

Large Language Models with Retrieval-Augmented Generation for Zero-Shot Disease Phenotyping

Will E. Thompson, MS; David M. Vidmar, PhD; Jessica K. De Freitas, PhD; Gabriel Altay, PhD; Kabir Manghnani, BS; Andrew C. Nelsen*, PharmD; Kellie Morland*, PharmD; John M. Pfeifer, MD, MPH; Brandon K. Fornwalt, MD, PhD; Ruijun Chen, MD; Martin C. Stumpe, PhD; Riccardo Miotto, PhD

Tempus Labs, Inc. Chicago, IL; *United Therapeutics Corporation, Silver Spring, MD | contact: will.thompson@tempus.com



TEMPUS

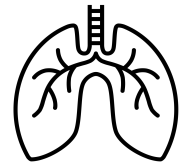
SUMMARY

- We designed a phenotyping approach centered around the prompting of large language models (LLMs) to identify patients with pulmonary hypertension (PH), a rare disease, from their clinical notes.
- To overcome the volume of clinical notes found within electronic health records (EHRs), our method harnesses both retrieval-augmented generation (RAG) and MapReduce to effectively analyze the complete patient documentation.
- We experimented with a number of different prompting techniques, including Chain of Thought (CoT) reasoning, as well as both an LLM and max function approach to aggregation. While there was significant variability in performance across choice of prompting design, we noticed higher mean performance and lower variance from the use of max function aggregation.
- Our method significantly outperforms physician logic rules (F1 score of 0.62 vs. 0.75) in identifying PH.

BACKGROUND

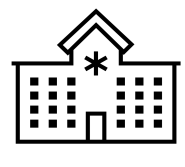
Identifying disease phenotypes from EHRs is critical for numerous secondary uses. The manual review of clinical notes is often time consuming, taking an average of 30 mins per patient file, and therefore not scalable. Further, encoding physician knowledge into rules is particularly challenging for rare diseases due to inadequate EHR coding. LLMs offer promise in text understanding but may not efficiently handle real-world clinical documentation. We investigated whether an LLM-based method enriched by RAG and MapReduce could provide a scalable alternative to physician logic rules.

EVALUATION DESIGN




Pulmonary Hypertension

PH is characterized by elevated pressure in the lungs and right side of the heart. It is broadly categorized as a rare disease with an estimated global prevalence rate of 1-3%.




Dataset

De-ID'd dataset from a medium-sized hospital system serving ~2.2 million people.



Large Language Model

[bison@001](#) is a version of PaLM available through Google Cloud's Vertex-AI offering.



Structured Phenotype Definition

One of our physicians reviewed patients within the training dataset to establish a rules-based algorithm for diagnosing PH using structured data from patient EHRs.

