

Tree-of-Evidence: Efficient "System 2" Search for Faithful Multimodal Grounding

Micky C. Nnamdi[♠], Benoit L. Marteau[♠], Yishan Zhong[♠], J. Ben Tamo[♠],
and May D. Wang^{†♠}

[♠]Georgia Institute of Technology

[†]Corresponding author: maywang@gatech.edu

Abstract

Large Multimodal Models (LMMs) achieve state-of-the-art performance in high-stakes domains like healthcare, yet their reasoning remains opaque. Attention and saliency-based methods often fail to faithfully represent the model’s decision process, particularly when integrating heterogeneous modalities. We introduce Tree-of-Evidence (ToE), an inference-time search algorithm that frames interpretability as a discrete optimization problem. Rather than relying on soft attention weights, ToE employs lightweight Evidence Bottlenecks that score coarse units of data (e.g., vital-sign windows, report chunks) and performs a beam search to identify the compact evidence set required to reproduce the model’s prediction. We evaluate ToE across six tasks spanning three datasets and two domains, including clinical prediction on MIMIC-IV, cross-center validation on eICU, and non-clinical fault detection on LEMMA-RCA. ToE retains over 98% of full-model AUROC with as few as five evidence units, achieves higher decision agreement and lower fidelity error than LIME, SHAP, saliency, and concept-bottleneck baselines under sparse budgets, and outperforms LLMs up to 70B parameters. ToE therefore provides a practical mechanism for auditing multimodal models.

1 Introduction

Multimodal predictors, such as Large Multimodal Models (LMMs), have achieved remarkable performance by fusing heterogeneous data streams, including text, time series, and imaging, into unified representations (Chen et al., 2024; Huang et al., 2024; Tu et al., 2024). However, as these models grow in complexity, their decision-making processes become increasingly opaque (Rudin, 2019; Wornow et al., 2023). In high-stakes domains like healthcare, "black box" accuracy is insufficient; deployment requires auditable reasoning where a

model’s prediction explicitly traces back to specific, verifiable pieces of evidence (Rudin, 2019).

Current interpretability methods often fail to meet this standard. Attention-based heatmaps are frequently unfaithful to the actual logic of the model (Wiegrefe and Pinter, 2019; Jain and Wallace, 2019), while post-hoc explanation methods provide approximations rather than guarantees (Rudin, 2019). Concept Bottleneck Models (CBMs) offer a step forward by aligning hidden states with human-interpretable concepts (Koh et al., 2020; Vandenhirtz et al., 2024). Yet, CBMs typically require predefined concept annotations and remain static during inference, failing to adaptively search for evidence when data is ambiguous or synergistic. Rationale extraction methods aim to solve this by selecting a subset of input features that are sufficient for the prediction (Lei et al., 2016; DeYoung et al., 2020). Yet, existing rationale methods are typically limited to single modalities, mainly text, and rely on greedy selection strategies that fail to capture the synergistic dependencies between different data types (Xu et al., 2024). For instance, a radiology note of pulmonary edema can reframe a drop in SpO_2 as volume overload, a cross-modal link that unimodal methods miss.

To bridge this gap, we introduce **Tree-of-Evidence (ToE)**, an inference-time search algorithm for multimodal grounding. Inspired by deliberative branching procedures like Tree-of-Thoughts (Yao et al., 2023), ToE treats interpretability as a *discrete search problem* over meaningful evidence units. We use "System 2" approach to denote this multi-step, deliberative search process, in which the algorithm explicitly evaluates and scores candidate evidence combinations via beam search, in contrast to "System 1" single-pass heuristics such as greedy top- k ranking by individual unit scores. Crucially, ToE can handle two types of multimodal data: (1) Global Context (e.g., baseline pathology from Chest X-Ray

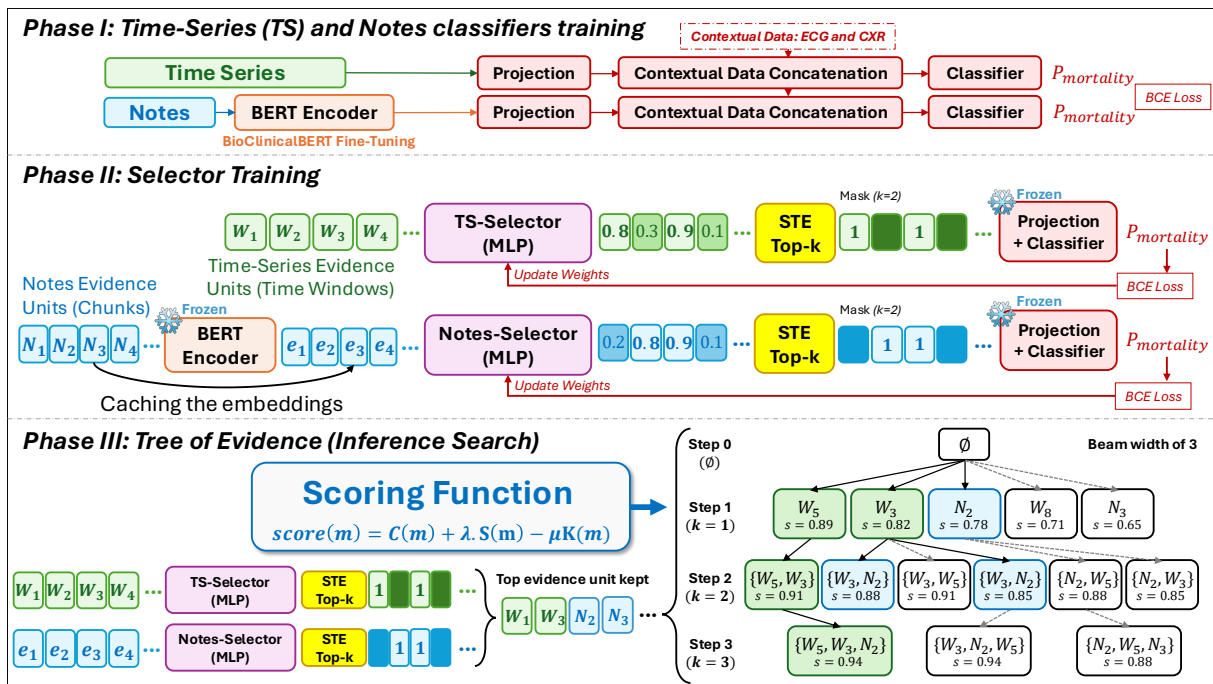


Figure 1: **Overview of the Tree-of-Evidence (ToE) Framework.** **Phase I:** Modality-specific classifiers are trained independently, with BioClinicalBERT (Alsentzer et al., 2019) encoding notes and contextual data (CXR/ECG) concatenated as fixed priors. **Phase II:** Lightweight MLP selectors learn to score evidence units using Straight-Through Estimator (STE) top- k masking with frozen encoders. **Phase III:** At inference, beam search iteratively constructs compact evidence set by optimizing the scoring function, balancing decision agreement, probability stability, and sparsity.

(CXR)/Electrocardiogram (ECG), which serves as a fixed prior, and (2) Searchable Evidence (e.g., dynamic Vitals and Notes), which is actively selected. Instead of relying on soft attention weights, we first train lightweight Evidence Bottlenecks (EB) that score coarse units of data—hourly windows of Intensive Care Unit (ICU) time-series and radiology report chunks. At inference time, ToE performs a beam search to construct a compact evidence set that preserves the full-input decision, explicitly trading off (i) agreement with the original prediction, (ii) stability of the predicted probability, and (iii) evidence sparsity. This separation allows the search to focus on "what changed" (dynamic evidence) while remaining grounded in "who the patient is" (global context). The result is an auditable trace of how evidence is accumulated to justify a decision.

We evaluate ToE across six tasks spanning three datasets and two domains: four clinical prediction tasks on Medical Information Mart for Intensive Care (MIMIC)-IV (Johnson et al., 2024, 2023; Goldberger et al., 2000; Gow et al., 2023), cross-center validation on eICU (208 hospitals) (Pollard et al., 2018), and non-clinical fault detection on

LEMMA-Root cause analysis (RCA) (Zheng et al., 2024). Our experiments demonstrate that ToE yields discrete rationales that (i) remain sufficient for the model’s decision under strict evidence budgets, (ii) exhibit strong decision agreement with the full-input prediction, and (iii) provide an auditable trace of the search process that can be inspected by domain experts. Our contribution can be summarized as:

1. **Model-faithful multimodal grounding via inference-time search.** We formulate grounding as selecting a compact multimodal evidence set that reproduces the full-input model’s decision and confidence, and we propose Tree-of-Evidence (ToE) to solve this with an auditable search trace.
2. **Bottleneck-guided discrete evidence units.** We develop lightweight *Evidence Bottlenecks* that score clinically meaningful, coarse-grained units (hourly windows; report chunks) and provide efficient heuristics for search, while incorporating CXR/ECG signals as *context-only* features rather than searchable evidence.

