Reasoning-Enhanced Healthcare Predictions with Knowledge Graph Community Retrieval

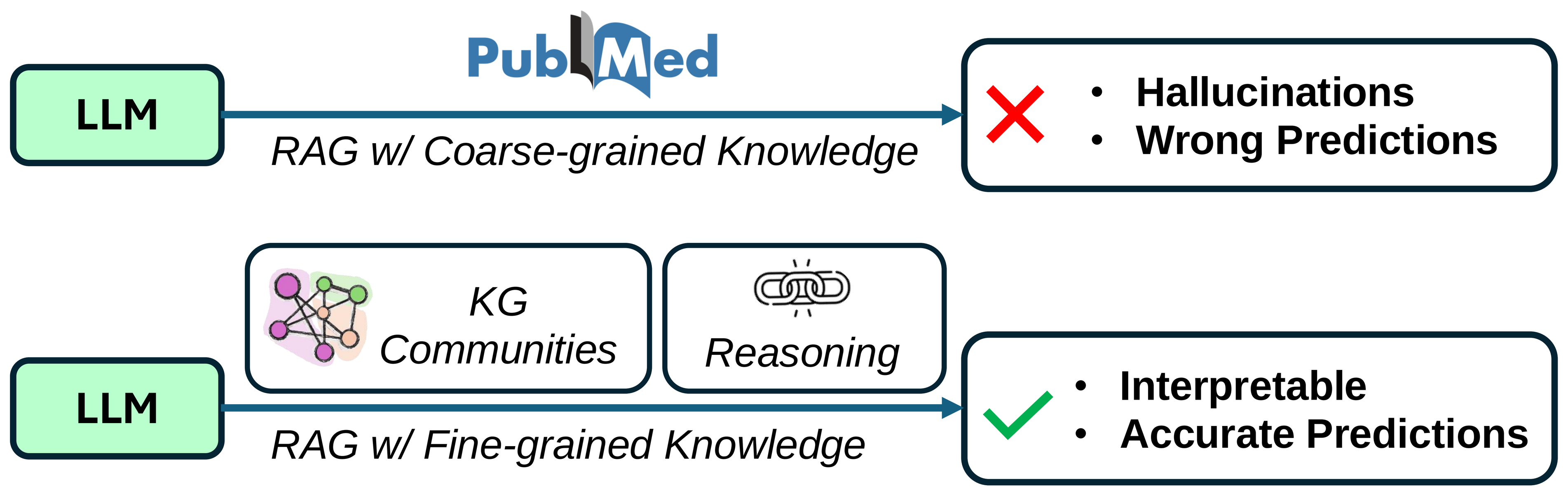
Pengcheng Jiang[1], Cao Xiao[2], Minhao Jiang[2], Parminder Bhatia[2], Taha Kass-Hout[2], Jimeng Sun[1], Jiawei Han[1]

