JudgEHR: LLM-Guided Cohort Inference for Clinical Decision Confidence Estimation

Kexuan Xin*, Guillaume Pelat†, Jonathan Vitale*, Shun Takeuchi†, Shlomo Berkovsky*, Amin Beheshti*‡
*Macquarie University, Sydney, Australia
Emails: xkxjane@gmail.com, {jonathan.vitale, shlomo.berkovsky, amin.beheshti}@mq.edu.au
†Fujitsu, Japan
Emails: pelat.guillaume@fujitsu.com, shun.takeuchi@fujitsu.com

***Abstract*—Clinical decisions, such as diagnosing conditions, prescribing medications, and recommending procedures, are rarely made with absolute certainty. Instead, they reflect probabilistic judgments shaped by evolving patient information and incomplete evidence. However, current EHR systems and knowledge graphs encode such decisions as deterministic triples, lacking a mechanism to represent the subjective confidence inherent in clinical reasoning. We present JudgEHR, a framework for clinical decision confidence estimation that leverages large language models (LLMs) to perform cohort-based collective inference over structured patient records by representing clinical events as knowledge graph triples and integrating them into LLM prompts. JudgEHR groups related clinical concepts into cohorts using LLM-driven relational inference, and then jointly evaluates the plausibility of all clinical decision triples within each cohort by considering patient visit history and background medical knowledge. We apply our method to the MIMIC-III dataset. Our statistical analysis shows that JudgEHR generates semantically consistent confidence scores, with similar concepts receiving closer values, whereas dissimilar replacements yield large confidence differences. Experiments on the MIMIC-III dataset show that incorporating the confidence into a zero-shot LLM-based pipeline improves relative AUROC by 14.6% and AUPRC by 21.8% on the mortality prediction task.**

***Index Terms*—Confidence Estimation, Large Language Models, Personalized Knowledge Graph, Electronic Health Records, Clinical Decision Support, Uncertainty Modeling**

I. Introduction

Clinical knowledge graphs (KGs) have become increasingly important for representing structured information extracted from electronic health records (EHRs), supporting downstream applications such as diagnosis prediction, medication recommendation, outcome forecasting and misinformation detection [1]–[4]. However, these graphs typically encode medical facts such as diagnoses, medications, or procedures as deterministic triples, without capturing the inherent uncertainty that characterizes real-world clinical decision-making.

In practice, clinicians rarely make binary, definitive decisions. Instead, medical reasoning involves probabilistic judgments based on incomplete and evolving patient data [5], [6]. For instance, a patient may show some but not all indicators for a condition, or be prescribed a treatment as a precaution rather than a certainty. Such situations call for an epistemic representation of confidence: not merely reflecting data noise or distributional uncertainty, but capturing how likely a clinical fact holds true, given the available evidence and patient-specific context.

Existing approaches to uncertainty modeling in knowledge graphs, such as probabilistic embeddings [7], [8] or truth discovery frameworks [9], primarily focus on generic data reliability or symbolic conflicts. However, these techniques still face challenges when applied to clinical settings. Most require access to large amounts of ground-truth labels or probabilistic supervision, which are rarely available in EHR data [1], [10]. For example, many clinical events in EHRs are only weakly annotated or inferred from billing codes, and gold-standard outcome labels (such as true diagnosis or disease progression) are typically unavailable outside of curated research cohorts. Furthermore, modeling the temporal relationships and evolving context across patient visits remains difficult for traditional approaches [11]. Our method, in contrast, enables unsupervised, context-aware confidence estimation by leveraging LLM-driven relational reasoning over current and historical visits, without relying on labeled data.

Designing confidence estimation mechanisms for clinical knowledge is inherently challenging [1], [12], [13]. Clinical data is highly heterogeneous, comprising diagnoses, medications, and procedures, each with distinct semantics, temporal dynamics, and measurement reliability [3], [10]. Furthermore, observations within the same visit are often interdependent, requiring joint reasoning over related triples. For example, the presence of a symptom may support or contradict a diagnosis [5], [14]. Supervised confidence estimation is further complicated by the scarcity of medical ground truth and the prevalence of weak or noisy labels in real-world EHRs [1], [15]. Finally, even if large language models (LLMs) offer powerful reasoning capabilities, their numeric estimates can be unstable and highly sensitive to prompt design, posing significant calibration and reliability challenges in safety-critical clinical applications [16]–[21].

To address these challenges, we propose **JudgEHR**, a dynamic confidence estimation framework for personalized clinical knowledge graphs. It provides fine-grained, interpretable confidence scores for clinical triples, integrating both patient-specific medical history and external medical knowledge.

‡Corresponding author.

979-8-3195-1202-4/26/$31.00 ©2026 IEEE
DOI 10.1109/ICDH72779.2026.00046

To handle heterogeneous and interdependent observations, we cluster semantically related clinical concepts and employ LLM-based collective inference for joint confidence scoring. To mitigate LLM calibration issues and reduce hallucination, we incorporate explicit confidence criteria in prompt design, leverage in-context examples as few-shot guidance, and aggregate cohort-level evidence to encourage consistent, context-aware scoring. JudgEHR operates fully unsupervised, enabling reliable, personalized confidence estimation without requiring expert-labeled data or gold-standard outcomes.

Our main contributions are as follows:

- We propose a novel epistemic-style structured confidence estimation task for structured clinical triples within patient-specific visits, reflecting context-sensitive diagnostic certainty.
- We propose a cohort-based collective inference framework that leverages LLM capability to jointly estimate confidence values, using relational prompts and anchor-based grouping strategies.
- We design an evaluation methodology for confidence estimation in the absence of labels, based on semantic perturbation and contextual plausibility, enabling robust comparison without ground truth.
- We demonstrate that incorporating confidence into clinical downstream tasks, such as 30-day mortality prediction on MIMIC-III, consistently improves performance over not using confidence, highlighting its practical utility.

