HERMES: Contrast-Aware Knowledge Graph Reasoning from Clinical Notes for Patient Outcome Prediction

Gia-Bach Nguyen[1,2], Tuan-Cuong Vuong[3], Trang Mai Xuan[3], Duy Quoc Ngo[4], Tien-Cuong Nguyen[5], Huan Vu[1], and Thien Van Luong[1]

[1] Business AI Lab, College of Technology, National Economics University, Vietnam
[2] Hanoi University of Science and Technology, Hanoi, Vietnam
[3] A2I Lab, Phenikaa School of Computing, Phenikaa University, Hanoi, Vietnam
[4] Department of Head and Neck Surgery, Vietnam National Cancer Hospital & Hanoi Medical University, Hanoi, Vietnam
[5] VNPT AI, VNPT Group, Hanoi, Vietnam
bach.ng2416130@sis.hust.edu.vn, {cuong.vuongtuan, trang.maixuan}@phenikaa-uni.edu.vn, duyyhn@gmail.com, nguyentiencuong@vnpt.vn, {huanv, thienlv}@neu.edu.vn

Abstract. Clinical predictive models often rely on structured Electronic Health Record data, such as time-series and procedure codes. While recent approaches have begun leveraging unstructured clinical notes, they typically encode them as flat sequences, which may lose explicit relational and temporal structure present in clinical narratives. In response, we propose HERMES, a graph-based framework that operates exclusively on clinical text while preserving clinical relationships. This approach builds on two key ideas. First, personalized Knowledge Graphs (KGs) are constructed through Large-Language-Model-guided extraction from clinical notes with Contrastive Logic Modeling that explicitly captures temporal dynamics and treatment failures and changes in outcomes. Second, a Graph Attention Network synthesizes patient representations through graph-based learning over the KGs. Experiments on MIMIC-III and MIMIC-IV for in-hospital mortality and 30-day readmission prediction show that HERMES consistently outperforms strong text-only baselines. Our findings demonstrate that explicit relational modeling with Contrastive Logic Modeling significantly advances predictive performance.

Keywords: Clinical Outcome Prediction · Knowledge Graphs · Large Language Models · Graph Attention Networks · Graph Neural Networks

1 Introduction

Electronic Health Records (EHRs) enable data-driven risk prediction, including in-hospital mortality and 30-day readmission forecasting. In the ICU, clinical evidence is documented in narrative notes capturing clinician reasoning, patient status, and treatment responses. However, many predictive pipelines prioritize

structured signals and treat notes as auxiliary inputs, underutilizing the clinical knowledge embedded in narratives.

Recent clinical language models (LMs) [15, 3] advance outcome prediction from notes, yet most approaches remain token-centric and implicitly represent clinical dependencies through sequence representations. Early prediction models treated clinical narratives as flat sequences, using Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers to represent token streams [15, 9]. Although long-context variants better accommodate lengthy ICU documentation [9], these approaches implicitly model clinical dependencies. Hierarchical encoders [4] and token-level graph models introduce local connectivity but remain token-focused. This paradigm cannot explicitly capture clinically meaningful relations, such as symptom-to-diagnosis evidence, treatment-to-response dynamics, and clinical contradictions. Consequently, models fail to expose the structured clinical logic central to adverse outcomes.

Knowledge-enhanced methods advance representations by injecting external medical knowledge from ontologies or curated bases [2, 6], yet they depend on structured codes rather than free-text. Traditional machine learning approaches, such as logistic regression with TF-IDF [12], remain competitive baselines due to simplicity and robustness, but treat text as bags-of-words and cannot capture temporal dependencies. Recent advances in Large Language Models (LLMs) have enabled strong performance in clinical information extraction [1]. Frameworks like EMERGE [17] and GraphCare [6] use LLMs to construct Knowledge Graphs (KGs) for prediction but depend on structured EHR data or external KGs.

In response, this work proposes HERMES, a text-only clinical prediction approach that combines narrative understanding with graph-based reasoning. HERMES differs in three key ways. First, it constructs personalized KGs exclusively from clinical text using LLM-guided extraction with lightweight Named Entity Recognition (NER) model, without external codes or knowledge bases. Second, it employs Contrastive Logic Modeling to prioritize clinical dynamics-temporal changes, treatment contradictions, and adverse outcomes that predict deterioration. Third, it achieves consistent gains over text-only baselines on two ICU benchmarks.

