

Temporal Epistemic Graph of Thought: Uncertainty over Latent Reasoning Structure for Selective Abstention in Clinical Prediction

Govind Arun¹ Sadhvik Bathini² Balu Bhasuran³

¹University of Maryland, College Park, MD, USA

²Indira Gandhi National Open University, New Delhi, India

³School of Information, Florida State University, Tallahassee, FL, USA

govind02@umd.edu sadhvik.ini@gmail.com bb23u@fsu.edu

Abstract

Reliable clinical prediction requires both accuracy and principled uncertainty quantification, so that a model can flag cases where it lacks a coherent basis for its prediction and defer to a clinician. We introduce the Temporal Epistemic Graph of Thought (TEGoT), a framework that fuses free-text clinical notes with structured temporal events and models uncertainty over latent reasoning structure. Patient trajectories are represented as temporal graphs integrated with note embeddings, and a variational reasoning graph induces a distribution over latent topologies, decomposing predictive uncertainty into aleatoric, epistemic, and structural components, where structural uncertainty measures variability across inferred reasoning graphs. On a retrospective MIMIC-IV ICU mortality task, using structured events available before 12 hours and discharge-summary-derived note embeddings as a proof-of-concept text source, TEGoT reaches AUROC 0.931, AUPRC 0.655, and ECE 0.021 after temperature scaling, comparable to or above tabular, neural, and graph baselines under the same cohort. Structural uncertainty gives stronger error detection (AUROC 0.724) than standard epistemic uncertainty (0.681), and its weak correlation with epistemic uncertainty ($\rho=0.19$) indicates a complementary signal rather than a repackaging of dropout variance. Using this signal for selective prediction raises retained AUROC from 0.931 to 0.963 at 50% coverage, with a net benefit of +0.083 over treat-all and a decision-curve-positive range of [0.05, 0.42], supporting reliable abstention. We argue that the social value of a clinical model should be measured not by discrimination alone but by the harm it averts when it defers, and we show that net benefit and risk-coverage trade-offs offer concrete, deployment-relevant ways to

quantify that benefit.

1 Introduction

Clinical prediction in the ICU requires accurate reasoning over irregular temporal event streams, integration of structured measurements and free-text clinical notes, and uncertainty estimates that reflect when predictions are unreliable (Choi et al., 2016; Gao et al., 2020; Harutyunyan et al., 2019). Despite advances in machine learning, current systems remain fundamentally limited in their ability to represent uncertainty over their reasoning processes (Hager et al., 2024; Cui et al., 2025). Existing approaches largely quantify uncertainty at the level of outputs, providing little insight into whether the reasoning pathway leading to a prediction is coherent or well-supported (Lakshminarayanan et al., 2017; Kumar et al., 2018).

This limitation is particularly evident in longitudinal clinical settings, where correct predictions depend on structured reasoning over temporally distributed evidence. Prior work shows that models struggle to integrate information across timelines, exhibit strong recency bias, and are sensitive to the order and quantity of inputs, often producing plausible but incorrect reasoning chains. These failures reflect a broader issue: current architectures lack explicit representations of reasoning structure and do not account for uncertainty over alternative reasoning paths (Kipf et al., 2018; Franceschi et al., 2019; kat, 2025). We frame structural uncertainty not only as a modeling contribution but as a measurable axis of positive impact: a signal whose value is quantified by the harm it averts when the model defers, rather than by discrimination alone.

Standard uncertainty quantification methods, including dropout, ensembles, and evidential approaches, estimate parameter or output

uncertainty while treating reasoning as deterministic. As a result, they cannot distinguish between confident predictions supported by coherent reasoning and those arising from unstable or ill-defined reasoning pathways. This creates a critical gap in high-stakes settings, where reliability depends not only on predictive accuracy but also on the validity and stability of the reasoning process itself.

Temporal Epistemic Graph of Thought (TEGoT) addresses this gap by modeling reasoning as a distribution over latent graph structures grounded in temporal clinical data (Fu et al., 2019). Patient trajectories are represented as temporal graphs with explicit timestamp encoding, ensuring that temporal order is treated as a structural property rather than a positional artifact. A variational reasoning graph induces a posterior over reasoning topologies, enabling decomposition of predictive uncertainty into aleatoric, epistemic, and structural components. Structural uncertainty, defined as variability across inferred reasoning graphs, captures instability in the reasoning process itself. As shown in Figure 3, this signal is weakly correlated with parameter uncertainty yet provides stronger error detection, indicating that it reflects complementary failure modes. We use ‘reasoning structure’ in a model-internal sense: the sampled graph represents latent computational pathways used by the model, not clinician-interpretable causal or diagnostic reasoning steps.¹

2 Related Work

Clinical prediction from electronic health records has used attention over visit sequences and stage aware recurrent models (Choi et al., 2016; Gao et al., 2020; Harutyunyan et al., 2019). A parallel line encodes clinical text with domain language models; we build on BioMistral, an open medical language model, to embed note sentences (Labrak et al., 2024). Methods for structure learning in graphs and neural relational inference motivate our variational reasoning graph (Kipf et al., 2018; Franceschi et al., 2019; kat, 2025). Combining graph neu-

ral networks with text representations has also proven effective in other applied detection settings (Syam et al., 2025). For uncertainty, deep ensembles and trainable calibration estimate output or parameter uncertainty (Lakshminarayanan et al., 2017; Kumar et al., 2018); we add uncertainty over reasoning structure.