II. Related Work

Knowledge Graph Uncertainty and Embedding-Based Methods. Uncertainty modeling in knowledge graphs (KGs) has attracted substantial attention, primarily through embedding-based methods such as TransE [22], RotatE [23], ComplEx [24], and holographic embeddings [25]. These approaches quantify triple plausibility through embedding distances, effectively capturing semantic relationships and uncertainty in structured data [7], [26], [27]. Probabilistic approaches like UKGE [28], ProPPR [29], and BEUrRE [8] further extend these concepts by explicitly modeling uncertainty via probabilistic representations. More recently, Zhu et al. [30] propose reasoning over uncertain knowledge graphs with statistical guarantees, advancing the formal treatment of KG uncertainty. Nevertheless, such methods primarily focus on static embeddings or general KG tasks, lacking adaptability to personalized or dynamically evolving contexts required in clinical decision-making scenarios.

LLMs for Confidence and Numeric Estimation. Recent advancements in large language models (LLMs) have facilitated numeric confidence estimation in various NLP tasks such as question answering and text generation [16], [18], [31]–[34]. Techniques like Chain-of-Thought prompting [35], [36], self-consistency aggregation [37], [38], and post-hoc calibration methods [39], [40] significantly improve confidence calibration and alignment with human intuition. While effective in open-ended textual or isolated inference contexts, these methods are not explicitly designed for structured relational scenarios or clinical specificity, limiting their utility in structured EHR data and clinical decision support systems. Meanwhile, LLMs are being increasingly adopted for broader digital health applications such as conversational patient engagement [41], further motivating the need for reliable confidence estimation in clinical contexts.

Clinical and Human-Centric Confidence Modeling. In clinical settings, uncertainty has been studied through linguistic analyses of physician notes, heuristic methods, and expert annotations [12], [42]–[44]. Historically, expert systems like MYCIN employed certainty factors based on manual annotations [45], and recent machine learning frameworks, such as InfEHR [46], mainly focus on predictive accuracy rather than explicitly quantifying clinician confidence. Beyond healthcare, modeling human decision confidence has been explored in educational settings [47], crowdsourcing [48], [49], and financial decision-making [50], typically relying on explicit confidence labels and extensive user interaction. While these approaches do not generalize easily to real-time clinical environments lacking explicit annotations or interaction labels.

III. Method and Evaluation Protocol

A. Task Definition

We define a novel task: **structured confidence estimation** over clinical decision triples. Given a patient's current visit v_t, we extract a set of structured assertions reflecting medical decisions made during the visit. Each assertion is represented as a triple (h, r, t):

- The head h is always the patient entity (e.g., *Patient_001*);
- The relation r denotes one of three clinically relevant types: *has_diagnosis*, *has_medication*, or *underwent_procedure*;
- The tail concept t is a UMLS-mapped concept of the corresponding type (e.g., *Flu*, *Amlodipine*, or *Colonoscopy*).

We refer to such $(\textit{Patient}, r, t)$ tuples as **clinical decision triples**. These represent structured proxies for clinician judgments made during a patient visit.

Our goal is to estimate a confidence score $s_i \in [0, 1]$ for each triple in the visit set $\mathcal{T}_t = \{(h_i, r_i, t_i)\}$, representing how confident a clinician might be in the truthfulness or contextual plausibility of the assertion, given the patient's condition at time t.

The confidence estimation is contextual and personalized: it should take into account visit history, co-occurring conditions, external medical knowledge, and concept correlations within the same visit. Formally, JudgEHR learns a mapping:

$$\mathcal{F} : (v_t, (h_i, r_i, t_i), \text{Context}) \rightarrow s_i \in [0, 1]$$

where *Context* includes previous visits of the same patient, concept-level background knowledge, and correlation-based interactions among triples within v_t.

The output is a **confidence-weighted visit graph**, where each clinical decision triple is annotated with a numeric score indicating its contextual reliability.

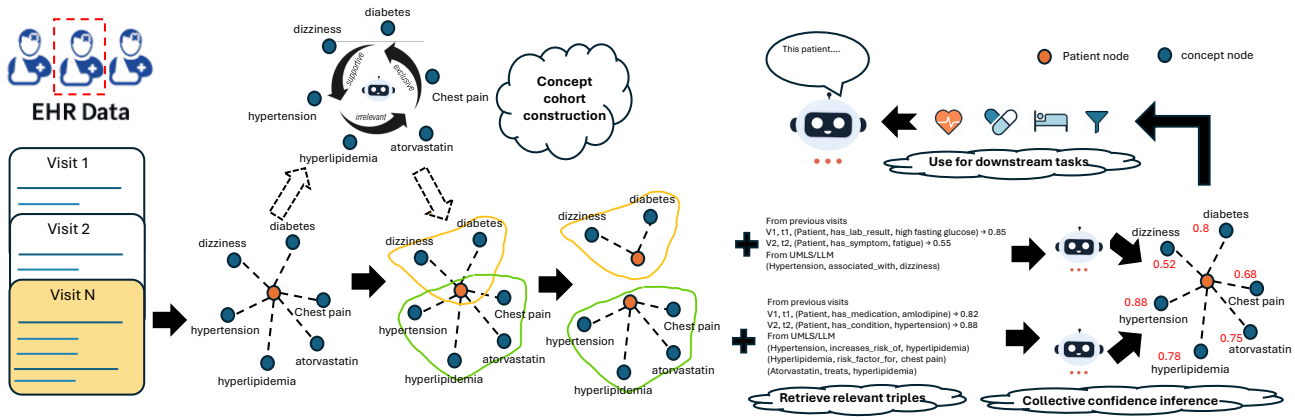


Fig. 1: Example of JudgEHR applied to MIMIC data. Given visit-level triples and patient history, the system constructs a concept graph, performs LLM-based reasoning, and produces confidence-weighted outputs.

