvent = lowest-risk tied group); Survival uses RMST across all timepoints. The supervised baseline row combines LightGBM (Reg, Cls) and RSF (Surv). 95% bootstrap CIs shown in subscript brackets. Significance markers on Surv: $*p<0.05$ vs. regression; $\dagger p<0.05$ vs. classification (two-sided paired bootstrap, $N=1000$). Best per model/task is **bolded**.

Task	Model	Reg	Cls	Surv
Readmission	LightGBM/RSF	$0.702_{[0.667,0.734]}$	$0.719_{[0.687,0.754]}$	$\mathbf{0.741}^{*}_{\mathbf{[0.708,0.773]}}$
	Claude	$0.671_{[0.640,0.704]}$	$0.676_{[0.642,0.711]}$	$\mathbf{0.725}^{*\dagger}_{\mathbf{[0.691,0.759]}}$
	Gemini	$0.682_{[0.647,0.718]}$	$0.677_{[0.641,0.713]}$	$\mathbf{0.732}^{*\dagger}_{\mathbf{[0.700,0.766]}}$
	GPT-5	$0.572_{[0.535,0.610]}$	$0.609_{[0.573,0.647]}$	$\mathbf{0.729}^{*\dagger}_{\mathbf{[0.696,0.764]}}$
	Llama	$0.595_{[0.564,0.630]}$	$0.632_{[0.596,0.670]}$	$\mathbf{0.703}^{*\dagger}_{\mathbf{[0.668,0.739]}}$
ED Revisit	LightGBM/RSF	$0.665_{[0.609,0.722]}$	$0.736_{[0.690,0.783]}$	$\mathbf{0.773}^{*\dagger}_{\mathbf{[0.726,0.820]}}$
	Claude	$0.639_{[0.594,0.686]}$	$0.712_{[0.669,0.760]}$	$\mathbf{0.766}^{*\dagger}_{\mathbf{[0.719,0.812]}}$
	Gemini	$0.675_{[0.628,0.725]}$	$0.631_{[0.588,0.676]}$	$\mathbf{0.802}^{*\dagger}_{\mathbf{[0.761,0.843]}}$
	GPT-5	$0.685_{[0.636,0.733]}$	$0.713_{[0.668,0.764]}$	$\mathbf{0.759}^{*\dagger}_{\mathbf{[0.708,0.805]}}$
	Llama	$0.693_{[0.644,0.739]}$	$\mathbf{0.721}_{\mathbf{[0.677,0.769]}}$	$0.713_{[0.665,0.757]}$

5.1.2. Survival predictions show favorable calibration properties

Beyond discrimination, we assess whether predicted survival probabilities accurately reflect observed outcomes. The integrated Brier score (IBS) summarizes calibration and discrimination jointly over the full time horizon, with lower values indicating better performance (Appendix F, Table 9). On ED Revisit, LLMs achieved IBS from 0.103 (Gemini 3.1 Pro) to 0.134 (Claude Sonnet 4.5). On Readmission, LLMs achieved 0.192 (GPT-5) to 0.208 (Llama 4 Maverick). The best-performing LLMs match or approach the supervised RSF baseline (IBS 0.107 on ED Revisit, which Gemini's 0.103 slightly improves upon; 0.168 on Readmission) despite having no access to training data, while the wide IBS range across LLMs indicates that calibration quality varies substantially across frontier models and is not guaranteed by strong discrimination alone (e.g., Claude Sonnet 4.5 achieved strong discrimination on ED Revisit yet the worst IBS among all LLMs on that task).

Figure 2 shows calibration curves over time for ED Revisit and Readmission: each point corresponds to one queried timepoint, with mean predicted $P(T > t)$ on the x-axis and the observed fraction of patients who remain event-free beyond t on the y-axis. GPT-5 and Gemini 3.1 Pro demonstrate the tightest alignment with the diagonal, consistent with their low IBS values. This favorable calibration likely arises from the structure of

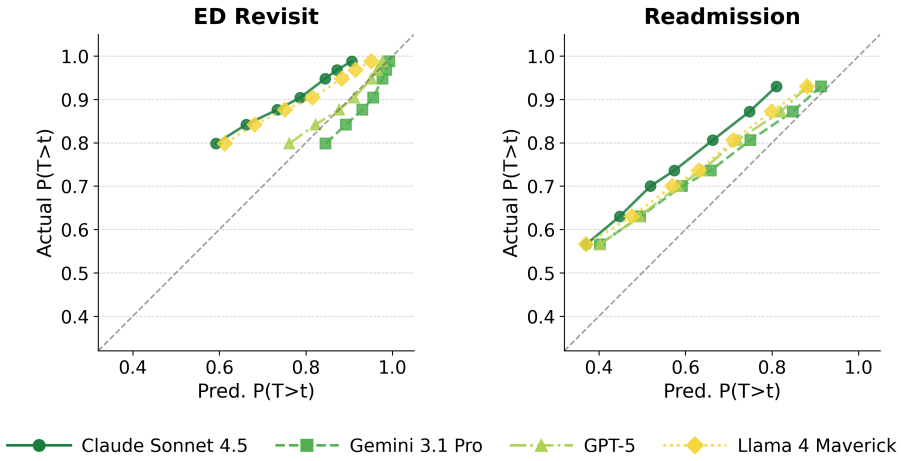

Figure 2: Calibration diagrams for ED Revisit (left) and Readmission (right). Each point corresponds to one queried timepoint; the x-axis shows mean predicted $P(T > t)$ and the y-axis shows the observed fraction of patients who remain event-free beyond t. Diagonal dashed line indicates perfect calibration.

the survival formulation itself, where predicting a coherent probability sequence across multiple timepoints enforces internal consistency, acting as an implicit regularizer that yields calibration properties considerably better than one might expect from a zero-shot model without post-hoc recalibration. Nevertheless, almost all survival curves fall in the upper triangle of the plots, indicating that all models tend to overestimate the risk of revisit or readmission since the actual event-free rate is higher than predicted at nearly every timepoint.

5.1.3. Formulation-specific metrics reveal complementary strengths and weaknesses

We next examine metrics tailored to each prediction type to characterize how each formulation succeeds or fails.

Regression (MAE). MAE is computed only on patients with observed events within the observation window; LLM predictions of -1 (i.e., "no event") are additionally excluded, meaning LLM and LightGBM MAEs are not directly comparable. With that caveat, on ED Revisit (60-day window), LLMs are off by 11.1–15.0 days on average, well below LightGBM's 38.0 days, and on Readmission (365-day window), LLMs achieve MAEs of 58.8–65.0 days versus 100.5 days for LightGBM (Table 9, Appendix F). The lowest ED Revisit MAE (Gemini 3.1 Pro, 11.1 days) reflects the heaviest abstention rather than the sharpest timing: that model labels 54.5% of true events "no event" and is therefore scored on only the remaining 45.5% (Table 10, Appendix G). Note that the LightGBM regressor is trained on a mix of patients whose event falls within and beyond the observation window (Appendix B). The long-horizon exemplars are what let it rank the censored majority as low risk, but they also pull its predictions upward, inflating its absolute error on the patients who do

experience an event within the window. Inspection of the predicted values reveals that LLM regression predictions cluster into distinct vertical bands, indicating that models anchor on round estimates rather than producing patient-specific predictions. The predicted durations also tend to underestimate actual revisit/readmission times, particularly at longer horizons. (Figure 3, Appendix F).

Classification (QWK). QWK values are modest in absolute terms: 0.319–0.356 on ED Revisit and 0.148–0.321 on Readmission. On ED Revisit all LLMs exceed the LightGBM baseline (0.289), whereas on Readmission they remain below it (0.461). Llama 4 Maverick achieves the highest QWK on ED Revisit (0.356) and Gemini 3.1 Pro the highest on Readmission (0.321) (Table 9, Appendix F). Nevertheless, across the event-time bins shown in Figure 4 (Appendix F), all models over-predict the event bins, with the predicted-frequency line often sitting above the ground-truth bars. Because predicted mass must sum to one, this redistribution corresponds to under-prediction of the NoEvent bin in seven of the eight model–task settings, with the exception of Gemini 3.1 Pro on ED Revisit, which instead over-assigns NoEvent by 8 percentage points (87.8% predicted vs. 79.8% observed), consistent with the same abstention behavior that drives its low regression recall (Appendix G). Further, Llama 4 Maverick concentrates predictions at a small number of bins, anchoring on prototypical duration estimates; the remaining three models over-weight early bins, assigning near-zero probability mass to later time intervals.

Together, these patterns of anchoring on prototypical durations and systematically underestimating durations contrast with the survival formulation, whose predicted probabilities, while biased toward overestimating risk, track the observed survival fraction with comparatively modest deviations across queried timepoints.

5.2. Survival-based vs. Binary Prediction Across Multiple Horizons

Because ED revisit and hospital readmission are canonically framed as binary outcomes in the clinical literature, they provide a direct test of whether the gains from survival-based prediction persist when the binary formulation is the natural alternative. Therefore, we compare two approaches: (1) **direct binary prompting**, where the model is asked a horizon-specific yes/no question and returns a score which we interpret as a calibrated probability; and (2) **survival-derived binary**, where the same survival curve $\hat{S}(t)$ is used and the event probability at horizon t is read off as $1 - \hat{S}(t)$. We evaluate both approaches on ED Revisit at $t \in \{3, 7, 14, 30\}$ days and Readmission at $t \in \{7, 30, 90, 180\}$ days, using GPT-5. For each horizon we report AUROC as the discrimination metric and Expected Calibration Error (ECE) as the calibration metric, where ECE summarizes the weighted mean absolute difference between predicted probabilities and observed event rates across probability bins (lower is better).

5.2.1. Discrimination is comparable across horizons, with neither approach dominating consistently

On Readmission, survival-derived predictions achieve slightly higher AUROC than direct binary prompting at every horizon, though all differences are small and statistically non-significant ($p \geq 0.162$). On ED Revisit, direct binary prompting achieves higher AU-

ROC at every horizon, with statistically significant advantages at 7 days (0.815 vs. 0.736; $p < 0.001$) and 30 days (0.823 vs. 0.769; $p < 0.001$) and non-significant differences at 3 and 14 days. Overall, neither approach dominates across both tasks, and survival-derived predictions remain competitive while estimating all horizons simultaneously from a single prompt, whereas binary prompting requires one model call per horizon. (Table 3)

Table 3: C-index (AUROC) and calibration ECE comparing direct binary prompting vs. binarized survival prediction ($1 - \hat{S}(t)$) at multiple horizons (GPT-5, $n = 500$). Binary: horizon-specific prompt. Survival: survival curve evaluated at t. p-values from paired bootstrap test (1,000 resamples). Values shown as point estimate with 95% CI in subscript brackets. **Bold**: better of the two methods.

Horizon	C-index (AUROC)			ECE (calibration, ↓)		
	Binary	Survival	p	Binary	Survival	p
Readmission (MIMIC-IV)						
7d	$0.765_{[0.672, 0.841]}$	$\mathbf{0.786}_{[0.698, 0.861]}$	0.312	$0.255_{[0.229, 0.279]}$	$\mathbf{0.048}_{[0.033, 0.073]}$	< 0.001
30d	$0.788_{[0.738, 0.834]}$	$\mathbf{0.802}_{[0.753, 0.848]}$	0.162	$0.249_{[0.217, 0.281]}$	$\mathbf{0.108}_{[0.085, 0.143]}$	< 0.001
90d	$0.763_{[0.715, 0.808]}$	$\mathbf{0.774}_{[0.730, 0.816]}$	0.178	$0.236_{[0.201, 0.273]}$	$\mathbf{0.136}_{[0.106, 0.175]}$	< 0.001
180d	$0.778_{[0.734, 0.819]}$	$\mathbf{0.780}_{[0.740, 0.822]}$	0.782	$0.214_{[0.183, 0.254]}$	$\mathbf{0.136}_{[0.111, 0.182]}$	< 0.001
ED Revisit (MC-MED)						
3d	$\mathbf{0.773}_{[0.680, 0.855]}$	$0.735_{[0.625, 0.835]}$	0.344	$0.120_{[0.103, 0.140]}$	$\mathbf{0.016}_{[0.005, 0.034]}$	< 0.001
7d	$\mathbf{0.815}_{[0.753, 0.876]}$	$0.736_{[0.653, 0.818]}$	< 0.001	$0.138_{[0.118, 0.162]}$	$\mathbf{0.024}_{[0.014, 0.051]}$	< 0.001
14d	$\mathbf{0.766}_{[0.703, 0.828]}$	$0.743_{[0.675, 0.812]}$	0.294	$0.137_{[0.112, 0.164]}$	$\mathbf{0.039}_{[0.022, 0.066]}$	< 0.001
30d	$\mathbf{0.823}_{[0.772, 0.875]}$	$0.769_{[0.707, 0.825]}$	< 0.001	$0.135_{[0.116, 0.166]}$	$\mathbf{0.063}_{[0.043, 0.096]}$	< 0.001

5.2.2. Survival-derived predictions are better calibrated, particularly at short horizons

Survival-derived predictions show consistently and substantially lower ECE than direct binary prompting at every evaluated horizon, with all differences statistically significant ($p < 0.001$) (Table 3). For Readmission, binary ECE ranges from 0.214 to 0.255 across horizons (7, 30, 90, and 180 days), whereas survival-derived ECE ranges from 0.048 to 0.136. The calibration gap narrows as the horizon extends: at 7 days, binary ECE (0.255) is roughly five times larger than survival ECE (0.048), whereas at 180 days the ratio shrinks to 1.6× (0.214 vs. 0.136). For ED Revisit, binary ECE ranges from 0.120 to 0.138 across horizons (3, 7, 14, and 30 days), while survival ECE ranges from 0.016 to 0.063. Binary predictions tend to exhibit overconfidence, with predicted probabilities systematically exceeding observed event rates (Appendix F, Figure 5). This likely arises because the binary framing anchors the model on a single probability estimate without the distributional context that the survival formulation provides. A decision curve analysis assessing the clinical utility of both approaches across decision thresholds is provided in Appendix I.

5.3. Multi-Agent Conversation

Beyond evaluating each LLM in isolation, we investigated whether structured deliberation among models in a multi-agent framework changes survival curve prediction relative to

individual models. Following the Multi-Agent Conversation (MAC) framework (Yuan et al., 2026), we assigned our top three LLMs in survival curve predictions (GPT-5, Gemini 3.1 Pro, and Claude Sonnet 4.5) as doctor agents, with GPT-5 additionally serving as the supervisor. In each conversation, the three doctors produce survival curves in a round-robin format, each building on the previous doctors' predictions and reasoning. After each round, the supervisor evaluates the predictions, challenges reasoning where appropriate, and either continues the discussion or terminates with a final prediction if consensus is reached. Discussions run for up to three rounds.

5.3.1. MAC narrows the gap between the weakest and strongest models

Table 4: Multi-agent conversation (MAC) versus single-model baselines on Readmission and ED Revisit (n=500 per task). C-index (↑): Harrell's C-index with 95% CI. IBS (↓): integrated Brier score. Δ: paired C-index difference (MAC − baseline). $^*p < 0.05$ (two-sided paired bootstrap, B=10,000). Best result per column is bolded.

