When Raw Data Prevails:

Are Large Language Model Embeddings Effective in Numerical Data Representation for Medical Machine Learning Applications?

Yanjun Gao, Skatje Myers, Shan Chen, Dmitriy Dligach, Timothy Miller, Danielle S. Bitterman, Matthew Churpek, Majid Afshar

Department of Biomedical Informatics, School of Medicine, University of Colorado Anschutz Medical Campus
LARK Language, Reasoning, Knowledge
Wisconsin University of Wisconsin–Madison
Loyola University Chicago
Dana-Farber Cancer Institute
Harvard Medical School

Introduction

- Tabular data constitutes a large portion of electronic health records (EHR) information.
- Traditional machine learning (ML) classifiers excel with raw data features.
- LLM-derived features for clinical predictive modeling remain unexplored.

Can LLM embeddings replace raw data features for medical ML applications?

Contributions

- A comprehensive study exploring LLM-generated numerical embeddings for medical ML.
- LLM embeddings show promise but need improvement to match raw data.
- LLM feature representation's efficiency and robustness are evaluated.

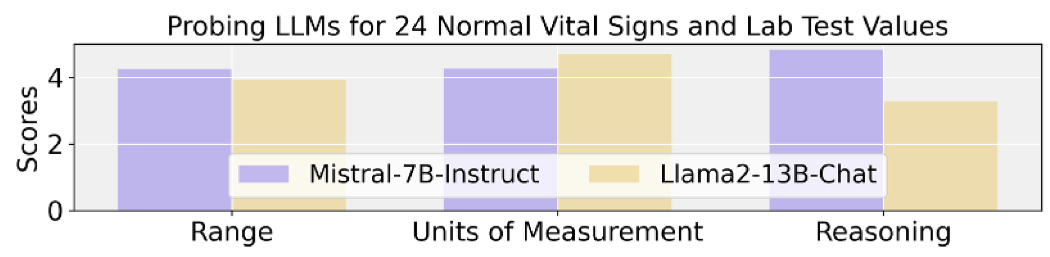
Does LLM Know Clinical Meanings of Numerical Data?

Example Question: What is the reference range and units of measurement for Systolic Blood Pressure? (Answer: 90-120 mm Hg.)

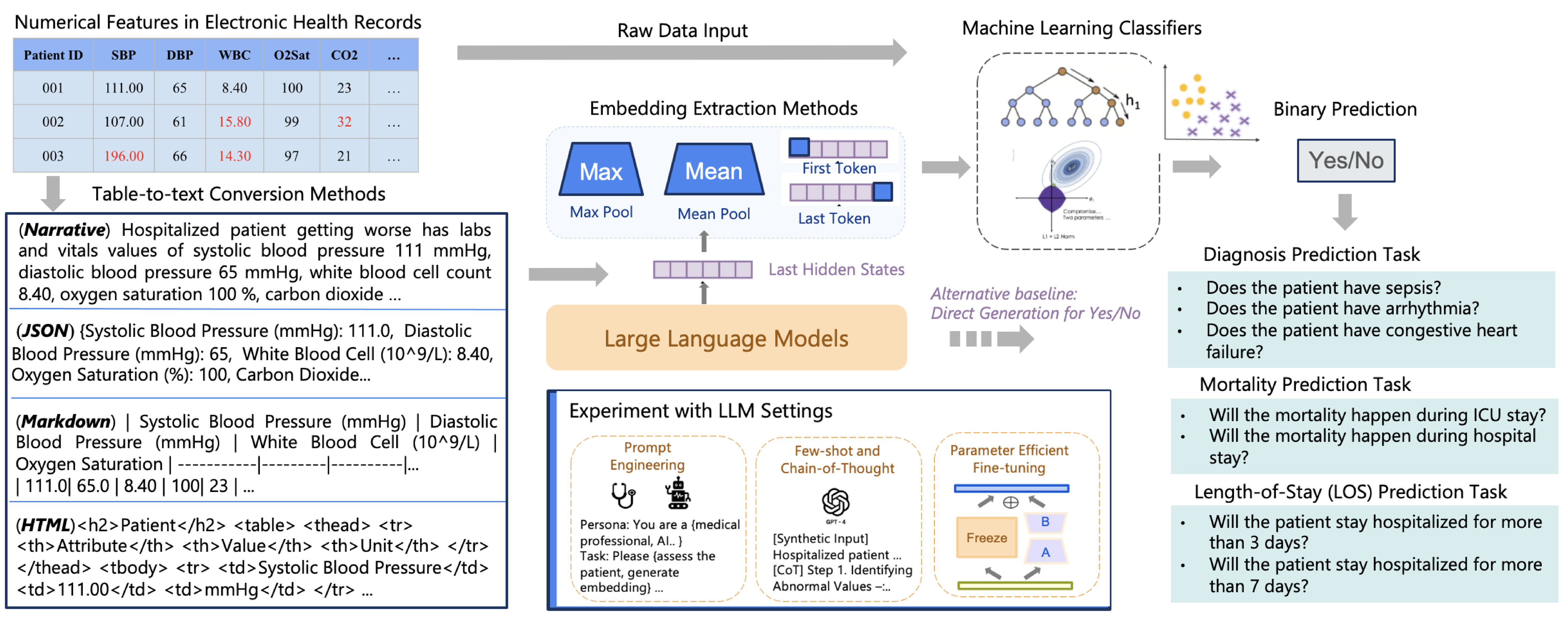
Likert Scale for Evaluation: (1) Strongly Disagree; (2) Disagree; (3) Neither Agree nor Disagree; (4) Agree; (5) Strongly Agree