B. Framework Overview

JudgEHR is a framework for estimating confidence over structured clinical observations within individual patient visits. Given a patient's current visit, the system first extracts a set of clinically relevant triples. These triples form a structured view of the visit, serving as the basis for contextual inference. The overall framework consists of three main stages:

- **Triple Extraction and Context Expansion:** Following [3], we begin by extracting structured triples from EHR records (e.g., diagnosis, medication), and augmenting them with relevant background knowledge from external sources such as UMLS and LLM-based concept retrieval.
- **Concept Cohort Construction:** Within each visit, we group the clinical decision triples into semantic cohorts based on the tails of them (i.e., the medical concepts) from pairwise relations inferred via LLM-based inference. Each cohort contains concepts that are supportive or exclusive with respect to at least one member, forming a semantically connected unit for collective inference.
- **LLM-based Confidence Scoring:** For each cohort, a context-aware and carefully crafted prompt is used to generate a numeric confidence score $s_i \in [0, 1]$, capturing the plausibility of each clinical decision triple in the specific cohort. Each prompt includes relevant triples of the concepts within the cohort, and LLM infers the confidence of the whole cohort of triples collectively. Next, the weighted visit graph can be used for different downstream tasks.

JudgEHR integrates symbolic structure, concept-level correlation, and language model inference into a unified pipeline, enabling the modeling of subjective decision certainty at scale across EHR records. Figure 1 illustrates an example.

C. Collective Inference over Visit Concept Cohorts

Building on the extracted decision triples and expanded background knowledge following [3], we perform cohort-based collective inference to capture semantic dependencies among visit concepts. Clinical concepts within a single patient visit often exhibit complex semantic dependencies. For example, the condition *Hypertension* may be supported by the prescription of *Amlodipine* (supportive), while a procedure like *Colonoscopy* may be clinically incompatible with a diagnosis of *Severe Coagulopathy* (exclusive). In contrast, concepts such as *Depression* and *Appendectomy* are generally unrelated (irrelevant) and should not influence each other's confidence. Estimating the confidence of each observation independently ignores such structure. To address this, we propose a collective inference strategy that groups related concepts together and estimates confidence scores in context.

1) Concept Interaction Cohorts construction with LLMs: Given a set of clinical concepts extracted from a visit, we first construct a visit-level *concept interaction cohort*. Each concept serves as a node, and we query an LLM to infer pairwise relations. For a pair (c_i, c_j), the relation is classified as: **Supportive** (c_j increases the plausibility/confidence of c_i), **Exclusive** (c_j conflicts with or weakens the plausibility of c_i), or **Irrelevant** (c_j is unrelated and should not influence the interpretation of c_i).

Accurately identifying such relations requires world knowledge and clinical reasoning. Recent work has shown that LLMs are effective at knowledge graph reasoning tasks, including relation classification, link prediction, and fact verification [51]–[54]. Inspired by these successes, we generate a structured prompt asking the LLM to classify the relation type based on general medical knowledge and reasoning. The core content of our relation inference prompt is shown in box III-C1. The LLM's classification result is used to determine whether a pair should be grouped together in the subsequent confidence estimation step. Only supportive or exclusive pairs are considered semantically relevant; irrelevant pairs are ignored to reduce noise and control group size.

Box III-C1: Relation Inference Prompt

Given the clinical concept pair:
A: Hypertension
B: Amlodipine
Based on medical knowledge and typical usage, what is the most likely relationship between A and B?
Possible options are:
- Supportive
- Exclusive
- Irrelevant
Please answer with the most appropriate label and a brief justification.

2) Greedy Cohort Assignment via LLM Inference: We group concepts into *inference-based cohorts* via a greedy, anchor-based strategy. Each unassigned concept is chosen as an anchor and compared with all remaining concepts. Those classified as *Supportive* or *Exclusive* with respect to the anchor are added to the same cohort, and assigned concepts are not revisited as future anchors.

Cohorts are not required to be fully connected: concepts may be grouped even if they are mutually *Irrelevant*. For instance, if c_1 relates to both c_2 and c_3 while c_2 and c_3 are irrelevant to each other, all three remain in one cohort. This implements an anchor-based closure that expands a cohort by anchor relations rather than enforcing pairwise connectivity.

The procedure is summarized in Algorithm 1.

Algorithm 1 Greedy Cohort Assignment via LLM-Based Pairwise Inference

Require: Concept set $\mathcal{C} = \{c_1, c_2, ..., c_n\}$
Initialize empty set of cohorts $\mathcal{K} \leftarrow \emptyset$
Initialize relation log $\mathcal{R} \leftarrow \emptyset$
Mark all concepts as unassigned
for each unassigned concept $c_i \in \mathcal{C}$ **do**
  Initialize new cohort $\kappa \leftarrow \{c_i\}$, mark c_i as assigned
  for each unassigned concept $c_j \in \mathcal{C} \setminus \kappa$ **do**
    Query LLM for relation type between c_i and c_j
    if relation $\in$ {**supportive**, **exclusive**} **then**
      Add c_j to κ, mark c_j as assigned
      Record $(c_i, c_j, \text{relation})$ in $\mathcal{R}$
    end if
  end for
  Add κ to cohort set $\mathcal{K}$
end for
return $\mathcal{K}$ as grouped concept cohorts, and $\mathcal{R}$ as inferred relations

This design offers several advantages. It reduces the number of LLM queries, and yields interpretable, semantically meaningful concept cohorts that support confidence estimation.

3) Usage for Confidence Inference: Each resulting concept cohort κ is processed jointly in the subsequent confidence estimation step. By presenting the full cohort as context to the LLM, we ensure that the confidence assigned to each triple accounts for relevant supporting or conflicting evidence within the same visit. This enables semantic consistency and mitigates hallucinated confidence in isolation.

