Early Risk Prediction with Temporally and Contextually Grounded Clinical Language Processing

Rochana Chaturvedi[1], Yue Zhou[2], Andrew Boyd[2], Brian T. Layden[2], Mudassir Rashid[3], Lu Cheng[3], Ali Cinar[3], Barbara Di Eugenio[2]

[1] Kellogg School of Management, Northwestern University, IL, USA
[2]University of Illinois Chicago, Chicago, IL, USA
[3]Illinois Institute of Technology, Chicago, IL, USA

Correspondence: rochana.chaturvedi@northwestern.edu

Abstract

Clinical notes in Electronic Health Records (EHRs) capture rich temporal information on events, clinician reasoning, and lifestyle factors often missing from structured data. Leveraging them for predictive modeling can be impactful for timely identification of chronic diseases. However, they present core natural language processing (NLP) challenges: long text, irregular event distribution, complex temporal dependencies, privacy constraints, and resource limitations. We present two complementary methods for temporally and contextually grounded risk prediction from longitudinal notes. First, we introduce HITGNN, a hierarchical temporal graph neural network that integrates intra-note temporal event structures, inter-visit dynamics, and medical knowledge to model patient trajectories with fine-grained temporal granularity. Second, we propose REVEAL, a lightweight test-time framework that distills LLMs' reasoning into smaller verifier models. Applied to opportunistic screening for Type 2 Diabetes (T2D) using temporally realistic cohorts curated from private and public hospital corpora, HITGNN achieves the highest predictive accuracy—especially for near-term risk—while preserving privacy and limiting reliance on large proprietary models. REVEAL enhances sensitivity to true T2D cases and retains explanatory reasoning. Our ablations confirm the value of temporal structure and knowledge augmentation, and fairness analysis shows HITGNN performs more equitably across subgroups.

1 Introduction

Modeling disease progression represents one of the most critical challenges in modern healthcare, with particularly high stakes for chronic conditions like Type 2 Diabetes (T2D), which affects half a billion people worldwide and continues rising (IDF, 2025). The rich information in free-text clinical notes available in EHRs provides an opportunity for NLP-based predictive modeling. These notes present two natural perspectives for understanding disease progression and suggest distinct yet complementary computational approaches. First, the **temporal progression** of symptoms, diagnoses, treatments, and interventions, which are natural for understanding the underlying disease pathology, requires structured modeling. Second, capturing the **semantic richness** of clinical notes calls for contextualized language understanding. These are particularly relevant for opportunistic screening, where patients with low socioeconomic status often miss routine care and timely screening (Danielson et al., 2023). In such settings, clinicians must act on all historical information available at a given visit to estimate the disease risk. Formally, given patient documents $d_1, d_2, \ldots, d_n$ ordered in a non-decreasing sequence of patient visits $1, 2, \ldots, n$, we want to estimate the risk at next visit $(n + 1)$.

To address these complementary modeling needs, this work proposes two novel methods that integrate domain-specific modeling with targeted representation learning techniques over clinical notes, optimized for low-resource settings. To address structured temporal modeling, we propose **Hi**erarchical **T**emporal **G**raph **N**eural **N**etwork (HITGNN), a dynamic model that captures patient state evolution with fine-grained temporal granularity and knowledge-enhanced structure. This fine-grained knowledge is crucial for understanding disease pathways. For example, elevated glucose **After** corticosteroid use likely indicates drug-induced hyperglycemia, not T2D, while the same glucose levels **Before** any steroid therapy may signal early T2D onset. This motivates our HITGNN framework, which leverages temporal graphs extracted from individual clinical notes using clinical temporal relation extraction (TREX) methods. TREX involves identifying entities such as clinical events (problems, treatments), time expressions

Transactions of the Association for Computational Linguistics, vol. 14, pp. 1742–1764, 2026. https://doi.org/10.1162/TACL.a.748
Action Editor: Yulan He. Submission batch: 2/2026; Revision batch: 4/2026; Published 7/2026.

or timex (e.g., dates), and their temporal relations (Before, After, and Overlap). HiTGNN augments these graphs with semantic information from a clinical knowledge base. These multi-layered event-temporal graphs are modeled with dynamic graph neural networks (GNNs) to capture patient timelines within and across visits. While GNNs are effective for relational reasoning, their performance depends on rich contextual embeddings; thus, we incorporate clinically pretrained language model (CPLM) embeddings and knowledge-graph embeddings to enhance input representations with rich semantics. This addresses the long-recognized need to incorporate causal and temporal patterns into diagnostic reasoning (Patil et al., 1981).

Complementing this structured approach, current Large Language Models (LLMs) offer strong localized and context-aware reasoning capabilities to model the underlying text semantics. However, they often struggle with long-range context integration and structured reasoning—limitations particularly pronounced in the clinical domain, where patient histories span multiple long notes. Additionally, real-world healthcare deployments face privacy and resource constraints, limiting the use of proprietary or large-scale models. Our second method, **Re**asoning with **Ve**rifier-**A**ided **L**abeling (REVEAL), is a scalable inference-time architecture preserving LLM-style reasoning in resource-constrained settings. It distills reasoning traces from a large LLM into a smaller one, scaling performance with interpretability without the full computational cost of training larger models.

Together, these complementary modeling paradigms—graph-based temporal reasoning and interpretable LLM inference—provide a holistic risk prediction framework grounded in structure and semantics. We evaluate them in the context of opportunistic screening of T2D as a representative use case, using a real private hospital (PH) corpus and a corpus curated from MIMIC-IV (Johnson et al., 2023). Additionally, due to the prevalent demographic biases (Meng et al., 2022; Zhou et al., 2025; Hall et al., 2015), we conduct a fairness analysis to better understand model behavior across demographic groups.

Contributions. (1) We present two complementary representation learning frameworks from longitudinal clinical notes, tailored for low-resource settings: (i) **HiTGNN**: the first application of clinical temporal relation extraction (TREX) for risk prediction that integrates intra-document temporal relations, inter-visit dynamics, and medical knowledge, enabling reasoning across both local event structures and longitudinal patient trajectories. (ii) **REVEAL**: an inference-time scaling framework where a smaller LLM validates predictions from a larger frozen LLM, inheriting interpretability and improving accuracy without full retraining.
(2) We demonstrate the translational value of temporally enriched representations in a real clinical application—opportunistic screening for T2D, especially in immediate-risk horizons, where intervention is most impactful.
(3) We curate rigorous datasets from private (PH) and public (MIMIC-IV) sources, excluding post-diagnosis inputs to prevent label leakage, and ensuring balanced cohorts for fair evaluation.
(4) Extensive ablations disentangle the impacts of temporal relations, KG augmentation, representation enrichment, and pipeline components, revealing corpus-dependent trade-offs.
(5) Our fairness analysis highlights how LLM-based systems exhibit instability and inconsistency in subgroup behavior, even with demographic conditioning, and that HiTGNN generally is fairer even without explicitly encoding demographic attributes.
These contributions establish a clinically grounded framework integrating concept abstraction, temporal structure, and semantic enrichment to support low-cost and privacy-conscious AI systems for real-world clinical decision support.

2 Related Literature

Computational Approaches for Disease Progression Prediction Despite significant advances in predictive modeling using electronic health records (EHR), many studies continue to rely predominantly on structured EHR data (Lipton et al., 2016; Singhal et al., 2023; Jiang et al., 2024), which is often noisy (Hersh et al., 2013) and incomplete (Capurro et al., 2014). Among efforts that leverage free-text, recent works either use only the last clinical note (Xu et al., 2023; Nguyen et al., 2024), relying on structured data to capture temporal trends; or use a coarser form of temporal modeling where each note is embedded using a language model (Huang et al., 2020), a bag-of-concepts (Chaturvedi et al., 2023), or topics (Ghassemi et al., 2015), and a BiLSTM models the representations from multiple notes across patient visits. While these studies explore various fusion strategies, a key gap

remains: no prior work models temporal relations both within a single clinical note and across multiple notes/visits to forecast long-term disease risk. Our work directly addresses this gap by: (1) leveraging the sequence of clinical notes for each patient, and (2) introducing multi-layered modeling to leverage temporal information extracted from each note and across the notes from multiple visits to track evolving health trajectories.

LLM in Healthcare The integration of Large Language Models (LLMs) into healthcare is a rapidly evolving field (Zhou et al., 2023; He et al., 2025). Key developments include impressive performance on medical question answering (QA) benchmarks (Singhal et al., 2025). However, outside of controlled QA settings, LLMs still struggle on tasks requiring complex reasoning for disease diagnosis in real clinical settings (Hager et al., 2024; Wang et al., 2024b), and often exhibit performance disparities across demographic groups (Zhou et al., 2025). Recent research has demonstrated that scaling test-time compute through verifier-guided search can significantly improve LLMs' reasoning capabilities and often outperforms simply scaling model size (Snell et al., 2025). However, scaling test-time compute of LLMs in low-resource real-world healthcare tasks is still underexplored. Our work contributes to this literature by exploring verifier-guided reasoning approaches tailored to low-resource clinical prediction tasks.

3 Data Curation

Our first dataset, the PH corpus, is curated from private data comprising adult patients (age $\geq$18 years) collected from a U.S.-based hospital between January 2010 and July 2021. We also curate a second corpus from MIMIC-IV (Medical Information Mart for Intensive Care, version 4) to study the generalizability of our methods. MIMIC-IV is a large, publicly available database comprising deidentified clinical notes from ICU encounters at Beth Israel Deaconess Medical Center in Boston from 2008–2019. We exclude diagnosis of other types of diabetes (e.g., Type-1 diabetes, gestational diabetes) using ICD codes (standardized diagnostic codes) to avoid conflating related yet distinct diagnostic trajectories. Therefore, our results focus on distinguishing chronic adult T2D—which accounts for 90–95% of diagnosed adult diabetes cases (Centers for Disease Control and Prevention, 2023)—from non-diabetes (NoD).

For patients in the NoD group, we include all available clinical notes except the last recorded encounter in the dataset, which is indicative of no diabetes. For the T2D cohort, an additional refinement step ensures that post-diagnosis data is excluded to prevent label leakage. While we use the date of the first recorded ICD code as a proxy for Type 2 Diabetes (T2D) onset, this signal is known to be noisy (Hersh et al., 2013) and often delayed relative to when the diagnosis is first documented in clinical notes. To mitigate this, we additionally process the T2D cohort's notes using a large language model (LLM) to identify the earliest explicit mention of a T2D diagnosis in unstructured text. This kind of additional filtering is crucial for ensuring realistic model evaluation. For example, Zhang et al. (2024) show that a widely used sepsis prediction model frequently issues risk alerts only after clinicians have documented suspicion, undermining its practical utility. A substantial fraction of patients show a mismatch between the recorded diagnosis date and an explicit textual mention of T2D across PH ($585/2{,}302 \approx 25\%$) and MIMIC-IV ($2{,}298/5{,}207 \approx 44\%$). Manual review of 150 notes from 30 randomly sampled patients in the T2D cohort of the final datasets found 1 violation in the PH corpus and no violations in MIMIC-IV.[1] Review of excluded notes shows perfect accuracy in identifying prior or current T2D diagnoses.