Mistral: The reference range for Systolic Blood Pressure is typically between 90-120 mmHg for adults. However, this can vary depending on factors such as age, gender, and health status. It is important to note that having a blood pressure within the reference range does not necessarily mean that it is optimal, and further evaluation may be necessary based on individual health factors.

Physician: Range: 5; Units of Measurement: 5; Reasoning: 5.



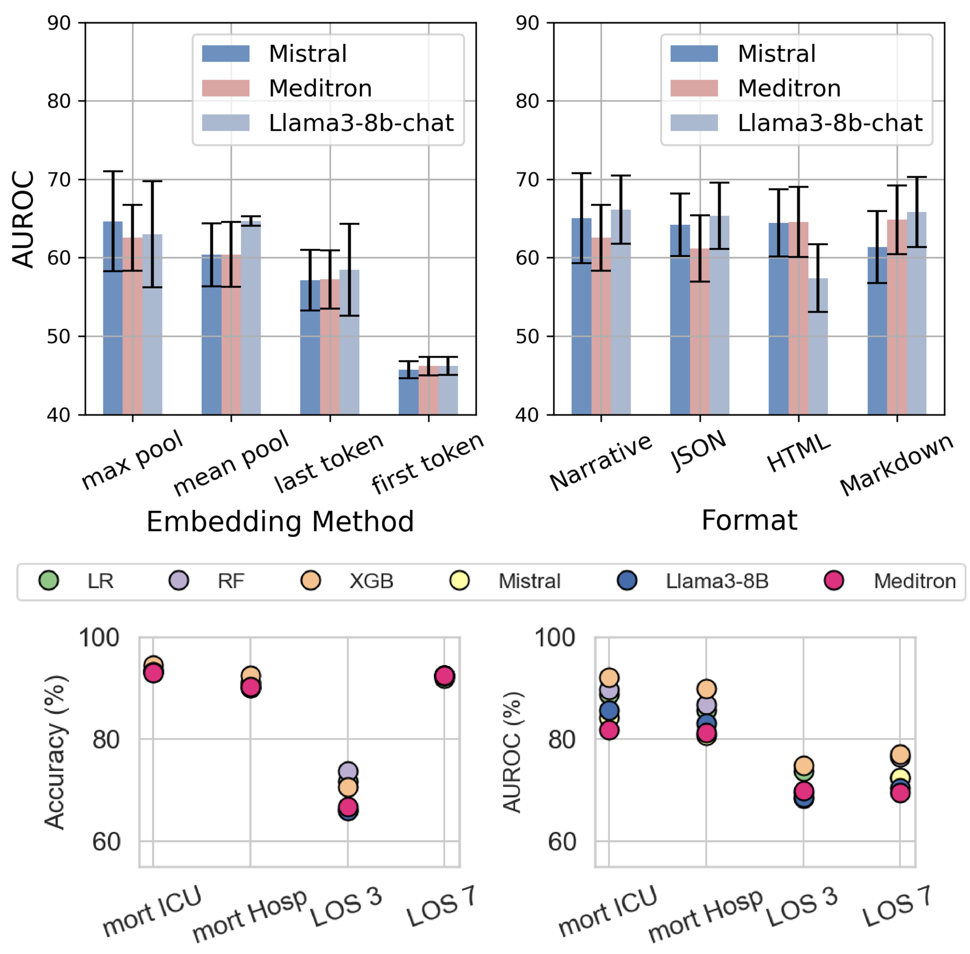
Method



- Table-to-text formats
- Embedding extraction formats
- LLM Setup:
  - Prompt engineering
  - Few-shot and Chain-of-thought
  - Parameter efficient fine-tuning
- ML Classifiers:
  - XGBoost
  - Logistic Regression
- Datasets:
  - Diagnosis prediction
  - MIMIC-Extract

