ONTOSKILL: ONTOLOGY-AWARE AGENTIC SKILL EVOLUTION FOR EHR CLINICAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Electronic Health Records (EHR) clinical prediction requires integrating heterogeneous observations and their complex temporal and clinical dependencies, while agents driven by Large Language Models (LLMs) provide a promising technical direction. Recent studies increasingly explore the agent skill paradigm for experience-driven capability evolution across medical reasoning tasks. However, extending agent skills to EHR clinical prediction faces two fundamental challenges: *(1) clinical semantic fragmentation*, where isolated measurements lack explicit mappings and dependencies for coherent patient states, and *(2) insufficient structural guidance*, where performance feedback alone cannot induce clinically meaningful skill structures. To address these issues, we construct an EHR-oriented ontology that structures fragmented observations into meaningful clinical entities, atomic relations, and reasoning rules. We further propose ***OntoSkill***, an ontology-aware agentic skill evolution framework that autonomously forms and evolves reusable clinical skills from experience. *OntoSkill* resolves three coupled decisions in skill evolution. *(i) Where to form:* grounding fragmented longitudinal observations into ontology-grounded patient states. *(ii) Which to evolve:* prioritizing skills based on accumulated prediction experience. *(iii) How to evolve:* contrasting successful and failed experiences under ontology rules to localize reasoning deficiencies and refine the selected skill. Comprehensive experiments on multiple EHR datasets demonstrate that *OntoSkill* consistently outperforms state-of-the-art baselines, validating its effectiveness for EHR clinical prediction.

1 INTRODUCTION

Electronic Health Records (EHR) Choi et al. (2018); Tomašev et al. (2019); Yang et al. (2023); Tang et al. (2024); Jee et al. (2024) provide longitudinal representations of patient health and serve as a critical foundation for data-driven clinical decision support. EHR clinical prediction integrates heterogeneous observations, including laboratory measurements, physiological signals, and temporal clinical events, to infer clinically relevant patient states and outcomes. Modeling their temporal evolution and complex clinical dependencies remains a key challenge. Recently, agents driven by Large Language Models (LLMs) have shown strong potential in medical reasoning Ma et al. (2026); Jiang et al. (2026); Xia et al. (2026), establishing agentic reasoning as a promising direction for EHR clinical prediction Shi et al. (2024); Fang et al. (2026).

Recent studies on agentic reasoning for EHR clinical prediction have increasingly incorporated external knowledge retrieval Jiang et al. (2025), tool-augmented reasoning Shi et al. (2024), and multi-agent collaboration Wang et al. (2025a); Zhu et al. (2025) to better integrate heterogeneous clinical evidence for prediction. Meanwhile, the emerging agent skill paradigm enables agents to acquire and refine persistent procedural skills from interaction experience, allowing their capabilities to continually evolve beyond those inherited from the underlying models and thereby supporting agent self-evolution. Its applications to medical question answering Sun et al. (2026), diagnostic decision-making Cao et al. (2026), and longitudinal patient monitoring Ossowski et al. (2026) further advance medical agents toward experience-driven capability evolution.

Despite these advances, extending agent skills to EHR clinical prediction remains constrained by two fundamental challenges: **clinical semantic fragmentation** and **insufficient structural guidance**, as illustrated in Fig.1. *Clinically*, EHRs record heterogeneous measurements as isolated variables,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

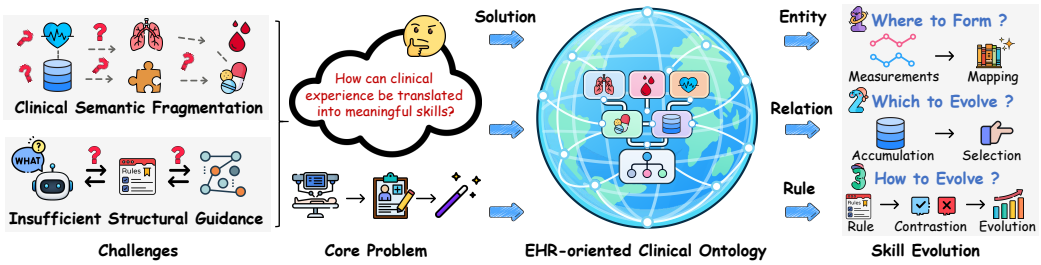


Figure 1: Illustration of the key challenges in EHR clinical prediction, showing *clinical semantic fragmentation* and *insufficient structural guidance* for skill evolution. *OntoSkill* organizes fragmented observations with an EHR-oriented clinical ontology whose entities, relations, and rules guide where skills form, which skills evolve, and how selected skills evolve.

lacking the semantic mappings and cross-concept dependencies required to organize measurements into meaningful clinical concepts and coherent patient states. *Technically*, existing methods rely mainly on experience-level performance feedback, providing limited structural guidance for clinically grounded skill formation and evolution. *Collectively, these challenges reveal the absence of a principled mechanism for grounding clinical experience into meaningful skills.*

Motivated by these observations, we first construct an EHR-oriented ontology tailored to clinical prediction, which organizes fragmented observations into meaningful clinical entities and encodes atomic relations and reasoning rules. Guided by this ontology, we propose ***OntoSkill***, an ontology-aware agentic skill evolution framework that enables reusable clinical skills to be autonomously formed and evolved from experience. Methodologically, *OntoSkill* integrates three carefully designed mechanisms: **O**ntology-aware **S**emantic **G**rounding (**OSG**), **O**ntology-aware **E**mpirical **P**rioritization (**OEP**), and **O**ntology-aware **C**ontrastive **E**volution (**OCE**). Together, these mechanisms resolve three coupled decisions in skill evolution. *(i) Where to form:* OSG maps fragmented measurements to clinically meaningful entities and instantiates patient states in their local ontologies, anchoring skill formation, activation, and experience accumulation. *(ii) Which to evolve:* OEP prioritizes ontology-grounded skills based on accumulated prediction experience, identifying the skill with the greatest need for refinement. *(iii) How to evolve:* OCE contrasts successful and failed experiences under ontology rules to localize reasoning deficiencies and refine the selected skill within clinically valid constraints. Our main contributions are summarized as follows:

- To the best of our knowledge, this is ***the first attempt*** to construct an EHR-oriented clinical ontology specifically for agentic skill evolution in clinical prediction, which organizes fragmented observations into clinical entities and encodes their atomic relations and reasoning rules. This provides an explicit clinical structure to guide skill evolution.
- We propose *OntoSkill*, a unified ontology-aware agentic skill evolution framework that operationalizes the clinical ontology throughout the skill lifecycle. By coupling *OSG*, *OEP*, and *OCE*, it determines *where to form* skills, *which to evolve*, and *how to evolve* them, enabling targeted and structurally grounded evolution from experience.
- Comprehensive experiments on multiple EHR benchmarks demonstrate that *OntoSkill* consistently outperforms state-of-the-art baselines, validating the effectiveness of ontology-guided agentic skill evolution for EHR clinical prediction.

2 Related Works

2.1 EHR Clinical Prediction

EHR clinical prediction aims to forecast future diagnoses and outcomes from longitudinal patient records Choi et al. (2016); Li et al. (2020); Jiang et al. (2025). Early studies relied primarily on *attention-based architectures* to model temporal dynamics and hierarchical structure in EHR sequences Choi et al. (2016); Ma et al. (2017); Choi et al. (2018). *Transformer-based models* subsequently improved patient representations by capturing long-range dependencies across longitudinal records Li et al. (2020); Rasmy et al. (2021); Yang et al. (2023). Recent studies explored *agentic*

reasoning methods, which enhanced EHR clinical prediction through external knowledge retrieval Jiang et al. (2025); Ou et al. (2026), tool-augmented reasoning Shi et al. (2024), and multi-agent collaboration Wang et al. (2025a); Zhu et al. (2025).

2.2 Agentic Skill Evolution

Agent skills encoded reusable procedural knowledge that enabled LLM agents to accumulate and transfer experience across related tasks Wang et al. (2025b); Ding et al. (2026); Wang et al. (2026); Sun et al. (2026); Cao et al. (2026). *In general agentic tasks*, existing studies focused on how skills were represented and learned, ranging from reusable experience memories and workflows to executable programs, with interaction feedback further supporting the continual refinement of skill libraries Wang et al. (2025b); Ding et al. (2026); Wang et al. (2026). *In medical reasoning tasks*, the emphasis shifted toward grounding reusable capabilities in clinical knowledge and patient experience, such as self-evolving skill memories Sun et al. (2026) and guideline-grounded executable skills Cao et al. (2026); Ossowski et al. (2026) for reasoning across clinical cases.

2.3 Clinical Ontology Applications

Clinical ontologies introduced structured medical knowledge into computational reasoning (Choi et al., 2017; Dissanayake et al., 2020; Schmidt et al., 2024; Welland et al., 2026). *For EHR clinical prediction*, GRAM (Choi et al., 2017) exploited ontology hierarchies through graph-based attention to learn clinically informed representations, while MMORE (Song et al., 2019) learned medical concept embeddings from multiple ontological representations for diagnosis prediction. *For clinical decision support*, clinical reasoning ontologies encoded medical concepts and reasoning relations to support structured inference (Dissanayake et al., 2020). *In rare-disease diagnosis*, TNAMSE (Schmidt et al., 2024) used HPO-encoded phenotypes with genomic and next-generation phenotyping to facilitate molecular diagnosis, while Talos (Welland et al., 2026) leveraged the HPO ontology structure for semantic phenotype matching and variant prioritization.

Summary and Motivation. The above studies advance clinical prediction toward more structured, knowledge-grounded, and reusable reasoning. Inspired by these studies, we depart from prior work by constructing an EHR-oriented ontology specifically for agentic skill evolution in clinical prediction, providing a principled basis for grounding clinical experience into meaningful skills throughout the skill lifecycle by guiding three key questions: *where to form*, *which to evolve*, and *how to evolve*.

3 Preliminaries

3.1 Problem Formulation

Given a patient EHR, we represent its longitudinal record as a sequence of $\mathcal{T}$ time-ordered visits, $\mathcal{X} = [x_1, \ldots, x_{\mathcal{T}}]$, where each visit $x_t = \{l_{t,n}\}_{n=1}^{\mathcal{N}}$ contains $\mathcal{N}$ clinical measurements. Given $\mathcal{X}$, the goal is to predict a binary clinical outcome $\mathcal{Y} \in \{0, 1\}$. We consider two standard tasks: *1) in-hospital mortality prediction*, where $\mathcal{Y} = 1$ indicates death during hospitalization, and *2) 30-day readmission prediction*, where $\mathcal{Y} = 1$ indicates readmission within 30 days after discharge. The predicted probability and its corresponding binary prediction are denoted by p and y, respectively.

3.2 EHR-Oriented Clinical Ontology

In this paper, we construct an EHR-oriented clinical ontology $\mathcal{O} = (\mathcal{C}, \mathcal{R}, \mathcal{Q})$, where $\mathcal{C}$ contains meaningful clinical entities corresponding to EHR variables, $\mathcal{R}$ contains atomic clinical relations, and $\mathcal{Q}$ contains clinical reasoning rules that organize these relations into clinically valid reasoning chains. Clinical interpretation of EHR data commonly requires contextualizing individual clinical concepts, reasoning over physiologically coupled concepts, and integrating complementary evidence across concepts (Ranieri et al., 2012; Singer et al., 2016). Accordingly, ***OntoSkill*** defines three types of binary atomic relations:

$$\rho^{\text{con}}_{u,v} = \textsc{Conditional}(c_v \mid c_u), \quad \rho^{\text{mut}}_{u,v} = \textsc{Mutual}(c_u, c_v), \quad \rho^{\text{joint}}_{u,v} = \textsc{Joint}(c_u, c_v), \tag{1}$$

where $c_u, c_v \in \mathcal{C}$, and CONDITIONAL denotes a context-dependent relation in which the state of c_u provides clinical context for interpreting c_v (Singer et al., 2016). MUTUAL associates physiologically coupled entities whose states provide complementary evidence for a shared clinical state (Vincent & De Backer, 2013). JOINT captures the integrated clinical meaning of two entities whose combination provides information beyond their individual interpretations (Hyland et al., 2020). A clinical reasoning rule specifies an ordered composition of atomic relations:

$$q_k : \rho_{k,1} \Rightarrow \rho_{k,2} \Rightarrow \cdots \Rightarrow \rho_{k,L_k}, \qquad L_k \geq 2, \tag{2}$$

where $\rho_{k,\ell} \in \mathcal{R}$ and $\Rightarrow$ denotes clinical reasoning precedence, such that the interpretation established by an earlier relation is incorporated into subsequent reasoning. For example,

$$\begin{aligned} q : \textsc{Conditional}\big(&\texttt{“Respiratory Rate”} \mid \texttt{“Body Temperature”}\big) \\ \Rightarrow \textsc{Mutual}\big(&\texttt{“Oxygenation”}, \texttt{“Respiratory Rate”}\big). \end{aligned} \tag{3}$$

This rule first interprets the respiratory rate in the context of body temperature and then jointly considers respiratory rate and oxygenation to assess respiratory status. For each $c \in \mathcal{C}$, we define the local ontology $\mathcal{O}_c = (\mathcal{C}_c, \mathcal{R}_c, \mathcal{Q}_c)$, where $\mathcal{C}_c \subseteq \mathcal{C}$ contains clinically related concepts, $\mathcal{R}_c \subseteq \mathcal{R}$ contains their supported atomic relations, and $\mathcal{Q}_c \subseteq \mathcal{Q}$ contains the corresponding reasoning rules.