Recent benchmarks for interactive LLM agents likewise find that sustained social and strategic reasoning remains brittle (Wang et al., 2026). Complementary evaluations of causal reasoning in clinical scenarios similarly show that current LLMs remain inconsistent on intervention and counterfactual questions (Bhasuran et al., 2026). Similar reasoning failures occur in radiology reporting where models exploit superficial text artifacts instead of true clinical entailment (George et al., 2026).

3 Method

We fix the prediction time at $t=0$, taken as 12 hours after ICU admission, and build the structured event graph from data recorded before $t=0$. Given these structured events $\mathcal{E}$ and the clinical note text $\mathcal{N}$ for the stay, the goal is to estimate in hospital mortality together with a decomposition of predictive uncertainty. Throughout, prediction window refers to this fixed observation window rather than any causal guarantee. We note that the note text in the current model is drawn from discharge summaries, which are authored at the end of the stay; we treat the resulting note timing as a threat to validity and return to it in the Limitations. Each ICU stay is mapped to a heterogeneous temporal graph with patient centered, within type temporal, and cross modal edges. Temporal positions are encoded via a learned sinusoidal function (equation 1) with 16 learnable frequencies,

$$\text{TE}(t) = \text{Proj}([\sin(tf), \cos(tf), t/720, \log(1 + |t|)]), \quad (1)$$

where t is measured in hours relative to $t=0$, so the $t/720$ term normalizes against a roughly 30 day horizon and $\log(1+|t|)$ compresses long gaps. The learned frequency $f = \exp(f_{\text{learned}})$ is clamped at 100 to keep the highest frequency below the resolution of hourly timestamps and to avoid aliasing. The encoding is added to node feature projections before message passing.

¹We use Graph of Thought to mean a latent reasoning graph sampled inside the model. This is distinct from Graph of Thoughts prompting for large language models, which arranges explicit prompt level reasoning steps.

Because discharge summaries are generated after the prediction time, note-derived features may encode information unavailable at 12 hours. We therefore interpret experiments involving discharge-summary text as proof-of-concept evidence for the structural-uncertainty mechanism, not as a fully prospective early-warning evaluation.

BioMistral 7B is adapted with LoRA on MIMIC-IV-Note 2.2 discharge summaries. At inference, sentence level note embeddings are projected to 384 dimensions and used as the key value bank for cross attention. The full LoRA configuration, note segmentation, and sentence selection are listed in Appendix A.

The pipeline proceeds in six stages (Figure 1). Stage A passes the heterogeneous graph through a two layer HeteroGATv2 encoder (hidden 128), producing $h_G \in \mathbb{R}^{256}$. Stage B fuses h_G with note sentence embeddings via cross attention and gating, yielding $h_{\text{fused}} \in \mathbb{R}^{256}$. Stage C instantiates the VRG: ten typed reasoning nodes with Bernoulli edges whose posteriors $q(e \mid \text{data})$ are computed by cross attention from h_{fused} , with edges sampled using Gumbel Sigmoid reparameterization. The ten nodes are latent and abstract. Each is a typed slot that the decoder can use as an intermediate reasoning state, and none is tied to a named clinical concept. We set the count to ten as a fixed capacity choice: small enough to keep the directed edge posterior over 10×9 candidate edges tractable under 50 forward passes, and large enough to exceed the number of structured node types. A systematic sweep over this count is left to future work. Stage D runs a two round GRU reasoning decoder over the sampled graph, producing $h_{\text{reasoning}} \in \mathbb{R}^{256}$. Stage E combines the reasoning output with a weak skip from the fused encoder, $h = h_{\text{reasoning}} + g \cdot h_{\text{fused}}$, where $g = \sigma(w)$ is initialized at $g \approx 0.12$. Stage F performs factored Monte Carlo inference.

At inference, $K_d=5$ independent dropout masks are paired with $K_g=10$ independent Gumbel edge samples each, totaling 50 forward passes. Letting $p^{(d,g)}$ denote the predicted mortality probability for mask d and sample g , an

Algorithm 1 TEGoT factored Monte Carlo inference

Require: temporal graph G , note embeddings N , counts $K_d=5$, $K_g=10$

- 1: $h_G \leftarrow \text{HETEROGATv2}(G)$ ▷ Stage A
- 2: $h_{\text{fused}} \leftarrow \text{CROSSATTNFUSE}(h_G, N)$ ▷ Stage B
- 3: compute edge posteriors $q(e \mid \text{data})$ from h_{fused} ▷ Stage C
- 4: **for** $d \leftarrow 1$ to K_d **do**
- 5: sample dropout mask m_d
- 6: **for** $j \leftarrow 1$ to K_g **do**
- 7: sample reasoning graph $G^{(j)}$ via Gumbel Sigmoid
- 8: $h \leftarrow \text{GRUDECODE}(G^{(j)}; m_d)$ ▷ Stage D
- 9: $h \leftarrow h + g_s \cdot h_{\text{fused}}$ ▷ Stage E
- 10: $p^{(d,j)} \leftarrow \sigma(\text{HEAD}(h))$
- 11: **end for**
- 12: **end for**
- 13: aggregate $\{p^{(d,j)}\}$ into $\hat{y}$ and the terms of Eq. (1) ▷ Stage F