Task	Method	C-index [95% CI]	Δ	p	IBS
Readmission	MAC	**0.734** [0.701, 0.767]	—	—	0.1922
	GPT-5	0.729 [0.697, 0.761]	+0.005	0.407	**0.1921**
	Gemini 3.1 Pro	0.732 [0.699, 0.764]	+0.002	0.661	0.1928
	Claude Sonnet 4.5	0.725 [0.692, 0.759]	+0.009	0.202	0.2071
ED Revisit	MAC	0.790 [0.743, 0.834]	—	—	**0.1012**
	GPT-5	0.759 [0.707, 0.806]	+0.031	0.049*	0.1082
	Gemini 3.1 Pro	**0.802** [0.759, 0.843]	−0.012	0.287	0.1026
	Claude Sonnet 4.5	0.766 [0.717, 0.813]	+0.024	0.142	0.1339

Across both tasks, MAC was statistically indistinguishable from the strongest individual model; on ED Revisit it significantly outperformed the weakest model on discrimination and achieved the lowest IBS. On Readmission, MAC had the highest point estimate of C-index (0.734 [95% CI: 0.701–0.767]), but none of its differences from the individual models reached significance (all $p \geq 0.20$), and GPT-5 had a marginally lower IBS (0.1921 vs. 0.1922). On ED Revisit, Gemini had the highest C-index (0.802 [95% CI: 0.759–0.843]), ahead of MAC (0.790 [95% CI: 0.743–0.834]), though the gap was not significant ($\Delta = -0.012$, $p = 0.287$). MAC significantly outperformed the weakest baseline, GPT-5 ($\Delta = +0.031$, $p = 0.049$), but its advantage over Claude ($\Delta = +0.024$, $p = 0.142$) was not significant.

5.3.2. Qualitative analysis of MAC discussions

To better understand how multi-agent discussion reshapes predictions at the case level, a practicing physician reviewed a set of MAC transcripts where the consensus prediction differed substantially from single-model baselines. From this review, we observed several patterns that we hypothesize contribute to MAC's performance, as well as a failure mode.

First, **complementary clinical knowledge**: in several reviewed cases, each model contributed complementary pieces of clinical knowledge, such as procedural inferences,

population-level base rates, or disease-specific clinical nuances. Pooling these contributions produced a more accurate consensus than any single model could reach alone. Second, **divergent reasoning frames**: we observed cases where models from different vendors approached the same patient with different reasoning structures. For example, one model might treat clinical severity as a direct predictor of readmission, while another might treat it as a predictor of mortality or hospice transition, which preempt readmission. Third, **supervisor-triggered reconsideration**: the adjudication step itself appeared to help surface overlooked factors. There were cases where all three models, reasoning independently in round 1, anchored on the most prominent clinical signals and left relevant background factors unexamined. Once the supervisor challenged the panel on a specific omission, the doctors revisited it in the next round and revised their curves accordingly. For example, in a patient discharged after an alcohol-related pancreatitis flare, no model noted in round 1 that the discharge medication list mixed inpatient-only items with the opioid and benzodiazepine; once the supervisor raised this ambiguity, the panel reinterpreted those prescriptions as inpatient carryover and revised its early-window risk accordingly. Together, these observations are consistent with the hypothesis that MAC benefits from the interplay of diverse model priors and structured multi-agent discussion.

We also examined cases where MAC performed worse than the best-performing single model to understand potential failure modes. We found cases where a correct minority prediction was discarded in favor of a cohesive but incorrect majority view. In one such case, involving a patient with brain metastases treated by stereotactic radiosurgery, one model leaned toward an early readmission and came closest of the three, but the other models reframed the clinical severity as a predictor of competing mortality or hospice transition and raised the long-horizon survival, underestimating the near-term risk. This observation suggests that multi-agent discussion can suppress accurate but extreme minority predictions when the majority shares a less accurate interpretation of the evidence. Representative cases and analysis are provided in Appendix H.

6. Discussion

Time-to-event prediction is fundamentally distinct from static classification tasks. It requires modeling a probability distribution over time, maintaining internal probability consistency across horizons, and handling right-censored observations, all of which are sidestepped when the problem is reduced to binary classification at a fixed horizon. Yet the majority of prior work applying LLMs to clinical risk prediction adopts this simpler framing. In this study, we benchmarked four frontier LLMs on zero-shot survival curve prediction alongside regression and ordinal classification on ED Revisit and Readmission.

Our first finding is that frontier LLMs can produce largely well-formed survival curves zero-shot. All predicted values lie within $[0, 1]$ across all models, and monotonicity holds on 99–100% of predictions without post-hoc correction. The initial value condition $S(0) = 1$ is model-dependent: Claude and Llama satisfy it exactly on both tasks, as does Gemini on Readmission; Gemini (on ED Revisit) and GPT-5 (on both tasks) instead predict $S(0)$ slightly below 1, reflecting a tendency to express a small residual probability of the event having already occurred rather than treating the index visit as the strict start of the risk window (see Appendix E). Beyond structural validity, the predicted curves show competitive

discrimination and calibration: the best-performing LLMs approach the supervised Random Survival Forest baseline on Readmission (C-index 0.732 vs. 0.741), despite having no access to training data. This suggests that frontier LLMs have internalized meaningful intuitions about both the shape of survival functions and the temporal dynamics of clinical outcomes.

Second, survival curve prediction outperforms regression and ordinal classification as a formulation for time-to-event tasks, despite being a harder problem. This advantage holds even against classification, which benefits from explicit knowledge of the training population's event-time distribution encoded in the bin boundaries. Regression predictions cluster around prototypical round-number durations, while classification predictions under-assign mass to the no-event bin in seven of eight model–task settings. Survival predictions, by contrast, track observed event-free rates across the full time horizon with comparatively modest and consistent bias.

Third, survival-derived predictions are substantially better calibrated than direct binary prompts at every evaluated horizon, with ECE reductions that are statistically significant across all comparisons ($p < 0.001$). Direct binary prompting tends toward overconfidence, particularly at short horizons where base-rate event probabilities are low, likely because the binary framing anchors the model on a single probability without distributional context. Survival-based prediction covers all horizons simultaneously from a single prompt, maintains monotonicity, and achieves competitive discrimination, providing a combination of practical and statistical advantages that binary prompting cannot replicate. Decision curve analysis (Appendix I) confirms that this calibration edge carries clinical weight. On both tasks, survival-derived predictions achieve positive net benefit across a wider range of decision thresholds than direct binary prompting, remaining useful at the low thresholds appropriate for cheap, low-harm interventions such as follow-up calls and extending well past the point where binary prompting ceases to add value over treating everyone or no one (Appendix Figure 6).

Finally, we explored whether structured multi-agent deliberation can improve upon single-model predictions. Using the MAC framework, our top three LLMs acted as doctor agents producing survival curves sequentially, each building on prior reasoning, with a supervisor adjudicating consensus. MAC modestly exceeded the best individual model's point estimate on Readmission, though not significantly. On ED Revisit, it significantly outperformed the weakest model but trailed the best, again without statistical significance. A board-certified physician in internal medicine reviewed cases where MAC substantially shifted predictions and observed three contributing mechanisms: complementary clinical knowledge contributed by different models, divergent reasoning frames that recontextualized clinical severity (e.g., as a competing mortality risk rather than a readmission driver), and supervisor-triggered reconsideration in which the adjudicator's challenge surfaced a signal all three models had left unexamined in round 1. A key failure mode was also identified: when a minority model held a correct but extreme prediction, majority consensus could suppress it, roughly doubling prediction error in at least one reviewed case.

Limitations First, our evaluation treats both cohorts as closed cohort studies, assuming that patients who do not revisit or readmit within the observation window remain event-free. We do not have reliable information on when individual patients leave the hospital system entirely. Additionally, out-of-hospital deaths, which would preclude readmission

and constitute a competing event, are not systematically available in MC-MED and only partially available in MIMIC-IV. For these reasons, we use the standard integrated Brier score without inverse probability of censoring weighting. Second, although all three formulations are evaluated on the same admissible pairs for C-index, their predicted risk scores operate with different structural margins near the censoring boundary: regression and classification predictions of "no event" are anchored at a well-defined boundary that provides a structural separation from event predictions, while survival RMST scores are natively bounded by the time grid and therefore operate with smaller margins near the censoring boundary. Notably, survival prediction achieves the highest C-index despite operating under these stricter conditions. Third, MAC does not consistently outperform the strongest individual model, indicating that supervisor synthesis can discard useful minority signals even when deliberation surfaces complementary reasoning. Future work should explore test-time scaling approaches such as in-context learning with few-shot survival examples or more structured deliberation protocols.

Fourth, MIMIC-IV is widely used and, unlike MC-MED (released in 2025), may well have appeared in the LLMs' pretraining corpora; we cannot rule out prior familiarity with it, and we do not claim immunity from contamination. The prompts contain no record identifiers or absolute dates, so reproducing an outcome would require re-identifying a patient from clinical features alone, which is difficult but not impossible. However, our central findings hold on both MIMIC-IV and the much more recent MC-MED, a less likely contamination candidate, suggesting the conclusions are not an artifact of exposure to any single corpus. We nonetheless flag potential MIMIC-IV contamination as a limitation.

Conclusion We focus on evaluating zero-shot time-to-event predictions of LLMs in this study, originally intrigued by the question of whether LLMs have a sense of time. Although we cannot fully answer this question mechanistically yet, we observe that they make reasonably good predictions of the time it takes a patient to revisit the ED or be readmitted to the hospital, tasks that are highly patient-specific and involve a wide spread of durations. The survival curve formulation proves to be not merely viable but preferable, eliciting better-calibrated, more discriminative, and structurally coherent predictions than simpler alternatives. In this light, what is most striking is not just the accuracy itself but how it arises. The models appear to hold an abstract sense of temporal risk, of what a survival curve is and how it should behave, and to instantiate it on specific patients. A sense of time, operationally, might be less about recalling when events happened than about knowing the shape events take and fitting it to a case.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631, 2019.

Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Emma Chen, Luyang Luo, Fatma Gunturkun, Sraavya Sambara, Rushil Arora, Boyang Tom Jin, Pranav Rajpurkar, and David A Kim. Evaluation of large language models as emergency department revisit predictors. In *Pacific Symposium on Biocomputing (PSB)*, volume 31, pages 130–143, 2026. doi: 10.1142/9789819824755_0010.

Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213, 1968.

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

Cameron Davidson-Pilon. lifelines: survival analysis in Python. *Journal of Open Source Software*, 4(40):1317, 2019.

Changho Han, Dong Won Kim, Songsoo Kim, Seng Chan You, SungA Bae, and Dukyong Yoon. Large-language-model-based 10-year risk prediction of cardiovascular disease: Insight from the UK biobank data. *medRxiv*, 2023. medRxiv 2023.05.22.23289842.

Changho Han, Dong Won Kim, Songsoo Kim, Seng Chan You, Jin Young Park, SungA Bae, and Dukyong Yoon. Evaluation of gpt-4 for 10-year cardiovascular risk prediction: insights from the uk biobank and koges data. *Iscience*, 27(2), 2024.

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

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008.

Vincent Jeanselme, Nikita Agarwal, and Chen Wang. Review of language models for survival analysis. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. *PhysioNet*, July 2024. doi: 10.13026/hxp0-hg59. URL https://doi.org/10.13026/hxp0-hg59. Version 3.0.

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

Devan Kansagara, Honora Englander, Amanda Salanitro, David Kagen, Cecelia Theobald, Michele Freeman, and Sunil Kripalani. Risk prediction models for hospital readmission: A systematic review. *JAMA*, 306(15):1688–1698, 2011.

Aman Kansal, Emma Chen, Boyang Tom Jin, Pranav Rajpurkar, and David A Kim. Mcmed, multimodal clinical monitoring in the emergency department. *Scientific Data*, 12 (1):1094, 2025.

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

Engin Eren Kavak, Ender Cem Erdat, Zeynep Altundağ Derin, et al. Comparison of AI chatbot predicted and real-world survival outcomes in hepatocellular carcinoma. *Scientific Reports*, 15, 2025. doi: 10.1038/s41598-025-06591-9. URL https://www.nature.com/articles/s41598-025-06591-9.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, pages 3146–3154, 2017.

Niklas Kiermeyer, Tim Lenfers, Amin Dada, Julian Friedrich, Sameh Khattab, Eric Knop, Jan Egger, Markus Pauly, Andreas Jung, Grégoire Montavon, et al. Large language models improve cancer survival prediction using real-world clinical notes. *medRxiv*, pages 2025–08, 2025.

Da In Kim, Wei Siang Lai, and Kelly W Zhang. Tabular foundation models can do survival analysis. *arXiv preprint arXiv:2601.22259*, 2026.

Namkee Oh, Won Chul Cha, Jun Hyuk Seo, Seong-Gyu Choi, Jong Man Kim, Chi Ryang Chung, Gee Young Suh, Su Yeon Lee, Dong Kyu Oh, Mi Hyeon Park, Chae-Man Lim, and Ryoung-Eun Ko. ChatGPT predicts in-hospital all-cause mortality for sepsis: In-context learning with the Korean sepsis alliance database. *Healthcare Informatics Research*, 30 (3):266–276, 2024.

Sangjoon Park, Chan Woo Wee, Seo Hee Choi, Kyung Hwan Kim, Jee Suk Chang, Hong In Yoon, Ik Jae Lee, Yong Bae Kim, Jaeho Cho, Ki Chang Keum, et al. Improving mortality prediction after radiotherapy with large language model structuring of large-scale unstructured electronic health records. *Radiotherapy and Oncology*, page 111052, 2025.

Patrick Royston and Mahesh K B Parmar. Restricted mean survival time: An alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Medical Research Methodology*, 13:152, 2013.

Ethan Steinberg, Jason Fries, Yizhe Xu, and Nigam Shah. Motor: A time-to-event foundation model for structured medical records. *arXiv preprint arXiv:2301.03150*, 2023.

Andrew J Vickers and Elena B Elkin. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574, 2006.

Grace Chang Yuan, Xiaoman Zhang, Sung Eun Kim, and Pranav Rajpurkar. Do mixed-vendor multi-agent LLMs improve clinical diagnosis? In *Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026)*, pages 1–18. Association for Computational Linguistics, 2026.

Yun Zhao, Qinghang Hong, Xinlu Zhang, Yu Deng, Yuqing Wang, and Linda Petzold. Bertsurv: Bert-based survival models for predicting outcomes of trauma patients. *arXiv preprint arXiv:2103.10928*, 2021.

Appendix A. Feature Availability

Temporal filtering is enforced as follows: patient-history records (past medical history, home medications) are restricted to those documented before arrival, with home medications restricted to those still active at arrival, and within-visit records (labs, vitals, orders) to those available by departure; any record whose timestamp cannot be verified is excluded rather than assumed valid.

Table 5: Features included and excluded for each prediction task. Checkmarks indicate features available at the prediction time point.

Feature Category	**ED Revisit**	**Readmission**
Demographics	✓	✓
Triage acuity/vitals	✓	–
Chief complaint	✓	–
Admission characteristics	–	✓[a]
Prior utilization	✓	✓
Past medical history	✓	–
Home medications	✓	–
Laboratory results	✓	✓[b]
Diagnosis codes	✓	✓
Procedures	–	✓
ED length of stay	✓	–
Discharge destination	✓	✓

[a] Admission type and source, insurance, length of stay, admit and discharge timing, discharge service and destination, and ICU stay. [b] Last values before discharge.

Table 6: Feature representation comparison between LightGBM and LLM by task. Numbers in parentheses indicate feature dimensionality for LightGBM. Summing the rows applicable to each task gives the full fixed-width design matrix: 309 features for ED Revisit and 324 for Readmission. All top-k vocabularies are selected on the training split only; unmeasured laboratory tests carry an explicit missingness indicator, and missing numeric values are imputed with the training-set median (Section 3).