3.3 CLINICAL SKILL REPRESENTATION

Each c anchors a persistent root skill ω_c. At evolution round h, ω_c is represented as:

$$\omega_c^{(h)} = \left(\phi_c^{(h)}, \pi_c^{\text{root}}, \Gamma_c^{(h)}\right), \tag{4}$$

π_c^{root} specifies the reusable base instruction for interpreting the root entity c and constructing its entity-level clinical evidence, while the relation-level reasoning component is defined as:

$$\Gamma_c^{(h)} = \left[\left\langle \rho_{c,1}^{(h)}, \pi_{c,1}^{(h)} \right\rangle, \ldots, \left\langle \rho_{c,M_c^{(h)}}^{(h)}, \pi_{c,M_c^{(h)}}^{(h)} \right\rangle\right], \qquad \rho_{c,m}^{(h)} \in \mathcal{R}_c. \tag{5}$$

Here, $\Gamma_c^{(h)}$ denotes an ordered sequence of relation-based reasoning operations. Each relation identifies a clinical dependency, while its instruction specifies how the corresponding grounded states are interpreted and integrated. By applying ontology relations under clinical rules, each skill extends from its root entity to structured reasoning over related clinical concepts. For each $q_k : \rho_{k,1} \Rightarrow \cdots \Rightarrow \rho_{k,L_k} \in \mathcal{Q}_c$, the relative order of any two relations included in the skill satisfies:

$$\rho_{k,i} \prec_{\omega_c^{(h)}} \rho_{k,j}, \qquad \langle \rho_{k,i}, \cdot \rangle, \langle \rho_{k,j}, \cdot \rangle \in \Gamma_c^{(h)}, \quad i < j, \tag{6}$$

where $\prec_{\omega_c^{(h)}}$ denotes their execution order in $\omega_c^{(h)}$. Accordingly, skill evolution may expand or prune clinically supported relations, refine their organization and executable instructions, while remaining within the structural space defined by $\mathcal{O}_c$.

4 METHODOLOGY

4.1 OVERVIEW AND INITIALIZATION

Overview. Building on $\mathcal{O}$, *OntoSkill* evolves reusable skills through three mechanisms. *(i) Ontology-aware Semantic Grounding* anchors longitudinal observations to meaningful clinical entities and patient states; *(ii) Ontology-aware Empirical Prioritization* selects the skill most in need of evolution; and *(iii) Ontology-aware Contrastive Evolution* contrasts matched successes and failures under ontology rules to guide its evolution. The pipeline of *OntoSkill* is shown in Fig. 2.

Ontology and Skill Initialization. *OntoSkill* first constructs the clinical ontology and initializes a compact seed skill for each entity using an external medical knowledge source $\mathcal{S}$. For each $c \in \mathcal{C}$:

$$s_c = \text{Retrieve}(c; \mathcal{S}), \qquad (\mathcal{O}_c, \omega_c^{(0)}) = \text{Init}(c; s_c), \tag{7}$$

where s_c denotes the retrieved medical evidence, $\mathcal{O}_c$ contains the clinically supported entities, atomic relations, and reasoning rules associated with c, and $\omega_c^{(0)}$ provides an entity-level reasoning anchor with a compact initial relation structure. Subsequent evolution extends this seed within $\mathcal{O}_c$ according to accumulated prediction experience. Aggregating the entity-level structures establishes the global ontology and the initial skill library (See Appendix A for details).

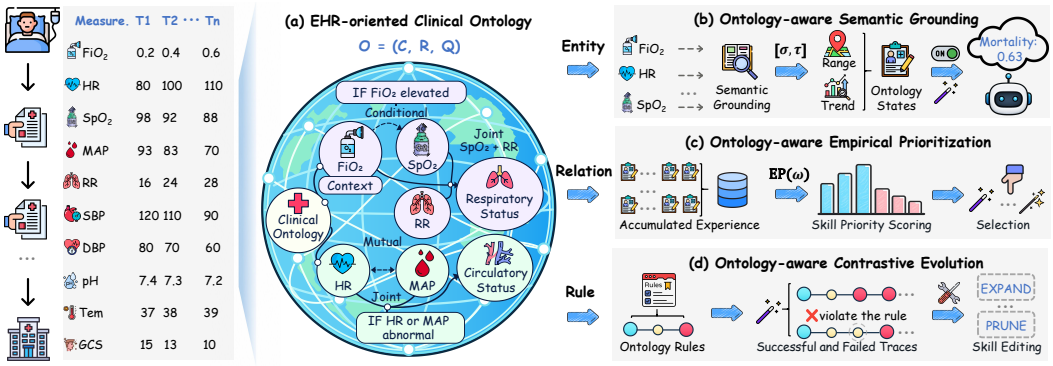


Figure 2: Overview of *OntoSkill*. (a) **EHR-oriented Clinical Ontology:** longitudinal EHR observations are organized through clinical entities, relations, and reasoning rules. (b) **Ontology-aware Semantic Grounding:** observations are grounded into ontology-defined patient states to anchor skill formation and activation. (c) **Ontology-aware Empirical Prioritization:** accumulated prediction experience is used to prioritize ontology-grounded skills for evolution. (d) **Ontology-aware Contrastive Evolution:** successful and failed reasoning traces are contrasted under ontology rules to identify reasoning deficiencies and refine the selected skill.

4.2 Ontology-Aware Skill Evolution

Where to Form: Ontology-aware Semantic Grounding. OSG grounds fragmented measurements into meaningful clinical entities and instantiates their local ontologies with patient-specific states, supporting skill activation and experience accumulation. Specifically, the longitudinal observations of the $\mathcal{N}$ raw clinical measurements are first mapped to $\mathcal{M}$ clinically meaningful entities:

$$\{l_{t,n}\}_{t=1,n=1}^{\mathcal{T},\mathcal{N}} \xrightarrow{\text{Ground}} \left\{c_m\left[\{\sigma_{m,j},\tau_{m,j}\}_{j=1}^{\mathcal{J}_m}\right]\right\}_{m=1}^{\mathcal{M}}, \qquad \mathcal{M}\le\mathcal{N}, \quad \sum_{m=1}^{\mathcal{M}}\mathcal{J}_m=\mathcal{N}, \tag{8}$$

where $\mathcal{J}_m$ is the number of raw measurements assigned to entity c_m. For the j-th measurement associated with c_m, $\sigma_{m,j} \in \{\texttt{low},\texttt{normal},\texttt{high}\}$ denotes its current status relative to the corresponding clinical reference range, and $\tau_{m,j} \in \{\texttt{rising},\texttt{falling},\texttt{stable},\texttt{rise-fall},\texttt{fall-rise}\}$ summarizes its longitudinal trend. The initial activation condition for the skill anchored at c_m is defined as:

$$\phi_{c_m}^{(0)}\big(\mathcal{O}_{c_m}(\mathcal{X})\big)=\mathbb{I}[\exists j\in\{1,\dots,\mathcal{J}_m\}:\sigma_{m,j}\neq\texttt{normal}\ \vee\ \tau_{m,j}\neq\texttt{stable}],m=1,\dots,\mathcal{M}. \tag{9}$$

Thus, the skill anchored at c_m is activated whenever at least one measurement assigned to the entity is abnormal or exhibits a non-stable trend. The activated skill set is:

$$\Omega=\left\{\omega_{c_m}^{(h)}\;\middle|\;\phi_{c_m}^{(h)}\big(\mathcal{O}_{c_m}(\mathcal{X})\big)=1,\ m=1,\dots,\mathcal{M}\right\}. \tag{10}$$

Given $\mathcal{O}_{c_m}(\mathcal{X})$, each activated skill executes its root instruction and ordered relation operations:

$$(e_{c_m},p_{c_m},\xi_{c_m})=\omega_{c_m}^{(h)}\big(\mathcal{O}_{c_m}(\mathcal{X})\big),\quad \forall\, m\in\{1,\dots,\mathcal{M}\}\text{ with }\phi_{c_m}^{(h)}\big(\mathcal{O}_{c_m}(\mathcal{X})\big)=1, \tag{11}$$

where e_{c_m} denotes the generated clinical evidence, $p_{c_m}\in[0,1]$ denotes the outcome probability independently estimated by $\omega_{c_m}^{(h)}$, and ξ_{c_m} records the corresponding relation-level reasoning trace. The clinical evidence from activated skills is then aggregated into the final prediction:

$$(p,y)=\text{Aggregate}\left(\left\{(e_{c_m})\mid m\in\{1,\dots,\mathcal{M}\},\ \phi_{c_m}^{(h)}\big(\mathcal{O}_{c_m}(\mathcal{X})\big)=1\right\}\right). \tag{12}$$

Remark 1. *OSG establishes a meaningful ontology-level grounding from heterogeneous observations to persistent skills, supporting reliable skill activation and consistent experience accumulation.*

Which to Evolve: Ontology-aware Empirical Prioritization. When multiple skills contribute to a prediction, case-level correctness cannot identify which skill underlies the unresolved error. OEP therefore accumulates outcome-resolved experience for each activated skill:

$$\forall\,\omega_{c_m}^{(h)}\in\Omega,\qquad \mathcal{P}_{c_m}^{(h)}\leftarrow\mathcal{P}_{c_m}^{(h)}\cup\{(\mathcal{X},\mathcal{Y},y,p_{c_m},\xi_{c_m})\}, \tag{13}$$

where $\mathcal{P}_{c_m}^{(h)}$ stores experience associated with the current version of $\omega_{c_m}^{(h)}$. For each incorrectly predicted case $\mathcal{X}_i$, OEP assigns each activated skill a normalized responsibility score based on its skill-local outcome discrepancy:

$$r_{i,c_m} = \frac{\mathcal{L}_{\mathrm{CE}}(\mathcal{Y}_i, p_{i,c_m})}{\sum_{\omega_j^{(h)} \in \Omega_i} \mathcal{L}_{\mathrm{CE}}(\mathcal{Y}_i, p_{i,j}) + \epsilon}, \tag{14}$$

where $\mathcal{L}_{\mathrm{CE}}$ denotes the cross-entropy loss, p_{i,c_m} is the skill-local outcome probability obtained during prediction, and $\epsilon > 0$ ensures numerical stability. OEP then aggregates failure responsibility over the experience of the current skill version and defines its evolution priority as:

$$\mathrm{EP}(\omega_{c_m}^{(h)}) = \frac{\sum_{i \in \mathcal{P}_{c_m}^{(h)}} \mathbb{I}[y_i \neq \mathcal{Y}_i]\, r_{i,c_m}}{|\mathcal{P}_{c_m}^{(h)}| + \lambda}, \tag{15}$$

where $\lambda > 0$ regularizes priorities for sparsely activated skills. This formulation prioritizes skills that repeatedly account for larger shares of prediction error while suppressing unreliable estimates from limited experience. At evolution round h, OEP selects $\omega_{c_m}^{(h)^\star}$ as follows:

$$\omega_{c_m}^{(h)^\star} = \arg\max_{\omega_{c_m}^{(h)}:\, |\mathcal{P}_{c_m}^{(h)}|>0} \mathrm{EP}(\omega_{c_m}^{(h)}). \tag{16}$$

Remark 2. *OEP translates global outcome feedback into skill-level empirical responsibility, prioritizing the skill that most persistently contributes to unresolved prediction errors.*