METHODS

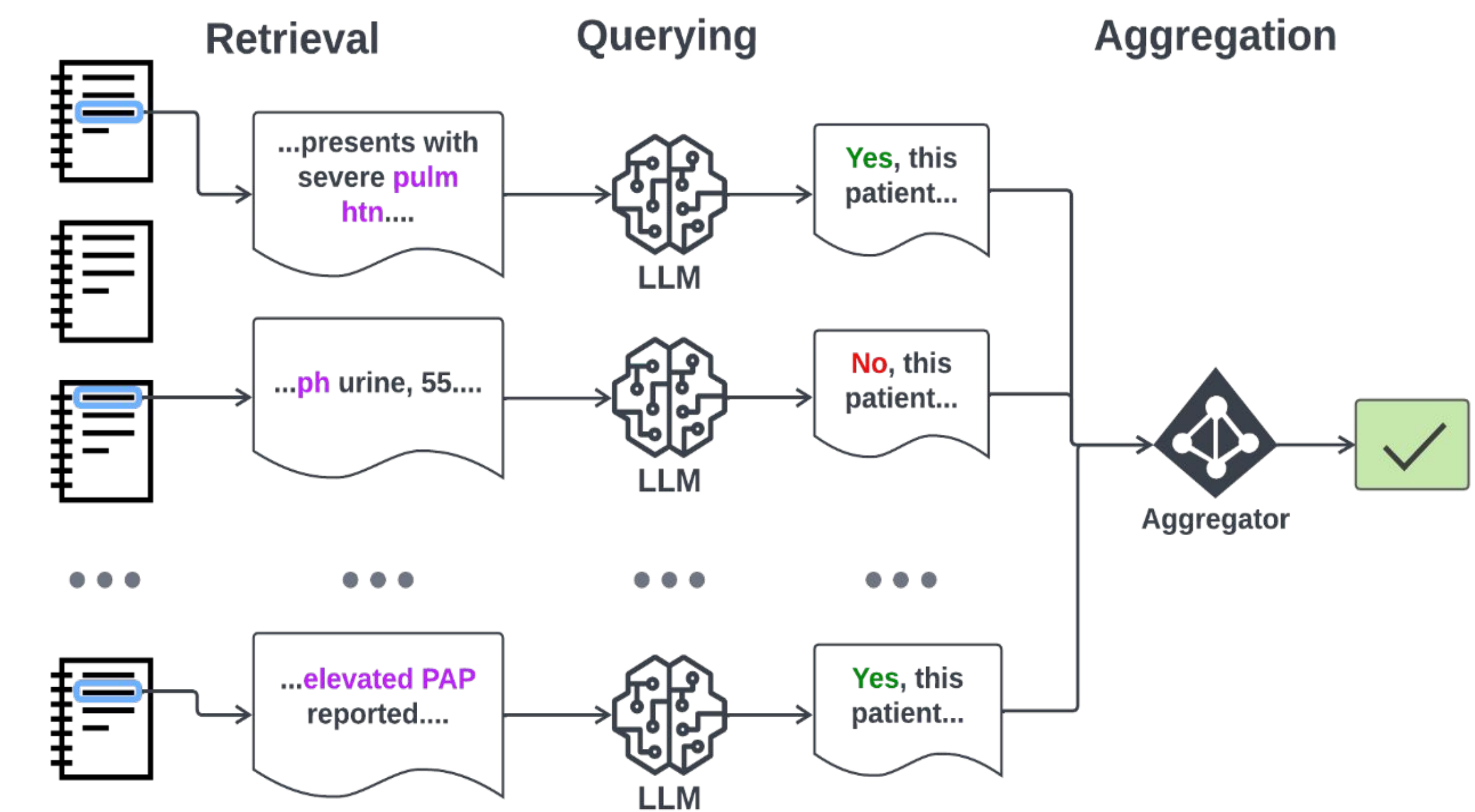


Figure 1. LLM Phenotyping Pipeline - The architecture consists of a Regex retriever followed by MapReduce for querying and aggregation.

RESULTS

Model	Aggregation	ECHO Exclusion	F1 Score
Structured	—	—	0.62
LLM	Max	Regex	0.73
LLM	Max	Prompt Amended	0.75
LLM	LLM	Prompt Amended	0.72

Table 1. Test Set Performance -Test set performance of three variants of the LLM-based phenotype compared to the structured phenotype.

Task:
Use the following pieces of context from a patient medical record to answer the question below. Provide your answer as one of {"Yes", "No", "Unsure"}.
Support your reasoning with evidence.

Question:
Does this patient have pulmonary hypertension?

Count a 'possible' case as a 'no'. Count a history of pulmonary hypertension as a 'yes'.

Context:
Patient 69 yo male came in due to concerns over shortness of breath ... h/o pulm htn ... recommended follow up in 1 month ...

Answer:
Let's think step-by-step.

Response:
Yes, the patient has a history of pulmonary hypertension.
This is evident from the line "h/o pulm htn" in the medical history. Final answer: yes.###

Multiple Choice

Steering

Chain of Thought (CoT)

Figure 2. Prompt Design - An illustrative example of a zero-shot prompt.

Prompt Design	Steering	CoT	Multiple Choice
A	✓	✓	✓
B	✓	✓	
C	✓*	✓	
D	✓		✓
E	✓		

Table 2. Prompt Variations - Descriptions of various prompting approaches explored during the evaluation. (*) prompt C was modified to remove the phrase "explain your reasoning".

ACKNOWLEDGMENTS

We thank Dana DeSantis for visualization and poster review.