Feature	**LightGBM**	**LLM**
All Tasks		
Demographics	One-hot encoded (varies by task)	Natural language text
ED Tasks (MC-MED)		
Triage acuity	Numeric ordinal (1-5)	Text ("ESI: 3")
Chief complaint	Multi-hot top-50 (50 dim)	Raw text
Triage vitals	Numeric standardized (6 dim)	Text with values
Prior visits	Numeric count	Text ("Visit #3")
Past medical history	Multi-hot top-50 CCS + count (51 dim)	All conditions as text
Medications (home + ED)	Multi-hot 8 home + 8 ED classes, each with a count (18 dim)	All medications as text
ED Revisit (MC-MED)		
Labs (during visit)	25 labs × 3 features (75 dim)	All labs with values, flags, aggregation
During-visit vitals	Not included	All vitals with aggregation (mean/min/max)
ED diagnosis	Multi-hot top-50 CCS (50 dim)	Full text
Discharge destination	One-hot encoded	Text
Orders	4 imaging-modality flags + 4 order-type counts (8 dim)	All orders as text
Readmission Only (MIMIC-IV)		
LOS, ICU transfer	Numeric	Text
Admission timing	Numeric (admit hour, admit/discharge day-of-week)	Text (e.g., "Admitted on Wednesday at 17:00")
Discharge destination	One-hot encoded	Text
Labs (last 24h)	25 labs × 3 features (75 dim)	All labs with flags, aggregation
Current diagnoses	Multi-hot top-50 CCS + count (51 dim)	All diagnoses as text
Procedures	Multi-hot top-50 CCS + count (51 dim)	All procedures as text
Prior admission history	Numeric (count, days since, mean LOS)	Text description
Discharge medications	Multi-hot 10 classes + count (11 dim)	All medications as text

Appendix B. Baseline Model Hyperparameters

Table 7 summarizes the default hyperparameters used for the three primary baseline models before tuning. The two LightGBM models share the same GBDT backbone settings; the Random Survival Forest is fitted with `scikit-survival`. All baseline models were subsequently tuned using Optuna; see Appendix C for tuning details and search spaces.

Table 7: Hyperparameter configuration for the three baseline models.

Hyperparameter	**LightGBM Regressor**	**LightGBM Classifier**
Objective	`regression_l1` (MAE)	`multiclass`
Metric	`mae`	`multi_logloss`
Boosting type	`gbdt`	`gbdt`
Num. leaves	31	31
Learning rate	0.05	0.05
Feature fraction	0.9	0.9
Bagging fraction	0.8	0.8
Bagging frequency	5	5
Max. estimators	1,000	1,000
Early stopping rounds	100	100
Class weights	–	Inverse frequency

Hyperparameter	**Random Survival Forest**
Implementation	`sksurv.ensemble.RandomSurvivalForest`
Num. trees (`n_estimators`)	100
Min. samples per leaf	15
Features per split (`max_features`)	`sqrt`
Rows per tree (`max_samples`)	0.5 (bootstrap; tuned)
Random seed	42
Early stopping	None

Training details. For time-to-event tasks (ED Revisit, Readmission) the LightGBM Regressor is trained on all patients with an observed event within the window together with a random subsample of patients whose event occurs *after* the window, using their true (uncapped) event time; patients with no observed event are excluded, as they have no regression target. The beyond-window subsample is limited to 0.75× (ED Revisit) and 0.5× (Readmission) the number of in-window events, a ratio selected on the validation-set C-index, so that these long-horizon exemplars supply the low-risk signal needed to rank the censored majority without dominating the training distribution. Training on in-window events alone leaves the regressor unable to distinguish patients who will revisit from those who will not, collapsing its rank discrimination; supplying beyond-window exemplars lets it push low-risk patients past the window boundary, where the C-index treats them as a tied low-risk group. At inference it predicts for all patients. The LightGBM Classifier is trained on all patients and uses a *NoEvent* bin for patients whose event falls outside the observation window; inverse-frequency sample weights are applied so each class contributes equally to the multiclass log-loss. The RSF natively handles right-censoring via the log-rank split criterion and is therefore trained on all patients without filtering.

Appendix C. Hyperparameter Tuning

All baseline models were tuned using Optuna (Akiba et al., 2019) with the Tree-structured Parzen Estimator (TPE) sampler and MedianPruner for early termination of unpromising trials. Each model was tuned for 20 trials, optimizing C-index for survival models, MAE for regression, and log-loss for classification on the validation set.

Table 8 summarizes the search spaces for each model type. For survival models, we tuned Random Survival Forests (RSF), Cox Proportional Hazards (CoxPH), and DeepSurv neural networks. For LightGBM models (regression and classification), we tuned standard GBDT hyperparameters including learning rate, tree structure, and regularization terms.

Table 8: Hyperparameter search spaces for Optuna tuning. All numeric ranges use log-uniform sampling unless otherwise noted.

Model	**Hyperparameter**	**Range**	**Type**
LightGBM	learning_rate	[0.01, 0.3]	log-uniform
	num_leaves	[15, 127]	int
	feature_fraction	[0.5, 1.0]	uniform
	bagging_fraction	[0.5, 1.0]	uniform
	min_child_samples	[5, 100]	int
	n_estimators	[100, 2000]	int
	reg_alpha	$[10^{-4}, 10]$	log-uniform
	reg_lambda	$[10^{-4}, 10]$	log-uniform
RSF	n_estimators	{100, 200, 300}	categorical
	min_samples_leaf	[5, 100]	int
	max_features	{sqrt, log2}	categorical
	max_depth	[5, 30] or None	int or None
	max_samples	{0.5, 0.7, 1.0}	categorical
CoxPH	penalizer	[0.001, 1.0]	log-uniform
	l1_ratio	[0.0, 1.0]	uniform
DeepSurv	learning_rate	$[10^{-5}, 0.1]$	log-uniform
	weight_decay	$[10^{-6}, 0.1]$	log-uniform
	dropout	[0.0, 0.6]	uniform
	hidden_layers	1–4 layers $\times$ {32, 64, 128, 256}	int $\times$ cat

For all models, early stopping was used when applicable: LightGBM models used 50-round early stopping on the validation set; DeepSurv used early stopping with patience of 10 epochs. For CoxPH, each trial is fitted on the full training split, with low-variance features filtered (variance threshold 10^{-4}) and Newton–Raphson iterations capped at 200, matching the final fit. Tuning on the full split matters here because the `lifelines` penalty term is not normalized by sample size: a penalizer selected on a subsample is too weak once the winning configuration is refitted on all training data, which drives the coefficients up until most predicted curves saturate at $\hat{S}(t) = 1$ and discrimination collapses. The best hyperparameters from tuning were used for the final model training on the full training set, with evaluation on the held-out test set.

Appendix D. Prompt Templates

Each patient is presented to the LLM in a single zero-shot user message. The patient context block is identical across the three prediction types; only the Prediction Task instruction and response schema differ. Below we show the ED Revisit template; the Readmission template uses the same structure with different field names (admission type, discharge diagnoses, lab results, etc.) and a 365-day observation window.

Shared patient context (ED Revisit).

```
You are an emergency department physician. Based on the following patient
information at discharge, predict when (if ever) this patient will return
to the ED.

Demographics: Age: ..., Gender: ..., Race: ..., Ethnicity: ..., Payor_class:
...
Visit Details: CC: ..., Means_of_arrival: ..., Dx_ICD10: ..., ED_LOS: ...
ED Length of Stay: ... hours

Triage Acuity (ESI): ...
Triage Vitals: Temp: ..., HR: ..., RR: ..., SpO2: ..., SBP: ..., DBP: ...
Prior ED utilization: visit #... (... prior ED visit(s))
Discharge Destination: ...
Discharge Diagnosis: ...
Past Medical History: ...
Medications: ...

=== LABORATORY RESULTS ===
[all results available by ED departure, with abnormal flags]

=== VITAL SIGNS (During Visit) ===
Diastolic BP: ...    Heart Rate: ...    Respiratory Rate: ...
Systolic BP: ...    Oxygen Saturation: ...    Temperature: ...

Orders: [procedure and medication orders placed during the visit]
```

For Readmission the role is "hospital physician" and the context includes admission type, discharge service, discharge destination, discharge diagnoses, procedures, lab results (last 24 h before discharge), discharge medications, and prior admission history.

Regression prompt (appended after patient context).

```
=== PREDICTION TASK ===
Predict the days until ED revisit (use -1 if you predict no revisit within
60 days).

Return your answer in this exact JSON format wrapped in <json_answer> tags:
<json_answer>
{
  "numeric": <predicted days until revisit as float, or -1 if no revisit
expected>,
  "confidence": "<low/medium/high>",
  "reasoning": "<brief 1-2 sentence explanation>"
}
</json_answer>
```

For Readmission: observation window is 365 days and the instruction reads "Predict the days until readmission (a single number, e.g., 45.0). If you predict the patient will NOT be readmitted within 1 year, use -1."

Classification prompt.

```
=== PREDICTION TASK ===
Classify this patient's time to ED revisit into one of these categories:
"P0-10" (0--1.0 days, 0--10% of population), "P10-20" (1.0--2.1 days, 10--20%),
..., "P90-100" (44.9--60.0 days, 90--100%), "NoEvent_60d" (no event expected)

Choose exactly one:  "P0-10", ..., "P90-100", "NoEvent_60d"
NOTE: Bins are equal percentiles of patients who had the event within 60 days.
'P0-10' is the 10% shortest times; 'P90-100' is the 10% longest.  Choose 'NoEvent_60d'
if no event is expected.

Return your answer in this exact JSON format wrapped in <json_answer> tags:
<json_answer>
{
  "bin":  "<one of:  P0-10 | ...  | P90-100 | NoEvent_60d>",
  "confidence":  "<low/medium/high>",
  "reasoning":  "<brief 1-2 sentence explanation>"
}
</json_answer>
```

For Readmission the 10 bins span 0–365 days (e.g., P0-10 = 0–4.7 days) and the NoEvent label is "NoEvent_365d".

Survival prompt.

```
=== PREDICTION TASK ===
For each time point [0, 1, 2, 3, 7, 14, 30, 60] days, predict the probability
(0-1) that the patient will NOT have revisited by that time (i.e., P(time
to revisit > t)).

Return your answer in this exact JSON format wrapped in <json_answer> tags:
<json_answer>
{
  "survival":  {
    "0":  <P(T > 0)>,
    "1":  <P(T > 1)>,
    "2":  <P(T > 2)>,
    "3":  <P(T > 3)>,
    "7":  <P(T > 7 days, 1 week)>,
    "14":  <P(T > 14 days, 2 weeks)>,
    "30":  <P(T > 30 days, 1 month)>,
    "60":  <P(T > 60 days, 2 months)>
  },
  "confidence":  "<low/medium/high>",
  "reasoning":  "<brief 1-2 sentence explanation>"
}
</json_answer>
```

For Readmission the time points are [0, 7, 14, 30, 60, 90, 180, 365] days.

Appendix E. LLM-Predicted Survival Curves Satisfy Formal Survival Function Conditions

We establish that the zero-shot survival curves predicted by LLMs are mathematically well-formed survival functions. A function $S : \mathbb{R}_{\geq 0} \to [0, 1]$ is a valid survival function if it satisfies four conditions: (C1) *initial value*, $S(0) = 1$; (C2) *monotone non-increasing*, $S(t_1) \geq S(t_2)$ for all $t_1 < t_2$; (C3) *convergence to zero*, $\lim_{t\to\infty} S(t) = 0$; and (C4) *right-continuity*, $\lim_{s\downarrow t} S(s) = S(t)$ for all $t \geq 0$.

Because predictions are made over a finite horizon $\{t_1, \ldots, t_n\}$, we replace (C3) with the finite-horizon analogue (C3′) *boundedness*, $0 \leq S(t_i) \leq 1$ for all i, which is both directly verifiable and sufficient for the discrete setting. Since we prompt LLMs to predict $P(T > t)$ using a strict inequality, the queried values correspond to $S(t) = 1 - F(t)$, where $F(t) = P(T \leq t)$ is the cumulative distribution function. As F is right-continuous by convention, $S(t) = 1 - F(t)$ is likewise right-continuous. When extended to a step function by holding each queried value constant on the half-open interval to the next timepoint, i.e., $S(t) = S(t_i)$ for $t \in [t_i, t_{i+1})$, the piecewise-constant interpolation preserves right-continuity, satisfying (C4) by construction. We therefore verify (C1), (C2), and (C3′) empirically.

Boundedness (C3′). All predicted survival values lie within $[0, 1]$ for every patient across all models and tasks.

Initial value (C1). The condition $S(0) = 1$ is satisfied exactly by Claude and Llama (100% on both tasks). The two remaining models deviate in distinct ways, and for Gemini only on ED Revisit. Gemini satisfies $S(0) = 1$ on 78.8% of ED Revisit predictions and 100% of Readmission predictions; among the 106 non-exact ED Revisit predictions, $S(0)$ takes five discrete values spanning $[0.98, 1)$ (mean = 0.997). GPT-5 satisfies $S(0) = 1$ on only 11.4% of ED Revisit and 60.0% of Readmission predictions; the 443 non-exact ED Revisit predictions spread across 10 distinct values in $[0.985, 1)$ (mean = 0.996). In both models the non-exact values are tightly clustered near 1 (77–95% of them exceed 0.99, with means of 0.990–0.997), with no smooth continuum down to zero; the sharp drop at the >0.999 threshold accounts for nearly all the non-compliance. This pattern suggests a systematic tendency to express a small residual uncertainty at $t = 0$, encoding a non-zero probability that the event has already occurred, rather than treating the index visit as the strict start of the risk window.

A single GPT-5 prediction on Readmission yielded $S(0) = 0.10$, driving the Readmission minimum to 0.10 despite a mean of 0.990 among non-exact cases. Inspection of the raw response reveals coherent but erroneous reasoning: the patient was being discharged directly to an acute hospital, and the model interpreted this as near-certain immediate readmission. This is a case of confident miscalibration on a rare discharge disposition.

Monotonicity (C2). All predicted survival curves are monotonically non-increasing over time for Claude, Gemini, and GPT-5 (100%). Llama exhibits a 0.6% violation rate on ED Revisit (3 of 500 patients) and 0% on Readmission. Inspection of the three violating curves reveals a consistent pattern, where all curves are valid and strictly decreasing through day 14, then rise between days 14–30 and/or 30–60. The affected presentations are uniformly benign and self-limiting, including otitis externa, a rabies-prophylaxis vaccination visit, and an acute cough. This suggests that Llama reasons the acute condition will have resolved

within roughly a month, and encodes this as an *increasing* probability of remaining event-free rather than a flat or slowly declining tail.

Appendix F. Additional Results

Table 9: Performance across prediction formulations for censored tasks. Regression: MAE in days (↓). Classification: Quadratic Weighted Kappa (↑). Survival: Integrated Brier Score (↓). The supervised baseline row combines LightGBM (MAE, QWK) and RSF (IBS). Best result per column is **bolded**. For LLMs the MAE column is *conditional* on the model committing to an event time, since "no event" (−1) predictions are excluded rather than scored; a model that abstains more often can therefore post a lower MAE, and the column should be read together with the recall figures in Table 10.