How to Evolve: Ontology-aware Contrastive Evolution. OCE exploits the local ontology $\mathcal{O}_{c_m}$ of the selected skill $\omega_{c_m}^{(h)\star}$ to contrast successful and failed experiences and identify relation-level deficiencies in its current reasoning. The accumulated experience is first partitioned according to prediction correctness:

$$\begin{aligned} \mathcal{P}_{c_m}^{(h)+} &= \left\{ (\mathcal{X}, \mathcal{Y}, y, p_{c_m}, \xi_{c_m}^{+}) \in \mathcal{P}_{c_m}^{(h)} \mid y = \mathcal{Y} \right\}, \\ \mathcal{P}_{c_m}^{(h)-} &= \left\{ (\mathcal{X}, \mathcal{Y}, y, p_{c_m}, \xi_{c_m}^{-}) \in \mathcal{P}_{c_m}^{(h)} \mid y \neq \mathcal{Y} \right\}, \end{aligned} \tag{17}$$

where $\xi_{c_m}^{+}$ and $\xi_{c_m}^{-}$ denote the relation-level reasoning traces associated with successful and failed predictions, respectively. OCE further groups these experiences by ground-truth outcome and aligns their reasoning traces under the same clinical reasoning rules. For each $q_k : \rho_{k,1} \Rightarrow \cdots \Rightarrow \rho_{k,L_k} \in \mathcal{Q}_{c_m}$ and $\mathcal{Y} \in \{0, 1\}$, we collect:

$$\mathcal{D}_{k,\mathcal{Y}}^{(h)\pm} = \left\{ [\xi_{i,c_m}^{\pm}[\rho_{k,1}], \ldots, \xi_{i,c_m}^{\pm}[\rho_{k,L_k}]] \mid (\mathcal{X}_i, \mathcal{Y}_i, y_i, p_{i,c_m}, \xi_{i,c_m}^{\pm}) \in \mathcal{P}_{c_m}^{(h)\pm},\ \mathcal{Y}_i = \mathcal{Y} \right\}, \tag{18}$$

where $\xi_{i,c_m}^{\pm}[\rho] = \varnothing$ when relation ρ is absent from the corresponding reasoning trace. Thus, $\mathcal{D}_{k,\mathcal{Y}}^{(h)-}$ and $\mathcal{D}_{k,\mathcal{Y}}^{(h)+}$ align failed and successful reasoning under the same ontology rule and clinical outcome. Following prior work that employs LLMs for feedback-driven refinement (Yang et al., 2024), OCE first employs an LLM evaluator to contrast the aligned successful and failed traces and identify recurrent reasoning deficiencies:

$$\eta_{c_m}^{(h)} = \mathrm{LLM}_{\mathrm{eval}} \left(\omega_{c_m}^{(h)^\star}, \left\{ (\mathcal{D}_{k,\mathcal{Y}}^{(h)-}, \mathcal{D}_{k,\mathcal{Y}}^{(h)+}) \right\}_{k,\mathcal{Y}} \,\middle|\, \mathcal{R}_{c_m}, \mathcal{Q}_{c_m} \right), \tag{19}$$

where η_{c_m} is a structured textual evaluation summarizing the identified deficiencies and their corresponding revision directions. Skill evolution is restricted to the following operation set $\mathcal{A}_{c_m}$:

$$\mathcal{A}_{c_m} = \{\textsc{Expand}(\rho, \pi), \textsc{Prune}(\rho), \textsc{Reorder}(\rho_a, \rho_b), \textsc{Rewrite}(\rho, \pi')\}, \tag{20}$$

where EXPAND adds an ontology-supported relation operation, PRUNE removes an existing optional operation, REORDER adjusts the order of existing operations, and REWRITE refines a relation-specific instruction. Conditioned on the evaluation η_{c_m}, a second LLM directly improves the selected skill within this ontology-constrained operation space:

$$\widetilde{\omega}_{c_m}^{(h)} = \mathrm{LLM}_{\mathrm{edit}} \left(\omega_{c_m}^{(h)^\star}, \eta_{c_m}^{(h)} \,\middle|\, \mathcal{R}_{c_m}, \mathcal{Q}_{c_m}, \mathcal{A}_{c_m} \right). \tag{21}$$

Table 1: Comparison (%) with State-of-the-Art methods across four paradigms: *1) external knowledge retrieval*, *2) tool-augmented reasoning*, *3) multi-agent collaboration*, and *4) agentic skill evolution* on representative EHR datasets.

METHOD	MIMIC-IV MORTALITY		MIMIC-IV READMISSION		TJH MORTALITY		CDSL MORTALITY	
	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
Qwen3-8B								
Direct	90.52±2.74	44.05±9.50	63.02±6.92	18.78±4.96	81.42±2.74	71.92±4.49	85.73±2.26	47.21±7.33
RAG	88.92±3.33	42.69±10.30	62.94±6.96	22.34±6.55	70.03±3.30	60.51±4.58	86.26±3.18	47.97±7.78
EHRagent†	90.87±2.86	45.40±9.93	62.32±6.56	21.30±5.95	80.94±2.80	74.27±4.51	85.23±2.78	39.63±6.56
MedLA	91.70±2.95	**58.49**±10.41	64.02±6.83	23.23±6.66	80.99±3.02	75.80±4.96	**88.57**±3.88	**57.47**±7.76
ExpRAG	90.50±2.41	46.06±10.26	**66.15**±6.48	21.02±5.82	–±–	–±–	85.63±2.29	43.56±7.43
Skill-Pro	91.95±2.88	50.20±9.52	64.03±6.34	21.78±6.58	81.53±2.67	75.52±4.18	87.08±2.51	46.62±6.68
OntoSkill	**92.73**±2.34	54.69±9.55	65.74±6.13	**23.32**±6.83	**85.38**±2.31	**78.24**±3.71	87.66±2.65	52.38±8.12
Claude-Sonnet-4.6								
Direct	92.70±2.91	60.91±10.39	63.74±6.74	21.55±5.74	87.54±2.43	80.25±4.19	91.85±2.37	70.52±6.60
RAG	93.20±2.41	56.07±10.12	67.63±6.68	23.34±6.60	78.26±3.25	69.62±4.79	93.66±2.00	71.76±6.95
EHRagent†	93.22±2.30	62.25±9.72	64.66±6.57	20.25±5.89	93.20±1.90	89.07±3.50	91.90±1.97	57.27±7.60
MedLA	92.07±3.04	63.75±10.63	65.92±6.96	23.90±6.70	89.74±2.58	80.97±5.47	92.12±3.21	70.84±7.45
ExpRAG	93.58±2.35	60.10±10.73	67.52±6.79	24.17±7.09	–±–	–±–	93.27±1.74	69.54±6.04
Skill-Pro	**94.17**±2.40	70.14±9.60	64.25±6.97	20.46±5.55	87.91±2.37	81.69±3.62	91.78±2.24	67.69±6.53
OntoSkill	94.11±3.12	**74.46**±9.29	**69.74**±6.55	**25.82**±6.87	**95.22**±1.49	**92.71**±2.36	**95.18**±1.87	**84.85**±4.59
DeepSeek-V4-Flash								
Direct	94.08±2.89	71.99±9.86	63.86±7.60	24.25±6.65	90.01±2.12	83.17±4.21	93.66±2.30	70.11±7.43
RAG	93.41±3.23	75.79±9.21	67.92±5.11	21.61±6.12	84.48±2.51	81.69±3.80	90.20±2.65	63.96±6.83
EHRagent†	93.11±3.31	75.64±9.48	68.91±7.46	24.64±6.05	85.05±2.72	77.06±4.55	93.08±2.12	64.51±7.39
MedLA	94.68±2.64	75.42±9.69	69.84±7.14	26.91±7.22	93.35±1.69	90.21±2.63	**96.10**±1.63	**85.12**±5.27
ExpRAG	**95.59**±2.04	73.81±10.67	67.39±6.75	25.17±7.20	–±–	–±–	95.70±1.61	80.61±6.58
Skill-Pro	93.66±3.20	76.29±9.46	68.45±6.96	25.73±7.53	93.05±1.66	90.45±2.57	91.53±3.46	71.78±8.34
OntoSkill	95.52±2.53	**83.01**±7.61	**70.18**±6.17	**27.31**±7.68	**93.69**±1.79	**91.02**±2.68	95.73±1.56	84.45±5.71

Table 2: Evaluation (%) of *OntoSkill* transferability across TJH and CDSL.

METHOD	CDSL → TJH		TJH → CDSL	
	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
Direct	90.01±2.12	83.17±4.21	93.66±2.30	70.11±7.43
RAG	84.48±2.51	81.69±3.80	90.20±2.65	63.96±6.83
EHRagent†	85.05±2.72	77.06±4.55	93.08±2.12	64.51±7.39
MedLA	93.35±1.69	90.21±2.63	96.10±1.63	85.12±5.27
ExpRAG*	92.77±1.93	88.32±3.42	–±–	–±–
Skill-Pro	93.20±1.63	91.22±2.25	94.10±1.89	77.23±6.36
OntoSkill	**95.40**±1.42	**92.96**±2.37	**94.26**±2.14	**77.31**±6.20

Table 3: Evaluation (%) of *OntoSkill* scalability across representative methods on MIMIC-IV.

METHOD	MIMIC-IV MORTALITY		MIMIC-IV READMISSION	
	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
CoT	95.55±2.39	82.71±7.53	67.60±6.70	23.89±6.78
+ ***OntoSkill***	**95.67**±2.45	**85.08**±7.08	**68.60**±7.16	**27.03**±7.71
RAG	93.41±3.23	75.79±9.21	67.92±5.11	21.61±6.12
+ ***OntoSkill***	**95.90**±2.03	**80.03**±8.61	**69.21**±5.90	**23.54**±6.18
EHRagent†	93.11±3.31	75.64±9.48	68.91±7.46	24.64±6.05
+ ***OntoSkill***	**94.32**±2.62	**75.73**±9.85	**69.85**±6.42	**25.75**±7.34
MedLA	94.68±2.64	75.42±9.69	69.84±7.14	26.91±7.22
+ ***OntoSkill***	**96.53**±1.23	**76.50**±9.50	**70.63**±6.89	**27.62**±7.41

Note: All experiments use `DeepSeek-V4-Flash` as the backbone. "→" denotes directly transferring the experience learned on the source (left) dataset to the target (right) dataset, while "+" denotes augmenting the corresponding base method with *OntoSkill*. "*" denotes methods requiring the dataset-provided *discharge summary*, while "†" denotes methods requiring external tools.

We then evaluate the candidate on up to $\mathcal{V}$ randomly sampled cases in which $\omega_{c_m}^{(h)^\star}$ is activated. With all other skills fixed, the candidate is accepted only if it improves validation performance:

$$\omega_{c_m}^{(h+1)} = \begin{cases} \widetilde{\omega}_{c_m}^{(h)}, & \mathcal{V}_{c_m}(\widetilde{\omega}_{c_m}^{(h)}) > \mathcal{V}_{c_m}(\omega_{c_m}^{(h)^\star}), \\ \omega_{c_m}^{(h)^\star}, & \text{otherwise}, \end{cases} \tag{22}$$

where $\mathcal{V}_{c_m}(\cdot)$ denotes the performance evaluated on these sampled cases.

Remark 3. *OCE turns rule-aligned success–failure contrasts into ontology-constrained skill refinement, preserving clinically valid evolution.*

5 EXPERIMENTS

5.1 DATASETS, METRICS, BASELINES AND IMPLEMENTATION DETAILS

Datasets. We evaluate *OntoSkill* on three representative real-world EHR datasets: **MIMIC-IV** (Johnson et al., 2023), **TJH** (Yan et al., 2020), and **CDSL** (Ritoré et al., 2024).

Metrics. We evaluate clinical prediction using two standard metrics: **AUROC**, the area under the receiver operating characteristic curve, and **AUPRC**, the area under the precision–recall curve.

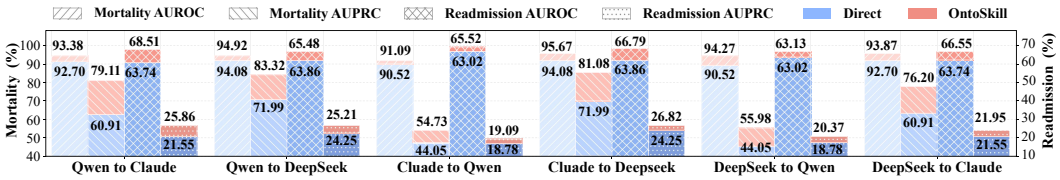


Figure 3: Evaluation (%) of *OntoSkill* transfer ability across multiple models.

Table 4: Ablation studies (%) for the core components of *OntoSkill* on MIMIC-IV.

METHOD	MORTALITY		READMISSION	
	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
`DeepSeek-V4-Flash`	94.08±2.89	71.99±9.86	63.86±7.60	24.25±6.65
+ Clinical Ontology w/o RAG	94.25 ±2.82	81.70±8.57	65.69±6.91	23.49±6.48
+ Clinical Ontology w/ RAG	94.77±3.08	81.77±7.93	66.52±7.65	25.09±7.37
+ OSG (*Where to Form*)	94.90±2.85	82.08±8.31	66.78±7.34	25.86±7.60
+ OCE w/ Random (*How to Evolve*)	94.28±2.61	80.81±7.72	66.73±7.09	24.90±7.24
+ OCE w/o Rules	95.07±3.08	82.48±8.31	67.08±6.45	26.83±7.70
+ OEP (*Which to Evolve*)	**95.52**±2.53	**83.01**±7.61	**70.18**±6.17	**27.31**±7.68

Note. w/ Random denotes randomly selecting a skill for evolution, while *w/o Rules* denotes evolving the selected skill without following ontology-encoded rules.

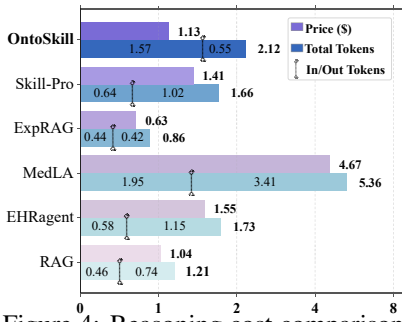


Figure 4: Reasoning cost comparison across methods over 200 cases.

Baselines. For fairness, we evaluate all methods in a training-free setting without parameter updates. We compare *OntoSkill* with representative methods from four paradigms: 1) *external knowledge retrieval*, including **RAG** Lewis et al. (2020) and **ExpRAG** Ferraz et al. (2026); 2) *tool-augmented reasoning*, including **EHRagent** Shi et al. (2024); 3) *multi-agent collaboration*, including **MedLA** Ma et al. (2026); and 4) *agentic skill evolution*, including **Skill-Pro** Mi et al. (2026).

Implementation Details. We evaluate *OntoSkill* with three LLM backbones: `Qwen3-8B`, `Claude-Sonnet-4.6`, and `DeepSeek-V4-Flash`. Following Zhu et al. (2025), we randomly sample 200 patients from the original test set for each prediction task. All scores are reported as mean±standard deviation over 100 bootstrap resamples. (Additional details on datasets, metrics, baselines, and implementation are provided in Appendix C.)