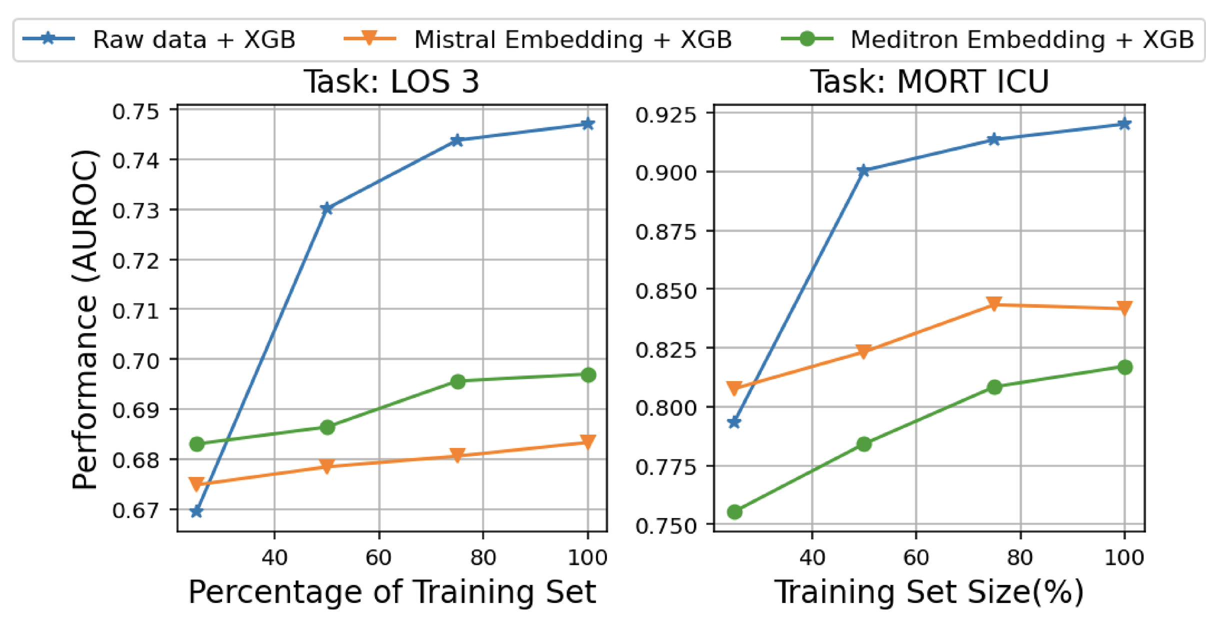
Results

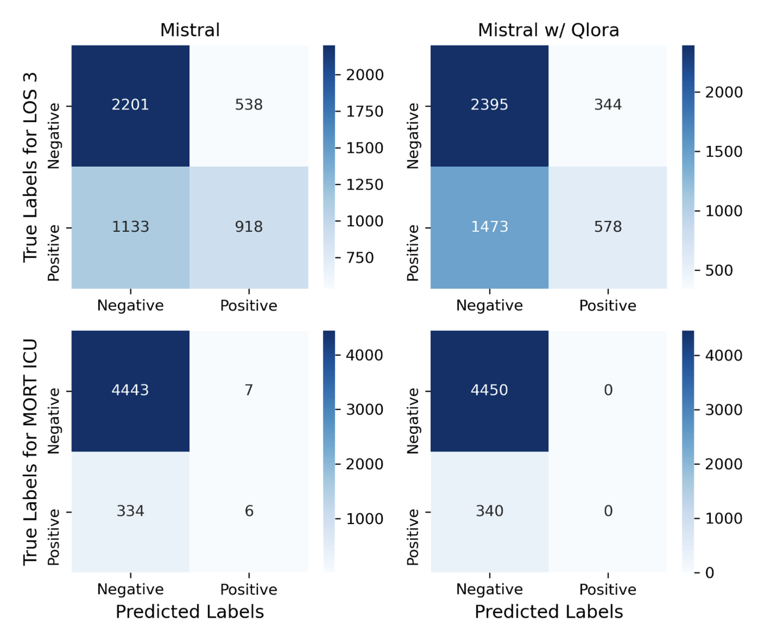
- Baseline
  - Raw data + ML Clasifiers
  - Random embedding + ML Classifiers
  - ClinicalBERT
- LLM Selection
  - General domain LLMs like Mistral-7B, Llama2 and Llama3
  - Clinical text LLM: Meditron