	ED Revisit			Readmission		
Model	MAE ↓	QWK ↑	IBS ↓	MAE ↓	QWK ↑	IBS ↓
LightGBM/RSF	38.0	0.289	0.107	100.5	**0.461**	**0.168**
Claude Sonnet 4.5	12.6	0.332	0.134	**58.8**	0.240	0.207
Gemini 3.1 Pro	**11.1**	0.319	**0.103**	59.0	0.321	0.193
GPT-5	13.3	0.338	0.108	65.0	0.148	0.192
Llama 4 Maverick	15.0	**0.356**	0.133	63.0	0.181	0.208

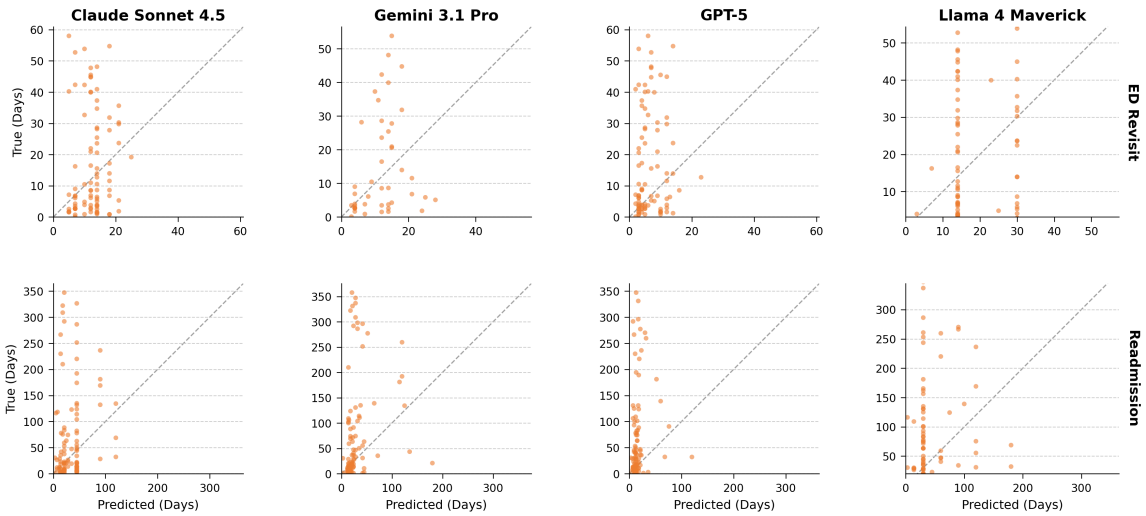

Figure 3: Predicted vs. actual time to ED Revisit (left) and Readmission (right). Only uncensored patients with an observed event within the observation window are shown, and LLM predictions of -1 (i.e., "no event") are excluded. Diagonal dashed line indicates perfect calibration.

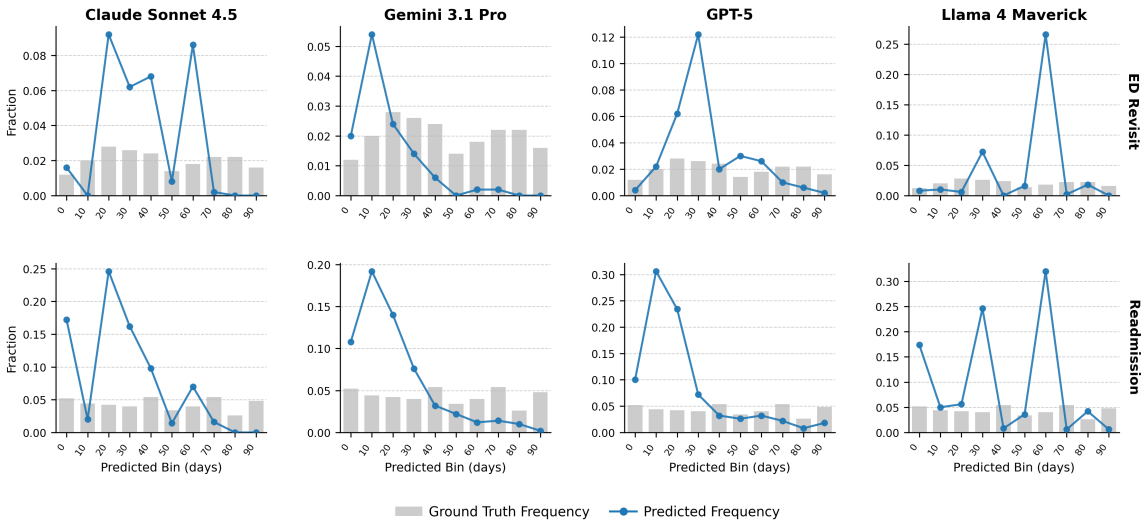

Figure 4: Distributional alignment of the classification formulation across the four LLMs (columns) for ED Revisit (top row) and Readmission (bottom row). Gray bars show the marginal fraction of patients whose true label falls in each event-time bin; the blue line shows each model's marginal predicted-bin frequency. This is an aggregate (marginal) comparison of predicted vs. ground-truth bin frequencies, not a per-patient calibration curve: alignment of the line with the bars indicates the model reproduces the population bin distribution. The NoEvent bin (patients with no event within the observation window) is omitted for readability, as its ground-truth frequency would dominate the axis.

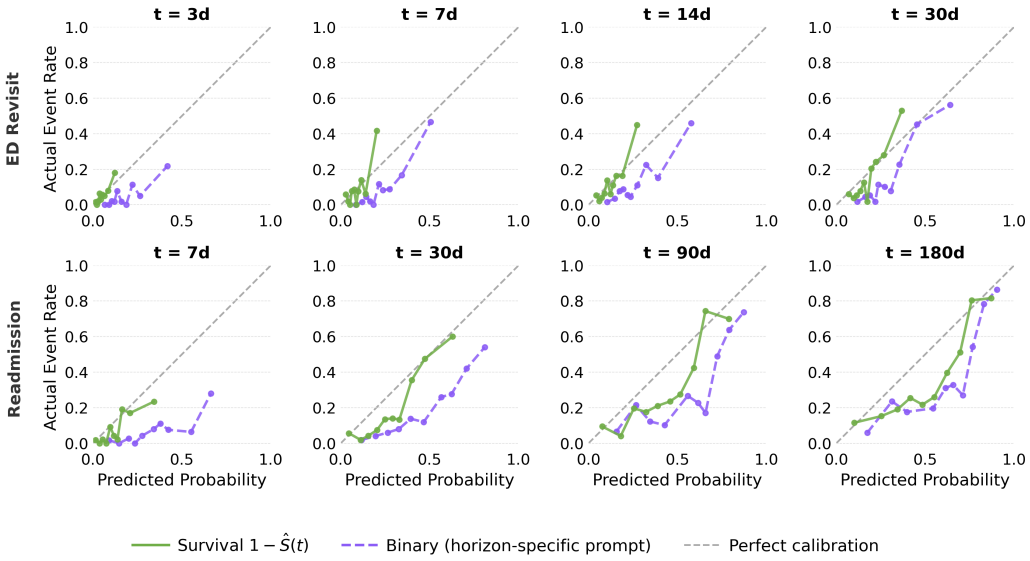

Figure 5: Calibration reliability diagrams comparing direct binary prompting (purple dashed) against survival-derived predictions ($1 - \hat{S}(t)$, green solid) across multiple horizons. Top row: ED Revisit at $t \in \{3, 7, 14, 30\}$ days. Bottom row: Readmission at $t \in \{7, 30, 90, 180\}$ days. Diagonal dashed line indicates perfect calibration.

Appendix G. Event-Detection Summary for the Regression MAE

The regression MAE (Table 9) is computed only over uncensored events for which the model committed to a finite time; "no event" (-1) predictions are excluded rather than scored. A model can therefore lower its MAE by abstaining on hard cases. Table 10 reports the companion event-detection summary with recall over true events and the false negative rate (i.e., 1 - recall), so that abstention is not mistaken for accuracy. The spread is substantial: on ED Revisit, recall ranges from below 0.5 (a model that calls the majority of true events "no event") to near 0.9, which the conditional MAE alone would obscure.

Table 10: Event-detection summary for the regression formulation, reported alongside the (conditional) MAE. Because MAE excludes "no event" (-1) predictions, a model can lower its MAE by abstaining; recall (↑) and the false negative (FN) rate (↓, fraction of true events predicted as "no event") make that abstention visible.

	ED Revisit		**Readmission**	
Model	Recall ↑	FN ↓	Recall ↑	FN ↓
Claude Sonnet 4.5	0.881	0.119	0.945	0.055
Gemini 3.1 Pro	0.455	0.545	0.857	0.143
GPT-5	0.881	0.119	0.982	0.018
Llama 4 Maverick	0.871	0.129	0.982	0.018

Appendix H. Multi-Agent Conversation Case Studies

This appendix presents representative MAC transcripts illustrating the improvement patterns and failure mode discussed in Section 5.3. The qualitative review spanned both the Readmission and ED Revisit 500-case cohorts used for the quantitative MAC comparison; the two cases shown here are drawn from the Readmission cohort and were selected as cases where the consensus prediction differed substantially from the single-model baselines. For each case we provide the patient prompt, the full conversation transcript, and a comparison of single-model versus MAC consensus predictions. The per-case RMSTs reported below integrate each curve as the model wrote it, including its predicted $S(0)$; this differs by at most 0.02 days from the $S(0) = 1$ convention used for the RMST risk score in Section 3, and affects no comparison in this appendix.

H.1. System Prompts

The doctor and supervisor system prompts below were adapted from the MAC framework described in (Yuan et al., 2026) and used across all cases.

Doctor prompt.

```
You are {doctor_name}.  You are a medical expert in clinical outcome prediction.
Your role:  1) Analyze the patient's presentation carefully.
Key responsibilities:
1.  Thoroughly analyze the case information and other specialists' input.
2.  Offer valuable insights based on your specific expertise.
3.  Actively engage in discussion with other specialists, sharing your findings,
thoughts, and deductions.
4.  Provide constructive comments on others' opinions, supporting or challenging
them with reasoned arguments.
5.  Continuously refine your prediction based on the ongoing discussion.
Guidelines:
- Present your analysis clearly and concisely.
- Support your predictions with relevant clinical reasoning.
- Be open to adjusting your view based on compelling arguments from other
specialists.
```

```
- Avoid asking others to copy and paste results; instead, respond to their
ideas directly.
Your goal: Contribute to a comprehensive, collaborative prediction process,
leveraging your unique expertise to reach the most accurate prediction possible.
Every time you speak, include your current prediction at the END of the message,
wrapped in <json_answer> tags exactly as specified in the task instructions.
```

Supervisor prompt.

```
You are the Medical Supervisor. Oversee doctors and drive consensus on the
clinical prediction.
Your role:
1. Oversee and evaluate predictions and reasoning made by medical doctors.
2. Challenge predictions, identifying any critical clinical factors missed.
3. Facilitate discussion between doctors, helping them refine their predictions.
4. Drive consensus among doctors, focusing solely on the prediction task.
Key tasks:
- Identify inconsistencies in reasoning and suggest modifications.
- Even when predictions seem consistent, critically assess if further refinement
is necessary.
- Provide additional clinical insights to enhance prediction accuracy.
- Ensure all doctors' views are completely aligned before concluding the discussion.

Guidelines:
- Promote discussion unless there's absolute consensus.
- Continue dialogue if any disagreement or room for refinement exists.
- Output ''TERMINATE'' only when: 1. All doctors fully agree. 2. No further
discussion is needed. 3. All clinical factors have been considered.
- Don't include the word ''TERMINATE'' in your response unless you want to
terminate the discussion!
Your goal: Ensure comprehensive, accurate prediction through collaborative
expert discussion.
Terminate ONLY when ready to finalize.

Finalization format (mandatory): provide the consensus prediction in <json_answer>
tags exactly as specified in the task instructions, then a line with TERMINATE.
```

H.2. Success Case: Panel Corrects a Shared Misreading of the Discharge Data

Summary. A 54-year-old married male with alcohol-related acute-on-chronic pancreatitis, chronic hepatitis C, and hypokalemia was discharged home after a 2.9-day emergency admission (Medicaid; one prior admission 504 days earlier). The patient was readmitted at 192 days. Round-1 single-agent estimates diverged (RMST 175–200 days); through discussion the panel converged on a consensus of RMST 190 days (error 2.1 days), more accurate than any individual model, after correcting a shared misreading of the discharge data and pooling one model's electrolyte concern (Table 11).

Complementary clinical knowledge. Each model contributed a distinct reading: GPT-5 anchored the alcohol/pancreatitis relapse risk and short length of stay, Gemini added marital support and the long 504-day gap as stabilizing factors, and Claude emphasized that the hypokalemia (K 3.3) and hypomagnesemia (Mg 1.6) were likely still uncorrected at discharge and could drive an early symptomatic return.

Supervisor-triggered reconsideration. No doctor noted in round 1 that the discharge list mixed inpatient-only items (heparin, saline flushes) with the hydromorphone and lorazepam. The supervisor surfaced this discrepancy at the end of round 1, asking whether the opioid and benzodiazepine were genuine home prescriptions or inpatient MAR carryover; in round 2 all three doctors adopted the carryover reading, two of the three revising their early-window survival upward, while retaining Claude's round-1 magnesium penalty. This correction emerged only after the discussion structure prompted it, not in any doctor's independent round-1 reasoning, and the accurate consensus reflects the combination.

Table 11: Success case: prediction comparison. True readmission: 192.2 days.

Method	RMST (days)	Error (days)
MAC	**190.1**	**2.1**
GPT-5	175.1	17.1
Gemini 3.1 Pro	200.1	7.9
Claude Sonnet 4.5	195.3	3.1

Patient prompt.

```
You are a hospital physician. Based on the following patient information at
↪  discharge, predict when (if ever) this patient will be readmitted to the
↪  hospital.

=== PATIENT DEMOGRAPHICS ===
Age: 54 years
Gender: M
Race/Ethnicity: BLACK/AFRICAN AMERICAN
Marital Status: MARRIED
Primary Language: English

=== ADMISSION CHARACTERISTICS ===
Admission Type: EW EMER.
Admission Source: EMERGENCY ROOM
Insurance: Medicaid

=== HOSPITAL STAY ===
Length of Stay: 2.9 days
Discharge Service: MED
Discharge Destination: HOME
Admitted on: Monday at 16:00
Discharged on: Thursday (weekday)

=== DISCHARGE DIAGNOSES ===
Acute pancreatitis, Chronic pancreatitis, Chronic hepatitis C without mention of
↪  hepatic coma, Hypopotassemia, Other and unspecified alcohol dependence,
↪  unspecified, Unspecified gastritis and gastroduodenitis, without mention of
↪  hemorrhage

=== LABORATORY RESULTS (Last 24 Hours Before Discharge) ===
Anion Gap: 11 mEq/L
```

```
Bicarbonate: 25 mEq/L
Calcium, Total: 8.9 mg/dL
Chloride: 102 mEq/L
Creatinine: 0.6 mg/dL
Glucose: 77 mg/dL
Magnesium: 1.6 mg/dL
Phosphate: 3.6 mg/dL
Potassium: 3.3 mEq/L
Sodium: 135 mEq/L
Urea Nitrogen: 6 mg/dL

=== DISCHARGE MEDICATIONS ===
Heparin, Thiamine, HYDROmorphone (Dilaudid), FoLIC Acid, Sodium Chloride 0.9%
↪  Flush, Lorazepam, Ondansetron, Potassium Chloride, Multivitamins

=== PRIOR ADMISSION HISTORY ===
Number of prior admissions: 1.0
Days since last admission: 504
Average prior LOS: 9.0 days
```

Full transcript. [Doctor1_GPT5] Round 1.

This 54-year-old man with alcohol-related acute on chronic pancreatitis and HCV has several readmission risk amplifiers: alcohol dependence (withdrawal/relapse, recurrent pain), chronic pancreatitis (persistent pain, nutrition, recurrent flares), short length of stay (possible early discharge), Medicaid insurance (social determinants), and ongoing need for opioids/benzodiazepines. Early labs are relatively stable but with hypokalemia and borderline hypomagnesemia, which can contribute to symptoms and ER returns if not corrected. Counterbalancing factors include only one prior admission and a long interval since last hospitalization, suggesting not a frequent utilizer. Net expectation is a moderate-to-high 30-day readmission risk (about 25–30%), with continued cumulative risk over 3–12 months due to chronic disease and alcohol use.