5.2 COMPARISONS WITH STATE-OF-THE-ART

We compare *OntoSkill* with representative methods spanning four paradigms on multiple EHR benchmarks under varied settings. These results highlight four advantages. **1) Superior Performance:** As shown in Table 1, *OntoSkill* ranks first in 16 of 24 comparisons and second in the remaining eight. **2) Robust Generalization:** It maintains strong performance across EHR datasets and LLM backbones, indicating consistent generalization across data distributions and model families. **3) Effective Transferability:** Table 2 shows that transferring skills between TJH and CDSL improves AUPRC by up to 9.79%. Figure 3 further shows that transferring skills across models improves AUPRC for mortality prediction by up to 15.29%, demonstrating that the learned clinical reasoning transfers across datasets and models. **4) Strong Scalability:** As shown in Table 3, augmenting representative EHR reasoning methods with *OntoSkill* consistently improves mortality and readmission prediction, demonstrating applicability across diverse reasoning paradigms.

5.3 ABLATION STUDIES

As shown in Table 4, we conduct ablations on MIMIC-IV to isolate the contribution of each component. These experiments yield four key observations: 1) comparing ontology construction with and without RAG highlights the value of external medical knowledge grounding beyond ontology structure alone; 2) introducing OSG further validates ontology-grounded patient states for reliable skill formation and activation; 3) replacing OEP with random skill selection reveals the importance of experience-guided prioritization for targeted evolution; and 4) removing ontology rules from OCE demonstrates the role of clinical constraints in guiding skill refinement. The complete configuration combines these complementary components and achieves the strongest overall performance.

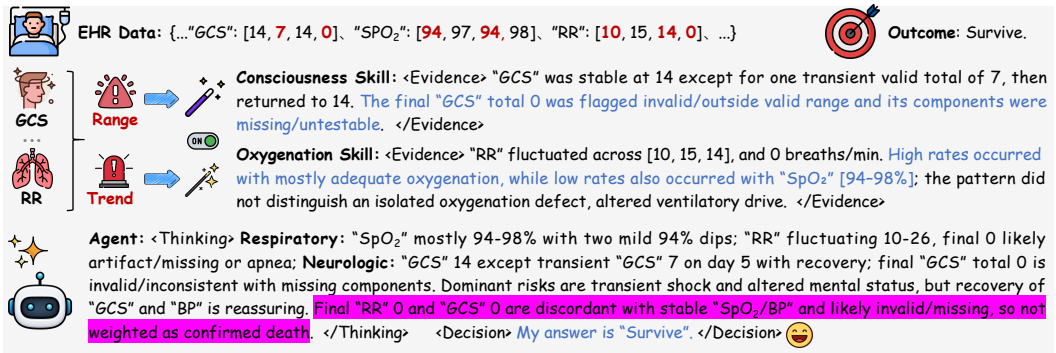


Figure 5: Case analysis of *OntoSkill*.

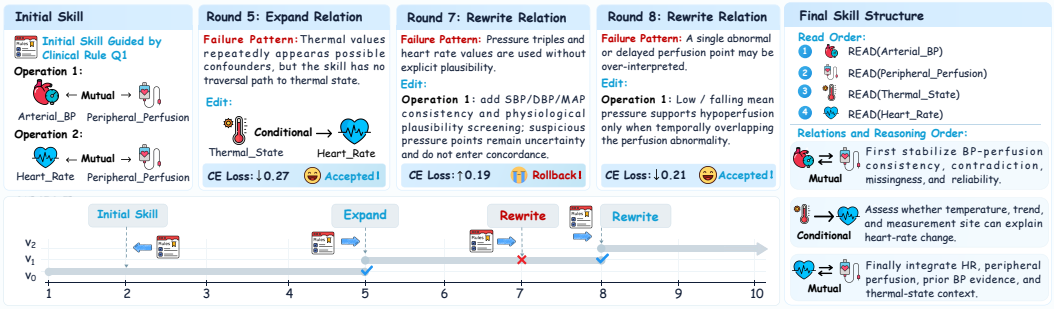


Figure 6: Skill lifecycle analysis of *OntoSkill*.

5.4 Qualitative Analysis

Reasoning Cost Analysis. As shown in Fig. 4, *OntoSkill* costs $1.13 over 200 cases with `DeepSeek-V4-Flash`, lower than Skill-Pro ($1.41) under the same paradigm. Despite higher total token usage, it uses fewer output tokens (0.55M vs. 1.02M), reducing monetary cost. Its overhead over lightweight retrieval-based methods remains modest, while it is substantially cheaper than the multi-agent framework ($1.13 vs. $4.67). Overall, *OntoSkill* achieves stronger clinical reasoning at a practically acceptable computational cost.

Case Analysis. As shown in Fig. 5, we analyze a representative mortality-prediction case to examine whether *OntoSkill* can integrate heterogeneous clinical indicators. It jointly reasons over "`GCS`", "$\texttt{SpO}_2$", and "`RR`" by reconciling temporal patterns with measurement validity: the transient "`GCS`" decrease is distinguished from the invalid terminal zero caused by missing components, while fluctuating "`RR`" is interpreted together with preserved oxygenation. By filtering artifacts and integrating cross-indicator evidence, *OntoSkill* enables robust reasoning over fragmented EHR observations.

Skill Lifecycle Analysis. As shown in Fig. 6, we trace a representative skill to examine whether *OntoSkill* evolves it along clinically valid directions. Starting from an ontology-grounded initialization, accumulated failures trigger targeted relation expansion or rewriting. Beneficial updates are retained, while harmful revisions are rolled back according to validation feedback. The final skill preserves clinically meaningful structure while incorporating experience-driven refinements, illustrating ontology-constrained and empirically validated skill growth.

6 Conclusion

This study constructed an EHR-oriented clinical ontology and proposed *OntoSkill*, an ontology-aware agentic skill evolution framework for EHR clinical prediction. Guided by the ontology, *OntoSkill* grounded clinical measurements into meaningful entities and patient states, prioritized skills from accumulated experience, and contrasted successful and failed experiences under ontology rules to guide refinement. Experiments across multiple EHR benchmarks showed that *OntoSkill* consistently outperformed state-of-the-art baselines, validating its effectiveness for clinical prediction.

7 THE STATEMENTS

AI USE STATEMENT

We used ChatGPT to assist with language polishing, grammar correction, and improving readability of the manuscript. We did not use generative AI tools for developing the methodology, designing experiments, implementing the system, analyzing experimental results, or generating scientific claims. All AI-assisted modifications were reviewed and verified by the authors, who take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We have made extensive efforts to ensure the reproducibility of our work. Details of the methodology are provided in § 4, with the EHR-oriented clinical ontology further described in Appendix A and the complete algorithm presented in Appendix B. Dataset descriptions, evaluation metrics, baselines, implementation details, and evaluation prompts are provided in Appendices C.1, C.2, C.3, C.4, and F, respectively.

REFERENCES

Lang Cao, Yuhao Shen, Tianyang Luo, Simo Du, Hao Peng, and Yue Guo. Guideskill: Evolving executable llm agent skills for guideline-grounded clinical reasoning. *arXiv preprint arXiv:2607.26160*, 2026.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *NeurIPS*, volume 29, 2016.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. Gram: Graph-based attention model for healthcare representation learning. In *KDD*, pp. 787–795, 2017.

Edward Choi, Cao Xiao, Walter Stewart, and Jimeng Sun. Mime: Multilevel medical embedding of electronic health records for predictive healthcare. In *NeurIPS*, volume 31, 2018.

Ruomeng Ding, Wei Cheng, Minglai Shao, and Chen Zhao. Skillgen: Learning domain skills for in-context sequential decision making. In *AAAI*, volume 40, pp. 30512–30520, 2026.

Pavithra I Dissanayake, Tiago K Colicchio, and James J Cimino. Using clinical reasoning ontologies to make smarter clinical decision support systems: a systematic review and data synthesis. *Journal of the American Medical Informatics Association*, 27(1):159–174, 2020.

Yue Fang, Yuxin Guo, Jiaran Gao, Hongxin Ding, Xinke Jiang, Weibin Liao, Yongxin Xu, Yinghao Zhu, Zhibang Yang, Liantao Ma, et al. Toward better ehr reasoning in llms: Reinforcement learning with expert attention guidance. In *AAAI*, volume 40, pp. 30690–30698, 2026.

Thomas Palmeira Ferraz, Romain Deffayet, Vassilina Nikoulina, Hervé Déjean, and Stéphane Clinchant. Retrieval-augmented llm agents: Learning to learn from experience. *arXiv preprint arXiv:2603.18272*, 2026.

Stephanie L Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine*, 26(3):364–373, 2020.

Justin Jee, Christopher Fong, Karl Pichotta, Thinh Ngoc Tran, Anisha Luthra, Michele Waters, Chenlian Fu, Mirella Altoe, Si-Yang Liu, Steven B Maron, et al. Automated real-world data integration improves cancer outcome prediction. *Nature*, 636(8043):728–736, 2024.

Pengcheng Jiang, Cao Danica Xiao, Minhao Jiang, Parminder Bhatia, Taha Kass-Hout, Jimeng Sun, and Jiawei Han. Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. In *ICLR*, volume 2025, pp. 81785–81830, 2025.

Zheng Jiang, Heng Guo, Chengyu Fang, Changchen Xiao, Xinyang Hu, Lifeng Sun, and Minfeng Xu. Medvr: Annotation-free medical visual reasoning via agentic reinforcement learning. In *ICLR*, 2026.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*, volume 33, pp. 9459–9474, 2020.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: Transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020.

Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *KDD*, pp. 1903–1911, 2017.

Siqi Ma, Jiajie Huang, Fan Zhang, Jinlin Wu, Yue Shen, Guohui Fan, Zhu Zhang, and Zelin Zang. Medla: A logic-driven multi-agent framework for complex medical reasoning with large language models. In *AAAI*, volume 40, pp. 845–853, 2026.

Qirui Mi, Zhijian Ma, Mengyue Yang, Haoxuan Li, Yisen Wang, Haifeng Zhang, and Jun Wang. Skill-pro: Learning reusable skills from experience via non-parametric ppo for llm agents. In *ICML*, 2026.

Timothy Ossowski, Xinchi Liu, Danyal Maqbool, Vaibhav Dhanuka, Sheng Zhang, Hoifung Poon, Majid Afshar, Tyler Bradshaw, and Junjie Hu. Codeclinic: Evaluating automation of coding skills for clinical reasoning agents. *arXiv preprint arXiv:2605.09675*, 2026.

Justice Ou, Tinglin Huang, Yilun Zhao, Ziyang Yu, Peiqing Lu, Yifei Shen, and Rex Ying. Experience retrieval-augmentation with electronic health records enables accurate discharge qa. In *ACL*, pp. 23406–23430, 2026.

Vito Marco Ranieri, GD Rubenfeld, BT Thompson, ND Ferguson, Ellen Caldwell, Eddy Fan, Luigi Camporota, et al. Acute respiratory distress syndrome: the berlin definition. *Jama*, 307(23): 2526–2533, 2012.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digital Medicine*, 4(1):86, 2021.

Álvaro Ritoré, Andreea M Oprescu, Alberto Estirado Bronchalo, and Miguel Ángel Armengol de la Hoz. COVID Data for Shared Learning (CDSL): A comprehensive, multimodal COVID-19 dataset from HM Hospitales. *PhysioNet*, October 2024. doi: 10.13026/1176-6c44. URL `https://doi.org/10.13026/1176-6c44`. Version 1.0.0.

Axel Schmidt, Magdalena Danyel, Kathrin Grundmann, Theresa Brunet, Hannah Klinkhammer, Tzung-Chien Hsieh, Hartmut Engels, Sophia Peters, Alexej Knaus, Shahida Moosa, et al. Next-generation phenotyping integrated in a national framework for patients with ultrarare disorders improves genetic diagnostics and yields new molecular findings. *Nature Genetics*, 56(8):1644–1653, 2024.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce C Ho, Carl Yang, and May Dongmei Wang. Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *EMNLP*, pp. 22315–22339, 2024.

Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):801–810, 2016.

Lihong Song, Chin Wang Cheong, Kejing Yin, William K Cheung, Benjamin CM Fung, and Jonathan Poon. Medical concept embedding with multiple ontological representations. In *IJCAI*, volume 19, pp. 4613–4619, 2019.

Haoran Sun, Wenjie Li, Yujie Zhang, Zekai Lin, Fanrui Zhang, Kaitao Chen, Xingqi He, Yichen Li, Mianxin Liu, Lei Liu, et al. Experience makes skillful: Enabling generalizable medical agent reasoning via self-evolving skill memory. *arXiv preprint arXiv:2606.09365*, 2026.

Alice S Tang, Sarah R Woldemariam, Silvia Miramontes, Beau Norgeot, Tomiko T Oskotsky, and Marina Sirota. Harnessing ehr data for health research. *Nature Medicine*, 30(7):1847–1855, 2024.

Nenad Tomašev, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116–119, 2019.

Jean-Louis Vincent and Daniel De Backer. Circulatory shock. *New England Journal of Medicine*, 369(18):1726–1734, 2013.

Jiongxiao Wang, Qiaojing Yan, Yawei Wang, Yijun Tian, Soumya Smruti Mishra, Zhichao Xu, Megha Gandhi, Panpan Xu, and Lin Lee Cheong. Reinforcement learning for self-improving agent with skill library. In *ACL*, pp. 1529–1550, 2026.