Our cohort construction strategy is designed with practicality and robustness in mind. First, any concept related to a cohort member via a supportive or exclusive relation is included, ensuring that no clinically relevant concept is overlooked. Second, the computational complexity is well-controlled: since each concept is queried only once as an anchor. Finally, by ignoring irrelevant concept pairs, the method is tolerant to noisy or uncertain relation predictions.

D. Context-Aware Confidence Scoring with LLMs

After forming concept cohorts κ_i, our goal is to estimate the contextual confidence of each clinical triple (h, r, t) associated with that visit. While cohorts are constructed based on semantic relations between concepts (e.g., *Flu*, *Fever*, *Sunstroke*), confidence estimation is performed at the triple level (e.g., *(Patient, has_diagnosis, Flu)*).

To model how confident a clinician might be in each assertion, we design a structured prompt that presents the entire cohort context to a large language model (LLM). For each triple within the cohort, we query the LLM to assign a numeric confidence score in $[0, 1]$, along with a short explanation. The scoring is guided by not only the well-crafted confidence scoring standard and several curated clinical examples for few-shot prompting, but also the evidence from:

- **Background knowledge:** UMLS and LLM-generated triples relevant to the concepts of this cohort;
- **Prior visits:** Historical triples involving relevant concepts, with existing confidence scores.

We emphasize that our prompt makes a clear distinction between *concepts* (e.g., "Fever") and *triples* (e.g., "Patient has_diagnosis Fever"), since each concept may participate in multiple assertions. This is especially important in multi-triple cohorts where a single medication or condition may appear in different clinical roles.

Although concepts in the same cohort are grouped via supportive or exclusive relations, we do not explicitly encode pairwise relations in the prompt. Instead, we rely on the LLM to reason holistically over the provided evidence and generate a contextualized confidence score for each target triple.

E. Evaluation without Supervision

Since ground-truth confidence labels are unavailable, we evaluate confidence consistency via *semantic replacement*. For each triple $(\textit{Patient}, r, t)$ with confidence s_{orig}, we replace t with an alternative concept t' from the same category (diagnosis/procedure/medication), covering both semantically near and far substitutions (e.g., *Flu* $\rightarrow$ *Cold* vs. *Diabetes*). We then score the replacement triple to obtain s_{rep} and compute the margin $\Delta_s = |s_{\text{orig}} - s_{\text{rep}}|$. A robust method is expected to show a *population-level* trend: clinically compatible, near replacements yield smaller margins on average, while distant or incompatible replacements yield larger margins.

TABLE I: Statistics of pre-processed EHR dataset. "#": "the number of", "/patient": "per patient".

	#patients	#visits	#visits/ patient	#conditions/ patient	#procedures/ patient	#drugs/ patient
MIMIC-III	35,707	44,399	1.24	12.89	4.54	33.71

IV. Experiments

A. Experimental Setup

Dataset and preprocessing. We conduct experiments on the MIMIC-III dataset [10] (statistics shown in table I). For each visit, we extract structured clinical triples from diagnoses, procedures, and medications, and represent each triple as $(\textit{Patient}, r, t)$, where $r \in \{\textit{has_diagnosis}, \textit{underwent_procedure}, \textit{has_medication}\}$ and t is a normalized clinical concept (e.g., UMLS-based). We retain visits containing at least two decision triples to support cohort construction and confidence scoring, and order visits by time using the MIMIC-III admissions metadata.

Background knowledge and inputs. Background medical knowledge is obtained from UMLS and LLM-generated clinical relations following prior work [3]. To comply with the MIMIC-III data use agreement, we only use de-identified structured triples as model inputs; no free-text notes or personally identifiable information are sent to external APIs/LLMs.

Implementation. We use GPT-4.1-mini as the base LLM for both relation inference and confidence scoring. We precompute background knowledge triples and pairwise concept relations, and cache intermediate results to reduce inference cost. We will release prompt templates, preprocessing/evaluation scripts, and experimental configurations to support reproducibility.

B. Unsupervised Evaluation via Semantic Replacement

Following the protocol in Section III-E, we further instantiate the replacement sets into four similarity buckets and compare methods. Each triple (h, r, t) is paired with several (h, r, t') replacements, where the new t' belongs to the same concept type (diagnostic, drug, etc) but varies in semantic similarity to t. For each pair, we calculate the confidence margin Δ_s to measure how JudgEHR distinguishes semantically consistent from inconsistent triples. Replacement concepts are divided into four uniform similarity buckets using the text-embedding-ada-002 model. For each bucket, mean confidence margins are computed and compared across methods. We analyze whether margins increase as semantic similarity decreases.

We benchmark JudgEHR against two baselines: (1) **Independent inference**, which assigns confidence to each triple separately by skipping the grouping step of our method; and (2) **Embedding-cluster grouping** replaces cohort reasoning with agglomerative clustering over concept embeddings. Each concept is encoded by *text-embedding-ada-002*, and average-linkage clustering with cosine distance groups concepts whose pairwise similarity exceeds 0.8. Each cluster serves as a cohort for collective confidence scoring.

TABLE II: Statistical Analysis of Confidence Generation Methods

Parameter	**Estimate (SE)**
Model: confidence_margin ∼ similarity × method + (1—concept)	
Baseline slope (JudgEHR)	-0.071 (0.006)
Final slopes:	
JudgEHR	$\mathbf{-0.071}$
cluster grouping	-0.059
independent	-0.038
Overall interaction test: $p < 0.001$ (methods differ significantly)	

1) Correlation Analysis: To rigorously quantify the trend significance, we fit a linear mixed-effects model (LMEM) with interaction terms: *confidence_margin* $\sim$ *similarity* $\times$ *method* + *(1—concept)*, where the random intercept *(1—concept)* accounts for repeated replacements of the same concept.