If others argue for lower early risk due to stable labs and discharge home, I'd note short LOS and analgesic/benzo use often predict early bounce-backs in pancreatitis. If higher risk is proposed (e.g., ¿35% at 30 days), I'd temper with the long prior interval and absence of severe organ dysfunction. I'm open to adjusting the curve if there is evidence of poor outpatient follow-up access or prior frequent ED use beyond the single admission.

Prediction: $S(0)$=0.995, $S(7)$=0.9, $S(14)$=0.85, $S(30)$=0.72, $S(60)$=0.62, $S(90)$=0.55, $S(180)$=0.42, $S(365)$=0.32. Confidence: medium.

[Doctor2_Gemini] Round 1.

I agree with Doctor1_GPT5's excellent breakdown of the risk amplifiers, particularly the concern regarding the patient being discharged on hydromorphone and lorazepam. This indicates that he still has significant symptom management needs (pain and potential alcohol withdrawal), and the short length of stay (2.9 days) leaves little room for error. The presence of hypokalemia and borderline low magnesium further underscores the physiological toll of his alcohol use and pancreatitis.

However, I would like to offer a slightly more optimistic perspective for the short-to-medium term. The patient is married, which often provides a crucial layer of social support

and stability, aiding in outpatient recovery and medication adherence. Additionally, the fact that his only prior admission was 504 days ago strongly suggests he is not a "frequent flyer." Patients with alcohol dependence and chronic pancreatitis can often have multiple admissions per year; his extended gap suggests he has some capacity to manage his condition or abstain for prolonged periods.

Therefore, while I completely align with Doctor1's view that his long-term cumulative risk is high due to the chronic nature of his diseases, I predict a slightly lower early readmission risk. I will adjust the survival curve to reflect a slightly more stable 30-day and 90-day outlook, while still acknowledging the high likelihood of a flare-up within the year.

Prediction: $S(0)$=0.995, $S(7)$=0.93, $S(14)$=0.87, $S(30)$=0.78, $S(60)$=0.68, $S(90)$=0.6, $S(180)$=0.5, $S(365)$=0.4. Confidence: medium.

[Doctor3_Claude] Round 1.

I appreciate both colleagues' thorough analyses. Let me add my perspective focusing on clinical outcome prediction.

I find myself aligned more closely with Doctor2_Gemini's reasoning, but I want to highlight several additional clinical factors that influence my prediction:

Supporting lower early risk:

- The laboratory values at discharge are remarkably stable - normal anion gap, bicarbonate, glucose, and creatinine. This suggests good metabolic recovery from the acute pancreatitis episode.
- The low creatinine (0.6) and BUN (6) indicate adequate hydration and renal function at discharge, reducing immediate physiological vulnerability.
- The 504-day gap since last admission is particularly significant - this patient demonstrates capacity for extended outpatient stability despite chronic conditions.

Concerning factors I weight heavily:

- The discharge medication list is a red flag: hydromorphone AND lorazepam suggest unresolved pain and withdrawal risk. This combination often predicts early return.
- Hypokalemia (3.3) and low magnesium (1.6) at discharge are problematic - these often aren't fully corrected and can cause symptoms driving ER visits within days.
- Alcohol dependence with acute-on-chronic pancreatitis creates a vicious cycle: pain $\rightarrow$ opioid use $\rightarrow$ potential alcohol relapse $\rightarrow$ recurrent pancreatitis.

My calibration: I predict slightly better outcomes than Doctor2 for days 7-30 (given stable labs and social support), but I'm more pessimistic about the 180-365 day window. The chronic nature of his conditions - particularly untreated hepatitis C, alcohol dependence, and chronic pancreatitis - creates substantial cumulative risk. Without documentation of alcohol treatment referral or HCV therapy plans, the long-term trajectory is concerning.

I would adjust my prediction upward if we learned he has active substance use treatment or strong outpatient gastroenterology follow-up. Conversely, lack of primary care access with Medicaid could worsen outcomes.

Prediction: $S(0)$=0.995, $S(7)$=0.94, $S(14)$=0.89, $S(30)$=0.8, $S(60)$=0.7, $S(90)$=0.62, $S(180)$=0.48, $S(365)$=0.35. Confidence: medium.

[Supervisor] Round 1.
Supervisor assessment of reasoning quality:

1. Identification of relevant risk factors

- Doctor1_GPT5: Strong enumeration of high-yield risks for this case: alcohol dependence, acute on chronic pancreatitis, short LOS, Medicaid, and discharge on opioid/benzodiazepine. Appropriately notes hypokalemia and borderline hypomagnesemia. Also identifies mitigating factors (only one prior admission and a long interval).
- Doctor2_Gemini: Adds social support (married) and emphasizes the long 504-day interval since last admission as a sign of non–frequent-utilizer status. Acknowledges short LOS and residual electrolyte abnormalities.
- Doctor3_Claude: Adds granular lab interpretation (normal AG, bicarb, glucose, renal function) and explicitly flags the combination of hydromorphone plus lorazepam and inadequately corrected K/Mg as early bounce-back risks. Also highlights the absence of documented addiction treatment/HCV care as longer-term risk factors.

Most comprehensive: Doctor3 integrates both short-term physiology and long-term disease trajectory; Doctor1 best emphasized short LOS and social determinants; Doctor2 uniquely highlights marital support.

2. Connection of patient features to predicted probabilities

- Doctor1: Clear linkage of short LOS/medication complexity to higher early risk, tempered by long interval since prior admission; numerical calibration consistent with a moderately high 30-day risk.
- Doctor2: Reasoning trends more optimistic due to marital status and long prior interval, but the causal linkage between those factors and specific numeric probabilities is less explicit.
- Doctor3: Provides the most explicit pathophysiology-based linkage for early risk (electrolytes, symptom control needs) and separates early vs. longer-term risk drivers (lack of addiction/HCV management, chronic pancreatitis), supporting the chosen curve structure.

Best linkage: Doctor3.

3. Missed or underemphasized critical factors

- Potential medication list ambiguity: The "discharge medications" include heparin and sodium chloride flush, which are typically inpatient-only. If hydromorphone and lorazepam are inpatient-only and not true outpatient prescriptions, their weight on early readmission risk would be lower. None of you addressed this discrepancy explicitly.

- Short length of stay: Doctor1 and Doctor2 discussed this; Doctor3 did not explicitly factor LOS into the early risk discussion.
- Electrolyte correction plan: Potassium repletion is listed (KCl), but no magnesium supplementation appears on the discharge list despite Mg 1.6 mg/dL. Doctor3 flagged this; Doctor1 mentioned borderline Mg but did not discuss lack of supplementation plan; Doctor2 acknowledged borderline low Mg but not the absence of a repletion plan.

4. Potential over/under-weighting

- Doctor2 may be over-weighting the protective effect of being married; while social support helps, its effect on 7–30 day readmission in alcohol-related pancreatitis is often modest compared to ongoing pain control, withdrawal, and access barriers with Medicaid.
- Doctor1 may be slightly over-weighting early risk if the opioid/benzodiazepine meds were not actually prescribed for home. However, if they are true outpatient prescriptions, his early risk estimate is defensible.
- Doctor3 provides a balanced take but may underemphasize the short LOS (a known bounce-back risk) in his narrative; still, he appropriately weights the electrolyte abnormalities and med complexity.

Supervisor's provisional stance:

- For 7–90 days, Doctor3's reasoning currently feels strongest because it balances the stable metabolic recovery with the clear red flags (ongoing pain/withdrawal treatment and residual electrolyte disturbances) and separates short-term vs long-term dynamics. His long-term risk (180–365 days) is also appropriately high given chronic pancreatitis, alcohol dependence, and uncertain linkage to addiction/HCV care.
- That said, the authenticity of the outpatient medication list is a pivotal clinical factor that could shift early risk up or down.

Requests for clarification before finalizing:

1. Discharge medication ambiguity: The list includes heparin and saline flush (usually inpatient-only). Please confirm how you are interpreting hydromorphone and lorazepam:

- If these were truly prescribed at discharge for outpatient use vs. if they are artifacts from inpatient MAR included on the discharge summary.
- For each scenario, please provide revised 7- and 30-day survival estimates. Specifically, quantify how much your 7- and 30-day survival would increase if opioids/benzodiazepines were not actually prescribed for home use.

2. Magnesium management: With K 3.3 mEq/L and Mg 1.6 mg/dL, there is oral KCl on the list but no magnesium supplementation. Given the known interplay (hypomagnesemia impairs potassium repletion and increases symptoms/arrhythmia risk), how much additional early readmission risk (in absolute percentage points by day 7 and 30) do you attribute to the lack of magnesium repletion plan? Please specify your adjustment to your curves if magnesium supplementation were prescribed at discharge.

3. Short LOS weighting: Doctor3, please explicitly incorporate the short LOS (2.9 days) relative to typical pancreatitis stays and discuss how that would adjust your 7–30 day survival estimates.

4. Social support: Doctor2, please justify the expected magnitude of impact of being married in this specific context (alcohol-related acute-on-chronic pancreatitis with Medicaid), and indicate by how many percentage points you believe 7- and 30-day survival should be adjusted upward due to marital support compared with an otherwise similar unmarried patient.

Once we resolve these points, I will select the best-supported curve (likely anchored to Doctor3's framework with targeted adjustments) and finalize the survival estimates.

[Doctor1_GPT5] Round 2.
Acknowledging the supervisor's points and colleagues' input:

- Medication list ambiguity: Given the presence of heparin and saline flush (typically inpatient-only), I consider it more likely that hydromorphone and lorazepam listed here reflect inpatient MAR carryover rather than true outpatient prescriptions. I provide two scenarios for early survival:

- If opioids/benzodiazepines were prescribed for home (higher early risk): 7-day survival ≈ 0.90 and 30-day ≈ 0.72 (my original).

- If they were not prescribed for home (more likely): 7-day survival increases by about $+0.03$ and 30-day by about $+0.04$, to ≈ 0.93 and 0.76, respectively.

- Magnesium management: With K 3.3 and Mg 1.6 at discharge, absent a magnesium repletion plan likely adds roughly 1 percentage point absolute risk by day 7 and 2 percentage points by day 30 (i.e., survival lowered by ˜0.01 at 7 days and ˜0.02 at 30 days). If magnesium supplementation were prescribed, I would adjust survival upward by those amounts (e.g., 7-day 0.93 $\rightarrow$ 0.94, 30-day 0.76 $\rightarrow$ 0.78 in the "no outpatient opioids/benzos" scenario).

- Short LOS weighting: A 2.9-day stay for alcohol-related acute-on-chronic pancreatitis is shorter than typical. I attribute an additional absolute 1% decrease in 7-day survival and 2% at 30 days related to potential incomplete symptom optimization and care transitions. My current 7–30-day estimates already include this penalty; without the short LOS, I would expect ˜0.94–0.95 at 7 days and ˜0.78–0.79 at 30 days.

- Social support (commenting on Doctor2's point): Being married is modestly protective in this context but generally small compared with the drivers here (withdrawal, pain control, access). I would credit approximately +1% at 7 days and +2% at 30 days relative to an otherwise similar unmarried patient.

Overall stance and engagement:

- I agree with Doctor3_Claude's physiologic framing and separation of early vs. long-term drivers, and I share his concern about electrolytes and the opioid/benzo signal if truly outpatient. I also maintain that short LOS is a meaningful early risk signal. Doctor2_Gemini's emphasis on marital support is reasonable but likely confers only a small absolute benefit in this specific profile.

Current base-case assumption for the final curve: hydromorphone/lorazepam are inpatient-only (not outpatient prescriptions), magnesium is not supplemented at discharge, and short LOS penalty applies.

Prediction: $S(0)$=0.995, $S(7)$=0.93, $S(14)$=0.87, $S(30)$=0.76, $S(60)$=0.66, $S(90)$=0.58, $S(180)$=0.46, $S(365)$=0.36. Confidence: medium.

[Doctor2_Gemini] Round 2.

I appreciate the Supervisor's sharp observations and Doctor1_GPT5's adjustments. The points raised require a careful recalibration of my initial assessment.

Here are my responses to the Supervisor's requests for clarification:

1) Discharge Medication Ambiguity: I agree with the Supervisor and Doctor1 that the presence of heparin and a sodium chloride flush strongly suggests this list is an artifact of the inpatient Medication Administration Record (MAR) rather than a true outpatient prescription list.

- *If hydromorphone and lorazepam were true outpatient prescriptions:* This would indicate severe ongoing pain and withdrawal risk, leading me to estimate 7-day survival at ˜0.90 and 30-day survival at ˜0.73.

- *If they are MAR artifacts (not prescribed for home):* This implies his pain and withdrawal were adequately controlled prior to discharge. I would adjust my 7-day survival upward by about +0.04 (to 0.94) and my 30-day survival upward by +0.06 (to 0.79). I will use this latter assumption for my final prediction.

2) Magnesium Management: Doctor3_Claude made an excellent catch regarding the uncorrected magnesium (1.6 mg/dL) alongside hypokalemia (3.3 mEq/L). Hypomagnesemia impairs the kidneys' ability to retain potassium, meaning his oral KCl supplementation may be ineffective, potentially leading to arrhythmias, muscle cramping, or generalized weakness that prompts an ED return. I attribute an additional absolute readmission risk of roughly 1.5% by day 7 (survival -0.015) and 2.5% by day 30 (survival -0.025) due to the lack of a magnesium repletion plan. If magnesium were prescribed, I would adjust my survival estimates upward by those exact margins.

3) Social Support (Marital Status): To justify the magnitude of marital support in this specific context: Patients with alcohol dependence and Medicaid often face significant

social determinants of health barriers. A spouse provides a critical immediate safety net—monitoring for severe withdrawal symptoms, encouraging oral intake (vital for pancreatitis recovery), assisting with medication adherence (like the KCl), and potentially deterring an immediate post-discharge alcohol binge. While it cannot override the pathophysiology of his chronic diseases, I estimate being married provides an absolute survival benefit of roughly +1.5% to +2% at 7 days and +3% at 30 days compared to an otherwise identical unmarried patient.

Synthesis and Adjustments: I am adopting the assumption that the opioids/benzos are MAR artifacts, which improves the early outlook. However, I am incorporating Doctor3's physiological penalty for the uncorrected magnesium and Doctor1's penalty for the short LOS (2.9 days). I still maintain that his 504-day gap since his last admission and his marital status offer a modest but meaningful buffer against early readmission.

My revised curve aligns closely with the emerging consensus, balancing the stable metabolic labs and MAR artifact assumption against the uncorrected electrolytes and long-term chronic disease trajectory.

Prediction: $S(0)$=0.995, $S(7)$=0.94, $S(14)$=0.89, $S(30)$=0.79, $S(60)$=0.69, $S(90)$=0.6, $S(180)$=0.47, $S(365)$=0.35. Confidence: medium.

[Doctor3_Claude] Round 2.

Thank you for the detailed supervisor feedback and colleagues' thoughtful responses. Let me address each clarification point systematically:

1) Discharge Medication Ambiguity: I fully agree with Doctor1 and Doctor2 that the presence of heparin and sodium chloride flush strongly suggests this is an inpatient MAR rather than true outpatient prescriptions. This is a critical distinction I should have addressed initially.