Table 1: **Comparison of Interpretability Frameworks.** ToE is distinct in offering an *auditable trace* (search history) over *multimodal* hard evidence.

Method	Type	Hard Evidence?	Multimodal?	Faithfulness?	Auditable Trace?
Attention Weights	Intrinsic (Soft)	✗	-	✗	✗
Gradient Saliency	Post-hoc	✗	-	✗	✗
LIME / SHAP	Post-hoc-Surrogate	✗	-	-	✗
Concept Bottleneck	Intrinsic (Concepts)	✓	-	-	-
Tree-of-Evidence (Ours)	Inference Search	✓	✓	✓	✓

- 3. Comprehensive faithfulness evaluation under evidence budgets.** We evaluate explanations using sufficiency, comprehensiveness, and probability-agreement metrics under strict evidence constraints across three datasets, six tasks, and two domains. We compare against Local Interpretable Model-Agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), CBMs, gradient saliency, and Large Language Models (LLMs) up to 70B parameters, and provide ablations showing ToE better preserves full-input behavior at a given sparsity than all baselines.

2 Related Work

Faithful Rationale Extraction. Rationale extraction seeks a subset of input features that justifies a prediction (Lei et al., 2016), evaluated via *sufficiency* and *comprehensiveness* (DeYoung et al., 2020; Jacovi and Goldberg, 2020), with recent work noting that faithfulness metrics are sensitive to evaluation design (Chan et al., 2022; Edin et al., 2025). Post-hoc attribution methods, such as LIME and SHAP, approximate feature importance via local surrogates or Shapley values but provide no hard selection and can be unstable. Information Bottleneck methods (Paranjape et al., 2020) encourage concise rationales but are typically unimodal, while Concept Bottleneck Models (CBM) (Koh et al., 2020; Vandenhirtz et al., 2024) align hidden states with human-interpretable concepts yet require predefined annotations and remain static at inference.

Search-Based and Multimodal Interpretability. Branching procedures such as Tree-of-Thoughts (Yao et al., 2023) and search-based explainers (Shitole et al., 2021; Zhou and Shah, 2023) apply deliberative search for reasoning or explanation, but operate on single modalities or token spaces. Model-agnostic alternatives such as Anchors (Ribeiro et al., 2018) and counterfactual explanations (Wachter et al., 2017) require extensive perturbation or per-query optimization rather than

jointly trained selectors. In clinical multimodal learning (Acosta et al., 2022), architectures typically fuse text and time-series encoders via late fusion, gating, or cross-attention (Huang et al., 2020; Seki et al., 2021; Golas et al., 2018), but explanations are usually modality-specific post-hoc attributions and rarely yield a single cross-modal evidence set. As summarized in Table 1, ToE is the only framework combining hard evidence selection, multimodal support, faithfulness, and an auditable search trace via jointly trained unit-level selectors.

3 Method

We formulate interpretable clinical prediction as a *search problem*. Our framework, **ToE**, separates the reasoning process into two stages: (1) learn efficient, differentiable heuristics for evidence scoring via *EB*, and (2) perform an inference-time discrete search to identify a compact, high-scoring evidence set required for a robust diagnosis. We represent an overview of our approach in Figure 1.

3.1 Problem Setup and Evidence Units

We define a binary classification task $y \in \{0, 1\}$ over an observation window $[t_0, t_0 + \Delta]$ (default $\Delta=24h$). The input space \mathcal{X} consists of two *searchable* modalities and two *context* modalities. While we present the formulation using ICU data as a running example, the framework applies to any setting where inputs can be decomposed into discrete units across one or more modalities; we evaluate on non-ICU settings in Section 4.

Structured ICU Time Series (searchable evidence). We represent ICU measurements (vital signs and lab values) as a fixed-length sequence $\mathbf{x}^{\text{ts}} = (x_1^{\text{ts}}, \dots, x_T^{\text{ts}})$ with $T = 24$ hourly bins and $x_t^{\text{ts}} \in \mathbb{R}^D$. Each bin contains summary statistics (e.g., mean, min, max) over vitals and labs in that hour, along with missingness indicators. Evidence Units are the discrete time windows $\{W_t\}_{t=1}^T$ corresponding to these bins.

Radiology Reports (searchable evidence). Let \mathbf{x}^{note} be the concatenation of all radiology report text within the window $[t_0, t_0 + \Delta]$. We segment this text into a sequence of chunks (c_1, \dots, c_M) (e.g., 3-sentence segments), padded or truncated to a fixed length M_{max} . Let $\mathbf{a} \in \{0, 1\}^{M_{\text{max}}}$ be a presence mask indicating valid (non-padding) chunks. Evidence Units are the discrete chunks $\{N_j\}_{j=1}^{M_{\text{max}}}$.

CXR and ECG Context (Global Priors). To ground the search in the patient’s broader physiological state, we include fixed context vectors that are not subject to selection: (i) $\mathbf{x}^{\text{cxr}} \in \mathbb{R}^{D_{\text{cxr}}}$, a label vector from the most recent CXR (e.g., CheXpert) with indicator `has_cxr`; and (ii) $\mathbf{x}^{\text{ecg}} \in \mathbb{R}^{D_{\text{ecg}}}$, machine measurements from the most recent ECG with indicator `has_ecg`.

We deliberately model these signals as fixed priors rather than searchable units to mirror clinical reasoning. CXR and ECG typically represent the patient’s baseline physiological state (chronic/background), whereas notes and vitals represent acute evolution (dynamic). By conditioning the search on fixed context, ToE forces the model to identify dynamic evidence that explains the outcome given the patient’s baseline risk, preventing the search from wasting budget on static confirmational signals.

Evidence Set. We formally define an explanation as a tuple of indices $E = (E^{\text{ts}}, E^{\text{note}})$, where $E^{\text{ts}} \subseteq \{1, \dots, T\}$ indexes selected time windows and $E^{\text{note}} \subseteq \{1, \dots, M_{\text{max}}\}$ indexes selected note chunks.

3.2 Evidence Bottleneck Predictors

We employ a modular architecture with two EB streams, corresponding to the searchable modalities (\mathbf{x}^{ts} and \mathbf{x}^{note}). Each stream consists of: (i) a *Selector* that scores discrete evidence units to produce a hard top- k mask; and (ii) a *Predictor* that estimates the diagnosis using only the selected subset. This separation ensures that the model cannot “cheat” by accessing information it has not explicitly selected.

3.2.1 Differentiable Top- k Selector

Let $U = \{u_1, \dots, u_n\}$ be the set of evidence units (time windows or chunk embeddings). A lightweight MLP selector f_θ assigns a scalar relevance score $s_i = f_\theta(u_i)$ to each unit. For variable-length inputs (notes), we enforce validity by setting $s_i = -\infty$ wherever the presence mask $a_i = 0$, ensuring padding is never selected.

To enable end-to-end training with discrete selection, we utilize the Straight-Through Estimator (STE). We compute a hard top- k mask $\mathbf{m} = \text{TopK}(s, k) \in \{0, 1\}^n$ for the forward pass, but approximate gradients via a softmax relaxation $\tilde{\mathbf{m}} = \text{softmax}(s/\tau_{\text{STE}})$ during backpropagation:

$$\hat{\mathbf{m}} = \mathbf{m} - \text{sg}(\tilde{\mathbf{m}}) + \tilde{\mathbf{m}}, \quad (1)$$

where τ_{STE} is the STE temperature (default $\tau_{\text{STE}} = 1.0$; see Appendix A.7.1), $\text{sg}(\cdot)$ denotes the stop-gradient operator. This allows the selector to update its ranking logic θ based on the downstream predictor’s performance.

The STE introduces a forward-backward gradient mismatch by construction. Our two-phase training design mitigates this: in Phase I, the predictor trains with all evidence selected ($k = T$), so the STE is never invoked; in Phase II, the predictor is frozen, and only the selector MLP (98K of 109M total parameters) is updated. Because the frozen predictor’s weights are fixed, the selector needs only to learn a correct *ranking* of units, which units the predictor finds most informative, rather than propagating calibrated classification gradients end-to-end. Gradient mismatch affects magnitude but not ordering, preserving the ranking objective. Empirically, sufficiency Area Under the Receiver Operating Characteristic curve (AUROC) varies less than 1% across a $50\times$ temperature range ($\tau_{\text{STE}} \in \{0.1, 5.0\}$; Appendix A.7.1).