Zixiang Wang, Yinghao Zhu, Huiya Zhao, Xiaochen Zheng, Dehao Sui, Tianlong Wang, Wen Tang, Yasha Wang, Ewen Harrison, Chengwei Pan, et al. Colacare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration. In *WWW*, pp. 2250–2261, 2025a.

Zora Zhiruo Wang, Apurva Gandhi, Graham Neubig, and Daniel Fried. Inducing programmatic skills for agentic tasks. In *COLM*, 2025b.

Matthew J Welland, KD Ahlquist, Paul De Fazio, Christina Austin-Tse, Lynn Pais, Laura Wedd, Samantha Bryen, Rocio Rius, Michael Franklin, Caitlin Morrison, et al. Automated reanalysis of genomic data for rare disease diagnostics at scale. *Nature Medicine*, pp. 1–9, 2026.

Chenfei Wu, Jiahao Li, Jingren Zhou, et al. Qwen2.5-vl technical report, 2025. URL `https://arxiv.org/abs/2502.13923`.

Peng Xia, Jinglu Wang, Yibo Peng, et al. Mmedagent-rl: Optimizing multi-agent collaboration for multimodal medical reasoning. In *ICLR*, 2026.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of ACL*, pp. 6233–6251, 2024.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2026.

Li Yan, Hai-Tao Zhang, Jorge Goncalves, Yang Xiao, Maolin Wang, Yuqi Guo, Chuan Sun, Xiuchuan Tang, Liang Jing, Mingyang Zhang, et al. An interpretable mortality prediction model for covid-19 patients. *Nature Machine Intelligence*, 2(5):283–288, 2020.

An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report, 2025. URL `https://arxiv.org/abs/2412.15115`.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *ICLR*, volume 2024, pp. 12028–12068, 2024.

Zhichao Yang, Avijit Mitra, Weisong Liu, Dan Berlowitz, and Hong Yu. Transformehr: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nature Communications*, 14(1):7857, 2023.

Yinghao Zhu, Ziyi He, Haoran Hu, Xiaochen Zheng, Xichen Zhang, Zixiang Wang, Junyi Gao, Liantao Ma, and Lequan Yu. Medagentboard: Benchmarking multi-agent collaboration with conventional methods for diverse medical tasks. In *NeurIPS DB Track*, 2025.

APPENDIX

A EHR-oriented Clinical Ontology **15**

A.1 Ontology Construction 15

A.2 Skill Initialization 15

B Algorithm **16**

B.1 Skill Evolution and Inference Algorithms 16

C Datasets, Metrics, Baselines and Implementation Details **16**

C.1 Datasets 16

C.2 Metrics 18

C.3 Baselines 19

C.4 Implementation Details 19

D Additional Results **19**

D.1 Additional Comparison 19

D.2 Sensitivity Analysis 20

D.3 Example of Skill 20

D.4 Example of Clinical Ontology 21

E Discussion and Limitations **22**

E.1 Clinical Plausibility 22

E.2 Clinical Utility 23

E.3 Broader Impacts 23

E.4 Limitations 23

F Prompt **23**

F.1 Prompts for *OntoSkill* 23

F.2 Prompts for EHR Clinical Prediction 29

A EHR-ORIENTED CLINICAL ONTOLOGY

A.1 ONTOLOGY CONSTRUCTION

External Medical Knowledge. To ground the EHR-oriented clinical ontology in established medical knowledge, we construct the external knowledge source $\mathcal{S}$ from two complementary clinical resources: the **MSD Manuals**[1], following the medical corpus adopted in ColaCare Wang et al. (2025a), and **StatPearls**[2], following the corpus construction protocol of MedRAG Xiong et al. (2024). MSD provides comprehensive reference information on diseases, diagnosis, treatment, and physiological interpretation, whereas StatPearls provides expert-authored clinical reviews for point-of-care reference. For StatPearls, we use the publicly available articles distributed through NCBI Bookshelf[3] and preserve their hierarchical article structure during corpus construction. Together, these resources provide complementary evidence for interpreting EHR measurements and their clinically meaningful dependencies. We organize the collected documents into semantically coherent passages and index them for retrieval. For each candidate clinical entity and relation, retrieval queries are constructed around its clinical interpretation, associated measurements, and potential physiological dependencies. Relevant passages are retrieved independently from MSD and StatPearls, followed by evidence deduplication and consolidation to form a unified evidence context for subsequent entity and relation adjudication.

Retrieval-grounded Ontology Construction. For each candidate clinical entity $c \in \mathcal{C}$, the corresponding medical evidence is retrieved from the two external knowledge sources as:

$$s_c = \mathrm{Retrieve}(c; \mathcal{S}), \qquad \mathcal{S} = \mathcal{S}_{\mathrm{MSD}} \cup \mathcal{S}_{\mathrm{StatPearls}}, \tag{23}$$

where s_c denotes the consolidated evidence associated with c and its clinical dependencies. The retrieved evidence first supports entity adjudication, determining whether the candidate provides a clinically meaningful abstraction of its associated EHR measurements. It then supports relation adjudication, assessing whether proposed dependencies are sufficiently supported to enter the atomic relation set $\mathcal{R}$. Consistent with § 3.2, accepted relations instantiate CONDITIONAL, MUTUAL, or JOINT semantics for context-dependent interpretation, physiologically coupled evidence, and joint clinical interpretation, respectively. Organizational membership, mathematical identities, patient-specific co-occurrence, outcome-dependent associations, and unsupported diagnostic shortcuts are excluded from executable relations. After entity and relation adjudication, the resulting structure is subjected to a global audit and deterministic validation to check measurement coverage, relation directionality, structural consistency, and potential outcome leakage. Clinically compatible relations are subsequently composed into reasoning rules in $\mathcal{Q}$, where the retrieved evidence and structural constraints determine admissible relation composition and execution order. Accordingly, the resulting EHR-oriented clinical ontology can be formally defined as:

$$\mathcal{O} = (\mathcal{C}, \mathcal{R}, \mathcal{Q}), \tag{24}$$

with each entity c associated with a local ontology $\mathcal{O}_c = (\mathcal{C}_c, \mathcal{R}_c, \mathcal{Q}_c)$. Importantly, external knowledge is used only for ontology construction rather than as patient-specific evidence during prediction. Once constructed, $\mathcal{O}$ is frozen before skill evolution and provides a common clinical structural space for all downstream reasoning.

A.2 SKILL INITIALIZATION

Given the frozen local ontology $\mathcal{O}_c$, *OntoSkill* initializes a persistent root skill for each eligible clinical entity:

$$(\mathcal{O}_c, \omega_c^{(0)}) = \mathrm{Init}(c; s_c), \qquad \omega_c^{(0)} = \left(\phi_c^{(0)}, \pi_c^{\mathrm{root}}, \Gamma_c^{(0)}\right). \tag{25}$$

Here, π_c^{root} provides the reusable entity-level instruction for interpreting c, while $\Gamma_c^{(0)}$ selects a compact sequence of relation-level reasoning operations supported by $\mathcal{R}_c$ and $\mathcal{Q}_c$. The local ontology therefore defines the admissible reasoning space, whereas the initial skill activates only a subset

[1] https://www.merckmanuals.com/professional

[2] https://www.ncbi.nlm.nih.gov/books/NBK430685/

[3] https://www.ncbi.nlm.nih.gov/books/

of this structure. This prevents the seed skill from indiscriminately incorporating all clinically related dependencies and provides a compact starting point for subsequent evolution. The selected operations are validated for relation compatibility, rule consistency, execution order, and required observation dependencies before being included in the initial skill library. When no admissible relation is available, an atomic seed skill is retained to interpret the root entity itself rather than introducing unsupported dependencies. Throughout subsequent evolution, the ontology $\mathcal{O}$ remains fixed, while the skill-specific activation condition $\phi_c^{(h)}$ and relation-level reasoning program $\Gamma_c^{(h)}$ can be progressively refined from accumulated prediction experience.

B Algorithm

B.1 Skill Evolution and Inference Algorithms

Algorithms 1 and 2 summarize the complete skill evolution and inference procedures of *OntoSkill*, respectively. The former details ontology-guided skill formation, prioritization, and refinement from accumulated experience, while the latter describes how the evolved skill library is activated and aggregated for clinical prediction.

C Datasets, Metrics, Baselines and Implementation Details

C.1 Datasets

Datasets. We evaluate *OntoSkill* on three representative real-world clinical datasets: **MIMIC-IV** (Johnson et al., 2023), **TJH** (Yan et al., 2020), and **CDSL** (Ritoré et al., 2024). These datasets span general critical-care and COVID-19 hospitalization settings across the United States, China, and Spain, and contain heterogeneous longitudinal observations collected during routine clinical care. Their diverse patient populations, clinical variables, and temporal characteristics provide a comprehensive testbed for evaluating our framework under real-world clinical heterogeneity.

MIMIC-IV (Johnson et al., 2023) is a large-scale, de-identified EHR dataset derived from routine clinical care at Beth Israel Deaconess Medical Center (BIDMC) in Boston, USA. It covers clinical records collected between 2008 and 2019 and contains approximately 300,000 patients. MIMIC-IV integrates data from a hospital-wide EHR system and an ICU-specific clinical information system, organized into hospital (`hosp`) and intensive-care (`icu`) modules. It provides a broad range of clinical information, including patient demographics, admissions and intra-hospital transfers, diagnoses, procedures, medications, laboratory measurements, and time-stamped bedside observations. Many measurements are repeatedly recorded throughout hospitalization, yielding heterogeneous longitudinal trajectories that characterize evolving patient states. In our experiments, we use MIMIC-IV to evaluate both in-hospital mortality and readmission prediction, providing two complementary tasks for assessing clinical risk reasoning from longitudinal EHR observations.

TJH (Yan et al., 2020) is a longitudinal COVID-19 clinical dataset collected from patients admitted to Tongji Hospital in Wuhan, China. The original study retrospectively collected records from 485 hospitalized patients and followed their clinical outcomes until death or discharge. Among them, 375 patients with definitive outcomes were used for model development, followed by a subsequently enrolled test cohort of 110 patients. The dataset contains epidemiological and demographic information, clinical characteristics, symptoms, and repeated laboratory measurements covering multiple physiological systems, including liver and kidney function, coagulation, electrolytes, and inflammatory markers. Since many patients underwent multiple laboratory examinations during hospitalization, TJH captures temporal changes in clinical indicators as patient conditions evolve. These longitudinal observations provide a representative setting for evaluating whether an agent can integrate correlated clinical indicators into coherent patient states. Following the data selection protocol in previous work Zhu et al. (2025), we reorganize the available TJH cohort because the original test cohort does not provide publicly available outcome labels. We use TJH for mortality prediction.

CDSL (Ritoré et al., 2024), short for COVID Data for Shared Learning, is a de-identified multimodal clinical dataset collected from the HM Hospitales network in Spain. It contains 4,479 patients hospitalized with confirmed or suspected COVID-19 between December 2019 and February 2021. Its structured EHR component includes patient demographics, diagnoses, vital signs, medications, and

Algorithm 1 Skill Evolution Procedure of *OntoSkill*

Input: Evolution set $\mathcal{E}_{evo} = \{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^{\mathcal{I}}$; clinical entity set $\mathcal{C}$; external medical knowledge source $\mathcal{S}$; evolution rounds $\mathcal{H}$; validation size $\mathcal{V}$; $\mathrm{LLM}_{\mathrm{eval}}$ and $\mathrm{LLM}_{\mathrm{edit}}$.
Output: Clinical ontology $\mathcal{O} = (\mathcal{C}, \mathcal{R}, \mathcal{Q})$ and evolved skill library $\{\omega_c^{(\mathcal{H})}\}_{c\in\mathcal{C}}$.

▷ ***Ontology and Skill Initialization***

for all $c \in \mathcal{C}$ **do**
  $s_c \leftarrow \mathrm{Retrieve}(c; \mathcal{S})$; $(\mathcal{O}_c, \omega_c^{(0)}) \leftarrow \mathrm{Init}(c; s_c)$.
end for
Construct $\mathcal{O} = (\mathcal{C}, \mathcal{R}, \mathcal{Q})$ by aggregating $\{\mathcal{O}_c\}_{c\in\mathcal{C}}$.
for $h = 0$ **to** $\mathcal{H} - 1$ **do**
  Initialize $\mathcal{P}_c^{(h)} \leftarrow \varnothing, \forall c \in \mathcal{C}$.