[1]University of Illinois Urbana-Champaign [2]GE HealthCare

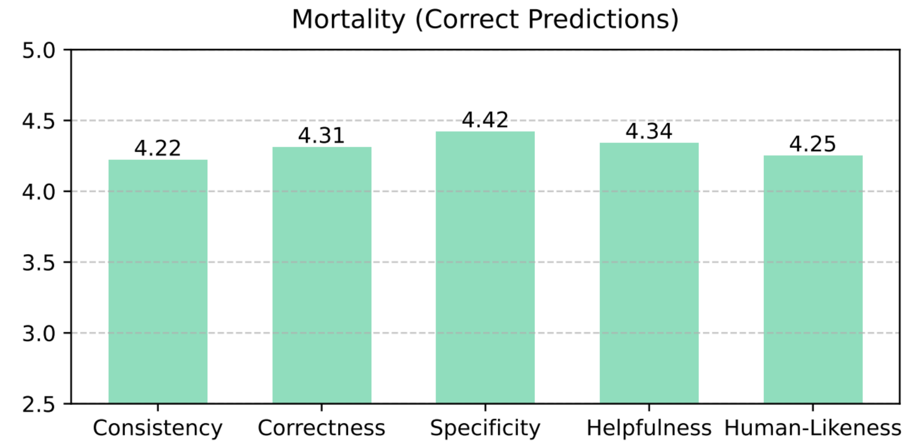
Paper Code

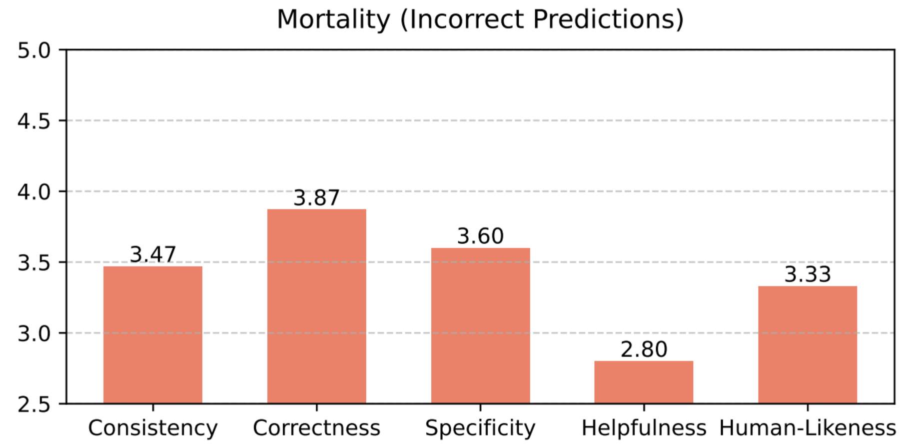
Introduction

- **Motivation**: LLMs hallucinate & struggle in healthcare due to coarse-grained knowledge and irrelevant retrieval.
- **Goal**: Enhance LLM predictions with fine-grained, context-relevant knowledge via knowledge graphs.
- **Solution Overview**: Introduce **KARE**, a framework that integrates hierarchical KG community retrieval and LLM reasoning.
- **Impact**: Up to 15% improvement on mortality and readmission prediction tasks across MIMIC-III/IV.



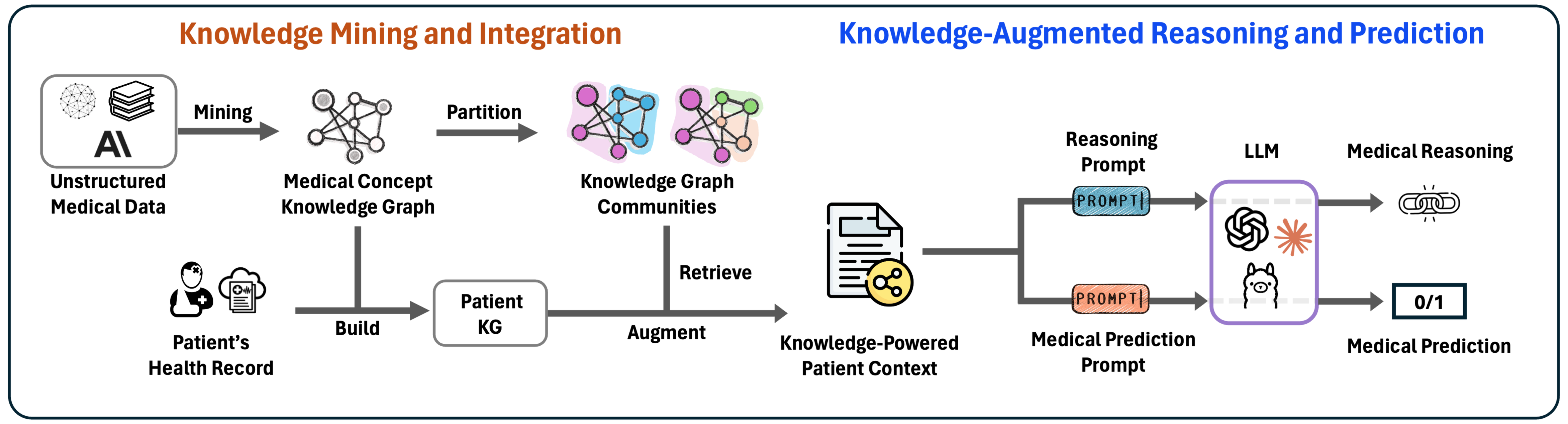
Expert's Evaluation of KARE's Clinical Reasoning





Correct predictions: High scores in specificity, helpfulness, and correctness.
Incorrect predictions: Quality drops, especially in helpfulness (2.80), but reasoning remains moderately consistent and correct.
Insight: KARE generates clinically valuable and interpretable reasoning, even under prediction errors.

KARE Framework



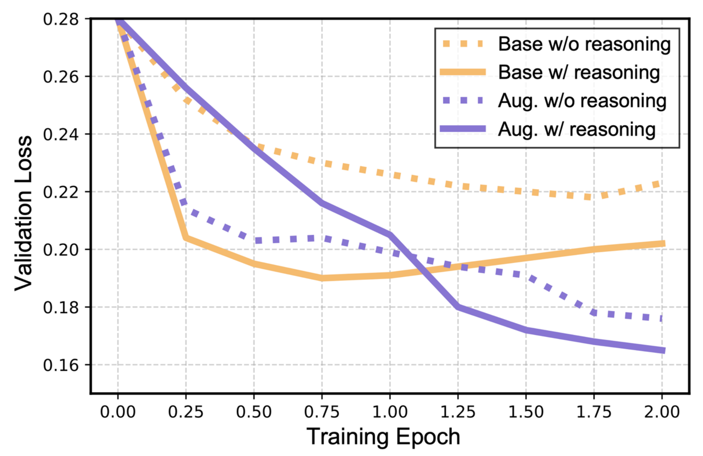
(Simplified Version for Illustration. Find the complete version in our paper.)