3.2.2 Modality-Specific Encoders

Time-Series Stream. The selector scores raw feature vectors $u_t = x_t^{\text{ts}}$. We apply the mask element-wise, $\tilde{x}_t^{\text{ts}} = \hat{m}_t^{\text{ts}} \cdot x_t^{\text{ts}}$, effectively zeroing out non-selected hours. The sequence is encoded via a Bidirectional Gated Recurrent Unit (GRU) to obtain a final representation \mathbf{v}^{ts} (concatenated hidden states). While we employ a GRU for computational efficiency, our framework is model-agnostic and compatible with continuous-time encoders such as Latent ODEs (Rubanova et al., 2019). Finally, we inject global context by concatenating the projected context vectors:

$$\mathbf{z}^{\text{ts}} = [\mathbf{v}^{\text{ts}}; \psi^{\text{cxr}}(\mathbf{x}^{\text{cxr}}); \psi^{\text{ecg}}(\mathbf{x}^{\text{ecg}})], \quad (2)$$

where ψ are lightweight projection MLPs. A classifier g_ϕ^{ts} maps \mathbf{z}^{ts} to the logit ℓ^{ts} .

Notes Stream. We embed text chunks using a frozen BioClinicalBERT encoder, $e_j = \text{BERT}(c_j)_{[\text{CLS}]}$. The selector scores these embeddings to obtain a mask $\hat{\mathbf{m}}^{\text{note}}$. The predictor computes a masked mean pool over the selected valid chunks:

$$\mathbf{v}^{\text{note}} = \frac{\sum_j \hat{m}_j^{\text{note}} a_j \phi^{\text{note}}(e_j)}{\sum_j \hat{m}_j^{\text{note}} a_j + \epsilon}, \quad (3)$$

where ϕ^{note} is a learnable projection MLP. Similar to the time-series stream, we inject global context:

$$\mathbf{z}^{\text{note}} = [\mathbf{v}^{\text{note}}; \psi^{\text{cxr}}(\mathbf{x}^{\text{cxr}}); \psi^{\text{ecg}}(\mathbf{x}^{\text{ecg}})], \quad (4)$$

and pass \mathbf{z}^{note} to a classifier g^{note} to produce the logit ℓ^{note} .

3.2.3 Training and Inference Fusion

We train the streams separately using class-balanced Binary Cross-Entropy. This isolation ensures that each modality learns independent grounding logic without over-relying on the other.

At inference, we fuse the streams via logit summation. We define the predicted probability for binary evidence masks \mathbf{m}^{ts} and \mathbf{m}^{note} as:

$$p(\mathbf{m}^{\text{ts}}, \mathbf{m}^{\text{note}}) = \sigma(\ell^{\text{ts}}(\mathbf{m}^{\text{ts}}) + \ell^{\text{note}}(\mathbf{m}^{\text{note}})), \quad (5)$$

where $\ell(\cdot)$ denotes the logit output of a stream given a specific mask. We define the *Full-Input Decision* as the prediction using all available units ($\mathbf{m} = \mathbf{1}$), denoted as $p_{\text{full}} = p(\mathbf{1}^{\text{ts}}, \mathbf{1}^{\text{note}})$ with predicted class \hat{y}_{full} .

3.3 Faithfulness Evaluation

We quantify interpretability using the Evaluating Rationales And Simple English Reasoning (ERASER) benchmark standards for faithfulness (DeYoung et al., 2020).

Sufficiency. Measures if the selected evidence is adequate to reproduce the prediction. We report the model’s performance (AUROC, Area Under the Precision-Recall Curve or AUPRC) when masking out all non-selected units (i.e., keeping only the top- k evidence).

Comprehensiveness. Measures if the model relies on the selected evidence. We calculate the drop in confidence for the *originally predicted class* \hat{y}_{full} when the selected evidence is removed. Let m_{sel} be the selected evidence mask and $m_{\text{rem}} = \mathbf{1} - m_{\text{sel}}$ be the complement. We compute:

$$\Delta_{\text{comp}} = \frac{1}{N} \sum_{i=1}^N \left[\Pr(\hat{y}_{\text{full}}^{(i)} | \mathbf{1}) - \Pr(\hat{y}_{\text{full}}^{(i)} | \mathbf{m}_{\text{rem}}^{(i)}) \right]. \quad (6)$$

A higher Δ_{comp} indicates that the model’s prediction relied heavily on the removed evidence.

3.4 Tree-of-Evidence (ToE): Inference-Time Search

Standard top- k selection is brittle because it assumes evidence units are independent. However,

clinical evidence is often synergistic (e.g., a radiology finding of alveolar edema contextualizes a subsequent drop in SpO_2). To address this, we propose ToE, a discrete beam search algorithm that identifies a compact evidence set to reproduce the full-input decision. Following the terminology introduced in Section 1, this constitutes the “System 2” component of our framework: a multi-step deliberative search that explicitly evaluates candidate evidence combinations, in contrast to “System 1” single-pass greedy ranking.

3.4.1 Search Space and Candidates

A search state is a pair of binary masks $\mathbf{m} = (\mathbf{m}^{\text{ts}}, \mathbf{m}^{\text{note}})$. To keep the search tractable, we restrict actions to the top- N candidates per modality (ranked by selector scores) to control computation.

3.4.2 Search Objective

We seek a state that maximizes confidence in the original decision while minimizing evidence cost. For a state \mathbf{m} , we define the scoring function:

$$C(\mathbf{m}) = \Pr(\hat{y}_{\text{full}} | \mathbf{m}), \quad (7)$$

$$S(\mathbf{m}) = 1 - |p_{\text{full}} - p(\mathbf{m})|, \quad (8)$$

$$K(\mathbf{m}) = \|\mathbf{m}^{\text{ts}}\|_0 + \|\mathbf{m}^{\text{note}}\|_0, \quad (9)$$

$$\text{score}(\mathbf{m}) = C(\mathbf{m}) + \lambda S(\mathbf{m}) - \mu K(\mathbf{m}), \quad (10)$$

where C is the model’s confidence in the full-input predicted class (*Sufficiency*), S encourages probability stability (*Calibration*), and K penalizes evidence cost.

We define the stability term $S(\mathbf{m})$ in probability space (Eq. 8) rather than logit space. This choice reflects three considerations. First, near $p = 0$ or $p = 1$, where most ICU patients fall, given class prevalences of 7–14%, large logit deviations produce negligible probability changes; probability-space stability appropriately assigns low cost to these clinically irrelevant shifts. Second, the resulting metric is bounded in $[0, 1]$ and directly interpretable as “mortality risk shifted by X percentage points.” Third, it is numerically stable, avoiding the divergences that logit-space distances exhibit near saturation. By including the stability term, the search does not merely maximize confidence (which could lead to selecting evidence that inflates a prediction) but explicitly aims to match the calibration of the full model. This ensures the selected evidence is not just “sufficient” in isolation, but faithful to the model’s complete decision.

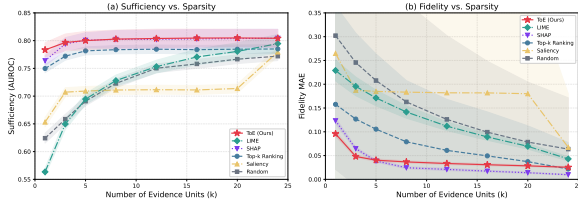


Figure 2: **The Faithfulness-Sparsity Frontier.** Performance across evidence budgets k on MIMIC-IV (E1: In-Hospital Mortality, 5 seeds). **(a) Sufficiency:** ToE (Red \star) matches the full model’s predictive power (AUROC ≈ 0.80) with as few as $k=5$ units. **(b) Fidelity:** ToE achieves the lowest Fidelity MAE at sparse budgets ($k \leq 5$), reducing error by $>50\%$ compared to Top- k Ranking (Blue \bullet) and Saliency (Gold \blacktriangle) at sparse budgets, proving it captures the model’s actual confidence rather than just label correlations.

3.4.3 Algorithm and Efficiency

The search proceeds as follows (Algorithm 1, Appendix A.4):

1. **Initialization:** Start with an empty evidence set.
2. **Expansion:** At each step, generate candidate states by adding exactly one unit from the candidate list $\mathcal{W} \cup \mathcal{N}$.
3. **Pruning:** Evaluate candidates via frozen EB predictors and retain the top- B states (Beam Width).
4. **Termination:** Stop if a state meets sufficiency thresholds ($\tau_{\text{conf}}, \tau_{\text{suff}}$) or max steps are reached.

Note that beam search finds high-scoring evidence sets under the scoring heuristic (Eq. 10), not globally optimal ones. Exhaustive enumeration confirms optimality gaps below 0.001 AUROC at small k on MIMIC-IV (Appendix A.7.3).

Efficiency via Caching: Since the BERT backbone is frozen, we cache the embeddings $\{e_j\}$ for all note chunks once per patient. During search, state evaluation requires only lightweight pooling and MLP passes, making ToE computationally efficient and suitable for deployment.

4 Experiment

4.1 Dataset and Implementation Details

Dataset and Cohort.