Finally, we construct a demographically matched test set for fair evaluation. We estimate T2D propensity scores (Rosenbaum and Rubin, 1985) from demographic variables, including age, gender, and race (and ethnicity in the case of the PH corpus; this is combined with Race in MIMIC-IV). This is followed by 1:1 greedy nearest-neighbor matching without replacement to select NoD controls.[2] The final PH corpus comprises 3332 patients (712 in the test set), and the final MIMIC-IV subset contains 5802 patients (291 in the test set).

Both datasets pose unique strengths and limitations: PH contains richer longitudinal histories with diverse note types, ideal for chronic disease modeling, but it is not public. In contrast, while MIMIC-IV is one of the most widely used, large

[1]The LLM frequently handles borderline cases by correctly inferring the absence of a T2D diagnosis from contexts such as negated mentions (e.g., "denies DM"), risk states ("pre-DM", "borderline DM"), action plans ("DM screening", "A1c ordered"), family history of diabetes, and alternate medication use (e.g., "metformin" prescribed for "morbid obesity").

[2]The data filtering details are provided in Appendix A.

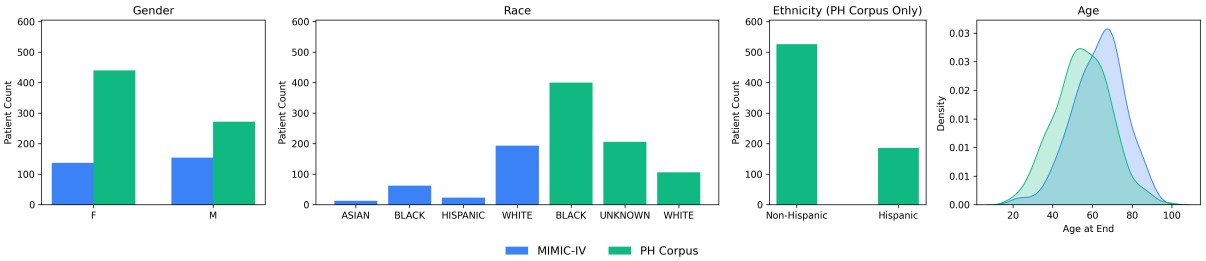


Figure 1: Demographic distribution in PH and MIMIC-IV test sets.

public datasets, it only focuses on ICU notes, omitting much of the hospitalization timeline, and has narrower temporal windows (84.5% patient records comprise single visits (Cui et al., 2025)). Further, MIMIC-IV notes are de-identified, lacking an essential temporal anchor—DATE—considered protected health information (PHI).[3] Figure 1 shows demographic differences: *females* are the majority in PH but a minority in MIMIC-IV; *Black* is the majority in PH, *White* in MIMIC-IV, with Asian as the minority. PH corpus records Hispanics under a separate ethnicity attribute. Therefore, evaluating methods on both datasets is important. However, with MIMIC-IV's limited temporal continuity, performance on this dataset likely represents a lower bound compared to richer data settings.

4 HiTGNN: Disease Prediction using Hierarchical Temporal Graphs

HiTGNN is a **Hi**erarchical **T**emporal **G**raph **N**eural **N**etwork for modeling fine-grained temporal information and external knowledge to predict T2D risk. We extract and refine temporal graphs from clinical notes (sections 4.1–4.3, example in Figure 2), and initialize graph nodes with complementary signals (Section 4.4) before modeling them across visit sequence (Section 4.5).

4.1 Extracting and Aligning Event-Temporal Graphs from Notes

We use state-of-the-art models from Chaturvedi et al. (2025) to extract temporal graphs from each note. Here, nodes are clinical entities (*Problem, Test, Treatment, Clinical Department, Clinical Occurrences, Evidential*), time expressions (*Date, Time, Duration, Frequency*)—circular nodes in Figure 2—and edges, their temporal relations (*Before*, *After*, *Overlap/Simultaneous*). We *align* extracted entities by linking and clustering multiple mentions into a single node (Section 4.2). We optionally denoise temporal edges by enforcing consistency constraints using TimeGraph and normalizing directed edges (Before/After) to Before (Section 4.3).

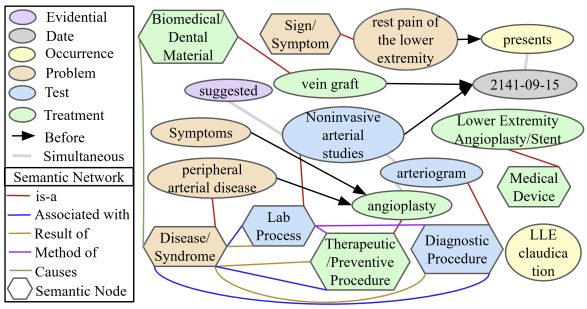


Figure 2: An example of an extracted graph from one clinical note of MIMIC-IV. Extracted entities (clinical events and time expressions) are circular nodes, color-coded by type, and have temporal edges (Before, Simultaneous). This is further augmented with the UMLS KG, which includes semantic nodes connected to the corresponding entity via the *is-a* relation and to each other via semantic relations (e.g., *causes*, *method of*).

4.2 Entity Linking/Normalization

Clinical Entities Next, we perform clinical entity alignment by clustering synonymous textual mentions into a single node via linking to an external knowledge graph. If any entity is of type *Problem/Treatment/Test/Clinical Department*, we use Metamap (Aronson and Lang, 2010) to map it to a unique concept in the Unified Medical Language System (UMLS) Metathesaurus (Bodenreider, 2004).[4] UMLS comprises a vast repository mapping synonymous medical terms to unique concept identifiers (CUI). Additionally, UMLS defines a Semantic Network that provides a higher-level abstraction over individual concepts by organizing

[3] Health Insurance Portability and Accountability Act (HIPAA), a U.S. law, considers dates as PHI. In MIMIC-IV notes, all PHIs are identified by a union of a rule-based and a neural approach, and replaced with three underscores '___' (Johnson et al., 2023).

[4] We do not link other entity types, as it leads to noise (Appendix C).

them into coarser semantic types (e.g., Disease or Syndrome, Diagnostic Procedure, Medical Device). For each extracted concept, we identify its corresponding semantic type using MetaMap and add it as an additional node in the graph (hexagonal nodes in Figure 2). A set of permissible semantic relations between semantic types is also pre-specified, which we incorporate as edges in the graph (e.g., Diagnostic Procedure *associated with* Disease or Syndrome, shown as a blue edge in Figure 2). Each concept entity node is connected to its associated semantic type node via *is-a* relations (red edges in Figure 2; e.g., stent *is-a* Medical Device), explicitly modeling the type hierarchy.[5] This yields a heterogeneous graph combining extracted temporal graphs with medically grounded abstractions.

Dates We normalize and align multiple textual mentions of the same date to a standard representation using Microsoft Recognizers-Text (Huang et al., 2017). All mentions of a date entity are clustered as a single node in the temporal graph.

4.3 Consistent Temporal Graph

We additionally evaluate temporal edge denoising via consistency enforcement using the TimeGraph algorithm (Miller and Schubert, 1990). This algorithm iteratively adds the most confident edges based on prediction probabilities to the final graph, and prunes those that introduce cycles. The resulting graph contains *Before*, *After*, and *Simultaneous* edges (black and grey edges in Figure 2). Following prior observations that temporal relations form inverse, redundant pairs (Mani et al., 2006), and inspired by work emphasizing canonical temporal ordering via *Before* relations (Chambers and Jurafsky, 2008), we flip the direction of *After* edges and relabel them as *Before*. This reduces label sparsity and simplifies temporal supervision without imposing additional modeling constraints. Table 1 presents descriptive statistics of the final graphs.[6]

4.4 Node Representation

To represent a node, we compute its corresponding textual span's embedding by concatenating the in-context BioMedBERT (Gu et al., 2021) embeddings of the first and last tokens with the span-width embeddings. This model is pretrained on biomedical corpora and demonstrated strong performance on domain-specific tasks.[7] We represent multiple linked mentions of an entity as a single node in each temporal graph, and initialize it with the mean of the embeddings of all linked mentions. To represent semantic type nodes, we average BioMedBERT embeddings over all tokens in their label (preferred name). We also use knowledge graph embeddings (KG-embeddings) for Concept Unique Identifiers (CUIs) derived from the UMLS knowledge graph for each linked event, as introduced by Maldonado et al. (2019). The final node representations are concatenated KG embeddings and text embeddings (BioMedBERT) over all mentions.

Parameter	PH Corpus	MIMIC-IV
#Patients	3332	5802
Avg. #tokens/doc	1209.2 (637.6)	454.5 (208.3)
Avg. #nodes/doc	114.7 (46.1)	74.8 (27.9)
Avg. #Edges/doc	261.1 (133.4)	158.2 (76.8)
#doc or visits	3.2(1.7)	2.3 (1.4)

Table 1: Mean tokens, nodes, and edges in the extracted temporal graph per document (doc), per patient, and number of documents per patient, with standard deviations in parentheses.

4.5 Hierarchical Temporal Modeling

We model irregular time-series of temporal graphs across patient visits using HiTGNN for Type 2 Diabetes prediction. Each visit is a graph $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$ with temporal (Before/After/Overlap) and semantic (UMLS) edges. HiTGNN captures intra-graph and inter-graph dependencies via a GNN and a recurrent neural network (RNN).

Intra-visit Modeling. Each graph is encoded with a multi-layer GraphSAGE encoder (with mean aggregation, residual connections, and layer normalization). Given initial features $\mathbf{h}_v^{(0)}$ of node v, $\mathcal{N}(v)$ the neighbors of v, and learnable parameters $\mathbf{W}_1^{(k)}, \mathbf{W}_2^{(k)}$, node updates at layer k are:

$$\mathbf{h}_v^{(k)} = \text{LayerNorm}\Big(\mathbf{W}_1^{(k)}\mathbf{h}_v^{(k-1)} + \mathbf{W}_2^{(k)}\frac{1}{|\mathcal{N}(v)|}\sum_{u\in\mathcal{N}(v)}\mathbf{h}_u^{(k-1)}\Big) + \mathbf{h}_v^{(k-1)}$$

Where $\mathbf{h}_v^{(k)}$ is the representation of node v at layer k. Updated node embeddings are mean-pooled to obtain the graph representation for a visit $\mathbf{g}_t$.

[5]The complete set of semantic types and relations is available at: https://www.nlm.nih.gov/research/umls/knowledge_sources/semantic_network.

[6]The mean number of tokens to get estimated document length is counted based on Llama3.1-8B tokenizer, as we also include comparisons with this model.

[7]We use the fine-tuned model version and width embeddings from Chaturvedi et al. (2025).