As shown in Table II, JudgEHR achieves the most pronounced negative slope ($\beta = -0.071$), indicating that its confidence margins decrease consistently as semantic similarity increases. The cluster grouping baseline exhibits a weaker trend ($\beta = -0.059$), while independent inference performs worst ($\beta = -0.038$). The overall interaction test ($p < 0.001$) confirms that these differences between methods are statistically significant.

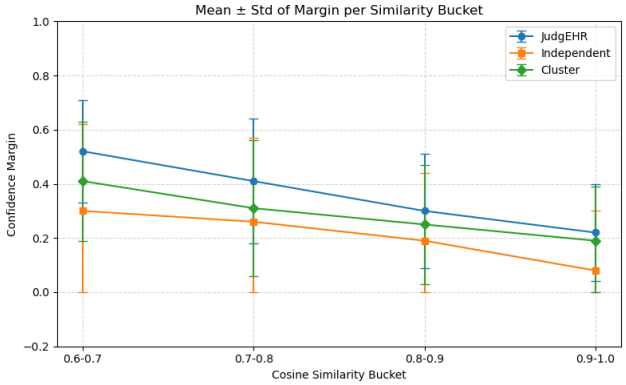


Fig. 2: Comparison of confidence margin across similarity buckets.

2) Confidence Margin Analysis.: Figure 2 visualizes the relationship between cosine similarity and confidence margins across buckets for all methods. It plots cosine-similarity buckets on the x-axis and the mean confidence margin per bucket on the y-axis; error bars show mean $\pm$ 1 std.

Consistent with our LMM analysis, all approaches demonstrate the expected negative trend: as semantic similarity

increases, confidence margins decrease. However, **JudgEHR** exhibits the steepest and most consistent decline, dropping from 0.52 to 0.22. The **clustering-based method** shows intermediate performance with a smaller standard variance, while **independent inference** displays the flattest trend and the highest standard deviations.

Importantly, JudgEHR maintains the most consistent performance with relatively stable standard deviations (0.18-0.23), while independent inference shows high variability (0.22-0.32) across all similarity buckets. The Independent method's high variance creates substantial uncertainty in confidence estimation, with standard deviations often exceeding the mean values. The pronounced slope in JudgEHR's trend line reflects its superior ability to discriminate between different levels of semantic similarity, directly supporting our hypothesis that cohort-based collective inference enhances contextual sensitivity in clinical confidence assessment.

3) Case Studies: We report three representative cases, each shown in a **pre-replacement** (target-centered) and **post-replacement** (replacement-centered) view. For readability, we display only selected tail concepts from the current and previous visits (previous-visit entries include confidence), omitting patient nodes and relation names.

Target: OTHER ANTIBACTERIALS in ATC (0.85) **Replacement:** POSTERIOR PITUITARY HORMONES (0.10)	
Pre-replacement	**Clinical Context**
Current visit	Pneumonia; Routine Chest X-ray
Previous visits	Chronic obstructive pulmonary disease and bronchiectasis (0.68); BETA-LACTAM ANTIBACTERIALS, PENICILLINS (0.82); Routine chest X-ray (0.61)
Post-replacement	**Clinical Context**
Current visit	—
Previous visits	—

TABLE III: Distinct clinical concepts: target supported by respiratory context; replacement lacks contextual support.

Distinct concepts (Table III). The target matches the infectious–respiratory context and receives high confidence, whereas the semantically distant replacement lacks supporting context in both current and prior visits, highlighting JudgEHR's preference for contextual coherence.

Both plausible (Table IV). Both concepts are clinically reasonable, but the target better matches the documented nutritional deficiency/anemia context, yielding a small margin (0.10) that reflects subtle contextual preference.

Near-synonyms (Table V). Target and replacement share nearly identical neurological context; the replacement is scored higher (0.73 vs. 0.65), suggesting that more specific hierarchical terms (e.g., TIA) can be preferred over broader categories.

C. Impact on Downstream Prediction Tasks

We conduct early-stage experiments to assess whether confidence scores can be successfully used to improve LLM predictions on clinical tasks, as compared to baseline approaches

Target: OTHER MINERAL SUPPLEMENTS in ATC (0.75) **Replacement:** INFLAMMATION AND INFECTION DETECTION (0.65)	
Pre-replacement	**Clinical Context**
Current visit	Nutritional deficiencies; Blood transfusion
Previous visits	Deficiency and other anemia (0.70); OTHER MINERAL SUPPLEMENTS in ATC (0.72)
Post-replacement	**Clinical Context**
Current visit	Nuclear medicine imaging (pulmonary)
Previous visits	Acute cerebrovascular disease (0.40); Magnetic resonance imaging (0.62)

TABLE IV: Related clinical concepts: small margin reflects partial plausibility for both, target remains more context-consistent.

Target: Acute cerebrovascular disease (0.65) **Replacement:** Transient cerebral ischemia (0.73)	
Pre-replacement	**Clinical Context**
Current visit	Magnetic resonance imaging; ANTITHROMBOTIC AGENTS
Previous visits	Headache; including migraine (0.59); Computerized axial tomography (CT) scan head (0.66)
Post-replacement	**Clinical Context**
Current visit	Magnetic resonance imaging; Platelet aggregation inhibitors excl. heparin
Previous visits	Headache; including migraine (0.59); Computerized axial tomography (CT) scan head (0.66)

TABLE V: Near-synonymous concepts: lower target confidence leads to an unstable ranking and a flip.

not feeding confidence scores to the LLM. In this study, we focus on the 30-day mortality prediction on MIMIC-III. Given a patient's visit history $(v_1, ..., v_n)$, the goal is to predict the survival status $y(v_n) \in \{0, 1\}$ using information from prior visits $(v_1, ..., v_{n-1})$. We deploy JudgEHR in a zero-shot pipeline, where past visit triples are fed into an LLM to predict the survival status. We compare three input variants: (1) using all triples without filtering (*no filtering* rows in Table VI), (2) including only triples with confidence above a threshold τ (*Thre-0.8*), and (3) selecting the top-k highest-confidence triples (*Top-10*). In all cases, we further test whether explicitly inserting the confidence value into the LLM prompt (*Prompt Conf*) improves reasoning. The independent inference variant is also assessed to examine the benefits of collective grouping. Table VI reports the results on 612 patients (10% of data) averaged over three runs, using $\tau = 0.8$ and $k = 10$ based on offline tuning on a separate set of 100 patients.