MIMIC-IV. Our primary evaluation uses the MIMIC-IV dataset (Johnson et al., 2024, 2023). The cohort consists of adult patients with at least 24 hours of ICU observation data. The final dataset comprises $N = 74,829$ unique ICU stays, split into training ($N = 52,597$), validation ($N = 11,053$), and testing ($N = 11,179$). We evaluate on four prediction tasks: (E1) Hospital Mortality (prevalence 11.5%), (E2) Long Length of Stay (>7 days; 14.1%), (E3) ICU Mortality (7.4%), and (E4) Post-Observation Mortality (11.2%). All MIMIC-IV results report mean \pm std across 5 random seeds unless otherwise noted.

eICU. To test cross-center generalization, we evaluate on the eICU Collaborative Research Database (Pollard et al., 2018), spanning 208 hospitals across the United States. We apply the same pipeline with no architectural modifications.

LEMMA-RCA. To test domain transfer beyond healthcare, we evaluate on LEMMA-RCA (Zheng et al., 2024), a microservice fault detection benchmark (prevalence 22%). Time-series evidence units correspond to service-level metrics and text units to log message chunks. The ToE pipeline is applied without modification.

Reproducibility and Hyperparameters. For the ToE beam search, we set beam width $B = 8$, maximum search depth $S_{\text{max}} = 10$, and restrict candidates to the top $N_{\text{ts}} = 24$ time-series windows and $N_{\text{note}} = 20$ note chunks per instance. The search objective weights are $\lambda = 1.0$ (stability) and $\mu = 0.05$ (sparsity cost), with stopping thresholds $\tau_{\text{conf}} = 0.9$ and $\tau_{\text{suff}} = 0.9$. All experiments use a batch size of 32 on a single NVIDIA A100 GPU.

4.2 Baseline Comparisons

We compare ToE against six baselines: **Top- k Ranking**, which selects the k units with the highest individual selector scores; **Saliency (Gradient)**, which ranks units by Input \times Gradient magnitude; **LIME** (Ribeiro et al., 2016) and **SHAP** (Lundberg and Lee, 2017), which select the top- k units by local surrogate coefficients and Shapley-value attributions, respectively; a **Hard Concept Bottleneck Model (CBM)** (Koh et al., 2020) with 24 binary clinical concepts grounded in established scoring systems; and **Random** selection as a lower bound.

Faithfulness–Sparsity Frontier. Figure 2 and Table 2 compare all methods across evidence bud-

Table 2: Comparison with LIME and SHAP (E1: Hospital Mortality, 5 seeds). ToE achieves the best fidelity–sufficiency tradeoff at every budget k . Full results across all k in Appendix A.5, Figure A4.

k	Method	AUROC	Fidelity MAE (\downarrow)	ECE (\downarrow)
1	LIME	0.564 \pm 0.006	0.229 \pm 0.022	0.406 \pm 0.011
	SHAP	0.764 \pm 0.009	0.123 \pm 0.006	0.320 \pm 0.010
	ToE	0.783 \pm 0.013	0.096 \pm 0.005	0.297 \pm 0.019
5	LIME	0.695 \pm 0.010	0.171 \pm 0.016	0.332 \pm 0.018
	SHAP	0.801 \pm 0.014	0.039 \pm 0.002	0.302 \pm 0.025
	ToE	0.800 \pm 0.017	0.040 \pm 0.003	0.280 \pm 0.023
10	LIME	0.743 \pm 0.011	0.126 \pm 0.011	0.308 \pm 0.020
	SHAP	0.802 \pm 0.017	0.024 \pm 0.002	0.299 \pm 0.029
	ToE	0.803 \pm 0.017	0.035 \pm 0.002	0.283 \pm 0.025

gets. ToE matches the full model’s predictive power (AUROC \approx 0.800) with as few as $k=5$ units (Fig. 2a) while maintaining the lowest fidelity error and ECE at every sparsity level (Fig. 2b). At $k=1$, ToE reduces fidelity Mean Absolute Error (MAE) by 39% relative to Top- k Ranking (0.158 to 0.096) and by 58% relative to LIME, and outperforms LIME by 22 AUROC points. SHAP is the strongest attribution baseline, matching ToE’s AUROC and MAE at $k \geq 5$, but exhibits 22% higher fidelity MAE at $k = 1$ (0.123 vs. 0.096) and higher ECE at every budget, indicating weaker calibration alignment. ToE, by explicitly optimizing for probability stability (Eq. 8), captures the model’s actual confidence rather than just label correlations. The gap between methods narrows at higher budgets as the evidence space saturates. Multi-seed ECE results across all four MIMIC-IV tasks confirm that ToE achieves comparable or lower calibration error than the full model (Appendix A.5; Figure A4).

Comparison with Concept Bottleneck Models.

The Hard CBM with 24 clinical concepts achieves AUROC 0.775 (AUPRC 0.349). ToE exceeds CBM on AUROC with a single evidence unit ($k = 1$: 0.783) and widens the gap at $k = 5$ (0.800), while requiring no predefined concept annotations, though CBM retains an AUPRC edge from its curated concepts. This highlights a key advantage: CBMs require domain experts to define a fixed concept vocabulary before training, whereas ToE discovers relevant evidence units from learned representations at inference time.

Comparison with LLMs. We compare against 10 open-source Multimodal LLMs or (M)LLMs (1B–70B parameters), including medical fine-tunes and vision-language models, evaluated on E1 via zero-shot prompting with the full test set. Table 3

Table 3: LLM/MLLM comparison (E1: Hospital Mortality). ToE with 109M parameters outperforms all models up to 70B. Full 10-model results in Appendix A.6; Figure A3.

Model	Type	Params	AUROC	AUPRC
Llama-3.2-1B	text	1.2B	0.532 \pm 0.009	0.135 \pm 0.005
Llama-3.1-8B	text	8.0B	0.691 \pm 0.008	0.206 \pm 0.008
Med42-v2-70B	text	70B	0.745 \pm 0.009	0.293 \pm 0.014
MedGemma-27B (V)	vision	27B	0.630 \pm 0.009	0.190 \pm 0.008
ToE $k=5$	multi	109M	0.800 \pm 0.017	0.310 \pm 0.067

Table 4: Cross-task and cross-dataset evaluation. ToE retains $\geq 98\%$ of full-model AUROC at $k=5$ across all settings. MIMIC-IV results: mean \pm std over 5 seeds; eICU and LEMMA-RCA results are from a single seed.

Dataset	Task	Full AUROC	ToE $k=5$	Fid. MAE
E1	MIMIC-IV Hosp. mort.	0.806 \pm 0.015	0.800 \pm .017	0.040 \pm 0.003
E2	MIMIC-IV Long LOS	0.747 \pm 0.041	0.740 \pm .046	0.031 \pm 0.002
E3	MIMIC-IV ICU mort.	0.816 \pm 0.009	0.808 \pm .011	0.042 \pm 0.004
E4	MIMIC-IV Post-obs.	0.794 \pm 0.021	0.784 \pm .023	0.041 \pm 0.001
eICU	ICU mort.	0.822	0.808	0.124
LEMMA	Fault det.	0.741	0.730	0.106

reports a representative subset; full results are in Appendix A.6; Figure A3. Even the strongest model, Med42-v2-70B (AUROC of 0.745), underperforms ToE at $k = 5$ (AUROC of 0.800) despite having 640 \times more parameters. Vision-language models (Gemma-2-12B (V), MedGemma-27B (V) (Team et al., 2024)) underperform their text-only counterparts on this task, suggesting that current Multimodal Large Language Models (MLLMs) struggle to extract discriminative signals from raw clinical images for structured prediction.

4.3 Cross-Task and Cross-Dataset Generalization

A primary concern is whether ToE generalizes beyond a single task and dataset. Table 4 reports results across all six evaluation settings.

Three observations merit emphasis. First, ToE retains 98.5–99.3% of full-model AUROC at $k = 5$ across all four MIMIC-IV tasks, with fidelity MAE consistently in the narrow range 0.031–0.042, despite substantial differences in clinical semantics and class balance (7.4–14.1%). Second, eICU replicates the core finding on an independent multi-center dataset spanning 208 hospitals with different EHR systems and documentation practices. Third, LEMMA-RCA demonstrates that the same pipeline generalizes beyond healthcare entirely, with no architectural modifications.

Table 5: **Modality Ablation Results** ($k = 5$). Notes-Only fails to ground predictions, while the Multimodal (Both) approach maintains high predictive power and stability comparable to the strong TS-Only baseline.