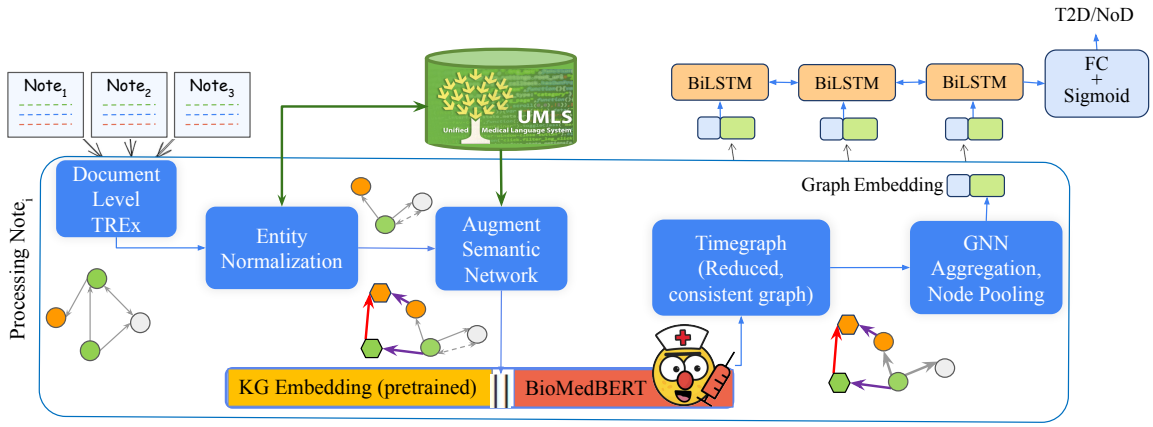


Figure 3: HITGNN architecture for longitudinal modeling of clinical trajectories, combining document-level temporal graphs with cross-visit aggregation and UMLS knowledge for T2D risk prediction.

Inter-visit Modeling: temporal dependencies across visits are modeled as:

$$\mathbf{z}_t = \text{BiLSTM}_\phi(\mathbf{g}_1, \ldots, \mathbf{g}_t), \qquad \mathbf{z}_T = [\overrightarrow{\mathbf{h}}_T ; \overleftarrow{\mathbf{h}}_1]$$

Where $\mathbf{g}_1, \ldots, \mathbf{g}_t$ is the sequence of graph encodings over visits. The final forward and backward hidden states from bidirectional LSTM ($BiLSTM$) are concatenated to represent a patient's trajectory. Finally, a fully connected network with sigmoid ($\sigma(\cdot)$) gives binary outcome ($\hat{y}$)—T2D or NoD (No Diabetes):

$$\hat{y} = \sigma(\mathbf{W} \cdot \mathbf{z}_T + \mathbf{b})$$

Figure 3 shows the model architecture, integrating UMLS knowledge graph, document-level temporal information, and cross-visit progression.

5 LLMs for Diabetes Risk Prediction

In this section, we use Large Language Models (LLMs) as reasoning engines to predict a patient's risk of developing T2D from the clinical notes.

5.1 Reasoning with Verifier-Aided Labeling (REVEAL)

Healthcare applications face significant resource constraints, which limit the deployment of LLMs; yet, maintaining reasoning capabilities is crucial for effective clinical decision-making. Recent advances in test-time compute scaling offer a promising solution: using smaller models enhanced by verification mechanisms, rather than relying solely on larger, computationally expensive ones. To this end, the REVEAL framework combines a reasoner LLM with a smaller verifier LLM, in three steps: (1) a reasoner LLM generates N reasoning paths with predictions; (2) a fine-tuned verifier evaluates the credibility of these paths and assigns a score from 0–1, and (3) predictions across different reasoning paths are aggregated using verifier scores.

Concretely, at *training time*, the reasoner model generates five stochastic outputs per example, each including a prediction ('true'/'false') and an explanation. A small LLM verifier, fine-tuned with LoRA (Hu et al., 2022), classifies each output as 'correct' or 'incorrect' using the prompt in Figure 10. The normalized probabilities of the output tokens serve as the verifier's confidence score. At *inference time*, the reasoner generates N ($N = 10$) diverse predictions. Each prediction contains a risk outcome ('true'/'false') and a corresponding explanation. The verifier scores them, and the outcome is selected by majority vote among the top-k highest-confidence predictions.

6 Experiments

This section presents a comprehensive empirical evaluation of HITGNN and REVEAL.[8]

6.1 Code and Data

The replication code is available at https://github.com/RochanaChaturvedi/temporal-nlp-clinical-risk. We also release non-identifiable metadata to reconstruct the MIMIC-IV T2D cohort, including patient identifiers, note timestamps, and demographic attributes. The full MIMIC-IV-T2D dataset will be released via PhysioNet (Goldberger et al., 2000).

[8] See Appendix B for prompts and experimental details.

6.2 Baseline Methods

Zero-shot Prompting: The model directly predicts T2D risk from clinical notes using deterministic decoding (temperature=0.0). We extract normalized token probabilities for 'true'/'false' predictions to compute AUC scores (see prompt in Figure 8). We use Llama3.1-8B and Llama3.2-1B, and also compare the performance on MIMIC-IV with prevailing large-scale models, DeepSeek-V3 (671B parameter Mixture-of-Experts model) and GPT-4o.[9]

Self-Consistency (Llama3.1-8B-SC): We generate N ($N = 10$) diverse predictions per patient using stochastic sampling (temperature=1.0) and take the majority vote for the final classification.

Supervised Fine-tuning (SFT): We fine-tune a 1B LLM using LoRA given the low-resource setting. The resulting model (Llama3.2-1B-ft) is trained to generate only 'true' or 'false' labels, without explanations, due to the lack of annotated reasoning data (see prompt in Figure 9).

CLSTM We also use concept-based CNN-LSTM baseline (Chaturvedi et al., 2023). Here, we model extracted entities (nodes of temporal graphs) with a CNN and max pooling, then process each document representation with an LSTM.

Both CLSTM and HiTGNN models utilize SpanTREx—a leading open-source end-to-end temporal relation extraction model from Chaturvedi et al. (2025). We also compare with their best model GraphTREx in section 7.4.2.

6.3 Evaluation Metrics

Both datasets are constructed with balanced case-control sampling to ensure sufficient representation of T2D patients. Consequently, data distribution does not reflect real-world prevalence ($\approx 11.6\%$ in the U.S.); therefore, absolute precision values may be optimistic. All models are evaluated under identical conditions; thus, relative comparisons remain valid. We emphasize ROC-AUC and T2D recall as prevalence-independent measures of discrimination performance. For completeness, we also report precision and recall for both groups, and the macro-average F_1.[10] We also use the following metrics for fairness evaluation:

Demographic Parity Difference (DPD) difference in the probability of positive prediction rates between a target group ($Z = z$) and the other groups ($Z \neq z$). It is defined as:

$$\text{DPD} = P(\hat{Y} = 1 \mid Z = z) - P(\hat{Y} = 1 \mid Z \neq z)$$

Where $\hat{Y}$ is the prediction (T2D $= 1$, NoD $= 0$), and Z is a sensitive attribute (e.g., gender, race).

Equal Opportunity Difference (EOD) difference in true positive rates (TPR) between the target group and others: $TPR_{Z=z} - TPR_{Z\neq z}$
Where $TPR_{Z=z} = P(\hat{Y} = 1 \mid Y = 1, Z = z)$.

7 Results and Discussion

Table 2 presents the main experimental results.[11] The first panel comprises LLM-based models. Among zero-shot LLM results, Llama3.2-1B fails to identify patients at risk of diabetes, classifying almost all patients into the 'no-risk' (NoD) cohort. The performance improves with the higher capacity 8B variant. While self-consistency (SC) has shown promising performance in other tasks and domains, Llama3.1-8B-SC performs poorly on both datasets. The Llama3.2-1B-ft (fine-tuned) demonstrates strong performance, achieving an AUC of 65.06% on PH and 64.41%—the second highest among all models—on MIMIC-IV. While this fine-tuned LLM excels as a classifier, it cannot provide reasoning or explanations for its outputs.[12] Our REVEAL model does not significantly outperform the Llama3.2-1B-ft. However, it preserves the reasoning output and improves upon the zero-shot 8B variant in terms of T2D group recall, highlighting its sensitivity in capturing true positives. For larger models, including DeepSeek-V3 with impressive clinical decision-making capabilities (Sandmann et al., 2025), the gains are less pronounced: its AUC is similar to the much smaller Llama3.1-8B, but with lower T2D recall. GPT-4o attains a higher AUC but exhibits worse T2D recall than both Llama3.1-8B and REVEAL, and an overall weaker performance than Llama3.2-1B-ft.

The second panel summarizes results with structured approaches. The CLSTM model provides a

[9] Not evaluated on PH corpus due to privacy constraints.

[10] ROC-AUC scores are computed from prediction probabilities. For LLMs, this is the probability of generating 'true' at the predicted token position. For LLM-SC, it is the 'true' token probability, averaged across all runs.

[11] We also perform a pairwise comparison of the top four models using bootstrap sampling to estimate whether the differences are statistically significant, based on non-parametric bootstrap resampling test (see Appendix Table D).

[12] Obtaining gold-standard reasoning data at scale for fine-tuning is costly and labor-intensive, especially from already overburdened clinical experts, limiting the feasibility of fine-tuning with reasoning.

Model	PH Corpus						MIMIC-IV					
	AUC	F_1	T2D		NoD		AUC	F_1	T2D		NoD	
			P	R	P	R			P	R	P	R
Llama3.2-1B (0-shot)	46.80	35.46	39.29	3.09	49.56	**95.22**	40.00	33.10	0	0	49.48	1.00
Llama3.1-8B (0-shot)	57.73	56.98	57.49	53.93	56.61	60.11	53.71	51.03	61.54	27.21	52.65	82.64
DeepSeek-V3 (0-shot)	†N/A	N/A	N/A	N/A	N/A	N/A	55.02	47.85	69.23	18.37	52.38	91.67
GPT-4o (0-shot)	N/A	N/A	N/A	N/A	N/A	N/A	57.75	51.81	**75.56**	23.13	54.07	92.36
Llama3.1-8B-SC	55.62	55.49	56.29	50.28	55.08	60.96	52.12	43.43	66.67	12.24	51.14	**93.75**
Llama3.2-1B-ft	65.06	61.62	59.53	**78.93**	68.75	46.35	64.41	59.17	63.89	46.94	57.38	72.92
REVEAL	57.98	65.24	66.77	60.96	64.08	69.66	53.13	51.60	55.21	36.05	51.79	70.14
CLSTM	68.24	65.62	63.19	76.69	**70.36**	55.34	63.08	59.62	63.96	48.30	57.78	72.22
HITGNN	**72.24**	**67.28**	**71.48**	58.43	64.85	76.69	**67.49**	**61.87**	65.81	**52.38**	**59.77**	72.22
#patients	712		356		356		291		147		144	

Table 2: Performance of HITGNN and REVEAL against different baselines. †Evaluation of external models on private data is not possible due to privacy restrictions on identifiable data.

strong baseline, showing higher (on PH) or comparable discrimination (on MIMIC-IV) to LLMs. HITGNN performs even better, attaining significantly higher AUC on the PH corpus, compared to other top-performing models (REVEAL, Llama3.2-1B-ft, and CLSTM). On MIMIC-IV, it also attains significantly higher AUC over REVEAL and CLSTM, and significantly higher T2D recall than REVEAL and Llama3.2-1B-ft. In contrast, no other model shows significant improvements over HITGNN across any metric on either corpus. HITGNN also outperforms both the larger models (DeepSeek-V3 and GPT-4o) across key metrics. We next present the results from additional analyses and ablations.