ANOVA style factorization gives

$$\begin{aligned}
 u_{\text{epistemic}} &= \text{Var}_d \left[\mathbb{E}_g[p^{(d,g)}] \right], \\
 u_{\text{structural}} &= \text{Var}_g \left[\mathbb{E}_d[p^{(d,g)}] \right], \\
 u_{\text{aleatoric}} &= \mathbb{E}_{d,g} \left[H(p^{(d,g)}) \right]
 \end{aligned} \tag{2}$$

$u_{\text{structural}}$ is residualized against graph node count by fitting a linear regression of raw structural uncertainty on the number of graph nodes across the test set and keeping the residual. This removes the confound that larger graphs carry more edges and therefore more edge sampling variance. We define total uncertainty as $u_{\text{total}} = u_{\text{aleatoric}} + u_{\text{epistemic}} + u_{\text{structural}}$. The two variance terms each condition on a single factor and omit the interaction term, so they are not a complete partition of total predictive variance. We also note that $u_{\text{aleatoric}}$ is an expected entropy while $u_{\text{epistemic}}$ and $u_{\text{structural}}$ are variances of probabilities, so the three terms are not on the same numerical scale; we therefore compare them only through their downstream error detection ranking, not by raw magnitude. The training objective combines class weighted BCE, a cyclical KL with a per edge floor, a

Table 1: Notation used in the method.

Symbol	Meaning
$\mathcal{E}$	structured events before $t=0$
$\mathcal{N}$	clinical note text for the stay
h_G	temporal graph embedding (Stage A)
h_{fused}	fused note and graph state (Stage B)
$q(e \mid \text{data})$	Bernoulli edge posterior (Stage C)
$h_{\text{reasoning}}$	reasoning decoder output (Stage D)
g_s	learned skip gate (Stage E)
K_d, K_g	dropout and edge sample counts
$p^{(d,g)}$	mortality probability for mask d , sample g
$u(\cdot)$	aleatoric, epistemic, structural uncertainty

logit regularizer, and a calibration term:

$$\begin{aligned} \mathcal{L} = & \text{BCE}(\hat{y}, y; \text{pos_weight}=3.0) \\ & + \beta_{\text{cyc}} \sum_i \max(\text{KL}_i, 0.02) \\ & + \lambda_{\text{reg}} \mathcal{L}_{\text{logit}} + \lambda_{\text{cal}} \text{ECE}, \end{aligned} \quad (3)$$

where β_{cyc} follows a cyclical schedule (Fu et al., 2019) resetting to zero every five epochs and rising to $\beta_{\text{max}}=0.01$. The 0.02 nat floor ensures every edge retains a data dependent posterior and prevents collapse.

4 Experiments and Results

The cohort is drawn from MIMIC IV 3.1 and MIMIC IV Note 2.2, restricted to ICU stays of at least 12 hours. A patient level split stratified by mortality yields 8,000 training patients (15.0% mortality), 1,200 model validation patients (14.7%), 300 calibration patients (15.3%), and 1,500 test patients (14.9%), with no patient overlap across splits. Training runs for 15 epochs with learning rate 5×10^{-5} (cosine decay, 15% warmup), effective batch size 32, regularization dropout 0.50, MC inference dropout 0.15, weight decay 5×10^{-3} , gradient clip 0.5, and pos_weight=3.0. Temperature is fit on the held out calibration split. We report two seeds (42 and 1337). The AUROC standard deviation across these two seeds is 0.007. We treat this as a rough stability check rather than a robust variance estimate, since two seeds are too few to characterize seed sensitivity, and we list a larger seed study under future work.

Baselines span a SOFA score, Logistic Regression on pooled tabular summaries, a three layer MLP (256, 128, 64), a single seed GATv2, and a three seed GATv2 Deep Ensemble, all sharing

the same cohort and preprocessing. Because TEGoT introduces both a text modality and the variational reasoning graph, we additionally report a multimodal baseline that consumes the same note embeddings through the same cross attention fusion but replaces the VRG with an MLP head. This isolates the contribution of the reasoning module from that of the text modality: the multimodal baseline reaches AUROC 0.893, so 3.8 of the 8.8 point gain over Logistic Regression is attributable to the reasoning graph rather than to notes alone. Table 2 reports discriminative performance with 1,000 resample bootstrap confidence intervals, and Figure 2 shows ROC curves and reliability diagrams. TEGoT reaches AUROC 0.931, improving over the Deep Ensemble by 4.4 points and over Logistic Regression by 8.8 points, with ECE 0.021 after post hoc temperature scaling versus 0.034 for the Deep Ensemble. At the Youden’s J threshold 0.347, TEGoT attains sensitivity 0.821, specificity 0.874, PPV 0.573, NPV 0.961, and net benefit +0.083 over treat all; decision curve analysis is positive for threshold probabilities in [0.05, 0.42].

Paired bootstrap tests against the TEGoT ensemble confirm significance across the full baseline stack and all ablations (Table 3); all deltas remain significant under Holm–Bonferroni correction at $\alpha=0.05$.