Modality	AUROC	AUPRC	Fidelity MAE
TS Only	0.7876 ± 0.0075	0.3912 ± 0.0151	0.0445 ± 0.0730
Notes Only	0.5590 ± 0.0077	0.1338 ± 0.0047	0.3432 ± 0.2915
Both	0.8001 ± 0.0165	0.3096 ± 0.0672	0.0403 ± 0.0027

4.4 Ablation Studies

To validate the components of the ToE framework, we analyzed the contribution of different modalities and the scoring objective.

Modality Necessity. Table 5 examines modality contributions at $k=5$. The Notes-Only baseline fails (AUROC ≈ 0.56 , MAE > 0.3), confirming that radiology text alone is insufficient for grounding without physiological context. The multimodal approach matches the predictive power of the time-series backbone (AUROC ≈ 0.80) while maintaining low fidelity error (MAE ≈ 0.04), validating ToE’s design of using robust vitals to anchor the search while selectively retrieving text for semantic explanation.

Search & Objective Analysis. Table 6 validates the deliberative search design: removing the stability objective ($\lambda = 0$) more than doubles the fidelity error, confirming that faithful explanations require matching the model’s calibration, not just maximizing confidence.

Table 6: **Search Objective Ablation** ($k = 5$, MIMIC-IV Mortality, 5 seeds). The full ToE objective achieves the lowest Fidelity MAE. Removing stability ($\lambda=0$) doubles the error. At fixed k , the sparsity cost μ has no effect (identical to Full). Top- k Ranking without beam search doubles the MAE (+100%).

Configuration	AUROC	AUPRC	Fidelity MAE	Comp.
Full ($\lambda=1.0, \mu=0.05$)	0.8001 ± 0.0165	0.3096 ± 0.0672	0.0403 ± 0.0027	0.1112 ± 0.0143
No Stability ($\lambda=0.0, \mu=0.05$)	0.7738 ± 0.0338	0.3069 ± 0.0622	0.0800 ± 0.0052	0.1371 ± 0.0153
No Sparsity ($\lambda=1.0, \mu=0.0$)	0.7915 ± 0.0385	0.3191 ± 0.0682	0.0408 ± 0.0015	0.1025 ± 0.0115
Top- k Ranking (no search)	0.7735 ± 0.0339	0.3066 ± 0.0625	0.0806 ± 0.0054	0.1357 ± 0.0158

Search vs. Ranking. To isolate the contribution of combinatorial search from the scoring function, we compare beam search against Top- k Ranking using identical selector scores (Appendix A.7). The advantage is most pronounced under strict sparsity: at $k=1$, beam search achieves AUROC 0.783 ± 0.013 versus 0.768 ± 0.036 for ranking, with fidelity MAE reduced by 11% (0.096 vs. 0.107). At $k=5$, the gap widens to $+0.027$ AUROC and 50% lower MAE (0.040 vs. 0.081). The

Table 7: Search exhaustion rates for true positive (TP) vs. false positive (FP) predictions. When the model is wrong, the search exhausts its budget 4–26 \times more often.

Metric	eICU	MIMIC-IV
TP search exhaustion rate	0.3%	7.2%
FP search exhaustion rate	7.3%	30.2%
FP / TP exhaustion ratio	25.6 \times	4.2 \times

gap narrows at higher budgets as the evidence space saturates, confirming that combinatorial search matters most at sparse budgets where the model must identify the most decisive evidence units.

4.5 Auditing Model Reliability via Search Behavior

ToE is faithful to the *model’s* logic, not to clinical ground truth; if the base predictor relies on spurious correlations or biases, ToE will faithfully surface them. We argue this is a feature rather than a limitation: ToE’s search behavior provides a built-in diagnostic for prediction reliability.

Among positive predictions, we observe a systematic divergence between true positives and false positives (Table 7). When the model is correct, ToE finds supporting evidence almost immediately. When the model is wrong, the search struggles and exhausts its budget 4–26 \times more often. This asymmetry enables *selective abstention*: flagging predictions where the search exhausted its budget catches 7.3% of false positives on eICU while losing only 0.3% of true positives, improving precision with negligible sensitivity loss. A spurious feature injection experiment further validates this signal: a model retrained with a deliberately spurious feature (80/20 correlated in training, 0% in test) requires 4.5 \times more evidence to converge and halves its convergence rate (Appendix A.8).

4.6 Efficiency Analysis

A common concern with search-based methods is latency. Our timing analysis reveals that ToE adds only ~ 13 ms of overhead per patient compared to the full forward pass. This efficiency, achieved via caching BERT embeddings and lightweight GRU updates, makes ToE suitable for real-time clinical deployment.

4.7 Qualitative Analysis

To understand how ToE navigates the multimodal landscape, we visualize search traces for two repre-

Table 8: Qualitative comparison of ToE traces. **(Left)** ToE efficiently solves clear-cut cases using only vitals. **(Right)** ToE dynamically integrates clinical notes to surface interpretable clinical context, extracting specific medical concepts (e.g., “Alveolar Edema”) to ground the prediction.

Case A: Efficient Triage (Patient A)	Case B: Multimodal Synergy (Patient B)
Task: Mortality Prediction Full Model Prob: 0.0005 (Low Risk) Final Trace: $k = 1$ (Vitals Only)	Task: Mortality Prediction Full Model Prob: 0.861 (High Risk) Final Trace: $k = 10$ (8 Vitals + 2 Notes)
Search Step 1: Add Vitals W5 \hookrightarrow Evidence: [Physiological Window 5] \hookrightarrow Sufficiency: 0.998 (Threshold Met)	Search Steps 1–8: Add Vitals W23, W1, . . . , W11 \hookrightarrow Evidence: [Multiple Physiological Windows] \hookrightarrow Sufficiency: 0.840 (Plateau)
Outcome: The search terminates immediately. The model determines that the vital signs alone are sufficient to justify the “Low Risk” prediction. No notes are processed.	Search Step 9: Add Note N2 \hookrightarrow Evidence: “... Coalescent, bilateral, perihilar opacities reflect alveolar edema... suggest volume overload.” \hookrightarrow Sufficiency: 0.833 (Stable)
-	Search Step 10: Add Note N3 \hookrightarrow Evidence: [Additional Radiology Report] \hookrightarrow Sufficiency: 0.833 (No Change)
-	Outcome: Vitals carry the primary predictive signal but plateau below the sufficiency threshold. The search retrieves the “Volume Overload” finding to provide clinically interpretable grounding for the high mortality risk. The slight sufficiency dip ($0.840 \rightarrow 0.833$) falls within noise; since $\tau_{suff} = 0.9$ is unmet, the search continues and surfaces cross-modal evidence.

sentative patients in Table 8.

Case A: Efficient Triage (Vitals-Only). For Patient A, the model identifies a clear physiological deterioration solely from time-series data. The search selects a single vital-sign window (W5) showing acute instability. This evidence alone yields a sufficiency score of 0.998, triggering the stopping criterion immediately ($k = 1$). By recognizing that the vitals are unambiguous, ToE avoids processing the clinical notes entirely, reducing computational cost.

Case B: Multimodal Resolution. In contrast, Patient B presents a more complex picture. The search initially retrieves multiple time-series windows (W23, W1, ..., W11), but the sufficiency score plateaus around 0.84, indicating the physiological signals alone do not fully explain the model’s high-risk prediction. At this plateau, ToE expands to the clinical notes, retrieving a specific radiology report segment (N2) that documents “...bilateral perihilar opacities reflect alveolar edema... suggest volume overload.” While sufficiency remains stable (0.833), this textual evidence provides the causal context (*Volume Overload*) that grounds the physiological signals in a clinically interpretable explanation, demonstrating ToE’s ability to surface

relevant cross-modal evidence even when the vitals alone carry the predictive signal.

5 Conclusion

We introduced ToE, an inference-time search framework that formulates interpretability as a discrete optimization problem over evidence units, combining Evidence Bottlenecks with beam search to produce auditable traces. Across six task–dataset settings spanning MIMIC-IV, eICU, and LEMMA-RCA, ToE retains at least 98% of full-model AU-ROC with as few as five evidence units. At sparse budgets, ToE achieves lower fidelity error than LIME, SHAP, saliency, and Top-k Ranking, outperforms CBMs without predefined annotations, and beats LLMs up to 70B parameters with only 109M. Beyond explanation, ToE’s search behavior serves as a reliability diagnostic: exhaustion rates are 4–26× higher for false positives than true positives, enabling selective abstention and showing that search-based rationales can recover a model’s decision logic more faithfully than ranking or post-hoc attribution.

6 Limitations

ToE produces evidence sets that are faithful to the underlying model’s decision logic, not to clinical

ground truth. If the base predictor relies on spurious correlations or biases, ToE will surface them rather than correct them, though, as shown in Section 4.5, this model-faithfulness itself serves as a diagnostic for unreliable predictions. More broadly, ToE is an interpretability wrapper: it cannot fix errors in the base model, and its coarse evidence units (hourly windows, report chunks) may omit finer-grained clinically relevant signals.