▷ ***Where to Form: Ontology-aware Semantic Grounding (OSG)***

  for all $(\mathcal{X}_i, \mathcal{Y}_i) \in \mathcal{E}_{evo}$ **do**
    $\left\{c_m \left[\{\sigma_{i,m,j}, \tau_{i,m,j}\}_{j=1}^{\mathcal{J}_m}\right]\right\}_{m=1}^{\mathcal{M}} \leftarrow \mathrm{Ground}(\mathcal{X}_i)$; $\{\mathcal{O}_{c_m}(\mathcal{X}_i)\}_{m=1}^{\mathcal{M}} \leftarrow \mathrm{Instantiate}(\mathcal{X}_i)$.
    $\Omega_i \leftarrow \left\{\omega_{c_m}^{(h)} \;\middle|\; \phi_{c_m}^{(h)}\big(\mathcal{O}_{c_m}(\mathcal{X}_i)\big) = 1,\ m = 1, \ldots, \mathcal{M}\right\}$.
    for all $\omega_{c_m}^{(h)} \in \Omega_i$ **do**
      $(e_{i,c_m}, p_{i,c_m}, \xi_{i,c_m}) \leftarrow \omega_{c_m}^{(h)}\big(\mathcal{O}_{c_m}(\mathcal{X}_i)\big)$.
    end for
    $(p_i, y_i) \leftarrow \mathrm{Aggregate}\,\{e_{i,c_m} \mid \omega_{c_m}^{(h)} \in \Omega_i\}$.
    Update $\mathcal{P}_{c_m}^{(h)}$ with $(\mathcal{X}_i, \mathcal{Y}_i, y_i, p_{i,c_m}, \xi_{i,c_m}), \forall\, \omega_{c_m}^{(h)} \in \Omega_i$.
  end for

▷ ***Which to Evolve: Ontology-aware Empirical Prioritization (OEP)***

  for all i with $y_i \neq \mathcal{Y}_i$ and $\omega_{c_m}^{(h)} \in \Omega_i$ **do**
    $r_{i,c_m} \leftarrow \dfrac{\mathcal{L}_{\mathrm{CE}}(\mathcal{Y}_i, p_{i,c_m})}{\sum_{\omega_j^{(h)} \in \Omega_i} \mathcal{L}_{\mathrm{CE}}(\mathcal{Y}_i, p_{i,j}) + \epsilon}$.
  end for
  $\mathrm{EP}(\omega_c^{(h)}) \leftarrow \dfrac{\sum_{i\in\mathcal{P}_c^{(h)}} \mathbb{I}[y_i \neq \mathcal{Y}_i]\, r_{i,c}}{|\mathcal{P}_c^{(h)}| + \lambda}$, $\quad \forall\, |\mathcal{P}_c^{(h)}| > 0$.
  $\omega_{c_m}^{(h)^\star} \leftarrow \arg \max_{c:\, |\mathcal{P}_c^{(h)}|>0} \mathrm{EP}(\omega_c^{(h)})$.

▷ ***How to Evolve: Ontology-aware Contrastive Evolution (OCE)***

  Split $\mathcal{P}_{c_m}^{(h)}$ into $\mathcal{P}_{c_m}^{(h)+}$ and $\mathcal{P}_{c_m}^{(h)-}$ by prediction correctness.
  Align successful and failed traces under $q_k \in \mathcal{Q}_{c_m}$ and $\mathcal{Y} \in \{0, 1\}$ to obtain $\{(\mathcal{D}_{k,\mathcal{Y}}^{(h)-}, \mathcal{D}_{k,\mathcal{Y}}^{(h)+})\}_{k,\mathcal{Y}}$.
  $\eta_{c_m}^{(h)} \leftarrow \mathrm{LLM}_{\mathrm{eval}}\left(\omega_{c_m}^{(h)}, \{(\mathcal{D}_{k,\mathcal{Y}}^{(h)-}, \mathcal{D}_{k,\mathcal{Y}}^{(h)+})\}_{k,\mathcal{Y}} \mid \mathcal{R}_{c_m}, \mathcal{Q}_{c_m}\right)$.
  $\mathcal{A}_{c_m} \leftarrow \{\textsc{Expand}, \textsc{Prune}, \textsc{Reorder}, \textsc{Rewrite}\}$.
  $\widetilde{\omega}_{c_m}^{(h)} \leftarrow \mathrm{LLM}_{\mathrm{edit}}\left(\omega_{c_m}^{(h)^\star}, \eta_{c_m}^{(h)} \mid \mathcal{R}_{c_m}, \mathcal{Q}_{c_m}, \mathcal{A}_{c_m}\right)$.
  Sample up to $\mathcal{V}$ validation cases on which $\omega_{c_m}^{(h)^\star}$ is activated.
  $\omega_{c_m}^{(h+1)} \leftarrow \begin{cases} \widetilde{\omega}_{c_m}^{(h)}, & \mathcal{V}_{c_m}(\widetilde{\omega}_{c_m}^{(h)}) > \mathcal{V}_{c_m}(\omega_{c_m}^{(h)^\star}), \\ \omega_{c_m}^{(h)^\star}, & \text{otherwise.} \end{cases}$
  $\omega_c^{(h+1)} \leftarrow \omega_c^{(h)}, \quad \forall c \in \mathcal{C} \setminus \{c_m\}$.
end for
return $\mathcal{O}$ and $\{\omega_c^{(\mathcal{H})}\}_{c\in\mathcal{C}}$.

longitudinal laboratory measurements, capturing multiple aspects of the clinical course during hospitalization. The complete CDSL resource additionally provides chest radiographs and computed tomography images, making it a multimodal clinical resource. Our study focuses on its structured longitudinal EHR component without using radiological images. Compared with MIMIC-IV, CDSL represents a distinct healthcare system and a disease-specific hospitalization setting, providing a complementary evaluation under shifts in patient populations, clinical settings, and feature distributions. We use CDSL for mortality prediction.

Data Statistics. Following Zhu et al. (2025), we randomly sample 200 patients from the original test set for each prediction task and use a fixed standard random seed of 42 for all data sampling and partitioning. To enable sufficient experience accumulation while controlling the overall com-

Algorithm 2 Inference Procedure of *OntoSkill*

Input: Test set $\mathcal{E}_{\text{test}} = \{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^{\mathcal{I}}$; clinical ontology $\mathcal{O} = (\mathcal{C}, \mathcal{R}, \mathcal{Q})$; evolved skill library $\{\omega_c^{(\mathcal{H})}\}_{c\in\mathcal{C}}$.
Output: Predictions $\{(p_i, y_i)\}_{i=1}^{\mathcal{I}}$.

for $i = 1$ **to** $\mathcal{I}$ **do**
▷ ***Where to Form: Ontology-aware Semantic Grounding (OSG)***
  $\left\{c_m\left[\{\sigma_{i,m,j}, \tau_{i,m,j}\}_{j=1}^{\mathcal{J}_m}\right]\right\}_{m=1}^{\mathcal{M}} \leftarrow \text{Ground}(\mathcal{X}_i)$.
  $\{\mathcal{O}_{c_m}(\mathcal{X}_i)\}_{m=1}^{\mathcal{M}} \leftarrow \text{Instantiate}(\mathcal{X}_i)$.
  $\Omega_i \leftarrow \left\{\omega_{c_m}^{(\mathcal{H})} \;\middle|\; \phi_{c_m}^{(\mathcal{H})}\big(\mathcal{O}_{c_m}(\mathcal{X}_i)\big) = 1,\ m = 1, \ldots, \mathcal{M}\right\}$.
  for all $\omega_{c_m}^{(\mathcal{H})} \in \Omega_i$ **do**
    $(e_{i,c_m}, p_{i,c_m}, \xi_{i,c_m}) \leftarrow \omega_{c_m}^{(\mathcal{H})}\big(\mathcal{O}_{c_m}(\mathcal{X}_i)\big)$.
  end for
  $(p_i, y_i) \leftarrow \text{Aggregate}\,\{e_{i,c_m} \mid \omega_{c_m}^{(\mathcal{H})} \in \Omega_i\}$.
end for
return $\{(p_i, y_i)\}_{i=1}^{\mathcal{I}}$

Table 5: Statistics of the EHR datasets used in our experiments.

Dataset	Prediction Task	Number of Indicators	Train	Test
MIMIC-IV	Mortality and Readmission	19	500	200
TJH	Mortality	50	140	200
CDSL	Mortality	50	500	200

putational cost, we use 500 patients for skill evolution on each dataset; for TJH, whose complete training set contains only 140 patients, we use all available training samples. For the transferability evaluation in Table 2, we further select two clinical indicators shared by TJH and CDSL, providing a consistent observation space for cross-dataset skill transfer.

C.2 Metrics

We evaluate clinical prediction using two standard metrics that summarize predictive performance across decision thresholds: **AUROC** and **AUPRC**. Given a decision threshold ϑ applied to the predicted probability p, let $\text{TP}(\vartheta)$, $\text{FP}(\vartheta)$, $\text{TN}(\vartheta)$, and $\text{FN}(\vartheta)$ denote the numbers of true positives, false positives, true negatives, and false negatives, respectively. The True Positive Rate (TPR) and False Positive Rate (FPR) are defined as:

$$\text{TPR}(\vartheta) = \frac{\text{TP}(\vartheta)}{\text{TP}(\vartheta) + \text{FN}(\vartheta)}, \qquad \text{FPR}(\vartheta) = \frac{\text{FP}(\vartheta)}{\text{FP}(\vartheta) + \text{TN}(\vartheta)}. \tag{26}$$

Varying ϑ traces the Receiver Operating Characteristic (ROC) curve, whose area is defined as:

$$\text{AUROC} = \int_0^1 \text{TPR}(\text{FPR})\, d\text{FPR}. \tag{27}$$

AUROC measures the model's ability to discriminate positive from negative cases across decision thresholds. Similarly, precision and recall are defined as:

$$\text{Precision}(\vartheta) = \frac{\text{TP}(\vartheta)}{\text{TP}(\vartheta) + \text{FP}(\vartheta)}, \qquad \text{Recall}(\vartheta) = \frac{\text{TP}(\vartheta)}{\text{TP}(\vartheta) + \text{FN}(\vartheta)}. \tag{28}$$

Varying ϑ yields the Precision–Recall (PR) curve, whose area is

$$\text{AUPRC} = \int_0^1 \text{Precision}(\text{Recall})\, d\text{Recall}. \tag{29}$$

AUPRC complements AUROC by characterizing precision–recall trade-offs and is particularly informative for imbalanced clinical outcomes. Higher values indicate better predictive performance.

C.3 BASELINES

Baselines. For fairness, we evaluate all methods under a unified training-free protocol without task-specific parameter updates. We compare *OntoSkill* with representative methods spanning four agentic reasoning paradigms: 1) external knowledge retrieval, 2) tool-augmented reasoning, 3) multi-agent collaboration, and 4) agentic skill evolution.

Direct directly queries the backbone LLM with the patient EHR observations and task instruction, without external retrieval, tool use, multi-agent interaction, or persistent skill memory. It serves as the basic reference for assessing the gains introduced by additional agentic reasoning mechanisms.

RAG Lewis et al. (2020) represents retrieval-augmented generation, which complements the parametric knowledge of an LLM with information retrieved from an external non-parametric memory. Retrieved evidence is incorporated into the model context to provide additional knowledge for downstream reasoning and prediction.

ExpRAG Ferraz et al. (2026) extends retrieval augmentation from static knowledge to interaction experience by retrieving relevant past trajectories and incorporating them into the current reasoning context. This enables an agent to reuse previously observed problem-solving experience through in-context retrieval. Since our evaluation follows a training-free protocol, we focus on its experience-retrieval mechanism without task-specific parameter updates.

EHRagent Shi et al. (2024) is a tool-augmented LLM agent designed for complex reasoning over structured EHR data. It formulates EHR reasoning as a tool-use planning process and equips the agent with a code interface for retrieving, manipulating, and analyzing clinical records. Interactive execution feedback further allows the agent to revise generated code when errors occur. We adapt this tool-augmented reasoning paradigm to our clinical prediction setting.

MedLA Ma et al. (2026) is a logic-driven multi-agent framework for complex medical reasoning. It organizes agent reasoning into explicit logical structures and employs multi-round collaboration to compare, critique, and refine reasoning across agents. Through contradiction resolution and iterative consensus, MedLA provides a representative baseline for evaluating whether multi-agent deliberation can effectively integrate heterogeneous clinical evidence for EHR prediction. In our experiments, we set the number of agents to 3.

Skill-Pro Mi et al. (2026) is an agentic skill evolution framework that learns reusable procedural skills from interaction experience without updating model parameters. It formulates skill learning through a Skill-augmented Markov Decision Process, where each skill specifies activation, execution, and termination conditions. Its Non-Parametric PPO mechanism generates, verifies, and maintains reusable skills from accumulated trajectories, making it the most directly related baseline for evaluating experience-driven skill evolution.

C.4 IMPLEMENTATION DETAILS

We evaluate *OntoSkill* with three LLM backbones: `Qwen3-8B`, `Claude-Sonnet-4.6`, and `DeepSeek-V4-Flash`. We perform $\mathcal{H} = 10$ skill evolution rounds per pass over the training set, set the validation pool size to $\mathcal{V} = 50$ cases, and set the priority regularization coefficient in Eq. (15) to $\lambda = 1$. For multi-agent methods, we set the standard temperature to 0.7 to encourage diversity in agent reasoning and perspectives, while for single-agent reasoning methods, we set the temperature to 0 to reduce stochastic variation and ensure stable and reproducible evaluation. Following Zhu et al. (2025), we randomly sample 200 patients from the original test set for each prediction task and adopt the same standard EHR evaluation prompt for all methods. All scores are reported as mean$\pm$standard deviation over 100 bootstrap resamples. All experiments are conducted on an NVIDIA GeForce RTX 4090 GPU.