Specifically, HERMES constructs personalized KGs from clinical text through a two-stage extraction pipeline with Contrastive Logic Modeling that emphasizes temporal dynamics. The pipeline extracts entities and their clinical relations while identifying critical patterns-contradictions between findings and treatments, treatment failures, and temporal changes-predictive of adverse outcomes. The KGs are then processed by a Graph Attention Network (GAT) [14], producing patient-level representations for outcome prediction.

We evaluate HERMES on MIMIC-III and MIMIC-IV for in-hospital mortality and 30-day readmission prediction. Across both datasets and tasks, the proposed text-only approach consistently outperforms text-only baselines, indicating that explicit relational modeling and attention-based aggregation complement sequence representations. Ablation results characterize the effects of

different Graph Neural Network (GNN) backbones and pooling strategies, supporting GAT with max pooling as the optimal encoder.

2 Methodology

We consider each patient's first ICU stay from their first visit. Each stay contains a set of clinical notes $\mathcal{N} = \{n_1, n_2, \ldots, n_{|\mathcal{N}|}\}$. Given $\mathcal{N}$, we aim to predict the patient's clinical outcome $\hat{y} \in \{0, 1\}$ for two clinical prediction tasks: in-hospital mortality and 30-day readmission. Throughout the framework, we use the concatenated clinical notes, $\mathcal{N}'$, as input to both NER and LLM-based extraction, leveraging the extended context window to preserve semantic relationships across the patient's clinical documentation.

Fig. 1 shows the overall architecture of our proposed HERMES framework.

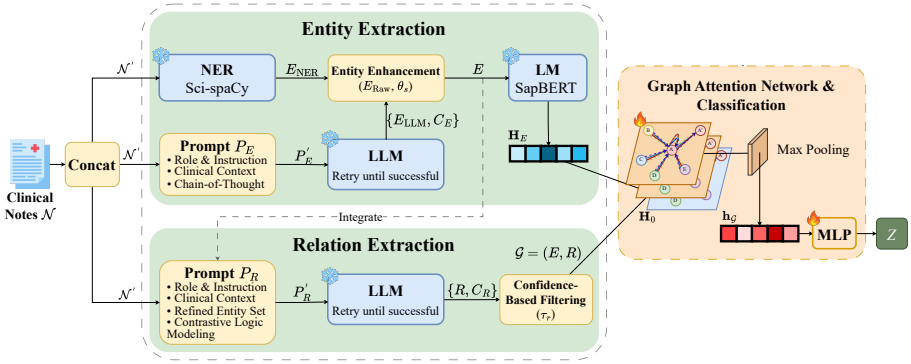

Fig. 1. Overall architecture of our proposed HERMES framework. The two extraction modules, Entity Extraction and Relation Extraction, process the input clinical notes $\mathcal{N}$, generating entities E and relations R for each personalized KG. The GAT then synthesizes KG representations for downstream prediction tasks.

2.1 Entity Extraction

Performing NER on unstructured data is challenging. We address this by combining LLM-guided extraction with clinical-domain spaCy NER [6] to maximize entity coverage. The raw extracted entity set is defined as:

$$E_{\text{raw}} = E_{\text{NER}} \cup E_{\text{LLM}}, \tag{1}$$

where E_{NER} and E_{LLM} represent the entity sets obtained from spaCy-based and LLM-guided extraction, respectively. Each entity is represented as a text string, and similarity comparisons are computed using cosine similarity based on text embeddings.

[6] https://allenai.github.io/scispacy/

Chain-of-Thought Reasoning Our LLM-guided extraction [1] uses the prompt template P_E (Fig. 2) that follows a chain-of-thought method [16], requiring the LLM to reason over extractions and assign confidence scores to each entity. Let C_E denote the set of entity confidence scores, where $C_E(e) \in [0, 1]$ represents the confidence score for entity e [13], which are incorporated into Eq. (2) and Eq. (3) to guide entity consolidation, prioritizing higher-confidence candidates during merging to form the final entity set E. We also preserve negation aspects (for example, "no pneumothorax", "denies pain") to maintain clinical semantics. Failed extractions are retried until successful, ensuring no entity sets are empty.