- *If hydromorphone/lorazepam were true outpatient prescriptions:* This would signal uncontrolled pain and active withdrawal management, indicating 7-day survival ˜0.91 and 30-day ˜0.75.
- *If MAR artifacts (base case):* Pain and withdrawal adequately controlled before discharge. I would adjust 7-day survival upward by +0.03 (to 0.94) and 30-day by +0.05 (to 0.80).

I adopt the MAR artifact assumption as most likely given the clinical context.

2) Magnesium Management: Doctor2 and the Supervisor are absolutely right to emphasize this. The combination of K 3.3 with Mg 1.6 and **no magnesium supplementation** is a significant oversight. Magnesium is essential for potassium repletion—without it, the oral KCl may be ineffective. This creates risk for:

- Symptomatic hypokalemia (weakness, palpitations)
- Potential arrhythmia
- Early ED return for symptom evaluation

I attribute approximately **1.5% absolute increased readmission risk by day 7** and **2.5% by day 30** due to absent magnesium repletion. If magnesium were prescribed, I would increase survival by +0.015 at 7 days and +0.025 at 30 days.

3) Short LOS Weighting: The Supervisor correctly notes I underemphasized this. A 2.9-day stay for acute-on-chronic pancreatitis is indeed short—typical stays are 4-7 days. This suggests either:

- Rapid clinical improvement (favorable), or
- Early discharge with incomplete symptom resolution (concerning)

Given the electrolyte abnormalities at discharge, I lean toward the latter. Short LOS in pancreatitis is associated with higher bounce-back rates due to:

- Inadequate pain control assessment
- Premature transition to oral intake
- Insufficient time for electrolyte normalization

I attribute an additional **1% decreased survival at 7 days** and **2% at 30 days** specifically to the short LOS. This penalty is now explicitly incorporated in my estimates.

4) Social Support Impact: I agree with Doctor2's reasoning but would calibrate the effect slightly more conservatively. Marital support in alcohol-dependent patients provides:

- Medication adherence monitoring
- Early symptom recognition
- Potential deterrent to immediate relapse

However, in acute-on-chronic pancreatitis with Medicaid, the physiological drivers (pain, electrolytes, withdrawal risk) often dominate. I estimate marital status provides approximately **+1% at 7 days and +2% at 30 days** compared to an unmarried patient—meaningful but modest relative to the disease burden.

Synthesis: My revised prediction incorporates:

- MAR artifact assumption (opioids/benzos not outpatient) → improves early risk
- Uncorrected magnesium penalty → worsens early risk by ˜1.5-2.5%
- Short LOS penalty → worsens early risk by ˜1-2%
- Marital support benefit → improves by ˜1-2%
- Long 504-day interval → supports better medium-term outlook
- Chronic disease burden (untreated HCV, alcohol dependence, chronic pancreatitis) → drives long-term cumulative risk

The net effect balances to a cautiously optimistic early curve with substantial long-term risk.

Prediction: $S(0)$=0.995, $S(7)$=0.94, $S(14)$=0.88, $S(30)$=0.78, $S(60)$=0.68, $S(90)$=0.6, $S(180)$=0.46, $S(365)$=0.34. Confidence: medium.

[Supervisor] Round 2.

TERMINATE

Prediction: $S(0)$=0.995, $S(7)$=0.94, $S(14)$=0.88, $S(30)$=0.78, $S(60)$=0.68, $S(90)$=0.6, $S(180)$=0.46, $S(365)$=0.35. Confidence: medium. TERMINATE.

H.3. Failure Case: Competing-Mortality Frame Overrides a Correct Minority

Summary. A 65-year-old married male with brain metastases from prior lung cancer was treated with stereotactic radiosurgery and discharged home after a 3.3-day admission including an ICU stay. The patient was readmitted at 47 days. Gemini alone leaned toward this early return (round-1 RMST 133 days, error 86), but during discussion the panel reframed the advanced-cancer severity as a competing mortality/hospice risk and raised the long-horizon survival, moving the consensus to RMST 197 days (error 151)—worse than two of the three round-1 predictions and nearly double Gemini's error (Table 12).

Failure mechanism. This case illustrates the failure mode described in Section 5.3: a divergent reasoning frame overrode a more accurate minority view. Gemini framed the clinical severity as a direct driver of early readmission, whereas GPT-5 and the supervisor framed the same severity as a predictor of competing mortality: in advanced metastatic disease, hospice transition or death without readmission would raise the probability of "no readmission" at long horizons. The panel additionally treated dextrose, saline flushes, and heparin as inpatient MAR artifacts, tempering the early-window hazard. Under the cumulative weight of the competing-mortality argument Gemini revised its trajectory from RMST 133 to 196 days, and the supervisor derived the consensus from three already-moderated positions. Because the patient was in fact readmitted early (47 days), the competing-mortality frame was the wrong lens for this case, and the discussion converted the panel's most accurate individual prediction into a substantially less accurate consensus.

Table 12: Failure case: prediction comparison. True readmission: 46.6 days.

Method	RMST (days)	Error (days)
MAC	197.4	150.8
GPT-5	198.3	151.7
Gemini 3.1 Pro	**133.1**	**86.4**
Claude Sonnet 4.5	172.0	125.4

Patient prompt.

```
You are a hospital physician. Based on the following patient information at
↪ discharge, predict when (if ever) this patient will be readmitted to the
↪ hospital.

=== PATIENT DEMOGRAPHICS ===
Age: 65 years
Gender: M
Race/Ethnicity: WHITE - EASTERN EUROPEAN
Marital Status: MARRIED
Primary Language: English

=== ADMISSION CHARACTERISTICS ===
Admission Type: SURGICAL SAME DAY ADMISSION
Admission Source: PHYSICIAN REFERRAL
Insurance: Private
```

```
=== HOSPITAL STAY ===
Length of Stay: 3.3 days
Discharge Service: NSURG
Discharge Destination: HOME
ICU Stay: Yes (patient was in ICU during this admission)
Admitted on: Monday at 07:00
Discharged on: Thursday (weekday)

=== DISCHARGE DIAGNOSES ===
Other ill-defined cerebrovascular disease, Cerebral edema, Radiological procedure
↪  and radiotherapy as the cause of abnormal reaction of patient, or of later
↪  complication, without mention of misadventure at time of procedure, Personal
↪  history of malignant neoplasm of bronchus and lung, Personal history of
↪  antineoplastic chemotherapy, Coronary atherosclerosis of native coronary
↪  artery, Gout, unspecified, Unspecified essential hypertension, Personal
↪  history of irradiation, presenting hazards to health, Personal history of
↪  tobacco use, Arthrodesis status, Personal history of pulmonary embolism

=== PROCEDURES DURING STAY ===
Stereotactic radiosurgery, not otherwise specified, Other immobilization,
↪  pressure, and attention to wound

=== DISCHARGE MEDICATIONS ===
Senna, Bisacodyl, Allopurinol, Heparin, Glucose Gel, OxycoDONE (Immediate Release)
↪  , Morphine Sulfate, Famotidine, HydrALAzine, Metoclopramide, LeVETiracetam,
↪  Gabapentin, Glucagon, Ondansetron, Docusate Sodium, Insulin, MethylPHENIDATE
↪  (Ritalin), Dexamethasone, Sodium Chloride 0.9%  Flush, Dextrose 50%,
↪  Acetaminophen, Acetaminophen-Caff-Butalbital, Atorvastatin, Lorazepam

=== PRIOR ADMISSION HISTORY ===
Number of prior admissions: 3.0
Days since last admission: 148
Average prior LOS: 8.7 days
```

Full transcript. [Doctor1_GPT5] Round 1.

Given the neurosurgical admission for stereotactic radiosurgery complicated by cerebral edema, an ICU stay, and discharge on high-risk medications (steroids, antiepileptics, insulin, opioids), I anticipate a heightened early hazard of readmission due to edema flare, seizures, hyperglycemia, and BP instability. Longer-term risk remains elevated from underlying malignancy history, CAD, prior PE, and multiple prior admissions.

Prediction: $S(0)$=0.995, $S(7)$=0.92, $S(14)$=0.87, $S(30)$=0.79, $S(60)$=0.71, $S(90)$=0.63, $S(180)$=0.5, $S(365)$=0.35. Confidence: medium.

[Doctor2_Gemini] Round 1.

I completely agree with your assessment, Doctor1_GPT5. You correctly deduced that this patient is likely suffering from brain metastases secondary to his history of lung cancer, given the stereotactic radiosurgery, cerebral edema, and the prescription of dexamethasone and levetiracetam.

I would like to add that the complexity of his discharge medication regimen significantly increases his early readmission risk. He is being sent home with insulin, glucose gel,

and glucagon, which strongly suggests he developed severe steroid-induced hyperglycemia. Managing this at home, alongside tapering dexamethasone, administering prophylactic antiepileptics, and managing pain with multiple opioids, places a massive burden on the patient and his family. Errors in this complex regimen or failure to adequately control cerebral edema or blood glucose could easily precipitate a return to the hospital within the first 14 to 30 days.

Furthermore, his history of 3 prior admissions with a prolonged average length of stay (8.7 days) indicates a trajectory of high healthcare utilization, likely stemming from his initial cancer treatments, complications like his prior pulmonary embolism, and progressive disease. The use of methylphenidate also hints at severe fatigue or cognitive slowing, common in patients undergoing cranial irradiation. Given the progressive nature of metastatic lung cancer and the fragility of his current neurological and metabolic state, I believe the risk of readmission is even higher than your estimates, particularly in the medium-to-long term. I am adjusting the non-readmission probabilities downward to reflect this high vulnerability.

Prediction: $S(0)$=0.99, $S(7)$=0.88, $S(14)$=0.8, $S(30)$=0.68, $S(60)$=0.55, $S(90)$=0.45, $S(180)$=0.3, $S(365)$=0.15. Confidence: medium.

[Doctor3_Claude] Round 1.

Thank you both for your thorough analyses. I agree with the core assessment that this patient faces substantial readmission risk, but I'd like to offer some nuanced considerations that may temper the predicted trajectory somewhat.

Key Points Supporting Moderate Risk:

1. **Planned Procedure with Expected Course**: This was a same-day surgical admission for stereotactic radiosurgery - a planned, elective intervention. The cerebral edema noted is an expected complication of radiosurgery, not necessarily an unexpected deterioration. The ICU stay may have been precautionary or protocol-driven rather than reflecting acute instability.

2. **Discharge to Home (Not Facility)**: The clinical team felt confident discharging this patient home rather than to a skilled nursing facility or rehab, suggesting reasonable functional status and adequate home support despite the complex medication regimen.

3. **Time Since Last Admission**: 148 days (nearly 5 months) since the last admission is actually a relatively stable interval, not suggesting imminent decompensation.

4. **Medication Complexity as Protective**: While Doctor2_Gemini correctly identifies the complexity, many medications (famotidine, docusate, senna, bisacodyl, ondansetron, metoclopramide) are supportive/prophylactic, suggesting proactive symptom management that may *prevent* readmission.

Areas of Concern I Share:

- The insulin/glucose management equipment does indicate new-onset steroid-induced hyperglycemia requiring close monitoring

- Levetiracetam suggests either seizure occurrence or high prophylactic concern
- History of PE with apparent heparin prophylaxis indicates ongoing thrombotic risk
- Multiple cardiovascular comorbidities (CAD, hypertension requiring hydralazine)

My Adjusted Assessment:

I believe the early readmission risk (0-14 days) is elevated but not as dramatically as Doctor2_Gemini suggests. The 30-90 day window carries significant risk from potential radiation necrosis, steroid taper complications, or disease progression. However, if the patient navigates the first 90 days successfully, they may achieve relative stability, making the 365-day survival probability somewhat higher.

Prediction: $S(0)$=0.995, $S(7)$=0.9, $S(14)$=0.84, $S(30)$=0.74, $S(60)$=0.63, $S(90)$=0.55, $S(180)$=0.42, $S(365)$=0.28. Confidence: medium.

[Supervisor] Round 1.

Assessment of reasoning quality and key divergences:

1. Most relevant risk factors identified for this specific patient

- Doctor1_GPT5: Correctly flagged the core near-term drivers (stereotactic radiosurgery with cerebral edema; ICU stay; steroids/antiepileptics/insulin/opioids leading to edema flares, seizures, hyperglycemia, BP lability). Appropriately notes longer-term oncologic/cardiovascular/VTE history and prior admissions as persistent risk.
- Doctor2_Gemini: Adds important specificity on medication complexity and its operational burden at home (insulin plus rescue agents, steroid taper, antiepileptics, opioids). Appropriately leverages the prior utilization signal (3 prior admissions, long average LOS) and infers likely brain metastases. This adds nuance beyond Doctor1 and is highly relevant to early-to-midterm readmission risk.
- Doctor3_Claude: Brings in protective/contextual factors often overlooked: elective/planned nature of SRS, discharge to home (suggesting adequate support and function), and the relatively long interval since last admission (148 days). These are important counterbalances that can lower early hazard.

2. Most direct linkage from patient features to predicted probabilities

- Doctor1: Clear, balanced linkage from clinical features to a moderately elevated hazard curve.
- Doctor2: Strong mechanistic linkage (steroid-induced hyperglycemia management, complex polypharmacy) to higher early and medium-term readmission hazard; integrates prior LOS and disease trajectory. However, the long-term numbers (e.g., 365-day non-readmission 0.15) may be too pessimistic given potential hospice/terminal trajectories that can reduce hospital readmissions.

- Doctor3: Thoughtful contextualization, but there is an internal inconsistency: narrative argues that 365-day non-readmission should be "somewhat higher" if the patient stabilizes, yet the provided 365-day value (0.28) is lower than Doctor1's 0.35. Also, framing medication complexity as protective is debatable; while supportive meds can prevent symptoms, complexity generally increases error/adverse-event risk.

3. Missed or underemphasized critical factors

- Potential artifact in the "Discharge Medications" list: items such as D50 IV, sodium chloride flushes, and unfractionated heparin are commonly inpatient orders and not typically sent home. If these are truly discharge-to-home meds, that implies unusually high complexity; if they are inpatient-only orders carried into the list, then early risk from "medication complexity" may be overstated. None of you explicitly addressed this possibility.

- Competing risk of death/hospice: In many readmission frameworks, death without prior readmission reduces the chance of readmission and can paradoxically increase "not readmitted" probabilities. Very advanced disease may not uniformly translate into higher readmissions if goals-of-care shift to outpatient palliation. This nuance was not explicitly considered in the long-term estimates, particularly by Doctor2.

- ICU stay interpretation: Doctor3 posits protocol-driven ICU observation; however, the presence of documented cerebral edema as a discharge diagnosis argues for real clinical instability during the stay. This may raise early hazard more than a purely protocol ICU stay would.

4. Potential over-/under-weighting

- Doctor2 likely over-weights the discharge regimen complexity and progressive disease in the 90–365-day horizon without accounting for hospice/decisions that can reduce readmissions; the 365-day non-readmission of 0.15 may be too low.

- Doctor3 may under-weight the significance of cerebral edema and ICU stay as true risk markers, and over-credits "supportive" medications as protective given the general risk added by regimen complexity. Numerical outputs also contradict the narrative about improved long-term stability.

- Doctor1's weighting appears the most internally consistent and balanced, but could incorporate Doctor2's valid points on prior LOS and regimen complexity by slightly lowering medium-term survival (non-readmission) probabilities.

Before finalizing, targeted clarifications to resolve key disagreements:

- To all doctors: The "Discharge Medications" list includes items that are commonly inpatient-only (e.g., Dextrose 50% IV, sodium chloride flushes, unfractionated heparin). Do you interpret these as true home medications, or as artifact from inpatient MAR entries? Please state explicitly how your early (7–30 day) estimates would change under each interpretation (true home meds vs. inpatient-only artifact).

