USING LARGE LANGUAGE MODELS FOR ENTITY EXTRACTION IN THE HEALTHCARE SECTOR

PURDUE UNIVERSITY
Daniels School of Business

SOHAM AGARWAL
agarw402@purdue.edu

MATTHEW A. LANHAM
lanhamm@purdue.edu

Abstract

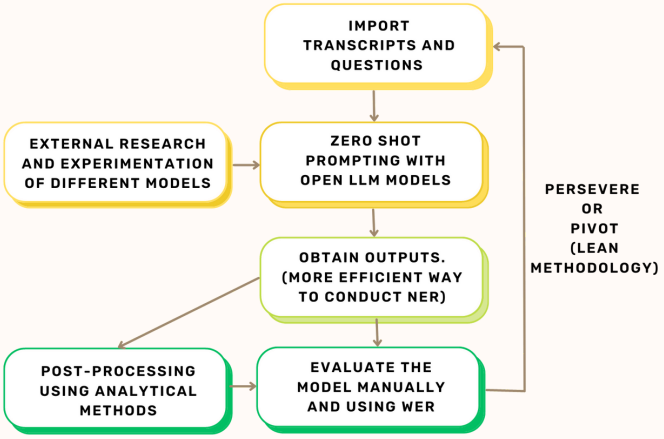
Utilizing AI in healthcare, we streamlined administrative tasks and upheld regulatory compliance, addressing physician shortages. Our study involved analyzing **2001 artificial transcripts** from a Kaggle competition to extract vital patient data. We employed various Large Language Models to optimize data processing, conducting A/B tests and adjusting variables to achieve the lowest possible Word Error Rate (WER).

Introduction

Healthcare is an industry with a dearth of physicians, **(2.47 physicians per 1000 people)**. The ability to automate decision processes can be facilitated by the use of open-source large Language models. Open-source models in healthcare streamline operations, ensure content accuracy, and enhance HIPAA compliance.

Our approach involved exploring various online models, from testing open LLM models on Amazon instances to using Hugging Face endpoints. Ultimately, we found that using APIs from companies offering cloud-based, open-source models was the most practical and cost-effective solution.

Abstract Decision Flow



Methodology and Evaluation

We explored various providers before choosing openrouter.ai for its efficient open-source model endpoints, which enabled direct model integration into our code. The development involved creating a code to query transcripts through their API, requiring multiple adjustments to the prompt template for optimal Named Entity Recognition (NER) outputs.

The post-processing phase was critical for achieving concise outputs. This included correcting errors, chaining models, and data cleaning, notably using regression for missing values, all contributing to a reduced WER. Factors like cost, API rate limits, and processing time for **12,006 rows** are crucial, particularly for companies with large-scale data analysis involving millions of records.

We benchmarked our process against Kaggle competition submissions, employing A/B testing for iterative improvements. Manual comparison with human transcription results provided valuable insights, guiding further code post-processing for accuracy in the expected outputs.

Key Insights and Results

Example outputs (right)

Question	Output
What is the patient's name?	Tina Will
What is the patient's age?	69
What is the patient's condition?	heart attack
What symptoms is the patient experiencing?	chest pain, vomiting, breathlessness
What precautions did the doctor advise?	None
What drug did the doctor prescribe?	None

Leader-board (down)

4	▲ 1	BAIMERS		0.53738

The **Mistral OpenOrca 7B** model significantly enhanced our results, aligning closely with our problem statements. Key improvements included capitalizing brand-named drugs, accurately imputing ages, and refining spacing and grammar in our outputs. These changes significantly reduced the Word Error Rate (WER) to **0.53738** on the test data, closely matching the ideal solution. Although other A/B tests **(around 20-25)** showed minimal impact, these focused modifications were crucial in achieving high accuracy. Consequently, our team secured **4th place among 95 teams** in a private Kaggle competition.

Open-sourced models allow for secure hosting on private servers but face limitations in language translation and API rates. Addressing these issues could pave the way for higher extraction accuracies.