1) Results: Across all settings, incorporating JudgEHR's confidence scores yields consistent improvements over the LLM baseline. Collective inference achieves the best performance when both confidence-based filtering and explicit confidence prompting are applied, reaching 13.4 AUPRC and 57.4 AUROC, which is an absolute gain of 7.3 AUROC points over the LLM-Base. Using all triples without filtering already improves stability, while confidence filtering (*Thre-0.8*) and selection (*Top-10*) further enhance precision by reducing noise

from low-confidence relations. Compared with independent inference, collective grouping consistently performs better, confirming the value of correlation context in confidence estimation.

TABLE VI: Zero-shot LLM methods for mortality prediction

Method	Conf. in Prompt	AUPRC	AUROC
LLM-Base	No	11.0	50.1
JudgEHR (collective)			
J-EHR	Yes	11.1	51.2
Thre-0.8	No	12.2	55.2
Thre-0.8	Yes	**13.4**	**57.4**
Top-10	No	11.3	52.3
Top-10	Yes	12.7	56.6
JudgEHR (independent)			
J-EHR	Yes	11.3	52.2
Thre-0.8	No	12.3	56.3
Thre-0.8	Yes	12.5	56.5
Top-10	No	11.1	51.1
Top-10	Yes	12.7	56.6

All methods are zero-shot without any training.

2) Case study: We show an example of system outputs for one patient in Table VII. In this example, many intensive procedures were performed on a patient, which led the baseline model to predict a high mortality risk. Adding the confidence values, some procedures show a lower confidence level. This does not indicate that performing them was inappropriate or clinically unjustified. Rather, we interpret this as suggesting that while these procedures remained valid within the patient's treatment plan, they were deemed by JudgEHR not the primary drivers influencing the current patient's clinical outcome. As a result, the severity of the case is evaluated to be much lower and we accurately predict the survival of the patient.

V. Discussion

JudgEHR provides a promising framework for context-sensitive confidence estimation in clinical knowledge graphs. By reasoning collectively over concept sets, the method achieves coherent and interpretable confidence level assessment. Our results demonstrate that collective inference enhances contextual consistency compared with independent or embedding-based alternatives, and it shows a positive effect on downstream tasks.

Several limitations should be noted. First, dependence on LLMs may introduce biases or calibration instability, especially for rare or ambiguous medical concepts. Second, cohorts formation through relation inference may overlook higher-order interactions among multiple entities. Third, although JudgEHR improves consistency and downstream task accuracy, its calibration quality and transportability to other datasets need to be examined in more depth. Overall, JudgEHR highlights the potential of LLM-guided collective reasoning for improving the reliability and interpretability of clinical AI, while emphasizing the need for extensive real-world evaluation.

Target: Patient with one visit **Label:** 0	
Procedures	cholecystectomy and common duct exploration (0.9), gastrostomy; temporary and permanent (0.8), respiratory intubation and mechanical ventilation (0.6), diagnostic bronchoscopy and biopsy of bronchus (0.9)
Base-LLM	The patient has multiple severe diagnoses and complications of surgical procedures. They underwent major surgical procedures. The complexity and severity of these conditions, combined with respiratory failure and the need for extensive medication indicate a high risk of mortality within 30 days despite discharge.
Prediction	1
JudgEHR-Thre-0.8	The patient has multiple serious diagnoses, all with high confidence levels. The patient underwent several major procedures, indicating significant invasive interventions. Medications reflect complex and intensive care. However, despite these severe conditions and interventions, the patient was deemed stable enough for discharge, which suggests a relatively controlled clinical situation. Therefore, the short-term mortality risk within 30 days is elevated but not definitive.
Prediction	0

TABLE VII: LLM output for one patient comparing LLM-Base and JudgEHR-Thre-0.8. We show a portion of the procedures performed. In particular, the removal of "respiratory intubation and mechanical ventilation" led to a switch to the correct predicted label.

This study harnesses de-identified MIMIC-III data under the PhysioNet agreement. While improving the performance of medical AI, we acknowledge ethical risks: LLM bias propagation to confidence estimates, demanding rigorous validation; necessity of human supervision for clinical interpretation; and potential misuse in sensitive contexts like insurance claims. Our use of GPT-4.1-mini involved a private Azure deployment with no prompt/completion storage or human review, prioritizing privacy of the patient data.

VI. Conclusion

We introduced JudgEHR, a framework for joint confidence estimation over clinical KGs using LLM-guided reasoning. Our method enables interpretable, context-sensitive confidence scoring for clinical decision triples, and shows promising performance on semantic replacement evaluation and zero-shot downstream tasks. Results highlight the value of collective inference and confidence-aware selection in improving clinical prediction. In future work, we plan to explore more sophisticated strategies for leveraging confidence scores in diverse downstream tasks, investigate their impact across broader clinical applications, and incorporate relationships across patients rather than focusing on each patient individually.