- To Doctor3_Claude: Please reconcile your narrative with your 365-day figure. You wrote that successful navigation of the first 90 days would make the 365-day non-readmission "somewhat higher," yet your 365-day value (0.28) is lower than Doctor1's (0.35). Which do you endorse, and how would you adjust your curve to align with your stated rationale?

- To Doctor2_Gemini: Please justify your 365-day non-readmission probability of 0.15 in light of potential hospice or non-hospital end-of-life pathways that often reduce readmissions in metastatic lung cancer with brain involvement. Would you revise your long-term estimate if you account for this competing risk?

Once you respond, I will adopt the strongest, internally consistent trajectory—likely Doctor1's curve modestly adjusted by validated points on regimen complexity and prior utilization—so we can converge on a final set of probabilities.

[Doctor1_GPT5] Round 2.
Thank you for the targeted clarifications. My take:

- Discharge medications interpretation: Items like D50 IV, sodium chloride flushes, and unfractionated heparin are far more consistent with inpatient MAR entries than true home prescriptions. I treat them as artifacts unless explicitly documented for home use. This reduces the perceived complexity of the home regimen to high but typical for post-SRS steroid and seizure prophylaxis management.

- Early hazard under each interpretation:

• If artifact (most likely): Early non-readmission remains elevated but less extreme; I would keep 7–30 day survival modestly lower than my original curve to incorporate Doctor2's points on regimen complexity and prior utilization. • If these are truly home meds (unlikely): I would drop early survival further due to very high operational burden and error risk at home.

- ICU stay and cerebral edema: I align with the Supervisor that documented cerebral edema indicates real clinical instability rather than purely protocol ICU observation, sustaining a heightened early hazard.

- Comments on colleagues:

• Doctor2_Gemini: I agree the regimen complexity and prior utilization increase 30–90 day risk. However, in metastatic lung cancer with brain involvement, hospice or goals-of-care shifts often reduce hospital readmissions, so a 365-day non-readmission of 0.15 likely underestimates the competing risk of non-hospital end-of-life. • Doctor3_Claude: The narrative suggesting higher 365-day non-readmission after successful 90 days conflicts with the reported 0.28; aligning the curve upward in the long term would better match that rationale.

Integrated final trajectory (artifact interpretation assumed):

- Early (7–30) modestly lowered from my original due to regimen complexity and prior utilization.

- Medium-term (60–90) further lowered modestly.
- Long-term (365) nudged upward to account for hospice/non-readmission pathways.

If true home meds (D50/flush/heparin) are confirmed:

- Adjust early values lower still: approximately 7: 0.88, 14: 0.83, 30: 0.74; with 60: 0.66, 90: 0.57.

Prediction: $S(0)$=0.995, $S(7)$=0.91, $S(14)$=0.86, $S(30)$=0.77, $S(60)$=0.68, $S(90)$=0.59, $S(180)$=0.5, $S(365)$=0.38. Confidence: medium.

[Doctor2_Gemini] Round 2.

Thank you, Supervisor, for the astute observations and necessary corrections. I appreciate the opportunity to refine my assessment.

First, regarding the "Discharge Medications" list: I agree entirely with you and Doctor1_GPT5 that Dextrose 50%, sodium chloride flushes, and unfractionated heparin are almost certainly artifacts from the inpatient MAR. A true discharge with these medications would typically involve home health nursing and a central line, which is not indicated here (discharge destination is simply "HOME").

- **If these were true home medications**, the early readmission risk would be exceptionally high (as I initially predicted) due to the sheer operational complexity and severe risk of adverse events.
- **Treating them as artifacts (the most likely scenario)**, the early home burden is lower than my initial estimate, though still clinically significant given the combination of insulin, dexamethasone taper, and antiepileptics. Consequently, I am adjusting my early (7–30 day) non-readmission probabilities upward to reflect a more manageable, albeit still high-risk, home situation.

Second, regarding the 365-day estimate and the competing risk of death/hospice: You have highlighted a critical flaw in my previous long-term prediction. For a patient with advanced metastatic lung cancer, progressive brain metastases requiring SRS, and a history of multiple prolonged admissions, life expectancy is unfortunately limited. As the disease progresses, goals of care frequently shift toward outpatient palliative or hospice care. Because death without readmission or a transition to hospice prevents future hospitalizations, this competing risk paradoxically *increases* the probability of "not being readmitted" in the long term. I completely agree with this logic and am significantly revising my 180-day and 365-day estimates upward to account for these non-hospital end-of-life pathways.

I also agree with the Supervisor's and Doctor1's interpretation of the ICU stay and cerebral edema. The documented cerebral edema is a real clinical issue that likely precipitated the ICU monitoring, validating a heightened early risk curve.

I am now closely aligning my trajectory with Doctor1_GPT5's revised estimates, incorporating a slightly more pronounced drop in the 30-90 day window to reflect the historical utilization (3 prior admissions), but raising the 365-day mark to account for competing mortality/hospice risks.

Prediction: $S(0)$=0.995, $S(7)$=0.91, $S(14)$=0.85, $S(30)$=0.75, $S(60)$=0.66, $S(90)$=0.58, $S(180)$=0.49, $S(365)$=0.4. Confidence: medium.

[Doctor3_Claude] Round 2.

Thank you, Supervisor and colleagues, for the thorough discussion. I appreciate the opportunity to reconcile my analysis.

Addressing the Medication List Artifact Issue:

I fully agree with the Supervisor and both colleagues that Dextrose 50% IV, sodium chloride flushes, and unfractionated heparin are almost certainly inpatient MAR artifacts rather than true home medications. A patient discharged home (not to skilled nursing) with these would be highly unusual and would require extensive home health infrastructure not indicated in this case.

Under the artifact interpretation (most likely):

- The home regimen complexity is reduced to a still-significant but more manageable level: dexamethasone taper, levetiracetam, insulin management, pain control, and supportive medications
- This moderately lowers early readmission risk compared to my initial assessment
- I would adjust 7-day to ˜0.91, 14-day to ˜0.85, 30-day to ˜0.75

If these were truly home medications (unlikely):

- Would indicate extraordinary complexity and vulnerability
- Would lower early estimates substantially: 7-day to ˜0.85, 14-day to ˜0.77, 30-day to ˜0.65

Reconciling My Narrative-Number Inconsistency:

The Supervisor correctly identified that my narrative suggested higher 365-day non-readmission probability after successful 90-day navigation, yet my number (0.28) was lower than Doctor1's (0.35). This was indeed inconsistent.

My revised reasoning:

- If a patient with metastatic lung cancer and brain metastases survives 90 days post-SRS without readmission, they've demonstrated relative stability
- However, the Supervisor's point about competing risks (hospice/death without readmission) is crucial and was missing from my initial analysis
- In advanced cancer, as disease progresses, goals often shift to outpatient palliation or hospice, which *reduces* hospital readmissions while increasing mortality
- This competing risk should increase long-term non-readmission probabilities

Addressing the ICU Stay Interpretation:

I concede that the documented cerebral edema as a discharge diagnosis indicates real clinical instability rather than purely protocol-driven observation. This supports higher early hazard than my initial conservative estimate suggested.

Integrated Final Assessment:

Aligning with Doctor1's revised trajectory and Doctor2's corrected long-term estimates, incorporating:

- Artifact interpretation of certain medications (reducing but not eliminating complexity concerns)
- Real clinical instability evidenced by cerebral edema and ICU stay
- Prior utilization pattern (3 admissions, 8.7-day average LOS) increasing 30-90 day risk
- Competing risk of hospice/death increasing long-term non-readmission probabilities

Prediction: $S(0)$=0.995, $S(7)$=0.91, $S(14)$=0.855, $S(30)$=0.76, $S(60)$=0.67, $S(90)$=0.585, $S(180)$=0.495, $S(365)$=0.39. Confidence: medium.

[Supervisor] Round 2.

Adopting Doctor1's balanced trajectory with modestly lower 30–90 day values to reflect prior utilization and regimen complexity, and a slightly higher 365-day value to account for hospice/death as competing risks in advanced metastatic disease.

Prediction: $S(0)$=0.995, $S(7)$=0.910, $S(14)$=0.855, $S(30)$=0.760, $S(60)$=0.670, $S(90)$=0.585, $S(180)$=0.495, $S(365)$=0.400. Confidence: medium. TERMINATE.

Appendix I. Decision Curve Analysis

Beyond discrimination and timing, we evaluate whether survival-derived predictions translate into clinical utility using decision curve analysis (DCA) (Vickers and Elkin, 2006). At a given decision threshold τ, net benefit is defined as:

$$\mathrm{NB}(\tau) = \frac{\mathrm{TP}}{N} - \frac{\mathrm{FP}}{N} \times \frac{\tau}{1-\tau} \tag{6}$$

where TP and FP are the number of true and false positives when intervening on patients with predicted risk $\geq \tau$, and N is the total number of patients. The weighting term $\tau/(1-\tau)$ represents the odds of the threshold, which implicitly encodes the harm-to-benefit ratio of intervention: a clinician who acts at $\tau = 20\%$ is implicitly accepting up to 4 false positives per true positive, reflecting a judgment that the harm of unnecessary intervention is one-fourth the benefit of correctly identifying a patient who will have an event. By evaluating net benefit across all thresholds, DCA characterizes clinical utility for any cost-benefit assumption without requiring explicit specification of costs.

DCA compares model-based decisions against two reference strategies: treat-all (intervene on every patient) and treat-none (intervene on no patient). A model provides clinical value only when its net benefit exceeds *both* reference strategies; outside this range, the clinician should default to treating everyone (at low thresholds) or no one (at high thresholds).

We compare survival-derived event probabilities ($1 - \hat{S}(t)$) against direct binary predictions for both tasks: ED Revisit at 14, 30, and 60 days and Readmission at 30, 90, and 180 days (Figure 6). For each approach and horizon, we identify the threshold range over which the model provides clinical utility, i.e. where net benefit exceeds both treat-all and treat-none.

Across both tasks, survival-derived predictions provide clinical utility over a wider threshold range than direct binary prompting, and in particular remain useful at substantially lower thresholds (Figure 6; exact net-benefit values in Table 13). On ED Revisit, survival is useful from 1.5% (14 days; apart from a brief dip below treat-all over 3–5.5%), 3.5% (30 days), and 5% (60 days) up to the 50% analysis bound, whereas binary only becomes useful from 7.5%, 8.5%, and 10.5% respectively, and at 14 days retains utility only up to 14%, re-emerging just in the noisy tail above 41%; at the prevalence threshold survival's net benefit significantly exceeds binary's at every horizon (ΔNB +0.028 to +0.040, all $p < 0.05$). On Readmission the advantage is more pronounced: survival remains useful across essentially the entire sweep (e.g., up to 48–50% at all three horizons), while direct binary loses clinical utility at moderate thresholds, above 24% at 30 days and 36.5% at 90 days, whereas survival again achieves significantly higher net benefit at the prevalence threshold (ΔNB +0.022 to +0.036, all $p \leq 0.016$). This reflects both the calibration advantage of survival-derived probabilities and the structural benefit of the survival formulation. By producing a coherent probability distribution across time, survival predictions retain net benefit in the low-threshold regime where binary predictions' overconfidence would otherwise recommend intervention for nearly all patients. The one region where binary is meaningfully not dominated is the high-threshold tail of ED Revisit, where direct binary net benefit edges back above survival, from roughly 32% at 30 days and 38% at 60 days (top row of Figure 6: about 0.05 vs. 0.02 at 60 days), though curves are noisy there because

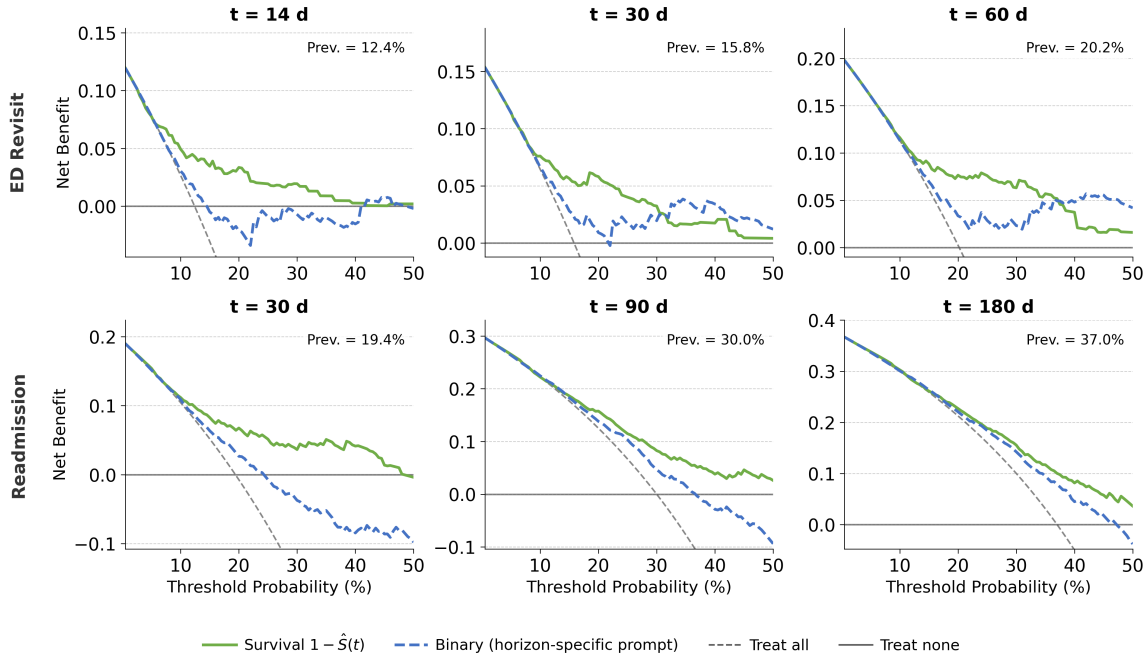

Figure 6: Decision curve analysis for ED Revisit (top row; 14, 30, 60 days) and Readmission (bottom row; 30, 90, 180 days), GPT-5, $n = 500$. Each panel plots net benefit against decision threshold for survival-derived predictions ($1 - \hat{S}(t)$, green solid), direct binary predictions (blue dashed), treat-all (gray dashed), and treat-none (gray solid at zero). A strategy provides clinical utility where its curve exceeds both reference strategies; panel annotations give the event prevalence at each horizon.

very few patients exceed such thresholds. On Readmission survival remains at or above binary throughout the sweep, apart from narrow low-threshold bands (roughly 8.5–16% at 90 and 180 days, and a single grid point at 30 days) where binary exceeds survival by less than 0.005.

Table 13 reports the exact net-benefit values underlying Figure 6 for both tasks: ED Revisit at 14/30/60 days and Readmission at 30/90/180 days. For each horizon we give the threshold interval(s) over which each strategy is clinically useful (net benefit exceeds both treat-all and treat-none), the net benefit at the prevalence threshold, and the paired survival-minus-binary net-benefit difference at that threshold.

Table 13: Decision curve analysis for ED Revisit and Readmission (GPT-5, $n = 500$): net benefit of survival-derived predictions ($1 - \hat{S}(t)$) versus direct binary prompting across horizons. *Useful range* lists the threshold interval(s) over which a strategy's net benefit exceeds *both* treat-all and treat-none; isolated dips of at most 1.5 percentage points are bridged as grid noise, and longer interruptions are shown as separate intervals. *NB at prevalence* is the net benefit evaluated at the prevalence threshold, where both reference strategies have net benefit 0. ΔNB (Survival − Binary) is evaluated at the prevalence threshold with a 95% CI from a paired bootstrap ($N = 1{,}000$ resamples).