[ROLE]
You are a clinical NLP specialist for patient knowledge graph construction.

[OBJECTIVE]
Extract a comprehensive set of clinical entities capturing diagnoses, evidence, and interventions from clinical text.

[COVERAGE]
Ensure coverage of: (1) Diagnoses and chief complaints; (2) Medications, procedures, lines, and devices; (3) Abnormal labs, vitals, and exam findings; (4) Temporal markers (e.g., "Post-op Day 1", "0400")
Target density: 10–20 entities per note segment. Prefer recall over precision.

[INPUT]
Clinical Context: {text}

[ENTITY TYPES]
PROBLEM: Diagnoses, symptoms, injuries
INTERVENTION: Medications, procedures, lines, tubes
OBSERVATION: Exam findings, qualitative states
MEASUREMENT: Labs and vitals with values
ANATOMY: Body parts and laterality
TEMPORAL: Timepoints and durations

[RULES]
• Expand abbreviations in brackets.
• Explicitly include negation (e.g., "No edema").
• Exclude generic terms unless clinically meaningful.
• Output format: entity_text | entity_type | confidence (0–1)

[EXAMPLE]
Input: "Neuro: MAE, PERRL. CV: SB 40s. No edema."
Output:
MAE [Moves All Extremities] | OBSERVATION | 0.98
PERRL [Pupils Equal Round Reactive to Light] | OBSERVATION | 0.98
SB [Sinus Bradycardia] | PROBLEM | 0.95
40s | MEASUREMENT | 0.90
No edema | OBSERVATION | 0.99

Fig. 2. Prompt P_E for entity extraction.

Entity Enhancement We normalize the raw entities (whitespace trimming, lowercasing) to unify their representation format. To consolidate the raw entity set E_{raw} and enhance entity quality, we implement a similarity-based merging algorithm. We define an entity scoring function:

$$\text{Score}(e) = C_E(e) + \text{bonus}(e), \tag{2}$$

where $C_E(e)$ represents the confidence value assigned by the LLM during extraction (the confidence for NER is fixed), and bonus(e) denotes an additional bonus term that accounts for extraction source and entity informativeness.

For each pair of entities $e_i, e_j \in E_{\text{raw}}$ where $\text{sim}(e_i, e_j) \geq \theta_s$, we apply:

$$\texttt{Merge}(e_i, e_j) = \begin{cases} e_i, & \text{if Score}(e_i) > \text{Score}(e_j) \\ e_j, & \text{otherwise} \end{cases}, \tag{3}$$

where θ_s represents the similarity threshold. This merging process consolidates similar entities while retaining the higher-scoring candidate according to Eq. (3), resulting in the final entity set E. We then employ a medical-domain language model [10] to obtain the dense representation for each entity in the set E:

$$\mathbf{H}_E = \text{TextEncoder}(E), \tag{4}$$

where $\mathbf{H}_E$ denotes the entity embeddings of the final set E obtained from the language model.

2.2 Relation Extraction

[ROLE]
You are a "clinical logic engine" for constructing patient knowledge graphs.

[OBJECTIVE]
Infer dynamic and causal relationships between extracted entities to form a coherent patient narrative.
Target density: 10–15 relations. Link each PROBLEM to an INTERVENTION or ANATOMY.

[INPUT]
Clinical Context: {text}
Entities: {entities}

[RELATION TYPES]
TIER 1: DYNAMICS
trends_to: Value change over time
contrasts_with: Conflict or contradiction
resolved_by: Successful intervention
worsened_by: Adverse cause or exposure
TIER 2: CLINICAL LOGIC
indicates: Sign or symptom implies diagnosis
managed_with: Problem treated by intervention
reveals: Test shows finding
located_in: Problem or device mapped to anatomy

[RULES]
• Ensure each entity participates in at least one relation.
• Explicitly encode temporal change with trends_to.
• Use contrasts_with for failed or contradictory evidence.
• Output format: entity1 | relation_type | entity2 | confidence (0–1)

[EXAMPLES]
Input: “Hct dropped from 30 to 24 despite 1u PRBC.”
Hct 30 | trends_to | Hct 24 | 0.98
1u PRBC | contrasts_with | Hct 24 | 0.95
1u PRBC | managed_with | Hct 30 | 0.90

Fig. 3. Prompt P_R for relation extraction.