7.1 Prediction Horizons

We compare performance across prediction horizons by dividing the time from last pre-diagnosis visit to diagnosis into 3-month windows, combining each T2D subgroup with all NoD controls to compute AUC (Figure 4).[13] The final window aggregates values above the 95^{th} percentile to avoid sparsity. Results for the four best-performing models (HITGNN, CLSTM, REVEAL, and Llama3.2-1B-ft) are presented. Across corpora, HITGNN outperforms other models in near-term prediction ($\leq$ 1.5 years), supporting use cases centered on more urgent risk stratification and timely preventive interventions, and reflects practical EHR constraints, where longitudinal completeness is rare. While CLSTM is not as effective in the short term, at longer horizons, it consistently achieves the highest AUC. Llama3.2-1B-ft also exhibits improvements over the longer term. HITGNN exhibits smoother performance degradation in MIMIC-IV and begins to match CLSTM in the long term. While REVEAL performance on PH shows consistent degradation over time, on MIMIC, it initially is low but achieves the best AUC within a $\approx 1-4$ year period before declining in the longer term.

[13] Appendix E Figure 11 shows trends for T2D group recall.

7.2 Fairness Analysis

			PH corpus					MIMIC-IV				
		G	GNN	CLS	RVL	ft	G	GNN	CLS	RVL	ft
Ethnicity	DPD	H	-0.12	-0.11	-0.05	-0.02					
		N	0.12	0.11	0.05	0.02					
	EOD	H	-0.08	-0.09	0	0.01					
		N	0.08	0.09	0	-0.01					
Gender	DPD	F	-0.04	-0.02	0.03	0.03	F	0.01	0.04	0.04	0.02
		M	0.04	0.02	-0.03	-0.03	M	-0.01	-0.04	-0.04	-0.02
	EOD	F	0	0	0.06	0.06	F	0.12	0.2	-0.10	0.19
		M	0	0	-0.06	-0.06	M	-0.12	-0.2	0.10	-0.19
Race	DPD	B	0.09	0.09	0.10	0.04	B	0.04	0.13	0.03	0.12
		W	0	-0.01	-0.09	-0.03	W	0.04	-0.04	-0.03	-0.04
		NI	-0.11	-0.10	-0.06	-0.03	H	-0.01	0.01	0.07	0.02
							A	-0.34	-0.32	-0.10	-0.31
	EOD	B	0.08	0.08	0.08	0.06	B	0.05	0.18	-0.06	0.16
		W	0	0.03	-0.09	-0.03	W	0.04	-0.06	-0.01	-0.05
		NI	-0.11	-0.11	-0.02	-0.03	H	-0.03	-0.06	0.15	-0.05
							A	-0.37	-0.33	-0.03	-0.32

Table 3: DPD/EOD for subgroups (G): H (Hispanic), N (non-Hispanic), F (Female), M (Male), B (Black), W (White), NI (Unknown), A (Asian). Models: GNN (HITGNN), CLS (CLSTM), RVL(REVEAL), and ft (Llama3.2-1B-ft).

Table 3 reveals consistent, corpus-dependent fairness trade-offs. In PH, all models exhibit negative DPD for Hispanic patients, with HITGNN showing the largest imbalance (DPD:-0.12, EOD:-0.08). For gender, HITGNN exhibits a mild imbalance against the Female majority (DPD -0.04; EOD 0), while LLM-based models RVL (REVEAL) and ft (Llama3.2-1B-ft) exhibit mild negative DPD and nontrivial EOD shifts (-0.06) against *Males*. Race-related disparities are most pronounced for the minority Unknown (NI) group. In contrast, the majority Black patients are mildly favored across models. In MIMIC-IV, disparities against Males are mild

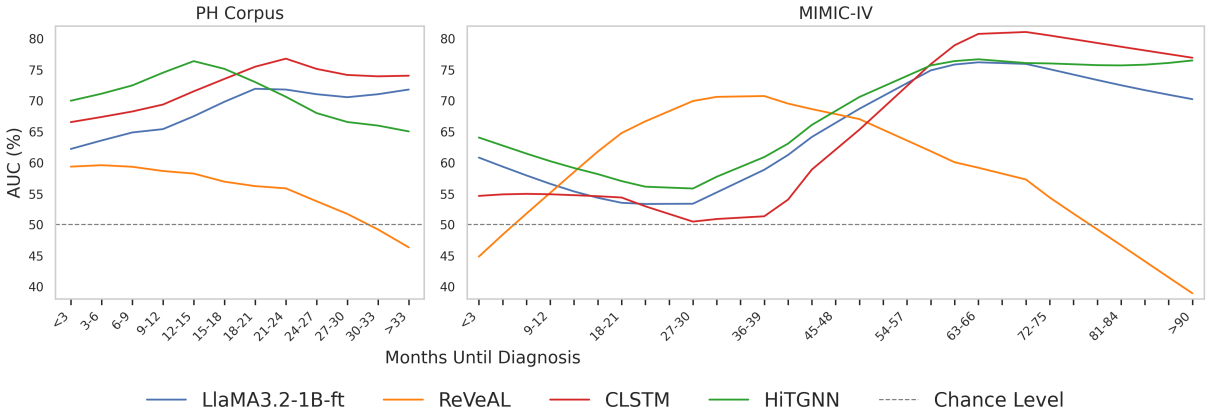


Figure 4: AUC as a function of the prediction horizon, evaluated over consecutive 3-month windows.

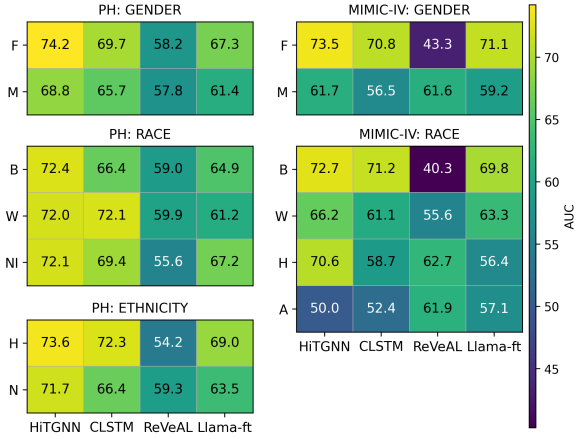


Figure 5: AUC evaluated at subgroup level.

in terms of DPD but substantial in terms of EOD, with RVL reversing direction to favor *Male* patients. Among race groups, Asian patients experience the strongest negative shifts in both DPD and EOD across most models. RVL strongly favors *Hispanics* and ft strongly favors *Black* (B). Across both corpora, CLSTM is comparable to or more imbalanced than HiTGNN.

HiTGNN achieves the best AUC across subgroups and best worst-case subgroup protection in PH (Figure 5). On MIMIC-IV, it again leads across groups for gender and race except for Asians, where it performs worst (50%). Here, Llama-ft offers the best worst-case race protection 56.4%, highlighting group-specific trade-offs between aggregate discrimination and extreme subgroup harm.

7.2.1 Demographic prior probing.

Although HiTGNN does not use protected attributes, LLM-based models' overall performance improves when demographics are provided, raising the question of whether such gains reflect legitimate population-level risk factors or reliance on spurious demographic shortcuts learned during pretraining. To disentangle these effects, we conduct demographic prior probing analyses. Specifically, we construct deliberately underspecified prompts that provide only a single protected attribute (race, ethnicity, or gender), with no clinical evidence, to isolate reliance on demographic priors. Probes are run on Llama3.1-8B—the model used for explanation generation—using greedy decoding to obtain deterministic, highest-probability completions.

Population data show that diagnosed diabetes prevalence in the U.S. is highest among non-Hispanic Black adults, followed by Hispanic and non-Hispanic Asians, and lowest among non-Hispanic White adults (Centers for Disease Control and Prevention, 2023), with higher risk in men than women (Kautzky-Willer et al., 2023). Llama-8B sometimes reflects these patterns, assigning a higher risk to Black, Hispanic, and Asians relative to White patients. However, it also generates self-contradictory explanations, alternately asserting that "as per CDC, Non-Hispanic whites have a higher prevalence of diagnosed diabetes compared to other racial and ethnic groups", and that "Type 2 Diabetes is higher among non-Hispanic blacks and Hispanics compared to non-Hispanic whites". For gender, it generates similar contradictions, variously asserting "women are at lower risk than men" and "men are generally at lower risk than women".

Appendix F Table 16 compares DPD and EOD with and without demographic conditioning, and finds that it affects fairness in corpus-dependent ways: in PH, it increases both the gender and racial imbalance. However, in MIMIC, it reduces gender imbalance but induces asymmetric racial shifts, substantially reducing bias toward Black patients while increasing it for White patients, and most strongly favoring Asian patients.

Overall, HiTGNN offers more stable and higher average subgroup-level discrimination on both corpora. LLM-based models improve worst-case protection for race on MIMIC-IV. Analysis of LLMs with and without demographic conditioning shows their sensitivity to protected attributes can substantially reweight subgroup outcomes. Demographic prior probing reveals that LLM gains from demographic conditioning should be interpreted cautiously, as they may result from unstable group-level shortcuts rather than individual evidence.

7.3 Computation Time and Resources

Model	Train-Time (minutes)	Test-Time (seconds)	Memory (GB)	#Train Params (Billion)	Input (#notes)
HiTGNN †	1.5	0.01	0.04	0.01	5
CLSTM†	12	0.02	0.05	0.05	5
Llama3.2-1B^{Ex}	pre-trained	6	2.4	N/A	2
Llama3.1-8B^{Ex}	pre-trained	6	15	N/A	2
Llama3.2-1B-ft	18 hours	0.2	2.42	0.001	2
REVEAL ^{Ex}‡	30 hours	62	17.42	0.001	2

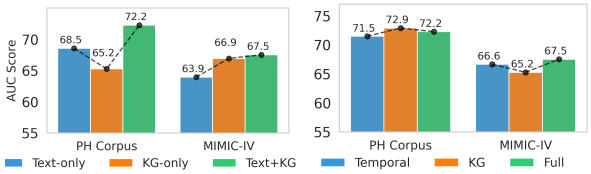
Table 4: Computational efficiency of models. Inference time is per patient. †HiTGNN and CLSTM rely on MetaMap for UMLS linking. In this study, MetaMap was run sequentially due to API constraints, making the one-time preprocessing pipeline—extraction, linking, TimeGraph, and embeddings lookup—cost ≈ 42 seconds per PH note and ≈ 15.6 seconds per MIMIC-IV note. This step can be substantially accelerated via parallelization using recent tools such as ParallelPyMetaMap (https://github.com/biomedicalinformaticsgroup/ParallelPyMetaMap). ‡The table excludes one-time data preparation costs for fine-tuning the REVEAL model; the training time is over 5 reasoning paths, while inference time is over 10. ^{Ex}These models generate explanations.

