# Large Language Models vs. Traditional Machine Learning for Clinical Risk Prediction: A Mechanistic Analysis of When Statistical and Semantic Models Fail

Aaron Ge[1], Jialong Wu[3], Jonas De Almeida[2], Bradley Maron[1]*

[1] Institute of Health Computing, University of Maryland, School of Medicine, Baltimore, MD, USA

[2] Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Maryland, USA

[3] Department of Computer Science, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, USA

## BACKGROUND

- Clinical prediction can be built on two paradigms: statistical models using structured numeric features versus semantic models that encode patient text with large language model (LLM) embeddings.
- It remains unclear when each paradigm excels or fails, and how input design (representation) and prompting influence performance.
- Understanding the mechanisms baehind disagreements between paradigms is critical for safe deployment in healthcare.

## OBJECTIVE

Predict 6 different binary endpoints (in-hospital mortality, ) using 458 engineered clinical features derived from the first 24 hours of ICU stay.

- **H1 LLM embeddings vs. XGBoost**: Optimized LLM embedding model will perform similarly to an optimized XGBoost model.
- **H2 Hybrid Model Benefit**: Models make different mistakes, so combining these models will improve performance.
- **H3 LLM Sensitivity**: Semantic system performance is sensitive to text serialization and LLM prompt task-alignment.

## METHODS

**1. Data Preparation**

- **Dataset:** ICU patient data with 458 clinical features from first 24 hours of admission.
- **Prediction Tasks:** In-hospital mortality, Vasopressor use, Mechanical ventilation, ICU length of stay >3 days, Total length of stay >7 days, 30-day readmission.

**2. Model Development**

- Numerical Models (NM)
- Semantic Models (SM)
  - **Text Serialization:** F1: Raw feature values, F2: Key-value pairs, F3: Narrative clinical summary.
  - **LLM Embeddings:** Multiple models compared: General-purpose LLMs, Medical-specific LLMs.

**3. Hybrid Model Fusion:**

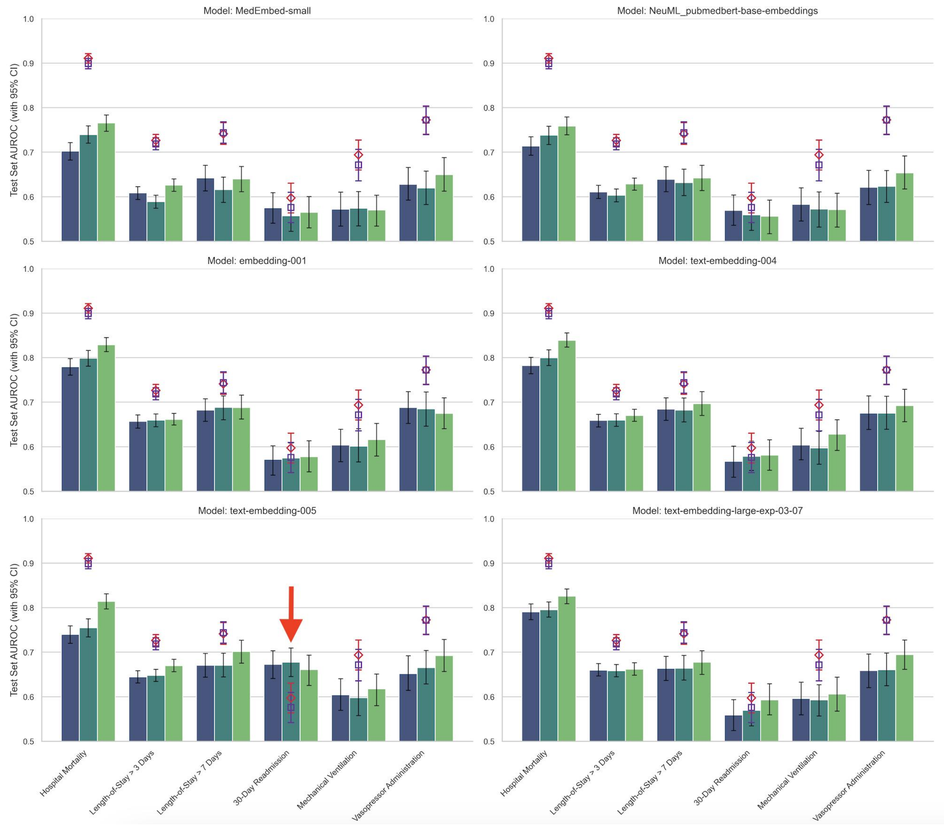
- **25+ fusion strategies systematically tested**

**4. Evaluation Framework**

- **Primary Metrics:** AUROC, AUPRC.
- **Statistical Analysis:** Stratified bootstrap (1000 iterations), 95% confidence intervals.