While entity extraction identifies key clinical concepts, patient health trajectories depend on how these concepts interact. We employ a second LLM-guided stage for relation extraction, aiming to reconstruct the underlying clinical logic and temporal semantics. We obtain the relation set R_{LLM} by using the prompt template P_R (Fig. 3) with $\mathcal{N}'$ and integrating the refined entity set E from the previous stage:

$$R_{\text{LLM}} = \texttt{LLM}(P_R, \mathcal{N}', E). \tag{5}$$

By anchoring relations to extracted entities, we ensure the KG remains grounded in the extracted nodes while allowing the LLM to infer clinical relations from full patient context.

Contrastive Logic Modeling Contrastive Logic Modeling captures clinically important nuances by distinguishing beneficial from harmful treatments. We instruct the LLM to model two relation types: (1) Dynamics capture temporal changes, clinical contradictions, and treatment outcomes-highly predictive of adverse events; (2) Clinical Logic represents static medical knowledge linking observations, diagnoses, interventions, and anatomy. Together, these ensure KGs encode both patient status and temporal evolution.

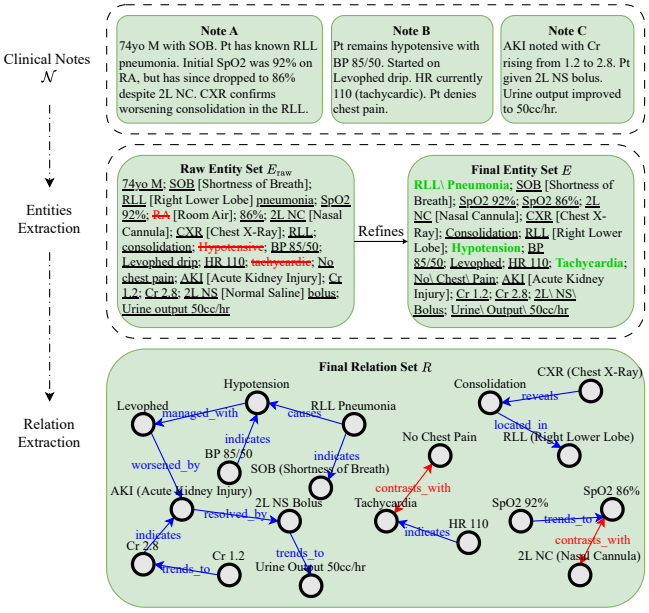

Fig. 4. Process of knowledge graph construction in the HERMES framework. Blue arrows indicate standard relations, while red arrows indicate contrastive relations. The contrastive relations are bi-directional by default.

Confidence-Based Filtering To ensure reliable relation extraction for downstream prediction tasks, our relation refinement process employs a confidence-based filtering mechanism [13] to remove ambiguous relations. Let C_R denote the set of relation confidence scores, where $C_R(r) \in [0, 1]$ represents the confidence score for relation $r \in R_{\text{LLM}}$. The final relation set R is defined as:

$$R = \{r \in R_{\text{LLM}} \mid C_R(r) \geq \tau_r\}, \tag{6}$$

where R denotes the relation set, and τ_r is the relation confidence threshold. R defines KG connectivity $\mathcal{G} = (E, R)$, transforming unstructured notes into a structured knowledge representation. Contrastive Logic Modeling shapes the KG topology, capturing clinical dependencies and temporal evolution.

The process of knowledge graph construction is illustrated in Fig. 4.

2.3 Representation Learning

We encode each patient KG using a two-layer GAT with residual connections and layer normalization, followed by graph pooling and an MLP classifier for outcome prediction. GAT learns attention weights over neighboring nodes, focusing on informative entities without explicitly using relation types.