Step 1: KG Construction & Indexing
→ Build a multi-source medical KG from EHRs, PubMed, and LLM-inferred links → Cluster semantically similar concepts → Detect and summarize hierarchical graph communities
Step 2: Patient Context Augmentation
→ Construct a patient-specific subgraph
→ Select relevant community summaries using node hits, coherence, and recency → Dynamically enrich EHR context
Step 3: Reasoning-Enhanced Prediction
→ Use an expert LLM to generate reasoning chains → Fine-tune a smaller LLM with both reasoning and label supervision → Predict outcomes with interpretable, step-by-step rationale

Ablation Studies on Training Components of KARE

Similar Patients	Retrieved Knowledge	Reasoning	MIMIC-III-Mortality Accuracy	Macro F1	Sensitivity	Specificity	MIMIC-III-Readmission Accuracy	Macro F1	Sensitivity	Specificity
✗	✗	✗	90.4	53.0	11.4	94.3	57.6	57.6	50.5	66.3
✗	✗	✓	93.1	58.4	15.8	97.5	65.5	64.7	62.3	67.7
✗	✓	✓	95.3	64.6	24.7	98.3	72.8	72.6	74.7	70.6
✓	✓	✓	93.6	61.3	18.4	98.6	73.9	73.7	76.7	70.7

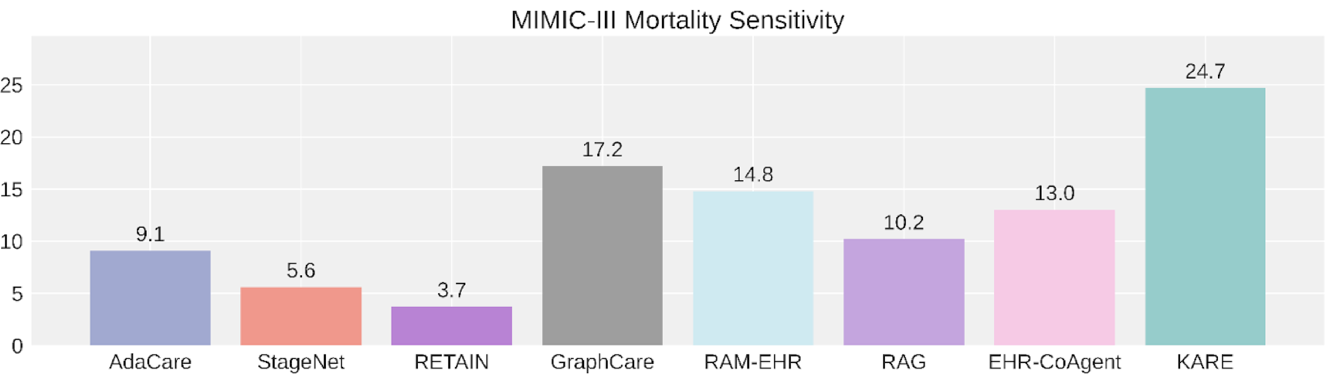
Similar Patients	Retrieved Knowledge	Reasoning	MIMIC-IV-Mortality Accuracy	Macro F1	Sensitivity	Specificity	MIMIC-IV-Readmission Accuracy	Macro F1	Sensitivity	Specificity
✗	✗	✗	92.2	83.1	65.0	96.2	56.1	46.7	23.1	76.2
✗	✗	✓	93.3	85.4	67.3	97.5	64.7	62.1	69.3	55.9
✗	✓	✓	93.8	89.6	74.5	98.8	72.2	71.9	81.1	64.0
✓	✓	✓	94.1	90.4	73.2	99.9	73.9	73.8	85.6	63.7