## RESULTS

**H1 LLM embeddings vs. XGBoost:** Human-engineered features still beat semantic embeddings.There is one notable exception.The text-embedding-005 model using the F3 representation did show a slight performance advantage over the XGBoost baseline for the 30-Day Readmission task.

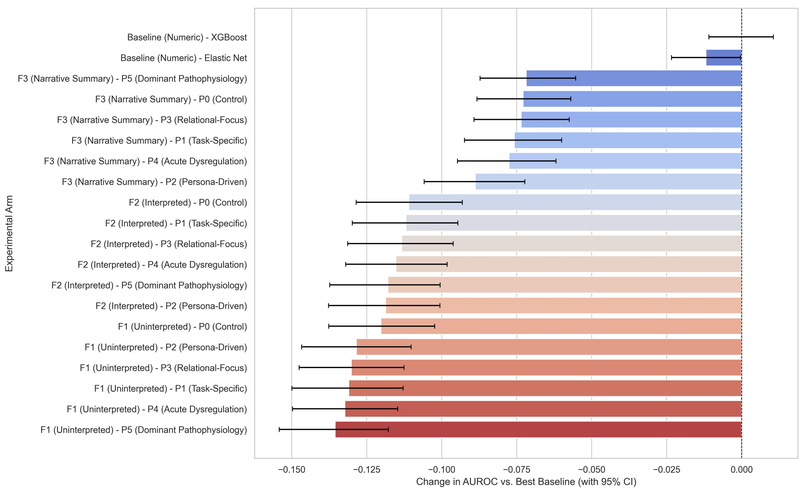



**H2 Hybrid Model Benefit:** Synergy is observed in 4 out of 6 tasks. For the task where the base models performed best (Mortality), fusion failed entirely.

| Prediction Task | Best Baseline Model | Best Baseline AUROC (95% CI) | Best Fusion Strategy | Best Fusion AUROC (95% CI) | Absolute Change |
|---|---|---|---|---|---|
| In-Hospital Mortality | Numerical | 0.9107 (0.8998-0.9213) | No Synergy Observed | 0.9100 (0.8968-0.9204) | -0.0007 |
| 30-Day Readmission | Semantic | 0.6776 (0.6539-0.7101) | No Synergy Observed | 0.6712 (0.6412-0.7020) | -0.0064 |
| **Vasopressor Intervention** | **Numerical** | **0.7724 (0.7405-0.8025)** | **Neural Network** | **0.7862 (0.7548-0.8176)** | **+0.0138*** |
| Ventilation Intervention | Numerical | 0.6939 (0.6602-0.7271) | Hierarchical | 0.7023 (0.6690-0.7356) | +0.0084 |
| Length of Stay > 3 Days | Numerical | 0.7265 (0.7140-0.7397) | Dynamic Weight | 0.7324 (0.7198-0.7452) | +0.0059 |
| Length of Stay > 7 Days | Numerical | 0.7438 (0.7203-0.7682) | Ensemble | 0.7499 (0.7268-0.7747) | +0.0061 |

**H3 Prompt Sensitivity:** Prompting is a Surprisingly Ineffective Strategy. There is no evidence that complex, task-aligned prompts provide any statistically significant benefit over a simple Control prompt (P0).

Figure 3: AUROC Improvement Over Best Baseline for Hospital Mortality




## CONCLUSIONS

- **Numerical Models Remain Strong:** Traditional numerical models outperform semantic embedding models for most prediction tasks, despite advanced prompt engineering and diverse data serialization techniques.
- **Narrative Data Shows Promise:** The format of data for LLMs is crucial. Succinct representations (F3) are more effective in capturing clinically relevant information than structured value lists.
- **SM and NM Have Distinct Failure Modes:** Synergy is possible, but it is conditional and not a general property of combining these two modalities. Numerical models do better on patients with invasive hemodynamic monitoring. Semantic models do better on patients with AKI.