Graph Attention Network. Let n be the number of entity nodes and d be the node embedding dimension. For $\ell \in \{0, 1, 2\}$, we define $\mathbf{H}_\ell \in \mathbb{R}^{n \times d}$ as the node feature matrix at GAT layer ℓ, whose i-th row is the node embedding $\mathbf{h}_i^{(\ell)} \in \mathbb{R}^d$. The input matrix $\mathbf{H}_0 \in \mathbb{R}^{n \times d}$ is initialized by incorporating the entity embeddings $\mathbf{H}_E$ (Eq. (4)) with the KG structure $\mathcal{G}$, where each node's initial representation is a vector from $\mathbf{H}_0$. For each node i, we define its neighbor set $\mathcal{S}(i)$ as the set of nodes connected to i in the KG, and we include a self-loop so that a node can retain its own information during message passing.

For each layer $\ell \in \{1, 2\}$, node i, and a neighbor $j \in \mathcal{S}(i)$, GAT first computes an unnormalized attention score $s_{ij}^{(\ell)} \in \mathbb{R}$:

$$s_{ij}^{(\ell)} = \text{LeakyReLU}\left(\mathbf{a}^T(\mathbf{W}\mathbf{h}_i^{(\ell-1)} \| \mathbf{W}\mathbf{h}_j^{(\ell-1)})\right), \tag{7}$$

where $\mathbf{W} \in \mathbb{R}^{d \times d}$ is a learnable weight matrix, $\mathbf{a} \in \mathbb{R}^{2d}$ is a learnable attention vector, and $\|$ denotes concatenation. These scores are normalized over the neighborhood of i using a softmax to obtain attention coefficients $\alpha_{ij}^{(\ell)} \in \mathbb{R}$:

$$\alpha_{ij}^{(\ell)} = \frac{\exp(s_{ij}^{(\ell)})}{\sum_{k \in \mathcal{S}(i)} \exp(s_{ik}^{(\ell)})}, \tag{8}$$

where the denominator sums over all neighbors k of i (including i itself). The coefficient $\alpha_{ij}^{(\ell)}$ controls how much information node i receives from node j.

Node representations are then updated by aggregating transformed neighbor features weighted by the attention coefficients:

$$\mathbf{h}_i^{(\ell)} = \phi\left(\sum_{j\in\mathcal{S}(i)} \alpha_{ij}^{(\ell)} \mathbf{W}\mathbf{h}_j^{(\ell-1)}\right), \tag{9}$$

where ϕ is a non-linear activation function, which is ReLU in our implementation. The framework uses multi-head attention with K heads and concatenates head outputs to form the layer output.

Let GAT$(\cdot)$ denote applying Eq. (9) to all nodes of the layer in parallel. We stack two GAT layers with residual connections and layer normalization:

$$\mathbf{H}_\ell = \text{LayerNorm}(\mathbf{H}_{\ell-1} + \text{GAT}(\mathbf{H}_{\ell-1}))\,, \tag{10}$$

where the initial embeddings are given by $\mathbf{H}_0$. After two layers, $\mathbf{H}_2$ contains the final node representations.

Pooling and Classification. To obtain the graph representation for the patient, we aggregate the node representations in $\mathbf{H}_2$ using max pooling:

$$\mathbf{h}_\mathcal{G} = \text{MaxPool}(\mathbf{H}_2)\,, \tag{11}$$

where $\mathbf{h}_\mathcal{G} \in \mathbb{R}^d$ is the graph representation for the patient. This representation is passed to a two-layer MLP classifier to produce the predicted probability $\hat{y}$:

$$\hat{y} = \sigma(\text{MLP}(\mathbf{h}_\mathcal{G}))\,. \tag{12}$$

Here, MLP$(\cdot)$ denotes a two-layer multilayer perceptron that outputs a scalar logit, and $\sigma(\cdot)$ is the sigmoid function for binary classification.

We optimize the model using binary cross-entropy (BCE) loss for both mortality and readmission prediction:

$$\mathcal{L}(\hat{y}, y) = -\frac{1}{M}\sum_{i=1}^{M}\left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right], \tag{13}$$

where M is the number of patients, $y_i \in \{0,1\}$ is the ground-truth label, and $\hat{y}_i$ is the predicted probability.

By learning over personalized KGs constructed from clinical notes, the GAT encoder integrates entity information with graph-structured clinical associations, producing patient representations that support accurate outcome prediction.