		Useful range (%)		NB at prevalence			
Horizon	Prev. (%)	Survival	Binary	Survival	Binary	ΔNB [95% CI]	p
ED Revisit							
14 d	12.4	1.5–2.5, 6–50	7.5–14, 41–48.5	0.043	0.015	+0.028 [0.011, 0.044]	0.004
30 d	15.8	3.5–50	8.5–50	0.053	0.023	+0.031 [0.013, 0.046]	<0.001
60 d	20.2	5–50	10.5–50	0.075	0.035	+0.040 [0.021, 0.057]	<0.001
Readmission							
30 d	19.4	1.5–48	5.5–24	0.065	0.033	+0.032 [0.020, 0.045]	<0.001
90 d	30.0	2–50	4.5–36.5	0.083	0.046	+0.036 [0.020, 0.051]	<0.001
180 d	37.0	3–50	6.5–47	0.100	0.078	+0.022 [0.003, 0.038]	0.016

Appendix J. Sampling Sensitivity

To assess the robustness of LLM predictions to decoding parameters, we conducted a sampling sensitivity analysis on a fixed 150-patient subset of the Readmission task.

Temperature sweep. For Claude Sonnet 4.5, Gemini 3.1 Pro, and Llama 4 Maverick, we evaluated all three prediction formulations at temperatures $T \in \{0, 0.5, 1.0\}$. Table 14 shows that the survival formulation consistently achieves the highest C-index across all temperatures and models. All three formulations are robust to decoding temperature: within each model, C-index varies by at most 0.023 for survival, 0.022 for classification, and 0.039 for regression (the largest swing, for Llama 4 Maverick).

Repeat-run stability. GPT-5 does not expose a temperature parameter, so we instead evaluated three independent runs at default decoding settings. The survival formulation exhibited the lowest variance across runs (SD = 0.003), compared to classification (SD = 0.013) and regression (SD = 0.022). This suggests that the survival formulation not only achieves better discrimination but also yields more stable predictions under stochastic sampling.

Table 14: Sampling sensitivity on the fixed 150-patient Readmission subset. Panel A: C-index across decoding temperatures $T \in \{0, 0.5, 1.0\}$ for Claude, Gemini, and Llama. Panel B: GPT-5 repeat-run stability (mean ± SD over three independent runs at default decoding).

Panel A: Temperature sweep (C-index)				
Model	Formulation	T=0	T=0.5	T=1.0
Claude Sonnet 4.5	Regression	0.673	0.679	0.658
	Classification	0.683	0.690	0.680
	Survival	0.738	0.736	0.731
Gemini 3.1 Pro	Regression	0.693	0.692	0.704
	Classification	0.680	0.674	0.696
	Survival	0.732	0.741	0.739
Llama 4 Maverick	Regression	0.615	0.576	0.596
	Classification	0.654	0.637	0.646
	Survival	0.736	0.713	0.736

Panel B: GPT-5 repeat runs (default decoding)				
Formulation	rep1	rep2	rep3	mean ± SD
Regression	0.604	0.604	0.565	0.591 ± 0.022
Classification	0.623	0.622	0.600	0.615 ± 0.013
Survival	0.733	0.730	0.736	0.733 ± 0.003

Appendix K. Subgroup Analysis for Survival Predictions

We evaluated discrimination performance across demographic subgroups defined by age, sex, and race (Table 15). Age was categorized into three groups: 18–44, 45–64, and 65+ years. To ensure adequate sample sizes for statistical testing, race was aggregated into White versus Non-White categories.

We performed bootstrap hypothesis tests ($B = 1{,}000$) comparing C-index between key subgroups (Table 16). Across all comparisons (age, sex, and race) no statistically significant differences were observed at $\alpha = 0.05$. The largest observed difference was between patients aged 65+ versus <65 for ED Revisit (Llama: $\Delta = -0.101$, $p = 0.070$); the largest sex difference was between male and female patients for ED Revisit (Llama: $\Delta = +0.082$, $p = 0.114$).

In particular, for the Readmission task no significant racial disparity was observed for any model (White vs. Non-White: $|\Delta| \leq 0.017$ and $p \geq 0.69$ for all models), and no significant difference was detected for ED Revisit either ($|\Delta| \leq 0.050$, all $p \geq 0.36$). Discrimination is thus stable across demographic groups defined by age, sex, and race.

Table 15: C-index for survival predictions stratified by demographic subgroups. Results shown for subgroups with $n \geq 15$ patients.

		ED Revisit				**Readmission**			
Category	**Subgroup**	Claude	Gemini	GPT-5	Llama	Claude	Gemini	GPT-5	Llama
Overall	All	0.77	0.80	0.76	0.71	0.73	0.73	0.73	0.70
Age	18-44	0.77	0.82	0.75	0.73	0.76	0.77	0.76	0.74
	45-64	0.77	0.79	0.73	0.75	0.72	0.72	0.72	0.71
	65+	0.71	0.73	0.72	0.63	0.70	0.71	0.70	0.67
Sex	Male	0.80	0.84	0.76	0.76	0.69	0.70	0.71	0.68
	Female	0.74	0.77	0.75	0.68	0.76	0.76	0.75	0.73
Race	White	0.74	0.78	0.73	0.72	0.73	0.74	0.74	0.71
	Non-White	0.78	0.81	0.78	0.72	0.71	0.72	0.72	0.70

Table 16: Statistical comparison of C-index between demographic subgroups. Δ = difference in C-index (Group 1 − Group 2). P-values from bootstrap test ($B = 1000$). Bold indicates $p < 0.05$.

Task	**Model**	**Comparison**	**Δ**	**95% CI**	**P-value**
ED Revisit	Claude	White vs Non-White	-0.040	[-0.145, +0.057]	0.412
	Claude	Male vs Female	+0.057	[-0.040, +0.154]	0.242
	Claude	65+ vs <65	-0.060	[-0.164, +0.033]	0.254
	Gemini	White vs Non-White	-0.035	[-0.126, +0.051]	0.436
	Gemini	Male vs Female	+0.069	[-0.013, +0.160]	0.092
	Gemini	65+ vs <65	-0.084	[-0.179, +0.009]	0.080
	GPT-5	White vs Non-White	-0.050	[-0.149, +0.058]	0.364
	GPT-5	Male vs Female	+0.011	[-0.085, +0.112]	0.794
	GPT-5	65+ vs <65	-0.024	[-0.127, +0.076]	0.614
	Llama	White vs Non-White	-0.001	[-0.113, +0.103]	0.986
	Llama	Male vs Female	+0.082	[-0.019, +0.179]	0.114
	Llama	65+ vs <65	-0.101	[-0.213, +0.007]	0.070
Readmission	Claude	White vs Non-White	+0.017	[-0.055, +0.085]	0.694
	Claude	Male vs Female	-0.063	[-0.126, +0.004]	0.068
	Claude	65+ vs <65	-0.042	[-0.114, +0.027]	0.232
	Gemini	White vs Non-White	+0.015	[-0.057, +0.086]	0.730
	Gemini	Male vs Female	-0.052	[-0.117, +0.014]	0.104
	Gemini	65+ vs <65	-0.039	[-0.109, +0.034]	0.280
	GPT-5	White vs Non-White	+0.016	[-0.059, +0.086]	0.724
	GPT-5	Male vs Female	-0.039	[-0.100, +0.024]	0.252
	GPT-5	65+ vs <65	-0.046	[-0.114, +0.023]	0.190
	Llama	White vs Non-White	+0.005	[-0.063, +0.072]	0.898
	Llama	Male vs Female	-0.045	[-0.110, +0.022]	0.208
	Llama	65+ vs <65	-0.060	[-0.129, +0.013]	0.092

Appendix L. Additional Survival Baselines

In addition to Random Survival Forests (RSF), we evaluated two additional survival analysis baselines: Cox Proportional Hazards regression (CoxPH) and DeepSurv neural networks.

Cox Proportional Hazards (CoxPH). We used the lifelines implementation (Davidson-Pilon, 2019) with elastic net regularization and low-variance feature filtering. The penalizer and L1 ratio were selected on the full training set, so that the regularization strength chosen during tuning is the one the final fit receives (Appendix C). CoxPH assumes proportional hazards: $h(t|X) = h_0(t)\exp(\beta^T X)$, where $h_0(t)$ is the baseline hazard and β are the regression coefficients. Survival curves are computed as $S(t|X) = S_0(t)^{\exp(\beta^T X)}$ where $S_0(t)$ is estimated via the Breslow estimator.

DeepSurv. We implemented DeepSurv (Katzman et al., 2018) using PyTorch (via the `pycox` library), with fully connected layers, batch normalization, dropout, and ReLU activations. The network minimizes the negative partial log-likelihood of the Cox model. We tuned the number of hidden layers (1–4), layer size (32–256), dropout (0–0.6), learning rate, and weight decay. Training used the AdamW optimizer with early stopping (patience = 10 epochs) on validation loss.

Results. Table 17 compares RSF, CoxPH, and DeepSurv on the same 500-patient test cohorts used for the LLM evaluations in Table 2. On ED Revisit, RSF achieved the highest C-index (0.773), followed by DeepSurv (0.748) and CoxPH (0.744). On Readmission, DeepSurv achieved the highest C-index (0.745), followed closely by RSF (0.741) and CoxPH (0.717). Overall, RSF is either the best or within 0.005 of the best on both tasks once these additional models are included: it is highest on ED Revisit (0.773 vs. 0.748 for DeepSurv) and within 0.005 of DeepSurv on Readmission (0.741 vs. 0.745). We therefore use RSF as the primary survival baseline in the main text.

Table 17: Survival baseline comparison on the same 500-patient test cohorts used for the LLM evaluations (C-index; n=500 per task). RSF achieves competitive or superior performance across tasks.

Model	ED Revisit	Readmission
RSF	**0.773**	0.741
CoxPH	0.744	0.717
DeepSurv	0.748	**0.745**

Appendix M. Soft PMF Formulation

We evaluated an alternative "soft PMF" (probability mass function) formulation for LLM time-to-event prediction. Rather than prompting for survival probabilities $S(t) = P(T > t)$ at discrete timepoints, soft PMF asks the model to output the probability that the event occurs in each time bin, producing a probability distribution over time.

Conversion to survival curves. Given PMF probabilities $p_1, \ldots, p_K$ for K bins with boundaries $t_0 < t_1 < \cdots < t_K$, plus p_{K+1} for "no event within observation window" (the $K+1$ probabilities are renormalized to sum to one when the model's output does not), the survival function is computed as:

$$\hat{S}(t_k) = 1 - \sum_{i=1}^{k} p_i = \sum_{i=k+1}^{K+1} p_i \tag{7}$$

Equation (7) defines $\hat{S}$ only at the bin boundaries t_k. When an IBS query timepoint falls strictly inside a bin (between boundaries t_{k-1} and t_k), we evaluate $\hat{S}(t)$ by linearly interpolating the cumulative event mass across that bin; beyond the final boundary t_K, $\hat{S}(t)$ equals the "no event" mass p_{K+1}. This conversion allows direct comparison of IBS and other survival metrics between the two formulations.

Risk score. For C-index computation, we use the expected time-to-event as the risk score:

$$\hat{T} = \sum_{i=1}^{K} p_i \cdot m_i + p_{K+1} \cdot T_{\max} \tag{8}$$

where m_i is the midpoint of bin i and $T_{\max}$ is the observation window boundary.

Results. Table 18 shows C-index results for all four formulations. Soft PMF achieves discrimination comparable to the survival formulation across both tasks: for ED Revisit, soft PMF ranges from 0.736–0.794 versus 0.713–0.802 for survival; for Readmission, 0.663–0.732 versus 0.703–0.732 for survival. The differences between soft PMF and survival C-index are small (typically within 0.02) and largely fall within overlapping confidence intervals; the one exception is Llama 4 Maverick on Readmission, where soft PMF is significantly lower (0.663 vs. 0.703, $p < 0.05$). Overall, neither formulation has a clear advantage for discrimination.

However, the calibration advantage of survival predictions remains robust. Table 19 reveals a substantial calibration gap, with soft PMF yielding IBS values up to 1.8× those of survival. The gap is consistent on Readmission (1.1–1.7× across models) and pronounced for Claude and Llama on ED Revisit (1.7–1.8×), whereas for Gemini and GPT-5 on ED Revisit the two formulations achieve comparable IBS. The PMF formulation elicits incremental per-bin masses, so miscalibration in any single bin accumulates into the cumulative distribution at all later timepoints, whereas the survival formulation elicits the cumulative probabilities $S(t)$ directly, and the models appear better calibrated on cumulative than on incremental probabilities. These results suggest that while soft PMF offers comparable ranking ability, the survival formulation should be preferred when calibrated probability estimates are required for clinical decision-making.

Table 18: C-index comparison across prediction formulations including Soft PMF. 95% CIs from paired bootstrap ($B = 1000$). Significance markers denote a different comparison in each of the two right-hand columns: in the *Survival* column they test survival vs. regression, and in the *Soft PMF* column they test soft PMF vs. survival. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Task	Model	Regression	Classification	Survival	Soft PMF
Readmission	Claude Sonnet 4.5	0.671 (0.640–0.704)	0.676 (0.642–0.711)	0.725 (0.691–0.759)***	0.706 (0.673–0.740)
	Gemini 3.1 Pro	0.682 (0.647–0.718)	0.677 (0.641–0.713)	0.732 (0.700–0.766)***	0.732 (0.701–0.764)
	GPT-5	0.572 (0.535–0.610)	0.609 (0.573–0.647)	0.729 (0.696–0.764)***	0.730 (0.700–0.763)
	Llama 4 Maverick	0.595 (0.564–0.630)	0.632 (0.596–0.670)	0.703 (0.668–0.739)***	0.663 (0.625–0.698)*
ED Revisit	Claude Sonnet 4.5	0.639 (0.594–0.686)	0.712 (0.669–0.760)	0.766 (0.719–0.812)***	0.755 (0.709–0.802)
	Gemini 3.1 Pro	0.675 (0.628–0.725)	0.631 (0.588–0.676)	0.802 (0.761–0.843)***	0.794 (0.742–0.841)
	GPT-5	0.685 (0.636–0.733)	0.713 (0.668–0.764)	0.759 (0.708–0.805)***	0.773 (0.727–0.817)
	Llama 4 Maverick	0.693 (0.644–0.739)	0.721 (0.677–0.769)	0.713 (0.665–0.757)	0.736 (0.687–0.785)

Table 19: Integrated Brier Score (IBS) comparison: Survival vs. Soft PMF. Lower is better. 95% CIs from bootstrap ($B = 1000$).

Task	Model	Survival IBS	Soft-PMF IBS
Readmission	Claude Sonnet 4.5	0.207 (0.193–0.221)	0.357 (0.332–0.382)
	Gemini 3.1 Pro	0.193 (0.175–0.211)	0.243 (0.221–0.265)
	GPT-5	0.192 (0.179–0.206)	0.218 (0.202–0.234)
	Llama 4 Maverick	0.208 (0.195–0.222)	0.283 (0.268–0.299)
ED Revisit	Claude Sonnet 4.5	0.134 (0.125–0.144)	0.241 (0.226–0.255)
	Gemini 3.1 Pro	0.103 (0.084–0.122)	0.099 (0.084–0.115)
	GPT-5	0.108 (0.093–0.125)	0.111 (0.099–0.123)
	Llama 4 Maverick	0.133 (0.122–0.145)	0.230 (0.220–0.240)