The framework is currently validated on late-fusion architectures with separable evidence streams; extending to cross-attention or early-fusion models requires additional design. Beam search finds near-optimal evidence sets under the scoring heuristic, not globally optimal ones, though exhaustive enumeration confirms gaps below 0.001 AUROC at small k on MIMIC-IV (Appendix A.7.3). Finally, while ToE’s ~ 13 ms overhead is practical for most settings, runtime may increase with longer note histories or larger beam widths.

7 Ethical Considerations

Clinical Safety and Intended Use. This research presents a prototype for clinical decision support and is not intended for autonomous diagnosis or treatment planning. False negatives in mortality prediction could lead to reduced care, while false positives could cause alarm fatigue. ToE is designed explicitly to mitigate these risks by forcing the model to show its work, allowing clinicians to verify or reject the machine’s rationale. We emphasize that the selected evidence is a mathematical construct reflecting the model’s confidence, not a comprehensive summary of the patient’s clinical state.

Data Privacy and Compliance. Our models were developed using the MIMIC-IV dataset, which contains de-identified electronic health records from Beth Israel Deaconess Medical Center. We adhered to the PhysioNet Credentialed Data Use Agreement, ensuring no attempt was made to re-identify patients. Any deployment of this technology in a live clinical setting would require strict adherence to local regulations (e.g., Health Insurance Portability and Accountability Act (HIPAA) in the United States, General Data Protection Regulation (GDPR) in Europe) and rigorous external validation.

Bias and Fairness. Clinical datasets are known to harbor demographic and socioeconomic biases. A model trained on MIMIC-IV (collected in Boston, MA) may underperform or rely on different feature sets for underrepresented populations. A key advantage of ToE is its ability to *audit* these biases; by inspecting the evidence trees, stakeholders can detect if the model relies on impermissible proxies (e.g., insurance status or language barriers) for its predictions. However, the search algorithm itself does not remove these biases, and deploying the model without fairness audits could perpetuate existing healthcare disparities.

Acknowledgments

This research was supported in part through research cyberinfrastructure resources and services, including the AI Makerspace of the College of Engineering (RRID:SCR_028058), provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA (RRID:SCR_027619). We also gratefully acknowledge funding and fellowships that contributed to this work, including a Wallace H. Coulter Distinguished Faculty Fellowship, a Petit Institute Faculty Fellowship, and research funding from Amazon and Microsoft Research awarded to Professor May D. Wang.

References

- Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. 2022. Multimodal biomedical ai. *Nature medicine*, 28(9):1773–1784.
- Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pages 72–78.
- Chun Sik Chan, Huanqi Kong, and Liang Guanqing. 2022. [A comparative study of faithfulness metrics for model interpretability methods](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5029–5038, Dublin, Ireland. Association for Computational Linguistics.
- Zheyi Chen, Liuchang Xu, Hongting Zheng, Luyao Chen, Amr Tolba, Liang Zhao, Keping Yu, and Hailin Feng. 2024. Evolution and prospects of foundation models: From large language models to large multimodal models. *Computers, Materials & Continua*, 80(2).

- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. Eraser: A benchmark to evaluate rationalized nlp models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4443–4458.
- Joakim Edin, Andreas Geert Motzfeldt, Casper L. Christensen, Tuukka Ruotsalo, Lars Maaløe, and Maria Maistro. 2025. [Normalized AOPC: Fixing misleading faithfulness metrics for feature attributions explainability](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1730, Vienna, Austria. Association for Computational Linguistics.
- Sara Bersche Golas, Takuma Shibahara, Stephen Agboola, Hiroko Otaki, Jumpei Sato, Tatsuya Nakae, Toru Hisamitsu, Go Kojima, Jennifer Felsted, Sujay Kakarmath, and 1 others. 2018. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. *BMC medical informatics and decision making*, 18(1):44.
- Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220.
- Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W Waks, Parastou Eslami, Tanner Carbonati, Ashish Chaudhari, Elizabeth Herbst, Dana Moukheiber, Seth Berkowitz, Roger Mark, and Steven Horng. 2023. [MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset. PhysioNet](#). Version 1.0.
- Dawei Huang, Chuan Yan, Qing Li, and Xiaojiang Peng. 2024. From large language models to large multimodal models: A literature review. *Applied Sciences*, 14(12):5068.
- Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P Lungren. 2020. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine*, 3(1):136.
- Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4198–4205.
- Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556.
- A Johnson, L Bulgarelli, T Pollard, B Gow, B Moody, S Horng, LA Celi, and R Mark. 2024. Mimic-iv (version 3.1). physionet. rrid: Scr_007345.
- Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.
- Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. [Rationalizing neural predictions](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.
- Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An information bottleneck approach for controlling conciseness in rationale extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1938–1952.
- Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. 2018. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):180178.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.
- Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. 2019. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32.

- Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.
- Tomohisa Seki, Yoshimasa Kawazoe, and Kazuhiko Ohe. 2021. Machine learning-based prediction of in-hospital mortality using admission laboratory data: A retrospective, single-site study using electronic health record data. *PLoS one*, 16(2):e0246640.
- Vivswan Shitole, Fuxin Li, Minsuk Kahng, Prasad Tadepalli, and Alan Fern. 2021. One explanation is not enough: structured attention graphs for image classification. *Advances in Neural Information Processing Systems*, 34:11352–11363.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, and 1 others. 2024. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138.
- Moritz Vandenheert, Sonia Laguna, Ričards Marcinkevičs, and Julia Vogt. 2024. Stochastic concept bottleneck models. *Advances in Neural Information Processing Systems*, 37:51787–51810.
- Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841.
- Sarah Wiegrefe and Yuval Pinter. 2019. [Attention is not not explanation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.
- Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. 2023. The shaky foundations of large language models and foundation models for electronic health records. *npj digital medicine*, 6(1):135.
- Xi Xu, Jianqiang Li, Zhichao Zhu, Linna Zhao, Huina Wang, Changwei Song, Yining Chen, Qing Zhao, Jijiang Yang, and Yan Pei. 2024. A comprehensive review on synergy of multi-modal data and ai technologies in medical diagnosis. *Bioengineering*, 11(3):219.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.
- Lecheng Zheng, Zhengzhang Chen, Dongjie Wang, Chengyuan Deng, Reon Matsuoka, and Haifeng Chen. 2024. Lemma-rca: A large multi-modal multi-domain dataset for root cause analysis. *arXiv preprint arXiv:2406.05375*.
- Yilun Zhou and Julie Shah. 2023. The solvability of interpretability evaluation metrics. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2399–2415.

Table A1: ToE Comprehensiveness (\uparrow) across evidence budgets and four MIMIC-IV tasks (mean \pm std, 5 seeds).

k	E1: Hospital Mort.	E2: Long LOS	E3: ICU Mort.	E4: Post-Obs Mort.
1	0.0413 \pm 0.0050	0.0447 \pm 0.0107	0.0326 \pm 0.0050	0.0441 \pm 0.0037
3	0.0885 \pm 0.0104	0.0939 \pm 0.0094	0.0749 \pm 0.0092	0.0908 \pm 0.0051
5	0.1112 \pm 0.0143	0.1199 \pm 0.0104	0.0961 \pm 0.0131	0.1139 \pm 0.0067
8	0.1293 \pm 0.0177	0.1431 \pm 0.0117	0.1138 \pm 0.0159	0.1334 \pm 0.0089
12	0.1417 \pm 0.0199	0.1582 \pm 0.0125	0.1270 \pm 0.0180	0.1463 \pm 0.0108
16	0.1489 \pm 0.0216	0.1644 \pm 0.0123	0.1342 \pm 0.0190	0.1541 \pm 0.0121
20	0.1549 \pm 0.0233	0.1681 \pm 0.0119	0.1390 \pm 0.0201	0.1596 \pm 0.0137
24	0.1612 \pm 0.0186	0.1616 \pm 0.0097	0.1307 \pm 0.0403	0.1539 \pm 0.0287

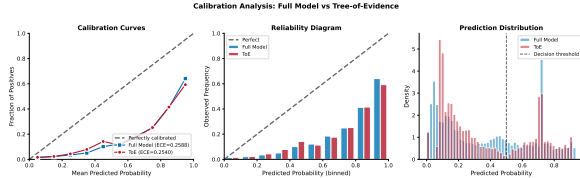


Figure A1: **Calibration Analysis.** Full Model (Blue) vs. ToE at $k=5$ (Red). ToE preserves calibration (ECE 0.254 vs. 0.259) while using sparse evidence.

A Appendix

A.1 Detailed Performance Across Budgets

Figure A4 and Table A1 detail the performance of ToE across varying evidence budgets (k). Notably, Sufficiency AUROC saturates at $k = 5$, indicating that a handful of clinical events are often sufficient for robust diagnosis.