3 Experiments

3.1 Experimental Setups

Evaluation Datasets We evaluate on MIMIC-III [8] and MIMIC-IV [7] for in-hospital mortality and 30-day readmission prediction. Following the experimental setup of EMERGE [17], we use the first 48 hours of each ICU stay,

concatenating notes from consecutive 12-hour segments. This setup enables controlled comparison and leverages the extended context window to capture clinical trajectories. The datasets statistics are shown in Table 1. $\textbf{Label}_{\text{Mortality}}$ and $\textbf{Label}_{\text{Readmission}}$ denote the number (and percentage) of positive labels for in-hospital mortality and 30-day readmission, respectively.

Table 1. Dataset statistics after preprocessing

Dataset	Split	Samples	$\textbf{Label}_{\text{Mortality}}$	$\textbf{Label}_{\text{Readmission}}$
MIMIC-III	Train	11,200 (70.00%)	1,538 (13.7%)	2,024 (18.1%)
	Val	1,600 (10.00%)	220 (13.8%)	289 (18.1%)
	Test	3,200 (20.00%)	440 (13.8%)	578 (18.1%)
MIMIC-IV	Train	11,200 (70.00%)	2,082 (18.6%)	2,741 (24.5%)
	Val	1,600 (10.00%)	297 (18.6%)	391 (24.4%)
	Test	3,200 (20.00%)	595 (18.6%)	783 (24.5%)

The two datasets differ substantially in documentation density: MIMIC-III contains 168,194 total notes (mean 10.5 per patient) with 221.5 words per note on average, while MIMIC-IV has 51,297 notes (mean 3.2 per patient) with 196.4 words per note. This translates to significantly denser extracted KGs, where MIMIC-III yields a mean of 133.7 entities and 1,263.4 relations per patient compared to 47.4 entities and 61.7 relations in MIMIC-IV. This reflects distributional differences between the two datasets.

Evaluation Metrics We evaluate using AUROC (threshold-independent discrimination), AUPRC (emphasizes positive-class performance under imbalance), and min(+P, Se) (minimum of precision and sensitivity). Higher values indicate better performance. In all results tables, **Bold** and Underlined denote best and second-best results, respectively.

Baseline Models We compare our approach against a diverse set of baseline methods. These include pre-trained clinical language models such as Clinical-BERT [15], PubMedBERT [3], and Clinical-Longformer [9]; Note-HCR [4], a hierarchical RNN-based model; and Logistic Regression with TF-IDF features [12], a traditional machine learning baseline that can process full documents but lacks semantic understanding. We further evaluate different GNN variants, including Graph Convolutional Network (GCN), GAT, and Relational Graph Convolutional Network (RGCN), to assess their effectiveness.

Implementation Details Experiments are conducted on a single NVIDIA RTX 3060 GPU (12 GB). We use SapBERT [10] for entity embedding and Mistral-Small-3.2-24B-Instruct[7] for LLM-guided extraction [1], both models are

[7] https://docs.mistral.ai/models/mistral-small-3-2-25-06

kept frozen. The GAT encoder has two layers with eight attention heads and ReLU activation, hidden dimension 128, and max pooling. The two-layer MLP classifier uses GELU [5] activation. Training uses AdamW [11] with learning rate 0.001, batch size 64, dropout 0.2, and patience 10. For extraction thresholds, we set entity similarity $\theta_s = 0.85$ and relation confidence $\tau_r = 0.7$. Results are reported as mean $\pm$ std via bootstrapping (1,000 resamples).

3.2 Experimental Results

We compare HERMES against representative text-only baselines on in-hospital mortality and 30-day readmission prediction in MIMIC-III and MIMIC-IV, as shown in Table 2.

Table 2. Performance comparison between HERMES and different baselines on MIMIC-III and MIMIC-IV datasets