VII. Ethical Considerations

This study utilizes the publicly available and de-identified MIMIC-III dataset, accessed in compliance with the PhysioNet

credentialing process and data use agreement. All patient data have been stripped of direct identifiers in accordance with HIPAA standards, and no attempt was made to re-identify individuals. Our work focuses on constructing personalized clinical knowledge graphs and estimating confidence scores for clinical concepts using LLMs. While this aims to improve interpretability and reasoning in medical AI, we acknowledge several potential ethical concerns. First, LLMs may introduce biases learned from pretraining data, which could propagate to confidence estimation; careful validation and fairness assessments are necessary before any real-world use. Second, any clinical interpretation of model outputs should be conducted under human supervision to prevent over-reliance on automated reasoning. Third, misuse of this technology—for example, generating misleading clinical statements with inflated confidence—could have harmful consequences, particularly if used in treatment-adjacent fields such as solving insurance claims.

Additionally, privacy issues must be considered in order to select the LLM powering JudgEHR. In this work, we use a private deployment of GPT-4.1-mini on Microsoft Azure, and opted out of both prompt/completion storage and human reviews by Azure/OpenAI staff. On-site LLM deployment should be considered if feasible; however, we acknowledge that many medical professionals or institutions do not have the infrastructure or capabilities to do so, and we believe our setup to be an adequate compromise between privacy protection and ease of use.

ACKNOWLEDGMENTS

AI-Generated Content Disclosure: In accordance with IEEE policy, we disclose the following use of AI systems in this work. GPT-4.1-mini (via private Azure deployment) was used as the core inference engine within the JudgEHR framework for relation classification and confidence scoring, as described in Sections III-C and III-D. Additionally, AI-based writing assistants were used for grammar and language editing of the manuscript.

REFERENCES

[1] R. Miotto, F. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records," *Scientific reports*, vol. 6, p. 26094, 2016.

[2] E. Choi, M. T. Bahadori, and J. Sun, "Gram: Graph-based attention model for healthcare representation learning," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 787–795.

[3] P. Jiang, C. Xiao, A. Cross, and J. Sun, "Graphcare: Enhancing healthcare predictions with personalized knowledge graphs," *arXiv preprint arXiv:2305.12788*, 2023.

[4] O. Clark and K. P. Joshi, "Real-time detection of online health misinformation using an integrated knowledgegraph-llm approach," in *2025 IEEE International Conference on Digital Health (ICDH)*. IEEE, 2025, pp. 111–120.

[5] A. S. Elstein and A. Schwartz, "Clinical reasoning in medicine," *Clinical reasoning in the health professions*, vol. 2, pp. 235–246, 2002.

[6] D. K. Owens, J. D. Goldhaber-Fiebert, and H. C. Sox, "Biomedical decision making: probabilistic clinical reasoning," in *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*. Springer, 2021, pp. 77–120.

[7] F. Salehi, R. Bamler, and S. Mandt, "Probabilistic knowledge graph embeddings," *ICLR*, 2019.

[8] X. Chen, M. Boratko, M. Chen, S. S. Dasgupta, X. L. Li, and A. Mccallum, "Probabilistic box embeddings for uncertain knowledge graph reasoning," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 882–893.

[9] B. Zhao, B. I. Rubinstein, J. Gemmell, and J. Han, "A bayesian approach to discovering truth from conflicting sources for data integration," *Proceedings of the VLDB Endowment*, vol. 5, no. 6, 2012.

[10] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[11] J. Chen, C. Yin, Y. Wang, and P. Zhang, "Predictive modeling with temporal graphical representation on electronic health records," in *IJCAI: proceedings of the conference*, vol. 2024, 2024, p. 5763.

[12] V. Bhise, S. S. Rajan, D. F. Sittig, R. O. Morgan, P. Chaudhary, and H. Singh, "Defining and measuring diagnostic uncertainty in medicine: a systematic review," *Journal of general internal medicine*, vol. 33, no. 1, pp. 103–115, 2018.

[13] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.

[14] C. Xiao, E. Choi, and J. Sun, "Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1419–1428, 2018.

[15] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu, "Clinical information extraction applications: A literature review," *Journal of biomedical informatics*, vol. 77, pp. 34–49, 2018.

[16] S. Kadavath, A. Askell, J. Kernion *et al.*, "Language models (mostly) know what they know," in *Advances in Neural Information Processing Systems*, 2022.

[17] A. Kriz, E. L. Janes, X. Shen, and T. Arbel, "Prompt4trust: A reinforcement learning prompt augmentation framework for clinically-aligned confidence calibration in multimodal large language models," *arXiv preprint arXiv:2507.09279*, 2025.

[18] Y. Tian, Y. Song, and Y. Z. Liu, "Justifying model confidence for answer verification in qa," in *ACL*, 2023.

[19] Y. Xia, P. H. L. de Araujo, K. Zaporojets, and B. Roth, "Influences on llm calibration: A study of response agreement, loss functions, and prompt styles," *arXiv preprint arXiv:2501.03991*, 2025.

[20] C. Yan, X. Fu, Y. Xiong, T. Wang, S. C. Hui, J. Wu, and X. Liu, "Llm sensitivity evaluation framework for clinical diagnosis," in *COLING*, 2025.

[21] T. Savage, J. Wang, R. Gallo, A. Boukil, V. Patel, S. A. A. Safavi-Naini, A. Soroush, and J. H. Chen, "Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment," *Journal of the American Medical Informatics Association*, vol. 32, no. 1, pp. 139–149, 2025.

[22] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," *Advances in neural information processing systems*, vol. 26, 2013.

[23] Z. Sun, Z. Deng, J.-Y. Nie, and J. Tang, "Rotate: Knowledge graph embedding by relational rotation in complex space," in *International Conference on Learning Representations*, 2019.

[24] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard, "Complex embeddings for simple link prediction," in *International Conference on Machine Learning*, 2016, pp. 2071–2080.

[25] M. Nickel, L. Rosasco, and T. Poggio, "Holographic embeddings of knowledge graphs," *AAAI Conference on Artificial Intelligence*, 2016.

[26] B. Yang, W.-t. Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," in *International Conference on Learning Representations*, 2015.

