Large Language Models and Traditional Machine Learning Methods: A Comparative Study on Medication Error Detection

Joseph Chopra, M2[1], Masum Shah Junayed, Graduate Student[2], Yiming Zhang, Graduate Student[2], Swapna Gokhale, PhD[2], Steven Demurjian, PhD[2], Thomas Agresta, MD[3]

University of Connecticut [1]School of Medicine, [2]Department of Computer Science & Engineering, [3]Family Medicine, Center for Quantitative Medicine, School of Medicine

INTRODUCTION

Medication reconciliation errors, or Med-Wreck, are highly prevalent[1] and substantial drivers of post-discharge patient adverse events.[2]

- **Med-Wreck Findings:**
  - **81% of patients' records at one hospital had ≥ 1 reconciliation error.[3]**
  - **39% of Med-Rec errors at another had potential for harm.[4]**
  - **Actual values are likely higher as these errors are difficult to track.**

Recent advances in machine learning (ML) models, especially large language models (LLMs), may provide a solution.

LLM Med-Rec Benefits	LLM Med-Rec Limitations
Could flag errors for faster, more accurate physician review.	Require large amounts of highly detailed training data.
Do not require data from external organizations to provide suggestions.	Patient privacy standards may limit medical applications of ML.

Project Components:

1. **Data generation: Created a standalone LLM-based tool to generate entirely synthetic, incorrectly reconciled patient medication lists to address the problem of limited data.**
2. **Model training and evaluation: Used these lists to train and test several ML models' ability to identify and flag Med-Rec errors.**

METHODS – DATA GENERATION

1. **Source "correct" medication lists derived from Synthea[5] pre-generated "The Coherent Dataset."[6]**
2. **Frequency information was missing from many Synthea lists, so a PaLM 2[7] LLM prompt was used to generate it.**
3. **PaLM 2 used to split "correct" lists into 3 "incorrect" lists, each seeded with reconciliation errors.**
   - **Prompt specified:**
     - **I. Number and formatting of incorrectly reconciled lists to generate.**
     - **II. 3 complete examples of what introducing reconciliation errors might look like.**

- **Final "incorrect" lists contained: list's source organization, medication name, dosage, frequency and route, and whether the medication info is correct.**

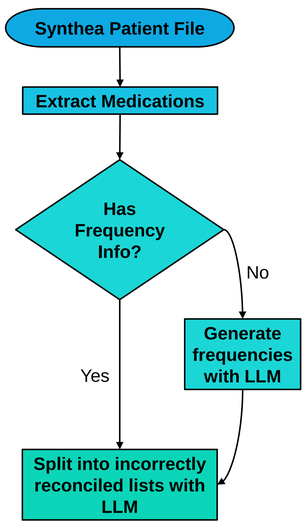

Figure 1. Generation of synthetic patient data with reconciliation errors, including correction for missing frequency information.

METHODS – MODEL TRAINING AND EVALUATION

Model Type	Specific Models Trained/Tested	Description
Machine Learning Classifier	Logistic Regression Random Forest XGBoost	Use mathematical optimization techniques to classify data into categories. Faster and more transparent than deep learning models, however they require manual tuning.
Deep Learning Model	ResNet50 VGG-16 FFNN	Like neurons in the human brain, these models consist of layers of interconnected "nodes," each with their own bias affecting how they weigh input and pass output to subsequent layers.
Modern LLM	BERT	One of the first LLMs, can classify text into categories or predict the most likely words to complete sentences.
	RoBERTa	"Robustly Optimized BERT Approach" – an optimized BERT.
	DistilBERT	Functions like BERT using a much smaller neural network.
	XLNet	Newer LLM that better understands relationships between words (e.g., recognizing "Yale New Haven Health" as a discrete concept instead of 4 loosely related ones).

LIMITATIONS

- **Homogenous Synthea source data:**
  - **Only ~100 unique medications throughout the Synthea dataset.**
    - **Many only in 1 specific dosage; none in the full range of possible dosages.**
  - **Data still far more heterogeneous than that generated using an LLM alone.**
- **Frequency information inconsistencies:**
  - **Many medications produced by Synthea lack frequency information.**
  - **Substitutions created using frequency-suggestion prompt were homogenous.**
- **Unrealistic reconciliation error distribution:**
  - **LLM prompt did not include information beyond the lists of medications themselves (e.g., diagnoses, transitions of care, etc.).**
  - **Less information made reconciliation errors less realistic.**
- **Challenging prompt engineering:**
  - **Quality of data produced by PaLM 2 was sensitive to small prompt changes.**
  - **Suggests inconsistent data quality across data set, and room for improvement.**
- **Implications for our models and results:**
  - **Generalizability limited due to models training and testing on unrealistic data.**
    - **Even if trained on real patient data, models require further fine-tuning.**

RESULTS AND DISCUSSION

- **Heterogeneous and unique generated data:**
  - **PaLM 2 consistently produced coherent reconciliation-error containing data.**
    - **Even with limited data access and only 3 examples provided.**
  - **PaLM 2 was also surprisingly adept at creating novel errors.**
    - **Some complex errors created entirely independently include: substituting similar sounding medications (e.g., ambien for amlodipine) and swapping diuretics.**
- **Impressive LLM performance:**
  - **Many models performed very well in flagging errors (Table 1).**
    - **Simplistic training data still limits conclusions.**
  - **LLMs outperformed older deep and shallow learning models (Table 1).**
    - **Suggests finding solutions to high training data requirements may be worthwhile.**
- **Implications:**
  - **Data quality suggests LLMs could help address the need for nuanced, yet HIPAA-compliant, patient data needed to train medical ML models.**
  - **Preliminary model efficacy at identifying Med-Wreck suggests AI may hold potential to aid in reducing these harmful, preventable errors.**
- **Future aims:**
  - **Training data improvements will help evaluate model performance in more realistic, heterogenous, and complex settings. Current goals include:**
    - **Replacing Synthea with real patient data.**
    - **Improving LLM data enhancement (better prompt engineering, providing more data about each patient, etc.).**

		Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
Large Language Models	BERT	90.85	86.55	94.70	78.15
	RoBERTa	88.54	84.79	92.84	76.86
	DistilBERT	87.12	82.56	89.95	78.16
	XLNet	88.45	85.21	92.94	77.02
Deep Learning Models	ResNet50	79.14	76.88	85.92	67.39
	VGG-16	77.92	73.97	83.49	66.48
	FFNN	78.05	74.76	83.72	67.20
Machine Learning Classifiers	Logistic Regression	69.82	66.42	74.16	61.45
	Random Forest	70.54	67.38	75.60	61.49
	XGBoost	69.16	66.84	74.97	60.14

Table 1. Comparative analysis of Large Language Models, Deep Learning Models, and Machine Learning Classifiers in identifying medication error. LLMs, especially BERT, have superior sensitivity, specificity, positive predictive value (PPV) and negative predictive value (PPV).

REFERENCES