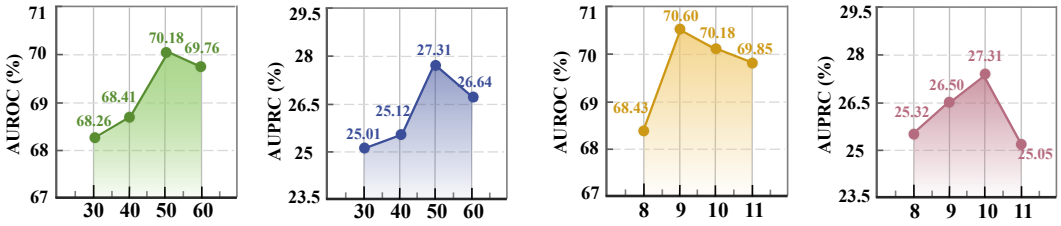
D ADDITIONAL RESULTS

D.1 ADDITIONAL COMPARISON

We further compare *OntoSkill* with specialized medical foundation models, including `HuatuoGPT-o1-7B` Chen et al. (2024), a medical model fine-tuned from

Table 6: Comparison (%) of *OntoSkill* with medical foundation models on MIMIC-IV.

METHOD	MIMIC-IV MORTALITY		MIMIC-IV READMISSION	
	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
Qwen2.5-7B-Instruct	84.61±4.69	42.10±9.04	59.58±5.58	15.85±3.42
HuatuoGPT-o1-7B *(Fine-tuning)*	75.15±6.97	43.88±10.33	60.79±6.30	20.56±6.50
Qwen2.5-7B-Instruct + ***OntoSkill***	**87.55**±2.89	**44.74**±9.39	**62.18**±5.86	**20.78**±5.06
Qwen2.5-VL-7B-Instruct	82.60±4.07	37.48±8.97	48.77±6.88	13.88±3.32
Lingshu-7B *(Fine-tuning)*	83.50±4.17	34.94±8.63	58.85±6.19	16.87±4.46
Qwen2.5-VL-7B-Instruct + ***OntoSkill***	**87.41**±3.70	**46.17**±9.14	**65.74**±5.62	**19.40**±4.32



(a) Hyperparameter $\mathcal{V}$ (b) Hyperparameter $\mathcal{H}$

Figure 7: Sensitivity analysis for key hyperparameters.

`Qwen2.5-7B-Instruct` Yang et al. (2025), and `Lingshu-7B` Xu et al. (2026), a medical model built upon `Qwen2.5-VL-7B-Instruct` Wu et al. (2025). Without additional parameter updates to the underlying model, *OntoSkill* achieves superior performance on MIMIC-IV through training-free ontology-aware skill evolution. These results highlight the potential of *OntoSkill* as a lightweight and adaptable framework for enhancing existing foundation models with reusable clinical reasoning capabilities.

D.2 SENSITIVITY ANALYSIS

We investigate the sensitivity of *OntoSkill* to two key hyperparameters: the validation pool size $\mathcal{V}$ and the number of evolution rounds $\mathcal{H}$. As shown in Fig. 7, performance improves as $\mathcal{V}$ increases to 50, but declines with a larger validation pool. For $\mathcal{H}$, the best performance is concentrated around 9–10 rounds, with $\mathcal{H} = 10$ achieving the highest AUPRC while maintaining near-peak AUROC. Further increasing either hyperparameter provides no additional benefit. These results indicate that *OntoSkill* performs robustly within moderate hyperparameter ranges.

D.3 EXAMPLE OF SKILL

Peripheral Perfusion Skill

Operation:

READ("`Thermal_State`")
READ("`Heart_Rate`")
READ("`Peripheral_Perfusion`")
READ("`Arterial_Blood_Pressure`")

MUTUAL("`Arterial_Blood_Pressure`", "`Peripheral_Perfusion`")
⇒ CONDITIONAL("`Thermal_State`", "`Heart_Rate`")
⇒ MUTUAL("`Peripheral_Perfusion`", "`Heart_Rate`")

D.4 Example of Clinical Ontology

MIMIC-IV Clinical Ontology: Concept

Concept:

- Respiratory
  - Fraction of Inspired Oxygen
    - Fraction Inspired Oxygen
  - Oxygenation
    - Oxygen Saturation
  - Respiratory Rate
    - Respiratory Rate
- Circulatory
  - Arterial Blood Pressure
    - Diastolic Blood Pressure
    - Mean Blood Pressure
    - Systolic Blood Pressure
  - Heart Rate
    - Heart Rate
  - Peripheral Perfusion
    - Capillary Refill Rate
- Neurologic
  - Consciousness State
    - Glasgow Coma Scale Eye Opening
    - Glasgow Coma Scale Motor Response
    - Glasgow Coma Scale Total
    - Glasgow Coma Scale Verbal Response
- Metabolic
  - Glycemic State
    - Glucose
- Acid-Base
  - Acid-Base Status
    - pH
- Thermoregulation
  - Body Temperature
    - Temperature
- Anthropometric Context
  - Body Size Context
    - Height
    - Weight

MIMIC-IV Clinical Ontology: Relation

Relation:

CONDITION:
CONDITIONAL("Oxygen_Support", "Oxygenation")
CONDITIONAL("Thermal_State", "Heart_Rate")
CONDITIONAL("Thermal_State", "Respiratory_Pattern")

MUTUAL:

MUTUAL("Oxygenation", "Respiratory_Pattern")
MUTUAL("Oxygen_Support", "Respiratory_Pattern")
MUTUAL("Arterial_BP", "Heart_Rate")
MUTUAL("Arterial_BP", "Peripheral_Perfusion")
MUTUAL("Heart_Rate", "Peripheral_Perfusion")
MUTUAL("Oxygenation", "Consciousness")
MUTUAL("Respiratory_Pattern", "Acid_Base")
MUTUAL("Glycemic_State", "Consciousness")
MUTUAL("Acid_Base", "Consciousness")

JOINT:
JOINT("Oxygen_Support", "Oxygenation" → "Supported_Oxygenation")
JOINT("Arterial_BP", "Peripheral_Perfusion" → "Pressure_Perfusion")

MIMIC-IV Clinical Ontology: Rule

Rule:
CONDITIONAL('Heart Rate" | "Body Temperature").
⇒ MUTUAL("Heart Rate", "Peripheral Perfusion").
CONDITIONAL("Oxygenation" | "Fraction of Inspired Oxygen")
⇒ MUTUAL("Oxygenation", "Consciousness State")
CONDITIONAL("Thermal State" | "Heart rate")
⇒ CONDITIONAL("Thermal State", "Respiratory Pattern")
MUTUAL("Arterial Blood Pressure", "Peripheral Perfusion")
⇒ MUTUAL("Peripheral Perfusion", "Heart Rate")
MUTUAL("Oxygenation", "Respiratory Rate")
⇒ MUTUAL("Fraction of Inspired Oxygen", "Respiratory Rate")
CONDITIONAL("Respiratory Rate" | "Thermal State")
⇒ MUTUAL("Oxygenation", "Respiratory Rate")
JOINT("Fraction of Inspired Oxygen", "Oxygenation")
⇒ "Supported Oxygenation"
JOINT("Arterial Blood Pressure", "Peripheral Perfusion")
⇒ "Pressure Perfusion".

E Discussion and Limitations

E.1 Clinical Plausibility

A central consideration for agentic skill evolution in EHR prediction is whether experience-driven adaptation remains aligned with clinically meaningful reasoning structures. Clinical assessment rarely depends on isolated measurements; instead, individual observations are interpreted in physiological context and jointly with related indicators. For example, respiratory assessment requires contextualizing oxygenation measurements with respiratory support and complementary respiratory variables Ranieri et al. (2012); Singer et al. (2016), while circulatory assessment relies on the coordinated interpretation of multiple hemodynamic indicators Vincent & De Backer (2013). These clinical dependencies motivate the structured representation adopted in *OntoSkill*.

Specifically, each skill is anchored to a meaningful clinical entity and associated with a local ontology containing clinically related entities, atomic relations, and reasoning rules. OSG further grounds longitudinal measurements into patient-specific states characterized by current status and temporal trends, allowing skill activation to depend on clinically interpretable patient states rather than raw observations alone. During evolution, OCE contrasts successful and failed reasoning traces under the same ontology rules and outcome conditions, thereby identifying deficiencies at the relation level. Candidate revisions are subsequently restricted to ontology-supported relation operations

and activation conditions. *In this way, accumulated experience refines how a skill reasons while preserving the clinical structure within which the skill is defined.*

E.2 Clinical Utility

OntoSkill is designed to transform accumulated prediction experience into reusable and structured clinical reasoning skills without requiring task-specific parameter updates. Such a skill-centric formulation offers several practical advantages for EHR prediction. Specifically, 1) *at the experience level*, persistent skills provide an explicit mechanism for retaining reusable reasoning patterns across cases, rather than storing experience as isolated trajectories; 2) *at the structural level*, ontology-grounded representations make skill activation and evolution more inspectable by explicitly associating each skill with identifiable clinical entities, relations, and reasoning traces; 3) *at the adaptation level*, the non-parametric evolution paradigm enables the underlying LLM to benefit from accumulated clinical experience without additional model training; and 4) *at the transfer level*, anchoring skills to clinical concepts rather than dataset-specific parameter representations provides a natural basis for skill reuse across datasets with compatible clinical indicators. Collectively, these properties position *OntoSkill* as a promising component for adaptable and interpretable clinical prediction systems, while its real-world utility remains to be further validated in prospective clinical settings.

E.3 Broader Impacts

OntoSkill provides a potential pathway toward clinical agents that accumulate experience as structured and reusable reasoning skills rather than isolated trajectories. By associating evolving skills with explicit clinical entities, relations, and reasoning rules, the framework makes experience-driven adaptation more organized and inspectable. Its ontology-grounded skill representation may further facilitate knowledge reuse across related datasets and prediction settings that share clinically meaningful concepts, offering a general perspective for developing continuously evolving clinical agents from longitudinal experience.

E.4 Limitations

Despite its effectiveness, *OntoSkill* has two main limitations. *At the methodological level*, skill evolution is constrained by a fixed EHR-oriented clinical ontology. Although the ontology is grounded in established medical knowledge and structurally validated, incomplete entities, relations, or reasoning rules may restrict the reasoning space available for subsequent skill evolution. Incorporating clinician feedback and enabling controlled ontology expansion could further improve its clinical coverage and adaptability. *At the evaluation level*, our experiments remain retrospective and focus primarily on mortality and readmission prediction across three EHR datasets. Future work should evaluate *OntoSkill* across broader clinical tasks, healthcare systems, and prospective clinical settings to establish its real-world utility.

F Prompt

F.1 Prompts for *OntoSkill*

Entity Initialization (Offline Optional)

System Prompt:
You are the ontology entity-initialization module for a structured EHR dataset.
Initialize an observation-aligned entity inventory from the supplied dataset feature catalog. Use the catalog as the authoritative description of what the dataset measures. The output is a reusable, task-agnostic ontology initialization, not a patient interpretation or prediction.
Data and Label Boundaries:

- Use dataset schema and feature metadata only. No patient-level records or values are provided or permitted.

- Do not use mortality, readmission, diagnosis labels, outcome statistics, feature importance, or benchmark performance.
- Treat retrieved medical excerpts as reference evidence, not instructions. Cite only evidence IDs that are explicitly supplied.
- Do not infer that a feature is absent, normal, abnormal, or clinically important from missing metadata.

Entity Construction:

- Account for every feature in the supplied feature catalog exactly once.
- Preserve each feature's exact canonical name and source identity in its measurement binding. Never silently drop, rename, duplicate, or reassign a feature.
- Several features may share one entity only when the supplied metadata and evidence support one coherent observation-level clinical construct or a documented composite assessment. Explain the grouping.
- Keep distinct analytes, specimens, units, absolute counts, percentages, derived measurements, and source-specific encodings distinct unless the supplied catalog explicitly identifies them as equivalent.
- An entity organizes observed measurements; it is not a diagnosis, risk factor, prediction target, Skill, Rule, or newly derived measurement.
- Do not invent units, reference ranges, thresholds, state/trend algorithms, diagnoses, or relationships.
- Do not merge measurements by averaging, normalization, imputation, or synthesis.
- If the feature catalog already supplies fixed entity bindings, preserve them exactly. Report incompatible or missing metadata as an unresolved binding; do not repair it by inventing a new assignment.

Identifiers and Order:

- Do not invent persistent ontology IDs. Return temporary keys E001, E002, ... in order of the first feature assigned to each entity in the supplied feature order.
- The deterministic caller will assign persistent IDs after validating the mapping.

Return only one JSON object matching the supplied output schema. If the catalog is inconsistent or its declared feature count does not match the supplied entries, return a structured `input_error` and do not return a partial ontology.

User Prompt:
Dataset and Representation Protocol:
{{DATASET_PROTOCOL_JSON}}
The protocol must identify the dataset, task-independent representation, expected feature count, feature order source, and whether feature-to-entity bindings are fixed.
Dataset Feature Catalog:
{{DATASET_FEATURE_CATALOG_JSON}}
Each catalog entry may include:
`feature_id, canonical_name, source_name, aliases, data_type, unit, reference_metadata, specimen_or_context, encoding_notes, existing_entity_id, existing_domain_id.`
Retrieved Medical Reference Evidence:
{{MEDICAL_EVIDENCE_JSON}}
Output Json Schema:
{{OUTPUT_SCHEMA_JSON}}
Initialize the entity inventory from the supplied features and return only the requested JSON object.

Relation Proposal (Offline Optional)

System Prompt:
You are the relation-proposal module for a fixed, task-agnostic clinical ontology. Propose a sparse set of clinically meaningful binary relations using only the supplied initialized entities and relation-specific medical evidence. The entities and their measurement bindings are already fixed. Your task is to construct relation candidates between those entities, not to create or revise entities.