Table 4 summarizes the resource usage across models (excluding GPT-4o and DeepSeek-V3, which are accessed via API). A single A100 GPU with 80 Gb RAM is used for all experiments (accelerated inference is used for Llama experiments). Training the HiTGNN model takes ≈ 1.5 minutes, and inference per patient takes ≈ 0.01 seconds. For CLSTM, training the model takes ≈ 12 minutes, and inference on one patient takes ≈ 0.02 seconds. For inference with the Llama model, the average time required with explanations is approximately ≈ 6 seconds. Fine-tuning the model takes around 18 hours, and inference with this version takes 0.2 seconds. REVEAL verifier training takes 30 hours (excluding training data preparation, which takes approximately 30 seconds per patient for obtaining explanations from Llama3.1-8B along 5 reasoning paths). Inference along 10 reasoning paths from this model takes ≈ 62 seconds (obtaining and verifying reasoning along 10 paths). The full Llama3.2-1B model requires ≈ 2.4 GB of memory, while Llama3.1-8B requires ≈ 15 GB. The LoRA adaptors require an additional 20 MB. In contrast, CLSTM takes 54 MB and HiTGNN is even lighter at 44 Mb, demonstrating impressive scalability in terms of both speed and memory usage. The SPANTREX model used by CLSTM and HiTGNN requires 0.44 GB of space.

7.4 HiTGNN Ablations

7.4.1 Node Embeddings and Subgraph Variations:

Figure 6 compares HiTGNN performance across input choices based on AUC scores. Corresponding T2D recall trends are discussed in Appendix E. Figure 6a shows comparisons across node representations. Node initialization with in-context BioMedBERT embeddings ('Text-only') outperforms UMLS-CUI embeddings ('KG-only') on the PH corpus but underperforms it on MIMIC-IV. Combining both ('Text+KG') gives the best performance on PH but stays comparable to 'KG-only' on MIMIC-IV (p-value=0.94).[14] Figure 6b compares subgraphs: 'Temporal' excludes the UMLS semantic network (KG), 'KG' excludes temporal edges, and 'Full' includes both. In PH, 'KG' performs comparably to 'Full' (p = 0.33). In MIMIC-IV, combining temporal edges and UMLS KG yields the best performance, but stays statistically comparable to 'Temporal' alone (p = 0.35).



(a) Embeddings (b) Subgraphs

Figure 6: Node embedding and subgraph ablations.

7.4.2 Temporal Relation Extraction Models:

Table 5 reports the performance of two recent state-of-the-art temporal relation extraction models, GRAPHTREX and SPANTREX (Chaturvedi

[14]P-values are reported based on bootstrap resampling test with 10,000 replicates.

et al., 2025), trained on the I2B2-2012 corpus (Sun et al., 2013) and evaluated on a small sample of the PH corpus. This sample comprises nine notes jointly annotated by an NLP expert and a physician following the I2B2 entity annotation guidelines, using the NarrativeTime annotation tool (Rogers et al., 2024). This tool enables faster and denser annotations, resulting in $\approx 280K$ relation pairs between $\approx 2K$ entities. We report partial span matching F_1 scores for events and timex, and exact match F_1 for temporal relation (TLink). Consistent with prior work, SPANTREX achieves slightly higher F_1 scores for Event and TimeEx extraction, while GRAPHTREX yields higher precision. For TLink extraction, GRAPHTREX substantially outperforms SPANTREX, with approximately 14% absolute gain in F_1. We also report details of the sample and the entity linking performance of Metamap (for clinical entities) and Recognizers-Text (for dates) in Appendix C. The results show high precision but lower recall, reflecting more fragmented clusters rather than noisy ones.

Despite GRAPHTREX achieving higher scores in TLINK F_1 over SPANTREX, Table 6 shows that downstream discrimination of HITGNN stays comparable (on PH) or better (on MIMIC-IV) with SPANTREX over GRAPHTREX-extracted graphs. Still, the GRAPHTREX-based HITGNN performs better (on PH) or comparable (on MIMIC-IV) to other baselines in Table 2.

Model	Event F_1	Timex F_1	TLink F_1
SPANTREX	**74.39**	**69.46**	38.00
GRAPHTREX	73.54	68.42	**51.97**

Table 5: Performance of state-of-the-art Temporal Relation Extraction models on PH subsample.

TREX Model	HITGNN AUC	
	PH	MIMIC-IV
SPANTREX	**72.24**	**67.49**
GRAPHTREX	71.81	64.27

Table 6: HITGNN AUC scores using SPANTREX vs. GRAPHTREX extracted temporal graphs.

7.4.3 Pipeline Component Ablations:

Table 7 presents an ablation study of the HITGNN pipeline, examining the contribution of entity linking, TimeGraph-based filtering, and flipping *After* edges to their inverse *Before* across corpora. For the PH corpus, the full pipeline (linking + TimeGraph + flipping) achieves the highest AUC (72.24), indicating that jointly leveraging external knowledge, enforcing temporal consistency, and temporal relation simplification improves the overall discrimination with temporal edges. In contrast, for MIMIC-IV, the highest AUC (67.49) is achieved using entity linking alone, while normalization via TimeGraph construction or label flipping does not consistently improve AUC. This contrast is consistent with the dataset characteristics discussed in Section 3, Table 1: MIMIC-IV has sparser event–temporal graphs, whereas the PH corpus contains denser graphs extracted from longer notes. In PH, TimeGraph and label flipping help mitigate noise in raw temporal relations, but in MIMIC, they may sparsify already weak but useful temporal signals.

Pipeline Configuration			PH Corpus	MIMIC-IV
Linking	TimeGraph	Flipping		
✓	✓	✓	**72.24**	64.00
✗	✗	✗	65.28	63.20
✗	✓	✗	62.46	63.05
✗	✓	✓	65.51	64.80
✓	✗	✗	64.84	**67.49**
✓	✗	✓	64.98	67.05

Table 7: HITGNN AUC variations with pipeline.

7.4.4 GNN Architecture Ablations:

Table 8 shows variations with GNN encoders.[15] GraphSAGE has the strongest performance across corpora. More expressive encoders (GAT, RGCN, HGTConv) do not provide further gains. HGT (Heterogeneous Graph Transformer) performs weakest, likely because its type-specific message passing and type-wise pooling introduce substantial sparsity and imbalance.[16] This suggests that HITGNN benefits more from stable neighborhood aggregation than highly parameterized models.

GNN Model	PH Corpus	MIMIC-IV
GraphSAGE	**72.24**	**67.49**
RGCN	66.09	65.75
GAT	65.99	64.73
HGT	54.67	59.62

Table 8: AUC across GNN variations in HITGNN.

[15]These configurations use mean pooling. With max pooling, performance drops, e.g., GraphSAGE AUC decreases to 68.02 on PH and 60.45 on MIMIC-IV.

[16]HGT parameterizes interactions over (source node type, relation type, target node type) triplets, leading to a combinatorial increase in edge types; due to computational constraints and to limit further sparsity, we exclude UMLS KG.

7.5 Clinical Analysis of REVEAL Explanations.

To better understand how REVEAL reasoning aligns with contemporaneous clinical reasoning, we perform a qualitative analysis of model explanations on a small set of cases from the PH test sample. A senior endocrinologist independently assessed risk based on the last two clinical notes and then reassessed after reviewing the REVEAL explanations over the same inputs. In all cases, explanations did not alter the clinician's original judgment, suggesting that they do not exert undue persuasive influence. In one case, the expert supports the model's judgment, in contradiction with the ground truth, possibly due to delayed onset or left censoring. These explanation-expert assessments are provided in Appendix G. They illustrate both cases where the model's explanations closely mirror expert judgment and cases where plausible-sounding explanations over-emphasize weak or irrelevant signals, highlighting important limitations of current LLMs in challenging clinical scenarios.

8 Conclusion

We address the challenge of clinical risk prediction from longitudinal notes by moving beyond static snapshots to dynamic patient representations. To this end, we present HITGNN, a temporally grounded event-centric GNN augmented with clinical KGs. HITGNN attains superior predictive performance, especially on the immediate-risk horizon, while remaining computationally efficient. Complementing HITGNN, we introduce REVEAL, a test-time LLM scaling method for interpretable predictions. By utilizing a smaller LLM to verify rationales from a larger LLM, REVEAL offers a promising balance between performance and explainability, while much larger models like GPT-4o exhibit limited zero-shot performance on this complex clinical task. By constructing temporally realistic, demographically matched cohorts, we avoid post-diagnosis data leakage and enable controlled fairness evaluation. Our fairness analysis finds that even without demographic conditioning, HITGNN demonstrates more stable and higher subgroup-level discrimination across datasets and groups, except for the Asian minorities. On the other hand, LLM assessments are sensitive to protected attributes and may rely on unstable demographic shortcuts.

Ablations examine the role of multiple information sources in HITGNN: enriching node representations with both contextualized and UMLS KG embeddings improves performance on PH but remains comparable to KG-only initialization on MIMIC-IV, temporal edges do not impact AUC but improve T2D recall on PH, while on MIMIC-IV they improve AUC but reduce recall. PH benefits from a full pipeline, whereas MIMIC-IV performs best with KG-linking but avoiding temporal edge pruning. We further find that HITGNN is robust to variation in upstream TREX quality, remaining better or comparable to other baselines. Overall HITGNN's downstream T2D performance reflects interactions among multiple upstream signals and corpus-specific characteristics.

While we currently use a hierarchical approach to aggregate temporal graphs within and across visits, the lack of publicly available cross-document temporal and coreference corpora limits explicit supervision to link events across notes; future works in this direction could enable better-connected patient timelines. HITGNN's modular design also enables component-level optimization—such as adopting newer biomedical linking tools (Kraljevic et al., 2021) and embeddings from larger language models. Improving subgroup fairness for underrepresented populations is another important direction for future work. For example, in MIMIC-IV, lower performance for the Asian subgroup may reflect both limited sample size and potential documentation bias; future work should explore cross-institutional data and targeted augmentation to support more equitable performance. Additional directions include developing alignment between graph-based reasoning and LLM explanations and extending these methods to other chronic diseases. More broadly, the proposed framework may also generalize to other domains requiring fine-grained temporal reasoning over longitudinal textual data, including financial forecasting and conflict prediction from news streams.

Ethics Statement

The use of the PH corpus was approved by the institutional review board (IRB) at the data hosting institution, and MIMIC-IV was reviewed by the IRB at Beth Israel Deaconess Medical Center. Both datasets were obtained after completing the required training and accessed under secure protocols in accordance with institutional and data use policies (Johnson et al., 2021). We use the

Azure OpenAI Service for GPT and Deepseek API hosted by togetherai for LLM inference on MIMIC-IV and opted out of human review due to restrictions on third-party processing of sensitive clinical data (https://physionet.org/news/post/gpt-responsible-use/).

The models presented in this work are intended as clinician-in-the-loop risk stratification or early-warning tools and should not be used as standalone diagnostic and medical decision-making tools without additional clinical validation.

Acknowledgments

The authors thank Sourav Medya for helpful feedback and Luke Lauridsen for assistance in preparing the annotated sample used in our additional analysis (Section 7.4.2). We also thank action editor Yulan He and reviewers for their helpful comments and suggestions. This work is also partially supported by the NSF under award IIS-2312862.

References

2025. *IDF Diabetes Atlas*, 11 edition. International Diabetes Federation, Brussels, Belgium.

Alan R. Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270.

Daniel Capurro, Erik van Eaton, Robert Black, and Peter Tarczy-Hornoch. 2014. Availability of structured and unstructured clinical data for comparative effectiveness research and quality improvement: a multisite assessment. *EGEMS*, 2(1).

Centers for Disease Control and Prevention. 2023. Centers for disease control and prevention. (2026) national diabetes statistics report. https://gis.cdc.gov/grasp/diabetes/diabetesatlas-statsreport.html. Accessed: 2026-01-25.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797.