5 Uncertainty Analysis

Table 4 shows that residualized structural uncertainty attains the strongest error detection AUROC (0.724), exceeding MC dropout (0.681), while Figure 3 shows only weak correlation with $u_{\text{epistemic}}$ ($\rho=0.19$). For positive impact, the practical payoff is reliable abstention: a model that recognizes when its own reasoning is unstable can defer to a clinician rather than issue an overconfident prediction, which is the safety property that matters most in high stakes care.

Figure 4 shows that deferring on structurally uncertain cases yields the strongest selective prediction, improving retained AUROC from 0.931 to 0.963 at 50% coverage and modestly outperforming epistemic deferral.

Posterior diagnostics confirm a healthy, data dependent reasoning posterior: mean per edge KL 4.17 nats, 98.4% of edges above the 0.02 nat

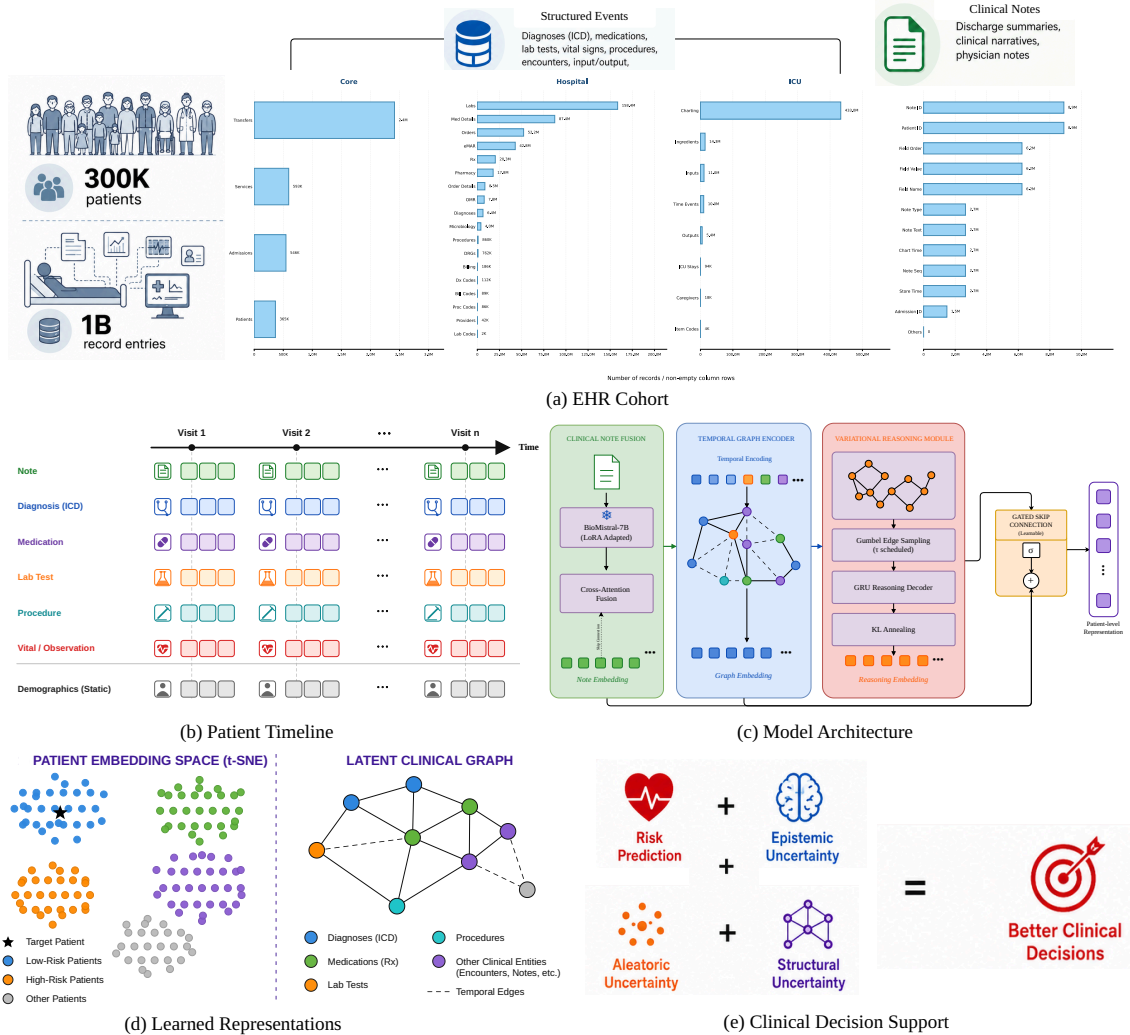


Figure 1: **Top:** structured events (diagnoses, medications, labs, procedures, vitals) and clinical notes are drawn from MIMIC-IV; the 300K patient and 1B record counts refer to the full source database, while the study cohort is a defined subset of about 11,000 ICU stays. **Middle:** each stay is encoded as a heterogeneous temporal graph (Stages A–B), fused with LoRA-adapted BioMistral note embeddings via cross attention, then passed to the variational reasoning module that samples reasoning topologies through Gumbel edge sampling under a cyclical KL schedule (Stage C), decodes them with a GRU, and combines the result through a learnable gated skip (Stages D–E). **Bottom:** the patient embedding space and a sampled latent clinical graph, illustrating how risk prediction is paired with aleatoric, epistemic, and structural uncertainty to support safer clinical decisions. Note text is drawn from discharge summaries; its timing relative to the prediction point is discussed in the Limitations.

