Agent-Augmented Sequence Modeling for Explainable ICU Diagnosis Prediction

AIM2 Course Project Proposal (Week 4)

Anya Sharma* Sriya Surapaneni Kristine Yang

Instructions (do not delete this box). This proposal is at most **2 pages of main text** (single-spaced, 12pt font, 1" margins). You may use a **3rd page for figures/tables** and have **unlimited references**. Keep the section headings below and replace the placeholder text with your own writing.

Hypothesis or Research Question

Can a multi-agent LLM reasoning layer improve primary ICD-10 diagnosis prediction accuracy on held-out MIMIC-IV ICU codes compared to a trained LSTM baseline alone, as measured by Top-1 accuracy and AUROC?

Background and Significance

Early intensive care unit (ICU) diagnosis prediction enables timely clinical decision-making and resource allocation. Machine learning models trained on ICU electronic health record time-series data can identify clinically meaningful patterns within the first hours of admission, allowing early prediction of diagnoses and outcomes and often outperforming traditional scoring systems like APACHE and SAPS [1] [2]. Additionally, recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) architectures have demonstrated strong performance on multivariate clinical time-series tasks by capturing temporal dependencies in physiological data [3] [4].

However, these models lack explicit clinical reasoning and may produce high-confidence errors when cases are ambiguous. Large language models (LLMs) have recently shown promise in medical reasoning by integrating clinical context, but they still remain unreliable when applied to structured physiological data [5] [6] [7].

This limitation has motivated hybrid approaches that combine predictive models with reasoning systems to improve clinical reliability. One such approach is MedGellan, which uses an LLM to generate reasoning-based clinical guidance that enhances diagnostic performance [8]. Multi-agent LLM systems further extend this by producing and evaluating multiple candidate explanations to improve decision quality, but their use as inference-time reasoning layer on top of trained clinical prediction models remains largely unexplored [9]. This project addresses that gap by evaluating whether a multi-agent reasoning layer can improve diagnosis prediction accuracy and interpretability without retraining the underlying model.

Dataset

MIMIC-IV is a publicly available ICU EHR dataset with de-identified longitudinal time-series data, covering 94,548 ICU stays for 65,366 patients (2008–2022) [2]. It includes vital signs, labs, medications, discharge summaries, and ICD-10 codes. Its high-frequency structure supports sequence modeling, and its size enables multi-class classification over the top-K diagnoses, with broad use in clinical ML providing a benchmark context [10].

Preprocessing filters the first 12 hours of ICU data from stays 12 hours, aggregated into hourly bins to produce 12 feature vectors per stay, with forward-filling for missing values. The target is the primary discharge diagnosis among the top-K ICD codes. Data are split 70/10/20 into train, validation, and test sets at the patient level to avoid leakage.

Preprint.

Baseline Model

We use an LSTM classifier as the baseline, a standard architecture for multivariate clinical time-series modeling widely applied to ICU outcome and diagnosis prediction [4]. Because the task predicts diagnosis from the first 12 hours of ICU data, the LSTM provides a strong sequence-based benchmark. Implemented in PyHealth as a single-label classifier, the model takes 12 hourly feature vectors as input, encodes them with one to two stacked LSTM layers, and applies a fully connected softmax layer to produce probabilities over K diagnoses, with predictions taken via argmax. Training uses cross-entropy loss, the Adam optimizer, and early stopping based on validation accuracy.

Proposed Methodology

The methodology for the Agent-Augmented model is as follows: at inference, the LSTM produces a ranked list of the top-3 diagnosis candidates with associated confidence scores. A separate LLM agent is instantiated for each candidate and provided with a structured natural-language summary of the patient's 12-hour clinical record, and the candidate diagnosis label. Each agent generates a brief clinical rationale assessing the plausibility of that diagnosis given the patient's presentation. A ranking agent then receives all three rationales (and the LSTM model's confidence scores for each candidate diagnosis) and selects a single final diagnosis as the system's prediction. No additional LSTM training is performed; the agent layer operates entirely at inference time.

Both systems (the baseline LSTM and the agent-augmented model) are evaluated on the held-out test set against the ground truth primary diagnosis. We report top-1 accuracy (exact match), top-3 accuracy (reported once as context, not as a system metric, just to characterize how often the right answer was even available to the agent), macro-averaged AUROC, and Precision@1. We additionally analyze the subset of cases in which the agent's final selection differs from the LSTM's top-1 prediction, to assess whether the reasoning layer produces systematic improvements or degradations across diagnosis categories.

While LLMs have been explored as standalone diagnostic tools, their use as a structured reasoning layer on top of a trained clinical prediction model is underexplored. This project investigates whether agent-based reasoning can meaningfully improve a deep learning system's predictions at inference time (without any retraining) offering a new approach to augmenting existing clinical models that, if effective, could generalize to other structured prediction tasks in healthcare.

Resources

All experiments will be implemented in Python using PyHealth for preprocessing and model construction and PyTorch for LSTM training. Models will be trained on Google Colab Pro with A100 GPU access; given the small LSTM architecture (1–2 layers, 12 timesteps) and dataset scope, standard GPU allocation should suffice. The agent layer will use the Gemini API via existing Colab credits, with additional credits purchased if full test-set inference exceeds the quota.

Challenges and Contingency Plans

Several practical challenges are anticipated. ICU data are often incomplete, and variables exceeding a predefined missingness threshold will be dropped and replaced with missingness indicators. If class imbalance among the top-K diagnoses destabilizes training, class-weighted cross-entropy loss will be used.

The agent-augmented model also introduces API cost constraints. If full test-set LLM inference is too expensive, evaluation will use a stratified random sample of 500 patients. If natural-language summaries yield inconsistent or clinically implausible rationales, the template will be simplified and further structured to reduce variability. If the agent layer does not improve over the LSTM baseline, the project will instead focus on override analysis, identifying diagnosis categories and clinical patterns where LLM reasoning systematically helps or harms performance.

References

[1] Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, June 2019.

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

[3] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports*, 8(1):6085, April 2018.

[4] Zachary C. Lipton, David C. Kale, Charles Elkan, and Randall Wetzel. Learning to Diagnose with LSTM Recurrent Neural Networks, March 2017. arXiv:1511.03677 [cs].

[5] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of GPT-4 on Medical Challenge Problems, April 2023. arXiv:2303.13375 [cs].

[6] Arya Rao, John Kim, Meghana Kamineni, Michael Pang, Winston Lie, and Marc D. Succi. Evaluating ChatGPT as an Adjunct for Radiologic Decision-Making. *medRxiv*, page 2023.02.02.23285399, February 2023.

[7] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, August 2023.

[8] Debodeep Banerjee, Burcu Sayin, Stefano Teso, and Andrea Passerini. MedGellan: LLM-Generated Medical Guidance to Support Physicians, September 2025. arXiv:2507.04431 [cs].

[9] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior, August 2023. arXiv:2304.03442 [cs].

[10] Sebastiano Barbieri, James Kemp, Oscar Perez-Concha, Sradha Kotwal, Martin Gallagher, Angus Ritchie, and Louisa Jorm. Benchmarking Deep Learning Architectures for Predicting Readmission to the ICU and Describing Patients-at-Risk. *Scientific Reports*, 10:1111, January 2020.