Dataset	Method	Mortality			Readmission		
		AUROC(↑)	AUPRC(↑)	min(+P, Se)(↑)	AUROC(↑)	AUPRC(↑)	min(+P, Se)(↑)
MIMIC-III	Clinical-LongFormer [9]	70.93±0.96	25.84±1.58	28.62±1.42	66.60±0.97	28.89±1.55	31.27±1.60
	ClinicalBERT [15]	70.99±2.44	26.95±3.77	29.58±3.86	67.72±2.35	30.50±3.43	33.04±3.26
	PubMedBERT [3]	67.43±2.74	23.72±3.09	27.95±3.65	64.37±2.49	26.88±2.89	30.69±3.47
	Note-HCR [4]	81.99±1.35	49.63±3.00	47.86±2.00	74.41±1.20	45.68±1.83	44.78±1.86
	LR + TF-IDF [12]	83.54±1.02	50.67±2.55	48.63±2.14	76.76±1.12	47.83±2.19	47.10±1.85
	HERMES (Ours)	**85.92±0.94**	**55.49±2.53**	**51.52±2.12**	**77.75±1.11**	**49.92±2.16**	**47.25±1.88**
MIMIC-IV	Clinical-LongFormer [9]	69.51±0.94	29.45±1.42	33.58±2.22	65.89±1.03	35.30±1.48	37.48±1.42
	ClinicalBERT [15]	75.51±2.01	43.95±4.01	43.47±3.47	69.57±2.11	44.60±3.49	44.52±3.12
	PubMedBERT [3]	74.05±2.14	43.37±4.07	44.02±3.43	70.13±2.09	44.70±3.71	44.20±3.27
	Note-HCR [4]	80.47±0.91	51.78±2.02	49.07±1.72	73.20±1.03	48.94±1.85	47.50±1.66
	LR + TF-IDF [12]	81.83±0.85	55.27±1.98	50.56±1.69	74.78±1.01	52.62±1.81	49.54±1.65
	HERMES (Ours)	**83.24±0.89**	**58.83±1.95**	**53.00±1.70**	**77.31±0.99**	**56.88±1.87**	**53.73±1.53**

Overall, HERMES achieves the best results across both datasets and tasks. Compared with the strongest non-graph baseline (LR + TF-IDF), HERMES improves AUROC/AUPRC by +2.38/+4.82 points on MIMIC-III mortality and by +1.41/+3.56 points on MIMIC-IV mortality, while also yielding consistent gains for readmission (+2.53 AUROC and +4.26 AUPRC on MIMIC-IV). These improvements suggest that explicitly modeling patient-specific relations extracted from notes provides a strong predictive signal.

It is also worth noting that the traditional ML baseline (TF-IDF + Logistic Regression) [12] remains quite competitive. This is likely due to its ability to pick up key prognostic cues from early ICU notes, such as clinician phrasing and care plans, especially when multiple note types are concatenated over 48 hours. Unlike HERMES, however, this performance is mainly driven by surface-level textual signals rather than capturing the underlying clinical and temporal relationships, such as evolving findings, treatment failures, and contradictions.

3.3 Ablation Studies

Comparing Different GNN Backbones We study the impact of the graph encoder by swapping the GNN backbone and pooling strategy while keeping the

framework fixed. We compare GCN, GAT, and RGCN under mean and max pooling, evaluated on the same prediction tasks, as shown in Table 3.

Table 3. Performance comparison of different GNNs in HERMES on MIMIC-III and MIMIC-IV datasets

Dataset	Method	Mortality			Readmission		
		AUROC(↑)	AUPRC(↑)	min(+P, Se)(↑)	AUROC(↑)	AUPRC(↑)	min(+P, Se)(↑)
MIMIC-III	GCN (mean)	82.31±1.02	45.88±2.57	44.74±2.16	76.50±1.08	45.23±2.26	44.63±1.88
	GCN (max)	83.61±1.05	52.02±2.59	49.16±2.12	77.61±1.09	49.84±2.17	**47.41±1.87**
	GAT (mean)	83.04±1.06	49.65±2.63	48.26±2.12	76.87±1.12	46.37±2.23	45.83±1.90
	GAT (max)	**85.92±0.94**	**55.49±2.53**	**51.52±2.12**	**77.75±1.11**	**49.92±2.16**	47.25±1.88
	RGCN (mean)	83.09±1.03	47.69±2.60	48.01±2.05	76.39±1.10	46.52±2.22	45.04±1.84
	RGCN (max)	83.14±1.04	49.52±2.44	46.48±2.15	76.44±1.12	48.08±2.14	46.43±1.88
MIMIC-IV	GCN (mean)	80.26±0.93	51.13±2.12	49.42±1.78	74.29±1.07	51.71±1.94	48.81±1.67
	GCN (max)	82.65±0.86	57.30±1.91	52.35±1.71	76.07±1.00	54.83±1.83	51.38±1.57
	GAT (mean)	80.71±0.93	51.67±2.12	48.49±1.84	74.04±1.05	49.27±1.95	47.76±1.68
	GAT (max)	**83.24±0.89**	**58.83±1.95**	**53.00±1.70**	**77.31±0.99**	**56.88±1.87**	**53.73±1.53**
	RGCN (mean)	80.06±0.91	49.29±2.12	46.60±1.79	72.27±1.03	45.86±1.93	45.04±1.73
	RGCN (max)	81.70±0.95	55.73±2.03	50.72±1.80	76.36±1.02	54.23±1.91	51.39±1.64