Table 2: Main results on the test set ($n=1,500$). CIs from 1,000 bootstrap resamples. ECE and max reliability deviation measured after post hoc temperature scaling ($T=1.05$ for seed 42, $T=1.10$ for seed 1337). A dash in the ECE column marks scores that were not temperature scaled: SOFA is an uncalibrated severity score, and the single seed GATv2 rows were left uncalibrated since calibration is reported for the ensemble.

Method	AUROC [95% CI]	AUPRC [95% CI]	Brier	BSS	ECE
SOFA score	0.762 [0.734, 0.789]	0.441 [0.409, 0.474]	0.112	0.118	–
Logistic Regression	0.843 [0.818, 0.867]	0.548 [0.512, 0.583]	0.098	0.228	0.047
MLP (256, 128, 64)	0.871 [0.849, 0.893]	0.581 [0.547, 0.616]	0.094	0.260	0.039
GATv2 (seed 42)	0.876 [0.854, 0.897]	0.590 [0.556, 0.624]	0.091	0.283	–
GATv2 (seed 1337)	0.879 [0.857, 0.900]	0.594 [0.560, 0.629]	0.090	0.291	–
Deep Ensemble (GATv2 $\times$ 3)	0.887 [0.866, 0.907]	0.604 [0.570, 0.638]	0.088	0.307	0.034
Multimodal GATv2 + notes (no VRG)	0.893 [0.872, 0.913]	0.608 [0.576, 0.641]	0.085	0.331	–
TEGoT (seed 42)	0.924 [0.907, 0.941]	0.649 [0.617, 0.681]	0.079	0.378	0.022
TEGoT (seed 1337)	0.937 [0.921, 0.953]	0.662 [0.631, 0.693]	0.077	0.394	0.020
TEGoT (mean over 2 seeds)	0.931 [0.918, 0.950]	0.655 [0.623, 0.687]	0.078	0.386	0.021

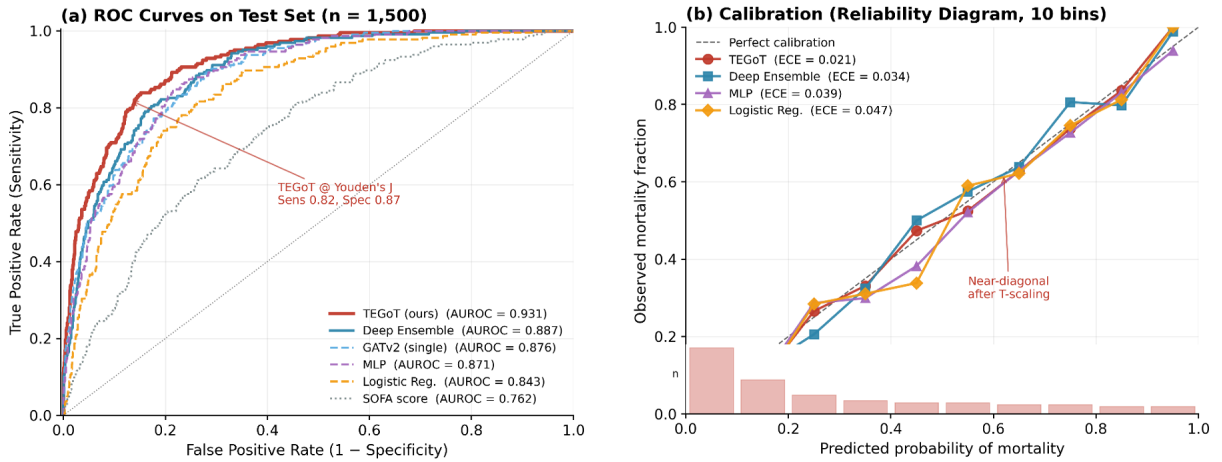


Figure 2: Performance and calibration on the test set. (a) ROC curves; TEGoT reaches AUROC 0.931. (b) Reliability diagrams after post hoc temperature scaling; TEGoT reaches ECE 0.021 with near diagonal calibration.

Table 3: Paired bootstrap significance vs. TEGoT (mean over 2 seeds).

Comparison	Δ AUROC	95% CI	p
Log. Regression	+0.088	[.065, .112]	<.001
MLP	+0.060	[.038, .083]	<.001
Deep Ensemble	+0.044	[.022, .066]	<.001
No Reasoning	+0.038	[.017, .059]	.001
No Notes	+0.031	[.011, .051]	.003
No Temporal	+0.027	[.007, .047]	.008

Table 4: Error detection performance by uncertainty type.