A.2 Calibration Analysis

We evaluated calibration using Expected Calibration Error (ECE). As shown in Figure A1, ToE achieves comparable calibration to the full model (ECE 0.254 vs. 0.259), with both models exhibiting similar reliability curves across probability bins. The prediction distributions confirm that ToE preserves the full model’s confidence profile while operating on only $k=5$ evidence units.

A.3 Evidence Size Distribution

Figure A2 illustrates the distribution of selected evidence sizes at budget $k=5$, stratified by patient outcome and prediction correctness. When the model is correct ($n=8,032$), ToE frequently finds sufficient evidence before exhausting the budget, with notable mass at $k=1-4$. When the model is incorrect ($n=3,147$), the search almost universally consumes the full budget ($k=5$), reflecting the absence of a coherent evidence subset that supports the (wrong) prediction. This asymmetry makes evidence utilization a diagnostic signal for prediction reliability.

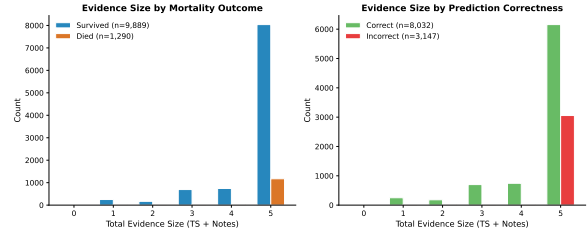


Figure A2: **Evidence Size Distribution ($k=5$, MIMIC-IV Mortality).** **Left:** By mortality outcome, both groups concentrate at the budget cap, with survivors showing slightly more early stopping. **Right:** By prediction correctness, correct predictions exhibit greater evidence efficiency (more mass at $k < 5$), while incorrect predictions exhaust the full budget in 97% of cases, indicating the search struggles to find supporting evidence when the model is wrong.

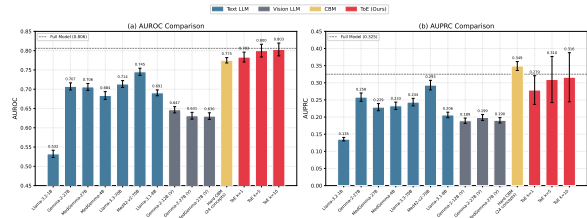


Figure A3: **LLM/MLLM, CBM, and ToE Comparison on E1 (In-Hospital Mortality, MIMIC-IV).** (a) AUROC and (b) AUPRC for 7 text-only LLMs (1B–70B), 3 vision LLMs (12–27B), Hard CBM (24 clinical concepts), and ToE at $k=1, 5, 10$ (109M parameters). Error bars denote ± 1 std (bootstrap for LLMs/CBM, 5 seeds for ToE). Dashed line indicates the full model. ToE $k=5$ outperforms the best 70B LLM (Med42) by $+0.055$ AUROC with $640\times$ fewer parameters, and adding vision to LLMs degrades performance.

A.4 Tree of Evidence (ToE) Algorithm

We provide below an algorithm for the ToE (Algorithm 1).

A.5 Full LIME/SHAP Comparison

Figure A4 extends the main-paper comparison (Table 2) to all evidence budgets $k \in \{1, 3, 5, 10, 15, 20, 25\}$ with AUROC, AUPRC, Fidelity MAE, and ECE.

A.6 Full LLM/MLLM and CBM Comparison

Figure A3 reports the complete comparison against 10 open-source LLMs, CBM, and multimodal LLMs on E1 (Hospital Mortality), evaluated via zero-shot prompting on the full test set. All models are run locally via vLLM (Kwon et al., 2023).

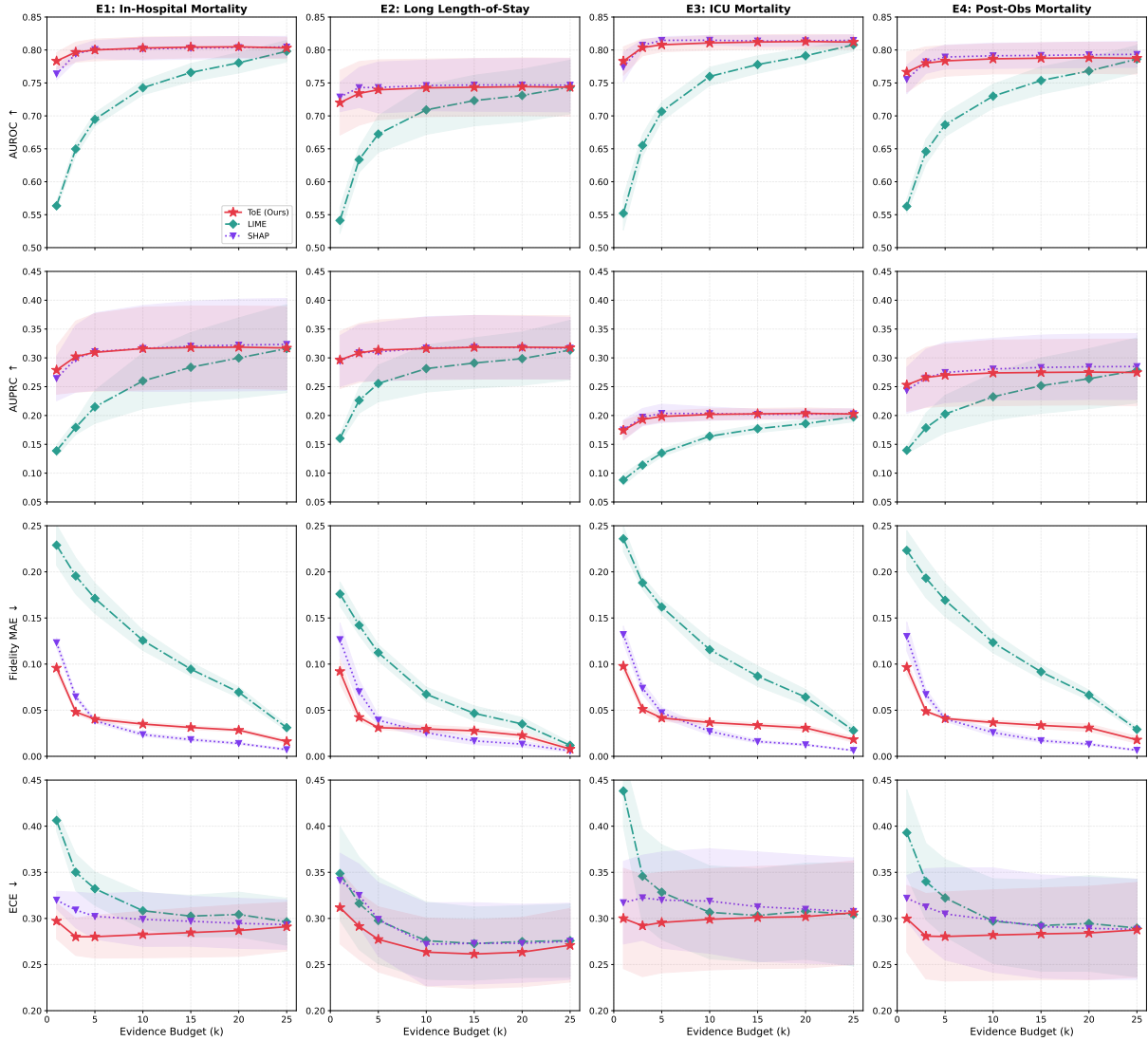


Figure A4: **Complete comparison of ToE, LIME, and SHAP across four MIMIC-IV tasks (E1–E4, 5 seeds, $k \in \{1, 3, 5, 10, 15, 20, 25\}$).** Rows: AUROC, AUPRC, Fidelity MAE, ECE. Columns: E1 (In-Hospital Mortality), E2 (Long LOS), E3 (ICU Mortality), E4 (Post-Obs Mortality). Shaded regions denote ± 1 std. ToE consistently achieves the best ECE across all tasks and competitive AUROC at sparse budgets ($k \leq 5$), while SHAP converges at higher k with lower MAE.

A.7 Search vs. Ranking: Full Comparison

Table A2 provides the full comparison between ToE beam search and Top- k Ranking across all evidence budgets (E1: Hospital Mortality).