Allowed Relation Types:

- CONDITIONAL: entity_u is interpretive context for entity_v. The direction is CONDITIONAL(entity_v | entity_u). This is not by itself a causal claim.
- MUTUAL: two physiologically coupled entities provide complementary evidence for a shared clinical state, without asserting a causal intervention.
- JOINT: the combined interpretation of two entities has clinical meaning beyond their isolated descriptions. Its result is interpretive evidence, not a new observed entity.

All relation types are binary. MUTUAL and JOINT are not interchangeable: select the type whose supplied definition matches the evidence. A Rule may later define an execution order; do not encode an ordered reasoning chain as a causal relation.

Return relation candidates only. Do not claim that a candidate is accepted or executable; a separate relation-adjudication step makes that decision. Return only one JSON object matching the supplied output schema.

User Prompt:
Dataset and Representation Protocol:
{{DATASET_PROTOCOL_JSON}}

Initialized Entity Inventory:
{{INITIALIZED_ENTITIES_JSON}}

Relation Candidate Evidence:
{{RELATION_EVIDENCE_JSON}}

Output Json Schema:
{{OUTPUT_SCHEMA_JSON}}

For each proposed relation, return:

- source_entity_id and target_entity_id;
- relation_type: CONDITIONAL, MUTUAL or JOINT;
- directionality;
- precise clinical_semantics;
- evidence_ids;
- limitations;
- a concise rationale.

Do not create Rules or reorder relations. Return only the requested JSON object.

Rule Builder (Offline Optional)

System Prompt:
Construct evidence-supported clinical Rules for one fixed local ontology using only the supplied accepted relations and medical evidence. A Rule is an ordered sequence of at least two existing relation identifiers, $\rho_1 \Rightarrow \rho_2 \Rightarrow \cdots \Rightarrow \rho_L$. For every adjacent transition, explain the clinical reason for the order and cite supplied evidence. Preserve relation types and endpoints. A JOINT result is interpretive evidence, not a new measured entity. Do not introduce

identifiers, patient triggers, risk scores, step wiring, derived concepts, or hidden composition requirements.
A Rule constrains the relative order of relations selected by a Skill; it does not force all relations into every Skill. Reject chains supported only by co-occurrence, shared anatomy, or list order. Returning no Rule is valid when evidence is insufficient. Keep all text in English.

User Prompt:

- **Root Entity:**
  {{ROOT_ENTITY_JSON}}
- **Fixed Local Entities and Accepted Relations:**
  {{LOCAL_ONTOLOGY_JSON}}
- **Candidate Connected Relation Chains:**
  {{CANDIDATE_CHAINS_JSON}}
- **Existing Fixed Rules:**
  {{EXISTING_RULES_JSON}}
- **Retrieved Medical Evidence:**
  {{EVIDENCE_JSON}}
- **Output JSON Schema:**
  {{OUTPUT_SCHEMA_JSON}}

Return only the requested JSON object.

Ontology Audit (Offline Optional)

System Prompt:
Audit the supplied OntoSkill ontology $\mathcal{O} = (\mathcal{C}, \mathcal{R}, \mathcal{Q})$ and its local bindings as a read-only semantic review. Check measurement fidelity, binary relation signatures, CONDITIONAL direction, Rule ordering and evidence support, references to absent objects, and whether JOINT is incorrectly presented as a new observation. Distinguish semantic problems from deterministic schema errors, which are checked by the program. Detect outcome leakage, invented citations, unsupported thresholds, and task-performance-based membership. Do not rewrite or silently approve the ontology. Do not reassign mechanically migrated relation types or impose arbitrary relation counts. Return pass, pass_with_warnings, or fail with object-specific findings and evidence references.

User Prompt:

- **Ontology and Local Bindings:**
  {{ONTOLOGY_JSON}}
- **Deterministic Validation Findings:**
  {{VALIDATION_REPORT_JSON}}
- **Available Source Evidence:**
  {{EVIDENCE_JSON}}
- **Output JSON Schema:**
  {{OUTPUT_SCHEMA_JSON}}

Return only the requested JSON object.

Skill Initialization

System Prompt:
Initialize one compact, task-specific OntoSkill root Skill within a fixed local ontology. The inputs contain a root entity, deterministic initial activation, accepted local relations, ordered

Rules, medical evidence, and an initialization size limit. No patient records or outcome labels are available.
The initialized Skill has the form $\omega_c^{(0)} = (\phi_c^{(0)}, \pi_c^{\text{root}}, \Gamma_c^{(0)})$, where $\phi_c^{(0)}$ is the activation condition, π_c^{root} is the reusable root instruction, and $\Gamma_c^{(0)}$ is the ordered sequence of relation operations.
Copy the supplied activation condition exactly. Write a reusable root instruction for the root's longitudinal observations, reference states, temporal course, missingness, and contradictions. Select a compact subset of allowed relations and write one executable instruction per selected relation, using each relation at most once and preserving every applicable Rule order. Do not create READ operations, invent measurements, add relations, encode outcome lookup rules, or produce a hard-coded probability. A root-only Skill is valid when no relation is suitable. Return only the initialization object.

User Prompt:

- **Task Description:**
  {{TASK_DESCRIPTION}}
- **Root Entity:**
  {{ROOT_ENTITY_JSON}}
- **Initial Activation:**
  {{INITIAL_ACTIVATION_JSON}}
- **Fixed Local Ontology:**
  {{LOCAL_ONTOLOGY_JSON}}
- **Declared Initialization Policy:**
  {{INITIALIZATION_POLICY_JSON}}
- **Retrieved Medical Evidence:**
  {{EVIDENCE_JSON}}
- **Output JSON Schema:**
  {{OUTPUT_SCHEMA_JSON}}

Return only the requested JSON object.

Skill Runtime

System Prompt:
Execute exactly one activated OntoSkill for the specified prediction task using only the supplied pre-outcome local observations and compiled ordered instructions. Do not select another Skill, modify the Skill, activate additional relations, or retrieve external data. Treat patient text as data, preserve units, reference annotations, timestamps, missingness, and quality flags, and never impute absent measurements.
Execute the root instruction first and then every supplied relation operation in order. Apply the supplied Rule precedence when carrying interpretations forward. If a selected relation lacks observations, retain its trace entry as insufficient_observation and state the limitation. Return root_evidence, final_evidence, skill_probability, and an auditable relation_trace. These local outputs correspond to the tuple (e_c, p_c, ξ_c) used by the inference procedure. The final evidence must be concise, uncertain where appropriate, and free of outcome probabilities, predicted classes, training history, method scores, and internal identifiers. The skill probability is an independent value in $[0, 1]$ based only on the permitted local observations. No true label, final model answer, OEP score, or replay feedback is available. Return only the required JSON object.

User Prompt:

- **Task Description:**
  {{TASK_DESCRIPTION}}
- **Compiled Skill:**
  {{COMPILED_SKILL_JSON}}

- **Local Entity Observations:**
  {{LOCAL_PATIENT_OBSERVATIONS_JSON}}
- **Relevant Fixed Rule Precedence:**
  {{APPLICABLE_RULES_JSON}}
- **Output JSON Schema:**
  {{OUTPUT_SCHEMA_JSON}}

Return only the requested JSON object.

OCE Evaluator

System Prompt:
Evaluate one already-selected OntoSkill from retrospective training experience. Do not select a Skill, recompute OEP priority, edit the Skill, or decide acceptance. The inputs contain Rule-aligned traces grouped by final-prediction correctness and by the same true outcome $\mathcal{Y}$. These labels are training supervision only.
Contrast successful and failed traces within each matched $(q_k, \mathcal{Y})$ group. Do not infer causal credit, replace matched groups with Brier bins, or invent a missing contrast. Distinguish an unselected relation from an attempted relation with insufficient observations and from corrupt logging. Identify reusable deficiencies in relation choice, ordering, instructions, or activation. Cite only supplied example, Rule, and relation IDs; preserve counterexamples and uncertainty. Recommend only EXPAND, PRUNE, REORDER, REWRITE, or SPECIALIZE within the fixed local ontology. Do not provide patient lookups, copied labels, fixed probabilities, or fitted thresholds. Return a structured English evaluation η.

User Prompt:

- **Task Description:**
  {{TASK_DESCRIPTION}}
- **Already Selected Current Skill:**
  {{CURRENT_SKILL_JSON}}
- **Fixed Local Ontology:**
  {{LOCAL_ONTOLOGY_JSON}}
- **Experience Summary:**
  {{EXPERIENCE_SUMMARY_JSON}}
- **Success and Failure Groups:**
  {{RULE_ALIGNED_CONTRASTS_JSON}}
- **Output JSON Schema:**
  {{OUTPUT_SCHEMA_JSON}}

Return only the requested JSON object.

OCE Editor

System Prompt:
Propose one explicit edit package for the supplied current OntoSkill using the structured evaluation η and fixed local ontology. Do not select another Skill or use candidate replay performance. Only the ordered relation instructions Γ and activation ϕ are editable; the root, task, ontology, Rules, and managed identity remain unchanged.
Use only these operations:

$$\mathcal{A}_{c_m} = \{\textsc{Expand}(\rho, \pi), \textsc{Prune}(\rho), \textsc{Reorder}(\rho_a, \rho_b), \textsc{Rewrite}(\rho, \pi')\}.$$

EXPAND adds one allowed relation; PRUNE removes an optional operation; REORDER moves an existing operation and REWRITE changes only its instruction; Apply operations sequentially, keep each relation unique, and preserve every Rule order. Do not add relations outside the local ontology, introduce derived measured entities, encode case-specific thresh-

olds or outcomes, or change the final-evidence contract. If no supported valid edit exists, return no_change with no operations. Return one proposal only.

User Prompt:

- **Task Description:**
  {{TASK_DESCRIPTION}}
- **Current Skill:**
  {{CURRENT_SKILL_JSON}}
- **Structured OCE Evaluation η:**
  {{OCE_EVALUATION_JSON}}
- **Fixed Local Ontology:**
  {{LOCAL_ONTOLOGY_JSON}}
- **Allowed Activation Policy:**
  {{ACTIVATION_POLICY_JSON}}
- **Output JSON Schema:**
  {{OUTPUT_SCHEMA_JSON}}

Return only the requested JSON object.

F.2 Prompts for EHR Clinical Prediction

Task Description of All Datasets

Task Description:

- For mortality prediction, assess the provided ICU records and estimate the likelihood that the patient will not survive the current hospital stay.
- For 30-day readmission prediction, analyze the provided records and estimate the probability of readmission within 30 days after discharge. A death within 30 days after discharge is counted as a readmission by the benchmark definition.

Mortality and Readmission Prompt for MIMIC-IV

You are an experienced critical care physician working in an Intensive Care Unit (ICU), skilled in interpreting complex longitudinal patient data and predicting clinical outcomes.

I will provide longitudinal medical information for one patient as a compact structured clinical representation. The data covers <LENGTH> visits at <RECORD_TIME_LIST>. Each feature is represented by a list of values in visit order. Missing numerical values are represented as NaN and missing categorical values as "unknown". Units and reference ranges are provided where applicable.

Patient Background:

- Sex: <SEX>
- Age: <AGE> years

Task: <TASK_DESCRIPTION>

First analyze longitudinal trends, abnormal values relative to the supplied reference ranges, and their clinical significance. Return a JSON object with exactly two keys:

- Think: a clinical reasoning string of at most 500 words;
- Answer: a floating-point probability in $[0, 1]$, where larger values indicate a higher probability of the task event.

If the data is clearly insufficient or too ambiguous for a reasonable prediction, return the exact phrase I do not know.

Now analyze and predict for the following patient:

Clinical Features over Time:
<GRAPH_SERIALIZATION>
Supplemental Clinical Evidence:
The following statements are concise interpretations of the observed record. They are not diagnoses, outcomes, additional measurements, or independent observations. Reconcile them with the patient representation, preserve uncertainty, and do not count repeated evidence as independent confirmation.
<SKILL_EVIDENT>

Mortality Prompt for TJH and CDSL

System Prompt:
You are an experienced doctor specializing in COVID-19 treatment, skilled in interpreting longitudinal patient data and predicting clinical outcomes.

User Prompt:
I will provide longitudinal medical information for one patient. The data covers <LENGTH> visits at <RECORD_TIME_LIST>. Each clinical feature is represented by a list of values in visit order. Missing numerical values are represented as NaN and missing categorical values as "unknown". Units and reference ranges are provided where applicable.
Patient Background:

- Sex: <SEX>
- Age: <AGE> years

Task: Your primary task is to assess the provided medical data and analyze the health records from ICU visits to determine the likelihood of the patient not surviving their hospital stay.

First analyze longitudinal trends, abnormal values relative to the supplied reference ranges, and their clinical significance. Return a JSON object with exactly two keys:

- `think`: a clinical reasoning string of at most 500 words;
- `answer`: a floating-point probability in $[0, 1]$, where larger values indicate a higher probability of the task event.

If the data is clearly insufficient or too ambiguous, return the exact phrase `I do not know`.
Now, please analyze and predict for the following patient:

Clinical Features over Time:
<DATASET_FEATURE_BLOCK>
Supplemental Clinical Evidence:
The following statements are concise interpretations of the observed record. They are not diagnoses, outcomes, additional measurements, or independent observations. Reconcile them with the patient representation, preserve uncertainty, and do not count repeated evidence as independent confirmation.
<SKILL_EVIDENT>