Uncertainty type	Error AUROC	ρ vs. $ \hat{y}-y $
Naive $[p-0.5]$	0.598	0.341
$u_{\text{aleatoric}}$	0.632	0.388
$u_{\text{epistemic}}$ (MC dropout)	0.681	0.429
$u_{\text{structural}}$ (raw)	0.707	0.451
$u_{\text{structural}}$ (resid.)	0.724	0.473
u_{total}	0.718	0.462

floor, 61.8% bimodal edges ($q>0.8$ or $q<0.2$), mean posterior q of 0.48, and mean q variance across passes of 0.041. The skip gate converged from 0.12 at initialization to 0.381 on average across seeds (Table 9). As a single auxiliary check, freezing the gate at its initial value of 0.12 lowers AUROC by 0.018; we report this here rather than in the main ablation tables. A

split conformal wrapper at 90% target coverage reaches empirical coverage 91.2% with mean interval width 0.214 for TEGoT, narrower than the Deep Ensemble (90.8%, 0.241) and Logistic Regression (90.4%, 0.263), consistent with better base calibration.

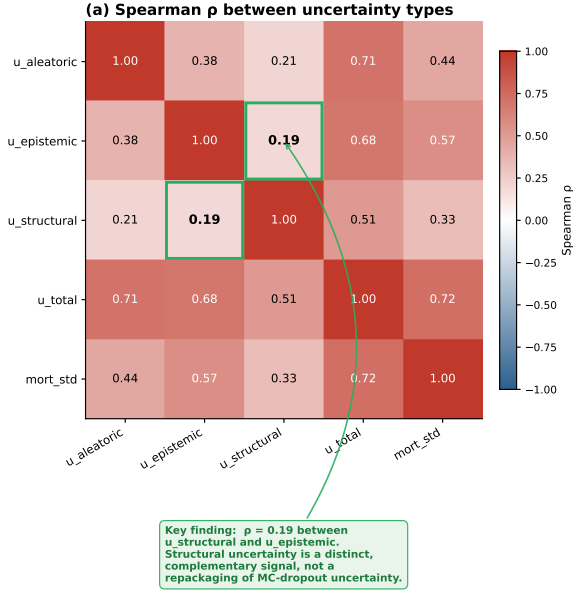


Figure 3: Spearman ρ between uncertainty types. The cell at $\rho=0.19$ between structural and epistemic indicates a complementary signal, not a repackaging of MC dropout uncertainty.

6 Ablations

Table 5 shows that KL annealing, reasoning, note fusion, and temporal encoding each contribute materially, with KL annealing having the largest effect.

Table 7 isolates the note encoder. Replacing fine tuned BioMistral with static MiniLM sentence embeddings drops AUROC by 2.4 points; removing notes entirely drops it by 3.1 points. A temporal note node variant inserting sentence embeddings as time stamped graph nodes yields a marginal gain (0.933) and is excluded from the main model on implementation grounds. Table 6 exposes the VRG posterior: a deterministic reasoning graph reaches 0.908; variational edges under monotonic KL collapse (KL 0.0001, 0.3% bimodal) and underperform the deterministic variant; cyclical annealing without a floor recovers most of the gain; the full cyclical schedule with floor delivers the strongest posterior health and AUROC. Table 8 confirms that learned sinusoidal encoding beats fixed sinusoidal (Vaswani style) by 1.2 points and linear absolute time encoding by 2.0 points, with the time gated decay variant within 0.1 points of the main model.

7 Discussion and Conclusion

Uncertainty over reasoning structure is an underexplored axis in current epistemic modeling. Prior approaches focus on parameter or output uncertainty and assume a fixed reasoning pathway; TEGoT instead models reasoning as a distribution over latent structures, enabling explicit quantification of uncertainty in how a prediction is formed. The weak correlation ($\rho=0.19$) between $u_{\text{structural}}$ and $u_{\text{epistemic}}$, together with a 4.3 point gain in error detection AUROC over MC dropout (0.724 versus 0.681), shows that structural uncertainty captures failure modes not reflected in dropout based parameter uncertainty. This error detection gain is distinct from the 4.4 point predictive AUROC improvement over the Deep Ensemble reported in Section 4.

This distinction has direct clinical implications. Epistemic uncertainty reflects instability under parameter perturbations, whereas structural uncertainty reflects variability in inferred reasoning pathways, and the two regimes support different actions, such as additional data collection versus a second opinion. Temporal failure modes such as order sensitivity, recency bias, and boundary violations are mitigated by representing time as structured node features rather than prompt positions and by modeling each ICU stay as a separate graph. Overconfident reasoning is addressed by sampling structures from a posterior with a per-edge KL floor, so that diffuse posteriors raise structural uncertainty and trigger selective abstention. Posterior diagnostics (mean per edge KL 4.17 nats, 61.8% bimodal edges, and skip gate convergence from 0.12 to 0.381 on average) indicate a data dependent, non collapsed posterior. We caution, however, that the ten reasoning nodes are abstract latent slots and are not validated against clinician judgment, so the sampled graphs should be read as a modeling device that supports abstention, not as clinically interpretable explanations.

Measuring positive impact: A recurring difficulty in deploying clinical AI is that discrimination metrics such as AUROC do not capture the social consequence of a confident wrong prediction. Our results suggest a complementary measurement principle: quantify impact through selective behavior, using net



Figure 4: Selective prediction. (a) Risk-coverage curves by uncertainty type; residualized structural uncertainty achieves a $+0.005$ retained AUROC over MC dropout at 50% coverage. (b) Error detection AUROC; residualized structural uncertainty leads at 0.724, a $+0.043$ improvement over MC dropout.