Rochana Chaturvedi, Peyman Baghershahi, Sourav Medya, and Barbara Di Eugenio. 2025. Temporal relation extraction in clinical texts: A span-based graph transformer approach. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25765–25788, Vienna, Austria. Association for Computational Linguistics.

Rochana Chaturvedi, Mudassir Rashid, Brian T. Layden, Andrew Boyd, Ali Cinar, and Barbara Di Eugenio. 2023. Sequential representation of sparse heterogeneous data for diabetes risk prediction. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 831–834. IEEE.

Hejie Cui, Alyssa Unell, Bowen Chen, Jason Alan Fries, Emily Alsentzer, Sanmi Koyejo, and Nigam H Shah. 2025. Timer: Temporal instruction modeling and evaluation for longitudinal clinical records. *npj Digital Medicine*, 8(1):577.

Kirstie K Danielson, Brett Rydzon, Milena Nicosia, Anjana Maheswaren, Yuval Eisenberg, Janet Lin, and Brian T. Layden. 2023. Prevalence of undiagnosed diabetes identified by a novel electronic medical record diabetes screening program in an urban emergency department in the us. *JAMA Network Open*, 6(1):e2253275–e2253275.

Marzyeh Ghassemi, Marco A. F. Pimentel, Tristan Naumann, Thomas Brennan, David A. Clifton, Peter Szolovits, and Mengling Feng. 2015. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in icu with sparse, heterogeneous clinical data. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 446–453. AAAI Press.

Ary L. Goldberger, Luis A. N. 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. *Circulation*, 101(23):e215–e220.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan

Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30(9):2613–2622.

William J Hall, Mimi V Chapman, Kent M Lee, Yesenia M Merino, Tainayah W Thomas, B Keith Payne, Eugenia Eng, Steven H Day, and Tamera Coyne-Beasley. 2015. Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review. *American Journal of Public Health*, 105(12):e60–e76.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2025. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Information Fusion*, 118:102963.

William R. Hersh, Mark G. Weiner, Peter J. Embi, Judith R. Logan, Philip R. O. Payne, Elmer V. Bernstam, Harold P. Lehmann, George Hripcsak, Timothy H. Hartzog, James J. Cimino, and Joel H. Saltz. 2013. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Medical Care*, 51:S30–S37.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Kexin Huang, Abhishek Singh, Sitong Chen, Edward Moseley, Chih-Ying Deng, Naomi George, and Charolotta Lindvall. 2020. Clinical xlnet: Modeling sequential clinical notes and predicting prolonged mechanical ventilation. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 94–100.

Wenhao Huang, Zijia Lin, Chris McConnell, and Börje F. Karlsson. 2017. Recognizers-Text: Recognition and resolution of numbers, units, and date/time entities expressed across multiple languages.

Pengcheng Jiang, Cao Xiao, Adam Richard Cross, and Jimeng Sun. 2024. Graphcare: Enhancing healthcare predictions with personalized knowledge graphs. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2021. MIMIC-IV. *PhysioNet*. Version 1.0.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.

Alexandra Kautzky-Willer, Michael Leutner, and Jürgen Harreiter. 2023. Sex differences in type 2 diabetes. *Diabetologia*, 66(6):986–1002.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A. Folarin, Angus Roberts, Rebecca Bendayan, Mark P. Richardson, Robert Stewart, Anoop D. Shah, Wai Keong Wong, Zina Ibrahim, James T. Teo, and Richard J.B. Dobson. 2021. Multi-domain clinical natural language processing with medcat: The medical concept annotation toolkit. *Artificial Intelligence in Medicine*, 117:102083.

Zachary Chase Lipton, David C. Kale, Charles Elkan, and Randall C. Wetzel. 2016. Learning to diagnose with LSTM recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Ramon Maldonado, Meliha Yetisgen, and Sanda M Harabagiu. 2019. Adversarial learning of knowledge embeddings for the unified medical language system. *AMIA Summits on Translational Science Proceedings*, 2019:543.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chungmin Lee, and James Pustejovsky. 2006. Machine learning of temporal relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 753–760.

Chuizheng Meng, Loc Trinh, Nan Xu, James Enouen, and Yan Liu. 2022. Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. *Scientific Reports*, 12(1):7166.

Stephanie A Miller and Lenhart K Schubert. 1990. Time revisited 1. *Computational Intelligence*, 6(2):108–118.

Tuan Dung Nguyen, Thanh Trung Huynh, Minh Hieu Phan, Quoc Viet Hung Nguyen, and Phi Le Nguyen. 2024. Carer-clinical reasoning-enhanced representation for temporal health risk prediction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10392–10407.

Ramesh S Patil, Peter Szolovits, and William B Schwartz. 1981. Causal understanding of patient illness in medical diagnosis. In *Computer-Assisted Medical Decision Making*, pages 272–292. Springer.

Anna Rogers, Marzena Karpinska, Ankita Gupta, Vladislav Lialin, Gregory Smelkov, and Anna Rumshisky. 2024. Narrativetime: Dense temporal annotation on a timeline. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12053–12073.

Paul R. Rosenbaum and Donald B. Rubin. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

Sarah Sandmann, Stefan Hegselmann, Michael Fujarski, Lucas Bickmann, Benjamin Wild, Roland Eils, and Julian Varghese. 2025. Benchmark evaluation of deepseek large language models in clinical decision-making. *Nature Medicine*, pages 1–1.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomašev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Pankhuri Singhal, Lindsay Guare, Colleen Morse, Anastasia Lucas, Marta Byrska-Bishop, Marie A. Guerraty, Dokyoon Kim, Marylyn D. Ritchie, and Anurag Verma. 2023. Detect: Feature extraction method for disease trajectory modeling in electronic health records. *AMIA Summits on Translational Science Proceedings*, 2023:487.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association: JAMIA*, 20(5):806.

Guanchu Wang, Junhao Ran, Ruixiang Tang, Chia-Yuan Chang, Chia-Yuan Chang, Yu-Neng Chuang, Zirui Liu, Vladimir Braverman, Zhandong Liu, and Xia Hu. 2024b. Assessing and enhancing large language models in rare disease question-answering. *Preprint*.

Yongxin Xu, Kai Yang, Chaohe Zhang, Peinie Zou, Zhiyuan Wang, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2023. Vecocare: Visit sequences-clinical notes joint learning for diagnosis prediction in healthcare data. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 4921–4929. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Shao Zhang, Jianing Yu, Xuhai Xu, Changchang Yin, Yuxuan Lu, Bingsheng Yao, Melanie Tory, Lace M. Padilla, Jeffrey Caterino, Ping Zhang, and Dakuo Wang. 2024. Rethinking human-ai collaboration in complex medical decision making: A case study in sepsis diagnosis. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S. Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Xian Wu, Yefeng Zheng, Lei Clifton, Zheng Li, Jiebo Luo, and David A. Clifton. 2023. A survey of large language models in medicine: Progress, application, and challenge.

Yue Zhou, Barbara Di Eugenio, and Lu Cheng. 2025. Unveiling performance challenges of large language models in low-resource healthcare: A demographic fairness perspective. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7266–7278, Abu Dhabi, UAE. Association for Computational Linguistics.

A Data Curation Details

A.1 PH corpus

We show the data-preparation steps in Table 9. The initial data is 164,910 patients, 15,140 of whom were diagnosed with Type 2 Diabetes, and 3,708,168 clinical notes. We exclude all post-diagnosis notes (these include notes after 3 days pre-diagnosis date). The 3-day pre-diagnosis window is considered to account for possible lag in transforming information from notes to ICD codes in structured tables. This filter ensures we only consider pre-diagnosis information, leaving us with around 3 million notes. We then exclude the rare demographic groups. These include race groups Native Hawaiian/Other Pacific Islander/American Indian/Alaska Natives/Asian, and unknown Gender/Ethnicity. We filter the data by note type and author, as it includes all providers and note types, some of which are uninformative. This subset of notes is determined with the help of a senior expert endocrinologist and is provided below:

Important Note Types Inpatient Progress Note, Progress Notes, Family Medicine Note, History and Physical Note, Emergency Department Note, Discharge Summary, Nutrition Note, Endocrinology Note, Specialty Pharmacy Services, General Eye Note, Diet Instructions, ED Notes, Outpatient Pharmacy Note, Discharge Note, Dialysis Rounding, Diabetes Education.

Important Author Types Unspecified, Anesthesiologist, Care Coordinator, Dentist, Dietician Student, Fellow, Licensed Clinical Social Worker, Licensed Practical Nurse, Medical Assistant, Medical Student, Mental Health Counselor, Midwife, Nurse Practitioner, Nursing Student, Occupational Therapist, Occupational Therapy Assistant, Occupational Therapy Student, Optometrist, Oral Surgeon, Pharmacist, Physical Therapist, Physical Therapy Assistant, Physical Therapy Student, Physician, Physician Assistant, Registered Dietitian.

We then remove notes shorter than 100 words, following previous work in health-risk prediction (Ghassemi et al., 2015). Upon manual inspection of some of the notes in this filtered subset from the pre-diagnosis date, we find that a diagnosis of T2D is already present but not entered in the structured records, or it is entered much later. As discussed previously, such ICD-coding errors are common (Hersh et al., 2013). This motivated us to apply an additional filter using a popular large language model (LLM), Llama3.1-8B. LLMs are exceptionally good at extracting explicit information already present in the text. We apply two different prompts (these are variations of the prompt in Figure 7), asking the model to identify if an explicit diagnosis is already present in the note of the T2D cohort (due to scalability challenges, we do not apply the LLM filter to the NoD cohort). Manual inspection of 50 samples reveals higher accuracy if the model answers yes to both prompts. We readjust the earliest diagnosis date accordingly and drop all notes on or after that date. This further reduces our T2D subset to 1717 patients. We split this data into a training-testing split in an 80:20 ratio, stratified by the diagnosis group.

Filtering Step	#NoD	#T2D	#Notes
Initial	164,910	15,140	3,708,168
Pre-diagnosis notes	164,910	4,138	3,146,140
Non-rare demographics	149,673	3,878	2,909,417
Important Note and Author-types	108,635	2,369	1,451,626
Note length	106,222	2,302	1,050,839
LLM for diagnosis date adjustment	106,222	1,717	1,039,043

Table 9: Data filtering steps and number of non-Diabetic (NoD) and T2D patients, and notes after each step for the PH corpus.

We then apply propensity-weighted matching to construct a treatment (T2D) and control (NoD) matched test sample (Rosenbaum and Rubin, 1985). We use the demographic attributes to match the covariates. For this, we first train a logistic regression model to estimate propensity scores (likelihood of being in the T2D group, given the demographic attributes) (the model achieves an AUC score of 73.40%). We use a greedy nearest-neighbor 1:1 matching without replacement to construct the final control group (NoD). For each treated sample (T2D), we find an unmatched control (NoD) with the closest propensity score, ensuring each control is used only once. Table 10 presents the covariate balance before and after the matching.

Covariate	Before	After
AGE	0.683	0.004
Binarized GENDER (M=1)	0.028	0
Binary RACE NI	-0.057	0
Binarized RACE W	-0.270	0.005
Binarized ETHNICITY (Y=1)	0.002	0

Table 10: Covariate balance using standardized mean differences (SMD) before and after propensity score matching for PH corpus test set.