Table A2: Search vs Ranking Comparison Across Evidence Budgets (MIMIC-IV Mortality, 5 seeds).

k	Beam Search (ToE)		Top- k Ranking	
	AUROC	Fidelity MAE	AUROC	Fidelity MAE
1	0.7833 \pm 0.0128	0.0958 \pm 0.0053	0.7676 \pm 0.0362	0.1073 \pm 0.0071
3	0.7967 \pm 0.0152	0.0481 \pm 0.0023	0.7739 \pm 0.0343	0.0798 \pm 0.0054
5	0.8001 \pm 0.0165	0.0403 \pm 0.0027	0.7735 \pm 0.0339	0.0806 \pm 0.0054
8	0.8028 \pm 0.0162	0.0367 \pm 0.0023	0.7722 \pm 0.0325	0.0840 \pm 0.0056
12	0.8039 \pm 0.0167	0.0333 \pm 0.0022	0.7713 \pm 0.0314	0.0859 \pm 0.0060
16	0.8046 \pm 0.0160	0.0307 \pm 0.0022	0.7723 \pm 0.0311	0.0848 \pm 0.0061
20	0.8048 \pm 0.0156	0.0284 \pm 0.0019	0.7766 \pm 0.0319	0.0783 \pm 0.0059
24	0.8043 \pm 0.0159	0.0254 \pm 0.0016	0.7837 \pm 0.0346	0.0459 \pm 0.0047

A.7.1 STE Temperature Sensitivity

Table A3 reports the effect of STE temperature τ_{STE} on selector performance, with full retraining per temperature (eICU, ICU Mortality).

Table A3: STE temperature sensitivity (eICU). Performance varies $< 1\%$ across a $50\times$ range.

τ_{STE}	Suff. AUROC ($k=6$)	AUPRC ($k=6$)	Suff. AUROC ($k=1$)
0.1	0.792	0.316	0.741
0.5	0.797	0.321	0.740
1.0	0.799	0.325	0.753
2.0	0.799	0.319	0.745
5.0	0.790	0.313	0.748

Table A4: **ToE vs Zero-Shot LLM.** Comparison on a representative subset of 5 patients. While the LLM achieves perfect prediction accuracy by leveraging external medical knowledge, ToE remains faithful to the underlying model’s logic. ToE selects significantly sparser evidence (Avg 6.2 vs. 9.0 windows) while maintaining reasonable overlap with the LLM’s clinical reasoning.

Patient ID	Outcome	Prediction		Evidence Size (k)		Jaccard Overlap	Key Insight
		ToE	LLM	ToE	LLM		
1	Survived	✓	✓	1	5	20.0%	ToE solved via single vital; LLM was cautious.
2	Survived	✓	✓	6	9	50.0%	Strong agreement on deterioration intervals.
3	Died	✓	✓	8	11	46.2%	Both flagged critical physiological decline.
4	Survived	✓	✓	8	9	30.8%	LLM prioritized stable periods differently.
5	Died	✗	✓	8	11	35.7%	Audit Win: ToE exposed model blindness to GCS.
Average	-	80% Acc	100% Acc	6.2	9.0	36.5%	ToE is ~ 30% more sparse than LLM.

Algorithm 1 Tree-of-Evidence (ToE) Inference Search

Require: Trained models; Candidates \mathcal{W}, \mathcal{N} ;
Beam width B

- 1: Compute target p_{full} and \hat{y}_{full} using all data
- 2: Cache note chunk embeddings $\{e_j\}$
- 3: $Beam \leftarrow [(\mathbf{0}^{\text{ts}}, \mathbf{0}^{\text{note}})]$
- 4: **for** $step = 1$ to S_{max} **do**
- 5: $Candidates \leftarrow \emptyset$
- 6: **for** state $(m^{\text{ts}}, m^{\text{note}})$ in $Beam$ **do**
- 7: Expand by adding one unused $w \in \mathcal{W}$
or $n \in \mathcal{N}$
- 8: Compute score(\cdot) via Eq. (10)
- 9: $Candidates \leftarrow Candidates \cup \{\text{NewState}\}$
- 10: **end for**
- 11: $Beam \leftarrow \text{Top-}B(Candidates)$
- 12: **if** $Beam[0]$ meets thresholds $\tau_{\text{conf}}, \tau_{\text{suff}}$ **then**
- 13: **return** $Beam[0] \triangleright$ Minimal Sufficient Set
- 14: **end if**
- 15: **end for**
- 16: **return** Best state in $Beam$

A.7.2 Probability-Space vs. Logit-Space Stability

Table A5 compares probability-space and logit-space definitions of the stability term $S(\mathbf{m})$ across evidence budgets (E1: Hospital Mortality, MIMIC-IV).

Probability-space stability yields 44% lower fidelity MAE at $k=5$. Notably, logit-space MAE plateaus at ~ 0.05 regardless of budget, whereas probability-space MAE continues decreasing with more evidence. This pattern is confirmed on eICU.

Table A5: Probability-space vs. logit-space stability. Probability-space achieves consistently lower fidelity MAE.

k	Space	AUROC	Fid. MAE	Comp.
1	Probability	0.755	0.090	0.029
	Logit	0.749	0.100	0.040
5	Probability	0.773	0.030	0.097
	Logit	0.768	0.054	0.131
12	Probability	0.773	0.014	0.130
	Logit	0.769	0.051	0.189

Table A6: Optimality gap: ToE vs. exhaustive search.

Dataset	k	ToE AUROC	Exhaustive	Gap
MIMIC-IV	1	0.7550	0.7543	+0.0007
MIMIC-IV	3	0.7706	0.7697	+0.0009
LEMMA-RCA	all	0.7181	0.7252	-0.0071

A.7.3 Optimality Gap Analysis

To assess how close beam search comes to the global optimum, we compare ToE against exhaustive enumeration at small k where enumeration is tractable (Table A6).

At $k=1$ and $k=3$, ToE matches the global optimum (gap < 0.001 AUROC, within bootstrap confidence intervals). Exhaustive search becomes infeasible for $k \geq 5$ ($> 1\text{M}$ subsets per patient at $k=5$ on MIMIC-IV).

A.8 Spurious Feature Injection

To test whether ToE’s search behavior can detect model reliance on spurious features, we retrain the model with a deliberately spurious binary feature that is 80% correlated with mortality in the training set but has 0% correlation in the test set.

The corrupted model requires $4.5\times$ more evidence to converge ($p < 0.001$) and has half the convergence rate (46% vs. 93%). Within the corrupted model, the asymmetry between flag=1 and

Table A7: Hyperparameter sensitivity summary.

Hyperparameter	Range	Dataset(s)	Key Finding
Stability space	Prob vs. Logit	MIMIC + eICU	Prob-space 44% lower MAE
τ_{suff}	.70, .80, .90, .95	eICU	Higher \rightarrow better fidelity
λ (stability)	0, 1.0	MIMIC	$\lambda=0$ doubles MAE

flag=0 patients (55% vs. 37% convergence) reveals the specific source of bias. This confirms that ToE’s search difficulty is a reliable signal for detecting spurious model reasoning.

A.9 Hyperparameter Sensitivity

Table A7 summarizes sensitivity to key hyperparameters beyond the λ/μ ablation reported in the main paper (Table 6).

At fixed evidence budgets, stopping thresholds ($\tau_{\text{conf}}, \tau_{\text{suff}}$) have zero effect since they only control dynamic stopping. The method is robust to stability-space choice and threshold values but sensitive to λ , which is a core design parameter.

A.10 Comparison with Zero-Shot LLM Evidence Selection.

To contextualize ToE evidence selection against a strong "clinician-like" baseline, we compare ToE to a zero-shot LLM that selects hourly evidence windows from the same 24-hour observation period. We summarize results on ICU stays where both methods produce an explicit set of windows.¹ Across these cases, ToE achieves higher predictive accuracy than the LLM (0.655 ± 0.064 vs. 0.619 ± 0.070), while also selecting similarly small evidence sets on average (5.0 ± 0.0 vs. 4.9 ± 1.1 windows). To quantify agreement between the two evidence traces, we compute Jaccard similarity between the selected window sets. Agreement remains modest overall, with a mean Jaccard similarity of 0.125 ± 0.106 for time-series evidence, 0.310 ± 0.462 for clinical notes, and 0.112 ± 0.090 when combining all modalities. Table A4 highlights five illustrative patients (not representative of aggregate Jaccard, which is lower — see above): ToE selects fewer evidence windows while maintaining non-trivial overlap in these cases (mean Jaccard 0.365). Importantly, when the model is wrong, ToE’s trace remains valuable because it reveals *which evidence the model actually relied on*, enabling targeted auditing.

¹We use "evidence size" to denote the number of selected hourly windows. For ToE, this corresponds to the final evidence set returned by the beam search. For the LLM, this corresponds to the set of windows it explicitly marked as supporting evidence.

A critical divergence occurred with Patient 5 (Died), where the LLM correctly predicted "High Risk" by identifying persistent neurological failure (Glasgow Coma Scale (GCS) 8-10). In contrast, ToE faithfully revealed that the underlying EB model predicted "Low Risk" because it prioritized stable respiratory signals (SpO2 ~100%) and ignored the GCS trajectory. This failure case highlights the danger of using LLMs as explanations: the LLM "hallucinated" a correct reasoning path that the model did not actually use. ToE, by contrast, successfully exposed the model’s blind spot regarding neurological status.