GAT with max pooling consistently performs best across both datasets and tasks, indicating that attention-based neighborhood aggregation is effective for highlighting clinically informative entities and relations. Across backbones, max pooling generally outperforms mean pooling, suggesting that a small set of high-signal nodes often drives the prediction.

4 Conclusion

This work proposes HERMES, a novel framework that constructs personalized KGs from clinical text through LLM-guided extraction with Contrastive Logic Modeling, which captures temporal dynamics, such as treatment contradictions, clinical outcomes, and patient changes-that predict deterioration. Experiments on MIMIC-III and MIMIC-IV demonstrate significant gains over text-only baselines for mortality and readmission prediction, validating explicit relational modeling for clinical outcome prediction. Future work may integrate multimodal data and refine the extraction process to strengthen predictive performance.

References

1. Agrawal, M., Hegselmann, S., et al.: Large language models are few-shot clinical information extractors. In: Proc. Conf. Empirical Methods Nat. Lang. Process. pp. 1998–2022 (2022)
2. Choi, E., Bahadori, M.T., et al.: Gram: graph-based attention model for healthcare representation learning. In: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. pp. 787–795 (2017)
3. Gu, Y., Tinn, R., et al.: Domain-specific language model pretraining for biomedical natural language processing. ACM Trans. Comput. Healthcare **3**(1) (2021)

4. Hashir, M., Sawhney, R.: Towards unstructured mortality prediction with free-text clinical notes. Journal of Biomedical Informatics **108**, 103489 (2020)
5. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv arXiv:1606.08415 (2016)
6. Jiang, P., Xiao, C., et al.: Graphcare: Enhancing healthcare predictions with personalized knowledge graphs. In: Proc. Int. Conf. Learn. Represent. (2024)
7. Johnson, A.E., Bulgarelli, L., et al.: Mimic-iv, a freely accessible electronic health record dataset. Scientific data **10**(1), 1 (2023)
8. Johnson, A.E., Pollard, T.J., et al.: Mimic-iii, a freely accessible critical care database. Scientific data **3**(1), 1–9 (2016)
9. Li, Y., Wehbe, R.M., et al.: A comparative study of pretrained language models for long clinical text. Journal of the American Medical Informatics Association **30**(2), 340–347 (2023)
10. Liu, F., Shareghi, E., et al.: Self-alignment pretraining for biomedical entity representations. In: Proc. Conf. North Amer. Chapter Assoc. Comput. Linguist. pp. 4228–4238 (2021)
11. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Proc. Int. Conf. Learn. Represent. (2019)
12. Mahbub, M., Srinivasan, S., et al.: Unstructured clinical notes within the 24 hours since admission predict short, mid & long-term mortality in adult icu patients. PLOS ONE **17**, e0262182 (2022)
13. Tian, K., Mitchell, E., et al.: Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In: Proc. Conf. Empirical Methods Nat. Lang. Process. (2023)
14. Veličković, P., Cucurull, G., et al.: Graph attention networks. In: Proc. Int. Conf. Learn. Represent. (2018)
15. Wang, G., Liu, X., et al.: A generalist medical language model for disease diagnosis assistance. Nature Medicine (2025)
16. Wei, J., Wang, X., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
17. Zhu, Y., Ren, C., et al.: Emerge: Enhancing multimodal electronic health records predictive modeling with retrieval-augmented generation. In: Proc. ACM Int. Conf. Inf. Knowl. Manag. pp. 3549–3559 (2024)