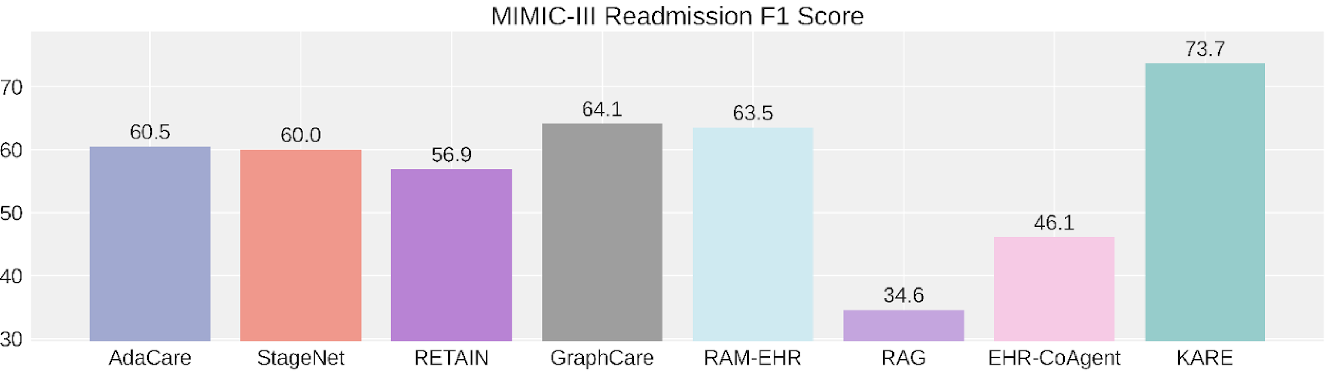


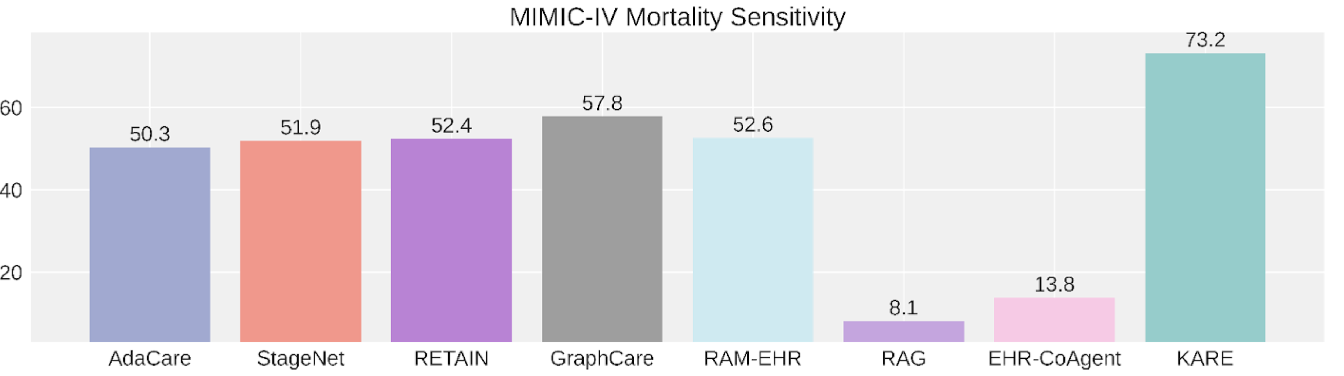
1.Both retrieved knowledge and reasoning chain significantly contribute to the performance gain
2.When the data is imbalanced (MIMIC-III-Mortality), similar patient retrieval hurts the performance
3.Without retrieved knowledge, the LLM could easily encounter the overfitting issue

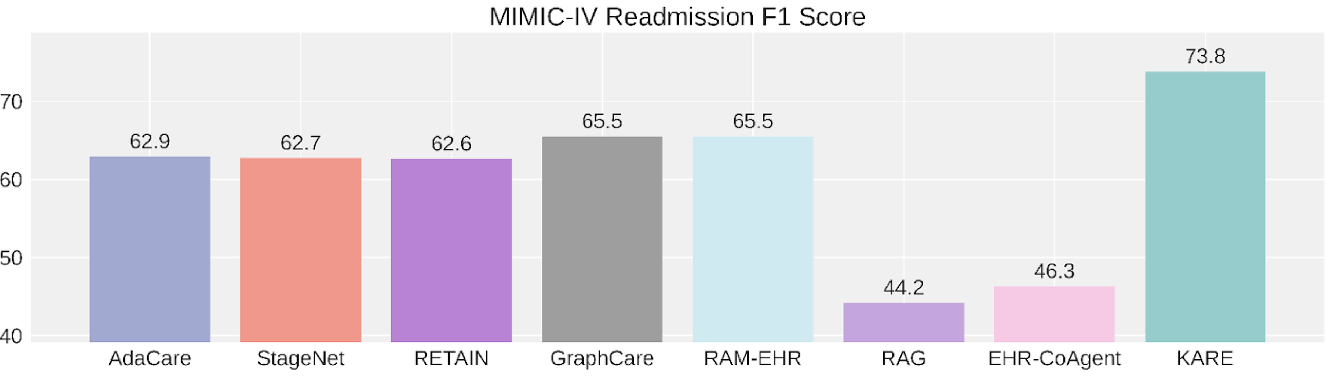
Performance on MIMIC-III/IV

KARE outperforms leading models by a large margin on mortality and readmission prediction tasks:









Future Directions

- RL-Driven Reasoning Optimization (i.e. R1-like)
- Interactive Clinical Feedback Loop
- Multi-task generalization (e.g., multi-label diagnosis)
- Scalable Community Retrieval

Email Patrick Jiang (pj20@illinois.edu) for further questions and discussions!