A.2 MIMIC-IV T2D Data

We select the treatment group as the subset of patients with ICD code E11 (Type 2 Diabetes). We exclude patients with ICD codes from related diseases (E08-E13, O24) to exclude other types of diabetes. We coarsify race as Hispanic if Race is Hispanic/Latino, BLACK/AFRICAN AMERICAN if race is either Black/African, and drop patients with race group listed as 'other'. Thereafter, we only retain pre-diagnosis notes. MIMIC-IV notes contain discharge summaries from which we retain "Chief Complaint", "History of Present Illness", and "Discharge Instructions" sections and drop all patients with more than 20 admissions following earlier works. We then drop duplicated notes for a patient, and notes that are too short (<.05 quantile) and too long (>0.95 quantile). Finally, we apply an LLM filter on the T2D subgroup to correct the diagnosis date. We use both Llama3.1-8B and GPT-4o and manual cleaning to construct a robust subset. Finally, we apply propensity-weighted matched sampling to get a balanced control group. The data filtering steps are detailed in Table 11. The AUC score of the logistic regression model to estimate the propensity score is 63.64%. The covariate balance before and after matching is presented in Table 12. The data is split into training and test in a 90:10 ratio.

Filtering Step	#Notes	#NoD	#T2D
Initial Data	4,756,326	180,640	
No other types of diabetes	4,578,799	178,524	
race not 'other'	4,033,399	156,513	
Assign cohort	4,033,399	142,412	14,101
pre-diagnosis notes	3,220,793	142,412	5,952
Drop if more than 20 admissions	235,727	59,566	5,847
No Duplicates	184,421	54,803	5,272
Notes length in .05-.95 quantile	175,201	54,170	5,207
No duplicates across sections	174,995	54,170	5,207
LLM + matched controls Filter	14,957	2,909	2,909

Table 11: Data filtering steps and number of non-Diabetic (NoD) and T2D patients, and notes after each step for MIMIC-IV.

Covariate	Before	After
AGE	0.29	0.001
Binarized GENDER (M)	0.12	-0.001
Binary RACE (B)	0.316	0
Binarized RACE (H)	0.110	0.001
Binarized RACE (W)	-0.334	0

Table 12: Covariate balance using standardized mean differences (SMD) before and after propensity score matching on MIMIC-IV test set.

Prompt

You are a helpful assistant who can read and identify required information from given clinical notes, either mentioned explicitly or in ICD code format.
Required information: a diagnosis of **Diabetes** for the given patient.
Please reply `True` if an explicit diagnosis is mentioned in the note.
Please answer `False` if:

- there is no diagnosis,
- the note negates the diagnosis (e.g., "no diagnosis of diabetes" or "diabetes: negative"),
- if the diagnosis is associated with someone other than the patient (e.g., family).

Do not output anything other than `True`/`False`.
NOTE: `[NOTE_TEXT]`
Output:

Figure 7: Identifying known T2D in a note.

B Implementation Details

B.1 Prompts

Figure 7 presents the prompt used to identify a known diagnosis of T2D from a clinical note. Prompts for LLM-based experiments are in Figures 8–10.

B.2 Experimental Settings

For supervised learning/fine-tuning experiments, a separate model is trained for each dataset.

GNN For GNN (and CLSTM) experiments, we use graphs from the last five notes (we experimented with 1,2,5,10). We train the models using 3-fold stratified cross-validation on the training set and apply early stopping based on fold-specific validation performance, saving the best model for each fold. Hyperparameters and GNN architectural choices are selected based on mean validation performance across folds and fixed before evaluation on the held-out test set. `SAGEConv` operator from PyTorch Geometric with default settings is used. We experiment with $1-5$ GNN layers and find the best performance with $k = 2$ layers for PH and $k = 1$ for MIMIC-IV. For aggregating the nodes for graph representation, we also experiment with max pooling and pooling only the nodes corresponding to document creation time (DCT/anchor

System Prompt: Reasoner

You are a helpful medical assistant. Based on the patient's information and the history of medical notes from one or more visits, assess if this patient has a likelihood of being diagnosed with Type 2 Diabetes in the near future.
Make the assessment based on **concrete** medical evidence and how the patient's condition has progressed. Do not solely rely on age as a risk factor.

 # Prediction Format:
Risk of Type 2 Diabetes: **[True/False]
Explanation: [Provide a brief explanation of the reasoning for the prediction.]

User Prompt

Case History:

Patient is `{age}` y.o. `{race}` `{ethnicity}` `{gender}`.

```
Note Date: {date1}
{note-type1} note
{note-text1}
```

```
Note Date: {date2}
{note-type2} note
{note-text2}
```

Note: `{ethnicity}` and `{note-type}` are excluded for the MIMIC-IV. `{ethnicity}` is also excluded for the PH corpus if the value is 'NI'.

Figure 8: T2D risk from a reasoning model.

System Prompt: Fine-tuned Reasoner

You are a helpful medical assistant. Based on the patient's information and the history of medical notes from one or more visits, assess if this patient has a likelihood of being diagnosed with Type 2 Diabetes in the near future.
You must respond with either:

- `true`
- `false`

No explanation is needed. Only output one of the two labels above.

Figure 9: T2D risk Prediction without explanation. User prompt is the same as Figure 8.

System Prompt: Fine-tuned Verifier

You are a medical reasoning assistant. Given a sequence of clinical notes from a patient and an analysis predicting the risk of an imminent Type 2 Diabetes (T2D) diagnosis, judge if the analysis is correct or incorrect.
You must respond with either:

- `correct`
- `incorrect`

No explanation is needed. Only output one of the two labels above.

User Prompt

Case History:

Patient is `{age}` y.o. `{race}` `{ethnicity}` `{gender}`.

```
Note Date: {date1}
{note-type1} note
{note-text1}
```

```
Note Date: {date2}
{note-type2} note
{note-text2}
```

```
Analysis
{Reasoner Output}
```

Figure 10: Fine-tuning the verifier model in the REVEAL to verify if the T2D prediction assessment was correct based on model input and output.

nodes); neither works well. In place of a BiLSTM to model multi-document graphs, we also test LSTM, simple attention-based aggregation of cross-visit graph representations, and transformer-based aggregation, but all underperform compared to BiLSTM.

LLM Due to computational constraints, we only fine-tune a smaller Llama3.2-1B model instead of the larger Llama3.1-8B variant, which exhausts memory even with a batch size of 1 on an A100 80GB GPU. For REVEAL, we use Llama3.1-8B as the zero-shot reasoner and Llama3.2-1B as the fine-tuned verifier. We experimented with 1, 2, 3, 5, 8, and 10 notes and found that including the last two notes yields the best performance. As the notes are long and detailed, experiments with additional notes cause the LLM to underperform. In contrast, as GNN uses a structured summary of the notes in the form of temporal graphs, it can incorporate additional information effectively and gives the best performance with five notes. We also observe a slight improvement in LLM accuracy when demographic details are included, while it reduces the performance of CLSTM and HITGNN.

All LLM fine-tuning experiments are conducted on Llama3.2-1B using a batch size of 1 due to the length of input sequences and memory limitations. For each dataset, 5% random held-out validation set is reserved for hyperparameter tuning, stratified by the T2D/NoD outcome. We train the model for 30 epochs using the AdamW optimizer with a learning rate of $6e^{-06}$ and a weight decay of 0.01.[17] A linear learning rate schedule without warm-up is applied. Gradient accumulation is used with a factor of 2 to stabilize updates, given the small batch size. The loss function is standard cross-entropy with mean reduction. In our setup, LoRA is applied to the query and value projection layers (q_proj and v_proj) within the transformer blocks. We use a rank of 8, a LoRA scaling factor (alpha) of 16, and a dropout rate of 0.1 during training. This configuration results in only a small fraction of the model parameters being updated ($850K/1.24B \approx 0.07\%$), making the training process efficient (31 minutes per epoch for PH corpus and 32 for MIMIC-IV) and tractable on a single A100 80GB GPU while still allowing the model to adapt to the binary classification task.

In the fine-tuning experiments, the pretrained Llama3.2-1B model initially performs near chance level, with a balanced accuracy of 54.50% on the validation set and 0.8% of outputs falling outside the valid label set. After fine-tuning, the best model achieves an accuracy of 70.50% on the validation set with no invalid predictions. On MIMIC-IV, the initial validation set accuracy is much lower at 44.00%, with 18.1% of predictions being invalid. This improves to 62.50% and zero invalid predictions after fine-tuning.

For proper supervision for REVEAL, we restrict training to cases where the 8B reasoner produces both 'true' and 'false' predictions across the five runs. This ensures that each case includes both valid and invalid reasoning paths, enabling the verifier to learn discriminative patterns. We have 5,550 training examples for the PH corpus and 9,180 for MIMIC-IV. A 5% random held-out validation set stratified by the 'correct/incorrect' label is reserved from each. After each epoch, the model is evaluated on the validation set, and the best-performing model checkpoint is saved. The verifier 1B model starts with a base accuracy of 43.2% and 1% invalid predictions on the PH corpus, and the accuracy improves to 84.9% and no invalid predictions in 30 epochs. For MIMIC-IV, the accuracy improves from 58.8% initially to 67.5%.

C Entity, Temporal relation, Coreference Annotation Details

Entity Type	Freq.
PROBLEM	373
TREATMENT	120
TEST	325
OCCURRENCE	489
CLINICAL_DEPT	41
EVIDENTIAL	90
DATE	214
TIME	64
DURATION	67
FREQUENCY	74

Relation Type	Freq.
Before	83,392
After	83,392
Simultaneous	112,776

Table 13: Annotation details for 9 PH corpus notes comprising 10,936 tokens.

Entity Type	P	R	F_1
PROBLEM	78.2	37.2	48.3
TREATMENT	54.6	48.3	50.7
TEST	47.0	63.3	48.6
CLINICAL_DEPT	29.2	26.5	27.5
DATE	98.7	67.2	77.0
Selected 5 Types (Micro)	77.2	52.9	61.9
All 10 Types (Micro)	77.2	40.2	52.2

Table 14: Coreference resolution performance.

[17] We experimented with learning rates of $5e^{-05}$ and $1e^{-06}$ that had lower performance.