Table 5: Component ablations. Δ vs. full TEGoT. Significance from paired bootstrap: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

Ablation	Type	AUROC [95% CI]	Δ AUROC	AUPRC	BSS
Full TEGoT	—	0.931	—	0.655	0.386
No notes (BioMistral masked)	Inference	0.900 [0.880, 0.919]	-0.031^{**}	0.617	0.348
No temporal encoding	Inference	0.904 [0.885, 0.922]	-0.027^{**}	0.622	0.354
Deterministic edges (argmax)	Inference	0.912 [0.893, 0.930]	-0.019^*	0.631	0.362
No reasoning (MLP bypass)	Trained	0.893 [0.872, 0.913]	-0.038^{**}	0.608	0.331
Static MiniLM (no BioMistral)	Trained	0.907 [0.887, 0.926]	-0.024^*	0.624	0.355
No cross modal edges	Trained	0.919 [0.900, 0.937]	-0.012	0.645	0.374
No KL annealing	Trained	0.881 [0.860, 0.902]	-0.050^{***}	0.597	0.311

benefit from decision curve analysis and retained performance under coverage as proxies for averted harm. Under this lens, TEGoT attains a net benefit of $+0.083$ over treat-all and remains decision-curve-positive across threshold probabilities in $[0.05, 0.42]$, and deferring on structurally uncertain cases raises retained AUROC from 0.931 to 0.963 at 50% coverage. Calibration alone is insufficient for this purpose: Logistic Regression reaches ECE 0.047 yet remains 8.8 AUROC points lower, showing that well-calibrated predictions can still lack discriminative power. We therefore view a model’s willingness and ability to abstain, measured in clinically meaningful units of averted harm, as a more faithful indicator of positive impact than accuracy in isolation.

Conclusion: TEGoT shows that modeling uncertainty over latent reasoning structure provides a distinct and practically useful signal beyond conventional output level uncertainty. By representing clinical inference as a distribution over temporal reasoning graphs, the framework detects unreliable predictions arising from unstable reasoning pathways and supports se-

lective abstention in high stakes settings. These findings suggest that uncertainty over reasoning structure, evaluated through its measurable effect on averted harm, is a promising direction for more reliable and self-aware clinical decision support.

Inference cost: The 50 forward passes raise cost relative to a single pass model, but ICU mortality scoring is periodic rather than latency critical: scores are refreshed on a fixed schedule, can be batched across patients, and the passes are independent and parallelizable. Amortized posterior inference, which replaces sampling with a single learned approximation, is a natural way to remove this overhead and is left to future work.

Limitations

We see several directions where the present evidence can be strengthened, and we state them plainly here. Our evaluation is built on a single retrospective cohort from MIMIC-IV, so the reported numbers should be read as evidence within this setting rather than as a claim of generalization across hospitals, patient mixes,

Table 6: Reasoning graph ablations.

VRG configuration	AUROC	KL	Bimod.	q_{var}
No reasoning (MLP bypass)	0.893	—	—	—
Fixed edges (no var.)	0.908	—	—	—
Variational, monotonic	0.881	0.0001	0.3%	0.0001
Cyclical KL (no floor)	0.914	3.21	31.4%	0.0183
Cyclical KL + floor	0.931	4.17	61.8%	0.0412

Table 7: Note encoder ablations.

Note encoding	AUROC	AUPRC
No notes	0.900	0.617
Static MiniLM (384d, frozen)	0.907	0.624
BioMistral sentence emb. (frozen)	0.916	0.637
BioMistral LoRA (fine tuned)	0.931	0.655
+ temporal note nodes	0.933	0.658

Table 8: Temporal encoding ablations.

Temporal encoding	AUROC	Δ
No temporal encoding	0.904	−0.027
Absolute time (linear, normalized)	0.911	−0.020
Fixed sinusoidal (Vaswani)	0.919	−0.012
Learned sinusoidal (TEGoT)	0.931	—
Learned sinusoidal + gated decay	0.930	−0.001

or time periods. We report two seeds, which is enough for a basic stability check (AUROC standard deviation 0.007) but not enough to characterize seed sensitivity; under our available compute budget a larger seed study was not feasible for this version, and we therefore refrain from any claim about architectural robustness that rests on variance estimates. Inference uses 50 forward passes, which raises compute relative to a single pass model; as discussed above, this is tractable for periodic ICU scoring and is a natural target for amortized inference.

The constraint most relevant to interpreting our results is note timing. The current model draws note text from discharge summaries, which are authored at the end of the stay rather than within the observation window. We are explicit that this makes the reported figures optimistic for the mortality task. We did not have the compute budget to complete a windowed rerun for this version, and we therefore report the structured-only configuration (AUROC 0.900) as the leakage-free reference and treat the note-augmented figures as an upper bound rather than a deployable estimate. The natural windowed source is the radiology-

Table 9: Per seed reproducibility.