[27] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: A survey of approaches and applications," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 12, 2017, pp. 2724–2743.

[28] X. Chen, N. Zhang, X. Li, Y. Feng, X. Chen, and H. Wang, "Embedding uncertain knowledge graphs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3363–3370.

[29] W. Y. Wang, K. Mazaitis, and W. W. Cohen, "Programming with personalized pagerank: a locally groundable first-order probabilistic logic," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 2013, pp. 2129–2138.

[30] Y. Zhu, J. Wu, Y. Wang, H. Zhou, J. Chen, E. Kharlamov, and S. Staab, "Certainty in uncertainty: Reasoning over uncertain knowledge graphs with statistical guarantees," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 8741–8763.

[31] Z. Lin, S. Trivedi, and J. Sun, "Generating with confidence: Uncertainty quantification for black-box large language models," *arXiv e-prints*, pp. arXiv–2305, 2023.

[32] J. Geng, F. Cai, Y. Wang, H. Koeppl, P. Nakov, and I. Gurevych, "A survey of confidence estimation and calibration in large language models," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024, pp. 6577–6595.

[33] X. Liu, T. Chen, L. Da, C. Chen, Z. Lin, and H. Wei, "Uncertainty quantification and confidence calibration in large language models: A survey," *arXiv preprint arXiv:2503.15850*, 2025.

[34] L. Aichberger, K. Schweighofer, M. Ielanskyi, and S. Hochreiter, "Improving uncertainty estimation through semantically diverse language generation," *arXiv preprint arXiv:2406.04306*, 2024.

[35] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[36] M. Nye, A. J. Andreassen, G. Gur-Ari, H. Michalewski, J. Austin, D. Bieber, D. Dohan, A. Lewkowycz, M. Bosma, D. Luan *et al.*, "Show your work: Scratchpads for intermediate computation with language models," *arXiv preprint arXiv:2112.00114*, 2021.

[37] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," in *The Eleventh International Conference on Learning Representations*, 2022.

[38] A. Wang, L. Song, Y. Tian, B. Peng, L. Jin, H. Mi, J. Su, and D. Yu, "Self-consistency boosts calibration for math reasoning," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 6023–6029.

[39] Z. Zhao, E. Wallace, S. F. Wang *et al.*, "Calibrate before use: Improving few-shot performance of language models," in *International Conference on Machine Learning*, 2021, pp. 12 697–12 706.

[40] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*. PMLR, 2017, pp. 1321–1330.

[41] B. Wen, R. Norel, J. Liu, T. Stappenbeck, F. Zulkernine, and H. Chen, "Leveraging large language models for patient engagement: the power of conversational ai in digital health," in *2024 IEEE International Conference on Digital Health (ICDH)*. IEEE, 2024, pp. 104–113.

[42] L. C. Nickels, T. L. Marshall, E. Edgerton, P. W. Brady, P. A. Hagedorn, and J. J. Lee, "Defining diagnostic uncertainty as a discourse type: A transdisciplinary approach to analysing clinical narratives of electronic health records," *Applied linguistics*, vol. 45, no. 1, pp. 134–162, 2024.

[43] N. Bodemer, Y. Hanoch, and K. V. Katsikopoulos, "Heuristics: foundations for a novel approach to medical decision making," *Internal and emergency medicine*, vol. 10, no. 2, pp. 195–203, 2015.

[44] K. Zou, Z. Chen, X. Yuan, X. Shen, M. Wang, and H. Fu, "A review of uncertainty estimation and its application in medical imaging," *Meta-Radiology*, vol. 1, no. 1, p. 100003, 2023.

[45] E. H. Shortliffe and B. G. Buchanan, "A model of inexact reasoning in medicine," *Mathematical Biosciences*, vol. 23, no. 3-4, pp. 351–379, 1975.

[46] J. Kauffman, E. Holmes, A. Vaid, A. W. Charney, P. Kovatch, J. Lampert, A. Sakhuja, M. Zitnik, B. S. Glicksberg, I. Hofer *et al.*, "Infehr: Resolving clinical uncertainty through deep geometric learning on electronic health records," *medRxiv*, pp. 2025–01, 2025.

[47] S. G. Upadhyay and K. M. Kelkar, "Predicting learner's confidence from their behaviour using a judgement questionnaire," in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*. IEEE, 2018, pp. 1–6.

[48] T. Takayanagi, R. Hashimoto, C.-C. Chen, and K. Izumi, "The impact and feasibility of self-confidence shaping for ai-assisted decision-making," *arXiv preprint arXiv:2502.14311*, 2025.

[49] Q. Li and P. K. Varshney, "Does confidence reporting from the crowd benefit crowdsourcing performance?" in *Proceedings of the 2nd International Workshop on Social Sensing*, 2017, pp. 49–54.

[50] A. E. Greenberg and H. E. Hershfield, "Financial decision making," *Consumer Psychology Review*, vol. 2, no. 1, pp. 17–29, 2019.

[51] Y. Zhu, X. Wang, J. Chen, S. Qiao, Y. Ou, Y. Yao, S. Deng, H. Chen, and N. Zhang, "Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities," *World Wide Web*, vol. 27, no. 5, p. 58, 2024.

[52] D. Shu, T. Chen, M. Jin, C. Zhang, M. Du, and Y. Zhang, "Knowledge graph large language model (kg-llm) for link prediction," *Proceedings of Machine Learning Research*, vol. 260, no. 62, 2024.

[53] Y. Tian, D. Song, Z. Wu, C. Zhou, H. Wang, J. Yang, J. Xu, R. Cao, and H. Wang, "Augmenting reasoning capabilities of llms with graph structures in knowledge base question answering," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 11 967–11 977.

[54] F. Shami, S. Marchesin, and G. Silvello, "Fact verification in knowledge graphs using llms," in *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025, pp. 3985–3989.