Model Pair	PH Corpus						MIMIC-IV					
	AUC	F_1	T2D		NoD		AUC	F_1	T2D		NoD	
			P	R	P	R			P	R	P	R
ReVeAL vs. Llama3.2-1B-ft	-0.07*	0.04	0.07***	-0.18***	-0.05*	0.23***	-0.11*	-0.08*	-0.09	-0.11*	-0.06*	-0.03
	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.05)	(0.04)	(0.06)	(0.06)	(0.03)	(0.05)
CLSTM vs Llama3.2-1B-ft	0.03	0.04	0.04**	-0.02	0.02	0.09**	-0.01	0	0	0.01	0	-0.01
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.01)	(0.02)	(0.01)	(0.01)	(0.02)
CLSTM vs ReVeAL	0.10***	0	-0.04	0.16***	0.06**	-0.14***	0.10	0.08	0.09	0.12*	0.06	0.02
	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.05)	(0.04)	(0.06)	(0.06)	(0.03)	(0.05)
ReVeAL vs. HiTGNN	-0.14***	-0.02	0.02	-0.16***	-0.07**	0.11***	-0.14**	-0.10**	-0.02	-0.16**	-0.11*	-0.16**
	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.05)	(0.04)	(0.05)	(0.06)	(0.05)	(0.06)
Llama3.2-1B-ft vs. HiTGNN	-0.07***	-0.06**	-0.05**	0.02	-0.03	-0.12***	-0.03	-0.03	-0.01	-0.05*	-0.02	-0.05*
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
CLSTM vs HiTGNN	-0.04*	-0.02	-0.02	0.0	-0.01	-0.03	-0.04*	-0.02	-0.02	-0.04	-0.02	0.00
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)	(0.03)

Table 15: Pairwise model performance differences for top four models overall (first model minus second), based on bootstrap resampling test. Statistical significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 13 shows the frequency of annotated entities and relations. This annotation effort required ≈ 16 person-hours each for entity and relation annotation, resulting in $1{,}857$ entities and $279,560$ relation pairs (Appendix C). Table 14 shows the performance evaluation of coreference resolution with entity linking using Metamap and Microsoft Recognizers-Text. While the performance on DATE is quite good, there is room for improvement across types. Except for TEST entities, the precision is high with lower recall, indicating the predicted clusters are more fragmented compared to the gold ones. The performance on selected entity types is higher than others, such as EVIDENTIAL, OCCURRENCE, which were not mapped in the HiTGNN pipeline as they cause further noise. For example, a matching string 'reports' will be considered coreferential by Metamap, whether it refers to the same reporting incidence (e.g., the same reporter or what is being reported). A manual inspection of clusters from 150 notes concurs with these observations and reveals some sources of error. E.g., recurrence of a symptom doesn't indicate coreference, but is marked so. Recognizers-text is unable to normalize entries such as 'pod #3', 'yesterday', 'this morning', with reference to the note date. Nevertheless, this linking still leads to improvement on the downstream tasks and provides a reliable foundation for further refinement.

D Statistical Significance

To assess whether the observed differences in performance metrics presented in Table 2 are statistically significant, we perform non-parametric bootstrap resampling and compare each pair of the top models—CLSTM, ReVeAL, HiTGNN, and Llama3.2-1B-ft. Specifically, for each metric (e.g., macro-F_1, AUC, precision, recall), we repeatedly sample subsets of the test set with replacement and compute the performance difference between model pairs over 10,000 bootstrap replicates. This generates an empirical distribution of differences for each metric and model pair. From this distribution, we computed the mean difference, the sample standard deviation, and a 95% confidence interval (CI) using the 2.5^{th} and 97.5^{th} percentiles. Statistical significance is determined based on the proportion of bootstrapped differences falling below or above zero, yielding a two-tailed p-value. Significance levels are reported using the convention: p<0.05 (*), p<0.01 (**), and p<0.001 (***). This approach makes no parametric assumptions and robustly quantifies uncertainty in model comparisons. The results are summarized in Table 15.

E Ablations using T2D Recall

Prediction Horizons. Figure 11 presents the trends for T2D group recall across the top four models (Llama3.2-1B-ft, ReVeAL, CLSTM and HiTGNN with varying prediction windows. HiTGNN performs comparably in the immediate prediction window for the PH corpus, and has more stable performance across horizons for MIMIC-IV. ReVeAL has poor performance on the PH corpus but shows better performance in the medium horizon (9 months-4.5 years) for MIMIC-IV.

Node Initialization and Subgraph Ablations. Figures 12a and 12b show the T2D recall trends across embedding and subgraph variations. In the PH corpus, text+KG node initialization consistently improves AUC and T2D recall, while KG-only edges yield the highest AUC; both temporal+KG improve T2D recall. In contrast, for MIMIC-IV, while both node embeddings and subgraphs lead to

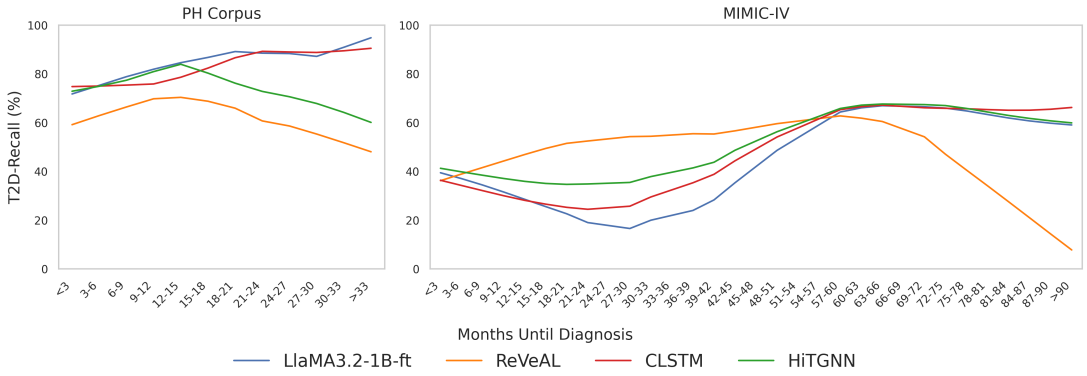
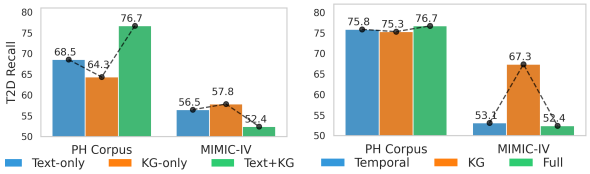


Figure 11: T2D recall across consecutive 3-month prediction horizons.



(a) Embedding ablations. (b) Subgraphs ablations.

Figure 12: HiTGNN ablations with T2D Recall.

the highest AUC, KG-only node initialization and KG-only edges yield the highest T2D recall at the standard decision threshold (0.5).

F LLM Fairness with and without Demographics

		PH Corpus			MIMIC-IV		
		G	Dem	no Dem	G	Dem	no Dem
Ethnicity	DPD	H	-0.05	-0.05			
		N	0.05	0.05			
	EOD	H	-0.06	-0.07			
		N	0.06	0.07			
Gender	DPD	F	0.05	-0.03	F	-0.01	-0.04
		M	-0.05	0.03	M	0.01	0.04
	EOD	F	0.05	0.02	F	-0.04	-0.07
		M	-0.05	-0.02	M	0.04	0.07
Race	DPD	B	0.09	0.01	B	0.04	-0.12
		W	-0.07	-0.02	W	-0.06	0.10
		NI	-0.07	-0.01	H	0.04	-0.01
					A	0.09	0.00
	EOD	B	0.13	0.03	B	-0.03	-0.23
		W	-0.08	-0.03	W	-0.08	0.12
		NI	-0.10	-0.02	H	0.09	0.09
					A	0.41	0.06

Table 16: DPD/EOD for Llama3.1-8B with and without demographic conditioning.

G Clinical Evaluation of REVEAL Explanations

Case 1: Expert-REVEAL agree against ground truth.

Ground Truth	Expert	REVEAL
Low	High	High

Expert Rationale: Family history of diabetes and existing obesity indicate an elevated risk.

REVEAL Explanation: The patient has a BMI of 33.4 kg/m^2 which falls into the obese category (Class 2 obesity). Her BMI is consistently high across the two visit notes. Obesity is a significant risk factor for the development of Type 2 Diabetes, as it is associated with insulin resistance and other metabolic disturbances that contribute to the disease. The patient's family history also includes a first-degree relative (mother) with diabetes mellitus type 2, which further increases her risk.

Case 2: Expert-REVEAL Agree with Ground Truth.

Ground Truth	Expert	REVEAL
High	High	High

Expert Rationale: Combination of African American, age, HTN, and high cholesterol raises my concern.

REVEAL Explanation: The patient is a [AGE-Redacted]-year-old black female with a history of hypertension and hyperlipidemia, both of which are significant risk factors for Type 2 diabetes. Additionally, the patient is taking pravastatin, a medication commonly used to treat high cholesterol, which is often prescribed to individuals at risk for Type 2 diabetes. While the patient's current weight is not mentioned, her use of medications for hypertension and hyperlipidemia, in conjunction with her age and ethnicity, contributes to a higher risk for developing Type 2 diabetes.

Case 3a: Expert-REVEAL Disagree		
Ground Truth	**Expert**	**REVEAL**
Low	Low	High

Expert Rationale: She has multiple factors elevating her risk, HTN, and some other comorbidities, but she is still relatively young. I agree with AI's explanation that she is at increasing risk, but I would not say she is at high risk yet. Not sure about the concern about pseudoephedrine. Also "Hypertriglyceridemia**: The patient's glucose level is elevated...potential insulin resistance" is inaccurate.

REVEAL Explanation: The patient's medical history and laboratory results from YEAR indicate several risk factors for Type 2 diabetes. 1. **Hypertriglyceridemia**: The patient's glucose level is elevated at 106 mg/dL (normal range: 70-99 mg/dL). While this is not extremely high, it is a subtle indication of potential insulin resistance. 2. **High CK levels**: The patient's creatine kinase (CK) level is elevated at 213 U/L (normal range: 24-195 U/L). Elevated CK levels can be a marker of metabolic syndrome, which is a cluster of conditions that increase the risk of developing Type 2 diabetes and cardiovascular disease. 3. **Hypertension**: The patient has a history of hypertension (HTN) and was treated with medication during her ED visit. Hypertension is a well-established risk factor for Type 2 diabetes. 4. **Family Medicine note**: The patient was instructed to stop taking pseudophedrine-containing products, which can raise blood pressure and potentially increase the risk of Type 2 diabetes. 5. **Age**: While age is a risk factor for Type 2 diabetes, the patient is 40 years old, which is relatively young for this diagnosis. However, combined with the other factors mentioned, it still increases her risk. 6. **Asthma and cough**: The patient has a history of asthma and was experiencing a cough at the time of her ED visit. These conditions may indicate an underlying inflammatory process that could be related to insulin resistance. In conclusion, while the patient's glucose level is not extremely high, the combination of elevated CK levels, hypertension, and other risk factors suggests that she is at a higher risk for developing Type 2 diabetes in the near future.

Case 3b: Expert-REVEAL Disagree		
Ground Truth	**Expert**	**REVEAL**
High	High	Low

Expert Rationale: Two strong risk factors, severe obesity and Hispanic ethnicity, should be weighted more.

REVEAL Explanation: The patient has a history of obesity, as documented in his notes from YEAR1 and YEAR2. However, despite his obesity, his laboratory results from YEAR2 show that his HbA1c (glycated hemoglobin) level is not elevated, indicating that his blood sugar control is within a normal range. Additionally, there is no mention of any other risk factors for Type 2 Diabetes, such as a family history of diabetes, hypertension, or a history of gestational diabetes. Furthermore, the patient's BMI, which was measured in YEAR2, is 55.7 kg/m^2, which is severely obese, but not significantly higher than his previous measurement. Given the absence of other risk factors and the patient's normal HbA1c level, the likelihood of him being diagnosed with Type 2 Diabetes in the near future is low. However, it is essential to continue monitoring his blood sugar levels and body weight to assess any potential changes in his risk profile.