Model	Sepsis AUROC (95% CI)	Arrhythmia AUROC (95% CI)	CHF AUROC (95% CI)	Average (95% CI)
Raw Data Features Baseline				
LogisticRegression	71.10 (67.01, 75.18)	74.40 (69.35, 79.56)	54.79 (47.74, 61.79)	66.76 (61.37, 72.18)
RandomForest	65.26 (61.79, 68.48)	53.07 (50.58, 55.80)	50.89 (49.01, 53.43)	56.41 (53.79, 59.24)
XGB	**71.17 (67.06, 75.11)**	**76.49 (71.32, 84.13)**	58.47 (51.36, 65.15)	**68.71 (63.25, 74.80)**
LLM embedding + XGB classifier				
Random	54.01 (49.89,58.44)	49.65(44.02,54.62)	50.02 (47.13, 52.29)	51.22 (47.01, 55.19)
Mistral-7b-Instruct$_{best}$	71.12 (67.54, 74.92)	68.00 (61.52, 73.93)	51.80 (44.48, 58.65)	63.40 (57.73, 68.77)
Llama3-8b-Instruct$_{best}$	63.84 (57.31, 69.87)	71.08 (65.69, 75.87)	**63.84 (56.77, 70.37)**	66.25 (60.15,72.35)
Llama2-13b$_{best}$	66.02 (61.64, 70.32)	58.62 (52.62, 64.46)	49.69 (48.83, 62.58)	58.11 (54.36, 65.79)
Llama2-70b-chat$_{best}$	68.57 (63.88, 71.53)	69.15 (67.08, 71.17)	53.87 (49.83, 58.52)	63.86 (60.93, 67.07)
Meditron$_{best}$	66.74 (62.30, 66.15)	72.26 (65.28, 77.43)	58.11 (50.64, 64.48)	63.90 (58.28, 65.45)
ClinicalBERT	58.80 (54.44, 63.04)	64.91 (61.84, 70.27)	49.67 (41.94, 57.51)	57.79 (52.74, 63.11)
LLM embedding + Logistic Regression classifier				
Random	49.58 (47.68, 51.12)	49.22 (48.09, 50.43)	49.36 (47.12 51.06)	49.39 (47.63, 50.87)
Mistral-7b-Instruct$_{best}$	62.61 (58.17, 66.95)	69.59 (64.67, 74.71)	48.98 (42.96,55.62)	60.39 (55.27, 65.76)
Llama3-8b-Instruct$_{best}$	66.54 (62.32, 70.62)	70.22 (64.82, 74.11)	63.52 (55.91,69.20)	66.76 (61.50, 72.02)
Llama2-13b-chat-hf$_{best}$	66.95 (62.82, 70.88)	66.04 (60.04, 71.22)	58.54 (52.09, 65.09)	63.84 (58.32, 69.06)
Llama2-70b-chat-hf$_{best}$	69.50 (65.37, 73.43)	68.11 (61.75, 70.57)	62.72 (56.47, 68.39)	66.78 (61.20, 70.80)
Meditron$_{best}$	66.91 (62.83, 71.09)	68.61 (63.49, 73.72)	57.60 (51.02, 63.89)	64.37 (59.11, 69.90)
ClinicalBERT	47.28 (43.07, 51.63)	44.62 (38.79, 50.29)	46.98 (42.96, 55.62)	46.29 (41.61, 52.51)

Discussion





- Raw data features provide **more informative** input for ML models than LLM embeddings.
- LLM embeddings show smaller performance gains with larger training sets than raw data.
- Zero-shot LLM embeddings **show promis**e but need efficiency improvements for practical use.
- Embeddings are still outperforming LLM direct generation on these tasks.

Conclusion

LLM embeddings cannot replace raw data features yet, but have strong potentials!

Acknowledgements

U.S NIH/NLM R00LM014308-02, R01LM012973-05, NHLBI R01HL157262-04, NIH-USA U54CA274516-01A1 and NIH-USA R01CA294033-01.