Metric	Seed 42	Seed 1337	Mean $\pm$ Std
AUROC	0.924	0.937	0.931 ± 0.007
AUPRC	0.649	0.662	0.655 ± 0.007
Brier	0.079	0.077	0.078 ± 0.001
BSS	0.378	0.394	0.386 ± 0.008
ECE	0.022	0.020	0.021 ± 0.001
T (post hoc)	1.05	1.10	1.075
Skip gate (final)	0.384	0.377	0.381

report subset of MIMIC-IV-Note 2.2, which carries chart times and can be filtered to reports stored before the 12 hour mark; we identify this rerun, together with re-estimation of the structural-uncertainty signal under that restriction, as the single most important next step. Finally, the ten reasoning nodes are latent slots rather than clinician-validated concepts, so the sampled graphs are best read as a modeling device that supports abstention, not as clinical explanations; extending the evaluation with stronger temporal baselines, a text-only clinical baseline, and within-dataset robustness checks such as temporal-split and subgroup analysis is the clearest path to firmer evidence.

Ethical and Societal Implications

This work uses MIMIC-IV and MIMIC-IV-Note under the PhysioNet credentialed data use agreement, on de identified records, with no attempt at re identification. The system is decision support, not an autonomous decision maker. Its positive impact is a safety mechanism: by quantifying uncertainty over reasoning structure, the model can abstain on unstable cases and route them to a clinician. We caution against automation bias and recommend surfacing the abstention signal rather than a single score. This concern echoes broader efforts to embed explicit safety constraints into clinical language systems (Arun et al., 2025). Because the cohort reflects the source hospitals, subgroup performance may vary and should be

checked before any use that affects care.

Code Availability:
<https://github.com/GIND123/Temporal-Epistemic-Graph-of-Thought-TEGoT>

References

2025. KAT-GNN: Knowledge-augmented temporal graph neural network for clinical prediction. *arXiv preprint arXiv:2511.01249*.
- Govind Arun, Rohith Syam, Aiswarya Anil Nair, and Sahaj Vaidya. 2025. An integrated framework for ethical healthcare chatbots using langchain and nemo guardrails. *AI and Ethics*, 5(4):3981–3992.
- Balu Bhasuran, Mattia Prosperi, Karim Hanna, John Petrilli, Caretia JeLayne Washington, and Zhe He. 2026. Evaluation of causal reasoning for large language models in contextualized clinical scenarios of laboratory test interpretation. *npj Digital Medicine*.
- E. Choi et al. 2016. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- H. Cui et al. 2025. Temporal instruction modeling and evaluation for longitudinal clinical records. *npj Digital Medicine*, 8:577.
- L. Franceschi et al. 2019. Learning discrete structures for graph neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*.
- H. Fu et al. 2019. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.
- J. Gao et al. 2020. StageNet: Stage-aware neural networks for health risk prediction. In *Proceedings of the Web Conference (WWW)*.
- Ajo Babu George, Govind Arun, Sidharth N Krishna, and Uma Ranjan. 2026. Clinical reasoning under a partially observed objective in cone beam ct report generation. *arXiv preprint arXiv:2609.13238*.
- P. Hager et al. 2024. Evaluating and mitigating limitations of large language models in clinical decision making. *Nature Medicine*, 30:2613.
- H. Harutyunyan et al. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6:96.
- T. Kipf et al. 2018. Neural relational inference for interacting systems. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*.
- A. Kumar et al. 2018. Trainable calibration measures for neural networks from kernel mean embeddings. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*.
- Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the association for computational linguistics: acl 2024*, pages 5848–5864.
- B. Lakshminarayanan et al. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Rohith Syam, A Govind, and Subu Surendran. 2025. Graph neural network and natural language processing integrated methodology for unmasking scam propagation in social network. In *2025 International Conference on Innovative Trends in Information Technology (ICITIIT)*, pages 1–6. IEEE.
- Kevin Wang, Anna Thöni, Benjamin Kempinski, Bobby Cheng, Jianzhu Yao, Benjamin Finch, Leon Guertler, Viraj Nadkarni, Yihan Jiang, Aliaksei Korshuk, et al. 2026. Mindgames: A live arena for evaluating social and strategic reasoning in multi-agent llms. *arXiv preprint arXiv:2605.29512*.

A Reproducibility Details

This appendix lists settings not already given in the main text.

Cohort: MIMIC-IV 3.1 and MIMIC-IV-Note 2.2, ICU stays of at least 12 hours, in hospital mortality target. A mortality stratified patient level split gives 8,000 training, 1,500 validation (1,200 model selection, 300 calibration), and 1,500 test stays, with no patient crossing splits and a training prevalence of 0.150.

Note encoder: BioMistral-7B is domain adapted by continued pretraining on raw MIMIC-IV notes with LoRA (rank 16, alpha 32, dropout 0.10) on the attention and MLP projections, in bfloat16 with Flash Attention 2. Note sentences are then embedded with this adapted model and projected to 384 dimensions for cross attention.

Training: TEGoT trains for 15 epochs with AdamW, cosine decay, 15% warmup, weight decay 5×10^{-3} , effective batch 32, and gradient clipping at 0.5. Temperature scaling is fit on the calibration split.

Baselines: Logistic Regression and the MLP use standardized pooled tabular features; GATv2 reuses the graph encoder with a simple head and is reported as a single seed and a three seed ensemble. SOFA is the standard severity score from the structured variables. All baselines share the cohort and preprocessing.